DOWN-MIXING USING COHERENCE SUPPRESSION

Alexander Adami, Emanuël A. P. Habets, Jürgen Herre

International Audio Laboratories Erlangen*
Am Wolfsmantel 33, 91058 Erlangen
alexander.adami@audiolabs-erlangen.de

ABSTRACT

A common problem in audio signal processing is to mix two or more signals into one sum signal. The mixing procedure, known as down-mixing, usually introduces some signal impairments, especially if two signals contain similar but phase shifted signal components. Summing up such signals results in severe comb-filter artifacts. In this paper, we propose a novel down-mix method which prevents comb-filter effects. This is achieved by suppressing the coherent signal parts of one input signal prior to mixing. During the down-mixing, a scaling gain is applied ensuring the preservation of overall signal energy. Furthermore, a phase-align extension to the down-mixer is introduced. The proposed down-mixer is evaluated qualitatively in comparison with two existing down-mix approaches as well as quantitatively by means of a distortion analysis.

Index Terms— Down-Mixing, Spatial Audio

1. INTRODUCTION

Since the introduction of stereo playback systems in the middle of the twentieth century, there has been a demand for backward compatibility to the more common mono playback systems. Nowadays, stereo playback systems are commonly used but there is a plurality of different multichannel playback content available in, for instance, 5.1, 7.1 and 22.2. This increases the wish for inter-system compatibility or at least stereo backward compatibility. Therefore, a down-mix mechanism is needed, which is able to mix a given number of channels into a lower number of channels without introducing artifacts.

Different methods have been proposed to achieve this down-mix task. Some of them are based on tree-structure-like down-mix approaches [1] and others perform the mixing all at once [2, 3]. To date, there are basically three principles of carrying out a down-mix:

- The passive principle, where the signals are simply summed up. For instance, the International Telecommunication Union (ITU) proposed a 5-to-2 channel time-domain matrixed down-mix [2]. Dolby Pro Logic II includes also a matrixed down-mix which applies phase shifts to the back-channels to decode them if necessary [3]. However, those simple mixing approaches can lead to down-mix artifacts such as comb-filters and signal dependent cancelations and amplifications due to misaligned correlated signal components in the signals to be down-mixed.
- The active post-processing principle, where signal impairments due to passive down-mixing are attempted to be compensated. Baumgarte et al., for instance, applied an adaptive equalizer to the passive down-mix signal [4], and Stoll et al. used additive correction terms to restore the impaired down-mix signal [5]. Those treatments, however, just scale (on a time-frequency tile basis) the already impaired signal such that the overall energy is preserved.
- The phase-align principle, where the signals are temporally or phase aligned prior to the mixing process. This has been proposed, for instance, to improve parametric stereo coders [6, 7, 8]. A continuous and robust phase-alignment is not an easy task and any misalignment will immediately result in comb-filter artifacts.

In this paper, we present a stereo-to-mono down-mix method which acts up to a new principle, where the down-mix signal is obtained by a superposition of a reference signal and those components of the second signal which are uncorrelated to the reference. This way, comb-filter artifacts can be avoided from the beginning. Furthermore, the reference signal is scaled prior to the mixing to preserve overall signal energy.

The remainder of this paper is structured as follows: in Sec. 2, the problem formulation and underlying signal model is given. In Sec. 3 two existing down-mix methods are briefly reviewed and in Sec. 4, our proposed method is introduced. A performance evaluation will be given in Sec. 5 followed by a conclusion in Sec. 6.

2. PROBLEM FORMULATION

In this work, we focus on the down-mixing of two signals in the short time Fourier transform (STFT) domain. The signals are denoted by $X_1(k,m)$ and $X_2(k,m)$, where k and m denote the discrete frequency and time indices. The proposed method aims at preventing artifacts from the beginning by suppressing the coherent signal components of $X_1(k,m)$ within $X_2(k,m)$ before down-mixing. Furthermore, prior to the actual mixing, $X_1(k,m)$ is scaled to meet a predefined energy constraint. The used signal model is given by

$$X_2(k,m) = W(k,m) \cdot X_1(k,m) + U(k,m), \qquad (1)$$

i.e., $X_2(k,m)$ is assumed to consist of the sum of a correlated and an uncorrelated signal part with respect to $X_1(k,m)$. The correlation between both signals is described by the filter coefficients $W(k,m)$. The uncorrelated signal component is represented by $U(k,m)$. We now define the desired down-mix signal as:

$$D(k,m) = G_{X_1}(k,m)X_1(k,m) + U(k,m), \qquad (2)$$

where $G_{X_1}(k,m)$ is a scaling factor to adjust the overall energy of the correlated signal part for each time-frequency bin such that the overall energy of the down-mix signal equals the sum of the energies

*A joint institution of the Friedrich-Alexander-University Erlangen-Nürnberg (FAU) and Fraunhofer IIS, Germany.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

of the individual input signals. An estimate of the desired down-mix signal is given by:

$$\widehat{D}(k,m) = \widehat{G}_{X_1}(k,m)X_1(k,m) + \widehat{G}(k,m)X_2(k,m), \quad (3)$$

where $\widehat{G}(k,m)$ is the suppression gain which suppresses the correlated signal component within $X_2(k,m)$. Given that model, our objective is to estimate $\widehat{G}_{X_1}(k,m)$ and $\widehat{G}(k,m)$. In the remainder of this paper, time and frequency indices will be omitted for brevity.

3. REVIEW AND ANALYSIS OF TWO EXISTING METHODS

Before we further introduce our proposed method, we briefly review two existing methods for a stereo-to-mono conversion and note their shortcomings.

3.1. Passive Down-Mix

The passive down-mix [2] is one of the most commonly used approaches and exhibits a very low computational complexity since it consists of a single addition, i.e.,

$$D_{\mathrm{P}} = X_1 + X_2. \quad (4)$$

The drawback of this approach becomes obvious when we analyze the energy of the resulting down-mix signal, which is given by:

$$\Phi_{D_{\mathrm{P}}} = \Phi_{X_1} + \Phi_{X_2} + 2\,\mathrm{E}\left\{|X_1|\,|X_2|\cos(\phi_{X_1} - \phi_{X_2})\right\}, \quad (5)$$

where $\Phi_Q = \mathrm{E}\left\{|Q|^2\right\}$ corresponds to the power spectral density (PSD) of a signal Q and $\mathrm{E}\{\cdot\}$ is the expectation value. Equation (5) shows that the energy of the down-mix signal depends not only on the individual signal energies, but also on the cosine of the phase difference between both signals (denoted by ϕ_{X_1} and ϕ_{X_2}). Consequently, if X_2 is a delayed version of X_1, the down-mix's signal energy varies across frequency and shows the typical comb-filter structure. This becomes even more visible, if we consider the signal model in (1) for the passive down-mix:

$$D_{\mathrm{P}} = X_1 + WX_1 + U = (1+W)X_1 + U. \quad (6)$$

If W implies a phase shift of π, the correlated part of the input signals is completely canceled. In general, using a passive down-mix always bears the risk of generating comb-filter or comb-filter-like artifacts if the signals are not uncorrelated with respect to each other.

3.2. Active Down-Mix

To mitigate comb-filter effects as generated by a passive down-mix, an adaptive post-scaling can be applied to the down-mix signal [4]. The active down-mix signal is given by

$$D_{\mathrm{A}} = (X_1 + X_2)\cdot G_{\mathrm{PS}}, \quad (7)$$

where $G_{\mathrm{PS}} = \sqrt{\frac{\mathrm{E}\left\{|X_1|^2\right\}+\mathrm{E}\left\{|X_2|^2\right\}}{\mathrm{E}\left\{|X_1+X_2|^2\right\}}}$ denotes the post-scaling gains. The scaling factor is derived by two power measures; the first before the actual down-mix process (corresponding to the numerator) and the second after the down-mixing (corresponding to the denominator). Short-time energies are used to realize the power measures. One drawback of this approach is that spectral notches can be restored effectively only if the frequency bands of the STFT are sufficiently small compared to the width of the notches. But the main drawback is that the post-scaling only restores the energy relations using an already impaired signal, i.e., if a down-mix suffers from comb-filtering, signal canceling has taken place already, and the post-scaling just amplifies the noisy residual signal to the energy level of the sum of the individual input signal energies. If we apply an active down-mix to the signal model in (1), it becomes:

$$D_{\mathrm{A}} = [(1+W)X_1 + U]\cdot G_{\mathrm{PS}}. \quad (8)$$

If W produces a phase shift of π, the correlated signal part of the input signal is completely canceled and only the uncorrelated signal part U is scaled by G_{PS} according to the corresponding signal energies. Since G_{PS} is real-valued, D_{A} exhibits the same phase as U.

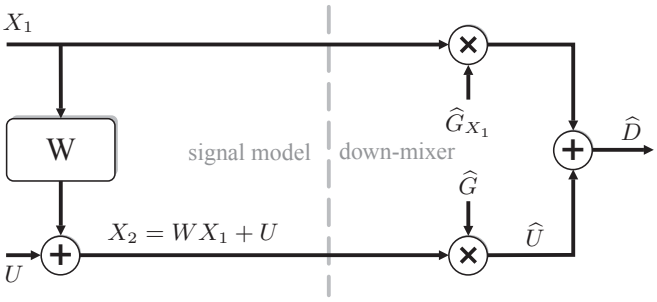

Fig. 1. Signal flow diagram of the down-mixer including the assumed signal model.

4. PROPOSED METHOD

In contrast to the previous approaches, we introduce an approach which aims at preventing the generation of comb-filters from the beginning. The section is sub-divided into the explanation of the coherence suppression, the energy scaling and a phase-align extension.

4.1. Coherence Suppression

The proposed method is based on suppressing coherent signal parts such that only a scaled reference signal (X_1) and a signal that is uncorrelated with respect to the reference signal are added together. In Fig. 1, the signal flow diagram of the down-mixer is depicted, including the assumed signal model. It should be noted that, given an estimate of W, we can in principle subtract $\widehat{W}X_1$ from X_2 to obtain an estimate of U. Through various experiments, we have found that W changes rapidly across time and frequency and direct subtraction of $\widehat{W}X_1$ does not yield the desired result. By using the gain function $\widehat{G}$, we can use approaches that are commonly used in single-channel speech enhancement to avoid audible artifacts.

The main task is to determine the signal part U which is uncorrelated to X_1. An estimate of this signal is obtained using

$$\widehat{U} = \widehat{G}\cdot X_2, \quad (9)$$

where we assume $\widehat{G}$ to be real-valued. The desired gains are determined by minimization of the mean squared error between $\widehat{U}$ and U, i.e., $\widehat{G} = \arg\min_G J(G)$ with $\widehat{G} \in \mathbb{R}$ and

$$\begin{aligned} J(G) &= \mathrm{E}\left\{\left|U - \widehat{U}\right|^2\right\} \\ &= \Phi_U(1 - 2\,G + G^2) + G^2\Phi_{WX_1}. \end{aligned} \quad (10)$$

To determine $\widehat{G}$, the partial derivative $\frac{\partial}{\partial G} J(G)$ is set to zero. The resulting term is solved for G, leading to the estimate

$$\widehat{G} = \frac{\Phi_U}{\Phi_U + \Phi_{WX_1}}, \quad 0 \leq \widehat{G} \leq 1. \tag{11}$$

Equation (11) can be written such that $\widehat{G}$ is only expressed by the power ratio of the signals U and WX_1, i.e.,

$$\widehat{G} = \frac{1}{1 + \frac{\Phi_{WX_1}}{\Phi_U}} = \frac{1}{1 + \Psi^{-1}}, \tag{12}$$

where $\Psi = \frac{\Phi_U}{\Phi_{WX_1}}$. In this work, the power ratio Ψ is estimated using the decision-directed estimator proposed in [9]:

$$\widehat{\Psi}(m) = \alpha_{dd} \frac{\left|\widehat{U}(m-1)\right|^2}{\Phi_{WX_1}(m-1)} + (1-\alpha_{dd}) P[\gamma_m - 1], \tag{13}$$

with $\gamma(m) = \frac{|X_2(m)|^2}{\Phi_{WX_1}(m)}$, $P[x] = \max(x, 0)$ and α_{dd} being a weighting factor. Since WX_1 is not directly observable, we need an estimate of the filter coefficients W to estimate the energy Φ_{WX_1}. This is done by minimizing the mean squared error between WX_1 and X_2, i.e. $\widehat{W} = \arg\min_W \mathrm{E}\left\{|X_2 - WX_1|^2\right\}$ which leads to

$$\widehat{W} = \frac{\mathrm{E}\left\{X_2 X_1^*\right\}}{\mathrm{E}\left\{|X_1|^2\right\}}. \tag{14}$$

An estimate of Φ_{WX_1} is then obtained by recursively averaging $\left|\widehat{W}X_1\right|^2$.

4.2. Energy Scaling

Since we suppressed the correlated signal part $\widehat{W}X_1$ of X_2, we need to scale X_1 to assure the resulting down-mix signal $\widehat{D}$ to be energy preserving. Hence, we like to find a scaling factor G_{X_1} such that

$$\Phi_{\widehat{D}} = \widehat{G}_{X_1}^2 \Phi_{X_1} + \Phi_{\widehat{U}} \stackrel{!}{=} \Phi_{X_1} + \Phi_{X_2}. \tag{15}$$

Solving (15) for $\widehat{G}_{X_1}$ leads to the desired scaling gains:

$$\widehat{G}_{X_1} = \sqrt{1 + \frac{\Phi_{X_2}}{\Phi_{X_1}} - \frac{\Phi_{\widehat{U}}}{\Phi_{X_1}}}. \tag{16}$$

4.3. Phase-Align Extension

The suppression is carried out by a real-valued gain function. Therefore, only the magnitude of X_2 is affected and its phase is left unchanged. It would be advantageous to additionally align the input signal's phases to reduce signal cancelation in cases where a suppression of coherent signal parts works insufficiently. Fortunately, we already have the information of the phase relationships implicitly due to estimating W. As an extension to the proposed down-mixer, we can easily extract the phase $\widehat{\theta}_W = \angle\widehat{W}$, where $\angle(\cdot)$ symbolizes the phase angle extraction operator, and apply it (with $j = \sqrt{-1}$ denoting the complex unit) to the suppression gains such that (12) becomes

$$\widehat{G} = \frac{1}{1 + \widehat{\Psi}^{-1}} \cdot e^{-j\widehat{\theta}_W}. \tag{17}$$

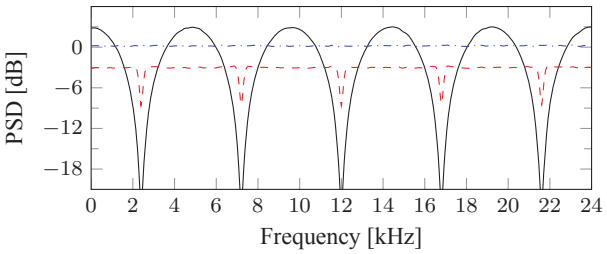

Fig. 2. PSD plot of a down-mixed white noise signal with an inter-channel delay of 10 samples: passive down-mix (solid), active down-mix (dashed, -3dB offset), proposed method (dash-dotted).

5. PERFORMANCE EVALUATION

For the evaluation of the proposed coherence suppression down-mixer, we used a sample rate of 48 kHz, a block size of 256 samples, and 128 samples between successive blocks. Every block was sine-windowed and transformed into frequency domain using a 512 point discrete Fourier transform. All expectation values were approximated by a first-order IIR filter:

$$\bar{\Phi}_X(k,n) = \alpha \cdot |X(k,n)|^2 + (1-\alpha)\bar{\Phi}_X(k,n-1), \tag{18}$$

where α is the forgetting factor with $0 < \alpha \leq 1$. The expectation values needed to determine $\widehat{W}$ were computed with an α corresponding to a time constant of 100 ms. For all other expectation values, a time constant of 30 ms was chosen. Within the decision-directed $\widehat{\Psi}$ estimation, a weighting with $\alpha_{dd} = 0.95$ is applied. We evaluate our proposed down-mixer in comparison to the two existing down-mix methods described in Sec. 3.

5.1. PSD of White Noise Signal with Inter-Channel Delay

Figure 2 depicts an estimate of the PSDs of the resulting down-mix signals originating from an active, passive and our proposed down-mixer. A two channel, double mono white noise signal served as input, where the right channel was 10 samples delayed. The passive down-mix shows the expected comb-filter envelope (solid line) and also the active down-mix (dashed line) reveals not compensable notches due to the given frequency resolution. Please note that an offset of -3 dB was applied for this plot. Our proposed down-mixer (dash-dotted line) produces a down-mix signal with a near flat PSD.

5.2. Signal with Opposite-Phase Component

Now, we want to qualitatively test how well signals with out-of-phase components are treated by the algorithms. The signal was taken from a movie audio track and consists of an opposite-phase speech component and some background noise. The speech is active roughly between seconds 1 to 1.75 and 2.5 to 3.75. In Fig. 3, the spectrograms of the left channel of the input signal and the output signals of the corresponding down-mixers are depicted. The image section only shows the region of interest where circled areas indicate speech activity. In the down-mix signal produced by the passive down-mixer, the speech components are almost completely canceled. The adaptive equalizer in the active down-mixer restores the original energy within the corresponding time-frequency bins, but based on the impaired passive down-mix signal. A clear loss of fine structure across frequency and time is visible. In the output signal of our proposed down-mixer, the speech component is completely preserved with no loss of fine structure.

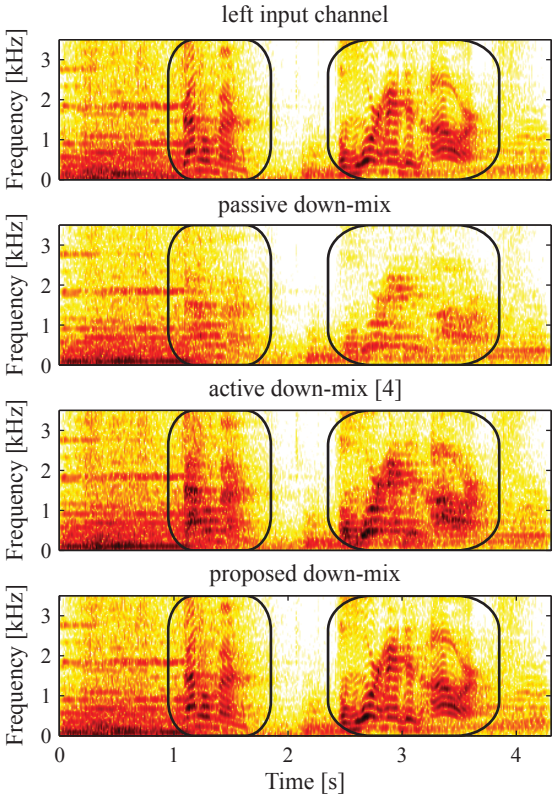

Fig. 3. Spectrogram of opposite-phase speech down-mix signals: left input channel, passive down-mix, active down-mix, proposed down-mix. Color range: 60 dB from white to black.

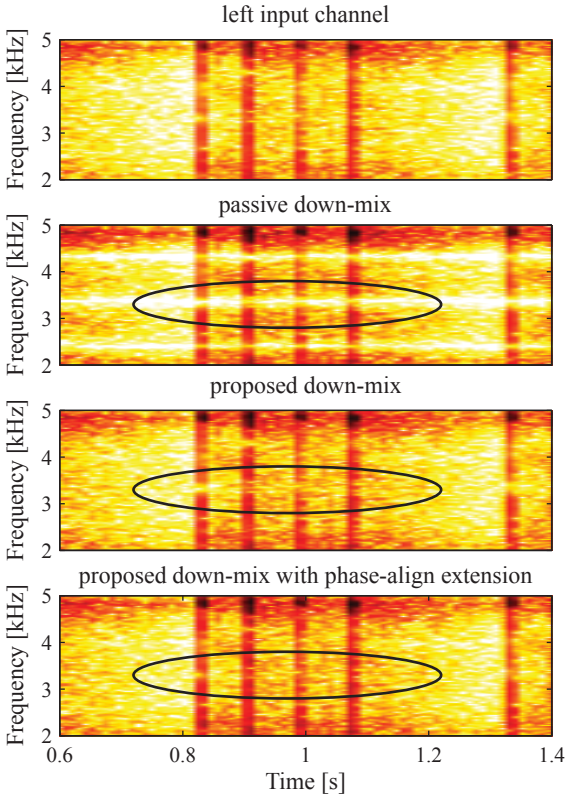

Fig. 4. Spectrogram of down-mix signals: left input channel, passive down-mix, proposed down-mix without phase-align extension, proposed down-mix with phase-align extension. Color range: 60 dB from white to black.

5.3. Benefit of the Phase-Align Extension

To show the benefit of the phase-align extension, we used a castanet signal with an inter-channel delay of 50 samples. The signal was processed by a passive down-mixer and our proposed down-mixer with and without the phase-align extension. The spectrograms of the left channel of the input signal and the down-mix signals are depicted in Fig. 4. The image section only shows the region of interest. As expected, the passive down-mix exhibits a comb-filter envelope with clearly visible notches. Our proposed down-mixer without the phase-align extension produces a signal without the typical comb-filter envelope but also shows some cancelations. This is due to estimation errors within $\widehat{G}$ which results in some remaining correlated signal components within $\widehat{U}$ which cause the cancelations. If the phase-align extension is used, those cancelations can be further reduced. In Fig. 4, a region is circled where this behavior can well be seen.

5.4. Distortion Comparison

For a quantitative evaluation, we computed the distortion of each down-mix with respect to an ideal down-mix D. The down-mix was synthesized according to the desired down-mix signal as given in (2). As distortion measure, the log-spectral distance was used which is given by $\Gamma = \sqrt{\frac{1}{k}\sum_k \left(\log(|D|^2) - \log(|D'|^2)\right)^2}$, where $D' \in \{D_\mathrm{P}, D_\mathrm{A}, \widehat{D}\}$. We used the same input signal as in Sec. 5.3 for the distortion analysis. In Table 1, the mean log-spectral distance values $\bar{\Gamma}$ are given for the considered down-mix approaches. Expectably, the passive down-mix produces the highest distortion value with about 26 dB. Since the active down-mix is limited in equalizing the spectral notches due to the given frequency resolution, the distortion value is still quite high: about 12 dB. With about 5 dB, the proposed down-mix exhibits the lowest distortion value. This result can be even improved a bit more using the phase-align extension.

Approach	$\bar{\Gamma}$ (dB)
Passive down-mix	25.85
Active down-mix	12.11
Proposed down-mix	4.92
Proposed down-mix (phase-align)	4.30

Table 1. Mean log-spectral distances of the considered down-mix approaches.

6. CONCLUSIONS

We proposed and evaluated a new down-mix approach based on the suppression of coherent signal components. Two existing down-mix approaches where used for comparison. The qualitative evaluation, using a white noise signal comprising an inter-channel delay, showed that comb-filter effects in the down-mix signal could be mitigated. Moreover, signals containing opposite-phase components keep their temporal and spectral fine structure. The proposed phase-align extension was able to reduce remaining cancelations in cases when the suppression works insufficiently. Finally, a quantitative comparison of the down-mix approaches showed that the proposed approach introduces the least amount of distortions.

7. REFERENCES

[1] J. Herre, K. Kjörling, J. Breebaart, C. Faller, S. Disch, H. Purnhagen, J. Koppens, J. Hilpert, J. Rödén, W. Oomen, K. Linzmeier, and K. S. Chong, "MPEG Surround – The ISO/MPEG Standard for Efficient and Compatible Multichannel Audio Coding," *Journal of the Audio Engineering Society*, vol. 56, no. 11, pp. 932–955, 2008.

[2] ITU-R BS.775-2, "Multichannel Stereophonic Sound System With And Without Accompanying Picture," 07/2006.

[3] R. Dressler. (05.08.2004) Dolby Surround Pro Logic II Decoder Principles of Operation. [Online]. Available: http://www.dolby.com/uploadedFiles/Assets/US/Doc/Professional/209_Dolby_Surround_Pro_Logic_II_Decoder_Principles_of_Operation.pdf

[4] F. Baumgarte, C. Faller, and P. Kroon, "Audio Coder Enhancement using Scalable Binaural Cue Coding with Equalized Mixing," in *116th Convention of the AES*, Berlin, Germany, 2004.

[5] G. Stoll, J. Groh, M. Link, J. Deigmöller, B. Runow, M. Keil, R. Stoll, M. Stoll, and C. Stoll, "Method for Generating a Downward-Compatible Sound Format," US Patent US 2012/0 014 526, 2012.

[6] M. Kim, E. Oh, and H. Shim, "Stereo audio coding improved by phase parameters," in *129th Convention of the AES*, San Francisco, 2010.

[7] Samsudin, E. Kurniawati, Ng Boon Poh, F. Sattar, and S. George, "A Stereo to Mono Dowmixing Scheme for MPEG-4 Parametric Stereo Encoder," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 5, 2006, pp. 529–532.

[8] W. Wu, L. Miao, Y. Lang, and D. Virette, "Parametric Stereo Coding Scheme with a New Downmix Method and Whole Band Inter Channel Time/Phase Differences," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 556–560.

[9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.