Wavelet Packet Filterbanks for Low Time Delay Audio Coding

Pierrick Philippe, François Moreau de Saint-Martin, and Michel Lever

Abstract—We study the application of wavelet packet filterbanks to low bit-rate transparent audio coding, taking the audio coders' delay requirements into account, and propose low-delay coders based on wavelet packet filterbanks. We first develop a method of comparison between filterbanks for perceptual audio coding by estimating the necessary bit-rate for a transparent compression. We use this comparison method in order to select the best filters for our audio compression scheme, from a large set of orthogonal and biorthogonal wavelets. Different wavelet filters may be used at different stages of the tree-structured decomposition with a constraint on the overall delay taken into account. The optimization is carried out with a simulated annealing procedure, proposing two wavelet packet filterbanks, exhibiting average and low delays. They are inserted in a complete audio coder that employs vector quantization and psychoacoustic models. The use of the proposed filterbanks leads to the design of a new bit allocation procedure, and considers psychoacoustic models. The use of the proposed filterbanks leads to the design of a new bit allocation procedure, especially for low-delay coding: with average delay, the quality per distortion is kept inaudible. Up until now the most used filterbanks in audio coding are the TDAC filters (time domain alias cancellation) [1] with 128 to 1024 subbands, and the ISO/MPEG filterbank [2] based on the pseudo-quadrature mirror filters (PQMF) [3] technique. For example, in the ISO/MPEG filterbank, the signal is split into 32 equally wide subbands, the lowest one corresponding to the information in the frequency band [0, 750] Hz for a 48 kHz sampling rate. For these frequency mappings, the time-resolution is the same for all frequency subbands, while it is often desirable to have higher time-resolution for high frequencies and higher frequency resolution for lower frequencies, as the human ear does.

The discrete wavelet transform and wavelet packets are efficient tools in source coding and have been widely used in still image and video compression. They consist of using a simple two-band splitting cell and of iterating this cell in a dyadic tree. Some previous papers [4]–[8] have shown the advantages of such filterbanks for audio coding. In most papers the choice of the filters is not thoroughly discussed, and no delay constraint is taken into account, making the transform unsuitable for standard transmission delay or postproduction requirements. This paper, we will only consider the delay due to the analysis and synthesis filters (we do not take into account implementation delays such as buffering ones). This delay has to be less than 10 ms, which means 480 samples at a 48 kHz sampling rate. The questions we address are the choice of the decompositions tree and the design of filters for each dyadic cell, while keeping the delay smaller than in the MPEG-1 system (481 samples). The resulting compression scheme is then to be compared to MPEG1 Layer 2.

Therefore, in Section II, we first review basic facts on perceptual audio coding and filterbanks. In Section III, we propose a way to compare the performances of two filterbanks in terms of perceptual coding efficiency. The optimization strategy presented in Section IV consists of first building a large library of filters with various characteristics, and then choosing the decomposition tree and the filters for each cell of the decomposition tree according to the comparison.
method of Section III. This optimization is carried out with a simulated annealing approach. The resulting filterbanks are then discussed in Section V. Compared to the filterbank in [2], the tree-structured filterbanks have frequency responses quite different from those of ideal subband filters, especially when strong delay constraints have to be taken into account. Therefore, the bit allocation procedure cannot, as in [2], consider filters as ideal filters: their frequency shape has to be known when optimizing the bit allocation. This issue will be considered in Section VI. We then conclude by presenting, in Section VII, audio coders based on wavelet packet filterbanks with this new bit allocation procedure providing nearly transparent compression at 80 Kb/s although no overall optimization of the coder has been carried out. This shows the high potential of wavelet packets for audio coding.

II. FILTERBANKS AND PERCEPTUAL AUDIO CODING

A. Psychoacoustics Properties

Modeling the ear is a difficult task that can be fruitful for audio compression: we know [9] that the ear can be modeled as a 25-channel filterbank covering the [0, 24] kHz spectra, named critical bands. The bandwidth of these channels grows approximately as $\log f$. Mimicking this decomposition on the so-called Bark scale with a digital filterbank could result in a useful representation for audio coding, as mentioned in [6].

We now briefly present the standard psychoacoustic model from [2]. This model intends to determine the kind of distortion that the human ear will not hear. It is based on signal frequency masking [10]. For a given audio frame, the signal spectrum is computed through the fast Fourier transform (FFT). The tonal and nontonal components of the signal are extracted, and for each of them the masking effect is computed. The overall masking curve is calculated by adding, along the critical bands, all the masking effects and taking the absolute curve of the ear into account. The transparency condition is satisfied if the reconstruction error spectrum lies under the frequency masking curve.

Although it is not a well-documented phenomenon, we have to take into account the fact that the masking effect does not act purely in the frequency domain. One could exploit the time masking effect, which means that a quiet sound arriving just after a loud one will not be heard at all, but practically it is not easy to take advantage of this phenomenon for bit-rate compression. From another point of view, and showing the limit of the frequency domain masking model: if the time resolution of the compression is not sufficient, the effect of the quantization noise in the time domain can be heard, for example as the phenomenon known as pre-echo. Therefore, we need synthesis filters with good time localization. The simplest way to achieve this is to limit the impulse response length of the equivalent synthesis filters. From another point of view, it is interesting to note that, in order to avoid pre-echo phenomena, it is more important to require good time resolution for the higher frequency bands than for the lower frequency bands, as performed by wavelet packet filterbanks.

B. Filterbanks and Wavelet Packets

The most popular filterbanks up till now have been the parallel “equal-bandwidth” filterbanks as depicted in Fig. 1 and the iterated dyadic “wavelet packet” filterbanks (Fig. 2). Subband coding consists of splitting the input signal into a set of subband signals by filtering and downsampling so as to extract the information corresponding to the frequency bands without introducing any redundancy. The reconstruction from the subband signals consists of the dual operations: up-sampling and synthesis filtering. With finite impulse response (FIR) filters it is also well known that alias cancellation and perfect reconstruction can be achieved by imposing some relations between the filters [11]. Actually, it is also possible to achieve perfect reconstruction with causal and stable infinite impulse response (IIR) filterbanks [12], but we consider only FIR systems, for which there are many flexible design algorithms available. Therefore, we will only consider FIR systems ensuring alias cancellation and perfect reconstruction.

We use a dyadic wavelet packet transform, because the flexible tree structure makes it possible, for example, to let the equivalent subbands follow the Bark scale. The scheme is then similar to the tree structure chosen in [6].

C. Required Properties in Filterbank Design

1) Perfect Reconstruction: With no quantization noise the system has to be a perfect-reconstruction scheme. An arrangement, where the reconstruction error is kept lower than 96 dB, can be termed as a near-perfect reconstruction one, since the signal is reconstructed with its original accuracy (16 b/sample).

2) Time-Frequency Localization: Frequency selectivity is an intuitive criterion for audio compression filterbank design, since it corresponds to the idea of coding in frequency sub-bands, and of adding the quantization noise independently in different subbands. However, filters’ lengths must be controlled so as to keep the overall delay acceptable. With the required values for the delay, bearing in mind the above remarks on the frequency aspects of the pre-echo phenomenon and as well as observing the fact that in such tree-structured filterbanks the time resolution is better for high-frequency components than for low-frequency components, the filter lengths are always such that no pre-echo problems appear. In addition, the complexity is kept low.

3) Energy Compaction: It is well known that a transform aims at concentrating the energy of the signal in a few subband signals. A classical and simple way of studying the statistical properties of a filterbank is the evaluation of the “coding gain” $G_c$ [13]. This aims to measure the improvement of the objective reconstruction quality, in terms of signal to noise ratio (SNR) for a given bit-rate, when comparing with a simple pulse code modulation (PCM) quantization and a subband quantization. For a two-band perfect reconstructing filterbank $G_c$ is given as [13, 14]

$$G_c = \frac{\sigma_x^2}{\|G_0\| \|G_1\| \sigma_{X,0} \sigma_{X,1}}$$

(1)

where $\|G_0\|^2$ and $\|G_1\|^2$ denote the energies of the synthesis filters, $\sigma_x^2$ the power of the input signal, and $\sigma_{X,0}^2$, $\sigma_{X,1}^2$
Fig. 1. Equal bandwidth filterbank (PQMF solution).

Fig. 2. Example of wavelet packet filterbank.

the power of the subband signals. We use it as in [15] with first-order autoregressive [AR(1)] signal models.

4) Regularity and Flatness: The regularity of a lowpass filter \( G \) is defined by considering a dyadic refinement scheme using this filter for the interpolation. If the refinement scheme converges toward a function, the regularity of the limit function is considered. In between, the refinement scheme implements the impulse response of the equivalent lowpass filters for one, two, etc. iterations, and the smoothness of these impulse responses is connected to the regularity of the limit function.

A necessary condition for regularity is vanishing moments (flatness) at the Nyquist frequency, i.e., that there exists \( K > 0 \) such that \( G(z) \) can be divided by \((1 + z^{-1})^K\). This property is a key point for the design of regular wavelet filters [16].

Here we will consider the regularity of wavelets in the Hölder sense [17] and estimate it using Rioul’s algorithm [17].

III. COMPARISON METHOD BETWEEN FILTERBANKS

One of the major issues in the choice of the filterbank for an application such as audio coding is the criterion stating whether or not a filterbank is better than another one in terms of compression efficiency. This is the point considered here.

A. Principle

We aim at comparing, for different filterbanks, the bit-rates that are necessary to ensure a transparent compression, achieved by the error spectrum lying under the spectral domain masking curve. A standard transform/quantization/coding scheme is used. The quantization is assumed to be uniform, with an adaptive step, varying with time and subband frequency. An entropic coding stage is added for noiseless bit-rate reduction, which takes care of the probabilistic distribution of the quantized values.

The method may be seen as a simple coding scheme based on the filterbank. The signal is transformed, quantized, and encoded, where the compression is performed in a transparent way. The transparency issue arises at the stage where the reconstruction error is generated, namely in the quantization stage. At this point, the choice of quantization steps has to be made under psychoacoustic constraints. For calculating the quantization steps, while minimizing the bit-rate under these constraints, we use a white noise model for the quantization error and a Laplacian model for the subband signal statistics, as justified below. The necessary bit-rate is then approached by the entropy of the quantized values. This estimated bit-rate is the comparison criterion between filterbanks.
B. Statistical Properties of the Subband Signals

The Laplacian assumption we will use for the computation of the quantization steps is permissible due to the following statistical analysis of subband signals. In general, a subband signal can be viewed as a zero-mean signal with a generalized Gaussian probability density function of the following form

\[ P_X(x) = \alpha \exp[-\frac{1}{2}(|x|/\gamma)^\gamma] \] \[ (18) \]

For \( \gamma = 1.0 \) and \( \gamma = 2.0 \) we have the special cases of Laplacian and Gaussian distributions, respectively. With real subband audio signals, \( \gamma \) has been estimated with the Kolmogorov–Smirnov test \[ (18) \] and varies in the range of 0.8 to 1.2. Working with Laplacian distributions allows explicit computations of the bit-rate for a certain quantization stepsize.

C. Computation of the Quantization Steps

The first issue is the determination of the quantization step in each subband. This computation consists of the following three steps:

• estimation of the quantization noise resulting from a quantization step;
• estimation of the resulting bit-rate;
• minimization of the bit-rate under psychoacoustic constraints.

The quantization noise is assumed white in each subband \( k \), with an energy \( N_k \), and the subbands are quantized independently. Thus, the quantization stage is modeled as adding a white noise process with energy \( E_k = \frac{N_k}{12} \). The psychoacoustic constraint is that the power spectral density of the resulting quantization noise has to be under the masking curve \( \psi(\omega) \) computed with model 1 from \[ (2) \] :

\[ \sum_{k=1}^{M} \frac{N_k}{12 D_k} |H_k(\omega)|^2 \leq \psi(\omega) \] \[ (2) \]

where \( M \) is the number of subbands, the \( H_k(\omega) \) represents the synthesis filters of the equivalent parallel filterbank, and \( D_k \) is the sampling rate of subband \( k \).

The estimation of the bit-rate is based on the statistical model of the \( k \)th subband signal. We assume that the subband signal is Laplacian with parameter \( \alpha_k \). For a given quantization stepsize \( \gamma_k \), we are able to explicitly calculate the first-order entropy of the quantized signals \( R(q_k) \). Let \( p_j \) be the probability that the quantization level \( j \gamma_k \) is selected:

\[ p_j = e^{-\alpha_k \gamma_k} \sinh\left(\frac{\alpha_k \gamma_k}{2}\right) \quad \text{for} \quad j \neq 0 \] \[ p_0 = 1 - e^{-\alpha_k \gamma_k/2}. \] \[ (3) \]

The bit-rate needed for subband \( k \) is then estimated as the entropy, which results in

\[ -\sum_j p_j \log_2 p_j = -(1 - e^{-\alpha_k \gamma_k/2}) \log_2 (1 - e^{-\alpha_k \gamma_k/2}) \]
\[ - e^{-\alpha_k \gamma_k/2} \log_2 \sinh\left(\frac{\alpha_k \gamma_k}{2}\right) \]
\[ + \frac{\alpha_k \gamma_k}{2(\log 2) \sinh\left(\frac{\alpha_k \gamma_k}{2}\right)} \]
\[ (5) \]

Thus, the bit-rate resulting from a uniform quantization followed by entropy coding can be approximated as a function

\[ R = \sum_{k=1}^{M} \frac{1}{D_k} \Phi\left(\sqrt{\frac{N_k}{\sigma^2_{X,k}}}\right) \] \[ (6) \]

where \( N_k \) and \( \sigma^2_{X,k} = 4/\alpha^2_k \) denote the local energy of the quantization noise and that of the signal, respectively, and where \( \Phi \) can be explicitly computed as

\[ \Phi(u) = -\left(1 - e^{-u}\right) \log_2 (1 - e^{-u}) \]
\[ - e^{-u} \log_2 \sinh u + \frac{u}{(\log 2) \sinh u}. \] \[ (7) \]

The determination of the quantization steps then consists of a constrained optimization:

\[ \min_{\gamma_k, 1 \leq k \leq M} \sum_{k=1}^{M} \frac{1}{D_k} \Phi\left(\sqrt{\frac{N_k}{\sigma^2_{X,k}}}\right) \quad \text{under constraint} \quad \forall \omega \]
\[ \sum_{k=1}^{M} \frac{N_k}{12 D_k} |H_k(\omega)|^2 \leq \psi(\omega). \] \[ (8) \]

Applying the wavelet transform to an audio signal allows computing the local energies in the subbands \( \sigma^2_{X,k} \). The final bit-rate \( R \) as defined in \[ (6) \] is minimized, while maintaining the noise inaudible by satisfying constraints \[ (2) \]. The constraints are turned into a cost function and a standard conjugate gradient algorithm \[ (19) \] provides the quantization steps.

D. Discussion

This method provides an estimate of the bit-rate needed for the transparent compression of an audio signal, although the bit-rate needed for auxiliary data and side information, as well as some interesting other properties such as delay, time-resolution, complexity, bit-rate reduction through vector quantization techniques, etc., are not taken into account. The bit-rate increase due to auxiliary data and the potential of bit-rate reduction through vector quantization techniques are likely to be the same for all similar filterbank transforms, hence it is justifiable to compare them without taking these elements into account. However, when choosing the optimal number of subbands, we will assume a given bit rate penalty for each subband in a reasonable way.

E. Experimental Validation

Experiments have shown that the formulation \[ (6) \] is valid, due to the estimation of the real entropy of the quantizer’s output symbols. We also verify that when the overall quantization noise is kept under the masking curve, very few degradations could be heard (i.e., near transparent coding). The final tests of the coders (see Section VII) are also an experimental validation of this comparison method. The actual necessary bit-rate for near transparent compression is higher than the entropy measured by the above comparison method, since the bit-rate estimation gives an averaged estimation and does not take into account isolated high bit-rate requests. However, the hierarchy between filterbanks is maintained and hence the main aim of this comparison method is achieved.
Another difficult point is the choice of the signal that is used for these comparisons. Of course, it is necessary to compare filterbanks with respect to the same signals. It is also important to deal with signals for which the psychoacoustic model is known to be reliable, imposing stationary signal segments. We have carried out optimization with signal segments from critical audio sequences known as Asajinder and Pitchpipe. These critical sequences are frequently used in digital audio codecs assessments. During the subjective tests of the resulting coders (see Section VII), different critical sequences have been used, and the results are almost as good with other sequences as with the training signals, showing that this choice is valid.

IV. WAVELET PACKET FILTERBANKS OPTIMIZATION

A. Principle

The problem consists of choosing an appropriate tree structure for the decomposition and of choosing the optimal wavelet filter for each cell of the decomposition tree (for example, tree in Fig. 2).

The optimization consists of minimizing the estimated bit rate provided by the comparison method described in Section III. The number of degrees of freedom for this optimization is huge. We develop new wavelet filter design algorithms for this study. Perhaps the optimal filters cannot be designed through existing techniques, but in any case, a “nearly-optimal” filterbank can be designed through existing algorithms. More precisely, we assume that an efficient solution can be found from a preset library of wavelet filters designed with various existing algorithms and various design specifications. This library will be described in the next subsection.

The optimization is then a discrete optimization issue. Two different strategies will be used for the choice of the decomposition tree and for the choice of the filter in each cell. The optimization is then an iterative process. First a decomposition tree is chosen so as mimic the Bark scale as precisely as possible. With this tree we then optimize the choice of the filters for each cell of the dyadic tree. Then the decomposition tree is optimized for this filters’ choice, and so on iteratively. This two-stage iteration technique leads to an efficient decomposition. We first describe the filter library and then at each stage: wavelet filter choice and decomposition tree structure.

B. Wavelet Filter Library

The first stage of the optimization consists of building a library of wavelet filters, so that the optimization problem becomes a discrete optimization problem. This library should be representative for all filters appropriate for audio coding. At this point, our assumption is that it is sufficient to consider various filters from the literature, that have been designed with respect to the above introduced criteria [frequency selectivity, coding gain with respect to AR(1) models, regularity, and flatness].

Riou’s algorithm [16] for the design of orthogonal wavelets with regularity and maximal selectivity takes as inputs the length of the filters, the transition bandwidth, and the flatness. We included in the filter library the filters designed with the following characteristics: lengths $L = 16, 20$, and $32$; transition of bandwidth $0.05, 0.1$, and $0.2$ (normalized frequency); flatness from zero (Smith and Barnwell filters [20]) to $L/2$ (Daubechies wavelets [21]). This results in a set of 111 filterbanks, which are representative of a large class of filterbanks. The other orthogonal wavelets we used are Caglar filters optimized with respect to coding gain and decorrelation [15], [22]; Onno filters [23] are optimized with respect to coding gain on AR(1) processes. For the design of biorthogonal wavelets, we used filters proposed by Cohen et al. [24], by Ikehara [25], and by Vetterli and Herley [26]. We have also designed filters with the algorithms by LeBihan et al. [27], and Moreau de Saint-Martin and Siohan [28].

C. Wavelet Filter Choice Strategy

The optimization strategy consists of a simulated annealing method [29]. For a given tree structure, each cell of the tree (elementary two-channel filterbank) should be associated with a selected two-channel filterbank from the library. In this huge discrete optimization problem, a simulated annealing strategy is well suited.

From a given configuration, a cell is selected, whose two-band filterbank may be changed. The cost function $R$ (the estimated bit rate) is computed for different changes of the filterbank for the current cell. From these possible values, probabilities of filterbank changes are computed. The change is carried out randomly according to these probabilities. Let us describe the algorithm more formally for a previously obtained decomposition tree.

1) At iteration $k$, the selected two band filterbanks are $F_k^0, \ldots, F_k^N$. The temperature is $T_k$. The cell which may change is $i$.

2) For each possible filterbank $f$ for site $i$, the resulting system delay is computed. If the delay is less than the constraint, the objective function $R$ is computed for the corresponding configuration, which is denoted as $U(f)$:

$$U(f) = R(F_{i-1}^k, \ldots, F_{i+1}^k, f, F_i^k, \ldots, F_N^k).$$

(9)

3) For each admissible $f$, selection probabilities $P(f)$ are computed as

$$P(f) = \frac{\exp \left( -\frac{U(f)}{T_k} \right)}{\sum_{\text{admissible } g} \exp \left( -\frac{U(g)}{T_k} \right)}.$$  

(10)

4) The new filterbank for the cell $i$ is selected randomly according to the probabilities $P$.

5) $T_k$ decreases to $T_{k+1}$ and a new site $i$ is selected.

If $T$ is slowly decreased, under certain conditions, the optimality of the algorithm can be proven [29]. In practice it leads at least to interesting configurations.
D. Tree Structure Choice Strategy

Let us assume now that the wavelet filters to be used in all cells are frozen: for example, we choose a filterbank for cells splitting 12 kHz wide signals, another one for 6 kHz wide and so on. The aim now is to optimize the decomposition tree by using the comparison method.

It is an algorithm with progressive construction of the tree. Our assumption is that a near-best tree with $M$ subbands can be obtained by adding an extra splitting cell to a near-best tree with $M-1$ subbands.

Thus we start from the basic one-cell tree and we look for the best two-cell tree. We can further split either the higher or the lower bands, and the selection is done by comparing the two resulting filterbanks. Thus we select the best three-band filterbank. We want to add a new splitting cell to this filterbank. There are three possibilities, which are compared in terms of estimated bit rate $R$. Interesting decomposition trees for various numbers of subbands are obtained this way.

In order to design an efficient coder, we also need to know the optimal number of subbands. This requires taking into account the auxiliary data due to the framing information, the bit allocation and so on. Two numerical values, namely 6 and 10 bits per frame and per subband are considered to take a reasonable amount of side information into account in the optimization process. This provides a fair comparison between the different filterbanks, and informs us about the optimal number of subbands.

V. Resulting Wavelet Packet Filterbanks

In this section, we analyze the filterbank resulting from the optimization presented above. This analysis applies to the following three points.

- Which is (are) the optimal decomposition tree(s)?
- What is the optimal way to deal with the delay and the lengths of the filters in different cells?
- What are the characteristics of the selected filters?

A. On the Tree Structure

Let us now review the main results of the optimization strategy in terms of the decomposition tree. The method is able to work as a “black box,” so that we can look directly at the optimal performances that can be achieved for a given number of subbands. With the extra estimated bit-rate due to the subband auxiliary data, we obtain the results depicted in Fig. 3. This suggests that the optimal number of subbands is around 16, although values between 10 and 24 yield similar performance.

We now intend to look at the best decomposition for a given number of subbands, e.g., 16. The results are presented in a compact way in Fig. 4. The curves are made in the following way: it is the line through points $(f_1; i)$, where $f_1$ is the upper frequency of the frequency band corresponding to subband $i$. The vertical axis thus denotes the subband index. The decomposition of an equal bandwidth filterbank would be represented as the diagonal of the rectangle. When the curve grows faster in low frequencies, this means that the lower-frequency subbands are decomposed more accurately. Together with the Haar filters, the filters used are Onno’s filters [23] of various lengths.

It is interesting to compare the resulting time-frequency decomposition with the Bark scale and the decomposition of the MPAC encoder [30]. Fig. 5 shows the normalized subband, which is the index of the subband divided by the total number of subbands, according to the frequency. Let us remember that all uniform band filterbanks, such as Musicam [2] or TDAC [1], would be represented as the diagonal of the rectangle. It is clearly seen that the kind of decomposition we obtain is much closer to the Bark scale than any other previous decomposition.

B. On Filters’ Lengths

As mentioned above, the filters’ lengths are a criterion which does not appear in the bit-rate estimated by the comparison method, but which has to be taken into account because of time-resolution (pre-echoes), and complexity. Also, the filter lengths are strongly limited by the delay constraint, so that the complexity is kept low, and the time-resolution...
is sufficient. It is interesting to see how the lengths have been dispatched along the cells. It is to be noticed that an orthogonal wavelet filter of length \( L \) in a cell at depth \( j \) has a contribution to the overall delay of \( 2^j(L - 1) \). Intuitively, one could expect decreasing length through the cell structure, otherwise the contributions to the delay increase very fast along the iterations. It was also thought that high performances in terms of coding gain or frequency selectivity would be more important in the first stages of the decomposition. Both arguments allow us to expect shorter filters along the iterations. In practice the choice of the simulated annealing algorithm more or less confirms our expectation as shown in Table I, where typical lengths of the filters for the different depths and for various delay constraints are presented. The lengths are more or less decreasing, but there are exceptions: For example, the first stage is quite often made of shorter filters than the second.

Quantitatively, it is of interest to compare the evolution of filters’ lengths along iterations with the results in [31], where Galand proposes letting the filter length be divided by two at each stage. This allows keeping the same attenuation by doubling the transition bandwidth, so that all equivalent filters have the same transition bandwidth and attenuation. We do not use any such intuition in our optimization. Table I shows that Galand’s intuition is confirmed as a trend, but is not the optimal choice.

C. On the Characteristics of Selected Filters

The discussion on the relevant criteria for filterbank design for various applications has been the major discussion point in filterbank applications in the last few years. We made contributions to the discussion around the synthesis criteria [32]. We wanted to know the best criterion for the design of efficient filters in audio coding. We worked with standard wavelet packet filterbanks (with the same wavelet filter in each cell), for a 25-band decomposition fitting the Bark scale. We will now compare the results in [32] with the filters that have been selected most often by the current optimization strategy.

1) Regularity: Rioul [33] has shown that the regularity of the wavelet filters is an important property for image compression. As the transforms for audio compression also use iterated filterbanks, we expected in [32] an influence of the regularity on the bit-rate. We obtained no correlation between regularity and bit-rate, as shown in Fig. 6. From the present point of view, with the use of different filters in different cells of the decomposition tree, the concept of regularity can not hold the same meaning as in [32] and [33]: It cannot be said that the smoothness of the impulse response of the equivalent low-pass filter is connected with the mathematical regularity of the wavelet filter. However it is interesting to note that the impulse response of the equivalent lowpass filter from the filterbank that will be presented in Fig. 12 is very smooth, as depicted in Fig. 7.

2) Selectivity and AR(1) Coding Gain: In [32] we obtained a better correlation between AR(1) coding gain and estimated bit-rate (cf. Fig. 9) than between frequency selectivity and estimated bit-rate (cf. Fig. 8). We note in Fig. 8 that even if the side lobe level is low, the bit-rate is not necessarily low as well. No correlation can therefore be noticed between the selectivity and the bit-rate. On the opposite we can see that Fig. 9 shows that the AR(1) coding gain seems to be well related to the bit-rate: as the coding gain increases the bit-rate decreases.

Among the filters that are selected through the optimization, all have good coding gain characteristics. To probe further,
Fig. 7. Equivalent lowpass impulse response for the wavelet packet filterbank depicted in Fig. 12.

Fig. 8. Estimated bit-rate versus frequency selectivity.

Fig. 9. Estimated bit-rate versus coding gain.

Another approach has also been explored, considering the same optimization, if the filter library is restricted to wavelet filters that are optimal in an AR(1) coding gain sense. Therefore we let the optimization algorithm run with a smaller wavelet filter library, made of Onno’s filters [optimal in an AR(1) coding gain sense]. The results are shown in Fig. 10. It shows that the restriction of the library to AR(1) coding gain optimal filters does not reduce the efficiency of the system. This might also be true with other subfamilies, for instance the Smith–Barnwell family, with various transition bandwidths, however this would make the family much larger than Onno’s family, and therefore the optimization of the filterbank longer.

3) Orthogonality and Phase Linearity: In the case of dyadic wavelets, it is well known that orthogonality and phase linearity cannot be maintained simultaneously for filter lengths larger than two. The orthogonality is a traditional property of the transforms used in signal processing, but for audio compression it is not necessary. For example, in image compression, when the scheme aims at maximizing the SNR the energy preservation property is very useful; in audio compression, SNR and exact energy preservation between time-domain and subband-domain are not really relevant. Looking for biorthogonal linear-phase wavelets, which do not ensure energy preservation, therefore seems to be promising. The results of our method of comparison with biorthogonal wavelets are similar to those with orthogonal wavelets. Actually it might also be possible to simultaneously relax both properties. For example, the wavelet filters proposed in [34] and [35] have neither orthogonality nor phase-linearity, but provide interesting results in terms of delay-versus-selectivity performance. Their application to audio coding could therefore be interesting in terms of compression efficiency and behavior with respect to pre-echoes.

D. Optimization Efficiency

Having carried out the simulated annealing algorithm, we want now to check whether this global optimization procedure provides significant improvement compared to filterbanks designed by hand. A natural way to design a wavelet packet decomposition for audio coding would have been to consider a tree fitting the Bark scale as well as possible, with 25 subbands,
and to place Daubechies’ filters according to the proposal by Galand. With the latter filterbank, the estimated bit-rate is 20% higher than the optimal solution for similar delay.

VI. BIT ALLOCATION PROCEDURE

Usual bit allocation procedures such as the one recommended in [2] are based on the assumption that the synthesis filters are ideal. This assumption is fair in the case of filterbanks with high frequency selectivity such as those in [2]. However, when using wavelet packets, the assumption no longer holds. The bit allocation must take into account the shape of the synthesis filters. Derived from the comparison method (see Section III), the bit allocation procedure consists of choosing the quantizer in each subband following an algorithm, that is in a way similar to [36].

The bit allocation issue can be stated as follows. Assume that a set of quantizers are available for each subband (scalar quantizers with 3, 5 . . . , levels, vector quantizers . . . ). The issue consists of the optimal choice of quantizers for each subband, according to

1) **fixed bit-rate constraint**, where the total bit-rate budget is fixed, and the allocation aims at maximizing the quality of the reconstructed signal;
2) **variable bit-rate**, where the total bit-rate has to be minimized under a quality constraint.

The bit allocation is based on the following observation: for a given choice of quantizers (one for each subband), the resulting spectrum of the reconstruction error can be estimated, under a white noise assumption, as

\[ s(\omega) = \sum_{k=1}^{M} N_k |H_k(\omega)|^2 \]  

where \( M \) denotes the number of subbands, \( N_k \) is the power spectral density of the quantization noise of the \( k \)th subband, and the \( H_k \) are the subbands synthesis filters frequency responses.

From this estimate, a perceptual distance metric is easily computed from the masking curve, e.g., as \( \max_{\omega} (s(\omega) - \psi(\omega)) \).

For each frame, the “variable rate” algorithm then consists of the following stages.

1) For each subband and for each possible quantizer, calculate the corresponding rate and distortion.
2) Sort all quantizers for each subband according to their rate-distortion properties.
3) Select an initial combination of quantizers: for each subband, consider that only this subband is quantized, and select the quantizer leaving the reconstruction error spectrum under the constraint at minimal cost.
4) Iteratively, apply the following procedure, quantizing subbands using all legitimate quantizers until the required quality is reached.
   a) Calculate the current perceptual distance.
   b) For each subband, switch to a coarser quantizer, estimate the error spectrum, calculate the new perceptual distance and rate. This means that for each subband a new rate and a new perceptual distance is calculated.
   c) Identify the subband providing the best ratio between perceptual distance decrease and bit rate increase.
   d) The new current combination is made by updating the quantizer of the selected subband.

This algorithm converges efficiently to a good combination. In practice, the given rate-distortion constraint is very well taken into account.

In case of a “fixed rate” allocation, the following stage is added to the algorithm:

5) Iteratively apply the following procedure until the required bit rate is reached:
   a) For each subband, switch to a coarser quantizer, estimate the error spectrum, calculate the new perceptual distance and rate;
   b) Look for the subband providing the best ratio between bit rate decrease and perceptual distance increase;
   c) the new current combination is made by updating the quantizer of the selected subband.

VII. EFFICIENT AUDIO CODERS

In this section, we present two simple and efficient encoding schemes based on wavelet packet decomposition. The encoding schemes are based on filterbanks designed by the optimization procedure proposed in this paper. One of them is optimized with a standard delay constraint (500 samples), and the other one is devoted to low-delay encoding and is optimized with a low-delay constraint (250 samples). These schemes also implement the bit allocation procedure presented in Section VI at a fixed bit-rate. We first present the quantization and lossless encoding stages, and then an overall description of the encoding schemes. The results of the subjective listening tests conclude this section.

A. Quantization and Lossless Encoding Stages

The set of available quantizers for each subband is made of scalar and vector quantizers. The bit allocation procedure described in Section VI leads to the optimal choice of quantizer for each subband. The quantization is carried out in a similar...
Fig. 12. Filterbank we use for conventional delay audio coding: the filtering delay is 464 samples.

way as in [2]. Blocks of subband 12 samples are first rescaled using a scale factor that has to be transmitted. This scale factor is computed and coded with the same method as described in the ISO/MPEG encoding scheme.

The resulting rescaled blocks are quantized either sample by sample, or as vectors of four samples (vector quantization of dimension 4) [37]. For each subband and each frame two quantization techniques and two lossless encoding techniques are available.

We use uniform scalar quantizers with three to 65 535 levels, designed as in [2] and vector quantizers designed with the Linde–Buzo–Gray algorithm [38]. The learning sequence was a set of samples following a Laplacian probability density function, since we have noticed in Section III that this is the best approximation of the subband signal statistic.

Vector quantization was added for reducing the cost in terms of bit-rate for improving the resolution for wide subbands. For example when a subband contains 256 samples, in scalar quantization the next possible resolution grows up the bit-rate of 256 bits (the next quantization step adds 1 bit/sample). While in a vector quantization scheme of dimension four, the next resolution adds only 0.25 bit/sample and for this example only 64 bits are added. Quantizers are set along the distortion axis according to the psychoacoustic model’s requirements [37], therefore ensuring a consistent set of quantizers.

Two lossless encoding strategies are used. A standard Huffman encoding is used for quantizers with few waveforms, between three and 15 levels for scalar quantizers. For vector quantizers the entropy-constrained vector quantization optimization combined with Huffman encoding improves the rate-distortion performances for small codebooks only. We then also use a generalization of the grouping described in [2]. This consists of grouping together successive quantization indices so as to build a so-called “meta-index.” For instance, five successive indices \( i_k \) from a three-level quantizer are encoded as \( \sum_{k=0}^{4} i_k 3^k \), reducing the cost to 1.6 bit/sample instead of 2. We have to choose among this available large set of quantizer and coding techniques the best for each subband signals and for each subband. With this set of quantizers and encoding techniques we obtain a large set of coding schemes well suited for the representation of subband audio signals.

B. Overall Description of the Encoding Schemes

The overall coder is depicted in Fig. 11. The final optimized choice of filterbanks is presented in Fig. 12 (standard delay case) and Fig. 13 (low delay case). The psychoacoustic model, which determines the level of admissible quantization noise in each subband, is derived from model 1 proposed in [2]. The bit allocation is carried out as described in Section VI.

We compare the performances of five coders by means of subjective evaluation: MPEG1-Layer II at 96 Kb/s \( (\text{T96}) \) and at 80 Kb/s \( (\text{T80}) \); the 464-sample delay scheme at 96 Kb/s \( (\text{D96}) \) and at 80 Kb/s \( (\text{D80}) \); the 218-sample delay scheme at 96 Kb/s \( (\text{D96}) \).

C. Results

The subjective tests have been performed on loudspeakers under the \( A-B-C \) rule rising critical audio sequences. The \( A-B-C \) methodology [39], known as \textit{triple-stimulus double-blind test with hidden reference}, is recommended by ITU-R for the subjective assessment of small impairments. It consists of presenting first the uncoded reference \( A \) and then two items \( B \) and \( C \), which are both compared to \( A \). Either \( B \) or \( C \) is the
encoded–decoded item, the other being the uncoded reference again (hidden because it is not known whether it is \( B \) or \( C \)). The listener has to choose \( B \) or \( C \) as the reference (giving to it the grade 5.0) and to assign a score to the other one. This score goes from 1.0 (very annoying distortion) to 5.0 (inaudible distortion). The mean opinion score (MOS) of both original and encoded signals are then compared and the quality of the compression scheme is given by the MOS between the reference and the encoded-decoded signals.

We used seven standard critical audio sequences for the following subjective tests: asajinder, pitchpipe, castanets, harpsichord, Tracy Chapman, Suzanne Vega, and Ornette Coleman. Tests have been carried out with ten expert listeners and their results are shown in Table II. The average \( \Delta \)MOS values are as follows: 0.8 for MPEG-1 layer 2 at 96 Kb/s; 0.9 for the 464-sample wavelet packet filterbank at 96 Kb/s; 1.4 for the 218-sample wavelet packet filterbank at 96 Kb/s; 1.5 for MPEG-1 layer 2 at 80 Kb/s; 1.8 for the 464-sample wavelet packet filterbank at 80 Kb/s. The Student test (see Table III) shows that the first two and last three are statistically equivalent. This means that if we compare the MPEG-1 layer 2 scheme with the 464-sample delay wavelet packet filterbank, the results are statistically equivalent for both coders at 80 and 96 Kb/s, although the MPEG-1 layer 2 scheme is used.

![Fig. 13. Filterbank we use for low delay audio coding: the filtering delay is 218 samples.](image)

**TABLE II**

<table>
<thead>
<tr>
<th>Item</th>
<th>M96 Mean</th>
<th>Std Dev.</th>
<th>W96 Mean</th>
<th>Std Dev.</th>
<th>D96 Mean</th>
<th>Std Dev.</th>
<th>M80 Mean</th>
<th>Std Dev.</th>
<th>W80 Mean</th>
<th>Std Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>1.470</td>
<td>0.953</td>
<td>2.100</td>
<td>0.924</td>
<td>2.890</td>
<td>0.858</td>
<td>2.820</td>
<td>0.559</td>
<td>3.050</td>
<td>0.762</td>
</tr>
<tr>
<td>C</td>
<td>1.670</td>
<td>0.585</td>
<td>2.020</td>
<td>0.938</td>
<td>2.190</td>
<td>1.026</td>
<td>2.180</td>
<td>1.110</td>
<td>3.070</td>
<td>0.776</td>
</tr>
<tr>
<td>A</td>
<td>1.810</td>
<td>1.003</td>
<td>0.780</td>
<td>1.082</td>
<td>1.970</td>
<td>1.550</td>
<td>2.670</td>
<td>0.506</td>
<td>1.670</td>
<td>1.307</td>
</tr>
<tr>
<td>Y</td>
<td>0.040</td>
<td>0.724</td>
<td>0.360</td>
<td>0.450</td>
<td>1.260</td>
<td>1.138</td>
<td>0.820</td>
<td>0.506</td>
<td>1.370</td>
<td>0.973</td>
</tr>
<tr>
<td>Z</td>
<td>0.320</td>
<td>1.319</td>
<td>0.740</td>
<td>0.809</td>
<td>0.090</td>
<td>0.606</td>
<td>0.590</td>
<td>0.771</td>
<td>1.650</td>
<td>1.094</td>
</tr>
<tr>
<td>T. Chapman</td>
<td>0.280</td>
<td>0.429</td>
<td>0.580</td>
<td>0.391</td>
<td>0.700</td>
<td>0.741</td>
<td>0.530</td>
<td>0.827</td>
<td>0.670</td>
<td>0.648</td>
</tr>
<tr>
<td>S. Vega</td>
<td>-0.240</td>
<td>0.898</td>
<td>-0.160</td>
<td>0.645</td>
<td>0.880</td>
<td>1.316</td>
<td>0.770</td>
<td>1.253</td>
<td>1.190</td>
<td>1.266</td>
</tr>
</tbody>
</table>
TABLE III
RESULTS OF STUDENT’S TEST FOR FIVE CODERS

<table>
<thead>
<tr>
<th>CODER</th>
<th>M96</th>
<th>W96</th>
<th>D96</th>
<th>M80</th>
<th>W80</th>
</tr>
</thead>
<tbody>
<tr>
<td>%</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Mean ΔMOS</td>
<td>0.8</td>
<td>0.9</td>
<td>1.4</td>
<td>1.5</td>
<td>1.8</td>
</tr>
<tr>
<td>Rank</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

in stereo-dependent mode, where the bit budget is shared between the right and left signals according to psychoacoustic requirements, while our codec uses a stereo-independent bit allocation. If we consider the 218-sample filterbank, its performances at 96 Kb/s is statistically equivalent to those of both other coders at 80 Kb/s. It is the price to pay for lower delay.

The results of these subjective tests show that efficient coders can be designed based on wavelet packet filterbanks, although the full optimization of such a coder is beyond the scope of this paper. In particular, an optimization of the psychoacoustic modulus considering the new bit allocation strategy would provide significant improvements.

VIII. CONCLUSION

In this paper, we have presented a new reliable method for comparing different filterbanks for perceptually optimized audio coding. This allows us to propose an original approach to the construction of the wavelet packet filterbanks, taking into account a delay constraint. The resulting filterbanks show that wavelet packets filterbanks are promising tools for low-delay audio compression.

Future work will include an overall optimization of the coder. In particular, the noise injection model is suboptimal when using the new bit-allocation procedure. It should be adapted, and a more promising way consists of using an improved ear model instead of the usual masking curve; The bit allocation procedure is able to follow rules given by an ear adapted, and a more promising way consists of using a

ACKNOWLEDGMENT

The authors would like to thank L. Mainard, P. Onno, J. B. Rault, and P. Siohan, C.C.E.T.T., and J. P. d’Alès de Corbet, University of Paris-Dauphine, for their valuable suggestions and for the fruitful discussions. They would also like to thank the anonymous reviewers, whose comments and suggestions helped them improve the manuscript.

REFERENCES


Pierrick Philippe was born in Vitré, France, in 1969. He received the Ph.D. degree from the University of Paris, Orsay, France, in November 1995. From 1992 to 1995, he was involved in the development of digital audio bit-rate reduction algorithms at the Centre Commun d’Etudes de Télédiffusion et Télécommunications (CCETT), a research center of the France Telecom Group, Cesson-Sévigné, France. From 1995 to 1997, he was a Research and Development Engineer at Innova Son, Vannes, France. Since 1997, he has been with the CCETT, where his current interests involve digital audio coding and multimedia indexing. He also works on the development of ISO/MPEG-4 and ISO/MPEG-7 normalization processes. Dr. Philippe is a member of the AES.


From 1993 to 1996, he was a Member of Technical Staff at the Centre Commun d’Etudes de Télédiffusion et Télécommunications (CCETT), a research center of the France Telecom Group, Cesson-Sévigné, France. His work concerned filterbanks and wavelets, with application to image compression, region-based image compression, audio compression, and transmission systems. In 1996, he joined the Ministère de l’Economie, des Finances et de l’Industrie, Paris, France, where he was involved in R&D programs and technical regulation in the area of audiovisual communication, multimedia, and consumer electronics. He is now with TDF-DO, Les Lilas, France.

Michel Lever was born in October 1959. He received the Ph.D. degree in 1987 from Rennes University, France, with a dissertation in speech coding. He carried out studies on speech processing for Matra Communication from 1986 to 1990, particularly low bit-rate applied to digital mobile telephone. He joined France Telecom in 1990, entering the the Centre Commun d’Etudes de Télédiffusion et Télécommunications (CCETT), a research center of the France Telecom Group, Cesson-Sévigné, France, where he has been involved in developments and studies related to the European standardization of a digital audio broadcasting system. He was in charge of the Audio Laboratory from 1995 to 1997. He was also a member of standardization groups within the ITU-R and the EBU. At the end of 1997, he joined the National Telecommunication Research Center (CNET, France Telecom), Lannion, France, where he is in charge of a department carrying out research and development for telephone services.