Journal of Computations & Modelling, vol.3, no.2, 2013, 95-110
ISSN: 1792-7625 (print), 1792-8850 (online)
Scienpress Ltd, 2013

Psychoacoustically-Driven Multichannel Audio Coding

Demetrios Cantzos[1]

Abstract

Multichannel audio rendering allows for the immersion of a listener into a rich acoustic scene, as compared to traditional stereophonic methods. Nevertheless, the excessive transmission and storage requirements of multichannel audio pose a considerable obstacle towards its widespread usage. A novel method is presented here with which a single channel of a multichannel signal is conventionally transmitted and stored while the remaining channels are resynthesized based on statistical conversion of the same, single channel-signal. The size of the parameters required for the resynthesis process is much smaller than the size of the original channel-signal for the same resulting audio quality.

Mathematics Subject Classification: 68P30

Keywords: Multichannel Audio, Gaussian Mixture, Line Spectral Frequencies, Residual Conversion, Coding

---

[1] Department of Automation, Technological Education Institute (TEI) of Piraeus, Athens, Greece.

Article Info: *Received* : April 21, 2013. *Revised* : May 24, 2013
Published online : June 20, 2013

1 Introduction

Multichannel audio is employed in a wide array of scenarios, such as Home Theater systems, cinemas and teleconferencing applications. The reason for this popularity is the high quality audio reproduction, delivered to the listener through multiple channel-signals. However, the storage and transmission costs for a multichannel audio signal are excessive as compared to traditional stereophonic audio. In this work, we propose a novel method that reduces the transmission or storage cost of multichannel audio and builds upon our previous work in audio resynthesis [3]. In audio resynthesis, one channel of a multichannel audio segment (target signal) can be recreated from another channel (source signal) of the same audio segment using a linear function determined by a small set of parameters. A similar framework is adopted in parametric coding techniques such as Binaural Cue Coding (BCC) [5], Intensity Stereo Coding (ISC) [6] or the latest Harmonics plus Individual Lines plus Noise model (HILN) [7]. Our method is different as we attempt to resynthesize the remaining channels' original waveform and not to recreate spatial audio cues that only approximate the original audio spectrum. The basic assumption upon which our method is built is that individual channels of the same multichannel recording exhibit similarities at the waveform level and this can be exploited to reduce coding or transmission overheads.

For simplicity, we assume that each channel corresponds to only one signal and vice versa. For the same reason, we also assume that only one signal (channel) needs to be recreated which we call the target channel or signal. The generalization to more than one target channels is straightforward. The source and target signals correspond to recordings of the same multichannel audio piece but they are obtained with microphones located at different places in the recording venue. We adopt the scenario presented in [9], where microphone recordings at various positions in an orchestra hall are taken. Our main goal is to recreate a single target recording (channel) by transmitting only a source recording and a small set of constant parameters. The size of these parameters will be only a

fraction of the source or target recording size.

The algorithm works in conjunction with a psychoacoustic model to convert the source channel to the target channel at a fine grain, incremental step. Specifically, the derivation of the conversion parameters is guided by a psychoacoustic criterion and the bitrate of these parameters is finely tuned according to a bit allocation scheme, described later. The statistical conversion itself is based on a Linear Predictive Coding (LPC) [11] scheme of Modified Discrete Cosine Transform (MDCT) coefficients. The MDCT domain is selected in order to ensure compatibility with modern transform codecs, although the parametric nature of LPC allows the whole algorithm to be applied directly even on PCM data.

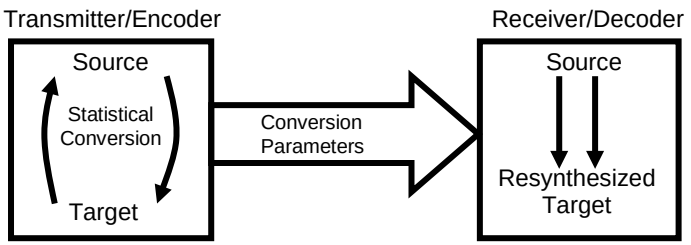

Figure 1: The audio resynthesis scheme. The transmitter has access to both source and target signals and derives the conversion parameters which are sent to the receiver. The receiver resynthesizes the target signal using the source signal and the conversion parameters.

The remainder of this paper is organized as follows. In Section II, we describe the core component of the algorithm which is the statistical conversion of features and residual vectors and the derivation of the conversion parameters set. Subsequently, methods on reducing the conversion parameters size via a sorting transformation are presented. Bitrate control of the conversion parameters size, based on a psychoacoustic model, is described in the last part of Section II. In

Section III, the results of the algorithm's performance are presented and evaluated in order to demonstrate the effectiveness of the method. In Section IV, concluding remarks on the algorithm are made.

2 Statistical Conversion

2.1 Pre-processing

The statistical conversion process is based on our previous work in [3]. Our starting point is a pair of source and target signals, taken from the same multichannel audio piece, with the source being one channel-signal and the target being another channel-signal (Figure 1). Note that the target signal is only available at the encoding side (transmitter), not at the decoding (receiver). At the encoder side, the source and target signals are transformed by the MDCT filterbank on a frame-by-frame basis. Each frame is pre-windowed with a Kaiser-Bessel window and adjacent MDCT spectral coefficients of each source or target frame are grouped into 32 subbands, similarly to a standard codec's approach, although other subband configurations are possible. After the MDCT filterbank, an LPC analysis is applied on the MDCT coefficients of each of the 32 subbands to extract the Line Spectral Frequency (LSF) [11] feature vectors and their corresponding residual vectors.

The LPC analysis of each MDCT subband group is performed on each frame separately, and not across all MDCT frames, leading to interframe independence and thus enabling us to accurately map a signal subband segment to the LSF or residual vectors and vice versa. This is important during the perceptually-driven conversion process as we will be able to modify the LSF or residual vectors of a particular MDCT frame without influencing adjacent frames due to the LPC analysis window overlap. Consequently, a minimum of two LSF and two residual vectors per subband and per frame (i.e. a minimum of two

overlapping LPC subframes) has to be produced to ensure perfect MDCT frame recovery during the inverse operation (LPC synthesis) at the decoder.

2.2 Conversion Function

After the LSF and residual vectors extraction from the source and target signals, the conversion function needs to be computed. Statistical conversion is based on the assumption that the source LSF or residual vectors of each subband are generated by a random process whose samples follow a diagonal Gaussian mixture model (GMM) pdf, given as

$$G(\mathbf{x}) = \sum_{k=1}^{K} p(C_k) \prod_{j=1}^{q} g(x^{(j)}; \mu_k^{(j)}, \sigma_k^{(j)}) \quad (1)$$

where C_k denotes the cluster (component) k, K is the number of clusters and $p(C_k)$ denotes the prior probability that the source vector $\mathbf{x}$ belongs to cluster k. The source vector is q dimensional and the jth coefficient is denoted by $x^{(j)}$. The mean and variance of each GMM cluster in (1) for coordinate j are noted as, $\mu_k^{(j)}$, and $\sigma_k^{(j)}$, respectively. The vector coefficients are considered to be independent and thus the vector pdf is the product of the q coefficient pdf's. The complete model parameters ($\mu_k^{(j)}$, $\sigma_k^{(j)}$, $p(C_k)$) for the LSF and residual vector GMMs can be estimated by an ML estimation algorithm as the one in [3], using LSF and residual vectors originating from pink noise, and are permanently stored. LSF or residual vectors conversion between the various source and target subband segments is implemented through a linear conversion function. The conversion function, $F(\cdot)$, acts on the source LSF/residual vector sequence $[\mathbf{x_1},...,\mathbf{x_n}]$ and produces a reconstructed vector sequence close in the least squares sense to the target LSF/residual vector sequence $[\mathbf{y_1},...,\mathbf{y_n}]$. Since we have selected a diagonal implementation, this function will act on the individual vector components and minimize the error

$$E = \sum_{t=1}^{n} \sum_{j=1}^{q} | \, y_t^{(j)} - F(x_t^{(j)})|^2, \tag{2}$$

as in [12]. To address this task we consider the function F as piecewise linear i.e.

$$F(x_t^{(j)}) = \sum_{k=1}^{K} P(C_k \mid \mathbf{x}_t)[v_k^{(j)} + \frac{u_k^{(j)}}{\sigma_k^{(j)}}(x_t^{(j)} - \mu_k^{(j)})] \tag{3}$$

for t =1,..,n and j =1,..,q. The conditional probability that a given vector belongs to cluster k, $P(C_k|\boldsymbol{x}_t)$, is given by Bayes' Rule. The unknown parameters set [$\boldsymbol{v}$,$\boldsymbol{u}$] can be found by minimizing (2) which reduces to solving a typical set of q independent least-squares equations [12] and hence the linear conversion function F is fully determined. We call [$\boldsymbol{v}$,$\boldsymbol{u}$] the *conversion parameters set*. These parameters are to be transmitted to the receiver or decoder in order to reconstruct the target LSF/residual data because these are the only parameters that are dependent on the particular source and target LSF/residual data. Note that a diagonal implementation is favored because the computation of the conversion parameters set is faster and, most importantly, the size of the parameters set itself is much smaller [12]. The remaining parameters of (3) are part of the LSF or residual mixture model and they are precomputed (once) and permanently stored during the mixture pdf estimation.

2.3 Sorting Transformation

A technique, based on our previous work [2], is adopted here that can significantly increase LSF and residual conversion accuracy, while reducing the conversion parameters size. We sort the source and target vector coefficients (LSF or residual) along each coordinate in ascending order. The motivation behind the sorting transformation is found in the form of the conversion function. The conversion function is a (piecewise) linear estimator that estimates the target data from the source data. Its optimal performance is achieved when the true relation

between the source and target data is linear along each coordinate. Therefore, this sorting technique allows us to reduce the number of mixture classes because the estimation is easier and consequently the number of conversion parameters is also reduced. Details of this scheme can be found in [2].

In order for the decoder to be able to use the (sorted) reconstructed LSF or residual coefficients and create the final subband signal, the original order of the reconstructed coefficients has to be known. The source data are available at the decoder and thus the original order of the source LSF and residual coefficients is known. Our focus is on deducing the original order of the target data and use that info to reorder the reconstructed data at the decoder end. We term this information, the *sorting information* and it is transmitted along with the conversion parameters. These two sets combined form the *transmitted parameters*.

The straightforward solution would be to transmit the original order of the target LSF and residual data as side information, along with the conversion parameters. At the decoder, the coefficients would be reconstructed one by one and a side index would determine where to place the particular LSF/residual coefficient. This scheme would require transmission of $n \cdot \log_2 n$ bits of information where $n+1$ is the number of elements being sorted (assuming n is a power of two). Instead of directly transmitting the sorting indices of the target data to the decoder, we can derive a sequence of minimum insertions and shifts that will take us from the source sorting indices to the target sorting indices. The reasoning behind this is that the source and target data have not identical but similar original position configurations and thus the target original positions could be inferred from the source original positions with fewer than $n \cdot \log_2 n$ bits of information. The steps of an algorithm [2] that allows us to transmit less information to the decoder without explicitly sending every index of the target column are:

1. The encoder checks if the source and target indices of the current row are the same. If yes, then a zero is transmitted. If no, then proceed to the next step.
2. The encoder looks in the target index of the current row and finds the position

(new row) of that index in the source column. The distance between the current row and the new row is transmitted. All values in the current row of the source column up to the new row are circularly shifted by one position towards the end of the column, so that the value of the new row replaces the value of the current row.

3. Repeat steps 1 and 2 until all rows of the target column have been traversed and the source column has been converted to the target column.

After the algorithm is completed, the source column has been converted to the target column. This lossless operation will enable us to send fewer bits at the decoder, especially after we perform entropy coding.

2.4 Bitrate Control

The sorting information, as described in the previous section, accounts for more than 85% of the transmitted parameters set. Therefore, it is essential to be able to accurately control the size of the sorting information in order to control the total bitrate of the transmitted parameters. It should be noted that bitrate control entails loss of conversion accuracy, i.e. in order to further reduce the sorting information, the reconstruction accuracy of the LSF or residual vectors is ultimately compromised. Naturally, this should happen in a fine step manner, so that fine grain scalability is achieved. A variant of our work in [2] is used in which prior to being sorted, the source vector coefficients X and target vector coefficients Y that participate in the derivation of the conversion function are modified according to

$$X' = |X| \tag{4}$$

$$Y' = Y \cdot \text{sign}(X) + c \cdot |X| \tag{5}$$

where c is called the *multiplier* and takes values from a predefined set of positive integers including zero (available to both the encoder and the decoder), and sign(·)

is the sign function which outputs +1 if the sign is positive and -1 if the sign is negative. The role of the multiplier is to increase the similarity between X' and Y' in terms of their sorting position indices so that, after sorting, the sorting information according to Section II-C will be less. The higher the multiplier is, the lower the size of the sorting information becomes -at the cost of decreased conversion accuracy- since X will dominate over Y in (5).

The role of the sign function is auxiliary and it only benefits residual conversion and not LSF conversion, since the LSFs are always positive. One can observe that the product $Y \cdot \text{sign}(X)$ is usually positive because the residual vectors X and Y have similar sign (especially in the low subbands), as they are taken from the same audio piece. Therefore, it is expected that the individual Y' residual vector coefficients of (5) will be, in most cases, positive. Inserting positive X', given by (4), and Y' vector coefficients in the conversion process (after they are sorted) increases the conversion accuracy at no extra transmission overhead since X and sign(X) are available at the decoder side. Note that the inversion of (5) to derive Y at the decoder is straightforward as long as c is known. In order to fully control the bitrate, we need a method to directly relate the multiplier c, for each subband and for either LSF or residual conversion, to the number of bits of the sorting information. A straightforward way to adjust the size of the sorting information to exactly M bits/coefficient by tuning c is as follows:

1. At the encoding side, set $c = 0$. After sorting X' and Y' along each coordinate, the encoder derives the sorting information and computes its maximum value, m, for all coefficients and coordinates. If $m \leq 2^M$, then store c and stop. Else, proceed to step 2.
2. At the encoding side, set $c = c + 1$. After sorting X' and Y' along each coordinate, the encoder derives the sorting information and computes its maximum value, m, for all coefficients and coordinates. If $m \leq 2^M$, then store c and stop. Else, go to the beginning of step 2.

Notice that as c increases, M decreases, along with the conversion accuracy.

The psychoacoustic model described next, helps to adaptively determine the appropriate values for *M* and maintain conversion accuracy at acceptable levels.

2.5 Psychoacoustic Model and Bit Allocation

A method for controlling the bitrate of the conversion parameters has been described in Section II-D. A problem remains, though, in determining which vectors (and of which subbands) will be converted and with how many bits of sorting information. The standard approach is to minimize a psychoacoustic distortion metric, such as the Noise-to-Mask ratio (NMR) [1]. An accurate method, based on our work in [4], for taking advantage of the available conversion function bitrate is to minimize the NMR by considering the effect of each of the 32 signal subbands separately. The difficulty in this task is that, by default, the NMR is calculated over a bark scale while our analysis runs across a linear frequency scale of 32 equidistant subbands. Nevertheless, by studying each subband separately, bit allocation becomes more efficient and yields considerable bitrate savings. The steps described below lead to a NMR variation matrix that is sensitive to the particular subband of a particular MDCT frame:

1. Starting with *T* MDCT frames, calculate the initial *T* NMR values between the source and target frames, and store them into a *T* x 1 vector, $NMRmat_0$.Set $i = 1$, $t = 1$.
2. For MDCT frame *t* and subband *i* of the source signal, replace the source LSF and residual vectors with the corresponding target vectors. Calculate the NMR for that frame and subtract $NMRmat_0[t]$ from it. Store that value to a 32 x *T* matrix NMRmat as the [*i*,*t*] entry. Set $i = i + 1$. If $i \leq 32$, go to step 2. Else, go to step 3.
3. Set $t = t + 1$, $i = 1$. If $t \leq T$, go to step 2. Else, terminate.

With the above method, we derive a 32 x *T* matrix of the NMR decrease that

each subband (out of a total of 32 subbands) of each MDCT frame (out of a total of *T* frames) incurs, as compared to the original NMR between the source and target MDCT frames. By picking the matrix indices for which the NMR decrease values are the largest, we know for which frames and subbands the conversion process would be most beneficial for the reduction of the total NMR of the reconstructed signal. Hence, for these indices we allocate more bits/coefficient for the sorting information, through the algorithm of Section II-D. For the indices with the smallest decrease, conversion does not even take place. This concept is similar to the water-filling bit allocation method of MPEG1-Layer 3 (MP3) [10].

3 Results

The algorithm is evaluated on a multichannel rock music recording wherein there is strong presence of vocals, bass and high frequency instruments. The number of rendering channels is 6 although, for convenience, the algorithm is tested on two channels only, i.e., one channel is the source signal and the other channel is the target. The extension to more channels is straightforward by choosing a different target signal in case a different channel needs to be resynthesized. The audio quality is quantified by means of the ITU-R BS.1387 PEAQ test, basic model [8], which emulates a subjective listening test. Its output is the Objective Difference Grade (ODG) value which ranges from -4 (“very annoying”) to 0 (“imperceptible distortion”). The PEAQ test results have been shown [13] to be highly correlated with the scores from a subjective listening test, especially at medium to high bitrates.

In Figure 2, the relation between the multiplier *c* and the size of the resulting sorting information is depicted. As expected, with the increase of the multiplier there is a decrease in the transmitted sorting information size, at the cost of increased conversion error. In Figure 3, the increase of the conversion error in a

specific frequency band, as a function of the multiplier, is shown. The error shown is the relative conversion error, i.e., the (Euclidean) distance between the converted and target vectors over the distance between the source and target vectors. If the multiplier increases excessively the conversion accuracy deteriorates at a point where audible distortion appears in the resynthesized channel. Fortunately, the psychoacoustic model described previously averts this by selecting the maximum multiplier that will result in an acceptable distortion level.

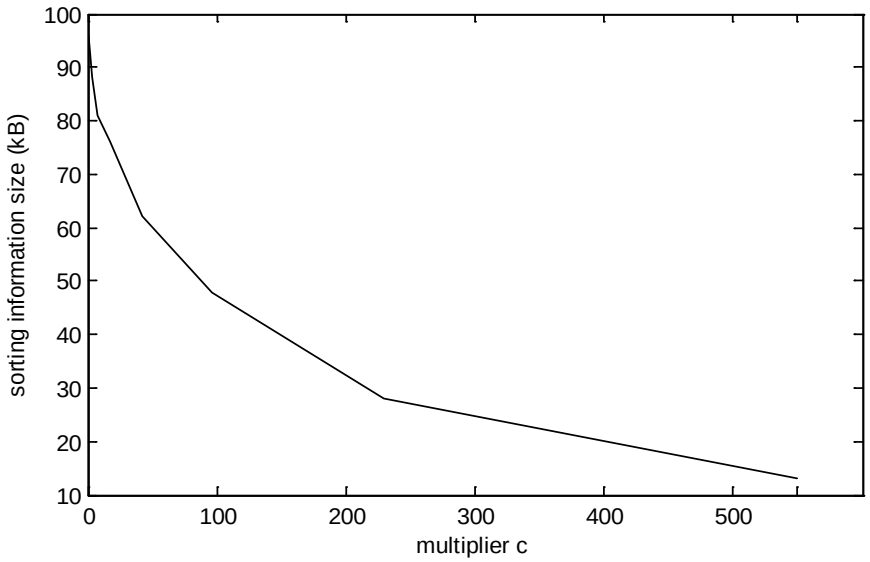

Figure 2: An example of the relation between the multiplier, *c*, and the size of the sorting information, in kB, for the residual vector conversion of subband 2 (690 Hz - 1380 Hz) of a random music piece.

In Figure 4, the overall audio quality of the resynthesized channel-signal is plotted against the bitrate of the total transmitted parameters. The reference signal against which the resynthesized signal is compared during the PEAQ test is the original target channel. Naturally, the audio quality improves as the available bitrate increases, i.e. from -4 to -0.5, corresponding to a bitrate from 60 kbps to

270 kbps. For comparison, the full bitrate of the original, uncompressed, channel-signal is typically around 800 kbps. This means that there is at least a 65% reduction in transmission size achieved by our algorithm. The net savings in kB scale up if, instead of one channel only, more channel-signals are resynthesized (e.g. 5 channels) using the same, predetermined, channel as a source signal.

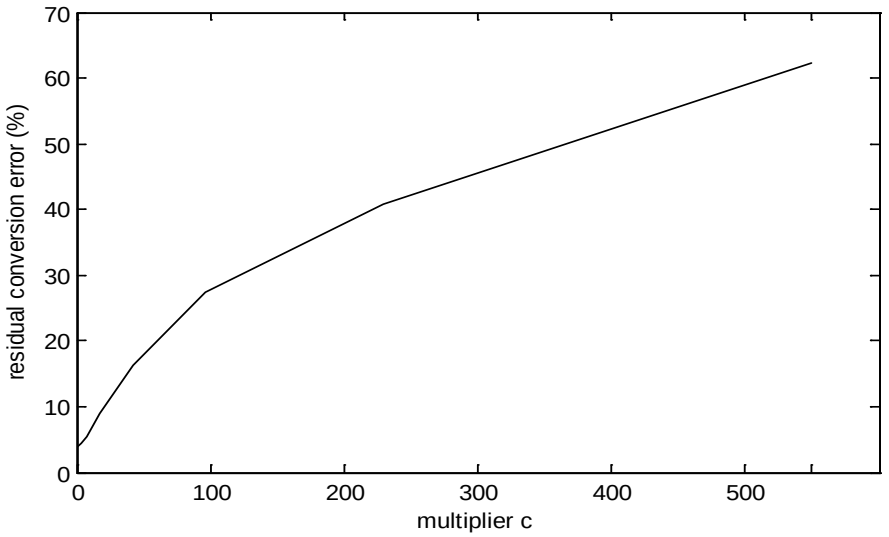

Figure 3: An example of the relation between the multiplier, *c*, and the residual conversion error for the residual vector conversion in the frequency range 690 Hz - 1380 Hz of a random music piece.

5 Conclusions

A novel multichannel coding algorithm was presented, based on statistical conversion of feature vectors and their residuals. It exploits the information redundancy that individual channels of the same multichannel audio piece naturally share in order to convert one channel-signal into another channel-signal with the minimum amount of parameters transmitted or encoded. The results show

that the transmission or encoding overhead reduction is at least 65% as compared to the size of the original channel-signal that is resynthesized. Further savings are attained in case the number of resynthesized channels grows and the same source channel-signal is employed. Future research on the psychoacoustic criterion used for bitrate control could yield even higher audio quality for the same size of transmission or encoding overhead.

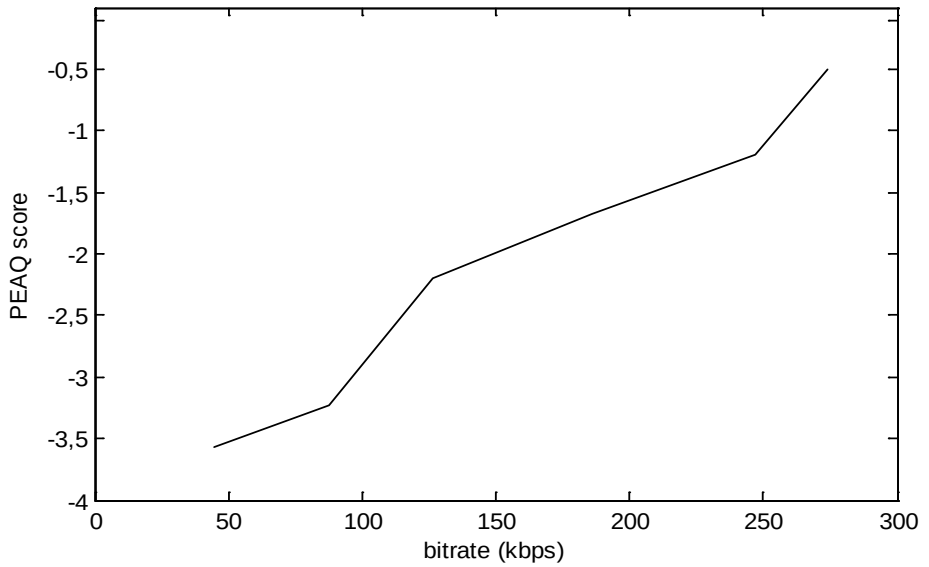

Figure 4: Audio quality (PEAQ) in relation to the total transmission bitrate of the resynthesized channel for the rock music multichannel recording. Higher score means better audio quality.

References

[1] K. Brandenburg, Evaluation of Quality for Audio Encoding at Low Bitrates, 82nd *AES Convention*, London, UK, (March, 1987), preprint 2433.

[2] D. Cantzos, A. Mouchtaris and C. Kyriakakis, Enhanced Multichannel Audio Resynthesis through Residual Processing and Features Alignment, *IEEE*

Proc. Int. Conf. Multimedia and Expo (ICME), Beijing, China, (July, 2007), 1267-1270.

[3] D. Cantzos, A. Mouchtaris and C. Kyriakakis, Multichannel Audio Resynthesis Based on a Generalized Gaussian Mixture Model and Cepstral Smoothing, *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (WASPAA), New Paltz, NY, (October, 2005), 215-218.

[4] D. Cantzos, A. Mouchtaris and C. Kyriakakis, Perceptually-Driven Scalable MDCT Enhancement of Compressed Audio Based on Statistical Conversion, *IEEE Proc. Int. Symposium on Multimedia* (ISM), Dana Point, CA, (December, 2011), 41-46.

[5] C. Faller and F. Baumgarte, *Binaural Cue Coding Applied to Stereo and Multi-channel Audio Compression*, 112th *AES Convention,* Munich, Germany, (2002), preprint 5574.

[6] J. Herre, K. Brandenburg and D. Lederer, Intensity Stereo Coding, 96th *AES Convention*, Amsterdam, Netherlands, (1994), preprint 3799.

[7] Information technology - Coding of Audiovisual Objects, Part 3: Audio, *ISO/IEC* 14496-3, (1999).

[8] Methods for Objective Measurements of Perceptual Audio Quality, *International Telecommunications Union*, Geneva, Switzerland, ITU-R Rec. BS.1387, (1999).

[9] A. Mouchtaris, S.S. Narayanan and C. Kyriakakis, Multi-resolution Spectral Conversion for Multichannel Audio Resynthesis, *IEEE Proc. Int. Conf. Multimedia and Expo* (ICME), Lausanne, Switzerland, (August, 2002), **2**, 273-276.

[10] P. Noll, *MPEG Digital Audio Coding Standards*, CRC Press LLC, 2000.

[11] L. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, New Jersey, 1993.

[12] Y. Stylianou, O. Cappe and E. Moulines, Continuous Probabilistic Transform for Voice Conversion, *IEEE Transactions on Speech and Audio Processing*, **6**(2), (March, 1998), 131-142.

[13] W.C. Treurniet and G.A. Soulodre, Evaluation of the ITU-R Objective Audio Quality Measurement Method, *J. Audio Eng. Soc*., **48**(3), (March, 2000), 164-173.