COMPARISON OF THE WAVELET DECOMPOSITION AND THE FOURIER TRANSFORM IN TCX ENCODING OF WIDEBAND SPEECH AND AUDIO

J-M. LeRoux†, R. Lefebvre††, J-P. Adoul††

† MATRA COMMUNICATION, Paris, France
†† University of Sherbrooke, Quebec, Canada

ABSTRACT

This paper reports on the specific contribution of the Wavelet Transform (WT) in the TCX coding model for audio signals. TCX, or "Transform Coded eXcitation", is a frame based coding algorithm that uses both time domain (linear prediction) and frequency domain (transform coding) approaches to exploit signal redundancies as well as frequency masking. While previous work on TCX used the Discrete Fourier Transform (DFT), the quality for highly non-stationary signals such as percussions was less than satisfactory. The WT has therefore been investigated as a compromise between time and frequency resolution.

1. INTRODUCTION

The demand for low bit rate wideband speech (50-7000 Hz) and audio is growing rapidly with the emerging applications of audio-video teleconferencing and multimedia. The main motivations for low bit rates are the need to minimize storage and transmission costs, along with the demand to transmit over channels of limited capacity [1].

While CELP coders produce high-quality speech at rates below 10 kb/s for the telephone bandwidth (300-3400 Hz) [2], and below 20 kb/s for wideband speech [3] [4], their complexity becomes exponentially large with bit rate to a point where they can no longer be implemented in real time using current DSP technology. Further, the underlying algorithm in CELP uses a source model that is too restrictive to allow efficient encoding of non speech-like signals such as music.

On the other hand, CELP coders incorporate a perceptual criterion in the form of a time-varying perceptual filter, which is used to transform the signals into a space that is more significant to the human ear. The distortion between the original and synthesis signals is measured in this so-called *perceptual space*.

The TCX coding model [5] has been introduced to alleviate the high complexity of CELP while retaining the masking properties of the coder. It has been shown to be more computationnally efficient than CELP, in particular for wideband speech and audio coding where large excitation codebooks are needed.

2. THE TCX CODING MODEL

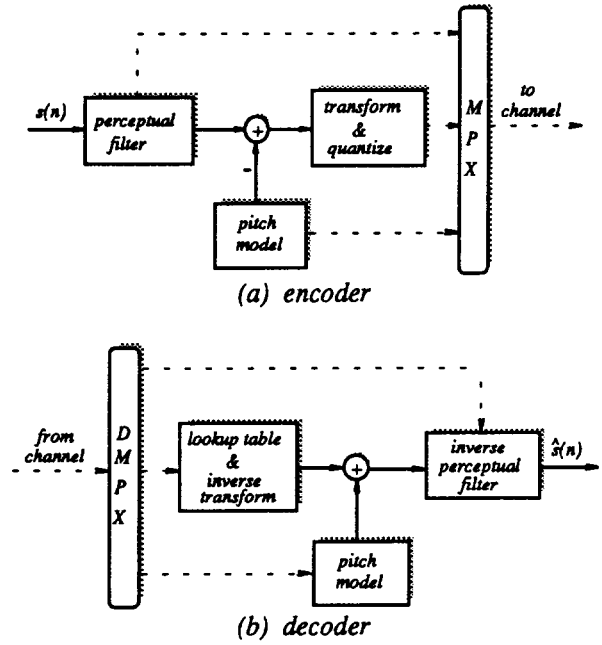

Figure 1. *TCX coding principle.*

Figure 1 shows the principle of TCX coding. The encoder operates in three basic steps. At each frame, the original signal *s(n)* is first perceptually filtered in such a way that the high energy portions of its spectrum are deemphasized, resulting in a flatter spectum. In this so-called *perceptual space*, the long-term redundancies are then removed by subtracting the contribution from a pitch model, which is typically a scaled, delayed version of the past synthesis in the perceptual space. Finally, the resulting signal is transform coded, with a mean-squared error (MSE) criterion. The transmitted information is thus the perceptual filter coefficients, the pitch model parameters (a set of delays and gains), and the result of the quantization in the transform domain.

At the decoder, these operations are performed in reverse order, as shown in Figure 1 (b). The quantization result is first recovered from a lookup table, and inverse transformed. Then , the so-called

 0-7803-2431-5/95 $4.00

"perceptual synthesis" is formed by adding back the pitch model contribution. Finally, the synthesis signal $\hat{s}(n)$ is obtained by filtering the perceptual synthesis through the inverse perceptual filter. It is this inverse filtering operation which carries out the noise shaping, so that coding noise will be properly masked.

The perceptual filter used in this work has a simple form, and is defined by Equation (1) :

$$W(z) = \frac{A(z)}{A(z/\gamma)} \qquad (1)$$

where $\gamma = 0.75$. The filter $A(z)$ is obtained by the autocorrelation method; its inverse, $1/A(z)$, models the short-term spectrum of the original signal.

In previous work on TCX, we used for the large part the Discrete Fourier Transform (DFT). For stationary signals having a non-flat power spectrum density (psd), the DFT is a natural choice to achieve good separation of the transform coefficients. With proper decimation and bit allocation strategies, the signal in the transform domain can be encoded effectively to achieve high compression ratios. Further, for a large class of audio signals, the short-term spectral envelope evolves slowly compared to the frame rate of the coder, so that predictive quantization can be applied to the amplitude spectrum to enhance performance. This motivated the use of the DFT with predictive quantization in [5], producing high-quality wideband speech at 16 kb/s, and 7 kHz bandwidth audio between 24 and 32 kb/s.

Some signals, however, fell short in terms of overall quality, namely sequences that are highly non-stationary such as percussions and transient signals. The problem is that the DFT of a non-stationary signal spreads over the whole spectrum, which makes compression in the transform domain a more difficult task. The wavelet transform (WT), a time-frequency analysis technique [6], has therefore been investigated as a way to better capture non-stationarities and localized events in the time domain.

3. WAVELET TRANSFORMS

Wavelets are a family of finite energy basis functions for which the frequency resolution Δf is no longer a constant as with the DFT, but rather varies with frequency. More precisely, the frequency resolution is inversely proportional to frequency (constant $\Delta f / f$). Hence, in constrast with the DFT which does not provide any time information about the signal within a frame, the WT coefficients carry an implicit time information as well as a frequency information. In particular, a concentration of energy in the time domain (such as a pulse) will produce, to some extent, a concentration of the WT coefficients.

In practice, the WT coefficients are obtained from an octave-band tree structure. The signal is decomposed in a number of subbands of unequal bandwidths by successively applying a set of basis lowpass and highpass filters (the highpass filter is actually the wavelet). At each level, the lowpass data from the previous level is further decomposed into a lowpass and a highpass signal, and then down-sampled by a factor of 2. The WT coefficients are formed by the highpass signals of each level together with the lowpass signal of the last level. Because of the decimation by 2 at each level, the WT of an N dimensional signal will produce N transform coefficients.

In our application, we used the orthogonal Daubechies wavelet with regularity 5, and 3 levels of decomposition [7].

4. BASIS FOR COMPARISON

The performance of the TCX coder described in Section 1 was evaluated in three different conditions: (1) the transform is a DFT; (2) the transform is the WT described in Section 3; and (3) the transform is the identity matrix. This third condition corresponds to quantizing the time domain signal without applying any transformation. It serves as a reference to determine the coding gain associated with the transform.

All other parameters of the coder were identical. The sampling rate is 16 kHz (7 kHz bandwidth). The frame length is 128 samples (8 ms). Filter $A(z)$, which also defines the perceptual filter $W(z)$ of Equation 1, has order $m = 16$. The pitch model in Figure 1 is a one-tap pitch predictor. The pitch delay can take on values 20 to 274 (fundamental frequency in the range 60 to 800 Hz). The delay is encoded in 8 bits, and the gain uses 4 bits. To avoid unstable filters, the pitch gain is saturated at a value of 1.2 before quantization.

Two basic experiments were conducted to compare the performance of the different transforms in the TCX coder. These are described below. The object of the first experiment was to determine the degree of "dispersion" of the transform coefficients. The second experiment evaluates the overall distortion of the coder with identical quantization strategies in the transform domain.

4.1 Experiment 1

The efficiency of a transform in a coder can be measured by the number of transform coefficients that

can be discarded (set to zero) without significant degradation. Only the significant coefficients need be encoded, thus reducing the required bit rate. An extreme example is that of a single non zero value. On the other hand, if the energy is spread uniformly over all coefficients, no coding gain can be achived by using the given transform.

To measure the "spread" over the transform coefficients, we used the following function:

$$\rho = \frac{E_{max}}{E_{avg}} \qquad (2)$$

where E_{max} and E_{avg} are defined as

$$E_{max} = \max_{k} \{t(k)^2\} \qquad (3)$$

$$E_{avg} = \frac{1}{K}\sum_{k=0}^{K-1} t(k)^2 , \qquad (4)$$

$t(k)$ is the result of the transform and K is the number of transform coefficients. Since $E_{max} \geq E_{avg}$, it follows that $1 \leq \rho \leq K$. A value of $\rho = 1$ means all $t(k)^2$ are equal (flat spectrum), while a value of $\rho = K$ means only one $t(k)$ is non zero. Hence, a large value of ρ indicates that the transform achieves a good separation in the transform domain.

In this experiment, the distribution of the spreading factor ρ defined by Equation (2) was computed for the WT, the DFT and the identity matrix, when used in the TCX coder. Two different audio signals were used: a castanet sequence, and a set of speech signals. The results are presented in Section 5.1.

4.2 Experiment 2

Using the same audio signals as in Experiment 1, the overall performance of the TCX coder is compared for the different transforms. To have the same quantization conditions for all transforms, the quantization is replaced by decimation. At each frame, the L smallest transform coefficients are set to zero and the other (K-L) are left intact. The corresponding signal-to-noise ratio (SNR) is computed. The results are presented in Section 5.2.

5. EXPERIMENTAL RESULTS

5.1 Results of experiment 1

Figures 2 and 3 show the results of the experiment described in Section 4.1. The transform used in each case is indicated on the histogram. When the identity matrix is used, the graph is labeled "time".

It can be observed, in the case of the castanet sequence (Figure 2), that the distribution of the spreading factor ρ is very similar for all three transforms. Assuming that ρ is a good criterion for estimating the number of coefficients that could be left uncoded without significant degradation, the results would suggest that for the castanet signal, the transform coding gain is negligeable for both the WT and the DFT. This is to be expected since this signal is mostly uncorrelated and non-stationary, as shown in Figure 6 (a).

In the case of speech signals, however, there is a significant advantage in using a non-trivial transform, as shown in Figure 3. The distribution of ρ spreads over higher values for the WT and the DFT than it does for the identity matrix. This suggests that in the case of speech, coding would be more efficient in the transform domain than in the time domain.

This experiment, however, does not allow to determine if any of the transforms (the DFT or the WT) has a definite advantage over the other.

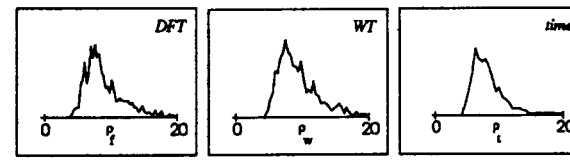

Figure 2. Distribution of ρ for castanets.

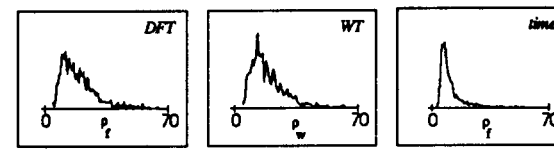

Figure 3. Distribution of ρ for speech.

5.2 Results of Experiment 2

Figures 4 and 5 show the results of the experiment described in Section 4.2. The SNR of the coder is shown as a function of the number of transform coefficients (in %) that are set to zero by the decimation.

For the castanet sequence (Figure 4), the decimation factor is varied between 20 and 95 percent. The WT yields the best performance overall, followed very closely by the time domain approach. The lowest SNRs are obtained with the DFT. On average, results obtained with the DFT are about 3 dB below that obtained with the WT. Hence, the WT has a significant advantage over the DFT in this case.

For speech (Figure 5), the decimation factor is varied between 33 and 98 percent. Note that the SNR of the

coder is much higher in this case; this is mostly due to the long-term predictor, which can account for most of the prediction gain in voiced regions. It can be seen that for speech, the time domain approach leads to considerably lower performance than with either the DFT or the WT. At the same time, the SNR obtained with the DFT and the WT are practically identical, as shown in Figure 5.

Hence, the use of the WT increases the coding gain in the case of a signal that is highly localized in time, while it performs as well as the DFT in the case of speech.

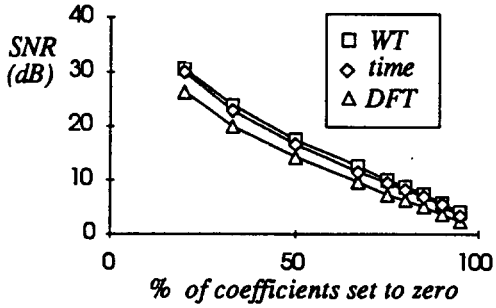

Figure 4. Result of decimation experiment for castanets.

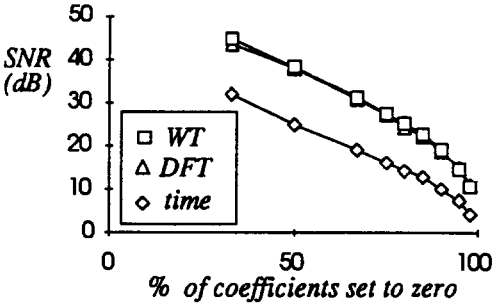

Figure 5. Result of decimation experiment for speech.

Finally, Figure 6 shows a segment of the castanet sequence. The original is shown, as well as the synthesis when using the DFT or the WT. In this particular example, the transform coefficients are actually quantized, according to [5]. The coder parameters, such as the bit assignments, are essentially the same for both transforms. The resulting bit rate is about 24 kb/s (1.5 bit per sample). It is obvious that the WT results in a synthesis that is closer to the original than with the DFT, and in particular that the attack is much more defined. Listening tests confirm this, showing that the WT makes the attacks sound more clear. In the case of speech, the quality of the synthesis for the DFT and the WT is comparable.

6. CONCLUSIONS

In this paper, we have presented a comparison between the DFT and the WT as used in the TCX coding algorithm. In this coding algorithm, a large portion of the bit rate can be allocated to the quantization of a perceptually enhanced signal, making it important to use an efficient quantization strategy. Transform coding allows, in general, important coding gains. The Fourier Transform provided excellent results in previous work on TCX, with the exception of highly non-stationary signals such as percussions. The use of the WT was therefore investigated. Results show that the WT produces a synthesis of better quality in the case of percussions in particular, and that a similar quality is achieved by the WT and the DFT in the case of stationary signals such as speech.

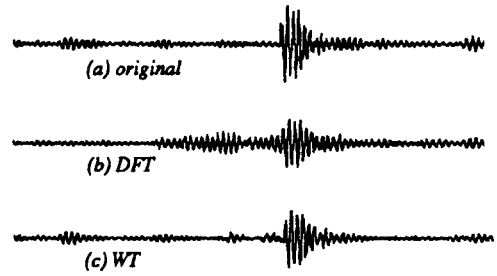

Figure 6. Segment of the castanet sequence for (a) the original,(b) the synthesis using the DFT, and(c) the synthesis using the WT.

7. REFERENCES

[1] P. Noll, "Wideband Speech and Audio Coding," *IEEE Comm. Mag.*, Vol. 31, No. 11, November 1993.

[2] M.R. Schroeder, B. Atal, "Code-Excited Linear Prediction (CELP): High-Quality Speech at Very Low Bit Rates," *IEEE Int. Conf. ASSP*, pp. 937-940, 1985.

[3] G. Roy, P. Kabal, "Wideband CELP Speech Coding at 16 kbits/sec," *Proc. IEEE Int. Conf. ASSP*, pp. 17-20, 1991.

[4] C. Laflamme, et al., "16 Kbps Wideband Speech Coding Technique Based on Algebraic CELP," *IEEE Int. Conf. ASSP*, pp. 9-12, 1991.

[5] R. Lefebvre, et. al, "High Quality Coding of Wideband Audio Signals using Transform Coded eXcitation (TCX)," *Proc. IEEE Int. Conf. ASSP*, 1994, pp. I-193-196.

[6] O. Rioul, M. Vetterli, "Wavelets and Signal Processing," *IEEE Sig. Proc. mag.*, October 1991.

[7] I. Daubechies, "Orthonormal Bases of Compactly Supported Wavelets," *Comm. in Pure and Applied Mathematics*, Vol. 41, No. 7, pp. 909-996, 1988.