CODING OF FINE GRANULAR AUDIO SIGNALS USING HIGH RESOLUTION ENVELOPE PROCESSING (HREP)

Florin Ghido[1,2], Sascha Disch[1,2], Jürgen Herre[1,2], Franz Reutelhuber[1], Alexander Adami[2]

[1] Fraunhofer Institut für Integrierte Schaltungen (IIS), Am Wolfsmantel 33, 91058 Erlangen, Germany
[2] International Audio Laboratories Erlangen, Am Wolfsmantel 33, 91058 Erlangen, Germany
Correspondence should be addressed to Sascha Disch (sascha.disch@iis.fraunhofer.de)

ABSTRACT

High Resolution Envelope Processing (HREP) is a new tool for improved perceptual coding of audio signals that predominantly consist of many dense transient events, such as applause, rain drop sounds, etc. These signals have traditionally been very difficult to code for perceptual audio codecs, particularly at low bit rates. Based on the gain control principle, HREP acts as a pre-/post-processor pair to perceptual audio codecs and preserves the temporal fine structure and subjective quality of applause-like signals. Subjective tests have shown a significant improvement in audio quality of around 12 MUSHRA points by HREP processing at 48 kbps stereo when used together with an MPEG-H 3D Audio codec. The new coding tool has been adopted as part of MPEG-H 3D Audio Second Edition.

Index Terms— Audio Coding, Gain Control, Envelope, Simultaneous Masking, Applause

1. INTRODUCTION

In perceptual audio coding, signals that predominantly consist of many dense transient events, such as applause, rain drop sounds, etc. have traditionally been very difficult to code, particularly at low bit rates [1]. This paper introduces a new coding tool called High Resolution Envelope Processing (HREP) which acts as a pre-/post-processor pair to perceptual audio codecs and preserves the temporal fine structure and subjective quality of coded applause-like signals.

2. BACKGROUND: TEMPORAL MASKING

Classic perceptual coders like MP3 [18] or AAC [2][19] are primarily designed to exploit *simultaneous masking*, but also have to deal with the temporal aspect of masking: noise is masked a short time prior to and after the presentation of a masker signal (*pre-* and *post-masking*) [9]. Post-masking is observed for a much longer time period than pre-masking (about 10-50 ms instead of 0.5-2 ms, depending on level and duration of the masker). Thus, also the temporal aspect of masking leads to constraints for perceptual coding schemes: to achieve transparency, the quantization noise must not exceed the spectral *and* the time-dependent masking threshold. In transform coders, e.g. using a Modified Discrete Cosine Transform (MDCT) of window length 2048, quantization noise will be spread over a block duration larger than 40 ms at 44.1 kHz sampling rate, being far from an allowable pre-masking time of 2 ms. For percussive, transient-like audio signals, this gives rise to a pre-echo [4]. Accordingly, applause signals, especially when coded employing parametric joint multi-channel coding [7], will sound less crisp and too noisy.

Many techniques have been proposed to avoid pre-echo artifacts in coded signals [6]. In [10], a pre-echo control increases the coding precision for spectral coefficients of the filter bank window that first covers the transient signal portion. Since this increases the amount of bits for the coding of transient frames, it cannot be applied in a constant bit rate coder. Moderate local variations in bit rate demand might however be accounted for by a bit reservoir [3][10].

The adaptive window switching approach proposed by Edler [5] adapts the size of the filter bank windows to the characteristics of the input signal. While stationary signal parts will be coded using a long window, short windows are used to code the transient part of the signal.

Temporal Noise Shaping (TNS) was introduced in [6] and achieves a temporal shaping of the quantization noise through open-loop predictive coding along frequency direction on time blocks in the spectral domain.

Another way to avoid the temporal spread of quantization noise is to apply a gain control to the signal prior to calculating its spectral decomposition and coding. Such, a temporally flattened signal is fed into the codec avoiding pre-echo problems. An inverse gain control post-processes the codec output and reinstates the original signal dynamics. The parameters of the gain modification are transmitted in the bit stream. In [8], Link proposed a gain control as an addition to a perceptual audio coder using gain modifications on the full-band time domain signal. Also, a frequency dependent gain control has been proposed: in his dissertation [11], Vaupel noticed that full-band gain control does not work well for signals which exhibit very different temporal envelopes in different spectral regions. For a frequency dependent gain control, he proposed a compressor/expander filter pair that can be dynamically controlled in its gain characteristics.

In the Scalable Sample Rate (SSR) profile of MPEG-2 Advanced Audio Coding (AAC) [2][16], gain control is

978-1-5090-4117-6/17/$31.00 ©2017 IEEE

used within a hybrid filter bank. A first filter bank stage (PQF) splits the input signal into four bands of equal width. Then, a gain detector and a gain modifier perform the gain control encoder processing. Finally, as a second stage, four separate MDCT filter banks with a reduced size (256 instead of 1024) split the resulting signal further and produce the spectral components that are subsequently used for coding.

3. HREP SIGNAL PROCESSING

The benefits of applying the proposed HREP scheme are two-fold: HREP relaxes the bit rate demand imposed on the encoder by reducing short time dynamics of the input signal; additionally, HREP ensures proper envelope restoration in the decoder's (up-) mixing stage, which is all the more important if parametric multi-channel coding techniques have been applied within the codec. In this case, the individual channel envelopes can be restored in a much finer granularity than originally provided by the parametric spatial coding scheme - an idea first introduced in the Guided Envelope Shaping (GES) *post-processing* tool in the MPEG Surround (MPS) decoder [12]. HREP adopts this idea but inserts the envelope flattening already at the encoder *before* down-mixing and perceptual coding. This becomes possible in MPEG-H 3D Audio, since, as opposed to MPS, the transmitted down-mixes are not supposed to be listened to directly but merely serve as transport channels.

At the encoder side, the tool works as a pre-processor that temporally flattens the signal for high frequencies while generating a small amount of transmitted side information (1-4 kbps for stereo signals). At the decoder side, the tool works as a post-processor that temporally shapes the signal for high frequencies, making use of the side information.

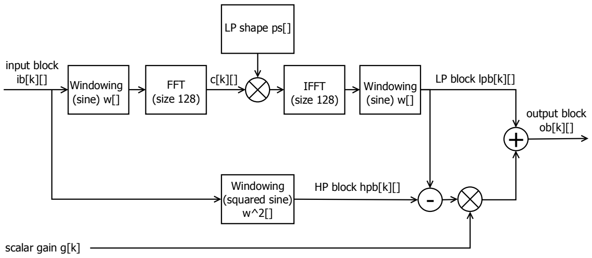

Figure 1: Detailed HREP signal flow in the encoder.

Figure 1 displays the signal flow inside the HREP encoder pre-processor. The pre-processing is applied by splitting the input signal into a low pass (LP) part and a high pass (HP) part, using a DFT to compute the LP part. Given the LP part, the HP part is obtained by subtraction in time domain. A time-dependent scalar gain is applied to the HP part, which is then added back to the LP part to create the pre-processed output.

The side information, estimated within an HREP analysis block (not depicted), comprises the beta_factor parameter, quantized and transmitted only once per frame using 3 bits, and the scalar gains $g[k]$, quantized using 3 bits and transmitted using adaptive entropy coding. For a frame of 1024 samples containing 16 blocks, the maximum side information has approximately $3 + 16 \times 3$ bits. The HREP analysis block may contain additional mechanisms that can gracefully lessen the effect of HREP processing on signal content ("non-applause signals") where HREP is not fully applicable. Thus, the requirements on applause detection accuracy are considerably relaxed.

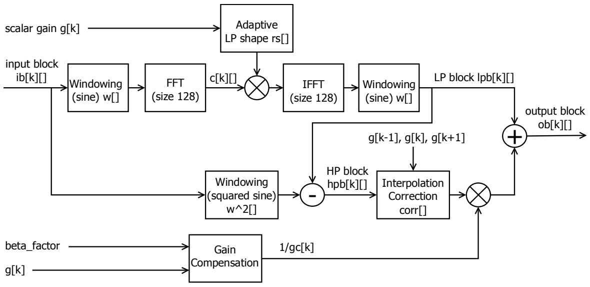

Figure 2: Detailed HREP signal flow in the decoder.

The processing at the decoder side is outlined in Figure 2. The side information on HP shape information and scalar gains are parsed from the bit stream (not depicted) and applied to the signal resembling a decoder post-processing inverse to the encoder pre-processing. It is applied by splitting the signal into a LP and a HP part by using a DFT to compute the LP part. Given the LP part, the HP part is obtained by subtraction. A scalar gain dependent on transmitted side information is multiplied to the HP part, which is added back to the LP part to create the output.

HREP should be combined with TNS in the codec. HREP will be activated throughout for sound textures containing densely spaced transients whereas TNS handles single prominent claps, sparsely distributed transients and note onsets. Accordingly, the actual TNS activation rate will be substantially reduced by application of the HREP tool, but still TNS adds noticeable perceptual benefit in case of sparsely distributed transients. TNS as an integral part of the encoder is applied *after* any down-mixing into transport channels and therefore cannot restore the individual channel envelopes at the decoder side like HREP does.

3.1. ENCODER OVERVIEW

The entire processing is done independently on each channel signal. Therefore, to simplify notation, the encoder and decoder processing is described only for one channel. The input signal s is split into blocks of size $N = 128$, which are half-overlapping, producing input blocks $\mathrm{ib}[k][i] = s[k \times \frac{N}{2} + i]$, where k is the block index and i is the sample position in the block k. A window w is applied to $\mathrm{ib}[k]$, in particular the sine window, defined as

$$w[i] = \sin\pi(i + 0.5)/N, \text{ for } 0 \le i < N.$$

Applying a DFT, the coefficients $c[k][f]$ are obtained as

$$c[k][f] = \mathrm{DFT}(\mathrm{w} * \mathrm{ib}[k]), \text{ for } 0 \le f \le \frac{N}{2},$$

where the $*$ operator means element-wise multiplication of two vectors. On the encoder side, in order to obtain the LP part, we apply an element-wise multiplication of $c[k]$ with the processing shape ps, which consists of the following:

$$\mathrm{ps}[f] = \begin{cases} 1, \text{ for } 0 \leq f < \mathrm{lp\_size} \\ 1 - \dfrac{f - \mathrm{lp\_size}{+}1}{\mathrm{tr\_size} + 1}, \text{ for } \mathrm{lp\_size} \leq f < \mathrm{lp\_size} + \mathrm{tr\_size} \\ 0, \text{ for } \mathrm{lp\_size} + \mathrm{tr\_size} \leq f \leq \dfrac{N}{2} \end{cases}$$

The LP shape is sent less frequently than the frame rate as a configuration setting, comprised of the lp_size parameter, representing the width of the low-pass region, and the tr_size parameter, representing the width of the transition region, both in DFT lines.

The LP block $\mathrm{lpb}[k]$ is obtained by applying an IDFT and windowing again as

$$\mathrm{lpb}[k][i] = w * IDFT(\mathrm{ps} * c[k]), \text{ for } 0 \leq i < N,$$

The HP block $\mathrm{hpb}[k]$ is then obtained by simple subtraction in time domain as

$$\mathrm{hpb}[k][i] = \mathrm{ib}[k][i] \times w^2[i] - \mathrm{lpb}[k][i], \text{ for } 0 \leq i < N.$$

Based on the gain $g[k]$ chosen by the encoder for each block, the value of the output block $\mathrm{ob}[k]$ is computed as

$$\mathrm{ob}[k][i] = \mathrm{lpb}[k][i] + g[k] \times \mathrm{hpb}[k][i], \text{ for } 0 \leq i < N.$$

Finally, the output signal is computed using the output blocks using overlap-add as

$$o\left[k \times \frac{N}{2} + j\right] = \mathrm{ob}[k-1]\left[j + \frac{N}{2}\right] + \mathrm{ob}[k][j], \text{ for } 0 \leq j < \frac{N}{2},$$
$$o\left[(k+1) \times \frac{N}{2} + j\right] = \mathrm{ob}[k]\left[j + \frac{N}{2}\right] + \mathrm{ob}[k+1][j], \text{ for } 0 \leq j < \frac{N}{2}.$$

3.2. DECODER OVERVIEW

The processing at the decoder mirrors the encoder, however there are several novel aspects which significantly improve the perceptual quality of the decoded signal.

3.2.1. Adaptive processing shape

For perfect reconstruction in the transition region, an adaptive reconstruction shape $\mathrm{rs}[k]$ must be used instead of the encoder processing shape ps for the computation of the LP block $\mathrm{lpb}[k]$, depending on the current gain $g[k]$ and defined for $\mathrm{lp\_size} \leq f < \mathrm{lp\_size} + \mathrm{tr\_size}$ as

$$\mathrm{rs[k]}[f] = 1 - (1 - \mathrm{ps}[f]) \times \frac{g[k]}{1 + (g[k] - 1) \times (1 - \mathrm{ps}[f])},$$
$$\mathrm{lpb}[k][i] = w * IDFT(\mathrm{rs}[k] * c[k]), \text{ for } 0 \leq i < N.$$

The adaptive processing shape ensures that the energy for each spectral bin f in the transition region is preserved.

3.2.2. Interpolation correction

The gains $g[k-1]$ and $g[k]$ applied on encoder side to blocks on positions $k-1$ and k are implicitly interpolated due to the windowing and overlap-add operations, and similarly for the gains $g[k]$ and $g[k+1]$. Introducing the auxiliary function $D(g_1, g_2) = \frac{g_1}{g_2} + \frac{g_2}{g_1} - 2$, on the decoder side the implicit interpolation of $1/g[k-1]$ and $1/g[k]$ produces an amplitude overshoot by a factor of up to $1 + D(g[k-1], g[k]) \times w^2[N/4] \times (1 - w^2[N/4])$.

To achieve perfect reconstruction above the transition region, a time-varying correction factor $1/\mathrm{corr}[i]$ is needed,

$$\mathrm{corr}[j] = 1 + D(g[k-1], g[k]) \times w^2[j] \times (1 - w^2[j]),$$
$$\mathrm{corr}\left[j + \frac{N}{2}\right] = 1 + D(g[k], g[k+1]) \times w^2[j] \times (1 - w^2[j]),$$

with $0 \leq j < \frac{N}{2}$. The factor ensures that the energy for each spectral bin f in the high frequency region is preserved.

3.2.3. Gain compensation

The core codec may introduce additional attenuation or smearing that can be modeled and compensated for by adjusting the gains $g[k]$ using the beta_factor$\in[0, 0.395]$. In the listening tests below, a fixed value determined by the bitrate was used. The gain compensation "expands" the value of the gains $g[k]$ around the neutral value of 1 by

$$gc[k] = g[k] + \mathrm{beta\_factor} \times (g[k] - 1).$$

Thus, if a gain $g[k]$ is smaller than 1, corresponding to a peaky transient event, $gc[k]$ will become even smaller.

3.2.4. Computation of the output blocks

Based on $gc[k]$ and $\mathrm{corr}[j]$, the value of the output block $\mathrm{ob}[k]$ at the decoder side is computed as

$$\mathrm{ob}[k][i] = \mathrm{lpb}[k][i] + \frac{1}{gc[k]} \times \frac{1}{\mathrm{corr}[i]} \times \mathrm{hpb}[k][i], \text{ for } 0 \leq i < N$$

and the final output signal is obtained using the output blocks with overlap-add exactly like in the encoder.

5. PERFORMANCE EVALUATION

The effect of HREP on perceptual quality has been assessed through MUSHRA [17] tests for two different scenarios.

Table 1: Listening tests overview.

Methodology	MUSHRA ITU-R BS.1534	
Test Conditions	"nohrep": Reference Quality Encoder "hrep": HREP enabled configuration	
Test Signals	1: ARLapplause 2: Exercise 3: HeavyApplause 4: Intro	5: Klatschen 6: MediumApplause 7: SallyBrown 8: applse
Bit Rate/Ch.	48kbps stereo (medium quality)	
No. of listeners	15	
Bit Rate/Ch.	128kbps stereo (high quality)	
No. of listeners	14	

Table 1 summarizes the listening test setups comprising stereo headphone listening at 48 kbps and 128 kbps based on an MPEG-H 3D Audio codec operating in frequency domain (FD) mode and using Intelligent Gap Filling (IGF) [15]. TNS was activated in all conditions. The beta_factor parameter was chosen 0.32 and 0.17, respectively. Solely applause items were tested, since HREP is only activated for such types of signal, e.g. using a dedicated classifier [21].

Figure 3 shows the absolute MUSHRA scores of the 48 kbps stereo test. Perceptual quality is in the "fair" to "good" range. Consistently, the "hrep" condition scores higher than the "nohrep" condition.

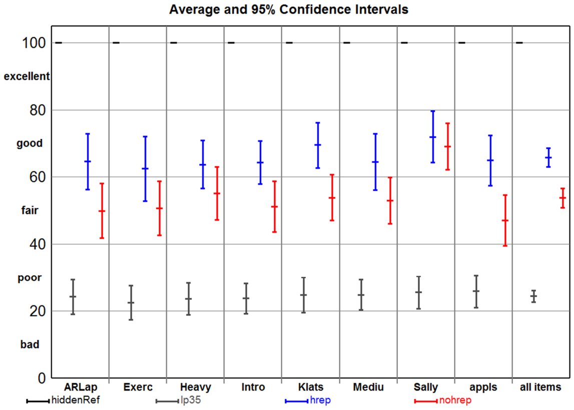

Figure 3: Absolute MUSHRA scores for 48 kbps stereo test.

Figure 4 depicts the difference MUSHRA scores. For all items except one, “hrep” scores significantly better than “nohrep”. Improvements ranging from 3 to 17 points are observed. Overall, there is a significant average gain of 12 points while none of the items is significantly degraded.

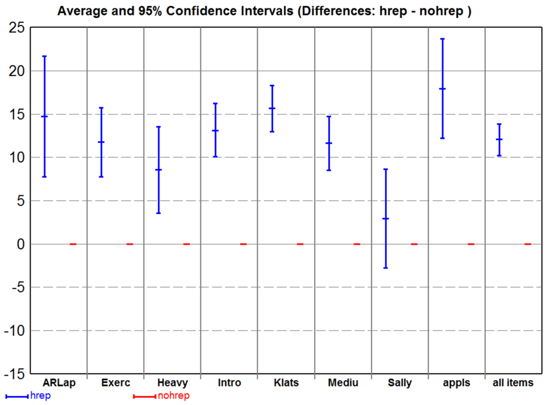

Figure 4: Difference MUSHRA scores for 48 kbps stereo test.

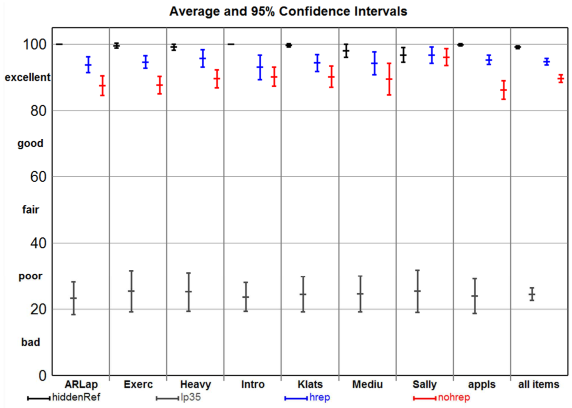

Figure 5: Absolute MUSHRA scores for 128 kbps stereo test.

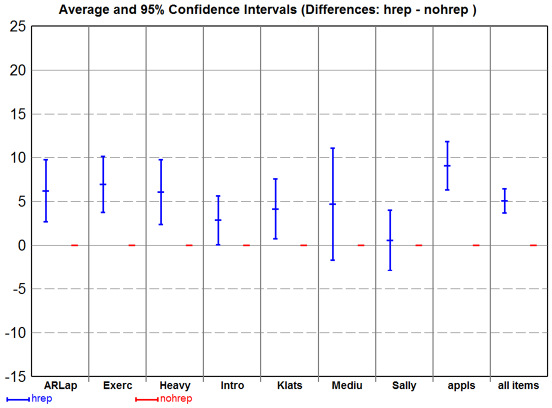

Figure 6: Difference MUSHRA scores for 128 kbps stereo test.

Figure 5 and Figure 6 show the absolute and the difference MUSHRA scores of the 128 kbps stereo test, respectively. In the absolute scores, all signals score in the range “excellent”. In the difference scores it can be seen that, even though perceptual quality is nearly transparent, for 6 out of 8 signals there is a significant improvement of 3 to 9 points, overall amounting to a mean of 5 MUSHRA points. None of the items is degraded significantly.

The computational complexity of the HREP processing is dominated by the calculation of the DFT/IDFT pairs that implement the LP/HP splitting of the signal. For a sampling frequency of 48 kHz, one real-valued DFT of size 128 with half-overlap consumes around 1.36 Mops and therefore a pair costs around 2.72 Mops per signal.

6. CONCLUSIONS

This paper presented a new tool called High Resolution Envelope Processing (HREP). HREP is a tool for improved perceptual coding of signals that predominantly consist of many dense transient events, such as applause, rain drop sounds, etc. HREP is based on the gain control principle and enhanced by a number of novel features that optimize the coding performance for application to applause-like signals.

The benefits of applying HREP are two-fold: HREP relaxes the bit rate demand imposed on the encoder by reducing short time dynamics of the input signal; additionally, HREP ensures proper envelope restauration in the decoder's (up-) mixing stage, which is particularly important if parametric multi-channel coding techniques are applied within the codec.

The computational complexity of HREP is moderate, around 2.72 Mops per signal. Subjective tests of HREP plus an MPEG-H 3D Audio codec have shown an improvement of around 12 MUSHRA points by HREP processing at 48 kbps stereo. Due to its merits, HREP has been included into the MPEG-H 3D Audio standard [12] within the Phase 2 developments [13] and is therefore contained in the Second Edition specification [14].

7. REFERENCES

[1] Demonstration CD-ROM on Audio Coding Artifacts: "Perceptual Audio Coders: What to listen for", *CD-ROM with tutorial information and audio examples*, AES publications http://www.aes.org/publications/technical/AudioCoding.cfm

[2] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, M. Dietz, J. Herre, G. Davidson, Oikawa: "MPEG-2 Advanced Audio Coding", *101st AES Convention*, Los Angeles, 1996.

[3] K. Brandenburg: "OCF - A New Coding Algorithm for High Quality Sound Signals", *Proc. IEEE ICASSP*, pp. 141-144, April 1987.

[4] J. D. Johnston, K. Brandenburg: "Wideband Coding Perceptual Considerations for Speech and Music", *in S. Furui and M. M. Sondhi, editors: "Advances in Speech Signal Processing",* Marcel Dekker, New York, 1992.

[5] B. Edler: "Codierung von Audiosignalen mit überlappender Transformation und adaptiven Fensterfunktionen", *Frequenz, Vol. 43*, pp. 252-256, 1989.

[6] J. Herre, J. D. Johnston: "Enhancing the Performance of Perceptual Audio Coders by Using Temporal Noise Shaping (TNS)", *101st AES Convention*, Los Angeles, 1996, Preprint 4384.

[7] Gerard Hotho, Steven van de Par, and Jeroen Breebaart: "Multichannel coding of applause signals", *EURASIP Journal of Advances in Signal Processing*, Hindawi, January 2008, doi: 10.1155/2008/531693.

[8] M. Link: "An Attack Processing of Audio Signals for Optimizing the Temporal Characteristics of a Low Bit-Rate Audio Coding System", *95th AES Convention*, New York, 1993, Preprint 3696.

[9] B. C. J. Moore: *An Introduction to the Psychology of Hearing*, Academic Press, London, 1989.

[10] ISO/IEC JTC1/SC29/WG11 MPEG, International Standard ISO 11172-3 "Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s".

[11] T. Vaupel: "Ein Beitrag zur Transformations-codierung von Audiosignalen unter Verwendung der Methode der 'Time Domain Aliasing Cancellation (TDAC)' und einer Signalkompandierung im Zeitbereich", *PhD Thesis*, Universität-Gesamthochschule Duisburg, Germany, 1991.

[12] ISO/IEC (MPEG-H) 23008-3:2015, "High efficiency coding and media delivery in heterogeneous environments - Part 3: 3D audio".

[13] ISO/IEC JTC1/SC29/WG11, "High efficiency coding and media delivery in heterogeneous environments - Part 3: 3D audio, AMENDMENT 3: MPEG-H 3D Audio Phase 2," 2015.

[14] ISO/IEC (MPEG-H) 23008-3:2017, Second Edition, "High efficiency coding and media delivery in heterogeneous environments - Part 3: 3D audio".

[15] S. Disch et al., "Intelligent Gap Filling in Perceptual Transform Coding of Audio", *141st AES Convention Proceedings*, Los Angeles, 2016.

[16] ISO/IEC 13818-7:1997, Information technology -- Generic coding of moving pictures and associated audio information -- Part 7: Advanced Audio Coding (AAC).

[17] ITU-R, Recommendation BS.1534-3 Method for the subjective assessment of intermediate quality level of coding systems, Geneva, 2015.

[18] K. Brandenburg, G. Stoll, The ISO/MPEG-Audio Codec: A Generic Standard for Coding of High Quality Digital Audio, *92nd AES Convention*, Vienna, 1992, Preprint 3336.

[19] J. Herre, M. Dietz: "Standards in a Nutshell: MPEG-4 High-Efficiency AAC Coding", IEEE Signal Processing Magazine, Volume 25, Issue 3, pp. 137 - 142, May 2008.

[20] Jürgen Herre, Kristofer Kjörling, Jeroen Breebaart, Christof Faller, Sascha Disch, Heiko Purnhagen, Jeroen Koppens, Johannes Hilpert, Jonas Rödén, Werner Oomen, Karsten Linzmeier, and Kok Seng Chong, "MPEG Surround – The ISO/MPEG Standard for Efficient and Compatible Multi-Channel Audio Coding", 122nd AES Convention, Vienna, Austria, 2007

[21] C. Uhle, "Applause sound detection," J. Audio Eng. Soc, vol. 59, no. 4, pp. 213–224, 2011.