A SIMPLE ADAPTIVE MATRIXING SCHEME FOR EFFICIENT CODING OF STEREO SOUND*

Maciej Bartkowiak, Tomasz Żernicki

Institute of Electronics and Telecommunications, Poznań University of Technology
Piotrowo 3A, 60-965 Poznań, Poland
phone: + (48) 616652862, fax: + (48) 6652572, email: mbartkow@et.put.poznan.pl
web: www.multimedia.edu.pl

ABSTRACT

A generalized scheme for mapping stereo audio channels onto coding channels is considered in the paper. An improved joint stereo transform-based perceptual audio coder is proposed that employs this scheme instead of MS coding. The discussed frequency-dependent matrixing scheme uses a free rotation of the two-dimensional signal space adapted to the dominant direction of arrival within particular frequency bands. This direction is determined using 2-D principal component analysis (PCA) in subbands or groups of transform coefficients. Experimental results show a significantly increased compression efficiency compared to traditional MS coding based on a static matrix. The proposed technique is also more compatible with mono mode since the new M channel represents signal energy that is maximally focused within each band, and the residual energy in S channel is minimized.

1. INTRODUCTION

1.1 Stereo audio production and encoding

Typical audio recording released in stereo or multichannel format represents many sound sources located in the acoustic space around the target virtual listener. These sounds are recorded using pairs of microphones or single direc-tional microphones. The final recording is mixed in the studio with the inclusion of spatial positioning effect processors. Most recording techniques and the practices of mixing result in stereo programme whose components are significantly correlated. In particular, the low frequency content in all channels is usually very similar, which is particularly observed in the low frequency range where the acoustic wave imposes similar pressure at the capsules of each microphone present at the recording session.

Efficient compression of stereo and multichannel audio usually involves some attempts at exploiting the inter-channel redundancy as well as exploiting human deficiencies in spatial hearing reflected by popular auditory phenomena such as phantom imaging. Traditional techniques commonly employed in joint stereo coding are MS matrixing and intensity coding. A new progress in this field is a recently developed binaural cue coding and parametric stereo coding technique.

1.2 M/S and intensity coding

Compression in MS mode consists in calculating two new signals x_M and x_S out of the input L and R components of the stereo pair (1)

$$x_M = \frac{1}{\sqrt{2}}\left(x_L + x_R\right), \quad x_S = \frac{1}{\sqrt{2}}\left(x_L - x_R\right) \tag{1}$$

Coding of these new signals is often more efficient than independent compression of x_L and x_R because similar components add in the x_M signal while they cancel each other in the x_S signal, so its entropy is also decreased and the required bit stream decreases significantly [1, 2]. At the decoder side, the matrixing procedure is reversed, which brings the possibility of accidental revealing the coding artefacts (e.g. unmasking the quantization noise) [3]. With perceptual coding in MS mode the psychoacoustic model is usually more complex than with independent channel coding, since it requires a conservative estimation of the masking profile on the worst-case basis.

Compression in intensity mode [4] is a simplified approach that rejects the residual component x_S carrying spatial information and relies on phantom imaging. Only the x_M component is encoded using an appropriate technique. At the decoder side both x_L and x_R signals are reconstructed by appropriate scaling of the spectral content of x_M usually represented by cosine transform coefficients grouped in separate *scaling bands*. A big advantage of intensity coding is a high compression gain achieved at the cost of reduced accuracy in auditory spatial imaging. In practice, this is very well tolerable for high frequencies, where spatial hearing relies on the volume envelope rather than on phase differences between spectral components.

All compression scenarios that operate in frequency domain allow to adaptively switch between various stereo modes depending on the frequency and the content of the signal. For example, MPEG-1 Layer-3, and MPEG-2/4 AAC apply MS mode for low and middle frequency range and intensity mode for high frequencies.

1.3 Binaural cue coding and parametric stereo

Recently proposed binary cue coding (BCC) and parametric stereo coding [5] is a more systematic exploitation of the properties of spatial hearing besides redundancy reduction in

* The research was supported by the Polish Ministry of Scientific Research and Information Technology

stereo programme. Several parameters are extracted from the input signal: interaural time difference (ITD), interaural level difference (ILD) and interaural coherence (ICC) which allow to resynthesize the realistic spatial information at the decoder side. Very high compression gain is achieved by parametric representation of all spatial information, while the spectral content is encoded using only one channel being a downmixed version of the stereo or multichannel source. Similarly to intensity coding, the main advantage of such approach is the number of bits required by the parameters being of order of magnitude less than the number required by encoding the difference signal, x_S. While claimed to be more truthful in recreating a wide spatial auditory image, BCC may still be considered as a generalized extension of the intensity coding strategy. Main points against parametric stereo are focused around its artificial sound, therefore it is rather predestined to low bit rate applications.

1.4 Adaptive matrixing

Our approach is a generalization of standard matrixing aimed at better decorrelation between the two resulting signals and higher energy compaction. It may be used together with MS-like coding, intensity coding or parametric stereo with a guaranteed improvement over the traditional methods. In the case of further treatment of both matrixed signals, a consistently better efficiency is observed over MS. In the case of intensity stereo or parametric stereo the only encoded channel M contains a better approximation of the monophonic content, because accidental cancelling of spectral components that are out of phase is avoided. Thus, a more representative and complete spectral content is used as a basis for a resynthesis of the auditory spatial image at the decoder side, which promises higher fidelity at the output.

2. PROPOSED TECHNIQUE

2.1 The principle

Signal matrixing may be considered as a mapping of vectors representing sample values from individual channels in a plane representing the virtual auditory space. In the case of standard MS matrixing this mapping is a projection of two-dimensional [x_L x_R] vectors onto the *m* and *s* orthogonal axes. A generalized matrixing may be thought of as an arbitrary invertible mapping. The matrixing considered in this paper is a free rotation of the [x_L x_R] vectors (2) which is adapted to the signal content through appropriate selection of the **H** matrix.

$$\begin{bmatrix} x_M \\ x_S \end{bmatrix} = \mathbf{H} \begin{bmatrix} x_L \\ x_R \end{bmatrix}, \quad \|\mathbf{H}\| = 1 \tag{2}$$

The aim of the transformation is maximum energy compaction into one channel, therefore the optimal rotation is one according to the direction of principal eigenvector of $\mathbf{R}_{xx}$, the covariance matrix of [x_L x_R] vectors. This direction corresponds to the direction of arrival as perceived by the listener.

Since the input signal is usually a mixture of sounds from various sources located in the virtual stereo space, it consists of many components with different correlation factors. Therefore, it is beneficial to apply different rotations **H** to those components of [x_L x_R] whose covariance matrices differ significantly. A separation of sources in the input signal may be achieved though independent component analysis (ICA), however for the purpose of coding, spectral decomposition may be used as well. In the general form, our matrixing may be a preprocessing step applied to the input signal independently from the following coder (fig. 1).

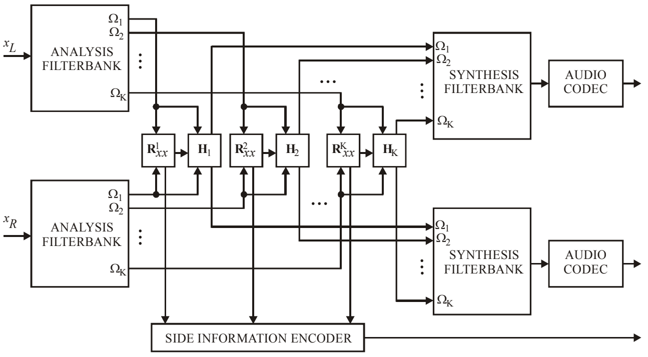

Fig. 1. General structure of the proposed matrix preprocessor

For the purpose of frequency-dependent decorrelation, input signal is decomposed into several non-overlapping subbands and processed on the frame-by-frame basis. In each subband Ω_k, the components of $\mathbf{H}_k$ are calculated as in (3).

$$\mathbf{H}_k = \begin{bmatrix} \cos(\varphi_k) & \sin(\varphi_k) \\ -\sin(\varphi_k) & \cos(\varphi_k) \end{bmatrix}, \tag{3a}$$

where

$$\varphi_k = \tfrac{1}{2} \tan^{-1} \frac{r_{LR} + r_{RL}}{r_{LL} - r_{RR}}, \tag{3b}$$

and $r_{..}$ are the elements of the covariance matrix $\mathbf{R}^k_{xx}$ (4).

$$\mathbf{R}^k{}_{xx} = \frac{1}{|\Omega_k|} \sum_{\underline{x} \in \Omega_k} \underline{x}\,\underline{x}^T = \begin{bmatrix} r_{LL} & r_{LR} \\ r_{RL} & r_{RR} \end{bmatrix}, \text{ where } \underline{x} = \begin{bmatrix} x_L \\ x_R \end{bmatrix} \tag{4}$$

Aplication of (2) in each subband leads to optimal rotation of the vector [x_L x_R] so that the resulting x_M contains most of the signal energy, while x_S contains the residual. Since the correlation between spectral partials in $\underline{x}$ may change in time, the values of φ_k should be updated from frame to frame. In order to reconstruct the LR signal at the decoder side, the information on $\{\varphi_k\}$ should be preserved and transmitted as side data.

The advantage of the new mapping over the traditional MS matrixing may be illustrated on a simplistic example (fig. 2). Consider a stereo signal that contains frequency components that do not overlap: a common part $X_C(f)$ is present in both channels, while parts $X_L(f)$ and $X_R(f)$ are panned hard left and hard right respectively (they appear only in one channel).

The [x_M x_S] pair resulting from standard mapping contains even more information than the input LR signal, therefore the perceptual model within the encoder will more likely select LR option, otherwise the compression efficiency would decrease. In the case of adaptive matrixing, all the spectral content is located in the M channel, and the S channel contains no residual at all. Of course this idealistic situation is not likely to happen in real world, mostly because L and R components in a stereo programme differ not only in intensity.

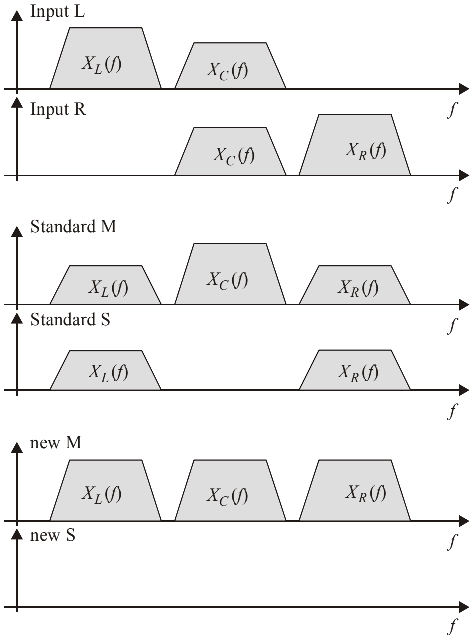

Fig. 2. An example showing the advantage of adaptive matrixing over traditional MS matrixing.

2.2 The significance of the filterbank

The most advantageous spectral decomposition of the input signal is such that components coming from differently spaced sources are separated into different subbands. On one hand, the narrower the subbands, the more effective is said separation, on another – high number of subbands is prohibitive in efficient coding due to high amount of side data. The problem may be solved either by using adaptive decomposition with variable filterbank or by a fixed filterbank with perceptually optimized spacing of the subbands. However, if full compatibility of the M channel with mono mode is expected, a special structure of the filterbank that avoids time-domain aliasing within the subbands may be required.

In our experiments very good results were achieved using an oversampled O^2-DFT-based 1024-channel filterbank, whose coefficients were grouped into ranges corresponding to critical bands. Since the latter partitioning corresponds to scaling bands within a perceptual coder, it is also suggested that frequency partitioning within the coder's own filterbank might be exploited for the purpose of matrixing.

2.3 Encoding of the $\{\varphi_k\}$ information

Since transformation (2) is fully invertible, it is possible to convert the decoded $[x_M\ x_S]$ vector back to LR representation, provided the information on $\{\varphi_k\}$ is available. The amount of side data representing the values of $\{\varphi_k\}$ should be marginal in order to benefit from the increased efficiency of subsequent signal compression. Therefore, quantization of the estimated optimal values to a low number of allowed angles is considered, enabling efficient representation of these values on a low number of bits. In our experiments very good results are obtained while limiting the number of allowed values to 16 (cf fig. 3) and applying predictive differential coding to reduce the bit rate.

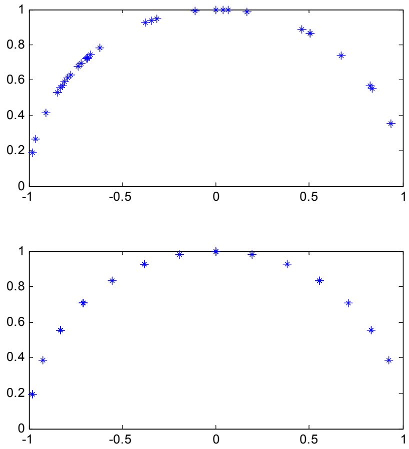

Fig. 3. Example set of directions $\{\varphi_k\}$ estimated from analysis of a test stereo recording (upper plot), and corresponding quantized values encoded in the side data (below). Here, the origin (0,0) represents the position of listener in the stereo field.

Experimental analysis of a series of recordings reveals that the original values of $\{\varphi_k\}$ sometimes change rapidly from frame to frame (fig. 4), which is not necessarily reflected by a significant change in the stereo image.

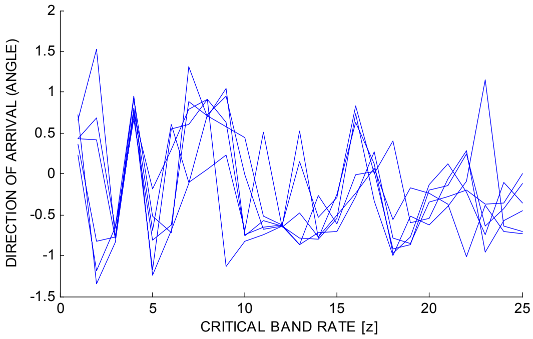

Fig. 4. Temporal changes in estimated directions. Here, values of $\{\varphi_k\}$ are ploted against CBR [Bark] as obtained from six consecutive audio frames of a test recording.

These fluctuations may be accounted for sensibility of the short-term analysis to phase. Rapid changes not only decrease the efficiency of said predictive encoding of $\{\varphi_k\}$, but also generate artefacts in the mapped M channel, if they are not properly compensated by reconstructed S channel (e.g. in case of intensity or parametric stereo encoding). In order to circumvent these undesirable effects, two additional stabilization operations are proposed:

- the estimation of $\mathbf{R}_{xx}$ is based on a long period by taking into account several past frames,

- calculated values of $\{\varphi_k\}$ are smoothed on a frame-by-frame basis through application of LMS algorithm (with low convergence constant μ).

Application of both operations results in very stable final values of $\{\varphi_k\}$ and very efficient compression without significant decrease of the matrixing gain.

3. A MODIFIED PERCEPTUAL CODEC

We propose a modified structure of a perceptual audio codec that employs the adaptive matrixing for improved efficiency (fig. 5).

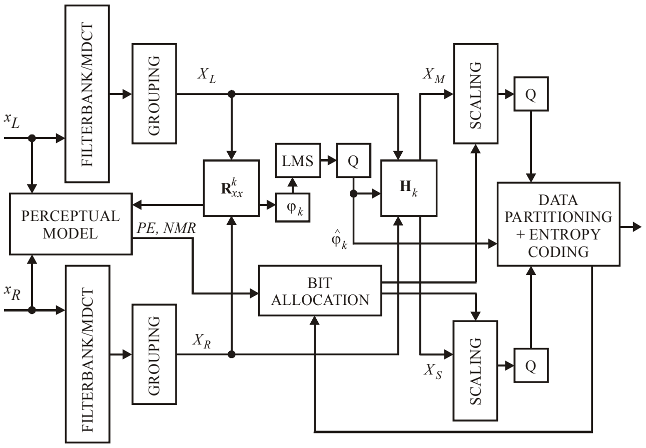

Fig. 5. The structure of the proposed encoder (simplified).

The input signal in both channels is transformed using a complex-MDCT filterbank with time domain aliasing cancellation. The coefficients are partitioned into scaling bands. Individual $\mathbf{R}^k_{xx}$ matrices are calculated for coefficients in consecutive scaling bands Ω_k, and respective optimal direction φ_k in each band is determined. The values of φ_k are subsequently smoothed, quantized and encoded. In order to ensure identical rotations at the encoder and decoder side, the reconstructed values are used in the encoder for final processing. Matrixed MDCT coefficients are obtained from respective $\mathbf{H}_k$ transformations, and subsequent mapping on real-valued MDCT domain through rejection of the imaginary part. These DCT coefficients are processed in a usual manner, i.e. they are appropriately scaled and quantized, which may be preceded by TNS and prediction tools, depending on the operating profile of the encoder (not shown in the picture). Since the proposed matrixing always results in lower entropy, the perceptual model no longer needs to consider LR mode as an alternative to joint stereo coding. The correlation parameters of $\mathbf{R}^k$ are passed to the model in order to take into account the interaural masking phenomena. This is achieved by appropriate weighting of the masking level difference factor [2].

4. EXPERIMENTAL RESULTS

The proposed matrixing scheme is tested experimentally together with the modified perceptual audio coder. The general conclusion from the simulations is that adaptive matrixing results in significantly reduced energy and perceptual entropy of the S channel and slight increase in the M channel, both with respect to standard MS matrixing. When combined with very efficient compression of the values of $\{\varphi_k\}$ (about 1.8-2.0 b/value), the proposed codec offers a significant increase in overall coding efficiency. In order to objectively assess the energy compaction offered by the adaptive matrixing, a measure is proposed based on signal RMS values:

$$\Delta_{MS} = 20\log\frac{U_M}{U_S} - \left|20\log\frac{U_L}{U_R}\right| \quad [\text{dB}] \qquad (5)$$

Table 1. Comparison of matrixing gain (5) between traditional and the proposed adaptive scheme

test track	Δ_{MS}	$\Delta_{\text{new}\,MS}$
symphon	5.6 dB	11.0 dB
triojazz	5.5 dB	12.1 dB
poprock	11.6 dB	16.2 dB
dance	14.9 dB	17.7 dB
hiphop	5.3 dB	9.3 dB
vocal solo	10.7 dB	14.4 dB

As shown in table 1, there is always a clear advantage of the adaptive matrixing over traditional matrixing. The gain is greatest for recordings with very wide stereo base containing several point sources, but even for very narrow recordings (like solo vocal performance) there is a gain of at least 4 dB.

5. CONCLUSIONS

The adaptive stereo matrixing scheme discussed in the paper shows a significant advantage in energy compaction over traditional channel mapping with constant matrix. Application of this scheme to perceptual coding is possible and results in increased compression efficiency. The slightly increased computational complexity is easily compensated by reduced complexity of the perceptual model. Usage of this mapping instead of simple downmixing in parametric stereo coding is also beneficial due to reduced energy of the rejected residual part of the signal.

REFERENCES

[1] J. D. Johnston, "Perceptual transform coding of wideband stereo signals," *Proc. ICASSP'89*, May 1989, pp. 1993 - 1996

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

[3] W. R. Th. ten Kate, P. M. Boers, A. Mäkivarita, J. Kuusama, K. E. Christensen, E. Sørensen, "Matrixing of bit rate reduced audio signals", *Proc. ICASSP-92*, San Francisco 1992, pp. II 205-208

[4] J. Herre, K. Brandenburg, D. Lederer, "Intensity stereo coding", *96th Audio Engineering Society Convention*, Amsterdam, 1994, Preprint no. 3799

[5] Ch. Faller, "Parametric coding of spatial audio", Proc. *DAFX'04*, Naples, Italy, October 5-8, 2004, pp. 151-156