Low bit-rate audio coding with hybrid representations

L. Daudet (1,2), P. Guillemain (1), R. Kronland-Martinet (1) and B. Torrésani (2)

(1) Laboratoire de Mécanique et d'Acoustique,
CNRS UPR7051, Marseille, France
(2) Laboratoire d'Analyse, Topologie et Probabilités,
Université de Provence, Marseille, France

Abstract

We present a general audio coder based on a structural decomposition : the signal is expanded into three features : its harmonic part, the transients and the remaining part (referred as the "noise"). The first two of these "layers" can be very efficiently encoded in a well-chosen basis. The noise is by construction modelized as a gaussian (colored) random noise. Furthermore, this decomposition allows a good time-frequency psycoacoustic modeling, as it dircetly provides us with the tonal and non-tonal part of the signal.

1 Introduction :

Because of its extraordinary variety and complexity, representing audio signals in a both efficient and precise way (ie. with the smallest amount of data and no audible distortion) is a challenging task for scientists and audio engineers. Within the last few years, the growth of digital audio industry has urged the developpement of efficient compression algorithms. Similarily as an image can be seen as a mixed ensemble of regular surfaces, edges and textures, sounds feature pseudo-periodic patterns, transients and stochastic-like components. Broadly used commercial algorithms rely on analyzing signals on short temporal segments and in each of them use Fourier-like transforms to split them into frequency bins. The trade-off between time and frequency resolution should not be kept identical between the fast-varying transients and the (pseudo-)stationnary harmonic tones, therefore such codecs have to switch between different operating modes. Although proved efficient for a broad variety of sounds, such a reprensentation is not fully satifying (especially at high compression ratios) : one can just imagine that there's no such tradeoff when two instruments are playing simultaneously, one of them with short notes, the other one with a long sustained note. A more adequate way to describe signals is thus to consider that the three kinds of features (harmonic, transients and stochastic-like) are simultaneously present at anytime. Similar approaches have already been proposed for analysis purposes, like Serra's [1] decomposition into deterministic and stochastic parts, or sinusoidal + noise + transients [2, 3]. The design of this analysis algorithm is critical, because the three kind of features are often difficult to distinguish. As far as data amount is concerned, this expansion of the signal into three parts, each of them having the same amount of data as the original signal, might not seem the optimal approach ; but now as we shall see each of these "layers" can be very efficiently represented in appropriate basis (now different for all of them).

2 Analysis algorithm

The analysis algorithm is represented on fig. 1. The original signal $s(t)$ is first split into time segments (of size 1024 samples for a 44.1 kHz sampling frequency), and then decomposed as follows:

$$s(t) = s_{tonal}(t) + s_{transients}(t) + s_{noise}(t) \quad (1)$$

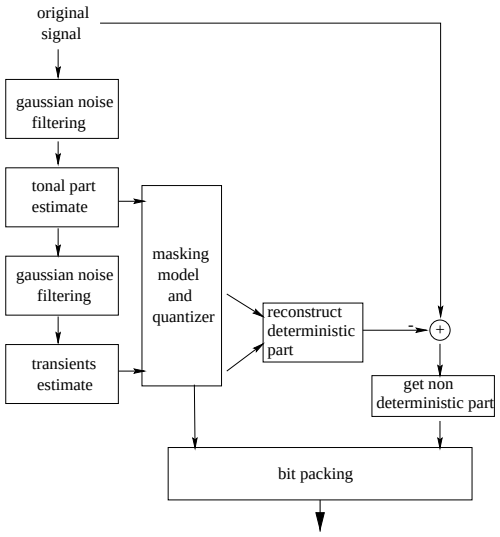

Figure 1: Block diagram of the encoder.

As these three elementary signals are not orthogonal to eachother (in the standard L_2 sense), much care has to be put into this separation algorithm. We here refer as "noise" the remainer when the tonal part and the transients have been removed. The main point here is that, for modelization purposes, it is necessary to make some assumptions about this noise : we'll here assume that it can be modelized as a gaussian random colored noise. Previous experiments [8] have shown that, for a large class of sounds, the non-deterministic part of a signal cannot be distinguished from a gaussian noise with the same power spectrum. Therefore, the main point of our analysis procedure is the estimation of the noise spectral properties.

2.1 Noise model

The baseline for modelizing the noise is here the claim that the ear will generally not be able to distinguish between two realizations of the same random process. Equivalently, to represent this noise it is sufficient to estimate the (possibly time-varying) parameters of this random process. As previously mentionned, if the probability density function (pdf) of the noise is a gaussian, this process is entierly determined by is power spectrum. Furthermore, psychoacoustic studies have shown that random noises are averaged over critical bands (in a way, this can be seen as a definition of the critical bands [5]): it is then only necessary to characterize the variance σ_k of the noise in each critical band:

$$s_{noise}(t) = \sum_{k=1}^{29} \sigma_k . W_k(t) \quad (2)$$

where W_k is a gaussian random generator with unit variance, that only (in the ideal case) contains frequencies in the subband k. The variances σ_k are here assumed constant over the whole segment.

2.2 Denoising algorithm

Starting from this model for the noise, we can design a "denoising" algorithm. We implement our filterbank with a wavelet packets decomposition [6], which frequency resolution closely matches the critical bands, as used for instance by Tewfik *et al.* in [7]. Within each subband k, we perform a hard thresholding of the coefficients [10] :

$$T_{t_k}(x) = \begin{cases} 0 & \text{if } |x| < t_k \\ x & \text{if } |x| \geq t_k \end{cases} \quad (3)$$

The threshold t_k is adaptively chosen such that the pdf of the coefficients below t_k (the series of which is seen as random process) is well approximated by a gaussian (for instance using a chi-square test). The filtered signal is reconstructed from the thresholded coefficients.

2.3 Tonal part estimate

As seen on figure 1, before the tonal part estimate, we perform a preliminary noise filtering, to reduce biases. This preliminary filtering is made in a very conservative way. The tonal part of the sound is then estimated with a lapped cosine transform, also called Malvar-Wilson wavelet transform [9]. The overlap between analyzed segments reduces artifacts caused by discontinuities . The locally harmonic parts of the signals will here give way to large coefficients. These are selected by thresholding as follows : Let c_k be the lapped DCT coefficents and $C = \max_k(|c_k|)$. The threshold is taken as a fraction of C, namely ρC, where ρ is a constant ($\rho < 1$), chosen according to the available bit-rate.

2.4 Transients estimate

After the tonal part removal, another noise filtering is performed, this time with released constraints on our gaussian test (the transients estimate is more likely to be corrupted by noise). Transients are detected through a standard dyadic wavelet transform (DWT) : fast-varying features will cause large DWT coefficients across scales. Here again, a thresholding relative to the maximum coefficient is applied.

2.5 Noise estimate

The remainer of the signal when the tonal part and the transients are subtracted represents the noise. Again, we perform a critical band-based wavelet packets analysis. The variances σ_k in each subband k are simply estimated from the sample variances of the coefficients.

3 Coding technique

For compression puposes, the selected DCT and DWT coefficients are encoded (ie. quantized) with a given number of bits, that depends on the target compression ratio. After this quantization, some efficient entropy coding is performed, like arithmetic coding. One also has to represent the set of indices of selected coefficients, and this is efficiently done using a run-lengh followed by entropy coding. Similarily, the variances σ_k of the noise in each subband are quantized and encoded.

4 Conclusion

We propose an original representation of audio signals, in a both accurate and compact way. The signal is split into its toanl part, transients and noisy part. The key feature of this algorithm is an explicit model of the noise, which allows a good estimate of the other parameters.

In addition to obvious compression purposes, this allows a better understanding of the internal structure of the musical sounds, and allows sound transformations (for instance if we want to increase the duration of a sound without changing the pitch we want transients to remain sharp). Furthermore, as we are directly given with a separation between the tonal and the non-tonal parts, psychoacoustic models can be simplified in this framework.

Preliminary results appeared encouraging, though extensive comparisons with existing coding algorithms still remain to be done. We deeply think that such an approach can be extremely promising for a large class of signals, especially at very low bit-rate (typically less than 32 bits/sec) where the percieved quality

of traditionnaly-encoded soundfiles falls drastically.

References

[1] Serra X., A System for Sound Analysis/Transformation/Synthesis based on a Deterministic plus Stochastic Decomposition, PhD Thesis, Stanford University (1989).

[2] Hamdy K., Ali M. and Tewfik H. , Low bit rate high quality audio coding with combined harmonic and wavelet representations, Proc. Int. Conf. Acoustics, Speech and Signal Processing, Atlanta (1996).

[3] McAulay R. and Quatieri T., Computationally efficient sine-wave synthesis and its application to sinusoidal transform coding, Proc. Int. Conf. Acoustics, Speech and Signal Processing (1998).

[4] Levine S., Audio Representations for Data Compression and Compressed Domain Processing, PhD Thesis, Stanford University (1998).

[5] Zwicker E. and Fastl H., *Psychoacoustics, facts and models*, Springer Verlag Eds (1990).

[6] Vetterli M. and Kovacevic J., Wavelets and subband coding, Prentice Hall (1995).

[7] Sinha D. and Tewfik H., Low bit rate transparent audio compression using adapted wavelets, IEEE Trans. Signal Processing, 41(6),(1993).

[8] Ystad S., Sound Modelling Using a Combination of Physical and Signal Models, PhD thesis, Université d'Aix-Marseille II(1998).

[9] Malvar H. S. and Staelin D. H., The LOT : transform coding without blocking effects. IEEE Trans. Acoust. Speech and Signal Proc. 37(4), (1989).

[10] Gersho A. and Gray R., Vector Quantization and Signal Compression, Kluwer Academic Publishers (1992).