HIGH QUALITY CODING OF WIDEBAND AUDIO SIGNALS USING TRANSFORM CODED EXCITATION (TCX)

R. Lefebvre *R. Salami* *C. Laflamme* *J-P. Adoul*

Department of Electrical Engineering, University of Sherbrooke,
Sherbrooke, Québec, Canada, J1K 2R1

ABSTRACT

This paper describes the application of Transform Coded Excitation (TCX) coding— an algorithm recently proposed by the authors— to encoding wideband speech and audio signals in the bit rate range of 16 kbits/s to 32 kbits/s. The approach uses a combination of time domain (linear prediction; pitch prediction) and frequency domain (transform coding; dynamic bit allocation) techniques, and utilizes a synthesis model similar to that of linear prediction coders as CELP. However, at the encoder, the high complexity analysis-by-synthesis technique is bypassed by directly quantizing the so-called target signal in the frequency domain. The innovative excitation is derived at the decoder by inverse filtering the quantized target signal. The algorithm is intended for applications whereby a large number of bits is available for the innovative excitation. In this paper the TCX algorithm is utilized to encode wideband speech and audio signals with 50-7000 Hz bandwidth. Novel quantization procedures including inter-frame prediction in the frequency domain are proposed to encode the target signal. The proposed algorithm achieves very high quality for speech at 16 kbits/s, and for music at 24 kbits/s.

1. INTRODUCTION

There is currently a growing demand for low bit rate wideband speech (50-7000 Hz) and audio coding for many applications such as audio-video teleconferencing and multimedia. ACELP coding has been successfully applied to obtain high quality wideband speech at 16 kbit/s and below [1]. However, the algorithm was fairly complex, and in order to achieve a single DSP chip implementation the codebook size had to be reduced resulting in a lower bit rate version with some quality degradation [2].

Recently [3], a new approach called Transform Coded Excitation (TCX) coding has been introduced as an efficient technique to encode the innovative excitation in CELP type speech coders at medium and high rates. The proposed algorithm utilizes time-domain linear prediction (LP) and pitch prediction (PP) analysis to determine the reconstructed signal. However, instead of using the computationally demanding analysis-by-synthesis techniques to determine the innovative excitation, the perceptually weighted signal with removed filter ringing and pitch correlations, better known as the target signal, is transformed and encoded in the frequency domain; it is then decoded and inverse transformed to extract the time-domain innovative excitation. The model is best applied to situations where a large number of bits is available for the excitation, such as medium-delay coding with backward LPC analysis, as was presented in [3], and in encoding wideband speech and audio signals as will be presented in this article.

The main advantage of TCX coding is its algorithmic simplicity (as the analysis-by-synthesis search procedure is eliminated). Simple scalar and vector quantization techniques are used and only one inverse filtering is needed to extract the innovative excitation. Further, quantizing the target in the frequency domain does not suffer from the defficiencies of traditional transform coding when applied directly to the original signal such as frame noise. In TCX the reconstructed signal is obtained by continous filtering which significantly reduces the framing discontinuities.

TCX coding can be seen as one possible approach to construct a target codebook at each frame of input signal. In CELP coding, the target codebook is constructed by filtering the entries of a fixed codebook through a time-varying (weighted) synthesis filter, which imposes the formant structure of the input signal on the target. Hence, the target codebook is adaptive, as its spectral characteristics evolve with the input signal. At each frame, the CELP coder selects the excitation that produces the reconstructed target (filtered excitation) closest to the target. In TCX coding, the fixed codebook is actually in the target domain. At each frame, the TCX coder first computes the quantized version of the target, and then calculates the corresponding excitation by filtering through the (zero-state) inverse weighted synthesis filter. Hence, there is no actual excitation codebook, in the sense that only one excitation per frame is computed and used. The target codebook in TCX can still be made adaptive, as in CELP, by the use of predictive encoding in the transform domain. In this case, the fixed codebook contains a set of predictive residuals. As will be shown later, successive amplitude spectra of the target are highly correlated, especially for some music signals, making predictive encoding of the target spectrum very effective.

The present paper concerns itself with the application of the TCX model to wideband audio signals (50–7000 Hz), speech and music, at rates ranging from 16 kbits/s to 32 kbits/s. At such high rates, the size of the innovative codebook of a CELP coder is exceedingly large, making analysis-by-synthesis virtually impossible to implement. The TCX model is attractive precisely because only one filtering is used to determine the innovative excitation, instead of one per innovative codebook vector. Most of the computational effort in TCX is due to the quantization of the target signal.

Section 2 presents the principle of TCX coding. In Section 3, we describe the quantization technique of the target signal, while Section 4 decribes the actual coding schemes used for both speech and music signals. Section 5 gives some

0-7803-1775-0/94 $3.00

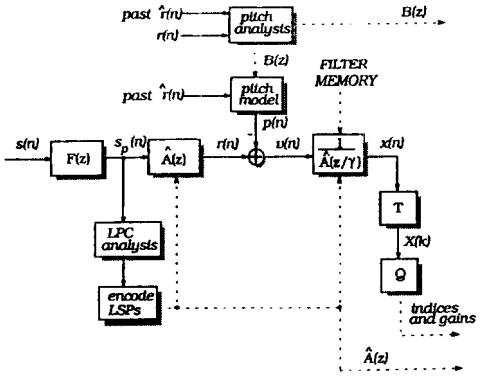

Figure 1. Principle of TCX coding.

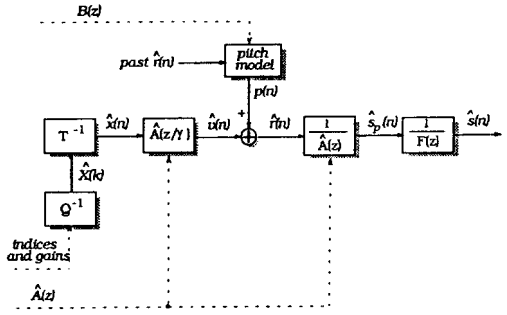

Figure 2. TCX decoding.

results and a discussion, and Section 6 the conclusions.

2. PRINCIPLE OF TCX CODING

Figure 1 shows the schematic diagram of the TCX coder. For each frame of N samples, the input signal $s(n)$ is first preemphasized with the filter $F(z) = 1 - \mu z^{-1}$, where $\mu = 0.5$, to increase the relative energy of the high-frequency components of the signal. This operation is of particular importance in wideband coding, to improve the higher order LP analysis, which is performed on the preemphasized signal $s_p(n)$ using the autocorrelation method. The LP coefficients of the filter $A(z)$ are quantized in the LSP domain [4]. Using the quantized version of the LPC filter, $\hat{A}(z)$, a residual signal $r(n)$ is computed. Closed loop pitch analysis is then performed to find the pitch delay and gain by minimizing the mean-square error between the weighted input speech and the past excitation $\hat{r}(n)$ filtered through the weighted quantized synthesis filter $1/\hat{A}(z/\gamma)$, with $\gamma = 0.8$ (the weighting filter is $W(z) = \hat{A}(z)/\hat{A}(z/\gamma)$). The pitch correlation is removed from the residual signal by subtracting the past excitation, with proper delay and gain, from the residual $r(n)$, to give signal $v(n)$. $v(n)$ is then filtered through $1/\hat{A}(z/\gamma)$ (with its initial states properly set) to give the target signal $x(n)$.

In traditional CELP coding, the next step involves filtering through $1/\hat{A}(z/\gamma)$ a set of innovative excitations from a codebook, to find the one that best matches the target $x(n)$. In TCX coding, the target signal is directly quantized, in the frequency domain, as shown in Figure 1, where the transformation T is a Fourier transform. The information transmitted by the coder is thus (1) the LPC parameters, in the form of quantized LSPs, (2) the pitch delay and gain, and (3) the result of the quantization of the complex target signal $X(k)$, namely a set of codebook indices I and gains g_c.

For the TCX coder to accomodate music signals, which can not in general be efficiently modeled by a unique set of equally spaced harmonics, the pitch predictor was only used on speech, and removed when coding music sequences. This resulted in a significant improvement in music quality. In this context, only the LPC coefficients (LSPs) and the quantized version of the target signal are transmitted.

At the decoder end (Figure 2), the quantized version of the complex target signal $\hat{X}(k)$ is inverse transformed to give the time-domain quantized target $\hat{x}(n)$. The innovative excitation $\hat{v}(n)$ is obtained by filtering the quantized target signal through the inverse weighted LPC filter, $\hat{A}(z/\gamma)$, with zero initial states. The pitch correlated excitation $p(n)$ is added back to the innovative excitation, and together they form the total excitation $\hat{r}(n)$, which produces the quantized preemphasized synthesis $\hat{s}_p(n)$ by filtering through $1/\hat{A}(z)$. Finally, the synthesis signal is found by filtering $\hat{s}_p(n)$ through the deemphasis filter $1/F(z) = 1/(1-\mu z^{-1})$.

3. QUANTIZATION OF THE TARGET

Replacing the low-energy, "white" innovation codebook of CELP with the high-energy, "colored" target codebook of TCX requires that the quantization procedure, denoted Q in Figure 1, be carefully designed to take into account the correlations between successive target signals. Since the target is obtained by subtracting the pitch correlations from the weighted input signal $s_w(n)$, it has to some extent the formant structure of $s_w(n)$. Hence, in the transform domain, successive amplitude spectra of the target will be correlated. This correlation will be particularly important in the case of music, where the target is simply the weighted input signal (after removing the zero-input response of $1/\hat{A}(z/\gamma)$). In general, the phase spectrum does not present inter-frame correlation as does the amplitude spectrum.

Figure 3 and Figure 4 show two consecutive amplitude spectra of the target for a voiced segment of male speech, and for an organ sequence, respectively. Recall that in the case of music, there is no pitch prediction, which allows larger frames to be used. It can be seen that the energy contour is roughly the same for the two consecutive speech frames, whereas the correlation is even greater in the case of music. In particular, for music, even the high-frequency fine structure is repeated from one frame to the next, for the example shown in Figure 4.

To exploit this redundancy, the phase and amplitude spectra of the target, respectively X_ϕ and X_a, are quantized separately. Differential quantization is used to encode X_a, while direct quantization is used for X_ϕ.

3.1. Quantizing the amplitude spectrum

The quantization procedure of amplitude spectrum at frame m, $X_a^{(m)}(k)$, is as follows. At each frame m, the amplitude spectrum $X_a^{(m)}(k)$ can be expressed as the sum of two terms

$$X_a^{(m)}(k) = \beta \hat{X}_a^{(m-1)}(k) + R(k), \qquad (1)$$

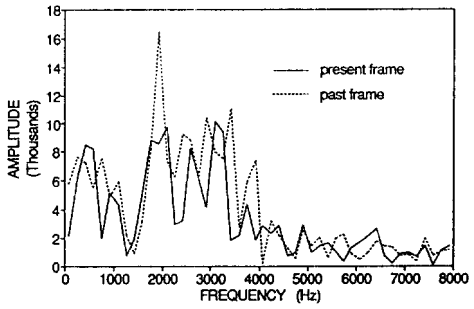

Figure 3. Two consective target amplitude spectra for a male speech sequence.

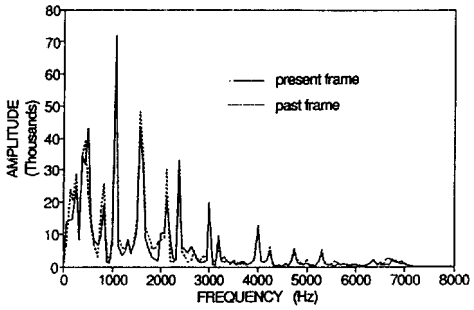

Figure 4. Two consective target amplitude spectra for an organ sequence.

where $\beta \hat{X}_a^{(m-1)}(k)$ is the predicted amplitude spectrum from past frame, and $R(k)$ is the prediction residual.

The best prediction gain $\hat{\beta}$ is first determined by selecting from a prediction gain codebook the one that minimizes the prediction error

$$E_p^2 = \sum_{i=0}^{K-1} \left[X_a^{(m)}(k) - \beta \hat{X}_a^{(m-1)}(k) \right]^2, \qquad (2)$$

where K is the number of amplitudes in the spectrum ($K = N/2$ where N is the frame size). Once the gain $\hat{\beta}$ is known, the prediction residual $R(k)$, given by

$$R(k) = X_a^{(m)}(k) - \hat{\beta} \hat{X}_a^{(m-1)}(k) \qquad (3)$$

is vector quantized. In general, a split-VQ approach is used, since the amplitude spectra are high-dimensional vectors. For each subvector, a gain and a shape are computed (gain-shape VQ). The total quantized amplitude spectrum is then given by

$$\hat{X}_a^{(m)}(k) = \hat{\beta} \hat{X}_a^{(m-1)}(k) + \hat{R}(k). \qquad (4)$$

The spectral prediction residual codebook, containing the reconstructed residuals $\hat{R}(k)$, is obtained using the k-means training algorithm. The training sequence of $R(k)$ vectors is obtained from a large database of speech or music signals.

The distribution of the prediction gain β was studied and found to lie between 0.2 and 1.2, for both speech and music. The maximum of this distribution lies around 0.75 for speech and around 0.9 for music. Recall that in the case of music, no pitch prediction is used, which increases the average correlation between successive amplitude spectra of the target.

3.2. Quantizing the phase spectrum

With R the total number of bits available to quantize the target, we have $R_\phi = R - R_a$, where R_ϕ and R_a are the number of bits to quantize the phase and amplitude spectra, respectively. The R_ϕ bits are dynamically allocated at each frame, as a function of the quantized amplitude spectrum $\hat{X}_a^{(m)}(k)$. This way, no side information is needed to transmit the bit allocation to the decoder, since it can be calculated locally.

The bit allocation algorithm is as follows. Let $\hat{X}_a$ be the quantized amplitude spectrum of the target and let K be the number of phases and amplitudes of the spectrum. Let R_ϕ be the number of bits to allocate to the phases. Finally, let r_k be the number of bits allocated to phase k, $k = 0, \ldots, K-1$.

- *Initialization:* set $r_k = 0$, for $k = 0, \ldots, K-1$.
- *Iteration:* Repeat R_ϕ times :
  1. Find the maximum amplitude of $\hat{X}_a$ at position k.
  2. $r_k = r_k + 1$ (allocate one bit to the kth phase).
  3. Divide $\hat{X}_a(k)$ by 2.

At the end, the r_k contain the bits allocated to the quantization of each phase.

A maximum of 7 bits was allowed for each phase. Experiments have shown that allocating 7 bits to each phase, at each frame, produced transparent quality speech and music, in all cases considered. Further, at the chosen encoding rates of 16 kbits/s and 24 kbits/s, the highest observed bit rate allocated to one given phase was 8 bits, using the bit allocation described above.

The phase spectrum $X_\phi(k)$ is then quantized with R_ϕ bits, following the bit allocation. Typically, it is this stage of the quantization procedure that requires the largest number of bits (in the order of 2 bits per phase, on average). The phases that are allocated a sufficient number of bits (for example, 2 bits or more) are separately scalar quantized; the other phases, that have been allocated fewer bits are block quantized. Many schemes are possible to vector quantize those phases. For instance, if 4 given phases were allocated respectively 0, 1, 0 and 1 bits, they could be block quantized using a 2-bit, 4-dimensional vector quantizer. This implies the use (and storage) of a multi-rate, possibly multi-dimensional, vector quantizer. We used a simpler scheme where all the phases that have been allocated 1 bit are submitted to a second simpler bit allocation, as follows. If there are K_l such phases, then the $K_l/2$ largest of those are allocated 2 bits, and the other $K_l/2$ (the smallest ones) are allocated 0 bit. All amplitudes corresponding to phases that are allocated 0 bit are set to zero. This results in dynamic decimation of the amplitude spectrum.

4. CODING SCHEMES FOR MUSIC AND SPEECH

The TCX coding algorithm was used to encode both speech and music files, in both configurations, namely with and without pitch prediction. Pitch prediction is used only in the case of speech. The pitch delay is encoded with 8 bits, with possible delays ranging form 40 samples (400 Hz) to 295 samples (54 Hz); the pitch gain is encoded with 4 bits.

Hence, a total of 12 bits are needed to encode the pitch parameters.

For speech, a frame length of $N = 96$ samples (6 ms) is chosen. This frame length is optimal for updating the pitch parameters, since smaller frames would result in an unnecessarily higher bit rate for updating the pitch parameters, and longer frames degrade the performance of the pitch predictor. As noted in Section 2, the value of γ in $W(z)$ is set to 0.8. At 16 kbits/s, there are 96 bits available at each frame. The 16 LPC parameters are quantized once every four frames with 48 bits, using split-VQ; the 16 LSPs are split into 7 subvectors of respectively 2, 2, 2, 2, 2, 3 and 3 LSPs, and each subvector is allocated 7 bits, except the last 3 LSPs which are allocated 6 bits. The LPC coefficients are linearly interpolated for the other three frames. This amounts to 12 bits per frame for the LPC coefficients. With already 12 bits used to transmit the pitch parameters, there are $(96 - 12 - 12)$ bits $= 72$ bits left at each frame to quantize the target signal. The 48-dimensional amplitude spectrum $X_a(k)$ is quantized as in Section 3.1, using 18 bits; the spectral prediction gain β is quantized with 5 bits, and the spectral prediction residual $R(k)$ is quantized with 13 bits (single 48-dimensional vector with 8-bit shape codebook, and 5-bit gain codebook). The phase spectrum $X_\phi(k)$ is quantized using the remaining 54 bits, as described in Section 3.2.

For music, a frame length of 256 samples (16 ms) is chosen, and no pitch prediction is used. Again, the value of γ in $W(z)$ is set to 0.8. At 24 kbits/s, 384 bits are available at each frame. The LPC coefficients are quantized and transmitted at each frame using 49 bits; the 16 LSPs are split into 7 subvectors of respectively 2, 2, 2, 2, 2, 3 and 3 LSPs, and each subvector is allocated 7 bits. Without pitch prediction, this leaves 335 bits to quantize the target vector. The amplitude spectrum of the target is quantized with 100 bits, as described in Section 3.1. A single spectral prediction gain β is used on the whole amplitude spectrum, and quantized on 5 bits. The spectral prediction residual is split into eigth 16-dimensional subvectors; the first 6 subvectors are quantized with 12 bits (8 bits for the shape and 4 bits for the gain - the gains are quantized in pairs with 8 bits per pair); the last two subvectors are quantized with 11.5 bits each (8 bits per shape, and 7 bits for that last gain pair, making it 3.5 bits per gain). The phase spectrum of the target is finally quantized with the remaining 235 bits $(384 - 49 - 100)$, as described in Section 3.2.

5. EXPERIMENTAL RESULTS

The TCX coder was tested on eight speech files, four male and four female, and eight music files, sampled at 16 kHz. The test music files contain a wide variety of sounds, ranging from organ, piano, orchestra, to a castanet sequence and also a capella singing (the often used Suzan Vega sequence). The encoding rate is 16 kbits/s for speech and 24 kbits/s for music.

With the coding scheme described in Section 4, the average SNR obtained for speech is 17.31 dB, with a minimum of 16.14 dB and a maximum of 18.42 dB. Note that because of the low value of γ in perceptual filter $W(z)$, the coder is not as good a waveform follower as with higher values of γ; this decreases the SNR values, in all cases, but increases the perceptual quality because of the noise masking properties of $W(z)$. With a γ value close to 1.0, the average SNR is about 19.0 dB, but the perceptual quality decreases significantly. The speech quality was judged very good by informal listeners.

The coding scheme used for music is also described in Section 4. In the case of music sequences, the SNR varies dramatically from one file to the other, ranging from 5 dB for the castanet sequence, to 14.5 dB for the organ sequence. This is to be expected since music is not in general as stationary as speech. For instance, the castanet sequence resembles a train of pulses, and any kind of predictive coder (either pitch prediction used on speech or spectral prediction used on music) will not be able to capture its highly temporally localized structure. Nevertheless, the perceptual quality of all music sequences was judged very good by informal listeners. In all cases, there is a perceptible difference between the original and synthesis signal, but the artifacts are not unpleasant, and in most cases are not detected when only the synthesis signal is listened to.

6. CONCLUSION

In this paper, we have proposed a wideband audio coder based on a new approach called TCX, which combines efficient time domain and frequency domain analysis to achieve the best perceptual reproduction of the original signal. The algorithm is based in part on the CELP model, with the main difference being that the target signal is directly quantized, in the frequency domain, and inverse filtered to give the innovative excitation, instead of using analysis-by-synthesis as in CELP. To make the target codebook adaptive, spectral prediction is used to encode the amplitude spectrum of the target. At each frame, most of the available bits are used to encode the target, with the phases requiring a larger portion of the bits than the amplitudes.

In the present work, two different schemes were used when encoding either speech or music, with pitch prediction being used only when encoding speech. A frame length of 6 ms was used for speech, whereas a frame length of 16 ms was used for music.

Informal listening test have shown that very high quality is obtained by the TCX coder, at 16 kbits/s for wideband speech and at 24 kbits/s music.

7. ACKNOWLEDGMENTS

The authors wish to thank the CITI (Center for Information Technologies Innovation; Industry and Science Canada) for the financial support provided for this research, especially Mr. Raymond Descout, Program Head of Multimedia Systems.

REFERENCES

[1] C. Laflamme, J-P. Adoul, R. Salami, S. Morissette, and P. Mabilleau, "16 kbps wideband speech coding technique based on algebraic CELP," *Proc. ICASSP'91*, Toronto, Canada, 14–17 May, 1991, pp. 177–180.

[2] R. Salami, C. Laflamme, and J-P. Adoul, "Real-time implementation of a 9.6 kbit/s ACELP wideband speech coder," *Proc. Globecom'92*, Orlando, Florida, Dec. 6–9, 1992, pp. 447–451.

[3] R. Lefebvre, R. Salami, C. Laflamme, and J-P. Adoul, "8 Kbits/s coding of speech with 6 ms frame-length," *Proc. ICASSP'93*, Minneapolis, Minnesota, April 27–30, 1993, pp. 612–615.

[4] K. K. Paliwal and B. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," *Proc. ICASSP'91*, Toronto, Canada, 14–17 May, 1991, pp. 661–664.