SUBBAND AUDIO CODING WITH SYNTHESIS FILTERS MINIMIZING A PERCEPTUAL DISTORTION

K. Gosse[1,3], *F. Moreau de Saint-Martin*[2], *X. Durot*[2], *P. Duhamel*[1] *and J.B. Rault*[2]

[1] Dept Signal, ENST, 46, rue Barrault, 75634 Paris cedex 13, France.
[2] CCETT, rue du Clos Courtel, B.P. 59, 35512 Cesson-Sévigné cedex, France.
[3] K. Gosse is currently with Motorola, Centre de Recherche, 3, avenue du Canada, 91958 Les Ulis-Courtaboeuf Cedex.

ABSTRACT

The design of filter banks for source coding purposes classically relies on the perfect reconstruction (PR) property. However, several recent studies have shown that taking the quantization noise into account in the design could yield noticeable reduction of the mean square reconstruction error. The purpose of this study is to show that perceptual improvement can also be obtained in the particular audio coding context by relaxing the PR constraint. In this context, the mean square error is not relevant any more, and we define a new perceptual distortion criterion, making use of a simplified ear model, the MPE (Mean Perceptual Error). Then, synthesis filters are optimized so as to minimize this MPE. Finally, this MMPE (Minimum MPE) filter bank is included in an audio coding scheme. Compared to the corresponding PR filter bank-based scheme by the means of POM (Perceptual Objective Measure), they show an improved audio quality.

1. INTRODUCTION

Filter banks (FB) are widely used in Audio Coding and they are classically chosen so that they achieve perfect reconstruction (PR) or almost perfect reconstruction in absence of quantization errors. However, recent work showed that this is not the optimal choice since subband signals are quantized. In this case, the reconstruction distortion can be reduced by tuning the synthesis filters, given the analysis filters and the quantization error amount. In [1], authors present matrix Wiener filters, the asymptotic solution for the synthesis filters minimizing the output MSE (Mean Square Error). Other studies also focus on the optimization of finite length synthesis filters, and highlight the improvement brought by such approach [2, 3, 4, 5]. Moreover, Gosse and Duhamel show that a joint optimization of both synthesis filters and subband quantizers brings further improvement in a rate-distortion sense [6, 7]. The resulting coding schemes are called MMSE schemes.

In this paper, we show a successful application of these ideas in the audio coding context: we propose an optimization algorithm of synthesis filters and quantizers, based on a perceptual error criterion, and we describe a new coding scheme benefitting from these optimized non-PR filters.

Indeed, the MSE does not seem to be a relevant criterion when dealing with Audio compression. In order to define a perceptual measure of the distortion introduced by the quantization process, we have to rely on a model of the human hearing system. Classically, the bit-rate allocation is optimized using the so-called masking curves, computed from a signal analysis in the frequency domain (Hertz scale). However, psychoacoustic results show that the Fourier transform is not an accurate model of the transformation performed by the inner ear, and consequently, we rather use a new model, based on the results in [8]. Yet, it is simplified so that it can be included in the perceptual criterion, the MPE (Mean Perceptual Error).

In a second step, we extend the work in [7], and we propose a joint optimization of the bit-rate allocation and the synthesis filters, so as to minimize this MPE. We denote the resulting filter bank as an MMPE FB.

Based on the MMPE filter design, our aim is to show the usefulness of relaxing the PR property in FBs for audio compression. In this purpose, we then propose a coding-decoding structure based on MMPE filter banks. In practice, it appears that the bit-rate allocation has to be adapted to each frame content according to local psychoacoustic requirements. We thus provide a new dynamic bit-rate allocation technique. It takes into account the filtering error occurring because of the non-PR analysis-synthesis scheme, and it is designed to deal with non-ideal filters. Finally, coding results are presented.

2. MODELLING THE HEARING SYSTEM AND MEAN PERCEPTUAL ERROR

The work in [8] intends to improve the rough model of the hearing system used in the MPEG1-Audio standard, relying on masking curves. Here, the time-frequency analysis performed by the ear is modelled by a bank of filters with good time-frequency localization. The numbers of filters should be very high since there are several millions hair cells in the inner ear, and 600 discernable frequencies. However, in practice, it is sufficient to consider fourty-nine 1-bark wide filters, centered at 0.5, 1, $\cdots$, 24.5 barks. These filters are complex FIR, they are depicted in fig.1. Note that each perceptual component can then be considered separately, since we mimic the critical bands with such a decomposition.

Now, denote z^r, the r^{th} component of the original signal filtered by the "ear transform" ($0 \leq r < 50$). Denote also $\hat{z}^r$, the corresponding component of the same signal, but after coding and decoding process. z^r and $\hat{z}^r$ can not be distinguished by the ear if the difference of their loudness does not exceed 1dB. This is referred to as the Just

Noticeable Difference (JND) in the litterature [9]:

$$10\log_{10}\left[\frac{\mathcal{E}|\hat{z}^r|^2}{\mathcal{E}|z^r|^2}\right] < 1dB \quad (1)$$

Under the independence assumption of z^r and of the quantization noise $\hat{z}^r - z^r$, the JND can be rewritten using a first order development of the log function:

$$\left[\frac{\mathcal{E}|\hat{z}^r - z^r|^2}{\mathcal{E}|z^r|^2}\right] < 0.26 \quad (2)$$

This gives us an upper bound for the quantization noise that can be injected in subband r without being audible. This bound is related to the variance of the signal in the same band: $P_r = \mathcal{E}|z^r|^2$. However, eq. (2) does not take into account the threshold in quiet, which represents the absolute sensitivity of the ear, varying from one critical band to another. For very high frequencies in particular, the amount of noise to be injected would be underestimated. Here, we handle this parameter, T_r, as an additional masker in each subband, by setting the upper bound for the noise to:

$$[T_r^\alpha + P_r^\alpha]^{\frac{1}{\alpha}} = \mu_r \quad (3)$$

In the following, α is set to 0.3, and eq. (2) becomes

$$\mu_r^{-1} \cdot \mathcal{E}|\hat{z}^r - z^r|^2 < 0.26 \quad (4)$$

In order to determine whether original and coded-decoded signals can be distinguished or not, the quality criterion expressed by eq. (4) has to be integrated over all subbands. For instance, [10] considers that eq. (4) should be verified for each subband r, and this choice would lead to a perceptual criterion of the form: $J_p^\infty = \max_r \mu_r^{-1} \cdot \mathcal{E}|\hat{z}^r - z^r|^2$ In our context, it seems easier to integrate the difference between original and coded-decoded signals along the frequency axis, by defining the perceptual criterion J_p:

$$J_p = \sum_r \mu_r^{-1} \cdot \mathcal{E}|\hat{z}^r - z^r|^2 \quad (5)$$

We thus assume that original and coded-decoded signals can not be distinguished if J_p is small enough. In consequence, J_p is the perceptual distortion to be minimized over the set of synthesis filters and quantizers, yielding MMPE coding schemes. Note that, by doing this, we simply replace condition J_p^∞ on the $\mathcal{L}^\infty$ norm on the signals by a similar one using the $\mathcal{L}^2$ norm (a multidimensional ellipsoid instead of a parallelipiped defined by the $\mathcal{L}^\infty$ norm).

3. JOINT OPTIMIZATION OF SYNTHESIS TRANSFORM AND QUANTIZERS

The expression of the MPE criterion does not rely on a particular structure for the filter bank (parallel, modulated or iterated). Here, we work with tree-structured analysis banks, and we optimize the coefficients of the equivalent parallel synthesis bank. Note that [7] gives more details for optimizing only part of the parameters at hand, thus reducing the overall complexity. Note also that the same formalism holds for parallel filter banks, of course; Moreover, the approach described in [11] deals with MMSE modulated filter banks, and its use here could enable to keep a low cost modulated structure on the optimized MMPE synthesis side.

3.1. MPE criterion

Let first define the notations used till the end of this paper. After analysis and quantization, quantized samples in subband k, $\tilde{y}_n^k$ are oversampled by a factor J^k, and reconstructed into $\hat{x}_n$ by synthesis filters $F^k(z)$ $(0 \leq k < J)$ to be optimized. $\hat{x}_n$ is then analyzed by the ear filters $E^r(z)$, yielding: $\hat{Z}^r(z) = E^r(z)\sum_{k=0}^{J-1} F^k(z)\tilde{Y}^k(z^{J^k})$

MPE expression in equation (5) also requires the expression of signal Z^r resulting from the analysis of the input signal $X(z)$ by $E^r(z)$. If we assume that perfect reconstruction can be achieved with synthesis filters $F^{\star k}(z)$, we express Z^r as the analysis by ear filters of a delayed version of the input signal $z^{-D}X(z) = \check{X}(z)$, here equivalently written as the filtering of the unquantized subband signals $Y^k(z)$ by $F^{\star k}(z)$.

Moreover, we model the quantization noise as an additive white process, decorrelated from one subband (of the coding filter bank) to another: $B^k(z) = \tilde{Y}^k(z) - Y^k(z)$, with variance $\sigma_{b^k}^2$.

Because of the additive quantization noise process, J_p splits into a filtering term, D_f, and a noise term D_b. It can be compactly expressed in the time-domain as:

$$J_p = \underbrace{\sum_{k=0}^{M-1}\sum_{n_1}\sum_{n_2} f_{n_1}^k\, f_{n_2}^k \sigma_{b^k}^2\, \alpha_{k,n_1,n_2}}_{D_b} + \underbrace{\sum_{k_1=0}^{M-1}\sum_{k_2=0}^{M-1}\sum_{n_1}\sum_{n_2}\left(f_{n_1}^{k_1} - f_{n_1}^{\star,k_1}\right)\left(f_{n_2}^{k_2} - f_{n_2}^{\star,k_2}\right)\beta_{k_1,k_2,n_1,n_2}}_{D_f} \quad (6)$$

In (6), coefficients α_{k,n_1,n_2} and β_{k_1,k_2,n_1,n_2} thus depend on the analysis bank, on correlations between signals and/or noises as well as on the model of the hearing system (coefficients e_n^r). Note that the filtering term D_f cancels with PR filters, whereas D_b tends towards zero at high bit-rates.

3.2. Parameters optimization

The minimization of the perceptual criterion J_p is performed under a bit-rate constraint, over synthesis filters and subband quantizers. As in previous work [7, 12], the optimization procedure is iterative, since optimizing each kind of parameters is easy when the other one is fixed.

If the quantizers, and thus noise variances, are fixed, finding optimal synthesis coefficients amounts in solving a set of linear equations. On the other hand, by relating $\sigma_{b^k}^2$, the noise variance in subband k, to the signal variance in the same coding subband by: $\sigma_{b^k}^2 = c_k\sigma_{y^k}^2 2^{-2R_k}$ (see [7] for more details), the criterion becomes:

$$J_p = D_f + \sum_{k=0}^{M-1}\left[\sum_{n_1}\sum_{n_2} f_{n_1}^k\, f_{n_2}^k \alpha_{k,n_1,n_2} c_k\right] 2^{-2R_k} \quad (7)$$

and it can be minimized over the set of subband bit-rates by classical methods. The analytic computation of optimal positive bit rates for each subband is given in [6] ;

Overall optimization algorithm:
1. Initialize filters to the PR case ;
2. Optimize the bit-rate allocation ;
3. Optimize the synthesis filters ;
4. Repeat steps **2.** and **3.** until convergence.
At each step, either the value of the perceptual distortion is reduced, or the process stops. Thus, the algorithm converges necessarily.

4. AUDIO CODING BASED ON OPTIMIZED SYNTHESIS TRANSFORM

4.1. Audio coding scheme

Let us briefly describe the basic parts of the audio coder used here, and depicted in figure 3. The analysis filter bank is a tree-structured filter bank with delay 464 samples, designed according to [13, 14]. This decomposition is described on figure 2. The quantization and coding schemes consist of both scalar and vector quantizations [15].

When using optimized filters in a real audio scheme, a new bit allocation procedure has to be used. Indeed, the bit allocation suggested by the analytic formulation can not take into account the local variations of signal statistics and of psychoacoustics requirements: it must be replaced by a dynamic one. On the other hand, the usual adaptive allocation techniques, such as the one in [16], assume that the analysis-synthesis system is perfect reconstructing and that the synthesis filters have high stopband attenuation. The MMPE filters are far from being ideal, so that we propose the following bit allocation algorithm, here based on masking curves in the frequency domain. For each frame, the filtering error is calculated, smoothed according to the Bark scale, and substracted from the masking curve. The spectrum of the reconstruction error is then estimated as in [13] and the bit allocation is done by the algorithm presented in [14]. This algorithm can also deal with bit allocation criteria based on the model of the hearing system.

The synthesis filters are optimized ones. More precisely, we observed in our first experiments that the quantization error power estimated with the optimal analytic computation of the bit-rates is quite different from the quantization noise injected by the adaptive allocation. Finally, better results are obtained when optimizing the filters for the noise amount measured in practice with a perfect reconstruction scheme.

Finally, there is a problem linked to the use of the MPE model for synthesis filter optimization. In fact, the perceptual criterion does not take into account very high frequencies, at which the ear is nearly deaf, but for which synthesis filters require some specifications. This yields numerical problems in the filter design. To solve this, we stabilized the distortion criterion by minimizing a weighted sum of MSE and MPE: ($J_p = \mathrm{MPE} + \epsilon \mathrm{MSE}$). Here, ϵ is set to 0.1.

4.2. Coding results

In this study, we compare the following coding schemes:
• the reference PR audio coding system,
• MMSE-J, the original MMSE scheme, with *joint* optimization of filters and quantizers,
• MMPE-J, the scheme relying on filters and quantizers minimizing *jointly* the perceptual criterion,
• MMSE, the synthesis filters optimized for the amount of quantization noise measured in the PR case, minimizing the mean square error,
• MMPE, the synthesis filters optimized for the amount of quantization noise measured in the PR case, minimizing the mean perceptual error.

The proposed coders were tested with the critical sequence known as `Asajinder`. The observed quantization noise levels are measured for a perfect reconstructing system. Then, the synthesis coefficients are optimized with respect to the MPE criterion. We work at bit-rate 128 kbit/s/channel

All these schemes are compared according to an objective measure of the subjective quality, POM (Perceptual Objective Measure [9]), which provides a figure which can be interpreted as the probability that an expert hears the difference between original and encoded-decoded signals.

Both MMSE-J and MMPE-J show poor performances. This is mainly due to the fact that calculating average bit-rate allocations is perceptually not accurate, and observed bit allocations are really different from analytically calculated ones.

With the new dynamic allocation procedure, and the optimization with respect to a given amount of quantization noise, we get the following probabilities of distinguishing original and coded-decoded signals:
• 44% for PR,
• 16% for both MMSE and MMPE schemes.
This clearly shows the improvement brought by optimized solutions over classical PR ones. We completed our study with informal tests that tended to conclude on the superiority of MMPE systems. However, MMSE and MMPE schemes have comparable performances. This means that the stationnarity assumption made for calculating MMPE filters is too strong in a perceptual point of view and that additional effort should concentrate on filters adaptation along time.

5. CONCLUSION

Subband decompositions that are not perfect reconstruction can be of interest for audio coding applications. Here, we optimize synthesis filter banks with respect to a psychoacoustic criterion, and we improve the quality of the coded-decoded signal. This certainly opens the way to various developments in this direction.

REFERENCES

[1] P. P. Vaidyanathan and T. Chen. Statistically optimal synthesis banks for subband coders. In *Proceedings IEEE Asilomar Conference*, November 1994.

[2] B.-S. Chen, C.-W. Lin, and Y.-L. Chen. "Optimal signal reconstruction in noisy filter banks : Multirate Kalman synthesis filtering approach". *IEEE Transactions on Signal Processing*, 43(11):2496–2504, November 1995.

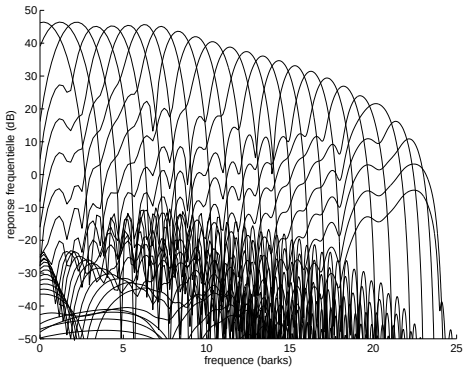

Figure 1. Frequency responses of the filters used for the "ear transform", $E^k(z)$ (1 filter out of 2 is represented)

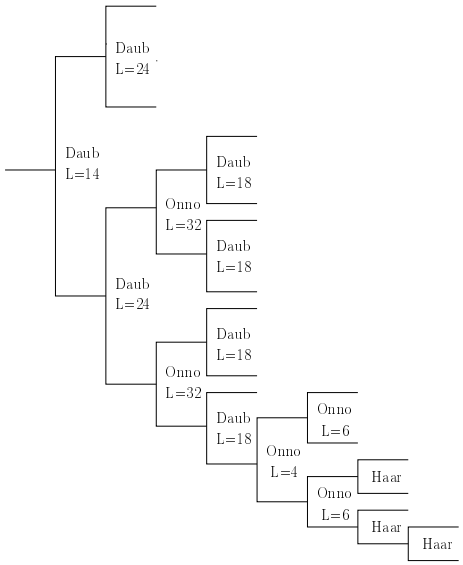

Figure 2. Analysis tree-structured filter bank used in the coding scheme

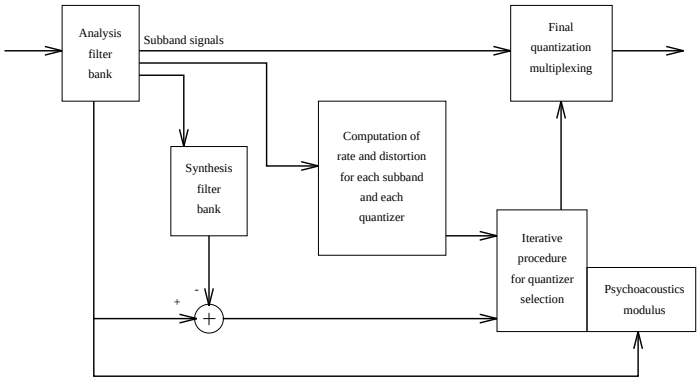

Figure 3. Overall representation the audio coding scheme

[3] J. Kovacevic. Subband coding systems incorporating quantizer models. *IEEE Transactions on Image Processing*, 4(5):543–553, May 1995.

[4] R. A. Haddad and K. Park. "Modeling, analysis and optimum design of quantized M-band filter banks". *IEEE Transactions on Signal Processing*, 43(11):2540–2549, November 1995.

[5] A. N. Delopoulos and S. D. Kollias. "Optimal filter banks for signal reconstruction from noisy subband components". *IEEE Transactions on Signal Processing*, 44(2):212–224, February 1996.

[6] K. Gosse and P. Duhamel. "Perfect reconstruction versus MMSE filter banks in source coding". *to be published in IEEE Transactions on Signal Processing*, April 1997.

[7] K. Gosse, F. Moreau de Saint-Martin, and P. Duhamel. "Filter bank design for minimum distortion in presence of subband quantization". In *Proceedings ICASSP'96*, Atlanta, USA, 1996.

[8] X. Durot and J.-B. Rault. A New Noise Injection Model for Audio Compression Algorithms. In *AES Convention*, Los Angeles, USA, nov 1996.

[9] C. Colomes, M. Lever, Y.F. Dehery, and G. Faucon. A perceptual objective measurement system (POM) for the quality assessment of perceptual codecs. In *AES Convention*, 1994.

[10] K. Brandenburg and T. Sporer. " "NMR" and "Masking flag" : evaluation of quality using perceptual criteria". In *AES Conference on Audio Test and Measurement*, pages 169–179, Portland, USA, 1992.

[11] K. Gosse, T. Karp, P. Duhamel, and A. Mertins. Modulated filter banks with minimum output distortion in presence of subband quantization. In *IEEE Asilomar Conference*, November 1996.

[12] K. Gosse, T. Karp, F. Moreau de Saint-Martin, P. Duhamel, and A. Mertins. "MMSE optimizations of modulated and tree-structured filter banks for efficient tradeoffs between rate, distortion and complexity". *submitted to IEEE Transactions on Signal Processing*, 1996.

[13] P. Philippe, F. Moreau de Saint-Martin, M. Lever, and J. Soumagne. "Optimal wavelet packets for low-delay audio coding". In *Proceedings ICASSP'96*, Atlanta, USA, 1996.

[14] P. Philippe, F. Moreau de Saint-Martin, and M. Lever. "Wavelet packet filter banks for low bit-rate audio coding". *submitted to IEEE Transactions on Speech and Audio Processing*, 1996.

[15] P. Philippe, M. Lever, L. Mainard, J.B. Rault, and J. Soumagne. "A Musicam-like VQ approach yields improved low bit-rate coding of audio signals". In *AES Convention*, San Francisco, USA, 1994.

[16] ISO/IEC CD 11172. "Coding of moving pictures and associated audio for digital storage media at up to 1.5 Mbits/s". Technical report, December 1991.