CENTER SIGNAL SCALING USING SIGNAL-TO-DOWNMIX RATIOS

Christian Uhle

Fraunhofer Institute for Integrated Circuits
Erlangen, Germany
christian.uhle@iis.fraunhofer.de

ABSTRACT

A novel method for scaling the level of the virtual center in audio signals is proposed. The input signals are processed in the time-frequency domain such that direct sound components having approximately equal energy in all channels are amplified or attenuated. The real-valued spectral weights are obtained from the ratio of the sum of the power spectral densities of all input channel signals and the power spectral density of the sum signal. Applications of the presented method are upmixing two-channel stereophonic recordings for its reproduction using surround sound set-ups, stereophonic enhancement, dialogue enhancement, and as preprocessing for semantic audio analysis.

1. INTRODUCTION

Audio signals are in general a mixture of direct sounds and ambient (or diffuse) sounds. Direct signals are emitted by sound sources, e.g. a musical instrument, a vocalist or a loudspeaker, and arrive on the shortest possible path at the receiver, e.g. the listener's ear or a microphone. When listening to a direct sound, it is perceived as coming from the direction of the sound source. The relevant auditory cues for the localization and for other spatial sound properties are interaural level difference (ILD), interaural time difference (ITD) and interaural coherence. Direct sound waves evoking identical ILD and ITD are perceived as coming from the same direction. In the absence of ambient sound, the signals reaching the left and the right ear or any other set of spaced sensors are coherent.

Ambient sounds, in contrast, are emitted by many spaced sound sources or sound reflecting boundaries contributing to the same sound. When a sound wave reaches a wall in a room, a portion of it is reflected, and the superposition of all reflections in a room, the reverberation, is a prominent example for ambient sounds. Other examples are applause, babble noise and wind noise. Ambient sounds are perceived as being diffuse, not locatable, and evoke an impression of envelopment (of being "immersed in sound") by the listener. When capturing an ambient sound field using a set of spaced sensors, the recorded signals are at least partially incoherent.

This paper discusses center signal scaling, i.e. the amplification or attenuation of center signals in audio recordings. The center signal is defined here as the sum of all direct signal components having approximately equal intensity in all channels and negligible time differences between the channels.

1.1. Applications

Various applications of audio signal processing and reproduction benefit from center signal scaling, e.g. upmixing, dialogue enhancement, and semantic audio analysis.

Upmixing refers to the process of creating an output signal given an input signal with less channels. Its main application is the reproduction of two-channel signals using surround sound setups as for example specified in [1]. Research on the subjective quality of spatial audio [2] indicates that locatedness [3], localization and width are prominent descriptive attributes of sound. Results of a subjective assessment of 2-to-5 upmixing algorithms [4] showed that the use of an additional center loudspeaker can narrow the stereophonic image. The presented work is motivated by the assumption that locatedness, localization and width can be preserved or even improved when the additional center loudspeaker reproduces mainly *direct* signal components which are panned to the center, and when these signal components are attenuated in the off-center loudspeaker signals.

Dialogue enhancement refers to the improvement of speech intelligibility, e.g. in broadcast and movie sound, and is often desired when background sounds are too loud relative to the dialogue [5]. This applies in particular to persons who are hard of hearing, non-native listeners, in noisy environments or when the binaural masking level difference is reduced due to narrow loudspeaker placement. The proposed method can be applied for processing input signals where the dialogue is panned to the center in order to attenuate background sounds and thereby enabling better speech intelligibility.

Semantic Audio Analysis (or Audio Content Analysis) comprises processes for deducing meaningful descriptors from audio signals, e.g. beat tracking or transcription of the leading melody. The performance of the computational methods is often deteriorated when the sounds of interest are embedded in background sounds, see e.g. [6]. Since it is common practice in audio production that sound sources of interest (e.g. leading instruments and singers) are panned to the center, center extraction can be applied as a pre-processing step for attenuating background sounds and reverberation.

1.2. Prior work

Related prior work on separation, decomposition or scaling is either based on panning information, i.e. inter-channel level differences (ICLD) and inter-channel time differences (ICTD), or based on signal characteristics of direct and of ambient sounds.

Methods taking advantage of ICLD in two-channel stereophonic recordings are the upmix method described in [7], the Azimuth Discrimination and Resynthesis (ADRess) algorithm [8], the upmix from two-channel input signals to three channels proposed by Vickers [9], and the center signal extraction described in [10].

The Degenerate Unmixing Estimation Technique (DUET) [11, 12] is based on clustering the time-frequency bins into sets with similar ICLD and ICTD. A restriction of the original method is that the maximum frequency which can be processed equals half the speed of sound over maximum microphone spacing (due to ambiguities in the ICTD estimation) which has been addressed in [13]. The performance of the method decreases when sources overlap in the time-frequency domain and when the reverberation increases. Other methods based on ICLD and ICTD are the Modified ADRess algorithm [14], which extends ADRess algorithm [8] for the processing of spaced microphone recordings, the method based on time-frequency correlation (AD-TIFCORR) [15] for time-delayed mixtures, the Direction Estimation of Mixing Matrix (DEMIX) for anechoic mixtures [16], which includes a confidence measure that only one source is active at a particular time-frequency bin, the Model-based Expectation-Maximization Source Separation and Localization (MESSL) [17], and methods mimicking the binaural human hearing mechanism as in e.g. [18, 19].

Despite the methods for Blind Source Separation (BSS) using spatial cues of direct signal components mentioned above, also the extraction and attenuation of ambient signals are related to the presented method. Methods based on the inter-channel coherence (ICC) in two-channel signals are described in [20, 7, 21]. The application of adaptive filtering has been proposed in [22], with the rationale that direct signals can be predicted across channels whereas diffuse sounds are obtained from the prediction error.

A method for upmixing of two-channel stereophonic signals based on multichannel Wiener filtering estimates both, the ICLD of direct sounds and the power spectral densities (PSD) of the direct and ambient signal components [23].

Approaches to the extraction of ambient signals from single channel recordings include the use of Non-Negative Matrix Factorization of a time-frequency representation of the input signal, where the ambient signal is obtained from the residual of that approximation [24], low-level feature extraction and supervised learning [25], and the estimation of the impulse response of a reverberant system and inverse filtering in the frequency domain [26].

1.3. Contribution of this work

The contribution of this work is a novel method for amplifying or attenuating the center signal in an audio signal. In contrast to previous methods, it considers both, lateral displacement and diffuseness of the signal components. Furthermore, the use of semantically meaningful parameters is discussed in order to support the user when operating an implementation of the method.

2. PROPOSED METHOD

The description of the proposed method is structured as follows: Section 2.1 describes the underlying signal model and the method and analyzes it for the case of input signal featuring amplitude difference stereophony. Section 2.2 discusses the more general case of mixing models featuring time-of-arrival stereophony. Section 2.3 gives intuitive explanations of the control parameters. Computational complexity and memory requirements are briefly discussed in Section 2.4.

2.1. Amplitude difference stereophony

The rationale is to compute and apply real-valued spectral weights as a function of the diffuseness and the lateral position of direct sources. The processing as demonstrated here is applied in the STFT domain, yet it is not restricted to a particular filterbank.

The Q channel input signal is denoted by $\mathbf{x}[n]$ with

$$\mathbf{x}[n] = [x_1[n] \cdots x_Q[n]]^T \; . \tag{1}$$

where n denotes the discrete time index. The input signal is assumed to be an additive mixture of direct signals $s_i[n]$ and ambient sounds $a_i[n]$,

$$x_l[n] = \sum_{i=1}^{P} d_{i,l}[n] * s_i[n] + a_l[n] \; , l = 1, ..., Q \tag{2}$$

where P is the number of sound sources, $d_{i,l}[n]$ denote the impulse responses of the direct paths of the i-th source into the l-th channel of length $L_{i,l}$ samples, and the ambient signal components are mutually uncorrelated or weakly correlated. In the following description it is assumed that the signal model corresponds to amplitude difference stereophony, i.e. $L_{i,l} = 1, \forall i, l$.

The time-frequency domain representation of $\mathbf{x}[n]$ is given by

$$\mathbf{X}(m,k) = [X_1(m,k) \cdots X_Q(m,k)]^T \; , \tag{3}$$

with time index m and frequency index k. The output signals are denoted by

$$\mathbf{Y}(m,k) = [Y_1(m,k) \cdots Y_Q(m,k)]^T \; , \tag{4}$$

and are obtained by means of spectral weighting

$$\mathbf{Y}(m,k) = G(m,k)\mathbf{X}(m,k) \; , \tag{5}$$

with real-valued weights $G(m,k)$. Time domain output signals are computed by applying the inverse processing of the filterbank.

For the computation of the spectral weights, the sum signal, thereafter denoted as the downmix signal, is computed as

$$X_{\mathrm{d}}(m,k) = \sum_{i=1}^{Q} X_i(m,k) \; . \tag{6}$$

The matrix of PSD of the input signal, containing estimates of the (auto-)PSD on the main diagonal, while off-diagonal elements are estimates of the cross-PSD, is given by

$$\Phi_{i,l}(m,k) = \mathcal{E}\{X_i(m,k)X_l^*(m,k)\} \; , \; i,l = 1...Q \; , \tag{7}$$

where X^* denotes the complex conjugate of X, and $\mathcal{E}\{\cdot\}$ is the expectation operation with respect to the time dimension. In the presented simulations the expectation values are estimated using single-pole recursive averaging,

$$\Phi_{i,l}(m,k) = \alpha X_i(m,k)X_l^*(m,k) + (1-\alpha)\Phi_{i,l}(m-1,k) \; , \tag{8}$$

where the filter coefficient α determines the integration time. Furthermore, we define the quantity $R(m,k;\beta)$ as

$$R(m,k;\beta) = \left(\frac{\sum_{i=1}^{Q} \Phi_{i,i}(m,k)^\beta}{\Phi_{\mathrm{d}}(m,k)^\beta} \right)^{\frac{1}{2\beta-1}} \; , \tag{9}$$

where $\Phi_{\mathrm{d}}(m,k)$ is the PSD of the downmix signal and β is a parameter which will be addressed in the following.

The quantity $R(m,k;1)$ is the signal-to-downmix ratio (SDR), i.e. the ratio of the total PSD and the PSD of the downmix signal. The power to $\frac{1}{2\beta-1}$ ensures that the range of $R(m,k;\beta)$ is independent of β.

Figure 1 illustrates the SDR for $Q=2$ as a function of ICC $\Psi(m,k)$ and ICLD $\Theta(m,k)$, with

$$\Psi(m,k) = \frac{|\Phi_{1,2}(m,k)|}{\sqrt{\Phi_{1,1}(m,k)\Phi_{2,2}(m,k)}} \ , \tag{10}$$

and

$$\Theta(m,k) = \frac{\Phi_{1,1}(m,k)}{\Phi_{2,2}(m,k)} \ . \tag{11}$$

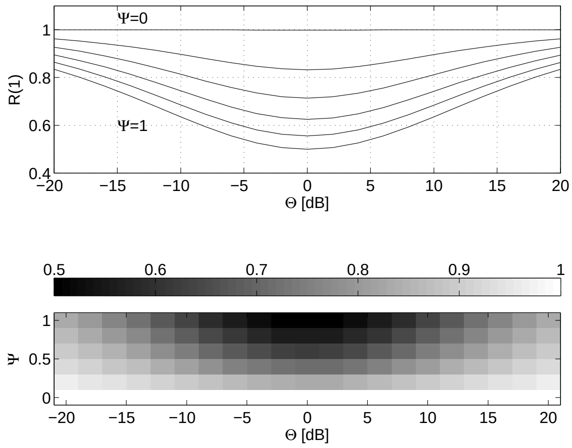


Figure 1: *Upper plot: SDR $R(m,k;1)$ for Q=2 as function of the ICLD $\Theta(m,k)$, shown for $\Psi(m,k) \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. Lower plot: SDR $R(m,k;1)$ for Q=2 as function of ICC $\Psi(m,k)$ and ICLD $\Theta(m,k)$ in color-coded 2D-plot.*

It shows that the SDR has the following properties:

1. It is monotonically related to both, $\Psi(m,k)$ and $|\log\Theta(m,k)|$.
2. For diffuse input signals, i.e. $\Psi(m,k)$=0, the SDR assumes its maximum value, $R(m,k;1)=1$.
3. For direct sounds panned to the center, i.e. $\Theta(m,k)=1$, the SDR assumes its minimum value R_{min}, where $R_{\text{min}} = 0.5$ for $Q=2$.

Due to these properties, appropriate spectral weights for center signal scaling can be computed from the SDR by using monotonically *decreasing* functions for the *extraction* of center signals and monotonically *increasing* functions for the *attenuation* of center signals.

For the extraction of a center signal, appropriate functions of $R(m,k;\beta)$ are, for example,

$$G_{c_1}(m,k;\beta,\gamma) = (1 + R_{\text{min}} - R(m,k;\beta))^{\gamma} \ , \tag{12}$$

and

$$G_{c_2}(m,k;\beta,\gamma) = \left(\frac{R_{\text{min}}}{R(m,k;\beta)}\right)^{\gamma} \ , \tag{13}$$

where a parameter γ for controlling the maximum attenuation is introduced.

For the attenuation of the center signal, appropriate functions of $R(m,k;\beta)$ are, for example,

$$G_{s_1}(m,k;\beta,\gamma) = R(m,k;\beta)^{\gamma} \ , \tag{14}$$

and

$$G_{s_2}(m,k;\beta,\gamma) = \left(1 + R_{\text{min}} - \frac{R_{\text{min}}}{R(m,k;\beta)}\right)^{\gamma} \ , \tag{15}$$

Figures 2 and 3 illustrate the gain functions (13) and (15), respectively, for $\beta=1, \gamma=3$. The spectral weights are constant for $\Psi(m,k)=0$. The maximum attenuation is $\gamma \cdot 6$dB, which also applies to the gain functions (12) and (14).

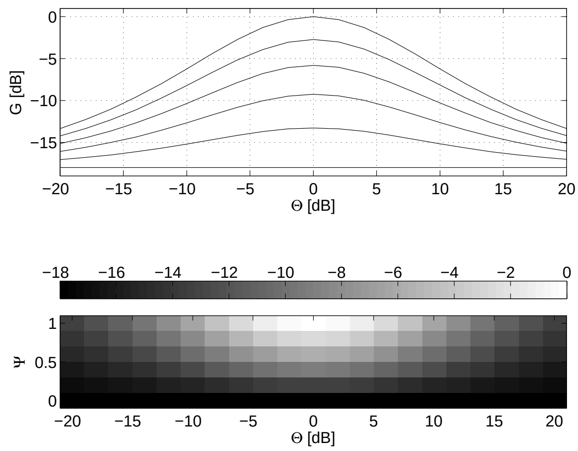


Figure 2: *Spectral weights $G_{c_2}(m,k;1,3)$ in dB as function of ICC $\Psi(m,k)$ and ICLD $\Theta(m,k)$.*

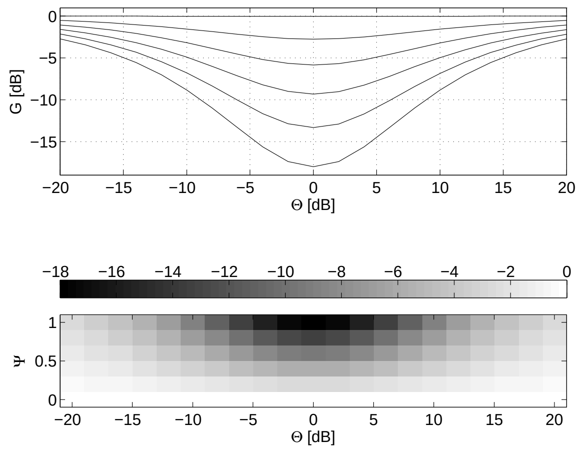


Figure 3: *Spectral weights $G_{s_2}(m,k;1,3)$ in dB as function of ICC $\Psi(m,k)$ and ICLD $\Theta(m,k)$.*

The effect of the parameter β is shown in Figure 4 for the gain function in Equation (13) with $\beta=2, \gamma=3$. With larger values for β, the influence of Ψ on the spectral weights decreases whereas the influence of Θ increases. This leads to more leakage of diffuse signal components into the output signal, and to more attenuation

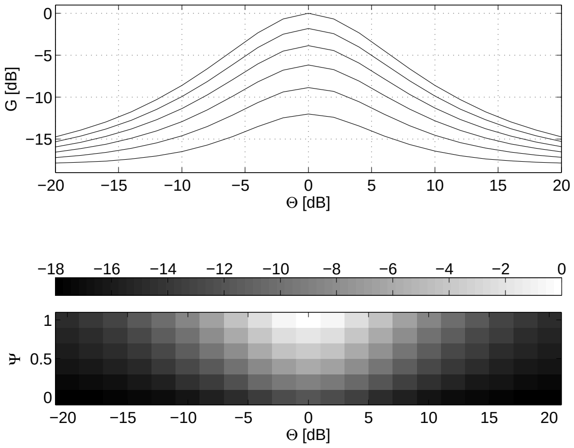


Figure 4: *Spectral weights $G_{c_2}(m, k; 2, 3)$ in dB as function of ICC $\Psi(m, k)$ and ICLD $\Theta(m, k)$.*

of the direct signal components panned off-center, when comparing to the gain function in Figure 2.

Post-procesing of spectral weights: Prior to the spectral weighting, the weights $G(m, k; \beta, \gamma)$ can be further processed by means of smoothing operations. Zero phase low-pass filtering along the frequency axis reduces circular convolution artifacts which can occur for example when the zero-padding in the STFT computation is too short or a rectangular synthesis window is applied. Low-pass filtering along the time axis can reduce processing artifacts, especially when the time constant for the PSD estimation is rather small.

2.2. Time-of-arrival stereophony

The derivation of the spectral weights described above relies on the assumption that $L_{i,l} = 1, \forall i, l$, i.e. the direct sound sources are time-aligned between the input channels. When the mixing of the direct source signals is not restricted to amplitude difference stereophony ($L_{i,l} > 1$), for example when recording with spaced microphones, the downmix of the input signal $X_\text{d}(m, k)$ is subject to phase cancellation. Phase cancellation in $X_\text{d}(m, k)$ leads to increasing SDR values and consequently to the typical comb-filtering artifacts when applying the spectral weighting as described above.

The notches of the comb-filter correspond to the frequencies $f_n = \frac{o f_s}{2d}$ for gain functions (12) and (13) and $f_n = \frac{e f_s}{2d}$ for gain functions (14) and (15), where f_s is the sampling frequency, o are odd integers, e are even integers, and d is the delay in samples.

A first approach to solve this problem is to compensate the phase differences resulting from the ICTD prior to the computation of $X_\text{d}(m, k)$. Phase difference compensation (PDC) is achieved by estimating the time-variant inter-channel phase transfer function $\hat{P}_i(m, k) \in [-\pi\ \pi]$ between the ith channel and a reference channel denoted by index r,

$$\hat{P}_i(m, k) = \arg X_\text{r}(m, k) - \arg X_i(m, k)\ , i \in [1, ..., Q] \setminus \text{r}, \tag{16}$$

where the operator $A \setminus B$ denotes set-theoretic difference of set B and set A, and applying a time-variant allpass compensation filter $H_{C,i}(m, k)$ to the ith channel signal

$$\tilde{X}_i(m, k) = H_{C,i}(m, k) X_i(m, k)\ . \tag{17}$$

where the phase transfer function of $H_{C,i}(m, k)$ is

$$\arg H_{C,i}(m, k) = -\mathcal{E}\{\hat{P}_i(m, k)\}\ . \tag{18}$$

The expectation value is estimated using single-pole recursive averaging. It should be noted that phase jumps of 2π occurring at frequencies close to the notch frequencies need to be compensated for prior to the recursive averaging.

The downmix signal is computed according to

$$X_\text{d}(m, k) = \sum_{i=1}^{Q} \tilde{X}_i(m, k)\ . \tag{19}$$

such that the PDC is only applied for computing X_d and does not affect the phase of the output signal.

2.3. Semantic meaning of control parameters

For the operation of digital audio effects it is advantageous to provide controls with semantically meaningful parameters. The gain functions (12) - (15) are controlled by the parameters α, β and γ. Sound engineers and audio engineers are used to time constants, and specifying α as time constant is intuitive and according to common practice. The effect of the integration time can be experienced best by experimentation. In order to support the operation of the method, descriptors for the remaining parameters are proposed, namely *impact* for γ and *diffuseness* for β.

The parameter *impact* can be best compared with the order of a filter. By analogy to the roll-off in filtering, the maximum attenuation equals $\gamma \cdot 6$dB, for $Q = 2$.

The label *diffuseness* is proposed here to emphasize the fact that then attenuating panned and diffuse sounds, larger values of β result in more leakage of diffuse sounds. A nonlinear mapping of the user parameter β_u, e.g. $\beta = \sqrt{\beta_u + 1}$, with $0 \leq \beta_u \leq 10$, is advantageous in a way that it enables a more consistent behavior of the processing as opposed to when modifying β directly (where *consistency* relates to the effect of a change of the parameter on the result throughout the range of the parameter value).

2.4. On computational complexity and memory requirements

The computational complexity and memory requirements scale with the number of bands of the filterbank and depend on the implementation of additional post-processing of the spectral weights.

A low-cost implementation of the method can be achieved when setting $\beta = 1$, $\gamma \in \mathbb{N}$, computing spectral weights according to Equation (12) or (14), and when not applying the PDC filter.

The computation of the SDR uses only one cost intensive nonlinear functions per sub-band when $\beta \in \mathbb{N}$. For $\beta = 1$, only two buffers for the PSD estimation are required, whereas methods making explicit use of the ICC, e.g. [7, 10, 20, 21, 23], require at least three buffers.

3. RESULTS AND DISCUSSION

This section illustrates the performance of the presented method by means of examples. First, the processing is applied to an amplitude-panned mixture of 5 instrument recordings (drums, bass, keys, 2 guitars) sampled at 44100 Hz of which an excerpt of 3 seconds length is visualized. Drums, bass and keys are panned to the center, one guitar is panned to the left channel and the second guitar is panned to the right channel, both with $|\text{ICLD}| = 20\text{dB}$. A convolution reverb having stereo impulse responses with an RT_{60} of about 1.4 seconds per input channel is used to generate ambient signal components. The reverberated signal is added with a direct-to-ambient ratio of about 8 dB after K-weighting [27]. Figure 5 shows spectrograms the direct source signals and the left and right channel signals of the mixture signal. The spectrograms are computed using an STFT with a length of 2048 samples, 50 % overlap, a frame size of 1024 samples and a sine window. Please note that for the sake of clarity only the magnitudes of the spectral coefficients corresponding to frequencies up to 4 kHz are displayed.

Figures 6 shows the input signal and the output signal for the center signal extraction obtained by applying $G_{c_2}(m, k; 1, 3)$. The time constant for the recursive averaging in the PSD estimation here and in the following is set to 200 ms. Figure 7 illustrates the spectrograms of the output signal. Visual inspection reveals that the source signals panned off-center (shown in Figure 5b and 5c) are largely attenuated in the output spectrograms. The output spectrograms also show that the ambient signal components are attenuated.

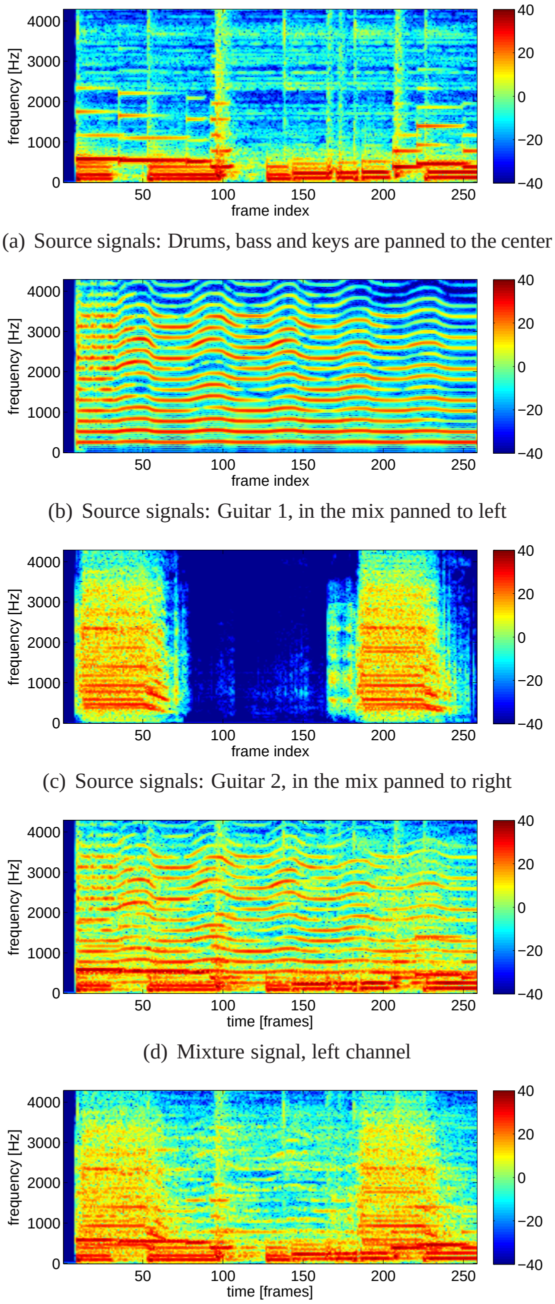
Figures 8 shows the input signal and the output signal for the center signal attenuation obtained by applying $G_{s_2}(m, k; 1, 3)$. The time signals illustrate that the transient sounds from the drums are attenuated by the processing. Figure 9 illustrates the spectrograms of the output signal. It can be observed that the signals panned to the center are attenuated, for example when looking at the transient sound components and the sustained tones in the lower frequency range below 600Hz and comparing to Figure 5a. The prominent sounds in the output signal correspond to the off-center panned instruments and the reverberation.

Informal listening over headphones reveals that the attenuation of the signal components is effective. When listening to the extracted center signal, processing artifacts become audible as slight modulations during the notes of guitar 2, similar to pumping in dynamic range compression. It can be noted that the reverberation is reduced and that the attenuation is more effective for low frequencies than for high frequencies. Whether this is caused by the larger direct-to-ambient ratio in the lower frequencies, the frequency content of the sound sources or subjective perception due to unmasking phenomena can not be answered without a more detailed analysis.

When listening to the output signal where the center is attenuated, the overall sound quality is slightly better when compared to the center extraction result. Processing artifacts are audible as slight movements of the panned sources towards the center when dominant centered sources are active, equivalently to the pumping when extracting the center. The output signal sounds less direct as the result of the increased amount of ambience in the output signal.

To illustrate the PDC filtering, Figure 10 shows two speech signals which has been mixed to obtain input signals with and without ICTD. The two-channel mixture signal is generated by mixing the speech source signals with equal gains to each channel and by adding white noise with an SNR of 10 dB (K-weighted) to the signal.

Figure 11 shows the spectral weights computed from gain function (13). The spectral weights in the upper plot are close to 0 dB when speech is active and assume the minimum value in time-frequency regions with low SNR. The second plot shows the spectral weights for an input signal where the first speech signal (Figure 10a) is mixed with an ICTD of 26 samples. The comb-filter characteristics is illustrated in Figure 11b. Figure 11c shows the spectral weights when PDC is enabled. The comb-filtering artifacts are largely reduced, although the compensation is not perfect near the notch frequencies at 848Hz and 2544Hz.



(a) Source signals: Drums, bass and keys are panned to the center

(b) Source signals: Guitar 1, in the mix panned to left

(c) Source signals: Guitar 2, in the mix panned to right

(d) Mixture signal, left channel

(e) Mixture signal, right channel

Figure 5: *Input signals for the music example. (a) Source signals panned to center, (b) source signal panned to left, (c) source signal panned to right, (d) left channel input signal, (e) right channel input signal.*

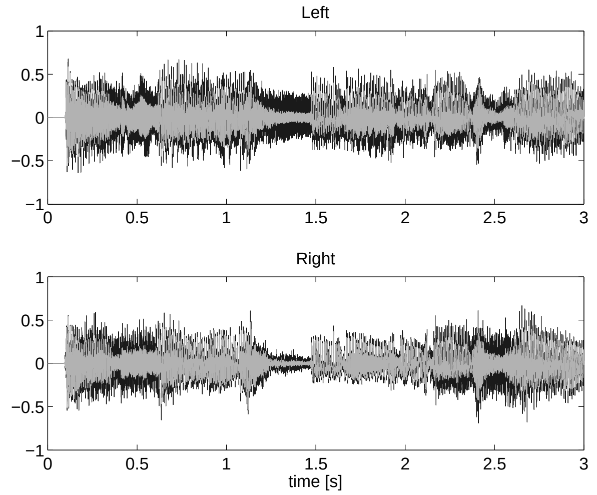


Figure 6: *Example for center extraction: Input time signals (black) and output time signals (overlaid in gray). Upper plot: left channel, lower plot: right channel.*

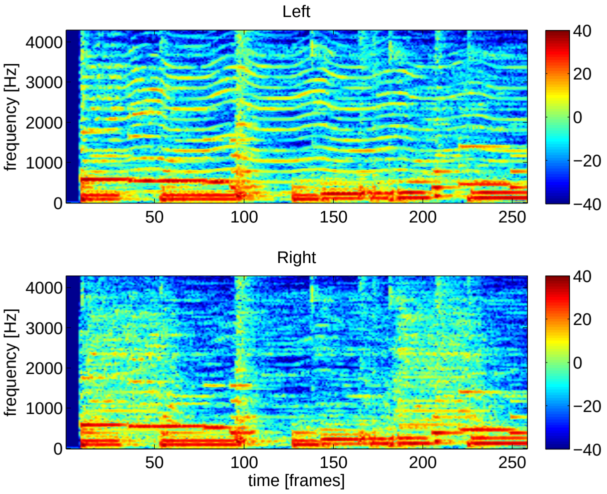


Figure 7: *Example for center extraction: Spectrograms of the output signals.*

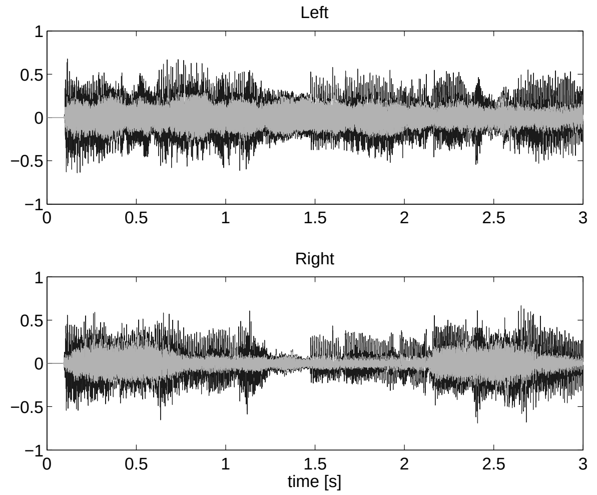


Figure 8: *Example for center attenuation: Input time signals (black) and output time signals (overlaid in gray).*

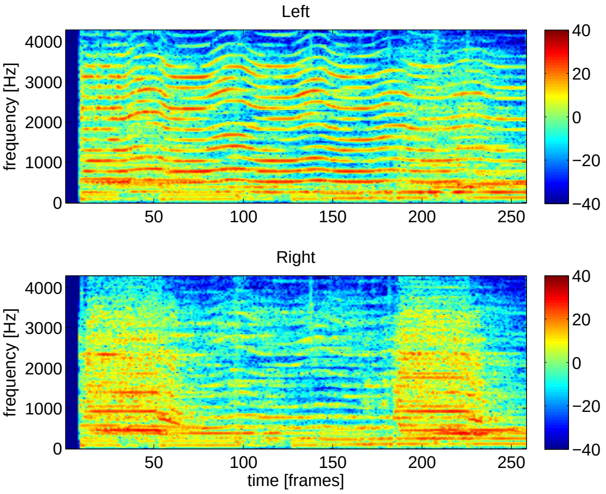
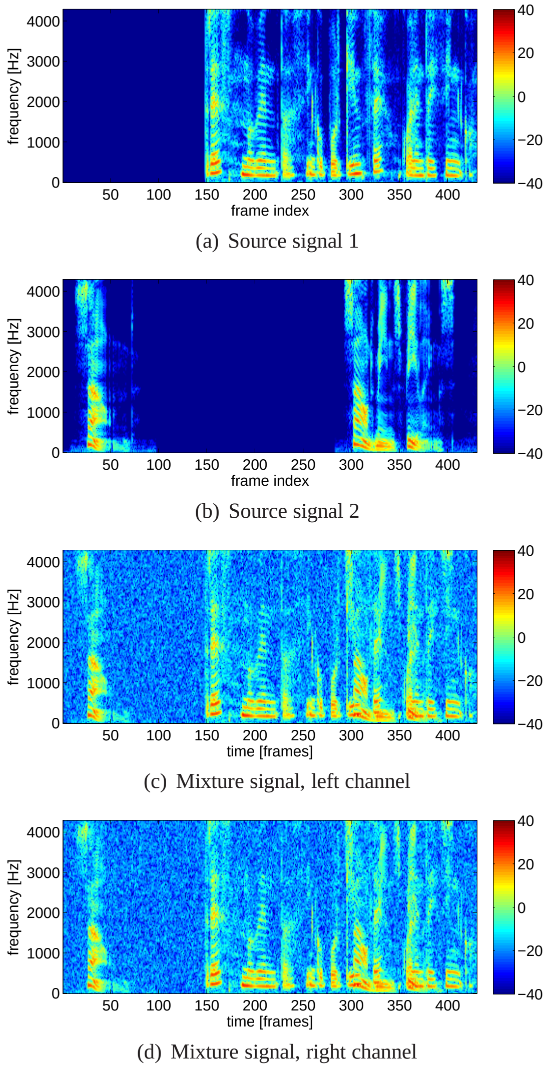


Figure 9: *Example for center attenuation: Spectrograms of the output signals.*

Informal listening shows that the additive noise is largely attenuated. When processing signals without ICTD, the output signals have a bit of an ambient sound characteristic which results presumably from the phase incoherence introduced by the additive noise. When processing signals with ICTD, the first speech signal (Figure 10a) is largely attenuated and strong comb-filtering artifacts are audible when not applying the PDC filtering. With additional PDC filtering, the comb-filtering artifacts are still slightly audible, but much less annoying.

Informal listening to other material reveals light artifacts, which can be reduced either by decreasing γ, by increasing β, or by adding a scaled version of the unprocessed input signal to the output. In general, artifacts are less audible when attenuating the center signal and more audible when extracting the center signal. Distortions of the perceived spatial image are very small. This can be attributed to the fact that the spectral weights are identical for all channel signals and do not affect the ICLDs. The comb-filtering artifacts are hardly audible when processing natural recordings featuring time-of-arrival stereophony for whom a mono downmix is not subject to strong audible comb-filtering artifacts. For the PDC filtering it can be noted that small values of the time constant of the recursive averaging (in particular the instantaneous compensation of phase differences when computing X_{d}) introduces coherence in the signals used for the downmix. Consequently, the processing is agnostic with respect to the diffuseness of the input signal. When the time constant is increased, it can be observed that (1) the effect of the PDC for input signals with amplitude difference stereophony decreases and (2) the comb-filtering effect becomes more audible at note onsets when the direct sound sources are not time-aligned between the input channels.



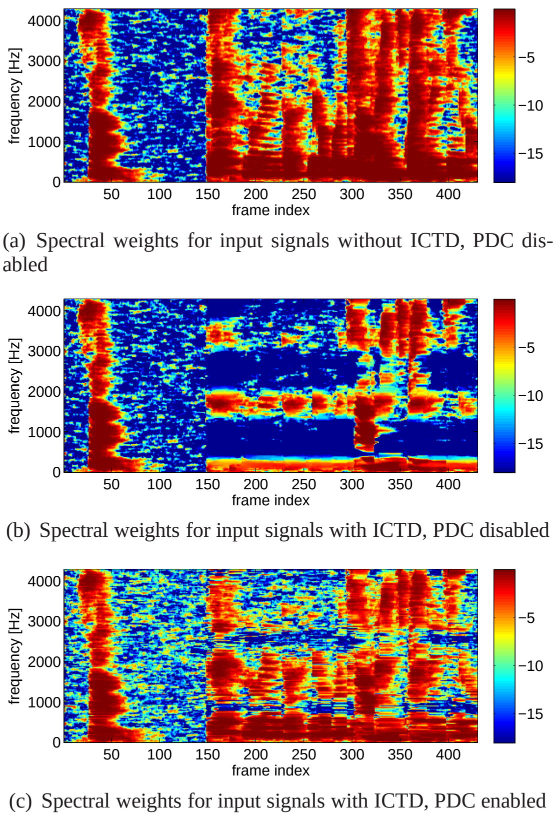
(a) Source signal 1

(b) Source signal 2

(c) Mixture signal, left channel

(d) Mixture signal, right channel

Figure 10: *Input source signals for illustrating the PDC. From top to bottom: First source signal, second source signal, left channel input signal, right channel input signal.*



(a) Spectral weights for input signals without ICTD, PDC disabled

(b) Spectral weights for input signals with ICTD, PDC disabled

(c) Spectral weights for input signals with ICTD, PDC enabled

Figure 11: *Spectral weights* $G_{c_2}(m, k; 1, 3)$ *for demonstrating the PDC filtering.*

4. CONCLUSION

A method has been presented for scaling the center signal in audio recordings by applying real-valued spectral weights which are computed from monotonic functions of the SDR. The rationale is that center signal scaling needs to take into account both, the lateral displacement of direct sources and the amount of diffuseness, and that these characteristics are implicitly captured by the SDR.

The processing can be controlled by semantically meaningful user parameters and is in comparison to other frequency domain techniques of low computational complexity and memory load.

The proposed method gives good results when processing input signals featuring amplitude difference stereophony, but can be subject to comb-filtering artifacts when the direct sound sources are not time-aligned between the input channels. A first approach to solve this is to compensate for non-zero phase in the inter-channel transfer function.

So far we have tested the method by means of informal listening. For typical commercial recordings, the results are of good sound quality but also depend on the desired separation strength.

5. REFERENCES

[1] International Telecommunication Union, Radiocomunication Assembly, “Multichannel stereophonic sound system with and without accompanying picture.,” Recommendation ITU-R BS.775-2, 2006, Geneva, Switzerland.

[2] J. Berg and F. Rumsey, “Identification of quality attributes of spatial sound by repertory grid technique,” *J. Audio Eng. Soc.*, vol. 54, pp. 365–379, 2006.

[3] J. Blauert, *Spatial Hearing*, MIT Press, 1996.

[4] F. Rumsey, “Controlled subjective assessment of two-to-five channel surround sound processing algorithms,” *J. Audio Eng. Soc.*, vol. 47, pp. 563–582, 1999.

[5] H. Fuchs, S. Tuff, and C. Bustad, “Dialogue enhancement - technology and experiments,” *EBU Technical Review*, vol. Q2, pp. 1–11, 2012.

[6] J.-H. Bach, J. Anemüller, and B. Kollmeier, “Robust speech detection in real acoustic backgrounds with perceptually motivated features,” *Speech Communication*, vol. 53, pp. 690–706, 2011.

[7] C. Avendano and J.-M. Jot, “A frequency-domain approach to multi-channel upmix,” *J. Audio Eng. Soc.*, vol. 52, 2004.

[8] D. Barry, B. Lawlor, and E. Coyle, “Sound source separation: Azimuth discrimination and resynthesis,” in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, 2004.

[9] E. Vickers, “Two-to-three channel upmix for center channel derivation and speech enhancement,” in *Proc. Audio Eng. Soc. 127th Conv.*, 2009.

[10] D. Jang, J. Hong, H. Jung, and K. Kang, “Center channel separation based on spatial analysis,” in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, 2008.

[11] A. Jourjine, S. Rickard, and O. Yilmaz, “Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures,” in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2000.

[12] O. Yilmaz and S. Rickard, “Blind separation of speech mixtures via time-frequency masking,” *IEEE Trans. on Signal Proc.*, vol. 52, pp. 1830–1847, 2004.

[13] S. Rickard, “The DUET blind source separation algorithm,” in *Blind Speech Separation*, S: Makino, T.-W. Lee, and H. Sawada, Eds. Springer, 2007.

[14] N. Cahill, R. Cooney, K. Humphreys, and R. Lawlor, “Speech source enhancement using a modified ADRess algorithm for applications in mobile communications,” in *Proc. Audio Eng. Soc. 121st Conv.*, 2006.

[15] M. Puigt and Y. Deville, “A time-frequency correlation-based blind source separation method for time-delay mixtures,” in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2006.

[16] Simon Arberet, Remi Gribonval, and Frederic Bimbot, “A robust method to count and locate audio sources in a stereophonic linear anechoic micxture,” in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2007.

[17] M.I. Mandel, R.J. Weiss, and D.P.W. Ellis, “Model-based expectation-maximization source separation and localization,” *IEEE Trans. on Audio, Speech and Language Proc.*, vol. 18, pp. 382–394, 2010.

[18] H. Viste and G. Evangelista, “On the use of spatial cues to improve binaural source separation,” in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, 2003.

[19] A. Favrot, M. Erne, and C. Faller, “Improved cocktail-party processing,” in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, 2006.

[20] J.B. Allen, D.A. Berkeley, and J. Blauert, “Multimicrophone signal-processing technique to remove room reverberation from speech signals,” *J. Acoust. Soc. Am.*, vol. 62, 1977.

[21] J. Merimaa, M. Goodwin, and J.-M. Jot, “Correlation-based ambience extraction from stereo recordings,” in *Proc. Audio Eng. Soc. 123rd Conv.*, 2007.

[22] J. Usher and J. Benesty, “Enhancement of spatial sound quality: A new reverberation-extraction audio upmixer,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, pp. 2141–2150, 2007.

[23] C. Faller, “Multiple-loudspeaker playback of stereo signals,” *J. Audio Eng. Soc.*, vol. 54, 2006.

[24] C. Uhle, A. Walther, O. Hellmuth, and J. Herre, “Ambience separation from mono recordings using Non-negative Matrix Factorization,” in *Proc. Audio Eng. Soc. 30th Int. Conf.*, 2007.

[25] C. Uhle and C. Paul, “A supervised learning approach to ambience extraction from mono recordings for blind upmixing,” in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, 2008.

[26] G. Soulodre, “System for extracting and changing the reverberant content of an audio input signal,” US Patent 8,036,767, Oct. 2011.

[27] International Telecommunication Union, Radiocomunication Assembly, “Algorithms to measure audio programme loudness and true-peak audio level,” Recommendation ITU-R BS.1770-2, March 2011, Geneva, Switzerland.