FINE-GRAIN SCALABLE AUDIO CODING BASED ON ENVELOPE RESTORATION AND THE SPIHT ALGORITHM

Heiko Hansen, Stefan Strahl

Carl von Ossietzky University Oldenburg
Department of Physics
D-26111 Oldenburg, Germany

Alfred Mertins

University of Lübeck
Institute for Signal Processing
D-23538 Lübeck, Germany

ABSTRACT

We present strategies for perceptual improvements of embedded audio coding based on psychoacoustic weighting and spectral envelope restoration. The encoding schemes exhibit fine-grain bitrate scalability via the set partitioning in hierarchical trees (SPIHT) algorithm. Weighting factors and envelope parameters are transmitted under careful consideration of the amount of side information. For low bitrates, where the number of actually transmitted waveform coefficients is low, missing coefficients are shaped w.r.t. the spectral envelope. In our approach, the envelope information is transmitted in form of band-wise values of the l_1-norm. Sets of standardized audio files as well as various audio data of contemporary music are encoded and the results are analyzed with objective measures of perceptual quality. The proposed coding scheme competes in perceptual quality with existing state-of-the-art fixed bitrate coders such as MPEG-2/4 AAC. For low bitrates, the proposed embedded coding envelope restoration (ECER) improves the perceptual audio quality notably.

Index Terms— progressive audio compression, noise shaping, embedded coding, scalability.

1. INTRODUCTION

State-of-the-art lossy audio coders, such as MPEG-2/4 AAC, MP3 or WMA provide perceptual transparent audio quality at fixed target bitrates between 48-128 kbps (kbits per second). This quality is achieved via sophisticated compression techniques that shape the inevitable quantization noise under consideration of psychoacoustic principles, and thereby conceal the perceptual distortion. In these audio coding schemes a fixed target and non-embedded code stream is commonly used, rendering it difficult to transcode an existing encoded signal to higher or lower bitrates. This is a drawback particularly for audio transmission via heterogeneous networks and wireless links with time-varying capacity. To overcome this drawback, several approaches for a certain bitrate scalability with varying quality have been presented (e.g. [1, 2, 3, 4, 5, 6]). For low bitrates, say, below 64 kbps, the perceived quality for both, scalable and fixed target bitrate coders, decreases markedly.

Recent attempts to mitigate the distortions for low bitrates are directed toward a restoration of waveform coefficients from parameterized quantities of the spectral envelope. Spectral band replication (SBR) approaches replicate high frequency components from the transmitted low-frequency bands under consideration of transmitted parameters of the spectral envelope ([7]). Again, these methods are designed for fixed target rates and cannot be easily applied for fine-grain scalable codecs that provide progressive transmission with bit-wise quality-embedded codestreams.

In this paper, we propose a restoration technique that can even be used if only a few, or even none, of the actual waveform coefficients have been transmitted. In the proposed scheme, in addition to band-wise weighting data (i.e. scale factors, [8]) the scalefactor-band wise values of the l_1-norm of the spectral envelopes are calculated and transmitted as side information. This side information is followed by a fully embedded bitstream which is created by the set partitioning in hierarchical trees (SPIHT) algorithm and describes the actual waveform of the signal. On the decoder side, the l_1-norm is used as a measure of the uncertainty on the spectral coefficients in a scale factor band. The magnitudes of already SPIHT-decoded coefficients are subtracted from the l_1-norm and reduce the uncertainty on the remaining coefficients. Spectral gaps are then filled with uniform noise in such a way that the total l_1-norm of the envelope is preserved and none of the introduced noise samples exceeds the uncertainty threshold of the SPIHT algorithm.

This paper is organized as follows. In the next section we develop the technical details behind the proposed scalable coding scheme. In the Section 3 we present our results and compare them to existing state-of-the-art fixed-rate codecs. We conclude in Section 4.

The work has been supported by the SFB-TR31 and the International Graduate School for Neurosensory Science and Systems, Carl von Ossietzky University Oldenburg. Contact Email: heiko.hansen@uni-oldenburg.de

978-1-4244-3298-1/09/$25.00

2. TECHNICAL DESCRIPTION

Lossy audio encoders typically consist of three essential building blocks: a time-frequency transform that maps the time domain signal frame-wise into the spectral domain, a quantization, bit allocation and bitstream formatting block, and a psychoacoustic model, which estimates the perceptual relevancy of the introduced quantization noise. In our scheme, the widely used modified discrete cosine transform (MDCT, e.g. [9, 10]) is used, where the number of channels per frame is set to 1024. The estimation of perceptual relevance is performed via the the MPEG-2/4 psychoacoustic model ([11, 12]). For an audio file sampled with CD quality (44.1kHz, 16Bit) the 1024 MDCT channels are grouped into 49 scale factor bands (SFBs), and for each SFB, the perceptual masking threshold is calculated with the psychoacoustic model. In our setup the SPIHT algorithm [13] is used as quantization and embedded bitstream forming procedure. Originally developed for image coding, SPIHT has also been frequently applied for audio coding (e.g. [1, 14, 15, 6]). To combine the three building blocks, the MDCT coefficients are weighted band-wise with the inverse masking threshold (e.g. [16, 17, 8]). As pointed out by Schuller et al. ([17]) the weighting can be seen as a normalization to its masking threshold and thus the level of perceptual noise is constant. After the weighting is applied, the time signal reconstructed from the weighted MDCT now shows a flat masking threshold, exemplified in Fig. 1. The masking threshold needs to be transmitted to the decoder as side information. This is done via the well-known scale-factor method (an instructive description is given in [18]). The differential indices of the scale factors are encoded via Huffman tables. In our approach, we use a dynamic encoding procedure where we determine for each frame whether the index differences between the scale-factor bands within the frame or the differences to the previous frame can be encoded more efficiently. The encoding of this information causes one bit overhead for each frame. The weighted MDCT coefficients are finally encoded with the SPIHT algorithm. SPIHT is a fully embedded bitplane encoding scheme. The basic idea is to transmit the waveform transform coefficients frame-wise and via a bitslicing technique, which starts from the most-significant bitplane and then transmits the other bitplanes with decreasing order. A frame-wise constant threshold T is progressively reduced (i.e. halved) in every iteration step. Values above the threshold are rated significant and are encoded in the layered bit stream. The partial ordering of the significant coefficients is accomplished by a tree-based mapping technique with a pre-defined tree structure. The decoded signal approximations are reconstructed in a bit-by-bit manner, where every bit adds or subtracts a fraction of the threshold T and thus increases the accuracy of the reconstructed signal. For simplicity, we use the tree as in [14] with $N = 4$, but note that improvements of SPIHT coding performance can be achieved

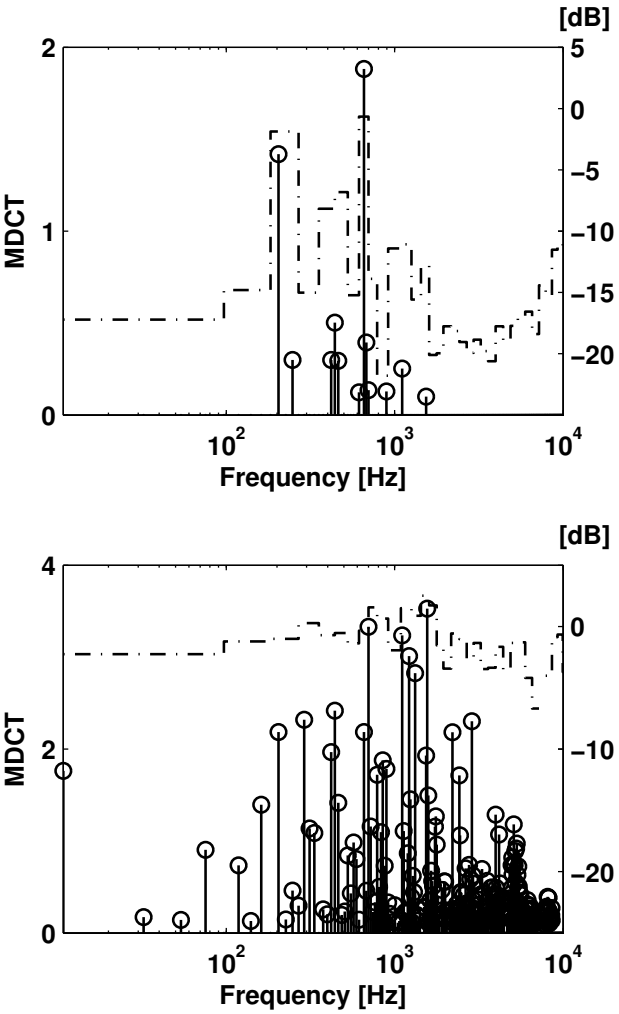

Fig. 1. MDCT coefficients and their masking thresholds. Upper panel: A typical set of MDCT coefficients (black stems) together with the associated masking threshold (dash-dotted line, in dB). Lower panel: The MDCT vector after weighting with the inverse threshold. The associated masking threshold of the time signal reconstructed from the weighted MDCT is included. See text for further explanations.

by using dynamic significance tree methods ([5, 19]). Subsequent to the decoding, the (weighted) MDCT approximations are re-scaled with the quantized masking threshold and the time signal is synthesized using the inverse MDCT and an overlap-and-add method ([9, 10]). This simple but elegant low-bitrate scalable coder (LSC) already provides fine-grain scalability down to a core bitrate of typically around 8 kbps, depending on the encoded audio data and the amount of overhead that needs to be transmitted. It is clear, however, that for bitrates below a certain limit, the loss in perceptual quality due to the quantization distortion is significant. At bitrates below, say, 60 kbps, too few of the perceptually

important waveform coefficients are decoded with reasonable accuracy to allow for transparent coding. Typically, modern audio coders tackle this problem by reducing the bandwidth of the audio file, and thus reducing the amount of information that needs to be transmitted. Clearly, this treatment reduces the perceptual quality. Dietz et al. ([7]) introduced spectral band replication (SBR) to mitigate the perceptual distortions at low bitrates. In the SBR approach, transmitted parameters of the spectral envelope are used to recreate, or restore, high-frequency waveform coefficients from the low frequency bands. However, to keep the flexibility of fine-grain scalability, reconstruction and/or restoration of waveforms needs to be a dynamic process. At a certain bitrate, only the non-transmitted waveform coefficients should be restored, under careful consideration of both the transmitted coefficients and additional spectral information. In our approach we keep the idea of the spectral envelope to derive the necessary spectral information. We use the band-wise l_1-norm of the weighted MDCT coefficients (X_k, with $k = 1, \ldots, 1024$) as measure of the uncertainty on the coefficients and thus also as a measure of the spectral envelope:

$$E(i) = \sum_{k_l(i)}^{k_u(i)} |X_k| \tag{1}$$

where $k_l(i)$ and $k_u(i)$ are the lower and upper indices for band i, respectively. For simplicity we choose the indices of the SFBs. The envelope is transmitted as side information in the same way as for scale factors, by using a lookup table and subsequent dynamic differential Huffman encoding. In our experiments with various audio data the amount of side information due to the quantized envelope increased by 30-90 percent. During the decoding procedure, the quantized envelope information $Q(E(i))$ per band is used as an indicator of the coefficient uncertainty. As quantized waveform coefficients $q_R(X_k)$ at bitrate R are recovered from the SPIHT bitstream, the uncertainty measure reduces as

$$A_{res}(i) = Q(E(i)) - \sum_{k_l(i)}^{k_u(i)} |q_R(X_k)| \tag{2}$$

Note that for a given bitrate some or all transform coefficients may have not been transmitted and thus the corresponding $q_R(X_k)$ can be zero. If the residual $A_{res}(i)$ is positive, the remaining gaps due to the non-transmitted transform coefficients are 'filled' with random uniform noise. The noise coefficients are scaled such that their corresponding l_1-norm for band i equals $\alpha A_{res}(i)$, with a global parameter $\alpha \geq 0.0$. In our experiments we achieved best results with $0.4 \leq \alpha \leq 0.6$. Another restriction on the noise coefficients is that none of the coefficients exceeds the actual significance threshold of the SPIHT algorithm. After re-weighting with the masking threshold, the audio signal is reconstructed in the known manner. The proposed method is called embedded coding envelope restoration (ECER) and the LSC equipped with ECER is named LSC ECER.

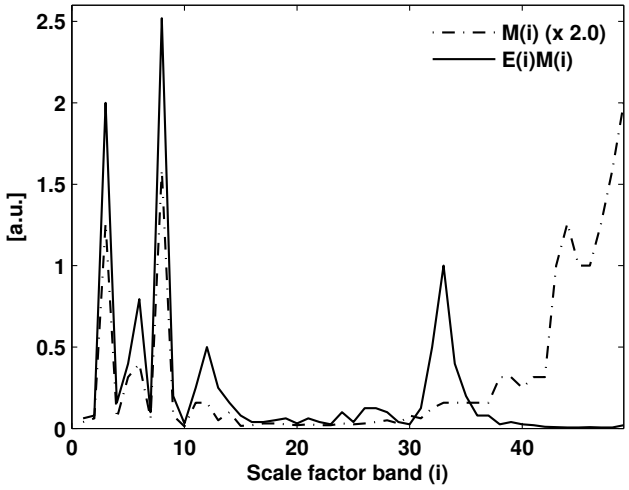

Fig. 2. Comparison of masking threshold and (non-weighted) spectral envelope: Shown are the masking threshold $M(i)$ (dash-dotted line) and the associated re-scaled envelope $E(i)M(i)$ (solid line). To illustrate the resemblance, $M(i)$ has been multiplied by a factor of 2.0. See text for further explanations.

In our effort to reduce the amount of side information, we also consider existing relationships between the masking threshold and the spectral envelope. Naturally, masking thresholds and (non-weighted) spectral envelopes, as a measure of spectral power, show high correlations for the lower bands. A typical example for this resemblance is shown in Fig. 2, where the masking threshold $M(i)$ and the re-scaled envelope $E(i)M(i)$ show a high resemblance for the lower bands. For high frequencies the threshold of hearing is the dominant masker. We recognize that generally, this resemblance is valid for the bands 1 to 33. Thereby, similar to the effects on the masking threshold stated above (exemplified in Fig. 1), the weighting with the inverse masking threshold results in a roughly constant value for $E(i)$. In one experiment, we set the transmitted spectral envelope to an estimated constant and transmit one value for the bands 1-33. The constant is estimated frame-wise by a linear regression of $M(i)$ and $E(i)M(i)$. Thereby, about 60 to 70 percent of the additional overhead due to ECER is saved. This experiment is called LSC ECER-base (short: ECER-base). The performance of the proposed schemes is discussed in the next section.

3. EXPERIMENTAL RESULTS

In this section, we evaluate the performance of the presented methods by means of objective perceptual quality measures. For this we encoded audio data with the proposed codecs

at different bitrates and calculated averages of the estimated quality. The first chosen dataset is the SQAM data base ([20]), which consists of 16 files of various sound and speech recordings. The second set consist of 12 files of contemporary pop music (including among others: Suzan Vega, Tracy Chapman, Robbie Williams, The Tiger Lillies, etc.). The measurement of objective perceptual quality has been performed with the PEAQ software tool ([21]). PEAQ provides the objective difference grade (ODG) as perceptual quality measure, where an audio file is rated continuously from 0.0 (imperceptible distortion to reference signal) to -4.0 (very annoying differences) ([21]). An ODG below -1.0 is said to have 'perceptible but not annoying', below -2.0 'slightly annoying' and below -3.0 'annoying' distortions.

The audio files have been encoded at bitrates between 24 and 128 kbps and ODG averages were calculated for both data sets. For the LSC ECER and ECER-base α was set to 0.5. To compare the performances with state-of-the-art audio codecs we used the MPEG-2/4 AAC reference implementation ([22]) at fixed target bitrates. Since the original introduction of the standard, AAC encoders have been continuously improved, and therefore we also used a state-of-the-art Nero AAC implementation (version 1.1.34.2, 2007) as an upper limit for the quality that can be achieved when no embedded bitstream is created. We further used the scalable AAC BSAC scheme which was provided with the AAC reference software ([2, 22]). AAC BSAC exhibits scalability starting from a base layer of 16 kbps, with several possible enhancement layer up to 64 kbps. The minimum layer step size is 1 kbps.

The results of the ODG measurements are summarized in Fig. 3. As one can see, LSC and LSC ECER generally outperform the AAC reference implementation for low bitrates. This is clearly visible for the SQAM database (Fig. 3, upper panel), where the LSCs show a superior quality to the AAC reference implementation. Furthermore, the ECER approaches generally improve the quality significantly in comparison to the pure LSC. ECER-base performs slightly below LSC ECER for bitrates under 64 kbps. For high bitrates the AAC reference implementation reaches a saturation in quality for some individual sounds, which explains the plateau-like slope. For the set of contemporary sounds (Fig. 3, lower panel) the LSCs show superior quality to the AAC reference for bitrates below 80 kbps. Again clearly visible is the increase in quality for LSC ECER and ECER-base compared to the pure LSC. Both ECER and ECER-base are of roughly the same quality, with a slight advancement of ECER-base for bitrates above 64 kbps. Apparently, the reduced overhead of ECER-base, compared to LSC ECER, compensates a presumed quality loss for low bitrates and leads to the quality increase for high rates, where the coefficient uncertainty is vanishing. AAC-BSAC, within its range of scalability, shows a poor quality, and can hardly compete with the LSCs at bitrates above 40 kbps.

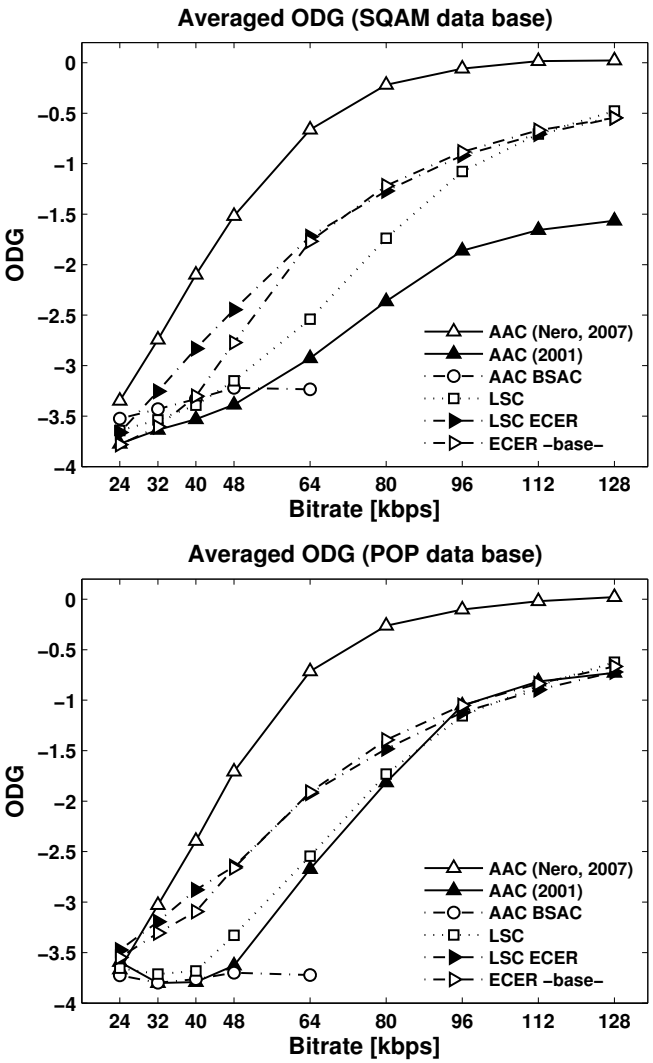

Fig. 3. Averaged ODG results. The audio files have been encoded at different bitrates with the proposed audio coding schemes and the AAC implementations. Upper panel: results for the SQAM audio data base. Lower panel: Similar analysis for a set of contemporary music.

4. CONCLUSIONS

We presented lossy audio coding schemes with fully embedded fine-grain scalability down to a core bitrate of about 8 to 16 kbps. The codecs are based on the appropriate weighing of the waveform coefficients with the (inverse) masking threshold and additionally, in the ECER approach, with the restoration of (non transmitted) coefficients from the spectral envelope, which were transmitted as band-wise values of the l_1-norm. Objective measurements of perceptual quality with datasets of standardized audio data and contemporary music revealed that the proposed codecs can compete with modern fixed-rate audio coders. The LSC ECER shows a significant

improvement in perceptual quality for bitrates below 64 kbps, in comparison to the pure LSC and the MPEG AAC reference implementation, and can measure up with state-of-the-art fixed target bitrate codecs. In fact, the additional amount of side information due to the encoded spectral envelope could be further reduced by using existing relationships between the masking threshold and the spectral envelope. The results are very promising, and it is assumed that a further tuning of ECER and the application of more sophisticated dynamic significance tree methods ([19]) will further increase the quality.

5. REFERENCES

[1] Z. Lu and W. A. Pearlman, "An efficient, low-complexity audio coder delivering multiple levels of quality for interactive applications," in *Proceedings of the IEEE Signal Processing Society 1998 Workshop on Multimedia Signal Processing*, 1998, pp. 529–534.

[2] S.-H. Park, K. Yeon-Bae, K. Sang-Wook, and S. Yang-Seock, "Multi-layer bit-sliced bit-rate scalable audio coding," in *AES 103th Convention*, NY, USA, Sept. 1997, preprint 4520.

[3] J. Li, "Embedded audio coding (EAC) with implicit auditory masking," in *Proc. ACM on Multimedia*, Nice, France, Dec. 2002, pp. 592–601.

[4] M. Raad, A. Mertins, and I. Burnett, "Scalable to lossless audio compression based on perceptual set partitioning in hierarchical trees (PSPIHT)," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, Apr. 2003, pp. V624–627.

[5] S. Strahl and A. Mertins, "An efficient fine-grain scalable compression scheme for sparse data," *SPARSE05*, 2005.

[6] C. Dunn, "Scalable bitplane runlenght coding," in *AES 120th Convention*, Paris, France, May 2006, Paper 6749.

[7] M. Dietz, L. Liljeryd, K. Kjorling, and O. Kunz, "Spectral band replication, a novel approach in audio coding," in *AES E-Library*. AES, April 2002, Audio Engineering Society, Paper number 5553.

[8] C. Dunn, "Aspects of scalable audio coding," in *AES 122th Convention*, Vienna, Austria, May 2007, Paper 7081.

[9] H. S. Malvar, *Signal Processing with Lapped Transforms*, Artech House, Boston, MA, 1992.

[10] S. Shlien, "The modulated lapped transform, its time-varying forms, and its applications to audio coding standards," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 4, pp. 359–366, July 1997.

[11] Moving Pictures Experts Group (MPEG), "Information Technology - Generic coding of moving pictures and associated audio information - Part 7: Advanced Audio Coding (AAC)," 2006, Fourth edition.

[12] Moving Picture Experts Group (MPEG), "MPEG-4 Audio Version 2 (Final Committee Draft 14496-3 AMD1)," *ISO/IEC/JTC1/SC29/WG11 N2803*, Jul 1999.

[13] A. Said and W. A. Pearlman, "A new fast and efficient image codec based on set partitioning in hierarchical trees," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 6, pp. 243–250, 1996.

[14] M. Raad and A. Mertins, "From lossy to lossless audio coding using SPIHT," in *Proc. of the 5th Int. Conf. on Digital Audio Effects*, Hamburg, Germany, Sept. 2002, pp. 245–250.

[15] S. Strahl, H. Zhou, and A. Mertins, "An adaptive tree-based progressive audio compression scheme," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA05)*, New Paltz, NY, USA, Oct. 2005, pp. 219–222.

[16] T. S. Verma and T. H. Y. Meng, "Sinusoidal modeling using frame-based perceptually weighted matching pursuits," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Washington, DC, USA, 1999, pp. 981–984, IEEE Computer Society.

[17] G. Schuller, B. Yu, D. Huang, and B. Edler, "Perceptual audio coding using adaptive pre- and post-filters and lossless compression," *IEEE Trans. Speech and Audio Processing*, vol. 10, no. 6, pp. 379–390, Aug. 2002.

[18] C. Bauer and M. Vinton, "Joint optimization of scale factors and Huffman code books for MPEG-4 AAC," *IEEE Transactions on Signal Processing*, vol. 54, no. 1, pp. 177–189, 2006.

[19] S. Strahl, H. Hansen, and A. Mertins, "A dynamic fine-grain scalable compression scheme with application to progressive audio compression," submitted to IEEE Trans. Audio, Speech, and Language Processing, Jan. 2009.

[20] EBU, "Tech 3253 - Sound Quality Assessment Material (SQAM)," Tech. Rep., European Broadcasting Union, April 1988.

[21] "ITU-R BS.1387 : Method for objective measurements of perceived audio quality," 11 2001, Recommendation BS.1387-1 (11/01).

[22] "ISO/IEC 14496-3, Information technology–coding of audio-visual objects–part 3: Audio," 2001.