Perceptual Coding of Speech Using a Fast Wavelet Packet Transform Algorithm

Benito Carnero and Andrzej Drygajlo
Signal Processing Laboratory
Swiss Federal Institute of Technology at Lausanne
CH-1015 Lausanne, SWITZERLAND
e-mail: carnero@lts.de.epfl.ch

ABSTRACT

This paper presents a new speech coding algorithm based on a fast wavelet packet transform algorithm and psychoacoustic modeling. The employed FFT-like overlapped block orthogonal transform allows us to approximate the auditory critical band decomposition in an efficient manner, which is a major advantage over previous approaches. Owing to such a decomposition of the original signal, we make use of the human ear masking properties to decrease the mean bit rate of the encoder.

1 Introduction

Wide-band speech compression is nowadays an active research area. The higher quality of wide-band speech is desirable for the extended communication tasks, e.g. audio-conference, loudspeaker telephony, multimedia, etc., and the promises of perceptual coding are significant [1].

Present audio compression systems used for this purpose are based on transform/sub-band coding using perceptual criteria which tend to concentrate the quantization noise energy in frequency regions, where it would be masked by perceptually preponderant signal components. Their major drawback is the large computational effort associated with sub-band decomposition (frequently uniform) and psychoacoustic modeling employing an additional FFT analyzer.

In this paper, we present an integrated approach to the design of the wide-band coder for speech signals sampled at 16 kHz, by incorporating the fast orthogonal wavelet packet transform algorithm [2] [3] and the multiresolution requirements of the psychoacoustic model into the design of the nonuniform decomposition filter bank.

2 Description of the algorithm

2.1 Orthogonal wavelet packet transform

The extension of the wavelet transform to wavelet packet decompositions allows flexible time-frequency transformations in signal analysis and coding [4]. Such analysis tools are traditionally implemented through the use of tree-structured filter banks. However, efficient algorithms, with computational loads close to FFT algorithms, have been proposed in [2] [3]. These *overlapped block orthogonal transforms* can be employed at different subsampling factors and provide simultaneously, in one block operation, all possible multiresolution time-frequency coefficients which could be also successively computed by equivalent tree-structured approaches.

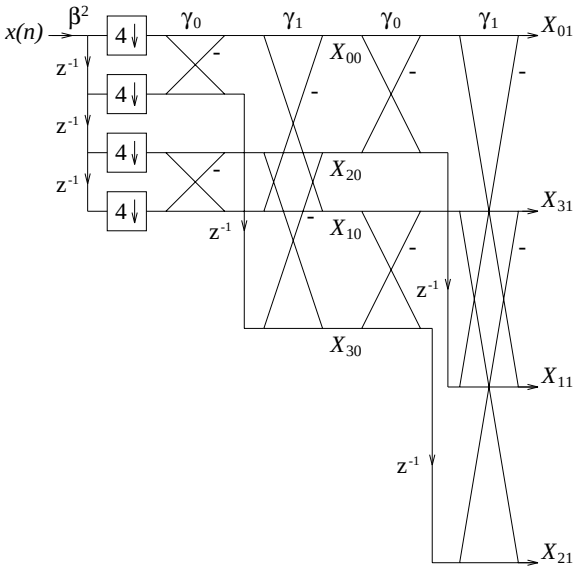

Figure 1: *Four-coefficient transform.*

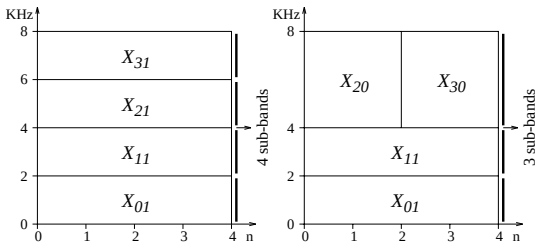

Figure 2: *Four-coefficient Time-frequency grid.*

For instance, a transform with $N = 4$ coefficients is represented in Figure 1. γ_i are the lattice parameters

defining the properties of the implemented filters (regularity, frequency selectivity, delay, etc.) and β is a normalization factor related to the γ_i. The number of stages of the transform is $\log_2 N = 2$. To each block of N input samples, N frame coefficients X_{ij} are simultaneously provided at each stage j with different time-frequency resolutions. Index i denotes the output coefficient number from 0 to $N-1$. Two examples of coefficient choices, as well as their corresponding time-frequency grids and number of sub-bands, are depicted in Figure. 2.

2.2 Critical band analysis

In order to approximate the 21-band Bark or critical band mapping performed by the human ear in the 0-8 kHz bandwidth [5], an overlapped block orthogonal transform has been developed with $N = 64$ frame coefficients (4 ms). The choice of the prototype filter for the transform, as well as its length, influences the separation of the sub-band signals. The Daubechies filters, due to their regularity property, are the ones which achieve the best separation when the number of frame coefficients N increases [6]. In this work, the employed prototype filters are of length 10. The chosen wavelet packet staircase approximation to the Bark scale bandwidths is represented in Figure 3. The approximation of critical band centers is plotted in Figure 4. In both figures, the psychoacoustically measured bandwidths and edges are also plotted.

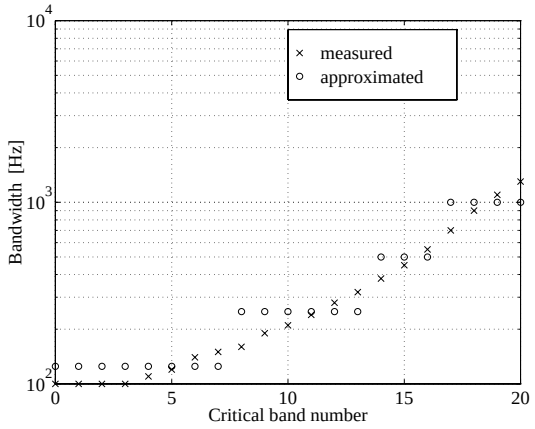

Figure 3: *Approximation of the critical bandwidths.*

The resulting time-frequency grid of this critical band decomposition is drawn in Figure 5, where it appears that higher frequency sub-bands possess several "temporal" coefficients in one frame.

2.3 Computation of masking thresholds

To obtain the staircase approximation of the Bark energy spectrum $\mathbf{A} = [A(0) \ldots A(20)]$, the l "temporal" coefficients in sub-band k are grouped according to

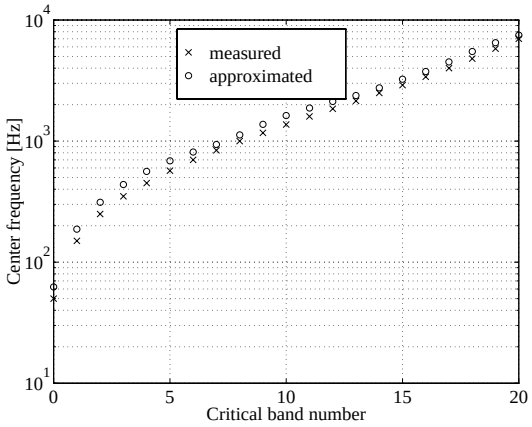

Figure 4: *Approximation of the critical band centers.*

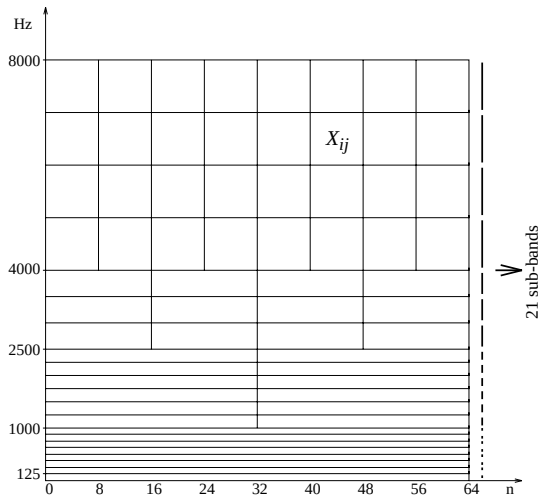

Figure 5: *Time-frequency grid of the transform.*

Table 1. The energy in each sub-band is calculated as

$$A(k) = \sum_i (X_{ij})^2 \quad , \quad k \in [0, 20] \, . \tag{1}$$

The resulting spectrum $\mathbf{A}$ must be convolved with the *spreading function* $\mathbf{B} = [B(-20) \ldots B(0) \ldots B(20)]$ given by

$$B(n) = 10^{B'(n)/10} \quad , \quad n \in [-20, 20] \tag{2}$$

where $B'(n)$ is the function

$$\begin{aligned} B'(n) \quad = \quad & 15.81 + 7.5(n + 0.474) - \\ & -17.5\sqrt{1 + (n + 0.474)^2} \ \text{[dB]}, \end{aligned} \tag{3}$$

that was defined in [7] and accounts for masking among critical bands. The resulting spread Bark spectrum $\mathbf{C}$ will be

$$C(k) = \sum_{j=0}^{20} A(j) \cdot B(k - j), \quad k \in [0, 20] \, . \tag{4}$$

k	l	i	j
[0,7]	1	[0,7]	5
8	2	8,9	4
9	2	10,11	4
10	2	12,13	4
11	2	14,15	4
12	2	16,17	4
13	2	18,19	4
14	4	20,21,22,23	3
15	4	24,25,26,27	3
16	4	28,29,30,31	3
17	8	32,33,34,35,36,37,38,39	2
18	8	40,41,42,43,44,45,46,47	2
19	8	48,49,50,51,52,53,54,55	2
20	8	56,57,58,59,60,61,62,63	2

Table 1: *Bark coefficient mapping with the overlapped block orthogonal transform.*

The *masking threshold* $M(k)$ in each critical sub-band is computed by scaling the spread spectrum by a *relative masking threshold* $O(k)$. This value depends on the tone-like or noise-like nature of the input signal in the processed block, which can be estimated using a spectral flatness measure (SFM) [8]. Several approaches were proposed in the literature to bypass the SFM calculation overhead [9] [6] [10]. The most interesting solution consists in using a composite relative masking threshold, based on the hypothesis that the signal is more tone-like in lower critical bands and noise-like in higher ones. These considerations lead to

$$O(k) = 10^{O'(k)/10}, \quad k \in [0, 20], \tag{5}$$

with

$$O'(k) = \begin{cases} 14.5 + k & , \quad k \in [0, 13] \\ 40.5 - k & , \quad k \in [14, 20] \end{cases}, \quad [\text{dB}] \ . \tag{6}$$

In order to account for the gain in each critical sub-band, $M(k)$ should still be normalized by $1/(DC_{gain}(j))$ of the corresponding transform stage j from which $A(k)$ was computed. Finally, the normalized masking threshold $\mathbf{M} = [M(0) \ldots M(20)]$ will be the staircase curve expressed by

$$M(k) = \frac{C(k)}{O(k) \cdot DC_{gain}(j)} \ , \quad k \in [0, 20] \ . \tag{7}$$

$M(k)$ represents the maximum inaudible noise energy which can be introduced in band k through the quantization process. An example of this procedure is presented in Figure 6, where $A(k)$ and $M(k)$ are successively computed. It clearly appears from these curves that the Bark energy in sub-bands 3, 4 and 6 lies below the masking threshold for the chosen example. Hence, the coefficients belonging to these bands are masked.

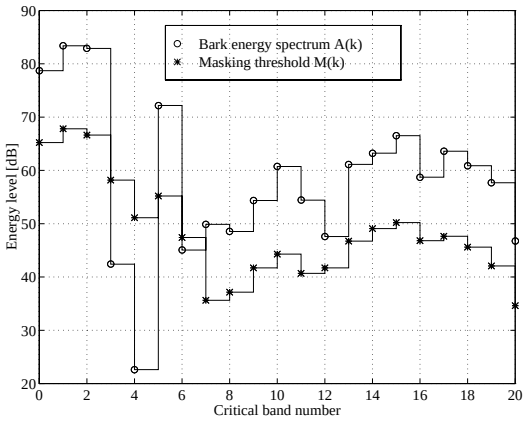

Figure 6: *Calculation of masking threshold* $M(k)$.

As defined in Table 1, each critical sub-band k, in one frame, contains l "temporal" transform coefficients. Thus, we can assume that the masking threshold $M(k)$ is equally shared by the l coefficients in the band and that the quantization noise contribution σ^2_{qij} of coefficient X_{ij} is given by

$$\sigma^2_{qij} = M_{ij} = M(k)/l \ . \tag{8}$$

M_{ij} is the maximum unperceived noise tolerated in the corresponding rectangle of the time-frequency grid in Fig. 5. No use of absolute hearing thresholds has been made in this work. In what follows, we will drop stage index j for clarity.

2.4 Quantization

If all the coefficients are uniformly quantized, then the quantization step of coefficient X_i is given by

$$\delta_i = \sqrt{12 \cdot \sigma^2_{qi}} \ . \tag{9}$$

Any coefficient X_i verifying the constraint $X_i^2 \geq \sigma^2_{qi}$ will have to be considered as unmasked and must be finely quantized. The rest of the coefficients can be ignored or coarsely quantized. The number of levels to quantize each coefficient in band k is calculated by Eq. 10, where $\lfloor \ \rfloor$ stands for "the integer part of".

$$L_i = \left\lfloor \frac{|X_i|}{\delta_i} + 0.5 \right\rfloor \tag{10}$$

The encoder will have to transmit, for each coefficient X_i, the following information: a masked/unmasked flag, L_i, δ_i and the sign of the coefficient.

The average bit rate prior to any entropic coding method has been evaluated to 43.3 Kbits/s and the segmental signal-to-noise ratio (SNR_{seg}) to 23 dB. At this bit rate, informal listening tests showed that the reconstructed signal was perceptually indistinguishable form the original one. At 37.3 Kbits/s the SNR_{seg} decreased to 22 dB and the quality of the coded signal was

still judged as excellent. These values were computed onto a set of 8 male/female sentences having a total duration of 30 seconds. A plot of the performance is shown in Figure 7. At lower bit rates, here 17.3 Kbits/s, the signal is distorted by a rumbling noise; however, its intelligibility is completely preserved.

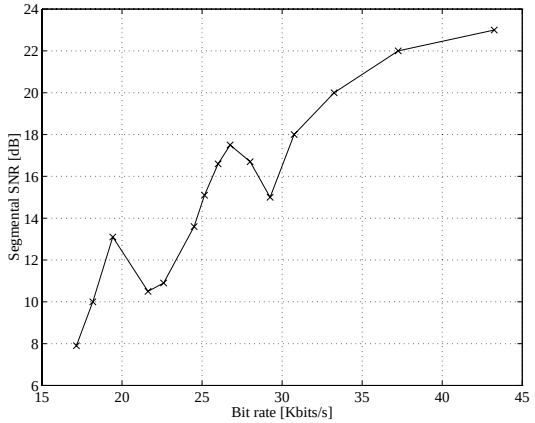

Figure 7: *Performance of the proposed encoder.*

3 Conclusions

We have presented a new approach to the problem of encoding speech signals using a perceptual model. The novelty of the algorithm proposed here lies in the integration of psychoacoustic modeling and a fast overlapped block orthogonal transform, having a complexity close to FFT algorithms. This integrated approach provides a fine temporal resolution towards higher frequencies, allowing an accurate encoding of fast signal transitions. The transform performs the Bark mapping necessary to the calculation of auditory masking thresholds, which account for a significant reduction of the coder bit-rate. The present system has been tested at 16-48 Kbits/s rates and shows a perceptually transparent quality at 43.3 Kbits/s. Though this bit rate is still considerably high, exploiting temporal correlations of the transform coefficients, as well as making use of entropic coding may lead to a considerable reduction of its value.

References

[1] P. Noll, "Digital Audio Coding for Visual Communications", *Proc. of IEEE*, vol. 83, pp. 925–943, June 1995.

[2] A. Drygajlo, "Butterfly orthogonal structure for fast transforms, filter banks and wavelets", *in Proc. of ICASSP'92*, vol. V, pp. 81–84, San Francisco, March 1992.

[3] B. Carnero and A. Drygajlo, "Fast short-time orthogonal wavelet packet transform algorithms", *in Proc. of ICASSP'95*, vol. II, pp. 1161–1164, Detroit, May 1995.

[4] M.L.Wickerhauser, *Adapted Wavelet Analysis: from Theory to Software*, A K Peters, Wellesley, MA, 1994.

[5] E. Zwicker and R. Feldtkeller, *Psychoacoustique, l'oreille, récepteur d'information*, Masson, Paris, 1981.

[6] D. Sinha and A. Tewfik, "Low Bit Rate Transparent Audio Compression using Adapted Wavelets", *IEEE Trans. on Sig. Proc.*, vol. 41, pp. 3463–3479, December 1993.

[7] M.R. Schroeder, B.S. Atal, and J.L. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear", *JASA*, vol. 66, pp. 1647–1652, December 1979.

[8] J.D.Johnston, "Transform Coding of Audio Signals Using Perceptual Noise Criteria", *Select. Areas in Comm.*, vol. 6, pp. 314–323, February 1988.

[9] D. Ştefănoiu, R. Kastantin, and G. Feng, "Speech coding based on the discrete-time wavelet transform and human auditory system properties", *in Proc. of EUROSPEECH'95*, vol. 1, pp. 661–664, Madrid, September 1995.

[10] M.Black and M. Zeytinoğlu, "Computationally efficient wavelet packet coding of wide-band stereo audio signals", *in Proc. of ICASSP'95*, pp. 3075–3078, Detroit, May 1995.