STEREO AND MULTICHANNEL AUDIO CODING WITH ROOM RESPONSE COMPENSATION FOR IMPROVED CODING TRANSPARENCY

Maciej Bartkowiak

Institute of Electronics and Telecommunications, Poznan University of Technology
Piotrowo 3A, 60-965, Poznan, Poland
phone: + (48 61) 6652864, fax: + (48 61) 6652 572, email: mbartkow@et.put.poznan.pl
web: http://www.multimedia.edu.pl

ABSTRACT

The paper presents a very simple enhancement of joint coding of stereo and surround channels within a perceptual audio codec. Moreover, the paper proposes two improvements to standard parametric stereo and spatial audio compression in order to avoid the smearing of transients in the process of channel downmixing. The improvements consist in compensating of inter-channel delays prior to mixdown, as well as additional encoding of the room response for realistic reconstruction of the stereo ambience.

1. INTRODUCTION

1.1 Spatial recording and its model

Modern stereo and surround audio programmes containing many sound sources are usually created from multichannel capture with two or more microphones, with additional help of spatial effect processors. These processors are used to enhance the natural ambience of the recording venue, or even to create an artificial ambience in the case of dry recording in a studio. Using several microphones is equivalent to sampling the three-dimensional acoustic wave produced by an ensemble of instruments in the same instant, but in different points in space (fig. 1).

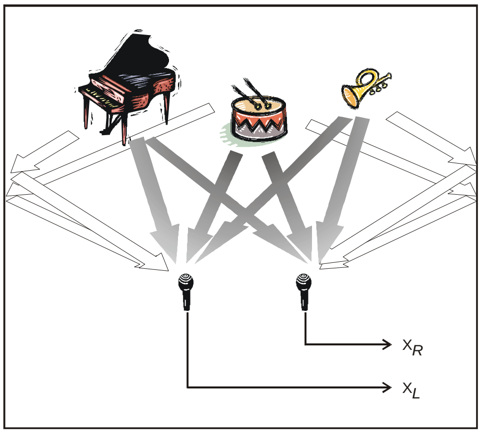

Figure 1. Stereo recording setup with some acoustic wave propagation paths shown (direct – shaded arrows, reflected – white arrows)

For each sound source, differences in signal magnitude and arrival time to the microphones are responsible for the impression of certain spatial location recreated at the listener site, according to spatial hearing cues: IID (inter-aural intensity difference) and ITD (inter-aural time difference, or corresponding inter-aural phase difference, IPD). Different wall reflection patterns creating certain ambience also captured by the microphones are responsible for the mutual incoherence of the recorded signals and invoke an impression of spaciousness and width.

A simplified model of the stereo recording (fig. 2) is proposed and features a separate signal path from each of the sources to each of the stereo channels which is decomposed into simple delays and a time invariant non-minimal phase impulse response to simulate early reflections. In this model a common impulse response for each channel is assumed (after compensating the delay differences), which does not represent the real situation, wherein the impulse response of an acoustic space significantly depends on the position of the signal source. Our experiments show however that even with this simplification it is possible to recreate a very realistic ambience of the original venue. A straightforward extension of this model to multichannel recordings is also possible.

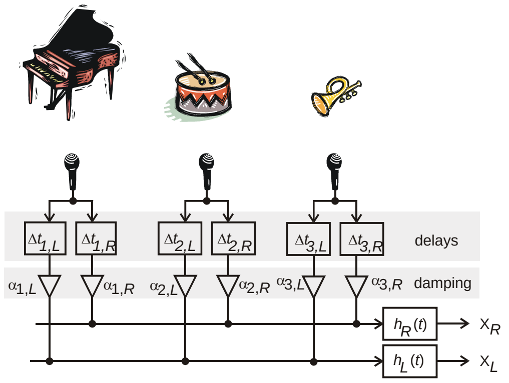

Figure 2. The proposed simplified model of the stereo recording

A rationale behind this model is that room response usually does not change much during the recording session. Since removing redundancy is a significant part of audio coding process, separating the reflection pattern from the content, if succeeded, reduces long term correlation, simplifies the signal and increases the efficiency of perceptual coding at the cost of additional transmission of the room response.

1.2 Traditional encoding of spatial audio

Perceptual audio coders apply various means of channel combination in the case of joint compression of stereo and multichannel programme [1]. Traditionally, simple M/S (middle/side) channel matrixing for low and middle frequency range is combined with intensity (fake stereo) coding of higher frequencies. M/S differential coding on one hand results in better efficiency due to energy compaction (most signal energy is concentrated in one channel M, leaving low-energy residual in the S channel), on the other hand it allows for better control over possible unmasking of the quantization noise. Our previous results showed that by introduction of optimal scaling coefficients it is possible to increase slightly the coding gain due to matrixing [2]. Our coefficients are calculated from two-dimensional PCA analysis of the $[x_L, x_R]$ signal channels pairs on the subband basis. Optimal (adaptive) matrixing in the form of free rotation of subband components in the stereo plane offers energy compaction gain 4÷6dB better than traditional matrixing at the cost of marginal side information. Unfortunately, minute temporal misalignment between channel signals result in significant phase shifts in high frequency components, which destroys the coding gain due to matrixing in this frequency range.

Intensity stereo coding exploits spatial hearing that relies on magnitude envelope rather than phase differences. Only one (M) channel signal is perceptually encoded together with separate scaling factors for stereo channels, and appropriate intensity panning is performed at the decoder side. This approach does not offer a truthful representation of the auditory scene, but it is very efficient in high frequency range.

1.3 Parametric stereo and surround coding

A new technique for parametric coding of the spatial information [3,4] has been proposed and recently standardised by ISO in an MPEG-4 Audio tool [5]. As an extension of intensity coding mode, additional side information is transmitted together with one monaural signal being a mixed down version of the original programme (channels simply added together). This side information (IID_b, IPD_b and ICC_b values estimated in 64 or 72 subbands) allows for more realistic reconstruction of the stereo field by means of appropriate matrix transformations on the decoded signal (fig. 3), which introduce temporal offsets and intensity scaling. A secondary decorrelated version of the signal is artificially produced using allpass filters [6]. Inter-channel coherence (ICC_b) parameters control how these two signals are combined in order to create the illusion of spatial width of the auditory image. Similar scenario is employed in multichannel audio coding, wherein appropriate sets of parameters for each of the surround channels are calculated with respect to a single downmix channel, and transmitted to the decoder. Parametric stereo (PS) coding offers tremendous progress in compression efficiency for low bit rates.

2. CODING WITH CHANNEL DELAY COMPENSATION

Temporal offsets between components of the stereo signals corresponding to individual instruments are particularly apparent on transients (fig. 4) which are important cues in spatial hearing.

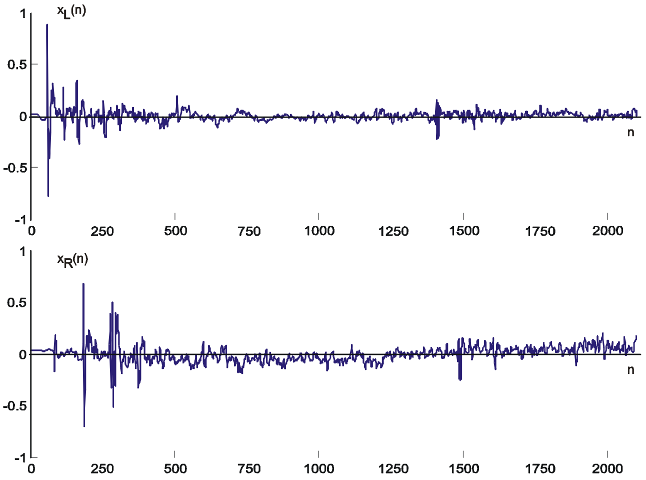

Figure 4. Signals recorded from left and right microphone show significant temporal offset [samples]

Simple channel mixdown practiced in standard joint stereo coding as well as in PS coding techniques do not take these shifts into account. Resulting monaural signal (fig. 5) exhibits significant smearing of the transients and unnatural coloration due to a well known comb filtering effect. Thus, it is not a good monaural representation of the spatial audio programme. Not only such signal is more complex and harder to encode (due to series of multiple transients), but in the

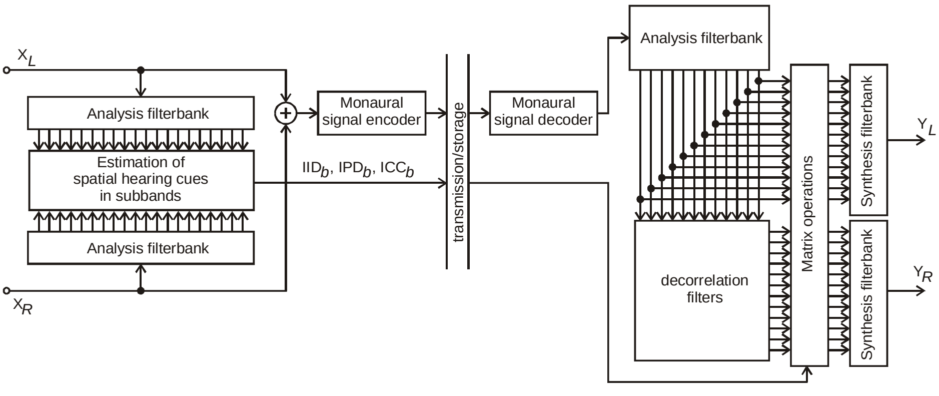

Figure 3 – The general diagram of parametric stereo coding

case of parametric stereo it is also practically impossible to reconstruct individual channel signals using linear transformations at the decoder side. In our opinion this is one of the reasons PS coding is not able to achieve transparent quality at high bit rates.

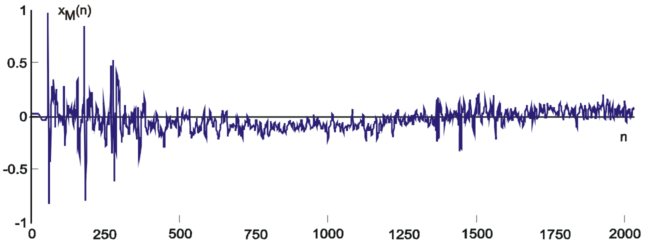

Figure 5 – Mixdown signal with smeared/multiple transients

Compensation of inter-channel time shifts is proposed to be applied prior to stereo or surround channel matrixing (in the case of joint stereo coding) or mixdown (in the case of parametric stereo). As individual instrument signals are usually not available, compensation is done on a subband basis. For efficient manipulation of signal components, a complex exponential modulated pseudo-QMF filterbank, as defined in [5], is applied according to (1).

$$x^k(n) = \sum_{m=0}^{639} x(64n - m)\, c(m)\, e^{j\frac{\pi}{128}(k+0.5)(2m+1)}, \qquad (1)$$

where c denotes the impulse response of the lowpass prototype. Such a filterbank provides a uniform decomposition of the signal into 64 subbands with very low aliasing (fig. 6) and analytic-like representation, allowing for advanced manipulation of the magnitude and phase, without artifacts.

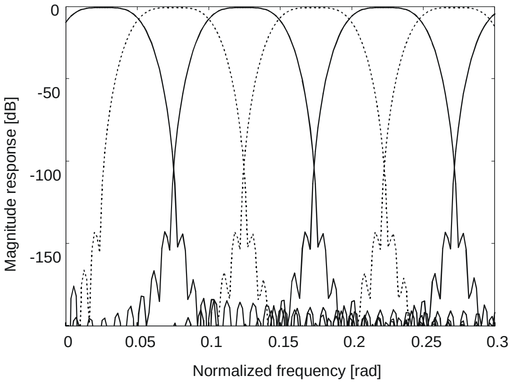

Figure 6 – Frequency response of the complex QMF filterbank (subbands 0...5 shown)

Time differences between stereo channels are easily determined and compensated through phase differences between corresponding subband samples (2).

$$\Delta\phi^k(n) = \arg\{x_L^k(n)\} - \arg\{x_R^k(n)\} \qquad (2)$$

Phase compensation is performed by half-way rotation of the complex vectors in the stereo plane (4). In order to avoid artifacts due to rapid phase changes between consecutive subband samples, the phase difference function is smoothed out using spline kernel ζ.

Final derivation of the compensated sum and difference signals requires prior calculation of the optimal weights which are obtained through PCA analysis of the covariance matrix (3,4), where φ_k denotes a rotation angle of the stereo vector in the signal plane, that maximizes the energy of x_M.

$$\begin{bmatrix} x_M^k(n) \\ x_S^k(n) \end{bmatrix} = \begin{bmatrix} \cos(\varphi_k) & \sin(\varphi_k) \\ -\sin(\varphi_k) & \cos(\varphi_k) \end{bmatrix} \underline{\tilde{x}}^k(n), \qquad (3)$$

and $\varphi_k = \frac{1}{2}\tan^{-1}\dfrac{r_{LR} + r_{RL}}{r_{LL} - r_{RR}}$, is calculated from (4).

$$\begin{bmatrix} r_{LL} & r_{LR} \\ r_{RL} & r_{RR} \end{bmatrix} = \mathbf{R}^k{}_{xx} = \frac{1}{N}\sum_{n=0}^{N-1} \underline{\tilde{x}}(n)\, \underline{\tilde{x}}^{*T}(n), \qquad (4)$$

and

$$\underline{\tilde{x}}^k(n) = \begin{bmatrix} \exp(0.5\, j\Delta\phi^k(n) * \xi) & 0 \\ 0 & \exp(-0.5\, j\Delta\phi^k(n) * \xi) \end{bmatrix} \underline{x}^k(n)$$

In the final step, all subband signals are combined together using a corresponding synthesis QMF filterbank in order to reconstruct the full band x_M (fig. 7) and x_S signals.

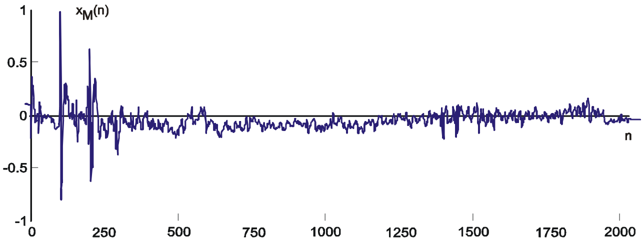

Figure 7 – Delay compensated $\boldsymbol{x_M}$ signal obtained from (3)

For joint stereo mode, both resulting signals are used for subsequent perceptual coding. However, due to very low energy of the difference component x_S, as well as its residual character, the perceptual model is usually inadequate. Therefore, x_S is encoded using a simplified transform codec, with bit allocation derived from quantization noise levels determined for simultaneously processed x_M.

In the case of PS-based coding, the x_S signal is not calculated at all. Only the mixdown x_M is transmitted, and the spatial cues are determined with respect to this reference. In fact, the IID and IPD parameters may be directly derived from the matrix coefficients used in (3) and the phase differences $\Delta\phi^k$, in (4), respectively.

3. ROOM RESPONSE CODING

In parametric stereo decoder, artificial ambience is generated by the application of allpass decorrelators or reverberators [3,4,6]. This allows for recreation of the synthetic incoherence of output left and right signal, necessary for invoking the impression of width. A common problem observed in current implementations of PS-based codecs is that the perception of such ambience is not very realistic and the decoded programme sounds "artificial".

Within the proposed modification, the process of decorrelation is replaced by a convolution with separately encoded impulse responses $h_R(n)$, $h_L(n)$, estimated from the original signal. The purpose of the impulse responses is to recreate the delay-compensated left and right signal from the mixdown monaural x_M. Since the natural acoustic properties of the recording venue are rather constant, the impulse responses are transmitted to the decoder as side information, and are updated infrequently.

In the encoder, the impulse responses $h_R(n)$, $h_L(n)$ are obtained from a special filter design procedure that involves cepstral analysis and optimizations. The description of the algorithm is skipped due to limited space.

In the modified parametric stereo decoder, the monaural signal x_M is convolved with $h_R(t)$ and $h_L(t)$ in order to obtain the two prototype left and right signals. Since the impulse responses are rather long and such convolution is computationally intensive, it may be performed in QMF domain, similarly to other reconstruction operations. The resulting prototype signals are subsequently scaled and phase shifted according to the parameters derived from IID and IPD values, as described in [5].

4. EXPERIMENTAL RESULTS

The proposed modifications is implemented and tested in Matlab environment. Since these enhancements deal with perceptual coding and their aim is to improve the perceived quality, a series of subjective listening tests is conducted according to the MUSHRA methodology [8]. For fair comparison, all coding scenarios use the same core encoder based on LC-AAC specification, implemented in Matlab. A test suite is prepared consisting of 7 pieces of music with various complexity and richness of their stereo image, including jazz and orchestral music, as well as a well-known holophonic recording. Instead of typical 3.5kHz and 7kHz anchors, a monophonic and a pseudo-stereo (fixed intensity panning) version of the recordings are used. These are compared with our technique (enhanced parametric), as well as with traditional joint stereo and parametric stereo scheme [5] at 48kb/s (fig. 8.), 64kb/s and 96kb/s. The listeners are instructed to judge the quality of the stereo image while disregarding the coding artifacts, however due to different coding efficiency of the compared techniques and the target bit rate being strictly fixed, certain quality differences are noticeable and may influence the score. The general conclusion is that in most cases the perceived soundstage width and transparency improved both with respect to traditional joint coding as well as parametric stereo coding. As expected, best results were obtained with audiophile recordings of acoustic music featuring few instruments. Perhaps the most interesting general observation from all experiments is that perceptual stereo coding with delay and room response compensation offers almost transparent stereo quality at sufficiently high bit rates (about 9.6 kb/s for the parametric part of the bitstream, including side information overhead required for delays and stereo impulse responses)

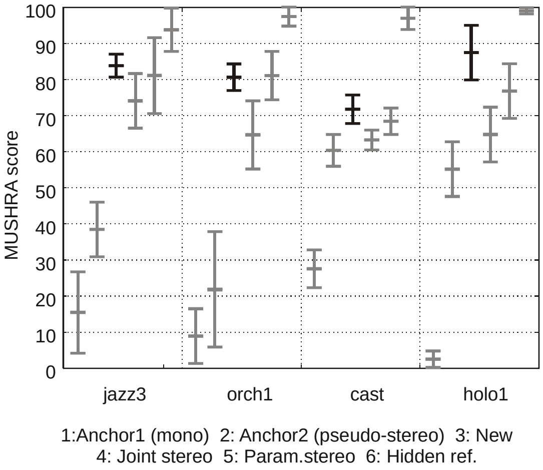

Figure 8 – Selected MUSHRA listening test results with 95% confidence interval.

5. CONCLUSIONS

The paper presents two simple enhancements applicable to joint stereo coding and parametric stereo coding as well as for encoding of multichannel programmes. By introduction of additional compensating delays into the process of channel matrixing or mixdown, a more appropriate monaural version of the spatial audio signal is created. This leads to significantly reduced residual signal in the case of joint stereo coding, or more transparent reconstruction of the spatial audio in the case of parametric stereo. Application of realistic room responses instead of artificial reverberation allows for more natural sounding ambience and thus expands the application range of parametric stereo coding to high quality uses. The proposed techniques are extensible to multichannel audio.

REFERENCES

[1] J. D. Johnston, A. J. Ferreira, "Sum-difference stereo transform coding ", Proc. *ICASSP'92*, March 1992, pp. 569 - 572

[2] M. Bartkowiak, T. Żernicki, "A Simple Adaptive Matrixing Scheme for Efficient Coding of Stereo Sound", *Proc. EUSIPCO'05*, Antlya, Turkey, 2005

[3] C. Faller, "Parametric Coding of Spatial Audio", in *Proc. DAFx'04*, Naples, Italy, October 2004, pp. 151-156

[4] J. Breebaart, S. v.d. Par, A. Kohlaursch, E. Schuijers, "Parametric Coding of Stereo Audio", *EURASIP Journal on Applied Signal Processing*, vol. 9, pp. 1305-1322, 2005

[5] ISO/IEC 14496-3:2001/AMD1, "Bandwidth extension"

[6] J. Engdegård, H. Purnhagen, J. Röden, L. Liljeryd, "Synthetic Ambience in Parametric Stereo Coding", *Proc. 166th AES Convention*, Berlin, Germany, May 2004

[7] D. Bees, P. Kabal, M. Blostein, "Application of Complex Cepstrum to Acoustic Dereverberation", in *Proc. Biennial Symp. Commun.*, Kingston, June 1990, pp. 324-327

[8] ITU-R document 10-11Q/33, "A Method of Subjective Listening Tests of Intermediate Audio Quality - Contribution from the EBU to ITU Working Party 10-11Q"