LOW COMPLEXITY BIT ALLOCATION ALGORITHM WITH PSYCHOACOUSTICAL OPTIMISATION

Marcos Perreau-Guimaraes, Madeleine Bonnet, Nicolas Moreau*

UFR Math-Info, Université René Descartes-Paris 5
45 rue des Saints-Pères, 75270 Paris cedex 06, France
perm,bonnet@math-info.univ-paris5.fr
*ENST, 46 rue Barrault, 75013 Paris, France
moreau@sig.enst.fr

10.21437/Eurospeech.1999-349

ABSTRACT

High quality music coders use an auditory masked threshold to account for the characteristics of the human ear. The masked thresholds calculated by these coders do not correspond to the theoretical threshold, solution of a non linear constrained deconvolution problem because of the huge complexity required. We present a binary allocation algorithm solving at the same time the deconvolution problem, while maintaining a tolerable complexity.

1. INTRODUCTION

In order to obtain significant bit compression rates, high quality audio coders take advantage of the masking effect in human auditory system. They preserve binary resources by avoiding the coding of information that cannot be perceived by human ear. The most often, an auditory model is used for spectral shaping of the reconstruction noise. As this modelisation theoretically requires very complex operations, existing coders make use of simplified auditory models. In this paper we present a new algorithm that approaches theory without increasing the complexity.

For each window of signal x (approximatively 20 ms), the auditory model must evaluate a spectral power density $S_X(f)$ estimation of the input signal. Then it must derive a curve, named masked threshold [8], $S_M(f)$ equivalent to a spectral power density. Let $S_Q(f)$ be the spectral power density of the reconstruction noise obtained by the coder. The masked threshold is defined by two conditions. Obviously if $S_Q(f) \leq S_M(f)$ (for each frequency f and for each signal window) then the coded/decoded signal $\hat{x}$ is not perceptually different from the original signal x. Moreover, the curve $S_M(f)$ must be as "high" as possible. Intuitively it must correspond to the greatest inaudible noise, in a sense that will be defined further.

The reconstruction noise is not, in general, an elementary sound as a sine wave or a narrow band noise. In order to take into account this character, the masked threshold is the solution of a complex optimisation problem. Some authors [7] have proposed to solve this optimisation problem (related deconvolution problems can also be seen in [6]) by means of classical numerical analysis tools but with a complexity that is not compatible with the applications (several hours for few seconds of signal). Existing coders do not solve this optimisation problem. They compute a curve that ensures transparency only for elementary signals as tones or narrow band noises.

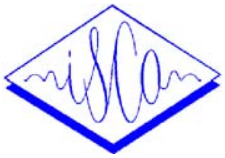
In fact, in a coding system, the problem is different. The bit rate is fixed by the application and the reconstructed signal subjective quality has to be optimised under this constraint. More precisely, let us consider a partition of the frequential axis corresponding to the coder sub-bands k (as for filter bank or transform coders). Let us note $\sigma_X^2(k)$ an estimate of the signal power, $\sigma_M^2(k)$ an estimate of the noise power allowed by the masked threshold and $\sigma_Q^2(k)$ the resulting noise power characterised by relations of the form [1]

$$\sigma_Q^2(k) = \sigma_X^2(k) g(b(k), k) \qquad (1)$$

where $b(k)$ is the number of bits allocated to the k^{th} quantizer, g is characteristic of the quantizer. The coding of music signals generally consists in minimising the maximum of $\sigma_Q^2(k)/\sigma_M^2(k)$ under the bit rate constraint $\sum_k b(k) \leq R$ as one is not sure to ensure the constraint :

$$\sigma_Q^2(k)/\sigma_M^2(k) \leq 1 \quad \forall k \qquad (2)$$

In audio coders, spectral noise shaping is performed by an algorithm allocating binary resources. This algorithm, known as "greedy", is derived from a gradient algorithm which minimises the distance between the masked threshold and the power spectral density of the reconstruction noise. The goal is to share the R available bits for encoding the current signal window in a way such that the maximum value of the noise to mask ratio $NMR(k)$ is as small as possible. The principle is simple : distribute iteratively each available bit



6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

to the sub-band that maximises the noise to mask ratio $NMR(k)$.

We now recall how to compute a masked threshold before to present a new algorithm that performs at once both the psycho acoustical optimisation and the binary allocation optimisation.

2. MASKED THRESHOLD EVALUATION

Computing a masked threshold requires several steps :

1. In order to take into account the ear frequential resolution, the audible frequencies axis can be partitioned non uniformly into 25 critical bands (1 Bark width by definition) [8]. The width of those critical bands increases from 100 Hz for low frequencies to 4500 Hz for high frequencies. Coders sometimes use subdivisions of these critical bands (49 bands of half-Bark width for MPEG number 2 model [5]). We will use here the generic term of basilar sub-bands for any subdivision of the critical bands.
   The basilar spectrum, an estimation of the signal power in each basilar sub-band j, is given by

$$\tilde{S}_X(j) = \sum_{k=b_l(j)}^{b_h(j)} S_X(k) \qquad (3)$$

   where $S_X(k)$ is the power spectral density estimation of the signal, $k = b_l(j)$ and $k = b_h(j)$ are respectively the first and the last coder sub-bands included in the basilar sub-band j.

2. The excitation $E_X(j)$ is obtained by a non linear convolution of the basilar spectrum with the spreading function $f_{etal}(i,j)$ which describes the influence of the basilar sub-band i over the basilar sub-band j

$$E_X(j) = \sum_i f_{etal}(i,j)\tilde{S}_X(i) \qquad (4)$$

3. The masking offset $av(j)$ depends on the tonal character of the signal. It is generally computed via the flatness spectral measure or via a predictive character measure of the signal [1]. The elementary masked threshold (only valid for elementary masked sounds) $E_Y(j)$ is then obtained by :

$$E_Y(j) = av(j)E_X(j) \qquad (5)$$

4. The evaluation of a masked threshold, valid for complex masked sounds, consists in finding the basilar spectrum, $\tilde{S}_M(j) = \tilde{S}_Q(j)$, of the reconstructed noise that minimises the binary resources $J(\tilde{S}_Q)$ and satisfies the non hearing (transparency) constraint

$$E_Q(j) \leq E_Y(j) \qquad (6)$$

where

$$E_Q(j) = \sum_i f_{etal}(i,j)\tilde{S}_Q(i) \qquad (7)$$

The number of bits necessary to encode the current window signal is estimated by

$$J(\tilde{S}_Q) = \sum_j w_j \frac{\tilde{S}_X(j)}{\tilde{S}_Q(j)} \qquad (8)$$

where w_j is the width of the basilar sub-band j.

This constrained optimisation problem can be viewed as a non linear constrained deconvolution problem. The criterion (6) is written in a convoluted domain although the bits are allocated in the non convoluted domain. The problem comes from the spreading of the powers inside the internal ear : the masking effect in a basilar sub-band depends on the signal power in neighbouring basilar sub-bands.

The auditory models used in music coders perform the first three steps. The fourth step is always neglected due to complexity purpose, as in the models described in [5].

3. JOINT OPTIMISATION

The classical method first computes the masked threshold without taking into account the binary resources limitation. Then, available bits are distributed to approach this threshold. The impairment of this method is to compute a simplified threshold, without any auditory optimisation. We notice the similarity between the auditory problem and deconvolution problems arising when a sub-band affects other sub-bands. For example, focusing on a wavelet coder, the work of [4] defines a method that performs deconvolution to take into account the strong filter bank overlapping : the addition of one bit in a sub-band strongly affects the noise power in several neighbouring sub-bands. Then, at each iteration, in order to select the sub-band which receives one bit he proposes to try all the possibilities for allocating one bit in a sub-band. The choice is done in the sense of the best improved quality. Quality is evaluated by a psychoacoustical distance between the original signal and the coded signal that could be obtained with each possibility. This closed loop algorithm needs to compute N auditory models at each iteration. It's far too complex for practical applications.

Now we will show how to substantially reduce this complexity by a more accurate analysis of the psychoacoustical criterion. We propose to improve two points of the previous algorithms : the frequency scale and the selection of the sub-band receiving one bit.

3.1. Computing in the basilar scale

The basilar frequency scale takes into account the lower hearing resolution in high frequencies. The use

of this scale drastically decreases the number of sub-bands without loss of resolution in psychoacoustical sense (MPEG number 2 model uses 49 basilar sub-bands corresponding to 1024 FFT coefficients). However, computing binary allocation in the basilar sub-bands requires to distribute afterward the bits into the coder sub-bands where the quantization stage takes place.

In what follows $\tilde{b}(j)$ is the mean number of bits allocated to the basilar sub-band j. Allowing $\tilde{b}(j)$ bits to a basilar sub-band j leads to give $w_j\tilde{b}(j)$ bits for all the coder sub-bands included in the basilar sub-band j. As for the standard algorithm, we define a table $\tilde{g}(\tilde{b}(j), j)$ giving the SNR in sub-band j.

3.2. Direct sub-band selection

The key of our algorithm is to directly select the basilar sub-band where the allocation of one bit leads to the greatest auditory gain. The psychoacoustical distance between x and $\hat{x}$ is relied to the inaudibility constraint (6). Most of the audible noise occurs in the basilar sub-band which maximises $E_Q(j)/E_Y(j)$. Then at each iteration the goal is to decrease this ratio in the basilar sub-band given by

$$i_0 = \arg\max_j\left(\frac{E_Q(j)}{E_Y(j)}\right) \tag{9}$$

It is equivalent to decrease $E_Q(i_0)$ which is the noise excitation in the basilar sub-band i_0.

The value of the excitation $E_Q(i_0)$ in the sub-band i_0 does not simply depend from the number of bits in i_0. The influence of a sub-band j over the sub-band i_0 is given by $f_{etal}(j, i_0)\tilde{S}_Q(j)$. The sub-band l that has most influence over $E_Q(i_0)$ is not always $j = i_0$. This is why allocating one bit to the basilar sub-band i_0 does not always lead to a significant decrease of $E_Q(i_0)$. In that case the basilar sub-band i_0 may be selected again at next iteration. This leads to a non efficient algorithm. It is the kernel of the deconvolution problem.

The goal is then to select the basilar sub-band l that maximises the bit allocation effect on $E_Q(i_0)$, without trying all possibilities. Excitation definition asserts that the basilar sub-band l that has the greatest influence over $E_Q(i_0)$ is

$$l = \arg\max_j(f_{etal}(j, i_0)\tilde{S}_Q(j)) \tag{10}$$

At the beginning of an iteration, the noise power in basilar sub-band j is given by

$$\tilde{S}_Q(j) = \tilde{S}_X(j)\tilde{g}(\tilde{b}(j), j) \tag{11}$$

The contribution to $E_Q(i_0)$ of the noise power in basilar sub-band j is then given by

$$f_{etal}(j, i_0)\tilde{S}_X(j)\tilde{g}(\tilde{b}(j), j) \tag{12}$$

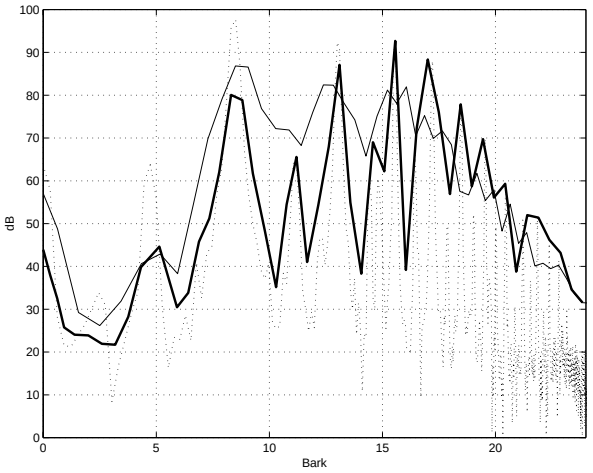


Fig. 1 – Comparison between masked thresholds obtained with MPEG number 2 model (normal line) and our algorithm (thick line) computed on a violin signal window (dashed line). This curves are shown in a Bark frequency scale.

If we add one bit to basilar sub-band j the noise power becomes

$$\tilde{S}_Q^+(j) = \tilde{S}_X(j)\tilde{g}(\tilde{b}(j) + 1, j) \tag{13}$$

and the contribution to $E_Q(i_0)$ is given by

$$f_{etal}(j, i_0)\tilde{S}_X(j)\tilde{g}(\tilde{b}(j) + 1, j) \tag{14}$$

Adding one bit to sub-band j leads to decrease $E_Q(i_0)$ by the difference between (12) and (14). Then the sub-band where to give one bit is the most efficient is given by

$$l = \arg\max_j(f_{etal}(j, i_0)\tilde{S}_X(j)[\tilde{g}(\tilde{b}(j), j) - \tilde{g}(\tilde{b}(j) + 1, j)]) \tag{15}$$

Due to the exponential fall of the spreading function, the selected sub-band l is always in the neighbouring of i_0. In general $l \neq i_0$ when

$$\tilde{S}_X(i_0)\tilde{g}(\tilde{b}(i_0), i_0) << \tilde{S}_X(j)\tilde{g}(\tilde{b}(j), j) \tag{16}$$

with j near i_0.

Figure 1 compares masked thresholds obtained with MPEG number 2 model and our algorithm. Here, what we call masked threshold in our algorithm is the noise power corresponding to bit allocation that just satisfies the inaudibility constraint. The threshold shape in MPEG number 2 model looks like an excitation shape. The second threshold is similar to a basilar spectrum without any spreading of the power nearby tonal spectrum lines. This illustrates the deconvolutional effect of our algorithm.

3.3. Distribution in coder sub-bands

A basilar sub-band includes one or several coder sub-bands. The $w_j\tilde{b}(j)$ bits allocated to a sub-band j are distributed to the w_j coder sub-bands included in the basilar sub-band j. This is performed by a classical allocation method that minimises the SNR in the basilar sub-band j.

4. IMPLEMENTATION AND RESULTS

In order to check its efficiency, our algorithm has been implemented into a TCX like coder [3] described in [2]. It is devoted to a sample frequency $f_e = 32$ kHz and a bit rate 64 kbit/s for high quality speech and music coding for multimedia applications. Our algorithm does not need the signal to mask ratios, outputs of classical auditory models. It requires efficient computation of basilar spectrum $\tilde{S}_X$, excitations E_X, E_Q and masking offset av. The spreading function $f_{etal}(j, i_0)$ is also necessary. We have developed an auditory model, based on MPEG number 2 model, computing those parameters.

4.1. Complexity

Our coder has been implemented with the Matlab software. The increase of complexity involved by our algorithm is about 30% regarding the standard method. This is to be compared with the closed loop algorithm where the increase of complexity about 100 times is not realistic. So, our algorithm does perform the non linear constrained deconvolution with a small increase of the complexity.

4.2. Results

Informal listening tests have been performed over a various corpus of speech and music sequences. As compared to the standard algorithm, the improvement brought by our algorithm is greater for music rather than speech sequences. It can be explained by the less tonal character of speech, as compared to music, leading to a less crucial role for deconvolution. It is due to the effect of the spreading function in the neighbourhood of tonal spectrum lines.

5. CONCLUSION

We have presented an algorithm that performs at once both the psycho acoustical deconvolution problem and the binary allocation optimisation. With this algorithm it is now possible to perform in practical applications the psycho acoustical optimisation required in psycho acoustical theory. The general character of this method allows its implementation on various coder schemes.

6. REFERENCES

[1] N. Jayant and P. Noll. Digital coding of waveforms. Prentice Hall, 1984.

[2] A. Jbira, N. Moreau, and P. Dymarski. Low delay coding of wideband audio (20 Hz - 15 kHz) at 64 kbps. Proc. Int. Conf. Acoust., Speech, Signal Processing, 1998.

[3] R. Lefebvre, R. Salami, C. Laflamme, and J.P. Adoul. High quality coding of wideband of wideband audio signals using transform coded excitation (TCX). Proc. Int. Conf. Acoust., Speech, Signal Processing, pages I–193–196, 1994.

[4] F. Moreau de Saint Martin. Banc de filtres et ondelettes. PhD thesis, Université de Paris 9, 1996.

[5] Norme internationale ISO/CEI 11172. Codage de l'image animée et du son associé pour les supports de stockage numérique jusqu'à environ 1,5 Mbit/s, 1993.

[6] John R. Sacha and Blane L. Johnson. A constrained iterative multiple operator deconvolution technique. JASA, pages 181–185, 1994.

[7] R. Veldhuis. Bit rates in audio source coding. IEEE J. on Selected Areas in Com., 10, no. 1:86–96, 1992.

[8] E. Zwicker and E. Feldtkeller. Psychoacoustique, l'oreille récepteur d'information. Masson, 1981.