PERCEPTUAL LOUDNESS COMPENSATION IN INTERACTIVE OBJECT-BASED AUDIO CODING SYSTEMS

Jouni Paulus

Fraunhofer IIS, Erlangen, Germany

ABSTRACT

Changing the rendering through interactivity in object-based audio coding may change the overall signal loudness. This paper proposes a method for estimating the change in the overall loudness using loudness information of the partial mixes and the rendering description. The method has been designed for a dialogue enhancement application scenario. The results of the method are compared with reference values from measurements, and the results match well with the mean absolute error of 0.11 LU. A subjective listening test is conducted for studying the amount of amplification applied by the test participants on a probe signal simulating the result of an interactive rendering when comparing it with a reference signal of the default mix. The average level adjustment reflects the change in the signal loudness through the modification.

Index Terms— audio, loudness, object-based coding, listening test

1. INTRODUCTION

Perceptual audio coding reduces the amount of information to be transmitted by utilizing a model of the human auditory system as the receiver. Highly successful examples of perceptual audio coding technologies include, e.g., the ISO/MPEG standards for MP3 and Advanced Audio Coding (AAC) [1]. Lately, there has been an increasing interest in object-based audio coding technologies in which the coding does not operate anymore only on the channel signals, but also on the semantic audio objects contained in these signals. This difference can be exemplified by considering coding a musical piece. The channel-based methods concentrate on the mixture signals to be played over the loudspeakers while the object-based methods focus on the individual instruments of the mix. Examples of object-based coding technologies include ISO/MPEG-D Spatial Audio Object Coding (SAOC) [2] and the object-based part of ISO/MPEG-H 3D Audio [3].

An object-based transmission may often include a description of a default output scene, i.e., the way the objects should be rendered into the output channels without any user interaction. This description may be transmitted in the form of object metadata detailing the geometrical locations of the objects as is done in the case of MPEG-H 3D Audio, or in the form of the semantically meaningful downmix signal or rendering presets as may be done with MPEG-D SAOC. In some applications, the end-user can change the rendering description from the default one. By changing the rendering, for example, by adjusting the levels of the objects, the perceptual loudness of the output signal may be affected.

This paper studies the loudness change in the specific application of dialogue enhancement offering a limited interactivity for the end-user. Section 2 describes the technical background in more detail and provides a brief overview of a method for estimating the loudness of signals. Section 3 derives a method for estimating the change in the loudness from the default output scene through interactivity. The method relies on loudness information from mixes of subsets of the objects (in the remainder of the paper these are referred to as partial downmixes). Section 4 describes tests for evaluating the method for the change estimation. These tests include objective measurements as well as a subjective adjustment experiment for comparing the estimated value to the compensation value preferred by listeners. Finally, Section 5 provides the conclusions of this paper.

2. BACKGROUND

The results of this paper can be applied regardless of the underlying mechanism for transmitting the audio object data, be it delivery of discrete object signals and mixing information, pre-mixed content stems, or a downmix and object side information. The concrete application example considered here employs the last of these principles and is known as Spatial Audio Object Coding for Dialogue Enhancement (SAOC-DE) [4]. This will be discussed next. The final part of this section describes the method ITU-R BS.1770-3 for estimating the perceptual loudness of a signal.

2.1. Spatial Audio Object Coding

An SAOC [2] encoder receives the original audio objects $\boldsymbol{S}$ (number of objects × number of samples) and gains $\boldsymbol{D}$ (number of downmix channels × number of objects) for mixing the objects into a downmix signal with

$$\boldsymbol{X}_{dmx} = \boldsymbol{DS}. \tag{1}$$

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

The encoder parametrizes the objects with their relative energy levels and optionally correlations between the objects in time-frequency tiles. This side information is sent in addition to the downmix signal. It is possible to apply a perceptual audio codec to the downmix signal and embed the side information into the created bitstream. A significant property of SAOC is that the downmix signal is a meaningful signal and can be listened to. This enables backward compatibility in which the legacy decoders incapable of decoding the SAOC information can still output the downmix signal. An SAOC decoder uses the provided side information, conceptually reconstructs the objects $\tilde{\boldsymbol{S}} \approx \boldsymbol{S}$, and then renders them according to the provided rendering gains $\boldsymbol{R}$ (number of output channels × number of objects) as

$$\boldsymbol{X}_{out} = \boldsymbol{R}\tilde{\boldsymbol{S}} \approx \boldsymbol{R}\boldsymbol{S}. \tag{2}$$

The rendering gain matrix $\boldsymbol{R}$ describes the way the objects should be assigned to the output channels and it is often specified at the decoder side, e.g., by the end-user.

2.2. Dialogue Enhancement

Recently, an amendment of the original SAOC standard was published customizing the capabilities of the technology for dialogue enhancement in broadcasting applications and named accordingly as SAOC Dialogue Enhancement profile (SAOC-DE) [4]. The main technological changes from the original standard include increasing the maximum number of downmix channels beyond two and restricting the rendering to be an in-place modification with equal number of downmix and output channels, the latter change being relevant for this paper.

The input objects $\boldsymbol{S}$ are grouped into two meta-objects of a foreground object (FGO) $\boldsymbol{S}_{FGO}$, e.g., the commentator, and a background object (BGO) $\boldsymbol{S}_{BGO}$, e.g., the stadium noises in a sports program. The downmixing matrix can be divided similarly into parts corresponding to the FGO $\boldsymbol{D}_{FGO}$ and the BGO $\boldsymbol{D}_{BGO}$. The downmix signal $\boldsymbol{X}_{dmx}$ is a sum of the partial downmixes (or stems) $\boldsymbol{X}_{FGO}$ and $\boldsymbol{X}_{BGO}$:

$$\boldsymbol{X}_{dmx} = \boldsymbol{X}_{FGO} + \boldsymbol{X}_{BGO} = \boldsymbol{D}_{FGO}\boldsymbol{S}_{FGO} + \boldsymbol{D}_{BGO}\boldsymbol{S}_{BGO}. \tag{3}$$

The rendering is limited in SAOC-DE so that only the relative levels of the meta-object downmixes $\boldsymbol{X}_{FGO}$ and $\boldsymbol{X}_{BGO}$ can be changed, and the output is

$$\begin{aligned} \boldsymbol{X}_{out} &= g_{FGO}\tilde{\boldsymbol{X}}_{FGO} + g_{BGO}\tilde{\boldsymbol{X}}_{BGO} && (4) \\ &\approx g_{FGO}\boldsymbol{X}_{FGO} + g_{BGO}\boldsymbol{X}_{BGO}, && (5) \end{aligned}$$

where $\tilde{\boldsymbol{X}}_{FGO}$ and $\tilde{\boldsymbol{X}}_{BGO}$ are reconstructions of the partial downmixes, and g_{FGO} and g_{BGO} are the rendering gains for the meta-objects. Setting $g_{FGO} > g_{BGO}$ amplifies the foreground relative to the background and symmetrically $g_{FGO} < g_{BGO}$ attenuates the foreground relative to the background. In SAOC-DE, these gains are determined from a single user input gain m_G with

$$\begin{aligned} g_{FGO} &= \min\left(1, m_G\right), \text{ and} && (6) \\ g_{BGO} &= \min\left(1, m_G^{-1}\right). && (7) \end{aligned}$$

A gain indicating FGO amplification with $m_G > 1$ is mapped into an attenuation of the BGO. Because of the mapping any input gain $m_G \neq 1$ will result into attenuation of one of the meta-objects and to a decrease in the overall signal level compared to the default mix (i.e., the downmix).

2.3. Loudness in broadcasting

In a broadcast scenario, one major source of annoyance for the end-users has been the varying levels between channels or between the main program and the advertisements. Recently, recommendations and regulations have been posted both in Europe [5] and USA [6] requiring loudness normalization of the programs. For example, the EBU R 128 [5] recommends that the average loudness of a program should be normalized to the target level of -23.0 LUFS measured using the method defined in ITU-R BS.1770-3 [7].

In interactive applications, such as dialogue enhancement, changing the mixing in the decoder affects also the overall loudness. Switching channels between a DE-processed and an un-processed one may re-introduce loudness jumps. Even though these jumps can be expected to be within the “Comfort zone” of -5.4 to $+2.4$ dB [8], it would be beneficial to minimize the changes and be closer to the ±0.5 LU deviation range allowed by EBU R 128 [5].

3. ESTIMATING LOUDNESS CHANGE

The discussion in this paper is based on the ITU-R BS.1770-3 method for estimating audio signal loudness [7], but may be applicable also with other methods. The BS.1770-3 method operates by applying a “K”-weighting (a high-pass and a shelving filter) on the signals, calculating the energy of the signals in short frames, and representing the values on a logarithmic scale. There is an optional gating operation on the per-frame loudness values omitting the low-energy frames from the calculation of average (“integrated”) loudness and it is enabled in EBU R 128 [5].

The loudness of a measurement interval is

$$L = c + 10\log_{10} E, \tag{8}$$

where $c = -0.691$ is a constant, and E is the sum of the weighted per-channel energies. The result of the calculation is in the scale of Loudness Units (LU) behaving similar to the decibel scale.

Let us denote the K-weighted energies of the partial downmixes $\boldsymbol{X}_{FGO}$ and $\boldsymbol{X}_{BGO}$ by E_{FGO} and E_{BGO}. Assuming that the FGO and BGO are independent, the K-weighted energy E_{dmx} of the downmix signal is the sum of their energies:

$$E_{dmx} = E_{FGO} + E_{BGO}. \tag{9}$$

From this follows that the loudness of the downmix is

$$L_{dmx} = 10\log_{10}\left(10^{L_{FGO}/10} + 10^{L_{BGO}/10}\right), \tag{10}$$

where L_{FGO} and L_{BGO} are the loudness values of the partial downmixes of FGO and BGO.

When the output signal $\boldsymbol{X}_{out}$ is obtained using a rendering setting of (4), i.e., applying gains g_{FGO} and g_{BGO} to the (reconstructions of the) partial downmixes of FGO and BGO, the loudness of the output can be estimated as

$$L_{out} = 10\log_{10}\left(g_{FGO}^2 10^{L_{FGO}/10} + g_{BGO}^2 10^{L_{BGO}/10}\right). \tag{11}$$

Comparing this with (10), the change in the loudness from the default mix to the rendered output is

$$\Delta L = 10\log_{10}\frac{g_{FGO}^2 10^{L_{FGO}/10} + g_{BGO}^2 10^{L_{BGO}/10}}{10^{L_{FGO}/10} + 10^{L_{BGO}/10}}. \tag{12}$$

This estimation requires the loudness values L_{FGO} and L_{BGO} of the partial downmixes to be available. In a broadcasting application, these may be estimated at the encoder side along with the program loudness and sent to the decoder. Compared to estimating the loudness from the decoder output signal, the advantages of the proposed method include it being computationally extremely light as it does not require an access to the output signals, and producing an accurate estimate without an integration delay. However, if gating is used or if the assumption of independence of the FGO and BGO does not hold, the estimate may deviate from the true value. Furthermore, time-varying signal characteristics may influence the perceptual relevance of the estimation result.

4. EVALUATION

For verifying the accuracy of the proposed method of (12) the values produced by it are compared with values estimated from signals in a small test. In addition to the objective evaluation, an adjustment test was conducted for collecting subjective data on the loudness change.

4.1. Test material

The test uses three items resembling sport broadcast content by consisting of a stereo background and a mono or a stereo dialogue object on top of that. The items are:

- “Football”: A football match program with a stereo background of stadium noises and a mono foreground commentator mixed to the middle of the stereo scene.

Item	L_{FGO} (LU)	L_{BGO} (LU)
Football	-16.9	-18.4
Formula	-20.0	-22.0
Wimbledon	-23.2	-22.6

Table 1. Partial downmix loudness values of the test items.

- “Formula”: A race track introduction clip with music, sound effects, and noises as the background and two commentators mixed into a stereo foreground.
- “Wimbledon”: A recording from a tennis match with stadium noise background and a dual-mono commentator foreground. The background is very loud in the beginning of the clip, but almost silent in the end.

All items are approximately 10 seconds in length, and the output channel configuration is stereo.

The test conditions for the computation methods are obtained by varying the FGO modification gain m_G from -20 dB to +20 dB in steps of 1 dB. The listening test uses a subset of the gains and includes the values {-18, -12, -6, 0, +6, +12, +18} dB. The 0 dB condition is included as a control point for assessing the loudness matching abilities of the test participants. For simulating the application scenario, the test items are processed with an SAOC-DE codec instead of using ideal mixes. The encoder uses 32-slot parameter frames, 28 parametric bands, and a 20 kbps residual signal with a 20-band bandwidth for each channel of the FGO. No perceptual coding is applied to the downmix signals.

4.2. Signal-based loudness change estimate

The loudness L_{dmx} of the downmix signals are estimated using the gated[1] measurement of ITU-R BS.1770-3 [7]. The loudness L_{out} of the SAOC decoder output is estimated similarly for each gain m_G, and the differences of these two measurements are shown with red dotted curve in Fig. 1. The quantization of the values is caused by the limited resolution (0.1 LU) of the implementation. It is worth noting that even with the extreme modifications up to ±20 dB, the maximum overall loudness changes are less than 4 LU.

4.3. Estimated loudness change

The precision of the proposed estimation method of (12) is tested by creating the partial downmix signals $\boldsymbol{X}_{FGO}$ and $\boldsymbol{X}_{BGO}$ and measuring their loudness values L_{FGO} and L_{BGO} according to ITU-R BS.1770-3 [7]. These values, shown in Table 1, are then inserted into (12) and the results are shown in Fig. 1 with the blue line. It can be seen that the results match the ones estimated from the signals very well: the mean absolute difference is 0.11 LU and the RMS-difference is 0.14 LU.

[1] In these items, the gated and un-gated values are identical.

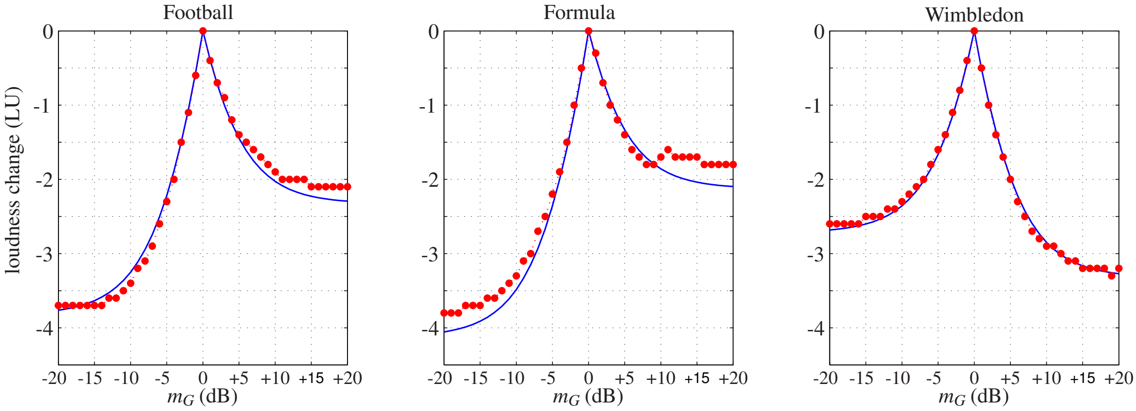

Fig. 1. The difference in the loudness between the downmix signal and the signal obtained by applying the FGO modification gain m_G. The red dots indicate the difference based on loudness estimates from the signals, and the blue line indicates the difference estimate produced by the proposed method.

4.4. Subjective adjustment test

The second part of the evaluations is a subjective listening test in which the participants needed to apply a gain on the SAOC-DE decoder output signal to match it with the default downmix signal as per the instructions. The motivation for this is to see how well the adjustments reflect the estimated loudness change. As the FGO in these items is speech, it is possible that the test subjects focus on the FGO level instead of the overall level. This would lead into smaller adjustments when the BGO has been attenuated.

The test participants were presented with a graphical user interface containing buttons for selecting either the reference or the probe signal to be played and a slider for adjusting the probe signal level. The signals were played back over headphones. The reference signal in the tests is the default downmix and the probe signal is the SAOC-DE decoder output.

In addition to the visual interface, a tactile interface was used. The listener could switch between reference and probe signals using the keyboard and adjust the level of the probe signal with a physical dial. The scale available was from -24 to $+24$ dB in 512 steps (0.0938 dB per step). When the adjustment was done, the listener could proceed to the next test item by pressing the dial downwards.

The items and conditions were presented in a random order and the starting level of the probe signal was randomly assigned to either end of the adjustment scale. It is worth noting that the user interface did not display any numerical values of the current level of the probe signal, nor had the slider any visible tick marks for visual anchoring. This was to ensure that the values set by the test participants were based on the acoustic information only.

The test task was defined as: *"You are sitting in front of your TV and zapping between two channels. Adjust the level of the probe-channel so that it is aesthetically optimal for your personal preference when zapping."* Direct instructions for loudness-matching the signals were omitted in order to get subjective data on actual behavior of the listeners.

4.5. Subjective test results

The test was taken by 11 participants. The mean adjustments with the 95% confidence intervals (using Student's t-distribution) are shown in Fig. 2. The plots contain also values derived from the estimated loudness changes shown earlier in Fig. 1 for a comparison. The following observations can be made:

- The values form a very rough V-shape and to correspond to the negatives of the estimated loudness change values.
- The subjective data is very noisy, especially in the conditions with large modifications in the mixing balance.
- The confidence interval of the 0 dB control condition overlaps with 0 dB adjustment in all items. This indicates that the test participants were able to match the loudness of the probe signal with the reference signal, providing confidence on their matching performance.
- In the "Formula" and "Football" items, the $m_G = +18$ dB condition deviates from the V-shape. This is not so drastic in the "Wimbledon" item. The differences may be partially explained by the temporal properties of the background (in the first two items it is rather static, but in the third item there are large level difference between the beginning and the end of the clip) and with the differing FGO/BGO balance in the default mix.
- The estimated loudness differences exceed the subjective adjustments. This suggests that the listeners apply conservative adjustments instead of attempting to match the two signals in overall loudness.

Discussions with the test participants revealed that they had different targets (focusing on the overall, the FGO, or the BGO loudness), which may explain the variation in the results. Some participants mixed the targets and matched the overall loudness when the FGO was attenuated and the FGO loudness when the BGO was attenuated. Both these make sense: the first is in line with the technical requirements of [5], while in the second the loudness of the main target of focus

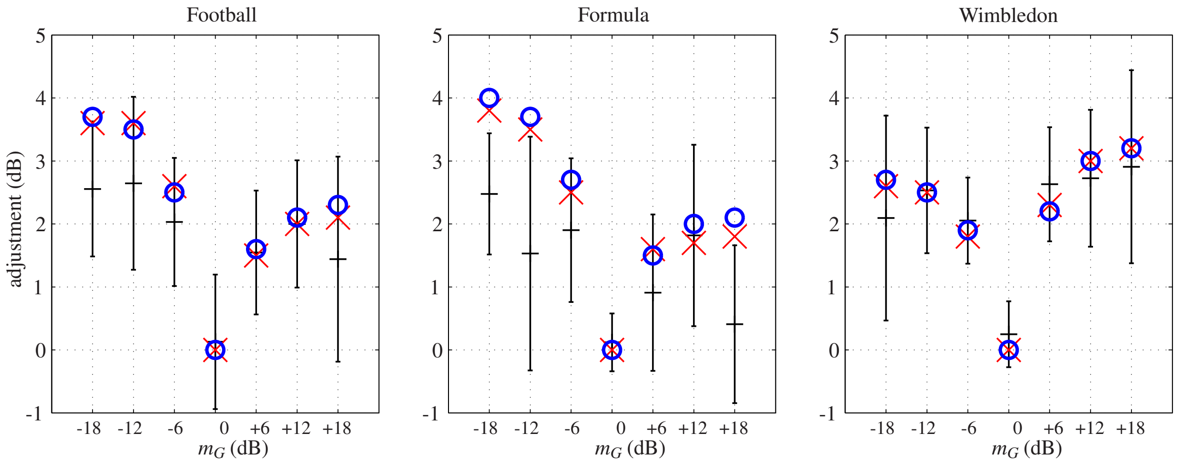

Fig. 2. The mean adjustment (black $-$) and the 95% confidence intervals using Student's t-distribution. The required signal amplifications derived from signal-based loudness estimation (red $\times$) and from the proposed estimation method (blue $\circ$) are given for comparison.

is kept constant. Focusing on different targets in different conditions partly explains the dip in the $m_G = +18$ dB condition: with smaller amplifications people match the overall loudness, but when the FGO becomes more dominant in the mixture, matching the FGO level becomes the main target.

5. CONCLUSIONS

Interactivity in object-based audio coding systems may lead to changes in the overall loudness of the output signal compared to the default mix. Especially in broadcasting applications, changes in the overall loudness are undesired. This paper proposes a method for estimating the loudness change by utilizing information of the loudness of the partial mixes of the objects and the output rendering information. An experiment focusing on dialogue enhancement application is described. The proposed parametric estimation method is demonstrated to produce results matching closely to values obtained via signal-based estimation. A subjective listening test was conducted. The test subjects adjusted the level of a probe signal (obtained by simulating interactivity) to make it aesthetically matching to a reference signal being the default mix. The average listener adjustment compensates approximately for the loudness difference between the reference and probe signals.

The proposed parametric loudness estimation method is able to estimate the change in the signal loudness accurately, and the subjective adjustment matches roughly the estimated loudness loss. Thus, estimating the loudness change with the proposed method and applying the corresponding compensation approximates closely the average adjustment by a listener. A compensation method based on the proposed estimation method is included in a DVB-specification and standardized in ETSI [9].

REFERENCES

[1] K. Brandenburg, "MP3 and AAC explained," in *Proc. of 17th International Audio Engineering Society Conference: High-Quality Audio Coding*, Florence, Italy, Aug. 1999.

[2] J. Herre *et al.*, "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.

[3] J. Herre *et al.*, "MPEG-H Audio - The new standard for universal spatial / 3D audio coding," *Journal of the Audio Engineering Society*, vol. 62, no. 12, pp. 821–830, Dec. 2014.

[4] J. Paulus *et al.*, "MPEG-D spatial audio object coding for dialogue enhancement (SAOC-DE)," in *Proc. of 138th Audio Engineering Society Convention*, Warsaw, Poland, May 2015.

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

[6] Advanced Television Systems Committee (ATSC), *Recommended Practice: Techniques for Establishing and Maintaining Audio Loudness for Digital Television (A/85:2013)*, Washington, D.C., USA, Mar. 2013.

[7] International Telecommunication Union, *ITU-R BS.1770-3: Algorithms to measure audio programme loudness and true-peak audio level*, Geneva, Switzerland, Aug. 2012.

[8] J. C. Riedmiller, S. Lyman, and C. Robinson, "Intelligent program loudness measurement and control: What satisfies listeners?" in *Proc. of 115th Audio Engineering Society Convention*, New York, Oct. 2003.

[9] ETSI TS 101 154 v2.2.1, *Digital Video Broadcasting (DVB); Specification for the use of Video and Audio Coding in Broadcasting Applications based on the MPEG-2 Transport Stream*, Geneva, Switzerland, Jun. 2015.