PERCEPTUAL SPEECH CODING USING TIME AND FREQUENCY MASKING CONSTRAINTS

Benito Carnero[1] *Andrzej Drygajlo*[1]

[1] Signal Processing Laboratory, Swiss Federal Institute of Technology of Lausanne, CH-1015 Lausanne, Switzerland
e-mail: carnero@lts.de.epfl.ch

ABSTRACT

This paper presents a new wide-band speech coding system based on a fast wavelet packet transform algorithm as well as a formulation of temporal and spectral psychoacoustic models of masking. The proposed 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 that used uniform filter banks. As a result of such a decomposition, the perceptually tuned time-frequency structure of the original speech signal is preserved. This allows us to make use of the temporal and spectral properties of the human auditory system to decrease the average bit rate of the encoder, while perceptually hiding the quantization error.

1. INTRODUCTION

Nowadays, wide-band speech compression is an active research area. The higher quality of wide-band speech is desirable for the extended communication tasks, such as audioconference, loudspeaker telephony, multimedia, etc. On the other hand, achieving bit rates as low as possible is a consequence of the ever increasing communication demand.

Currently, the ITU standard for wide-band speech coding, G.722 [1], uses sub-band adaptive differential pulse code modulation and produces excellent speech quality at 64 kbit/s. Large efforts have been made to reduce this bit rate, while trying to preserve the same perceptual quality [2]. All new issues in wide-band speech coding use the G.722 standard as a reference for performance comparison. They are mainly based on transform/sub-band coding, using perceptual criteria that 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. Furthermore, no use is made of temporal masking models.

In this paper, we present an integrated approach to the design of a wide-band coder for speech signals sampled at 16 kHz. It merges, fast orthogonal wavelet packet transform algorithms matched to auditory critical band decomposition, with models of temporal and spectral psychoacoustic masking constraints.

2. DESCRIPTION OF THE ALGORITHM

The block diagram of the proposed algorithm is represented in Fig. 1. The encoder is mainly composed of an analysis transform that performs simultaneously an approximated auditory time-frequency decomposition, as well as a uniform decomposition. The former provides the coefficients to be encoded and also used for the calculation of a masking threshold, while the latter is employed to estimate the tonality of the current analysis frame. Then, spectral as well as temporal models of masking are successively computed from the auditory transform coefficients and a final masking threshold is calculated. Next, the auditory coefficients are uniformly quantized according to the obtained masking threshold. The decoder is much simpler, compared to the encoder, since it is simply composed of a de-quantization step, plus reconstruction using a synthesis transform.

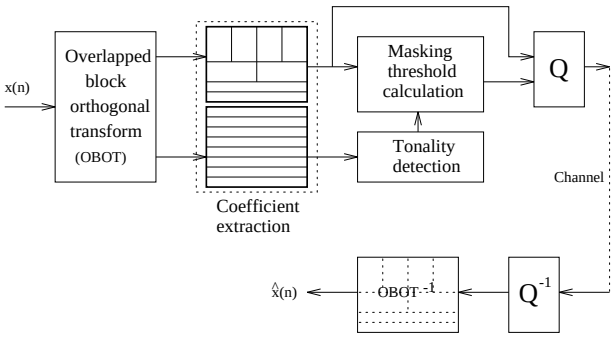

Figure 1. Block diagram of the codec

2.1. Orthogonal wavelet packet transform

Recently, wavelet and wavelet packet transforms have emerged as powerful and elegant solutions for nonstationary signal analysis and coding [3], where they allow for flexible time-frequency decompositions. Such analysis tools are traditionally implemented using tree-structured filter banks. In a previous work, efficient algorithms, with computational loads close to FFT algorithms, have been proposed [4, 5]. These *overlapped block orthogonal transforms* provide, in one block operation, all possible multiresolution time-frequency coefficients computed successively by tree-structured approaches.

Several studies have highlighted the nonuniform temporal and spectral resolutions of the human ear [6]. The spectral decomposition performed in the cochlea follows a Bark scale. Its related filters are called critical band filters. In the 0-8 kHz bandwidth, there are 21 critical bands that can be approximated with an overlapped block orthogonal transform with a frame lenght of $N = 64$. This corresponds to a temporal duration of 4 ms. The choice of the prototype filter of the transform, as well as its length, influences the separation of the sub-band signals. The Daubechies filters are the ones which achieve the best separation when the number of frame coefficients N increases, because of their regularity property [7]. Here, the prototype filter is of length 10. The resulting wavelet packet approximations to the Bark scale center frequencies (critical band number) and to the critical bandwidths are represented in Fig. 2.

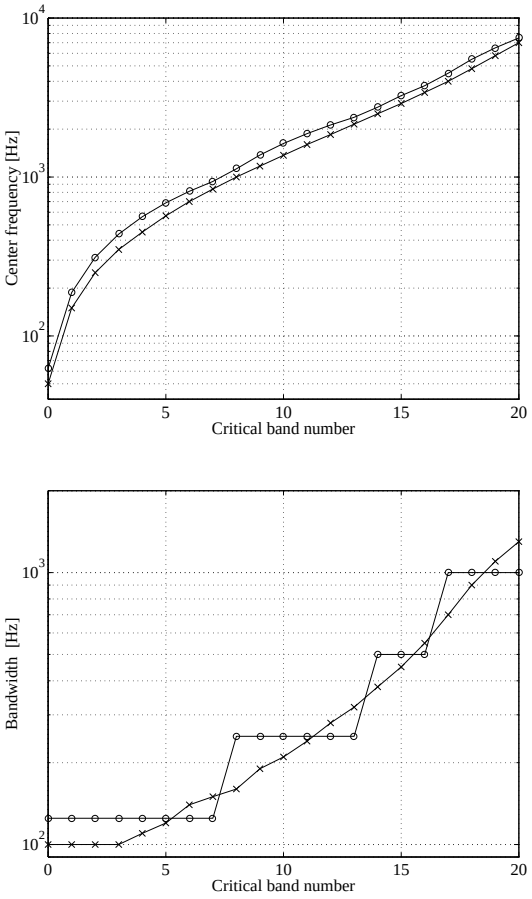

Figure 2. Approximation of the critical band decomposition. Top: center frequencies. Bottom: bandwidths. '×' are reference values and 'o' are approximation values.

2.2. Calculation of the masking threshold

The masking threshold in each critical band determines the tolerable quantization noise that can be introduced in that band without being perceived by the human ear [8]. The time-frequency resolution of employed transform can be described in terms of the grid that is shown in Fig. 3. The energy in each critical band is measured by calculating the Bark spectrum, whose staircase approximation $A(k)$, with $0 \leq k \leq 20$, is computed as

$$A(k) = \frac{1}{l} \sum_{i} (X_{ij})^2 \ . \qquad (1)$$

Here, X_{ij} is a transform coefficient, corresponding to one rectangle of the time-frequency grid; i is the coefficient number, $0 \leq i \leq 63$; j is the transform stage from which X_{ij} is chosen; and, l represents the number of "temporal" coefficients in critical band k, which is one row of the grid. The computed wavelet packet transform has $b = \log_2 64 = 6$ stages. The Bark spectrum calculation is done according to the values given in Table 1. A relative masking threshold

sub-band k	l	coefficient i	stage 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 orthogonal transform.

can be obtained for each critical band k, depending on the tonality nature of the input signal within a processed frame. It is a provisory masking threshold before considering spreading and is computed as a negative shift of the Bark spectrum level. Since the employed overlapped block transform already introduces some spectral overlapping, we have estimated the relative shift to be -20 dB for tonal blocks and -10 dB for noise-like blocks. An approach similar to that proposed in [8] has been employed to estimate intermediate tonality cases, by defining the relative shift to be

$$O(k) = -\alpha \cdot 10 - (1 - \alpha) \cdot 20, \quad [\text{dB}] \ . \qquad (2)$$

The parameter α is a spectral flatness measure, estimated over the sub-band signal energies and obtained at the maximal spectral resolution (decomposition depth) of the wavelet packet transform. Hence, the relative masking threshold becomes

$$A'(k) = A(k)/O'(k) \ , \qquad (3)$$

with $O'(k) = 10^{\frac{O(k)}{10}}$.

The deflection energy induced by a sound onto the basilar membrane spreads along its length. This corresponds to the simultaneous response of neighboring bands to the

sound located in one given critical band. Such energy spreading rises the tolerated noise floor $A'(k)$ by an amount that can be computed by convoluting $A'(k)$ with a spreading function expressed as

$$B(n) = a + \frac{v+u}{2}(n+c) - \frac{v-u}{2}\left(t + (n+c)^2\right)^{1/2} \quad , \quad (4)$$

in dB [9]. This approach is based on the hypothesis that masking can be considered as additive. $B(n)$ is a triangle-shaped curve. v represents the lower slope in dB/Bark, u the upper slope in dB/Bark, t the peak flatness, and c and a are compensation factors needed to satisfy $B(0) = 1$ (see Table 2). The parameter n varies from -20 to 20 Bark. The raised spread threshold is given by

$$C(k) = \sum_{m=0}^{20} A'(m) \cdot B'(k-m), \quad k \in [0, 20], \quad (5)$$

after the conversion $B'(n) = 10^{\frac{B(n)}{10}}$. In our context, we have estimated v to be 30 dB/Bark, u to be -25 db/Bark and t to be 0.3. Finally, the *frequency masking threshold* or tolerable noise contribution σ^2_{qij} associated with coefficient X_{ij}, in critical band k, is given by

$$M_{ij} = \sigma^2_{qij} = C(k) \ . \quad (6)$$

This simultaneous masking approach allowed us to obtain an average bit rate of 41.5 kbit/s prior to any entropic coding method. A similar simplified approach has already been presented in [10], at 43.3 kbit/s, using fixed relative shifts $O'(k)$.

In order to further decrease the bit rate without degrading perceptually the speech signal, a non-simultaneous or temporal masking factor is estimated over the different transform coefficients within a given critical band. Again, some amount of temporal masking is already introduced by the temporal overlapped nature of the transform basis functions. Temporal masking is maximum for signals close in frequency and within the same critical band. Furthermore, non-simultaneous masking is a nonlinear perceptual phenomenon [6]. Thus, a worst-case masking model must be adopted. This kind of masking can also be interpreted in terms of energy spreading across time. A model of temporal masking can be formulated by parameterizing a new spreading function $D(n)$ similar to Eq. 4. In this paper, non-simultaneous masking is also considered as additive.

The maximal temporal resolution takes place in the high-frequency critical bands, as represented in the grid of Fig. 3. This corresponds to a minimum grid step (mgs) of 0.5 msec or 8 samples. The employed spreading function $D(n)$ has parameters $v = 20$ dB/mgs (40 dB/ms), $u = -10$ dB/mgs (-20 dB/ms), and $t = 0.3$ (see Table 2), which is in accordance with psychoacoustic curves presented in [11]. A convolution is performed in each critical band k between its squared coefficients X^2_{ij} and a re-sampled version of the temporal spreading curve $D'(n)$, with $D'(n) = 10^{\frac{D(n)}{10}}$. This re-sampling is necessary since the number of temporal coefficients and the grid step is different in each critical band. The procedure is shown in the top of Fig. 3, where

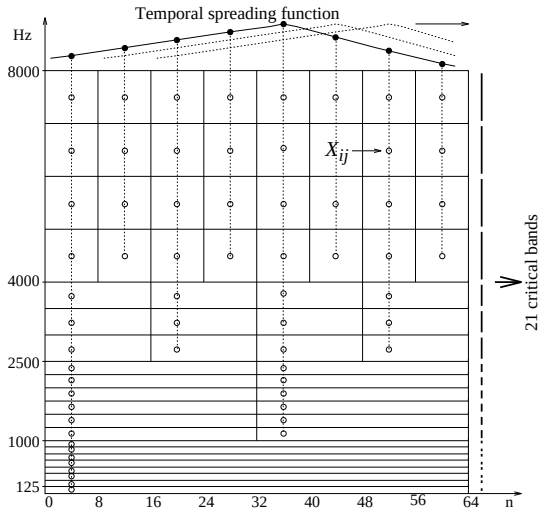

Figure 3. Time-frequency grid of the transform and calculation of temporal masking.

the vertical dotted lines stand for the re-sampling operation. The result of this convolution is an estimation of a temporal spread energy $T_{ij} \geq X^2_{ij}$, from which a *temporal masking factor* is defined as

$$\beta_{ij} = T_{ij} / X^2_{ij} \geq 1 \ . \quad (7)$$

A factor $\beta_{ij} = 1$ means that no temporal masking has been induced by neighbor coefficients onto X_{ij}. On the other hand, a higher β_{ij} shows that some amount of temporal masking is occurring. Thus, the time-frequency masking threshold in critical band k can be estimated as by rising the frequency masking threshold $C(k)$ by the amount β_{ij}:

$$M_{ij} = \sigma^2_{qij} = C(k) \cdot \beta_{ij} \ . \quad (8)$$

Spreading function	v	u	t	c	a
spectral	30 $\frac{\text{dB}}{\text{Bark}}$	-25 $\frac{\text{dB}}{\text{Bark}}$	0.3	0.09	27.39
temporal	20 $\frac{\text{dB}}{\text{mgs}}$	-10 $\frac{\text{dB}}{\text{mgs}}$	0.3	0.19	7.75

Table 2. Parameters of the spreading functions (mgs = minimum grid step).

2.3. Quantization

Stage index j can now be dropped for clarity. If all the coefficients are uniformly quantized, then the quantization step of coefficient X_i is given by

$$\delta_i = \sqrt{12 \cdot \sigma^2_{qi}} \quad (9)$$

Any value $|X_i(k)|$ which lies over σ_{qi} has to be considered as unmasked and must be finely quantized. On the other hand,

coefficients below σ_{qi} can be ignored or coarsely quantized. The number of levels required to quantize each coefficient in band k is calculated by Eq. 10, where $\lfloor\rfloor$ stands for "the integer part of".

$$L_i(k) = \left\lfloor \frac{|X_i(k)|}{\delta_i} + 0.5 \right\rfloor \qquad (10)$$

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

3. MAJOR RESULTS

By taking into account a temporal masking model, an average bit rate reduction of more than 5% has been obtained with respect to the method which considered only frequency masking. Thus, perceptually transparent coding has been achieved at 39.4 kbit/s, prior to the use of any lossless compression procedure. Furthermore, if the amount of tolerable quantization noise σ_{qi}^2 is increased, such that it becomes more and more perceptible, the bit rate can be gradually reduced without affecting drastically the intelligibility of the decoded speech signal. This result has been verified at a 15 kbit/s rate and is mainly due to the fine temporal representation of the higher frequencies by the wavelet packet transform, which preserves abrupt signal transitions.

The quality of the encoded speech signals at 39.4 kbit/s has been found subjectively equivalent to that of the G.722 standard at 64 kbit/s. However, the processing delay of the proposed algorithm, of about 34.5 ms, is large compared with the 3 ms of G.722 and may cause echo problems when used in some telecommunication systems.

4. CONCLUSIONS

We have presented a new approach to the problem of encoding wide-band speech signals using a perceptual model. The novelty of the algorithm proposed here lies in the use of a fast overlapped block orthogonal transform, which has allowed the formulation of a spectral and temporal masking model. Particularly, the consideration of temporal masking has led to a bit rate gain of more than 5% over previous coding schemes making use of only simultaneous masking. This improvement is achieved without perceptual degradation in the quality, compared to the original speech.

REFERENCES

[1] P. Mermelstein, "G.722, A New CCITT Coding Standard for Digital Transmission of Wideband Audio Signals", *IEEE Communications Magazine*, vol. 26, pp. 8–15, January 1988.

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

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

[4] 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.

[5] 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.

[6] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, Springer-Verlag, Berlin, 1990.

[7] 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.

[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] W. A. Deutsch, A. Noll, and G. Eckel, "The Perception of Audio Signals Reduced by Overmasking to the Most Prominent Spectral Amplitudes (peaks)", *in 92th. AES Convention, Preprint 3331*, Vienna, 1992.

[10] B. Carnero and A. Drygajlo, "Perceptual Coding of Speech Using a Fast Wavelet Packet Transform Algorithm", *in Proc. of EUSIPC0-96*, pp. 1661–1664, Trieste, Italy, September 1996.

[11] T. Sporer, U. Gbur, J. Herre, and R. Kapust, "Evaluating a Measurement System", *JAES*, vol. 43, pp. 353–362, May 1995.