MULTICHANNEL AUDIO UPMIXING BASED ON NON-NEGATIVE TENSOR FACTORIZATION REPRESENTATION

J. Nikunen, T. Virtanen

Tampere University of Technology
Korkeakoulunkatu 1, 33720 Tampere, Finland
joonas.nikunen@tut.fi, tuomas.virtanen@tut.fi

M. Vilermo

Nokia Research Center
Visiokatu 1, 33720 Tampere, Finland
miikka.vilermo@nokia.com

ABSTRACT

This paper proposes a new spatial audio coding (SAC) method that is based on parametrization of multichannel audio by sound objects using non-negative tensor factorization (NTF). The spatial parameters are estimated using perceptually motivated NTF model and are used for upmixing a downmixed and encoded mixture signal. The performance of the proposed coding is evaluated using listening tests, which prove the coding performance being on a par with conventional SAC methods. Additionally the proposed coding enables controlling the upmix content by meaningful objects.

Index Terms— Spatial audio coding, Object-based audio coding, Non-negative tensor factorization

1. INTRODUCTION

The audio coding research have recently focused on spatial audio coding (SAC), where one or few discretely encoded signal channels are transmitted with additional spatial cues for synthesis of multiple channels. The existing SAC algorithms are mostly based on binaural cue coding principles [1]. Algorithms include for example parametric stereo coding [2] and coding of multichannel audio [3]. The spatial synthesis from a downmixed signal is based on adjusting the level, time delay and decorrelation of the time-frequency blocks used for spatial parameter estimation. The parametrization relies on assumption of non-overlapping sound sources in the frequency domain or momentary dominance of certain sound source in perception of direction. Another degree of parametrization in audio coding is using objects having interpretable structure, for example audio objects based on harmonic components of instruments [4]. In this paper we focus on incorporating SAC with an object-based model for spatial parametrization.

We propose a new object-based SAC algorithm utilizing audio spectrogram parametrization by non-negative tensor factorization (NTF) algorithm, which estimates object spectra and their spatial parameters simultaneously. The NTF spatial parametrization is used for recovering the multichannel signal from a downmixed and perceptually encoded stereo signal by filtering the downmix short-time Fourier transform (STFT) in Wiener filtering manner using the NTF model as a time-frequency filter kernel.

The proposed approach relies on object parametrization from the mixture signal in a blind sound separation manner by using non-negative matrix factorization (NMF) and its extension to multidimensional data by NTF [5]. The NMF algorithm with various extensions has been intensively studied for blind sound source separation [6, 7]. The separation is based on ability of NMF algorithm to find and model repetitive structures from audio signals using a single object. The NMF objects usually represent sound structures such as individual notes of an instrument, chords or drum hits.

In addition to object separation, the advantage of NTF representation for SAC is that it utilizes long-term redundancy present in an audio signal by using a single object to describe repetitive sound events. Additionally, the NTF signal model estimates the spatial parameters and the object spectrum simultaneously allowing utilization of inter-channel redundancy in representation of the spatial information. Such non-redundant spatial representation is efficient with respect to coding and bitrate performance. In comparison to most existing SAC methods, NTF allows representing overlapping frequency content of the objects, which enables better spatial synthesis and separation of such simultaneous sound events.

The rest of the paper is organized as follows. In Section 2 the novel method for object-based spatial audio encoding and decoding utilizing NTF for signal parametrization is proposed. In Section 2.1 a perceptually motivated NMF cost function [8] is extended for NTF and multichannel observations. The upmix filtering framework for spatial synthesis with NTF is proposed in Section 2.2. The estimation of NTF parameters optimized for the upmix operation is proposed in Section 2.3 and the quantization end encoding of the parameters is shortly revised in Section 2.4. The results from a listening test are provided in Section 3.

2. PROPOSED METHOD FOR SPATIAL AUDIO CODING

The encoding and decoding of the proposed SAC algorithm are illustrated in Figures 1 and 2, respectively. The encoding starts by calculating the STFT of each input signal channel to obtain a magnitude spectrogram tensor. The masking level is estimated from the input signal and the NTF algorithm with perceptual weighting is applied to the spectrogram tensor to obtain an object-based spatial parametrization of it. The input signal is downmixed to stereo and perceptually encoded. The encoded downmix is STFT analyzed and used as an additional time-frequency weighting to optimize the NTF spatial parametrization for the upmix filtering operation. The estimated spatial parameters are quantized and entropy encoded.

The decoding to recover the multichannel signal starts by decoding of the downmix and calculating its STFT. The spatial parameters are decoded and dequantized. The upmixing is done by filtering the downmix STFT in a Wiener filtering manner where the channel dependent filter kernels are obtained from the NTF model. The time-domain signals are synthesized using the phases obtained from the downmix STFT analysis for every upmixed channel.

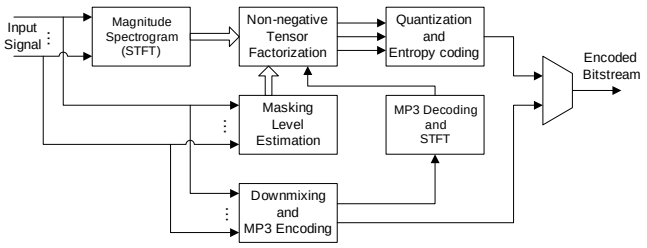

Figure 1: Block diagram of the encoding part of the proposed coding algorithm.

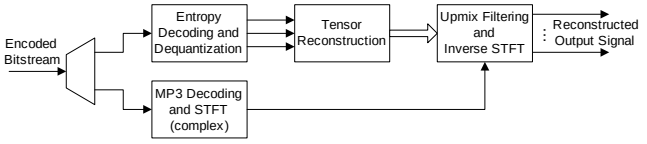

Figure 2: Block diagram of the decoding part of the proposed coding algorithm.

2.1. Non-negative Tensor Factorization

In this section the NTF representation of multichannel magnitude spectrogram tensor is introduced and we extend the perceptual weighting proposed in [8] for multichannel observations with NTF. We will use the following notation. Tensors are denoted by capital bold letters and a single entry of rank-j tensor $\mathbf{X}$ is denoted as $\mathbf{X}_{i_1,i_2,\ldots i_j}$.

For a multichannel time-domain audio signal $x(n,c)$, of sample index $n = 1,\ldots N$ and in channels $c = 1,\ldots,C$, absolute values of its STFT are denoted by $\mathbf{X}_{k,t,c}$, where $k = 1,\ldots,K$ is the positive DFT frequency bin index and $t = 1,\ldots,T$ is the STFT frame index. STFT is calculated using frame length of $N = 2(K-1)$ samples and consecutive frames are overlapping by $N/2$ samples, Hanning window function is used.

The NTF signal model for approximating spectrogram tensor $\mathbf{X}$ of rank three can be written as a product of three matrix entries summed over the decomposition objects r as

$$\mathbf{X}_{k,t,c} \approx \hat{\mathbf{X}}_{k,t,c} = \sum_{r=1}^{R} \mathbf{B}_{k,r}\mathbf{G}_{r,t}\mathbf{A}_{r,c}, \tag{1}$$

where R is the number of NTF objects used for the approximation. Each column of $\mathbf{B}$ contains the DFT spectrum of an object. The corresponding row of $\mathbf{G}$ represents its gain in each STFT frame, and the corresponding row of $\mathbf{A}$ represents the channel-dependent gain of the object. The NTF model has been found to produce good results in sound source separation in [5]. The NTF model (1) constitutes from a set of fixed object spectra that have time-varying gain and a channel-dependent gain.

2.1.1. Perceptually Motivated Weighting for NTF

The cost function to be minimized in finding the NTF approximation is the noise-to-mask ratio (NMR) [9], which evaluates perceptual quality of encoded audio by determining audibility of encoding artefacts based on masking phenomenon invoked by the desired signal content. The perceptually motivated NMF algorithm, minimizing the NMR of the approximation by multiplicative updates of the model parameters was proposed in [8].

We propose to extend the NMR cost function for NMF [8] to be used with the NTF signal model (1). The masking level for each frame is estimated in Bark band domain and the masking level conversion to any desired DFT frequency resolution is given in [8]. We will denote the masking level for the each time-frequency point in each input channel c by a tensor $\mathbf{W}_{k,t,c}$. The NMR measure equals to squared Euclidean distance of the original and NTF spectrogram weighted by the masking level and can be defined as

$$c_{\mathrm{NMR}} = \sum_{c=1}^{C}\sum_{t=1}^{T}\sum_{k=1}^{K} \mathbf{W}_{k,t,c}(\mathbf{X}_{k,t,c} - \sum_{r=1}^{R}\mathbf{B}_{k,r}\mathbf{G}_{r,t}\mathbf{A}_{r,c})^2. \tag{2}$$

2.1.2. Estimation of the Perceptually Motivated NTF

The estimation of the NTF model minimizing NMR (2) is achieved by iterative multiplicative updates, which can be derived using same principles as in [6]. The update rules are given as

$$\begin{aligned}
\mathbf{B}_{k,r} &\leftarrow \mathbf{B}_{k,r}\frac{\sum_t\sum_c \mathbf{W}_{k,t,c}\mathbf{X}_{k,t,c}\mathbf{G}_{r,t}\mathbf{A}_{r,c}}{\sum_t\sum_c \mathbf{W}_{k,t,c}\hat{\mathbf{X}}_{k,t,c}\mathbf{G}_{r,t}\mathbf{A}_{r,c}},\\
\mathbf{G}_{r,t} &\leftarrow \mathbf{G}_{r,t}\frac{\sum_k\sum_c \mathbf{B}_{k,r}\mathbf{W}_{k,t,c}\mathbf{X}_{k,t,c}\mathbf{A}_{r,c}}{\sum_k\sum_c \mathbf{B}_{k,r}\mathbf{W}_{k,t,c}\hat{\mathbf{X}}_{k,t,c}\mathbf{A}_{r,c}},\\
\mathbf{A}_{r,c} &\leftarrow \mathbf{A}_{r,c}\frac{\sum_k\sum_t \mathbf{B}_{k,r}\mathbf{G}_{r,t}\mathbf{W}_{k,t,c}\mathbf{X}_{k,t,c}}{\sum_k\sum_t \mathbf{B}_{k,r}\mathbf{G}_{r,t}\mathbf{W}_{k,t,c}\hat{\mathbf{X}}_{k,t,c}},
\end{aligned} \tag{3}$$

where $\hat{\mathbf{X}}_{k,t,c}$ is the reconstructed NTF model evaluated according to (1) before each update.

The complete NTF algorithm is as follows. First the entries of matrices $\mathbf{B}, \mathbf{G}$ and $\mathbf{A}$ are initialized with random values uniformly distributed between zero and one. The decomposition matrices are then iteratively updated by applying the updates (3) for each of the matrices at a time. A fixed number of iterations is used.

2.2. Object-based Spatial Upmixing Using NTF Model

In this section we will propose an object-based spatial upmixing method for recovering the multichannel signal. We will derive the upmixing model only for stereo downmix, but it can be defined for any other desired channel configuration of the downmix encoding.

The original multichannel time-domain signal $x(n,c)$ is downmixed to stereo by

$$l(n) = \sum_{c\in\mathcal{L}} x(n,c), \qquad r(n) = \sum_{c\in\mathcal{R}} x(n,c), \tag{4}$$

where $l(n)$ and $r(n)$ are the left and right channel respectively. For a 5.1 speaker configuration $\mathcal{L}$ contains front and rear left channel with center and low-frequency extension, $\mathcal{R}$ contains respectively the right side counterparts. The downmixed time-domain signal is perceptually encoded and is available at the decoder. The decoded downmix signal is STFT analyzed using same analysis parameters (window length, etc.) to obtain complex downmix STFT spectrogram $\mathbf{L}_{k,t}$ and $\mathbf{R}_{k,t}$ for left and right stereo channels respectively.

The object-based multichannel signal model is used for upmixing the downmixed STFT as follows. The STFT of the upmixed

signal is obtained as

$$\mathbf{Y}_{k,t,c} = \begin{cases} \mathbf{L}_{k,t} \dfrac{\sum_{r=1}^{R} \mathbf{B}_{k,r}\mathbf{G}_{r,t}\mathbf{A}_{r,c}}{\sum_{c\in\mathcal{L}}\sum_{r=1}^{R} \mathbf{B}_{k,r}\mathbf{G}_{r,t}\mathbf{A}_{r,c}}, & \text{(5a)} \quad c \in \mathcal{L}, c \notin \mathcal{R} \\ \mathbf{R}_{k,t} \dfrac{\sum_{r=1}^{R} \mathbf{B}_{k,r}\mathbf{G}_{r,t}\mathbf{A}_{r,c}}{\sum_{c\in\mathcal{R}}\sum_{r=1}^{R} \mathbf{B}_{k,r}\mathbf{G}_{r,t}\mathbf{A}_{r,c}}, & \text{(5b)} \quad c \in \mathcal{R}, c \notin \mathcal{L} \\ \dfrac{1}{2}\left[\text{(5a)} + \text{(5b)}\right] & c \in \mathcal{L}, \mathcal{R} \end{cases} \tag{5}$$

Note that $\mathbf{L}_{k,t}, \mathbf{R}_{k,t}$ and $\mathbf{Y}_{k,t,c}$ are complex-valued. Time-domain signals are obtained by inverse STFT and overlap-add. The above can be viewed as filtering the downmix to multiple channels using a time-varying Wiener filter.

Similar filtering methods are widely used in reconstruction of source signals when NMF or NTF is used for sound source separation, for example [7]. The proposed method allows synthesizing only selected objects. In this case the summation in numerator of (5) is evaluated over desired group of objects $r \in \mathcal{O}$.

2.3. NTF Parameter Optimization for the Upmix Filtering

Taking into account the filtering operation (5) the NMR cost of NTF model becomes

$$c = \sum_{c=1}^{C}\sum_{t=1}^{T}\sum_{k=1}^{K} \mathbf{W}_{k,t,c}(\mathbf{X}_{k,t,c} - |\mathbf{Y}_{k,t,c}|)^2 \tag{6}$$

where $\mathbf{Y}$ is evaluated according to (5). The difference between NTF cost functions (2) and (6) is that the latter takes into account the downmixing process and particularly the case where sound events from different spatial positions are overlapping or are closely separated in time and frequency. Such case will introduce cross-talk to the upmixed channels if cost function (2) and updates (3) are used. This equals to filtering undesired downmix STFT details to the upmixed channels.

We propose to approximate the upmixing cost function (6) and reduce the cross-talk of the upmixed channels by giving bigger weighting for time-frequency bins in which the downmix STFT has high magnitude with respect to the NTF channel sum in equation (8). This is achieved by replacing $\mathbf{W}_{k,t,c}$ in (3) with

$$\hat{\mathbf{W}}_{k,t,c} = \begin{cases} \mathbf{W}_{k,t,c}\dfrac{\mathbf{L}_{k,t}}{\hat{\mathbf{L}}_{k,t}}, & c \in \mathcal{L}, c \notin \mathcal{R} \\ \mathbf{W}_{k,t,c}\dfrac{\mathbf{R}_{k,t}}{\hat{\mathbf{R}}_{k,t}}, & c \in \mathcal{R}, c \notin \mathcal{L} \\ \mathbf{W}_{k,t,c}\dfrac{1}{2}\left(\dfrac{\mathbf{L}_{k,t}}{\hat{\mathbf{L}}_{k,t}} + \dfrac{\mathbf{R}_{k,t}}{\hat{\mathbf{R}}_{k,t}}\right), & c \in \mathcal{L}, \mathcal{R} \end{cases} \tag{7}$$

where

$$\hat{\mathbf{L}}_{k,t} = \sum_{c\in\mathcal{L}}\sum_{r=1}^{R} \mathbf{B}_{k,r}\mathbf{G}_{r,t}\mathbf{A}_{r,c}, \quad \hat{\mathbf{R}}_{k,t} = \sum_{c\in\mathcal{R}}\sum_{r=1}^{R} \mathbf{B}_{k,r}\mathbf{G}_{r,t}\mathbf{A}_{r,c} \tag{8}$$

The weights $\hat{\mathbf{W}}$ need to be updated after every NTF iteration due the change of $\hat{\mathbf{L}}_{k,t}$ and $\hat{\mathbf{R}}_{k,t}$.

The practical implementation of the proposed weighting is done by first using cost function (2) and parameter updates (3) for several hundreds of iterations to estimate an initial NTF model and then change to (7) to optimize the NTF model to the given downmix. Even though we cannot proof that algorithm described above would minimize (6), the experiments with the implementation have shown to produce desired result.

The behavior of the cost function (6) measuring the upmixed signal NMR was investigated to prove its decrease with the proposed weighting. The evaluation of (6) was done first with the update rules (3) and then changing to the proposed updates (7). The cost function was averaged over the whole test set described in Section 3 and the resulting cost is illustrated in Figure 3. Detailed encoding settings are given in Section 3. The decrease of the upmix filtering NMR cost is evident when switching to the proposed updates at iteration 500.

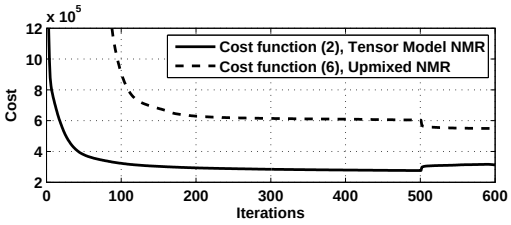

Figure 3: Behaviour of cost function (6) with the NTF algorithm updates (3) for the first 500 iterations and with updates (7) for successive 100 iterations.

2.4. Quantization and Encoding of the NTF Parameters

The NTF model derived in Section 2.3 is quantized and entropy coded and sent as a side information for spatial synthesis by the proposed upmix filtering. We use the quantization framework proposed in [10], which applies a non-uniform quantization for object spectrum $\mathbf{B}_{k,r}$ and gains $\mathbf{G}_{r,t}$ in such way that more quantization levels are assigned for smaller parameter values. For quantization of the channel gain parameter $\mathbf{A}_{r,c}$ we use uniform quantization.

In [10] the frequency of occurrence of quantized values of $\mathbf{B}_{k,r}$ and $\mathbf{G}_{r,t}$ was gathered from a large test set resulting to distributions having high probability of zeros and rest of the quantization levels had relatively small probability. Such distributions of quantized values can be effectively utilized for reducing the output bitrate by entropy coding. In the case of proposed SAC algorithm we calculated the entropy of each individual model parameter $\mathbf{B}_{k,r}, \mathbf{G}_{r,t}$ and $\mathbf{A}_{r,c}$ to estimate the final bitrate after entropy coding.

3. EVALUATION

In this section we will present listening test results of the proposed SAC algorithm when evaluated using multiple stimuli with hidden reference and anchor (MUSHRA) [11] methodology and comparing the coding quality to MP3 surround [3] at similar bitrates. Test samples used were the MPEG multichannel evaluation samples.

The listening test was run in Nokia Research Center listening room, which is fully conformant with ITU-R BS.1116-1 [12]. Speakers were set up according to ITU-R BS.775. 3.5kHz low-pass filtered original was used as a lower anchor. A lower anchor with spatially reduced quality was deemed unnecessary since all listeners were experts. 10 listeners participated in the test. Listeners were instructed to grade the samples taken into account all coding artefacts including spatial sound image.

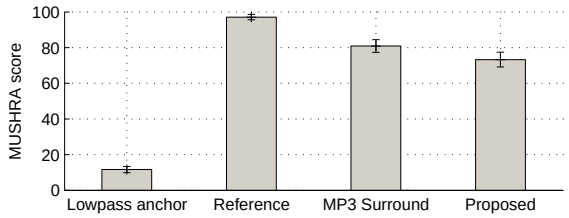

Figure 4: Listening test mean score and 95% confidence intervals.

Encoding parameters were chosen as follows, the window length was set to $N = 880$ samples, which equals to 20ms with the sampling frequency $F_s = 44.1$ kHz. The length of the NTF segment was chosen to 15 seconds. The number of NTF algorithm iterations were set to 500 with updates (3), and 100 for the optimization stage with updates (7). The number of bits for representing each NTF parameter with quantization described in Section 2.4 was defined by preliminary listening tests with professional listeners. Quantization evaluation resulted to using $n_b = 4, n_g = 4$ and $n_a = 6$ bits per parameter for $\mathbf{B}_{k,r}, \mathbf{G}_{r,t}$ and $\mathbf{A}_{r,c}$ respectively.

The stereo core encoding algorithm was MP3 at 96 kbps and the target bitrate with the NTF upmixing side information was 128 kbps. The number of NTF objects was determined by allocating the remaining bitrate after downmix encoding to the NTF model. The bitrate reduction by entropy coding was estimated as described in Section 2.4. The resulting number of objects $R = 64$ corresponds to a NTF bitrate of 26.0 kbps after quantization and averaging the estimated entropy coding bitrate reduction over the whole test set.

The listening test results are given in Figure 4. The results indicate that with similar bitrates the proposed coding attains slightly lower mean score than the compared SAC method, MP3 Surround. However, the score difference is small and barely not statistically significant, which indicates that coding performance achieved by the proposed algorithm is comparable to the existing SAC methods. Both tested SAC methods do not achieve transparency compared to the hidden reference, but the overall quality level can be considered to be moderate and suitable for coding of multichannel audio for consumer applications. The listening results prove that the proposed SAC method can be used for coding of multichannel audio with at bitrates equivalent to 128kbps or similar.

The proposed algorithm achieved coding performance comparable to conventional SAC approaches and additionally the proposed upmix filtering allows manipulation of the upmix content by NTF objects, corresponding to meaningful sound events. The NMF and NTF signal decomposition models have been shown to achieve promising results in blind sound source separation [5, 6, 7]. Combining the proposed coding with separation would produce SAC with possibility to control the content of the upmix for example by instruments. The sound separation performance of the proposed SAC method was informally evaluated by k-means clustering of the NTF objects with spectral and time-gain based features. The separation performance was determined to be promising and comparable to separation results achieved in [5, 6].

4. CONCLUSION

We proposed a novel method for spatial audio coding (SAC) by using non-negative tensor factorization (NTF) for deriving an object-based spatial upmixing model. The spatial synthesis was done in Wiener filtering manner using NTF representation as a time-frequency filtering kernel and we proposed an experimental algorithm for estimating the NTF parameters found to minimize the filtering cost. The listening test showed that the proposed SAC algorithm achieved the performance of conventional spatial audio coding methods, but additionally enabling the control of the upmix by objects in blind sound separation manner. The future work will include more extensive evaluation of sound source separation performance of the proposed method.

5. REFERENCES

[1] F. Baumgarte and C. Faller, "Binaural Cue Coding-Part I: Psychoacoustic Fundamentals and Design Principles," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 6, pp. 509–519, 2003.

[2] E. Schuijers, J. Breebaart, H. Purnhagen, and J. Engdegård, "Low complexity parametric stereo coding," in *Proc. of 116th AES Convention, Berlin, Germany*, 2004.

[3] J. Herre, C. Faller, C. Ertel, J. Hilpert, A. Hoelzer, and C. Spenger, "MP3 Surround: Efficient and compatible coding of multi-channel audio," in *Proc. of 116th AES Conv.*, 2004.

[4] E. Vincent and M. Plumbley, "Low bit-rate object coding of musical audio using Bayesian harmonic models," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1273–1282, 2007.

[5] D. FitzGerald, M. Cranitch, and E. Coyle, "Non-negative Tensor Factorisation for Sound Source Separation," in *Proc. of the Irish Signals and Systems Conference, Dublin, Ireland*.

[6] T. Virtanen, "Monoaural Sound Source Separation by Non-negative Matrix Factorization With Temporal Continuity and Sparseness Criteria," *IEEE Trans. on Audio, Speech and Language Process.*, vol. 15, pp. 1066–1074, 2007.

[7] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. on Audio, Speech and Language Process.*, vol. 18, no. 3, pp. 550–563, 2010.

[8] J. Nikunen and T. Virtanen, "Noise-to-Mask Ratio Minimization by Weighted Non-negative Matrix factorization," in *Proc. of ICASSP '10*, Dallas, USA, 2010.

[9] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, C. Colomes, M. Kheyl, G. Stoll, K. Brandenburg, and B. Feiten, "PEAQ - The ITU Standard for Objective Measurement of Perceived Audio Quality," *Journal of AES*, vol. 48, pp. 3–29, 2000.

[10] J. Nikunen and T. Virtanen, "Object-based Audio Coding Using Non-negative Matrix Factorization for the Spectrogram Representation," in *Proc. of 128th AES Conv.*, London, UK, 2010.

[11] ITU-R BS. 1534-1, "Method for the subjective assessment of intermediate quality level of coding systems," *Int. Telecomm. Union Radiocomm. Assembly*, 2003.

[12] ITU-R BS. 1116-1, "Methods for the Subjective Assessment of Small Impairments in Audio Systems including Multichannel Sound Systems," *Int. Telecomm. Union Radiocomm. Assembly*, 1994.