# INTEGRATED SPEECH ENHANCEMENT AND CODING IN THE TIME-FREQUENCY DOMAIN

*Andrzej Drygajlo*[1] *Benito Carnero*[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 addresses the problem of merging speech enhancement and coding in the context of an auditory modeling. The noisy signal is first processed by a fast wavelet packet transform algorithm to obtain an auditory spectrum, from which a rough masking model is estimated. Then, this model is used to refine a subtractive-type enhancement algorithm. The enhanced speech coefficients are then encoded in the same time-frequency transform domain using masking threshold constraints for quantization noise. The advantage of the proposed method is that both enhancement and coding are performed with the transform coefficients, without making use of the additional FFT processing.

## 1. INTRODUCTION

Usually speech enhancement is used as a preprocessing stage for speech coding in noisy environments. In many coding situations, however, the enhancement system as well as the coding system use some kind of spectral mapping in order to decorrelate and separately process signal components. Thus, taking advantage of a common processing structure an integrated approach should reduce the complexity of the whole system. On the other hand, the theory of the quantization of noisy sources states that the optimum quantization solution, in a minimum mean square error sense (MMSE), for a Gaussian signal and an additive Gaussian noise is composed of an optimum MMSE filter, followed by a separate optimum MMSE quantizer [1]. This suggests that enhancement and coding could be appropriately performed in a spectral domain after the noisy signal has undergone only one transform operation.

The incorporation of masking constraints in the quantization process of spectrally-transformed signals has led to important gains in terms of bit rate, while preserving signal components that are perceptually relevant signals [2]. In an analogous manner, the application of auditory models to speech enhancement, based on spectral subtraction, provided valuable means of reducing residual "musical noise" artifacts, while maintaining intelligibility of the enhanced signal [3]. Indeed, sub-band coders calculate a noise masking threshold for the speech quantization stage, which is similar to the one used for the adaptation of the subtraction parameters in subtraction-based speech enhancement algorithms.

In this paper, we present an integrated approach to the enhancement and coding of speech signals sampled at 16 kHz, by incorporating fast orthogonal wavelet packet transform algorithms to perform an auditory representation. Such a system should control bit rate, while maintaining acceptable speech intelligibility and quality. It should also lead to a decreased computational load, compared to the use of separate coding and enhancement solutions.

## 2. DESCRIPTION OF THE ALGORITHM

The clean speech signal $s(n)$ is assumed to be corrupted by an additive independent noise $d(n)$, giving the noisy signal $y(n) = s(n) + d(n)$. The block diagram of the proposed enhancement/coding solution is presented in Fig. 1. The noisy input signal $y(n)$ is decomposed with the help

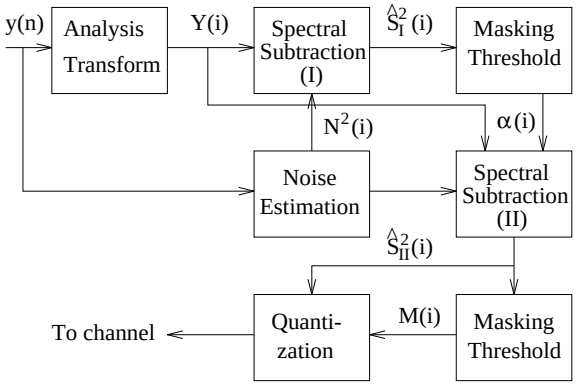


**Figure 1. Proposed enhancement/coding system.**

of an *overlapped block orthogonal transform* presented in Section 2.1. Then, the noisy transform sequence $Y(i)$ is enhanced by a subtractive-type system described in Section 3. to obtain the rough speech estimate $\hat{S}_I^2(i)$, which is further used to calculate a masking threshold, as described in Section 2.2. Using this first rough masking threshold, a new subtraction rule is applied to compute $\hat{S}_{II}^2(i)$ which depends on masking. This is achieved by assuming that the high-energy frames of speech will partially mask the input noise (high masking threshold), reducing the need of a strong enhancement mechanism. On the other hand, frames containing less speech (low masking threshold) will undergo an overestimated subtraction. Finally, $\hat{S}_{II}^2(i)$ is quantized by re-computing its related masking threshold. The employed quantization procedure is explained in Section 4. Noise estimation is performed during speech pauses.

### 2.1. Overlapped block transform

The implemented orthogonal decomposition possesses an efficient algorithm, with a computational load close to FFT algorithms, as it has been described in [4, 5]. In order to approximate the 21-band Bark or critical band mapping

$(0 \leq k \leq 20)$ performed by the human ear in the 0-8 kHz bandwidth, an overlapped block orthogonal transform has been developed with $N = 64$ frame coefficients (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, due to their regularity property, are the ones which achieve the best separation when the number of frame coefficients $N$ increases [6]. The time-frequency grid resulting from the chosen analysis transform appears in Fig. 2. Coefficient $Y(i)$ is the analysis transform coefficient corresponding to a rectangle of the time-frequency grid, as detailed in Table 1.

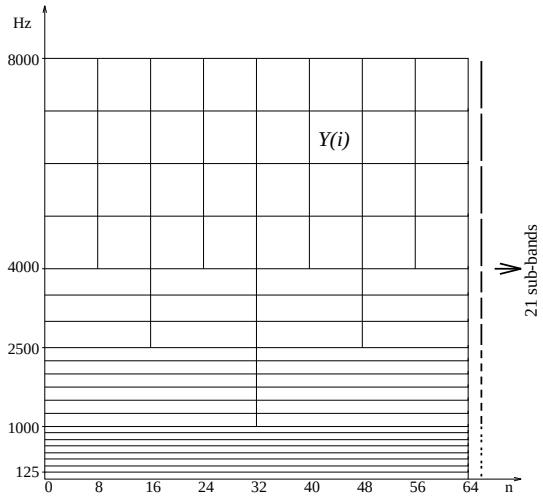


**Figure 2. Time-frequency grid of the analysis transform.**

### 2.2. Masking threshold calculation

The masking threshold determines the tolerable noise, either from enhancement or quantization, that can be introduced in each critical band without being perceived by the human ear [2]. Since the time-frequency grid of the chosen Bark decomposition appears as shown in Fig. 2, to obtain the staircase approximation of the Bark spectrum $A(k)$, a coefficient grouping is done according to the values given in Table 1. This calculation is described by Eq. 1. The value $l$ represents the number of "temporal" coefficients in the approximated critical band $k$.

$$A(k) = \frac{1}{l} \sum_{i} (Y(i))^2 \quad , \quad k \in [0, 20] \, . \tag{1}$$

A relative masking threshold can be defined in each critical band $k$ depending on tonality nature of the input signal in the processed block. It corresponds to a negative shift of the Bark spectrum level. Since the employed overlapped block transform introduces some spectral overlapping, we have estimated the relative shift to be -20 dB for tonal blocks and to be -10 dB for noise-like blocks. A similar approach to that proposed in [2] has been employed to estimate intermediate tonality cases, by defining the relative shift as

$$O(k) = -\gamma \cdot 10 - (1 - \gamma) \cdot 20, \quad [\mathrm{dB}] \, . \tag{2}$$

The parameter $\gamma$ is a spectral flatness measure, estimated over the sub-band signal variances. It is obtained by a uniform decomposition that reaches the maximal spectral resolution (decomposition depth) of the orthogonal transform.

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

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

This operation is simply an extension of the wavelet packet decomposition and is computed at the same time. Hence, the relative masking threshold becomes

$$A'(k) = A(k)/O'(k) \, , \tag{3}$$

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 neighbor bands to the sound located in one given critical band. Such energy spreading will rise the tolerated noise floor $A'(k)$ by an amount that can be computed by convolving $A'(k)$ with a spreading function expressed as [7]

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

in dB. This approach assumes that masking is 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 $B(0) = 1$. The parameter $n$ varies from -20 to 20 Bark. The raised spread threshold is given by

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

after the conversion $B'(n) = 10^{\frac{B(n)}{10}}$. In our context, we have estimated that $v = 30$ dB/Bark, $u = -25$ db/Bark and $t = 0.3$ (see Table 2). Finally, the frequency masking threshold or tolerable noise contribution $\sigma_q^2(i)$ of coefficient $Y(i)$, in critical band $k$, is given by $M(i) = \sigma_q^2(i) = C(k)$ .

This masking approach, in noise-free conditions allowed us to obtain an average bit rate of 43.3 Kbit/s, prior to any entropic coding procedure, as it has already been presented in [8].

| $v$ | $u$ | $t$ | $c$ | $a$ |
|---|---|---|---|---|
| 30 dB/Bark | -25 dB/Bark | 0.3 | 0.09 | 27.39 |

**Table 2. Parameters of the spreading function.**

## 3. SPECTRAL SUBTRACTION

No use of the FFT is made in the approach presented here. Hence, the spectral subtraction algorithm has to be justified in this context. Spectral subtraction is a well-known single-channel enhancement method used to reduce background noise [9]. If $d(n)$ is assumed to be uncorrelated with $s(n)$ and both are locally stationary sequences, over short time frames, it can be stated that $\Gamma_y(\omega,m) = \Gamma_s(\omega,m) + \Gamma_d(\omega,m)$; where $\Gamma_y(\omega,m)$, $\Gamma_s(\omega,m)$ and $\Gamma_d(\omega,m)$ are the short-time Power Spectra Densities (stPSD) of frame $m$ for $y(n)$, $s(n)$ and $d(n)$. Therefore, the clean speech stPSD can be estimated with the spectral subtraction rule

$$\hat{\Gamma}_s(\omega,m) = \Gamma_y(\omega,m) - \overline{\Gamma_d(\omega)} \ , \tag{6}$$

where $\hat{\Gamma}_s(\omega,m)$ is estimated over the last $L$ frames, during speech pauses, as

$$\overline{\Gamma_d(\omega)} = \frac{1}{L}\sum_{i=0}^{L-1} \Gamma_d(\omega, m-i) \ . \tag{7}$$

In the solution presented here, noise subtraction is performed in a time-frequency transform domain. This approach is derived from a generalized Wiener filtering solution formulated in [10] for unitary transforms. It states that, for zero mean signals, the optimum scalar filter, that weights the output of channel $i$ of the transform, can be expressed as

$$W_i = \frac{\sigma^2_{S_i}}{\sigma^2_{S_i} + \sigma^2_{D_i}} = 1 - \frac{\sigma^2_{D_i}}{\sigma^2_{Y_i}} \ . \tag{8}$$

The values $\sigma^2_{[]_i}$ are variances in the transform domain of the different signals. Then, the MMSE-estimated variance of the transformed speech signal in channel $i$ is given by

$$\hat{\sigma}^2_{X_i} = W_i\, \sigma^2_{Y_i} = \sigma^2_{Y_i} - \sigma^2_{D_i} \ , \tag{9}$$

which is similar to the spectral subtraction technique suggested in [9] and expressed by Eq. 6. Thus, PSD's in the Fourier domain have been replaced by variances in the new transform domain. This last formulation is valid for stationary signals. If processing is done on a frame-by-frame basis, where signals can be considered as stationary during the frame duration, the variances in the different transform channels can be estimated by squaring the transform coefficients (i.e.: $\sigma^2_{D_i} \rightarrow D_i^2$, $\hat{\sigma}^2_{X_i} \rightarrow \hat{X}_i^2, ...$). In that case, the frame index $m$ can be dropped. With the proposed transform, we can make use of Eq. 7 and Eq. 9 to write that

$$\hat{S}^2(i) = Y^2(i) - \overline{D^2(i)} \ . \tag{10}$$

The above expression is the one which is used to estimate $\hat{S}_I^2(i)$, in the subtraction block (I) of Fig. 1.

The basic magnitude spectral subtraction described in [9] suffers from a residual noise with a perceptually annoying musical structure. To reduce this effect, a parametric-type approach, using an overestimation factor $\alpha$ and spectral flooring $\beta$, was proposed in [11]. In this last method, $\alpha$ depends on the SNR of the frame $m$. With the proposed transform, this approach can be expressed as

$$\Delta(i) = Y^2(i) - \alpha\overline{D^2(i)} \tag{11}$$

$$\hat{S}^2(i) = \begin{cases} \Delta(i) & \text{, if } \Delta(i) > \beta\overline{D^2(i)} \\ \beta\overline{D^2(i)} & \text{, otherwise .} \end{cases} \tag{12}$$

A further improvement of this approach, to reduce residual noise, is based on the introduction of auditory masking properties and is proposed in [3]. In this last approach, both $\alpha(\omega,m)$ and $\beta(\omega,m)$ are made dependent on frequency and the current frame. Furthermore, they can be adapted based on the computation of a masking threshold. In this paper, a similar approach is proposed for the enhancement scheme. Parameter $\alpha$ has been also made dependent on the masking threshold level of transform channel $i$ and updated at each new frame. $\beta$ has been maintained at a fixed level of 0.001. The rule that governs the variation of $\alpha(i)$ is similar to the one that was proposed in [11] to render $\alpha$ dependent on the SNR of the frame. With the help of the adapted $\alpha(i)$ and Eq. 11, a refined spectral subtraction is performed to obtain $\hat{S}^2_{II}(i)$.

Prior to quantization, the coefficients of the transformed clean speech are estimated as

$$\hat{S}(i) = \text{sign}\,(Y(i))\,\sqrt{\hat{S}^2_{II}(i)} \ , \tag{13}$$

where the sign of the estimated coefficient is the sign of the noisy signal.

## 4. QUANTIZATION

If all the enhanced coefficients are uniformly quantized, then the quantization step of coefficient $\hat{S}_{II}(i)$ is given by

$$\delta(i) = \sqrt{12 \cdot \sigma_q^2(i)} \tag{14}$$

Any value $|\hat{S}_{II}(i)|$ which lies over $\sigma_q(i)$ will have to be considered as unmasked and must be finely quantized. On the other hand, coefficients below $\sigma_q(i)$ can be ignored or coarsely quantized. The number of levels to quantize each coefficient is calculated by

$$L(i) = \left\lfloor \frac{|\hat{S}_{II}(i)|}{\delta_i} + 0.5 \right\rfloor \ , \tag{15}$$

where $\lfloor\rfloor$ stands for "the integer part of". The encoder will have to transmit, for each coefficient $\hat{S}(i)$, the following information: a masked/unmasked flag, $L(i)$, $\delta(i)$ and the sign of the coefficient.

## 5. MAJOR RESULTS

The spectrogram of the original French sentence "Un loup s'est jeté immédiatement sur la petite chèvre", by a male speaker, is shown in Fig. 3(a). It has been corrupted with additive white Gaussian noise at a SNR of 5 dB, producing the spectrogram of Fig. 3(b). After the noisy signal has been enhanced and coded with the algorithms proposed above, it was decoded to give the reconstructed speech appearing in Fig. 3(c). Each spectrogram has 80 dB dynamics, from its maximum down to lower values. Fig. 3(c) shows that some vertical-line artifacts appear in the spectrum of the enhanced signal. They are perceived as a slight musical noise. However, its annoyance is much lower than with classical spectral subtraction. Furthermore, a good tradeoff between this residual noise and distortion due to enhancement can be achieved by means of the adaptation of $\alpha(i)$ in the transform domain.

As said in Section 2.2., the initial bit rate of the perceptual coding system, in the absence of noise, was an average of 43.3 Kbit/s. When speech is corrupted with noise, this bit rate increases as the SNR decreases. This suggests that

the masking model of the coder is very sensitive to input noise and even to the resulting residual noise, after speech has been enhanced. This particular problem needs a further study.

## 6. CONCLUSIONS

We have presented a new approach to the problem of the simultaneous enhancement and coding of wide-band speech in a time-frequency domain using a perceptual model. The novelty of the algorithm proposed here lies in the use of the same fast overlapped orthogonal transform both for the enhancement and coding processing. Before making use of any entropic coding method, the average bit rate obtained with this system is of about 43.3 Kbit/s, varying with the SNR of the input signal. A compromise among quality, intelligibility and bit rate is achieved by an adaptive tuning of a parametric subtraction rule. At high SNRs, coding is perceptually transparent. At low SNRs, residual musical noise is reduced while trying to preserve speech intelligibility.

## REFERENCES


[1] E. Ayanoğlu, "On optimal quantization of noisy sources", *IEEE Transactions on Information Theory*, vol. 36, pp. 1450–1452, November 1990.

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

[3] N. Virag, "Speech enhancement based on masking properties of the auditory system", *in Proc. IEEE ICASSP'95*, pp. 796–799, Detroit, May 1995.

[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] 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] 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.

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

[9] S. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", *IEEE Trans. on ASSP*, vol. 27, pp. 113–120, April 1979.

[10] W.K. Pratt, "Generalized Wiener Filtering Computation Techniques", *IEEE Transactions on Computers*, vol. c-21, pp. 636–641, July 1972.

[11] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise", *in Proc. IEEE ICASSP'79*, pp. 208–211, April 1979.


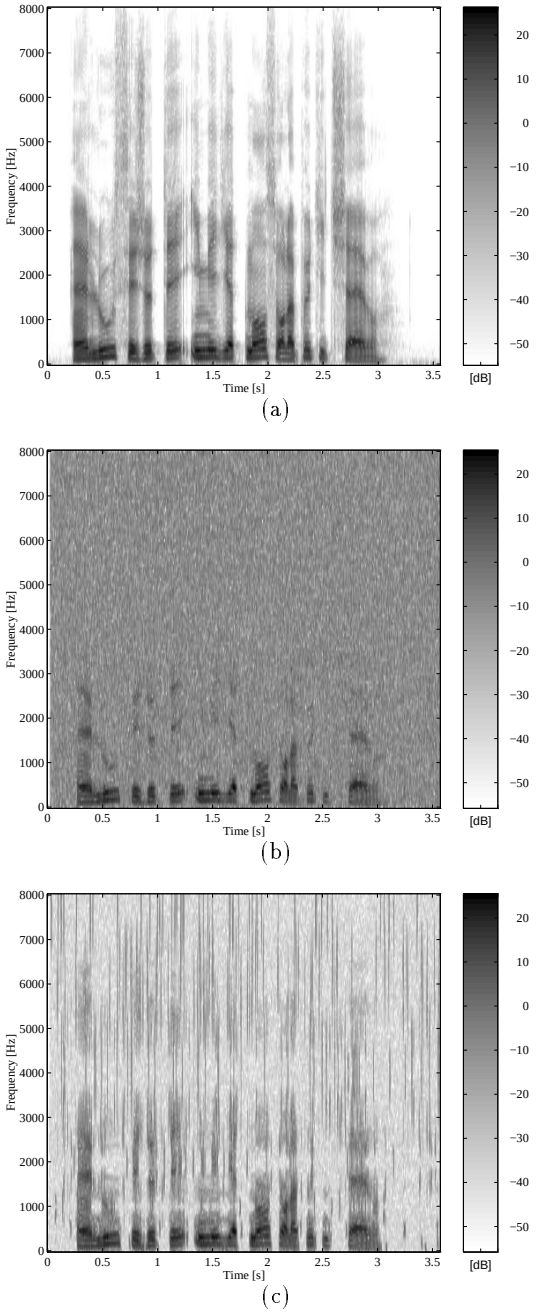


**Figure 3. Signal spectrograms. (a) Original speech. (b) Corrupted signal at 5 dB SNR. (c) Enhanced decoded signal.**