Parameter Domain Loudness Estimation in Parametric Audio Object Coding

Jouni Paulus
Fraunhofer Institute for Integrated Circuits IIS, and
International Audio Laboratories Erlangen
Erlangen, Germany
jouni.paulus@iis.fraunhofer.de

***Abstract*—Parametric audio object coding employs principles of informed source separation for obtaining object reconstructions from the mixture signal used in the transport enabling flexible output signal rendering into output scenes unknown at the encoder. Information of the object level in the rendered output is important for loudness and dynamic range control applications, e.g., in broadcast. This paper proposes a method for estimating the object level in an arbitrary output scene based on the downmix signal level that is then projected through the combined un-mixing and rendering matrix. This avoids explicit reconstruction of the objects only for the level estimation offering computational complexity savings. In the evaluations, the proposed method shows a high estimation accuracy with a root-mean squared error of 0.26 LUFS (loudness units relative to full scale) compared to 3.7 LUFS of the baseline with object reconstructions.**

***Index Terms*—object-based audio, parametric audio object coding, dynamic range control, loudness, spatial audio object coding, DRC, SAOC**

I. Introduction

In perceptual audio coding, a model of the human auditory system is used to reduce the amount of data needed for encoding so that the impact on perceptual audio quality is minimal. In many audio coding schemes used today, perceptual audio coding is used to encode the channel-based signals, each containing a mixture of audio objects ready for playback. Recently, more focus has been given to object-based audio coding, in which the individual audio objects are encoded and combined into channel signals only at the rendering stage of the decoding. The benefit of object-based audio coding is that it allows for greater flexibility in rendering, thus enabling many interactivity applications and playback over various loudspeaker setups. A prominent example of an audio coding system supporting both channel- and object-based coding is ISO/MPEG-H 3D Audio [1], [2], recently deployed in broadcast applications.

In some applications the storage or bandwidth constraints do not allow for transportation of the individual audio objects and parametric audio object coding [3] is used instead. Parametric audio object coding is an application of informed source separation: a number of audio objects are downmixed into a normally smaller number of transport channel signals, and parametric side information for decoder-side object separation is computed and encoded along the mixture channel. The total encoded size of the mixture and side information combined is less than the encoded size of the individual objects. The decoder uses the side information and allows for access to the audio objects in the mixture, either for manipulating the mixture (e.g., in MPEG-D Spatial Audio Object Coding (SAOC) [4]–[7]) or up-mixing the mixture for playback with a larger number of output channels (e.g., MPEG-H SAOC 3D [8]).

Dynamic Range Control (DRC) is a term referring to various audio processing methods modifying the dynamic range of an audio signal [9]. Examples of DRC include dynamic range compression reducing variation in the dynamic range, dynamic range expansion operating as the opposite of the compression, and a ducker controlling the level of a signal depending on the level of another signal. A prominent application for DRC in broadcast is controlling the overall signal loudness in order to comply with the recommendations, such as EBU-R128 [10]. The DRC processing is implemented as time-varying gains applied on the signals, either in a broadband or a sub-band manner. Many modern audio coding schemes support DRC metadata that allows the processing in the decoder to be applied depending on the user settings or other control information. The DRC metadata can be either the time-varying gain to be applied or a description of how the gains can be computed from the signals. This latter alternative, referred to as parametric DRC, requires less data to be transmitted, but requires information on the object level to be available either as an additional side information or computed from the object signals. If the playback system allows interactivity and content personalization, it is difficult or even impossible to transmit the object level information for all possible output rendering scenarios. The effect of user-interactivity on overall loudness in the case of MPEG-D SAOC-DE was discussed in [11]. Decoding the individual object signals from the mixture only for loudness estimation is computationally inefficient. This paper proposes a method for combining parametric audio object coding and parametric DRC in a computationally efficient way where the object loudness estimation and DRC processing take place mainly in the parametric representation domain.

The International Audio Laboratories Erlangen is a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg and Fraunhofer IIS.

ISBN 978-90-827970-1-5

II. Background

The method proposed in this paper can be applied in various parametric audio coding methods, but the following discussion uses the principles of the codecs of the SAOC family (MPEG-D SAOC [4], [5], MPEG-D SAOC-DE [6], [7], and MPEG-H SAOC 3D [1], [2]) as an example.

A. Spatial Audio Object Coding (SAOC)

A number of input audio signals $s_m, 1 \leq m \leq M$ are given. Each of these signals contains a single audio channel. If an audio object consist of multiple channels, it is possible to indicate that those channels belong to one object. The audio signals are transformed into a time/frequency representation using a hybrid-QMF filter bank (a quadrature mirror filter bank with an additional oddly-modulated Nyquist filter bank to increase the frequency resolution of the lowest bands). The output of the filter bank is complex-valued samples with a reduced temporal resolution (one sample per each 64 time-domain samples), but with an access to individual frequency bands (including the Nyquist filtering stage: 71 bands). These bands are grouped into a lower number of non-uniformly spaced parameter bands (up to 28 bands). Additionally, a number of consecutive samples (often 16 or 32) are grouped into frames, and it is assumed that the signals are stationary within these frames. In the following, the description focuses on one parameter band, but the operations are performed in an identical way for each band. The complex-valued samples of all M input signals are denoted with $\boldsymbol{S}$ of size $M \times L$, where L is the length of the frame. This tiling reduces resolution on both time and frequency axes and allows for reduction in the amount of side information needed. Parametric side information is extracted from the input signals in these tiles for the purpose of being able to reconstruct the objects at the decoder.

The most relevant side information parameters used in SAOC are Object Level Difference (OLD) describing the energy of signal relative to the maximum energy within a parameter tile and Inter-Object Correlation (IOC) describing the normalized correlation coefficient between two signals. Together these parameterize the covariance matrix $\boldsymbol{E}$ of size $M \times M$ between the input signals. Note that due to the normalization of the OLD parameter the absolute level of the signals is not contained in the parameterization.

In addition to the parametric information, an N-channel downmix signal is created from the input object signals using a linear instantaneous combination of

$$\boldsymbol{X} = \boldsymbol{D}\boldsymbol{S}, \tag{1}$$

where the downmixing matrix $\boldsymbol{D}$ of size $N \times M$ contains the mixing weights. In the case of SAOC-DE, the downmixing weights are designed such that the resulting signal is a meaningful signal, which can be listened to as such. In SAOC 3D this requirement is not present, but it is still a possibility.

The side information and the downmix signal are transmitted or stored.

The decoding uses a similar time/frequency transform on the downmix signal and then uses the side information for computing an un-mixing matrix $\boldsymbol{G}$ of size $M \times N$ in each parameter time/frequency-tile. The un-mixing matrix can be used to create parametric reconstructions $\tilde{\boldsymbol{S}}$ of the original input signal with

$$\tilde{\boldsymbol{S}} = \boldsymbol{G}\boldsymbol{X}. \tag{2}$$

The object un-mixing matrix $\boldsymbol{G}$ can be computed from the object covariance matrix $\boldsymbol{E}$ and the downmixing matrix $\boldsymbol{D}$ with

$$\boldsymbol{G} = \boldsymbol{E}\boldsymbol{D}^*(\boldsymbol{D}\boldsymbol{E}\boldsymbol{D}^*)^\dagger, \tag{3}$$

where $x^\dagger$ denotes a regularized matrix pseudo-inverse.

A rendering matrix $\boldsymbol{R}$ of size $N_{ren} \times M$ is then used to combine the object reconstructions to an output scene with

$$\boldsymbol{Y} = \boldsymbol{R}\tilde{\boldsymbol{S}}. \tag{4}$$

The un-mixing and rendering stages are combined into one operation for computational efficiency, resulting in

$$\boldsymbol{Y} = \boldsymbol{R}\boldsymbol{G}\boldsymbol{X}. \tag{5}$$

The rendered output scene is then transformed back into time-domain using a synthesis filter bank.

A problem of (5) is that the un-mixing matrix $\boldsymbol{G}$ provides a minimum mean squared error solution for the un-mixing task, meaning that this solution has in many cases some estimation errors. Due to these errors, the estimated energies of all objects are slightly lower than they should be in reality: the lower the energy of an object in the mixture is, the higher the relative estimation error is. Some parametric audio object coding systems address this by creating the final output signal as a combination of the "dry" path signal of (5) and an additional "wet" path signal adding the missing energy or decorrelation. In SAOC 3D, the output is obtained with

$$\boldsymbol{Y} = \boldsymbol{P}\begin{bmatrix}\boldsymbol{Y}_{dry}\\ \boldsymbol{Y}_{wet},\end{bmatrix} \tag{6}$$

where $\boldsymbol{P}$ is a mixing matrix defined as concatenation of two sub mixing matrices

$$\boldsymbol{P} = \begin{bmatrix}\boldsymbol{P}_{dry} & \boldsymbol{P}_{wet}\end{bmatrix} \tag{7}$$

contents of which depend on the decoding mode and are given in [2]. The dry and wet signals in (6) are defined as

$$\boldsymbol{Y}_{dry} = \boldsymbol{R}\boldsymbol{G}\boldsymbol{X}, \tag{8}$$

being identical to (5), and

$$\boldsymbol{Y}_{wet} = \boldsymbol{M}_{post}\boldsymbol{X}_d. \tag{9}$$

The decorrelator output signals $\boldsymbol{X}_d$ are obtained with

$$\boldsymbol{X}_d = decorr(\boldsymbol{M}_{pre}\boldsymbol{Y}_{dry}), \tag{10}$$

and $\boldsymbol{M}_{pre}$ and $\boldsymbol{M}_{post}$ are decorrelation pre- and post-processing matrices defined in [2]. The decorrelation function $decorr(\cdot)$ provides signals with similar spectral energy distribution and similar overall energy as the input signals, but which are statistically mutually independent and independent from the inputs.

ISBN 978-90-827970-1-5

B. Dynamic Range Control and Loudness Estimation

The principle of a DRC processing has been the same from the early description of an automatic analogue dynamic range control system of [12] and the first digital DRC implementation of [13]: estimate the level of the input, determine an output gain based on the desired processing characteristics, and apply the gain to a signal after some temporal smoothing. For a more complete description, the reader may refer to [9], [14], [15]. The level estimation is traditionally based on signal peak or RMS levels. The first is more appropriate for limiting applications, while the latter "reflects our perception of loudness of the signal" [14]. A more complex estimate of signal loudness can also be used, as done, e.g., in [16].

The input signal level is compared to a DRC mapping curve returning a gain that should be applied to the signal. Different input/output mapping curves define different DRC characteristics. Having a different DRC curve, e.g., depending on the playback environment (high-end home theater vs. ear buds of a mobile device in a noisy bus), allows a better end-user experience. One possibility is to generate the DRC gains at the encoder side, transport them as additional metadata, and apply them in the decoder. This requires some additional transport capacity. An alternative is to generate the DRC gains at the decoder side, in which case only a parametric description of the input/output mapping is needed, to be either transported as a part of the metadata or provided entirely from the decoding device.

The problem of parametric DRC in conjunction with parametric audio codecs is obtaining the level estimates of the objects. The level information could be included as additional metadata, but this requires again transport capacity and it cannot account for interactivity. The second alternative is to decode the objects from the coded representation and to perform normal level estimation. This has a significant additional computational cost since the full object reconstructions are rarely needed for moderate interactive semantic manipulation of the mixture signal, e.g., dialogue enhancement. The proposed method addresses both these aspects and is able to provide level estimates without additional metadata and a lower computational complexity than performing a full object reconstruction.

Loudness refers to measure that approximates the perceived level of a sound by humans. There are various computational ways for this approximation, but this document focuses on a definition similar to ITU-R BS.1770-4 [17] with the following main steps:

1) The input signals are weighted with a K-weighting filter (a combination of a high-pass filter and a shelving filter at high frequencies).
2) The energy of the signals is computed in frames of 400 ms with 300 ms overlap.
3) A weighted sum of the signal energy over the channels is computed. The per-channel weight depends on the assumed physical location of the channel.
4) The energies are transformed into the decibel domain.
5) Optional gating of frames with low energy, removing them from the computation of temporal average.

The result is the loudness of the signal in LUFS (loudness units relative to full scale). The reason for focusing on BS.1770 is that it is used in recommendations and requirements in broadcast applications, e.g., EBU-R128 [10].

III. Proposed Method

Fig. 1 shows a block diagram of a parametric audio object coding system including the proposed method. The proposed changes are added into the decoding process and include the following steps:

1) The downmix signal energy level is estimated in the form of a covariance matrix.
2) This energy is projected using the object un-mixing and rendering information into per-object level estimates in the rendered output.
3) The object levels are used as an input to a DRC algorithm producing one or more DRC gains to be applied to the objects, e.g., ducking of an object depending on the level of another object or compressing the dynamic range of the dialogue content for improved intelligibility.
4) The obtained gains are used to modify the rendering matrix before this is combined with the un-mixing matrix and applied to the downmix signal for obtaining the final output.

These steps will be discussed in more detail in the following.

A. Downmix Level Estimation

The downmix signal $x_n, 1 \leq n \leq N$ with N channels is transformed into a time/frequency representation and split into frames, similar to the processing performed in the encoder. The downmix signal channels in the frequency band F are denoted with $\boldsymbol{X}_f, f \in F$. In each frame, the un-normalized covariance matrix of the downmix signal channels

$$\boldsymbol{E}_{dmx} = \begin{bmatrix} c_{1,1}, & \ldots, & c_{1,N} \\ \vdots, & \ddots, & \vdots \\ c_{N,1}, & \ldots, & c_{N,N} \end{bmatrix} \tag{11}$$

is computed. The matrix elements are obtained with

$$c_{i,j} = \Re\left\{ \frac{1}{L} \sum_{f \in F} \sum_{k=1}^{L} \boldsymbol{X}_f(i,k) \boldsymbol{X}_f^*(j,k) \right\}, \tag{12}$$

where L is the length of the frame, $\Re\{\cdot\}$ returns the real part of the complex number, and x^* is the complex conjugate of x. A simpler, but more inaccurate alternative to this is to compute only the energies of each of the downmix signals, i.e., only the main diagonal entries $c_{n,n}$.

If the signal loudness level is of interest instead of the energy level, the frequency-dependent K-weighting of BS.1770 can be easily accounted for by including the weighting as $W(f)$ into (12) resulting in

$$c_{i,j} = \Re\left\{ \frac{1}{L} \sum_{f \in F} W^2(f) \sum_{k=1}^{L} \boldsymbol{X}_f(i,k) \boldsymbol{X}_f^*(j,k) \right\}. \tag{13}$$

ISBN 978-90-827970-1-5

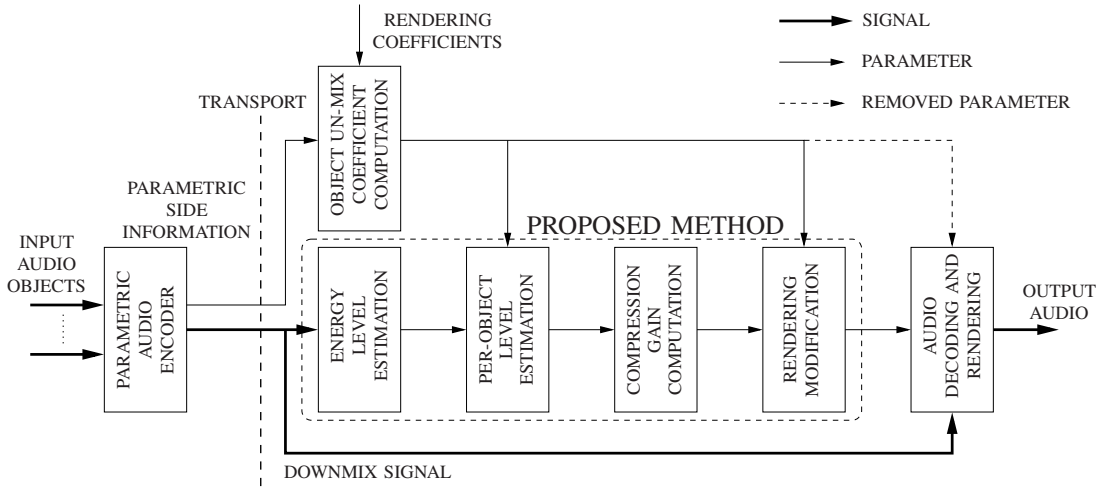

Figure 1. A block diagram of a parametric audio object coding system including the proposed method inside the dashed region. The proposed method estimates the downmix signal level and projects it using the un-mixing and rendering information into per-object level estimates. These are used in a parametric DRC algorithm for obtaining DRC gains to modify the rendering. This modified rendering replaces the original one, denoted also with a dashed arrow.

B. Per-object Level Estimation

The downmix signal covariance matrix can be projected into a parametric approximation of the covariance matrix of the channels of a rendered object with

$$\boldsymbol{E}_{obj} = \boldsymbol{R}_{obj}\boldsymbol{G}\boldsymbol{E}_{dmx}\boldsymbol{G}^{*}\boldsymbol{R}_{obj}^{*}, \tag{14}$$

where the per-object rendering matrix $\boldsymbol{R}_{obj}$ is obtained from the full rendering matrix $\boldsymbol{R}$ by setting the coefficients not associated with the target object obj into zero, i.e.,

$$\boldsymbol{R}_{obj}(i,j) = \begin{cases} \boldsymbol{R}(i,j), & \text{if } j \in obj, \\ 0, & \text{otherwise} \end{cases}. \tag{15}$$

The main diagonal entries of the matrix $\boldsymbol{E}_{obj}$ provide approximations of the target object level in each of the output channels.

In the case of SAOC 3D output of (6), an approximation of the target object covariance in the output channels can be obtained with

$$\begin{aligned}\boldsymbol{E}_{obj} &= \boldsymbol{P}_{dry}\boldsymbol{R}_{obj}\boldsymbol{G}\boldsymbol{E}_{dmx}\boldsymbol{G}^{*}\boldsymbol{R}_{obj}^{*}\boldsymbol{P}_{dry}^{*} \\ &\quad + \boldsymbol{P}_{wet}\boldsymbol{M}_{post}diag\left(\boldsymbol{M}_{pre}\boldsymbol{R}_{obj}\boldsymbol{G}\boldsymbol{E}_{dmx}\boldsymbol{G}^{*}\boldsymbol{R}_{obj}^{*}\boldsymbol{M}_{pre}^{*}\right)\boldsymbol{M}_{post}^{*}\boldsymbol{P}_{wet}^{*},\end{aligned} \tag{16}$$

where $diag(\boldsymbol{X})$ sets all matrix entries not on the main diagonal to 0. One should note, that (16) assumes the two signal paths to be statistically independent, which is not always the case in reality.

An alternative to this assumes that the parametric object covariance matrix $\boldsymbol{E}$ is available and computes a parametric covariance matrix of the rendered target object with

$$\boldsymbol{E}_{obj}^{ref} = \boldsymbol{R}_{obj}\boldsymbol{E}\boldsymbol{R}_{obj}^{*}. \tag{17}$$

The parametric estimate of the covariance matrix of the parametric output signal is

$$\boldsymbol{E}_{obj}^{param} = \boldsymbol{R}_{obj}\boldsymbol{G}\boldsymbol{D}\boldsymbol{E}\boldsymbol{D}^{*}\boldsymbol{G}^{*}\boldsymbol{R}_{obj}^{*}. \tag{18}$$

We can compute correction scaling factors for the estimated output energies from these with

$$g_i = \boldsymbol{E}_{obj}^{ref}(i,i)/\boldsymbol{E}_{obj}^{param}(i,i). \tag{19}$$

These gains can then be used to scale the output energy estimates of (14) with

$$\boldsymbol{E}'_{obj}(i,i) = g_i\boldsymbol{E}_{obj}(i,i). \tag{20}$$

C. Loudness Estimation

An approximation of the loudness level similar to BS.1770 can now be obtained by integrating the level estimates over target frequency range F, temporal range T, and the channels i, and transforming the result into decibel domain with

$$L_{obj} = c + 10\log_{10}\sum_{i,f\in F,t\in T}\boldsymbol{E}_{obj}(i,i)(t,f), \tag{21}$$

where $c = -0.691$ is an offset constant defined in [17]. Please note, that the loudness is obtained from the downmix signal level through a parametric projection without time-domain object reconstruction.

D. Parametric DRC

Having estimates of the loudness levels of one or more objects in the mixture signal, dynamic range control processing can be performed. The object levels are used as the input to the DRC algorithm determining a gain to be applied to an object. The exact way of determining the gain depends on the design of the DRC algorithm and is beyond the scope of this paper. Some alternatives are described in [9], [15], [18].

As suggested in [18], applying the gains to the individual object signals can be implemented in an efficient way by modifying the rendering matrix $\boldsymbol{R}$ before computing the output with (5). For example, applying the gain g_{obj} to the object obj can be achieved by the modification of

$$\boldsymbol{R}_{DRC}(i,j) = \begin{cases} g_{obj}\boldsymbol{R}(i,j), & \text{if } j \in obj, \\ \boldsymbol{R}(i,j), & \text{otherwise} \end{cases}, \tag{22}$$

ISBN 978-90-827970-1-5

Table I
ROOT MEAN SQUARED ERROR (IN LUFS) OF PER-FRAME LEVEL ESTIMATES.

	baseline	**plain**	**complete**
FGO	4.5	1.7	0.25
BGO	2.6	1.3	0.28
mean	3.7	1.5	0.26

and using this in the rendering of the output with (5) or (6).

IV. RESULTS

The proposed loudness estimation method is evaluated in an SAOC framework that uses a hybrid-QMF filter bank, 28 parameter bands, a frame length of 32 QMF-slots, and object parameterization with IOCs and OLDs, both quantized with fine quantization. Using only the dry path reconstruction of (5), the functionality is close to MPEG-D SAOC-DE [6]. The estimation is tested using 11 stereo items of 10 s in length, each consisting of a stereo background and a single-channel speech foreground in the center of the audio scene. The simulated rendering attenuates the background content by 6 dB compared to the downmix signal. The ground truth is the level estimated from the original component signals after applying the rendering. The *baseline* estimation reconstructs and renders the objects in signal-domain and estimates the level from these signals. The *plain* proposed method uses the level estimates (14), and *complete* proposed method uses the corrected estimates of (20).

Each method provides a level estimate for both the background and foreground objects in each time frame. The object frames with less than -50 LUFS absolute level are excluded from the evaluation. Table I shows the root mean squared error of the three tested methods. It can be seen that the baseline method using object reconstructions has quite a large estimation error, in addition to the higher computational complexity. The plain version of the proposed method performs with a slightly higher accuracy, and it can be applied in methods in which only the un-mixing matrix $\boldsymbol{G}$ is available and the per-object normalized covariance matrix $\boldsymbol{E}$ is not. Including the correction term (19) in the estimation process yields level estimates with an accuracy that should be good enough for most practical applications.

V. CONCLUSIONS

This paper has described a method for estimating the loudness of an object or group of objects at the decoder of a parametric audio object coding system without a full un-mixing and rendering of the object from the transport mixture. The proposed method does not need additional side information, but can be implemented with the existing parametric un-mixing information. The proposed method is also able to consider the decoder-side signal personalization and interactivity in the estimation. In the evaluations, the proposed method shows a high estimation accuracy with a root-mean squared error of 0.26 LUFS compared to 3.7 LUFS of the baseline method with object reconstructions, while having lower computational complexity.

REFERENCES

[1] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H 3D Audio - The new standard for coding of immersive spatial audio," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 770–779, Aug. 2015.

[2] International Organization for Standardization, *ISO/IEC DIS 23008-3 Information technology – High efficiency coding and media delivery in heterogeneous environments – Part 3: 3D Audio*, Geneva, Switzerland, 2015.

[3] J. Herre and L. Terentiv, "Parametric coding of audio objects: technology, performance and opportunities," in *Proc. of Audio Engineering Society 42nd International Conference on Semantic Audio*, Ilmenau, Germany, Mar. 2011.

[4] International Organization for Standardization, *ISO/IEC 23003-2:2010 Information technology – MPEG audio technologies – Part 2: Spatial Audio Object Coding (SAOC)*, Geneva, Switzerland, 2010.

[5] J. Herre, H. Purnhagen, J. Koppens, O. Hellmuth, J. Engdegård, J. Hilpert, L. Villemoes, L. Terentiv, C. Falch, A. Hölzer, M. L. Valero, B. Resch, H. Mundt, and H.-O. Oh, "MPEG Spatial Audio Object Coding - the ISO/MPEG standard for efficient coding of interactive audio scenes," *Journal of the Audio Engineering Society*, vol. 60, no. 9, pp. 655–673, Sep. 2012.

[6] International Organization for Standardization, *ISO/IEC 23003-2:2010/Amd.3 Information technology – MPEG audio technologies – Part 3: Spatial Audio Object Coding (SAOC) Amendment 3: Dialogue Enhancement*, Geneva, Switzerland, 2014.

[7] J. Paulus, J. Herre, A. Murtaza, L. Terentiv, H. Fuchs, S. Disch, and F. Ridderbusch, "MPEG-D Spatial Audio Object Coding for Dialogue Enhancement (SAOC-DE)," in *Proc. of 138th Audio Engineering Society Convention*, Warsaw, Poland, May 2015.

[8] A. Murtaza, J. Herre, J. Paulus, L. Terentiv, H. Fuchs, and S. Disch, "ISO/MPEG-H 3D Audio: SAOC 3D decoding and rendering," in *Proc. of 139th Audio Engineering Society Convention*, New York, New York, USA, Oct. 2015.

[9] U. Zölzer, *Digital Audio Signal Processing*. John Wiley & Sons, Ltd., 2008.

[10] European Broadcasting Union (EBU), *Recommendation R128 - Loudness normalisation and permitted maximum level of audio signals*, Geneva, Switzerland, Jun. 2014.

[11] J. Paulus, "Perceptual loudness compensation in interactive object-based audio coding systems," in *Proc. of 23rd European Signal Processing Conference*, Nice, France, Aug. 2015, pp. 579–583.

[12] B. A. Blesser, "Audio dynamic range compression for minimum perceived distortion," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 1, pp. 22–32, Mar. 1963.

[13] G. W. McNally, "Dynamic range control of digital audio signals," *Journal of the Audio Engineering Society*, vol. 32, no. 5, pp. 316–327, May 1984.

[14] J. C. Schmidt and J. C. Rutledge, "Multichannel dynamic range compression for music signals," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Atlanta, Georgia, USA, May 1996, pp. 1013–1016.

[15] P. Dutilleux and U. Zölzer, "Nonlinear processing," in *DAFX: Digital Audio Effects*, U. Zölzer, Ed. John Wiley & Sons, Ltd., 2002, ch. 7.

[16] J.-M. Jot, B. Smith, and J. Thompson, "Dialog control and enhancement in object-based audio systems," in *Proc. of 139th Audio Engineering Society Convention*, New York, New York, USA, Oct. 2015.

[17] International Telecommunication Union, *ITU-R BS.1770-4: Algorithms to measure audio programme loudness and true-peak audio level*, Geneva, Switzerland, Oct. 2015.

[18] International Organization for Standardization, *ISO/IEC 23003-4:2015 Information technology – MPEG audio technologies – Part 4: Dynamic Range Control*, Geneva, Switzerland, 2015.

ISBN 978-90-827970-1-5