Audio Engineering Society

Convention Paper 10578

Presented at the 152nd Convention
2022 May, In-Person and Online

This paper was peer-reviewed as a complete manuscript for presentation at this convention. This paper is available in the AES E-Library (http://www.aes.org/e-lib) all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Low Complexity Methods for Robust Stereo-to-Mono Down-mixing

Pallavi Maben[1,2], Christian Borß[1], Bernd Edler[1], and Guillaume Fuchs[1]

[1]*Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany*
[2]*Currently affiliated to Ittiam Systems Pvt. Ltd., Bengaluru, India*

Correspondence should be addressed to Guillaume Fuchs (guillaume.fuchs@iis.fraunhofer.de)

ABSTRACT

Stereo to mono down-mix is a key component of parametric stereo coding to drastically reduce the bit rate, but at the same time it is also an irreversible process that is a potential source of undesirable artifacts. This paper aims to reduce typical distortions induced by down-mixing, such as signal cancellation, comb filtering or unnatural instabilities. Two down-mixing methods are designed with different trade-offs between natural timbre and energy preservation based on simple rules that ensure low complexity. The results of a listening test show that both the proposed methods have a substantial advantage over the passive down-mix, while being very competitive compared to more computationally demanding active down-mixing approaches. The proposed methods are, therefore, particularly well suited to low complexity stereo coding schemes, such as those required for communication applications.

1 Introduction

In many audio applications such as digital audio broadcasting, streaming and storage, audio is widely used in the stereo (two-channel) format. For efficient joint-stereo coding, it is necessary to exploit the cross-channel redundancies and irrelevancies between the stereo channels [1]. In recent years, a technique called Parametric Stereo (PS) has been developed, where the stereo signal is coded as a mono down-mixed signal along with additional stereo parameters modelling the stereo image [2]. This guarantees a minimum bit-rate for coding, and also ensures backward compatibility with legacy mono coding schemes. However, obtaining an efficient monophonic down-mix for stereo audio is a challenging task, due to potential problems such as phase cancellation, comb-filtering, energy mismatch, and signal coloration. Besides, communication applications require low computational complexity processing on limited resources. They also demand low delays to support natural conversation and low memory requirements to reduce costs.

This paper presents two memory-efficient and robust stereo-to-mono down-mixing methods with low computational complexity, which aim at mitigating typical problems in conventional down-mixing. In these methods, the down-mix is computed adaptively in the frequency domain by a weighted combination of the input stereo channels. Different trade-offs can be made between preserving the original timbre and energy.

This paper is organized as follows. An overview of the spatial parameters used in PS coding is presented in Section 2. Some conventional down-mix methods and

their limitations are discussed in Section 3. Section 4 presents the new down-mix methods and their advantages. Performance evaluation based on a listening test is presented in Section 5, followed by the conclusion in Section 6.

2 Overview of parametric stereo coding

Parametric Stereo (PS) is a technique designed for joint coding of stereophonic audio. It generates a monophonic down-mix of the two-channel signal and associates compact spatial side information. This results in significant bit-rate savings compared to coding the stereo channels independently, as the parameters require a reasonably small bit-rate overhead. PS coding employs a dedicated filterbank applying a time-to-frequency domain transform, like an STFT (Short Time Fourier Transform), exploited for generating spatial parameters for different frequency bands and for generating the down-mixed mono signal to encode.

The spatial parameters used in PS coding describe the relationship between the stereo channels. These parameters are computed at the encoder side after a time-frequency analysis of the input stereo channels and are then used for modelling the spatial image which is resynthesized at the decoder side. A set of parameters are calculated for each non-overlapping frequency sub-band b regrouping adjacent frequency coefficients from index k_b to index $k_{b+1}-1$. The spatial parameters generally used are:

- The **Interchannel Level Difference (ILD)**, which is an energy ratio between the channels and is defined as:

$$ILD[b] = 10\log_{10}\frac{\sum_{k=k_b}^{k_{b+1}-1} L[k]L^*[k]}{\sum_{k=k_b}^{k_{b+1}-1} R[k]R^*[k]} \quad (1)$$

  where $L[k]$ and $R[k]$ are the frequency coefficients of the right and left channels respectively.

- The **Interchannel Phase Difference (IPD)**, which represents the amount of phase misalignment between the channels [3], and is a consequence of the time-delay between the two channels.

$$IPD[b] = \angle\left(\sum_{k=k_b}^{k_{b+1}-1} L[k]R^*[k]\right) \quad (2)$$

  IPDs can be seen as an alternative to the Interchannel Time Difference (ITD) by approximating the delay and its associated constant slope of phase along frequency by piecewise constant phase differences.

- The **Interchannel Coherence (IC)**, which is responsible for the width of the stereo image, is defined as the normalized cross-correlation between the channels [4]:

$$IC[b] = \frac{|\sum_{k=k_b}^{k_{b+1}-1} L[k]R^*[k]|}{\sqrt{(\sum_{k=k_b}^{k_{b+1}-1} L[k]L^*[k])(\sum_{k=k_b}^{k_{b+1}-1} R[k]R^*[k])}} \quad (3)$$

3 Conventional down-mix methods

3.1 Passive down-mix

One of the most commonly used down-mix methods is the classical passive down-mix, which is obtained by a simple time-invariant averaging of the two stereo channels, and can be expressed in the frequency domain as follows :

$$M[k] = \frac{L[k]+R[k]}{2} \quad (4)$$

where $M[k]$ represents the frequency coefficient k of the mono down-mixed signal. The energy in the obtained down-mix is given by

$$\sum_{<k>}|M[k]|^2 = \frac{\sum_{<k>}|L[k]|^2 + \sum_{<k>}|R[k]|^2 + 2\Re(\sum_{<k>} L[k]R^*[k])}{4} \quad (5)$$

where $\sum_{<k>}$ denotes the sum over all frequency bins.

Ideally, the target energy of the down-mix should be equal to half the sum of the energies of the two channels [5] and also independent of ILD, IPD and IC:

$$S_D = \frac{\sum_{<k>}|L[k]|^2 + \sum_{<k>}|R[k]|^2}{2} \quad (6)$$

However, the passive down-mix satisfies the above criterion only if ILD = 0dB, IPD = 0 rad and IC = 1, i.e. L = R. In the general case, the energy of the passive down-mix is under-estimated and is largely dependent on the relation between the channels. Due to this dependency, passive down-mix energy can fluctuate greatly especially when the channels are coherent. In the most critical scenario when the channels are nearly of the same level, out-of-phase and coherent (ILD = 0dB, IPD = π rad, IC = 1), the signal components of the

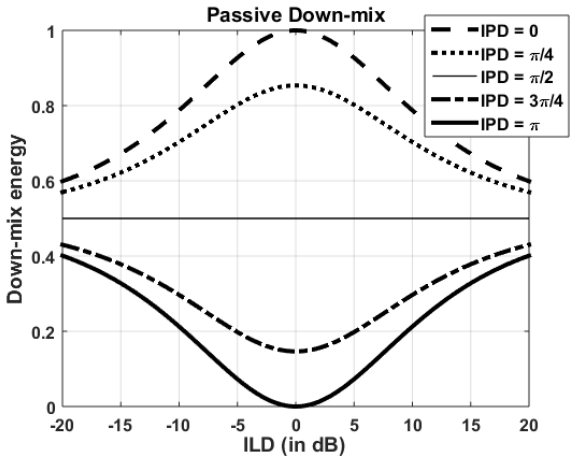


Fig. 1: Energy variation of passive down-mix for IC=1; energy is not preserved for a large number of cases.

passive down-mix are completely cancelled out, as shown in Figure 1. Moreover, in case of completely incoherent channels (IC $= 0$), the passive down-mix shows a systematic energy attenuation of 3dB. Since signal cancellation happens only when coherent components of the two channels are added, for the analysis from now on, we focus on the case where IC $= 1$.

3.2 Down-mix methods based on phase-alignment

There exists a family of down-mix approaches which addresses the signal cancellation problem of the passive down-mix by phase-aligning the stereo channels prior to the mixing process [1, 6, 7]. These methods undoubtedly avoid the problem of signal cancellation due to phase mismatch between the channels. Nevertheless, none of these approaches are fully suitable as the phase rotation of the channels can lead to an unnatural mixing of the input channels, and can create severe instabilities and block effects, especially when phase jumps occur between adjacent time frequency bands.

3.3 Down-mix methods based on coherence analysis

In the existing literature, more evolved down-mix techniques have been proposed, some of which are based on the observation that signal cancellation during down-mixing occurs only for time-frequency components which are coherent between the two channels. The method in [8] is based on suppressing the coherent signal parts of one of the two stereo channels before down-mixing. In [9], the phase alignment is performed only for the coherent components, along with a regularization step in order to reduce abrupt changes in phase adjustment coefficients between adjacent frames. However, both techniques are computationally demanding. In [8], the filter coefficients for suppressing the coherent signal parts and the scaling factors to adjust the overall energy in the down-mix need to be computed per time-frequency bin, whereas in [9] a covariance matrix between the channels has to be computed.

4 New Down-mix Methods

The new down-mix methods are based on a controlled energy-equalization of the sum signal (L+R), which is complemented by the L-channel (which is arbitrarily chosen) in case of signal cancellation. The down-mix is performed in frequency domain and has the following general form:

$$M[k] = W_1[k](L[k] + R[k]) + W_2[k]L[k] \tag{7}$$

where $W_1[k]$ and $W_2[k]$ are the weights that determine the relative amounts of the sum component $(L[k]+R[k])$ and the left-channel $(L[k])$ in the down-mix. Figure 2 illustrates the principle of the new down-mix methods. The radius of the circle denotes the energy level that is desired in the down-mix according to equation (6). We have seen from the passive down-mix that the sum term (L+R), by itself, is unable to preserve the desired energy in the down-mix for coherent channels, especially at large phase-shifts. Hence, by using additionally the

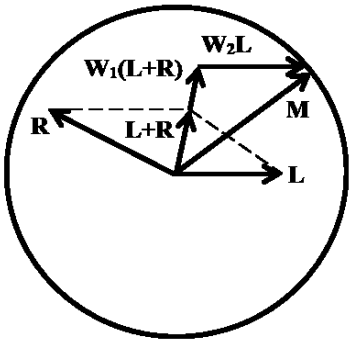


Fig. 2: Illustration of energy preservation principle of the new down-mix methods.

weighted L-channel as the second component, it is possible to reach closer to this target level.

Based on the selection of the weights $W_1[k]$ and $W_2[k]$, two variants of the new down-mix can be achieved, and these are discussed next.

4.1 Weighted Active Down-mix

In the first approach, the down-mix $M[k]$ is derived from cross-fading between an initial down-mix prototype $D[k]$ and the left channel $L[k]$

$$M[k] = \lambda[k]D[k] + (1-\lambda[k])\,L[k] \quad , \tag{8}$$

where the cross-fading coefficient $\lambda[k]$ and the initial down-mix prototype $D[k]$ are defined as follows.

$$D[k] = \frac{\sqrt{|L[k]|^2+|R[k]|^2}}{\sqrt{2}\,|L[k]+R[k]|}\,(L[k]+R[k]) \tag{9}$$

$$\lambda[k] = \frac{|L[k]+R[k]|}{|L[k]|+|R[k]|} \tag{10}$$

The initial down-mix prototype $D[k]$ normalizes the sum signal $L[k]+R[k]$ to the energy of the input signals whereas the cross-fading coefficient $\lambda[k]$ eliminates the pole for coherent out-of-phase input signals.

Inserting (9) and (10) into (8) yields the down-mix weights $W_1[k]$ and $W_2[k]$ according to equation (7).

$$W_1[k] = \frac{\sqrt{|L[k]|^2+|R[k]|^2}}{\sqrt{2}\,(|L[k]|+|R[k]|)} \tag{11}$$

$$W_2[k] = 1 - \frac{|L[k]+R[k]|}{|L[k]|+|R[k]|} \tag{12}$$

Figure 3 shows the energy response of the weighted active down-mix for different ILDs and IPDs. Overall, the new down-mix preserves the target energy better than the passive down-mix under many conditions while guaranteeing a natural mixing of the input channels. In addition, the new active down-mix does not show a systematic attenuation in case of incoherent input channels. Due to this natural mixing, the weighted active down-mix is more likely to preserve the timbral quality of the original signal. However, it comes with a trade-off that energy preservation is not guaranteed for the whole range of ILDs and IPDs, and especially for the specific case where the channels are coherent and out of phase, which is a generally rare case. This is illustrated in Figure 3.

4.2 Energy Preserving Down-mix

The constraint on the energy preservation imposed by the weighted active down-mix is insufficient, and the target energy is not reached over all possible level and phase differences. The second variant of the new down-mix tries to overcome this by using a hard constraint on the weights. A brief description of the computation of the weights $W_1[k]$ and $W_2[k]$ is given below:

- First, the weight $W_1[k]$ is computed as

$$W_1[k] = \frac{1}{|L[k]+R[k]|}\sqrt{\frac{|L[k]|^2+|R[k]|^2}{2}} \tag{13}$$

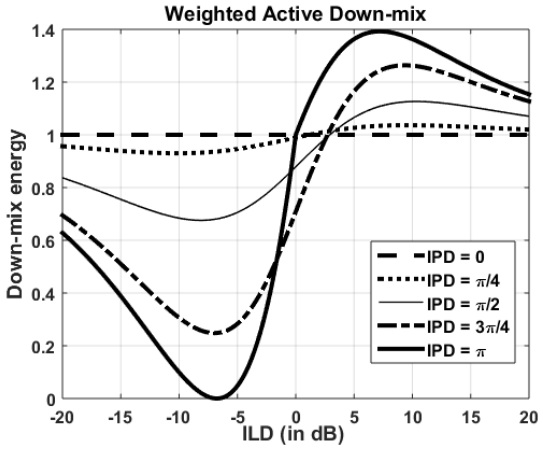


Fig. 3: Energy variation of weighted active down-mix for IC = 1; energy is well preserved for a majority of the cases, except for the rare case of fully coherent and out-of-phase, or nearly out-of-phase channels.

To avoid the weight $W_1[k]$ from increasing drastically, for large phase-shifts, an exponential normalization is applied to limit the maximum value of $W_1[k]$ to unity, according to:

$$W_1'[k] = \begin{cases} \frac{1}{\sqrt{2}} + (1-\frac{1}{\sqrt{2}})\exp(\frac{\frac{1}{\sqrt{2}}-W_1[k]}{1-\frac{1}{\sqrt{2}}}), & \text{if } W_1[k] > \frac{1}{\sqrt{2}} \\ W_1[k], & \text{otherwise} \end{cases} \tag{14}$$

- The weight $W_2[k]$ is then calculated so as to fulfill the following energy preservation criterion:

$$|W_1'[k](L[k]+R[k]) + W_2[k]L[k]|^2 = \frac{|L[k]|^2+|R[k]|^2}{2} \tag{15}$$

Solving equation (15) leads to a quadratic equation in $W_2[k]$, further leading to two possible roots. Out of these, the positive root is considered as the solution to $W_2[k]$:

$$W_2[k] = \frac{1}{|L[k]|}\left(\sqrt{\frac{|L[k]|^2+|R[k]|^2}{2}} - W_1'[k]|L[k]+R[k]|\right) \tag{16}$$

Thus, the energy preserving down-mix ensures that the target energy of the down-mix is preserved over the full range of level and phase differences. However, this strict energy preservation is found to affect the timbral quality of the original signal.

5 Experimental Results

A multi-stimulus listening test was conducted to assess the performance of the two proposed down-mix techniques in comparison with the down-mix based on coherence suppression [8] and the passive down-mix. However, it differed from the MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) test, as no hidden reference was included, because the test conditions being the mono down-mixed signals were compared against the original stereo signal. Also no low-(3.5kHz low-pass filtered) and mid-(7kHz low-pass filtered) anchors were used because of the absence of a good reference down-mix, from which these anchors could be derived. A low-delay STFT-based coding framework was used for generating the different active down-mixes. The frame size used was 20ms with a hop size of 10ms and with 25% overlap between adjacent windows. No source coding was used and the signals, after down-mixing, were directly assessed. The listeners were asked to evaluate the basic audio quality of the test conditions. The test items for the listening test were chosen from a set of the most critical audio test materials, which are expected to audibly stress the underlying system. A total of 10 stereo items which were found to be critical for down-mixing in general, were used for the study. They are listed in Table 1. Eight expert listeners participated in the test.

Figure 4 shows the results of the listening test with the mean and 95% confidence intervals of the absolute scores (MUSHRA points) indicated for each test item as well as over all 10 test items. Overall, the quality of the two new down-mix methods is significantly better than that of the passive down-mix. Out of the two new methods, the weighted active down-mix has an

Table 1: Description of test items used for subjective quality evaluation.

Item no.	Item name	Description
1	Austin	Speech immersed in background noise (clip from movie “Austin Powers”)
2	Fatboy	Robotic voice from a vocoder (beginning of “Kalifornia” by Fatboy Slim)
3	Music	Artificially panned music item (using sc03 from MPEG test set)
4	bruckner	Orchestral piece (from one of Bruckner’s symphonies)
5	glock	Glockenspiel (MPEG test set)
6	harmonics	Tonal signal with harmonic frequencies (MPEG test set)
7	lightning	Guitar strumming (track from “Lightning Crashes” by rock-band Live)
8	si01	Harpsichord playing an arpeggio (MPEG test set)
9	si02	Castanets (MPEG test set)
10	sm01	Bagpipes (MPEG test set)

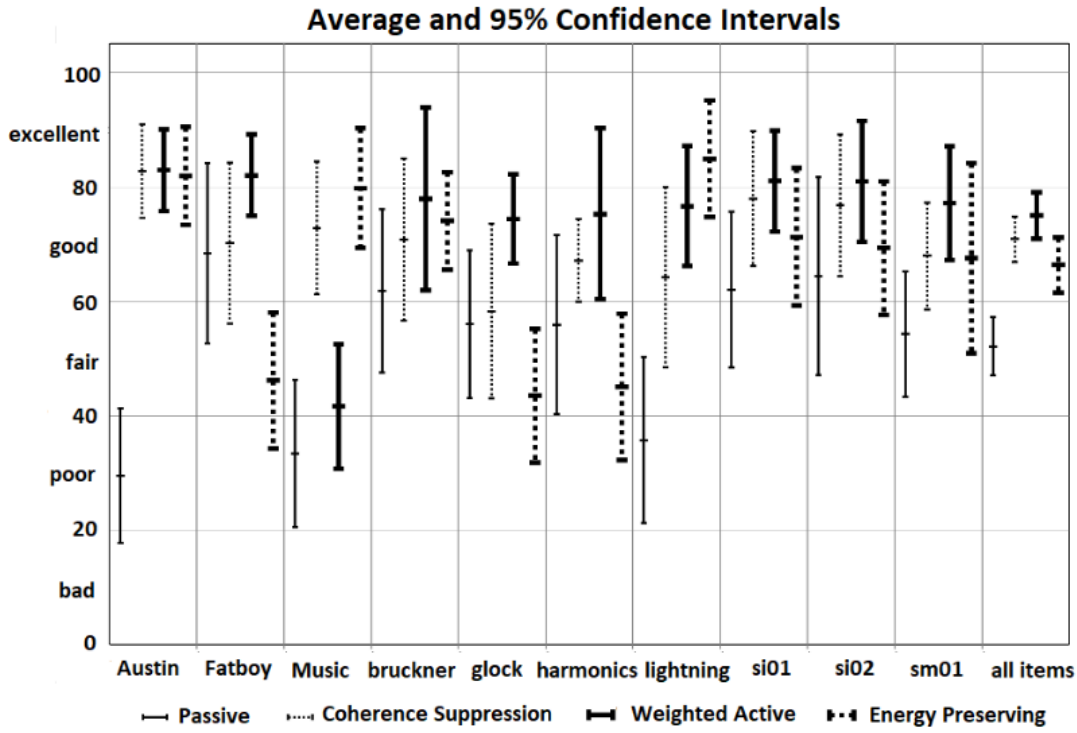


Fig. 4: Subjective quality evaluation showing significantly better audio quality for the new down-mix methods compared to passive down-mix.

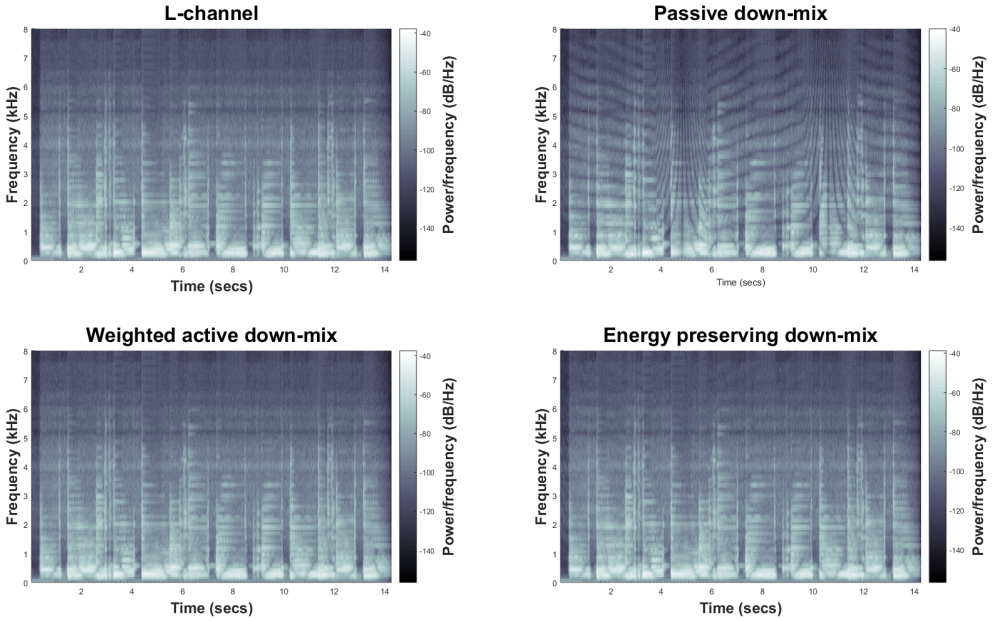


Fig. 5: Illustration of comb filtering - one of the drawbacks of passive down-mixing, which can be overcome by the new down-mix methods.

upper hand in terms of audio quality. For some specific items (Fatboy, glock, harmonics) the quality improvement is significant. The weighted active down-mix also ranks higher than the down-mix using coherence suppression [8], in terms of quality.

An example for a critical item for the passive down-mix is **lightning**, for which, at regular intervals, certain frequency components remain in-phase, while other frequency components become out-of-phase. Performing passive down-mix on such a signal, where the stereo channels are simply summed up, causes certain frequency components to add constructively, and other frequency components to add destructively, resulting in comb-filter artifacts, as illustrated in Figure 5. On the other hand, the new down-mix methods prove their capability to avoid cancellation of out-of-phase signal components, and avoid perceivable spectral bias.

In addition, in order to test the effect of stereo panning on the down-mix, an artificial item, **Music** was created with complete panning from left to right, and out-of-phase channels, owing to the fact that the performance of most down-mix methods shows a degradation when IPD $= \pi$ rad, as illustrated in the earlier sections. This artificial item was created by considering a short segment of a music item as the base for for both stereo channels, followed by the introduction of an ILD variation between the channels, starting from -20dB to +20dB and back to -20dB. This serves to test any variations in the energy level in the resulting down-mix at different ILDs. This item was designed for testing the worst case scenario for the passive down-mix and both the newly proposed down-mixes. It was found to be critical for the passive down-mix and the weighted active down-mix, both of which exhibit a null-point at a certain range of ILDs for out-of-phase channels. Hence a variation in the energy level can be observed when listening to the respective down-mixed signals, which is also illustrated in Figure 6. The energy preserving down-mix on the other hand, ensures perfect energy preservation over the entire range of ILDs, and thus maintains a uniform energy level.

6 Conclusion

In this paper, a new family of low complexity methods for stereo-to-mono down-mixing has been presented. The new methods are able to solve typical problems in

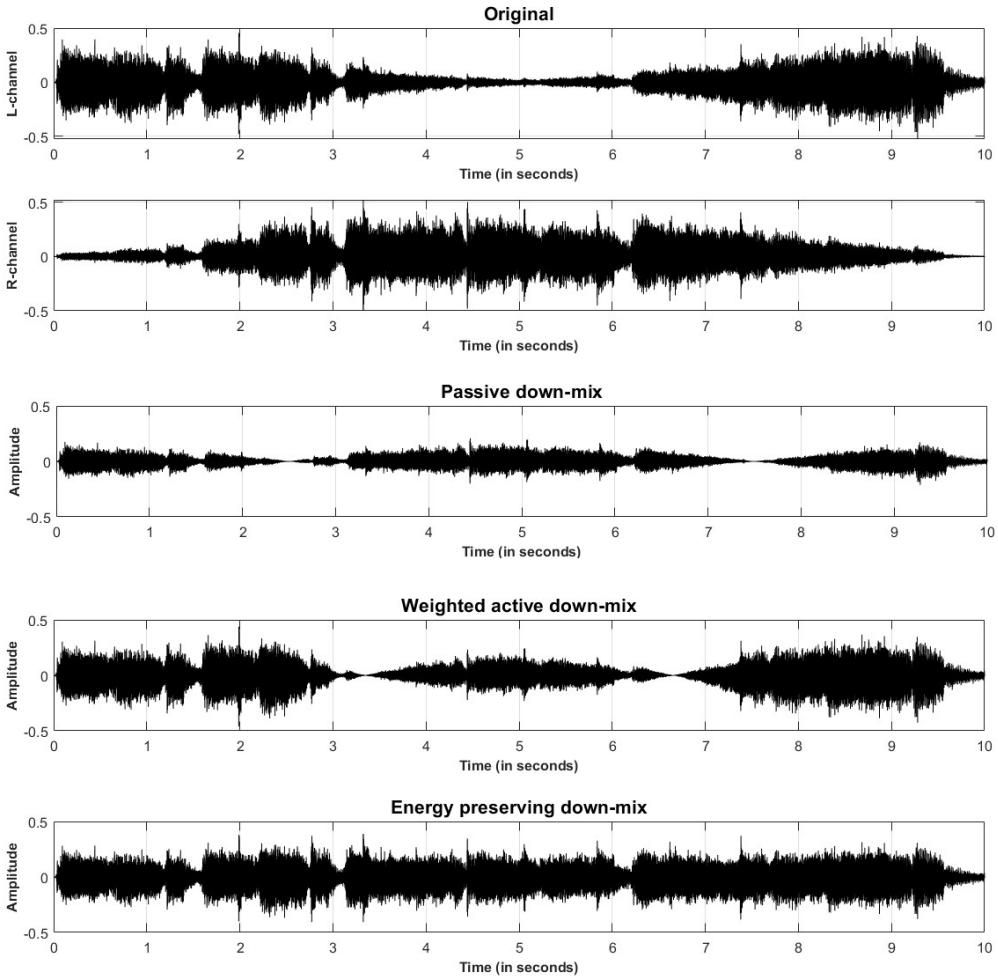


Fig. 6: Illustration of signal cancellation in passive and weighted active down-mix for certain ILDs for out-of-phase channels, which can be overcome by using the energy preserving down-mix.

conventional down-mixing, such as signal cancellation, coloration and comb-filter artifacts for out-of-phase signal components. Subjective listening test evaluation shows that the proposed methods are significantly better than the traditional passive down-mix in terms of audio quality. Moreover, they are computationally less-complex, making them suitable for communication applications.

This paper highlights a trade-off between down-mix energy preservation and timbral quality as seen in the two variants of the new down-mix. The weighted active down-mix, on one hand, ensures a more smooth and natural mixing of the stereo channels, and is more relaxed with respect to the energy preservation criterion, and hence preserves with great fidelity, the original timbre. On the other hand, the energy preserving down-mix ensures, under any condition, a perfect energy preservation in the down-mix. This, though advantageous, requires strict conditions to be imposed on designing the weights leading to a more strenuous down-mixing, and thereby affecting the naturalness of the timbre. Therefore, in a nutshell, it is necessary for a good down-mix to strike a good balance between preserving the target energy and the timbral quality.

References

[1] Hoang, T. M. N., Ragot, S., Kövesi, B., and Scalart, P., "Parametric stereo extension of ITU-T G. 722 based on a new downmixing scheme," in *Multimedia Signal Processing (MMSP), 2010 IEEE International Workshop on*, pp. 188–193, IEEE, 2010.

[2] Wu, W., Miao, L., Lang, Y., and Virette, D., "Parametric stereo coding scheme with a new downmix method and whole band inter channel time/phase differences," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 556–560, IEEE, 2013.

[3] Kurniawati, E., Poh, N. B., Sattar, F., George, S., et al., "A Stereo to Mono Dowmixing Scheme for MPEG-4 Parametric Stereo Encoder," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 5, pp. V–V, IEEE, 2006.

[4] Breebaart, J., van de Par, S., Kohlrausch, A., and Schuijers, E., "Parametric coding of stereo audio," *EURASIP Journal on Applied Signal Processing*, 2005, pp. 1305–1322, 2005.

[5] et al., B., "Processing of Multi-Channel Signals," 2008, uS Patent 7,343,281 B2.

[6] Kurniawati, E., Ng, B. P., Sapna, G., Sattar, F., et al., "A Subband Domain Downmixing Scheme for Parametric Stereo Encoder," in *Audio Engineering Society Convention 120*, Audio Engineering Society, 2006.

[7] Baumgarte, F., "Audio mixing using magnitude equalization," 2012, uS Patent 8,090,122.

[8] Adami, A., Habets, E. A., and Herre, J., "Downmixing using coherence suppression," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2878–2882, IEEE, 2014.

[9] Vilkamo, J., Kuntz, A., and Füg, S., "Reduction of spectral artifacts in multichannel downmixing with adaptive phase alignment," *Journal of the Audio Engineering Society*, 62(7/8), pp. 516–526, 2014.