MULTICHANNEL AUDIO DECORRELATION FOR CODING

Soledad Torres-Guijarro, Jon Ander Beracoechea Álava, F. Javier Casajús-Quirós

Dept. of Signals, Systems and Radiocommunications
Polytechnic Univ. of Madrid
marisol@gaps.ssr.upm.es

Luis I. Ortiz-Berenguer

Dept. of Audiovisual and Communication Engineering
Polytechnic Univ. of Madrid
lortiz@diac.upm.es

ABSTRACT

Within digital audio codification, the processing of multi-channel signals has become one of the main fields of research. Current work on the subject look for effective ways to exploit the existing redundancy between the different channels in order to reduce the codification binary rate. This work studies the Karhunen-Loeve Transform (KLT) as a method of decorrelating multi-channel signals prior to coding. Results on codification via AAC are reported.

1. INTRODUCTION

Commercial multichannel audio technologies have evolved from stereo to 5.1 and 10.2, and from theatres to home. There are also new technologies that call for a even higher number of channels, such as wavefield synthesis (WFS, [1]), that tries to recreate 3D audio fields by means of an array of loudspeakers.

The audio material used in these multichannel systems is diverse. The 5.1 sound track of a film is usually obtained by mixing music, dialogs and ambience sounds (doors, steps, …). The resulting multichannel generally shows high correlation between certain channel pairs (L-Ls, R-Rs, C-R, C-L), and low correlation between the rest. Music recordings in 5.1 format for DVD-Audio and SACD show a much higher correlation between any pair of channels. When the recording is made with the objective of sampling the acoustic field with a microphone array with the aim of reproducing it by means of WFS, for instance, the correlation between any pair of channels may be very high. We are focusing on this last kind of recordings composed of a wide number of highly correlated channels.

Two widely used multi-channel codification systems are Dolby AC-3 and MPEG Advanced Audio Coding (AAC). AC-3 is the audio standard chosen for high-resolution television (HDTV), and it is able to compress 5.1 audio signals using 384 kbits/s. AAC is the most powerful multi-channel codification system within the family of MPEG coders at the moment. It is able to compress 5.1 audio signal using 320 Kbits/seg without apparent loss of quality. Both schemes use perceptual models to hide coding distortions. Although they are very powerful systems that support a high number of channels, they are optimized to work with 5.1 recordings. For instance, the operation mode called 'Joint stereo' in AAC tries to exploit the correlation between symmetric channel pairs (by means of Mid/Sum Coding or Intensity Coding), but is unable to eliminate the existing correlation between the rest of the channels. Our aim is to exploit this correlation to reduce the transmission bandwidth while keeping a high quality in the reconstructed signal.

2. MULTICHANNEL RECORDINGS

Original (non-compressed) multi-channel audio recordings are difficult to obtain. For this reason we are employing dry sources and the impulse-response recordings of the varecoic chamber in Bell Labs [2]. This chamber can adjust its reverb time and simulate different audio source locations. The impulse responses are measured in a 22-microphone linear array. Figure 1 shows the position of the 22 microphones and source locations. This arrangement is designed to study the "virtual acoustic opening" [3], which is the kind of application that generates a high order-high correlation multichannel we are interested in.

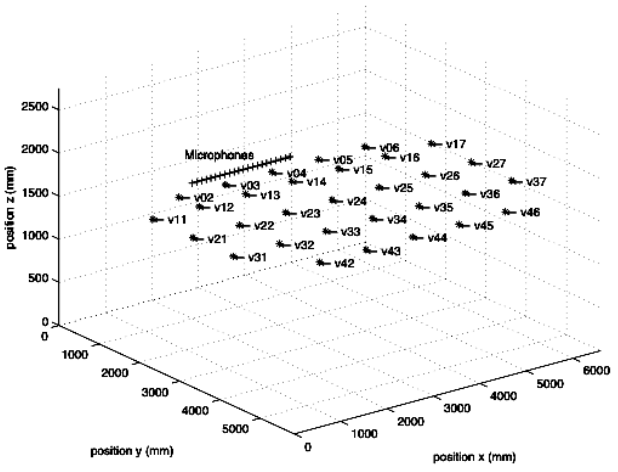

Figure 1: *Position of the microphones and the speakers at the varecoic chamber*

A concert recording has been simulated with a T60 reverberation time of 0,28 seconds, and three monophonic sources sampled at 44.1 kHz corresponding to the tracks of the theme "Worldbeat" (two guitars and a solist voice). The three sources are located in

positions v31, v37 and v24 of the chamber. The resulting mix corresponds to 5 seconds of natural sounding, 22-channel pop music. In order to measure the correlation between different channels in this mix, we can define the normalized cross-covariance matrix of the multichannel. Each element of this symmetric matrix represents the cross-covariance of the corresponding pair of channels, normalized such a way that the covariance of any channel is 1. The normalized cross-covariance matrix of "Worldbeat", plotted in figure 2, shows a high correlation between a significant number of channel pairs, and range from 0.52 to 0.01 besides the main diagonal.

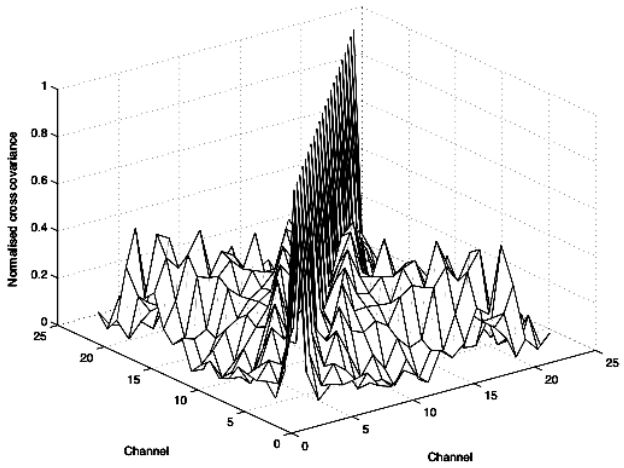

Figure 2: *Normalised cross-covariance of multichannel recording*

Minimizing the correlation between channels should optimize a subsequent coding block. Several strategies have been employed up to now in this direction: discrete cosine transform (DCT), Fourier Transforms, and interchannel lineal prediction, which was shown to be quite inefficient [4]. A recently proposed approach [5] suggest the use of the Karhunen-Loeve transform (KLT).

3. DECORRELATION VIA KLT

The KLT is, from a theoretical point of view, the optimal decorrelation method. As shown in figure 3, it consists on a matrix computation of the form MV=U, where V contains the original multichannel signal, M is composed of the eigenvectors associated to the covariance matrix of V, and U contains the decorrelated signals, that we will call eigenchannels.

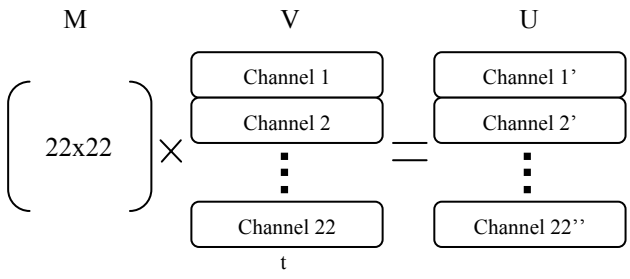

Figure 3: *Karhunen-Loeve transform*

The ability of the KLT to remove the interdependencies between channels can be evaluated computing the cross-covariance matrix of the eigenchannels, and is shown in figure 4. The values off the principal diagonal are below 10^{-5}, which shows the decorrelating efficiency of the KLT.

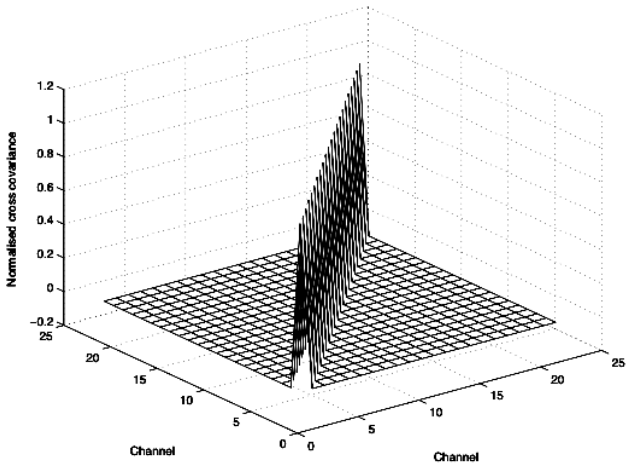

Figure 4: *Normalised cross-covariance of decorrelated multichannel recording*

Other interesting properties of KLT are, on one hand, that it is very easy to invert: thanks to the fact that $M^{-1}=M^{T}$, you only need to transpose M and multiply it by U to recover the original channels. On the other hand, the eigenchannels produced by the transform are audio-like signals from a perceptual point of view, ordered on power basis.

Figure 5 shows channel energies (in percentage) before and after KLT (using 11 channels instead the original 22 in order to reduce computational complexity). As we can see, before KLT most channels have a similar energy; the small differences are due to the position of the microphones related to the sound sources and the acoustical properties of the chamber. However, after KLT most of the energy is concentrated in a few eigenchannels, suggesting several possible coding strategies that will be seen in the following section.

4. CODING STRATEGIES AND QUALITY MEASURES

In order to determine a quality measurement procedure to evaluate the degradation introduced by the proposed coding schemes, it should be kept in mind that the target application of this kind of multi-channel recording is the replication of the original acoustic field after transmission or storage. This means that the quality evaluation should include information about the ability of the whole system, coding included, to preserve spatial information. This leads to subjective tests carried out in a controlled chamber with an appropriate loudspeaker arrangement. The complexity of this approach suggests starting with an objective quality measurement to obtain a general idea of the degradation introduced by the coding process. We have specifically used an implementation of the basic version of recommendation ITU-R BS. 1387-1 [6]. This module measures the perceptual difference between the original and processed

signal by means of the so called Objective Difference Grade (ODG), a figure between 0 and –4, where 0 means "no perceptible degradation" and –4 means "very annoying degradation". The results of this quality measurement applied to the original and reconstructed multichannel on a channel by channel basis will be used from now on to compare different coding strategies. In the subsequent computations, we will reduce the multichannel order by a factor of 2 for complexity reasons, by discarding one out of two adjacent channels. The obtained results can be directly extrapolated to the 22 channel-case.

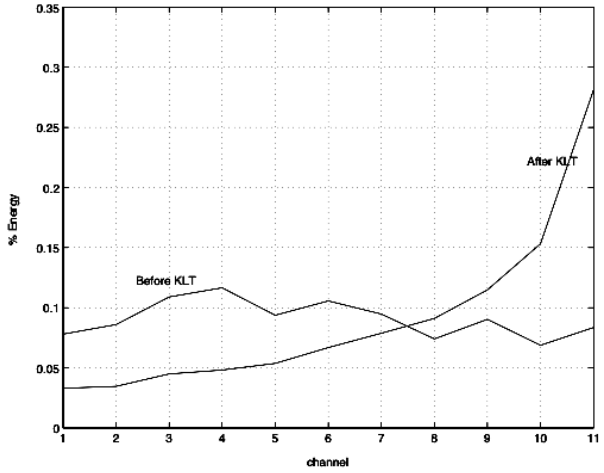

Figure 5: *Energy percentage before and after KLT*

Hence, the KLT gives 11 eigenchannels ordered by power. This suggests a first coding strategy that consists of direct elimination of eigenchannels. This approach is clearly applicable to audio streaming for transmission through internet. When bandwidth restrictions apply, those eigenchannels with smaller energy can be discarded for transmission. Figure 6 shows the mean degradation of the quality of the perceived signal depending on the number of preserved eigenchannels. In each iteration we add one eigenchannel to the inverse KLT computation (beginning with the most energetic one) and then we calculate the mean ODG of the 11 channels after reconstruction.

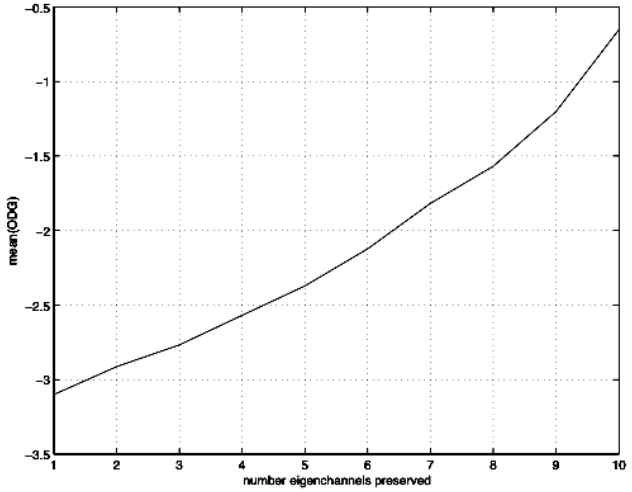

Fig 6: *Mean ODG versus n° of preserved eigenchannels*

As we can see, the relationship between quality and number of preserved eigenchannels follows a monotonic increasing law. This property assures that the quality will not suffer a quick fall as we discard eigenchannels. However, it is clear that, with this approach, the quality suffers from a noticeable degradation even if only one eigenchannel is discarded.

It is also important to remark that we have used the same decorrelating matrix M for the entire audio signal (5 sec). One should think that the decorrelation efficiency of the KLT transform might increase if the transform coefficients are periodically recalculated. To verify this extreme, we have performed the same experiment just reported but for different KLT computation periods. Figure 7 shows the results for KLT recalculation periods corresponding to 2048, 8192, 131072 and 252000 samples (this last period corresponds to the length of the signal, 5.7 sec.). This figure shows that there is no need for updating the KLT coefficients more often than every 5 seconds, at least for "Worldbeat". As a consequence, the coding overhead due to the need of transmitting this coefficients to the decoder will be negligible.

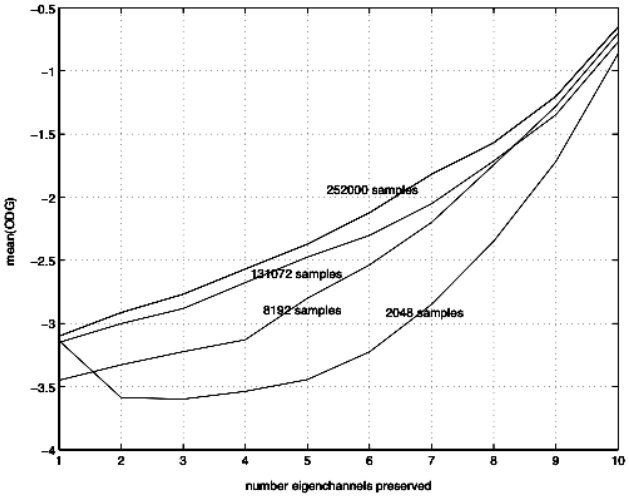

Figure 7: *Mean ODG depending on the n° of preserved eigenchannels, for different KLT computation periods*

Another possible coding approach makes use of the fact that the eigenchannels preserve, generally speaking, the temporal, spectral and perceptual characteristics of audio signals. This makes them suitable to be coded with standard perceptual coders. In the following experiments, a MPEG AAC coder will be employed. Direct codification of the original multichannel with 64 Kbps/channel via AAC gives a mean ODG across the 11 channels of –1.9, while the cascade of KLT+AAC+KLT^{-1}, with the same rate per channel applied this time to the eigenchannels, gives a mean ODG of –1.5, showing the viability of perceptually code the eigenchannels.

Another advantage of coding the eigenchannels instead of the multichannel itself is the possibility of adjusting the bit rate of each eigenchannel depending on its power. Before addressing the bit distribution among channels, we have studied the behaviour of the AAC coder in terms of ODG as a function of

the bit rate. Figure 8 shows the obtained values when AAC is applied to one of the channels of "Worldbeat" (for the rest of the channels, the results are very similar).

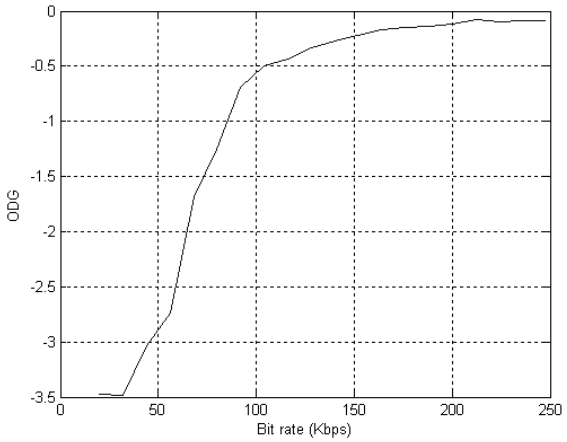

Figure 8: *ODG for one channel coded at different bit rates*

As we can see, the ODG versus bit rate is highly unlinear, presenting a kind of saturation effect at high very high bit rates, and a severe quality drop at low bit rates. This behavior suggests employing a minimum bit rate for all the channels, whatever its energy is, and to distribute de remaining bits attending to the energy percentage of the channel.

The proposed bit rate distribution is expressed in the following equation: $b_i = b_{min} + E_i/E_t \cdot b_r$, where b_i represents the bit rate assigned to eigenchannel i, b_{min} is the minimum bit rate for all the channels, E_i/E_t is the energy percentage of eigenchannel i and b_r is the remaining bit rate after assigning the minimum value to every eigenchannel (bit rate pool).

In the following example, a total bit rate of 704 Kbps, equivalent to 64 kbps/channel (times 11 channels) is to be distributed. Figure 9 shows the mean ODG obtained when fixing b_i to any value between 0 and 64 kbps (1 Kbps steps), and distributing the corresponding bit rate pool among the eigenchannels following a linear law with the energy percentage, as the preceding equation says.

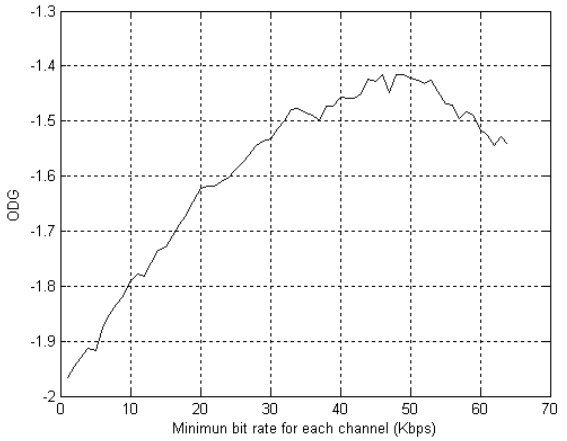

Figure 9: *Mean quality after distributing bits after eigenchannel energy*

The figure shows that selecting a minimum bit rate around 45 kbps/channel, and distributing the remaining bit rate after eigenchannel energy, the mean ODG is 0.1 above the situation where all the channels are coded with exactly 64 Kbps, represented by the last point of the graph. Moreover, if we keep in mind that direct codification of the original multichannel with 64 Kbps/channel via AAC gives a mean ODG of –1.9, we can conclude that decorrelating the multichannel via KLT previous to AAC coding gives a mean ODG improvement of at least 0.5 for a total bit rate equivalent of 64 kbps/channel.

5. CONCLUSIONS

The KLT has proven to be an efficient tool to decorrelate multichannel audio signals prior to coding. The energy-ordering of the decorrelated channels allows discarding smaller eigenchannels with a smooth quality degradation, making a streaming coding strategy feasible. On another hand, perceptual codification of decorrelated channels leads to quality improvement of the reconstructed audio signals, compared to direct codification of the multichannel, for the same total bit rate.

Future work will focus on evaluating the effect of the above coding strategies on spatial perception. Instead of using a channel-to-channel degradation measure averaged over multiple channels, as in this contribution, we will try to reconstruct the sound field at a listener position via room simulation and HRTFs. This approach will allow us to consider additional masking effects related to multiple signals superposition and room reverberation.

REFERENCES

[1] De Vries, D., Boone, M. M., (1999), "Wave field synthesis and analysis using array technology", Proc. 1999 Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York, Oct. 17-20, pp. 15-18.

[2] http://www.bell-labs.com/org/1133/Research/Acoustics/VarechoicChamber

[3] Härmä, A. (2002), "Coding Principles for Virtual Acoustic Openings" AES 22nd Internacional Conference, pp. 159-165.

[4] Kuo, S. and Johnston, J. D. (2001), "A Study of why Cross Channel Prediction in Not Applicable to Perceptual Audio Coding " IEEE Signal Processing Letters Vol. 8, Nº 9 September 2001.

[5] Yang, D., Ai, H., Kyriakakis, K, Jay Kuo, C. C. (2000), "An interchannel redundancy removal approach for high-quality multichannel audio coding", AES preprint 5238, AES 109th Conv., Los Angeles, Sept. 2000.

ITU Recommendation BS.1387 (2002) "Method for objective measurements of perceived audio quality".