A CANDIDATE CODER FOR THE ITU-T'S NEW WIDEBAND SPEECH CODING STANDARD

*Juin-Hwey Chen**

Voxware, Inc.
Princeton, New Jersey, USA

ABSTRACT

This paper presents AT&T's candidate coder for the ITU-T's new wideband speech coding standard at 16, 24 and 32 kb/s. This coder achieves high speech quality with a low coder complexity. The basic idea of the coder is to perform closed-loop pitch prediction on perceptually weighted speech, and then quantize the prediction residual using perceptually based transform coding techniques. A first version of the coder based on DFT was thoroughly tested and submitted to the ITU-T in February 1996, and it was selected as one of two surviving candidates to advance to the next phase. A revised version based on MDCT was later submitted in October 1996. Both versions are described in this paper.

1. INTRODUCTION

The International Telecommunication Union-Telecommunication Standardization Sector (ITU-T), formerly known as the CCITT, is currently in the process of standardizing wideband speech coding at 16, 24, and 32 kb/s. Two modes of coders are being considered: Mode A is a low-delay coder, and Mode B is a low-complexity coder. The sampling rate is 16 kHz, and the nominal signal bandwidth is from 50 to 7000 Hz. This standard is expected to be finalized by mid-1998.

In February 1996, eight candidate coders were submitted to the ITU-T: four for Mode A and four for Mode B. Only two of the eight showed good enough performance to be admitted to the selection phase. One of these two surviving candidates is AT&T's Mode B candidate [1], which is a variation of the Transform Predictive Coding (TPC) algorithm reported in [2], [3]. This candidate coder uses the Discrete Fourier Transform (DFT) for transform coding of the pitch prediction residual. The coder performs quite well for clean speech signals, although its performance for music and noisy speech is not good enough.

To improve music coding performance, we changed the coder structure and replaced the DFT by the Modified Discrete Cosine Transform (MDCT) [4]. This revised version [5] was submitted to the ITU-T in October 1996. Due to space limitation, only high-level descriptions of these two TPC versions are given in this paper. More details can be found in [1] and [5].

The rest of the paper is organized as follows. Section 2 describes the DFT-based TPC candidate. Section 3 describes the MDCT-based TPC candidate. Sections 4 and 5 discuss the coder's performance and complexity, respectively.

2. DFT-BASED TPC CANDIDATE CODER

The block diagrams of the encoder and decoder of the DFT-based TPC candidate coder are shown in Figs. 1 and 2, respectively. Certain aspects of this coder are somewhat similar to the TCX coder [6], but there are also important differences.

The coder has a frame size of 20 ms (320 samples) and a subframe size of 4 ms (64 samples). A 16^{th}-order LPC analysis is performed once per frame. The Line-Spectrum Pair (LSP) coefficients are quantized to 49 bits using split VQ [7]. The quantized LSP coefficients are linearly interpolated for each subframe and converted to the LPC predictor coefficients.

The LPC prediction error filter in Fig. 1 removes the short-term redundancy from the input speech. The LPC prediction residual further goes through a shaping filter to produce a perceptually weighted speech signal. The shaping filter consists of a bandwidth-expanded LPC synthesis filter in cascade with a bandwidth-expanded second-order spectral tilt filter [8]. Similar to CELP coding, for each subframe, the zero-input response of the shaping filter is subtracted from the weighted speech to obtain the target vector for pitch prediction.

A three-tap pitch predictor is used. Once every frame, a pitch period represented by 8 bits is extracted from the LPC prediction residual in an open-loop fashion. For each subframe, the pitch period is linearly interpolated, and the 3 pitch taps are vector quantized to 6 bits using a closed-loop codebook search. The search is done in such a way that when the previously quantized LPC residual is filtered by the three-tap pitch synthesis filter and then by a shaping filter with zero memory, the output vector is closest to the target vector for pitch prediction. The output vector corresponding to the best set of pitch taps is subtracted from the target vector for pitch prediction. The resulting closed-loop pitch prediction residual is the target vector for transform coding.

The transform processor of Fig. 1 performs the 64-point FFT of this target vector and the subsequent normalization. The FFT coefficients are first normalized by the magnitude response of the shaping filter. The root-mean-square (RMS) values, or gains, of

* Work performed while the author was working for AT&T Labs - Research.

the normalized FFT coefficients over three frequency bands are then calculated. The gain of the 0-1250 Hz band is scalar quantized to 5 bits, and the gains of the 1250-4000 Hz and 4000-8000 Hz bands are vector quantized to 7 bits. The three quantized gains are used to further normalize the FFT coefficients. The real and imaginary parts of the final normalized FFT coefficients are directly quantized using VQ with adaptive bit allocation. The DC term is scalar quantized. Two-dimensional VQ is used for frequency components up to 4000 Hz. Above 4000 Hz, 4-dimensional VQ is used.

The quantized LPC spectrum is used to derived a noise masking threshold function [9], and it is also treated as the initial spectrum of coding noise. Based on this threshold function and initial noise spectrum, an approximation of the perceived noise loudness is calculated using a simplified version of the method outlined in [9]. A "greedy" adaptive bit allocation algorithm assigns one bit at a time to the frequency with the largest approximated noise loudness and then reduces the estimated noise power at that frequency by a pre-determined amount. The approximated noise loudness at that frequency is updated and the bit assignment process is repeated until all available bits are exhausted. There is no need to transmit the bit allocation result as side information, since the TPC decoder can perform the same adaptive bit allocation based on quantized LPC coefficients. The TPC candidate coder changes its bit-rate between 16, 24, and 32 kb/s by simply changing the number of bits allocated to the transform coding of the closed-loop pitch prediction residual.

At 16 kb/s, the bit allocation algorithm only allocates bits to frequencies below 4000 Hz. At 24 and 32 kb/s, all frequencies are allowed to receive bits. For each frequency above 4000 Hz which receives zero bits, the corresponding transform coefficient is synthesized as follows. The phase is random between 0 and 2π. The magnitude is determined by the signal-to-masking ratio (SMR), which is the ratio of the quantized LPC spectrum and the noise masking threshold. The magnitude is 0 dB where SMR > 5 dB, and it is -3 dB elsewhere. The high-frequency synthesis processor of Fig. 1 performs this operation.

The inverse transform processor first multiplies the quantized or synthesized FFT coefficients by the quantized gain and the magnitude response of the shaping filter. Then, it performs 64-point inverse FFT to obtain the quantized closed-loop pitch prediction residual, which is passed through the inverse shaping filter. The result is added to a pitch prediction vector to obtain the quantized LPC prediction residual vector. This vector is used to update the memory of the pitch predictor and the shaping filter used in the zero-input response vector calculation.

The decoder in Fig. 2 duplicates many encoder operations described above. A long-term postfilter, an LPC synthesis filter, and a short-term postfilter are added in order to synthesize the output speech from the quantized LPC prediction residual.

3. MDCT-BASED TPC CANDIDATE CODER

The encoder and decoder block diagrams of the MDCT-based TPC candidate coder are shown in Figs. 3 and 4, respectively. The main changes include the simplification of the encoder structure, the replacement of DFT by MDCT, and the changes in gain calculation and quantization. These are described below.

The encoder structure of Fig. 3 is much simpler than that of Fig. 1. The pitch predictor of Fig. 1 implicitly operates on *dt*, the quantized LPC prediction residual, although the distortion measure of the closed-loop codebook search for the three pitch taps is a mean-squared error (MSE) in the weighted speech domain. This means the shaping filter is involved in the codebook search, which means relatively high complexity.

In contrast, the pitch predictor in Fig. 3 directly operates on *wq*, the quantized version of the weighted speech signal *w*. Thus, the shaping filter is not involved in the codebook search, and even the zero-input response calculation is eliminated. This reduces the computational complexity. It also makes the encoding operation conceptually easier to follow — basically the input speech is passed through the weighting filter, the pitch-predicted component is subtracted, and then the residual signal is transform coded. The decoder simply reverses this process to get the output speech. The decoder structures in Fig. 2 and Fig. 4 are almost identical, except that the location of the inverse shaping filter has been moved. The simpler encoder structure in Fig. 3 makes the integration of MDCT much easier.

We replaced the 64-point FFT in TPC by a 128-point MDCT with a sine window and 50% overlap. Our ultimate goal is to use MDCT with adaptive window switching [10]. Such a technique is known to achieve high coding gain while avoiding the so-called "pre-echo" distortion [10]. However, we did not complete this window switching scheme in time before the ITU-T's October 1996 submission deadline. Therefore, our submitted candidate coder only uses a fixed 128-point short window.

Instead of 3 gain bands, we used 12 gain bands in this coder version, with the spacing roughly proportional to the Bark scale. A Karhunen Loeve Transform (KLT) with a fixed set of basis vectors (designed off-line) is applied to the logarithmic (dB) values of the 12 gains. The first KLT coefficient is scalar quantized to 5 bits. The second and the third KLT coefficients are vector quantized to 3 bits, and the fourth through the sixth KLT coefficients are vector quantized to 4 bits. The remaining KLT coefficients are set to zero. Applying inverse KLT gives the 12 quantized log-gains, which are then converted back to the linear domain. Although this scheme has not been fully optimized yet, it already gives better performance than the simple 3-band approach mentioned in Section 2.

4. PERFORMANCE

We have tested the DFT-based TPC candidate coder in a very extensive formal subjective listening test, following the qualification test plan specified by the ITU-T. The Mean Opinion Score (MOS) results showed that this TPC coder fully met all ITU-T's performance requirements for coding clean speech signals under single encoding, tandeming, input level variation, and frame erasure conditions. In fact, for the majority of the conditions, the MOS of this coder exceeded the requirements by a statistically significant margin. However, the coder did not meet the performance requirements for some of the conditions for coding music and for speech in background noise.

We did not have a chance to test the MDCT-based TPC candidate in an MOS test. Our informal listening showed that this coder gave noticeable improvements for some music signals,

but not all. Overall, the improvement for music coding is not as much as we had hoped. We think this is because the MDCT with a short 128-point (8 ms) window is not able to provide the same high coding gain as an MDCT with a much longer (e.g. 40 ms) window. For most speech signals, the output of the MDCT-based TPC candidate sounds roughly the same as that of the DFT-based TPC candidate, although it sounds slightly worse occasionally.

5. COMPLEXITY

We estimate that with careful code optimization, both versions of the TPC candidate can run full-duplex under 15 MIPS on a 16-bit fixed-point DSP. In fact, without any hand optimization of assembly code, our compiled C simulation code can run a full-duplex TPC candidate coder using only about 1/3 of the CPU power of either a 150 MHz SGI Indy workstation or a 133 MHz Pentium PC. To put things in perspective, a full-duplex 16 kb/s G.728 LD-CELP coder or a full-duplex 8 kb/s G.729 CS-ACELP coder (both operate at half the 16 kHz sampling rate used by TPC) takes about 60% of the CPU on a 150 MHz SGI Indy.

6. CONCLUSION

We have developed two versions of a candidate coder for the ITU-T's new wideband speech coding standard. The coder is one of the two surviving candidates. It produces high speech quality with a low coder complexity.

ACKNOWLEDGMENT

I would like to thank David Kapilow for contributing to the C simulation code and Dongmei Wang for her design tools for KLT-based gain quantization. I would especially like to thank Cheng-Chieh Lee for his help in the preparation of this paper, for doing the fixed-point DSP implementation of TPC, and for taking over TPC algorithm refinement after I left AT&T.

References

[1] "High-level Description of AT&T's Mode B Candidate for the ITU-T Wideband (7 kHz) Speech Coding Standard," AT&T contribution to ITU-T SG15/Q6 meeting, Rome, Italy, February 1996.

[2] J.-H. Chen, "Low-Complexity Wideband Speech Coding", *Proc. IEEE Workshop on Speech Coding*, pp. 27-28, Annapolis, Maryland, September 1995.

[3] J.-H. Chen and D. Wang, "Transform Predictive Coding of Wideband Speech", *Proc. ICASSP-96*, pp. 275-278, Atlanta, Georgia, May 1996.

[4] J. P. Princen et al., "Subband/Transform Coding Using Filter Bank Designs Based on Time Domain Aliasing Cancellation," *Proc. ICASSP-87*, pp. 2161-2164, Dallas, Texas, April 1987.

[5] "Detailed Description of AT&T's Mode B Candidate for the ITU-T Wideband (7 kHz) Speech Coding Standard," AT&T contribution to the ITU-T SG15/Q6 meeting, Baltimore, Maryland, October, 1996.

[6] R. Lefebvre, et al., "High Quality Coding of Wideband Audio Signals Using Transform Coded Excitation (TCX) Coder," *Proc. ICASSP-94*, pp. 193-196, Adelaide, Australia, April 1994.

[7] K. K. Paliwal and B. S. Atal, "Efficient Vector Quantization of LPC Parameters at 24 bits/frame", *Proc. ICASSP-91*, pp. 661-664, Toronto, Canada, May 1991.

[8] E. Ordentlich and Y. Shoham, "Low-Delay Code-Excited Linear-Predictive Coding of Wideband Speech at 32 kbps," *Proc. ICASSP-91*, pp. 9-12, Toronto, Canada, May 1991.

[9] M. R. Schroeder, et al., "Optimizing Digital Speech Coders by Exploiting Masking Properties of the Human Ear," *J. Acoust. Soc. Amer.*, pp. 1647-1652, December 1979.

[10] A. Sugiyama, et al., "Adaptive Transform Coding with an Adaptive Block Size ATC-ABS," *Proc. ICASSP-90*, pp. 1093-1096, Albuquerque, New Mexico, April 1990.

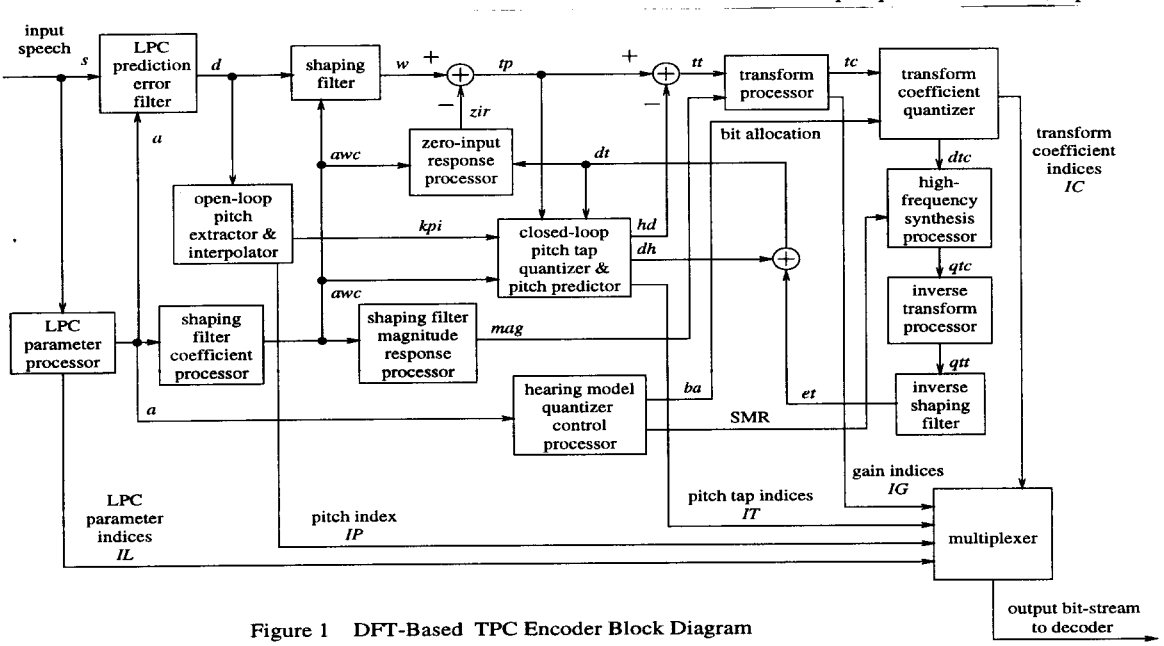

Figure 1 DFT-Based TPC Encoder Block Diagram

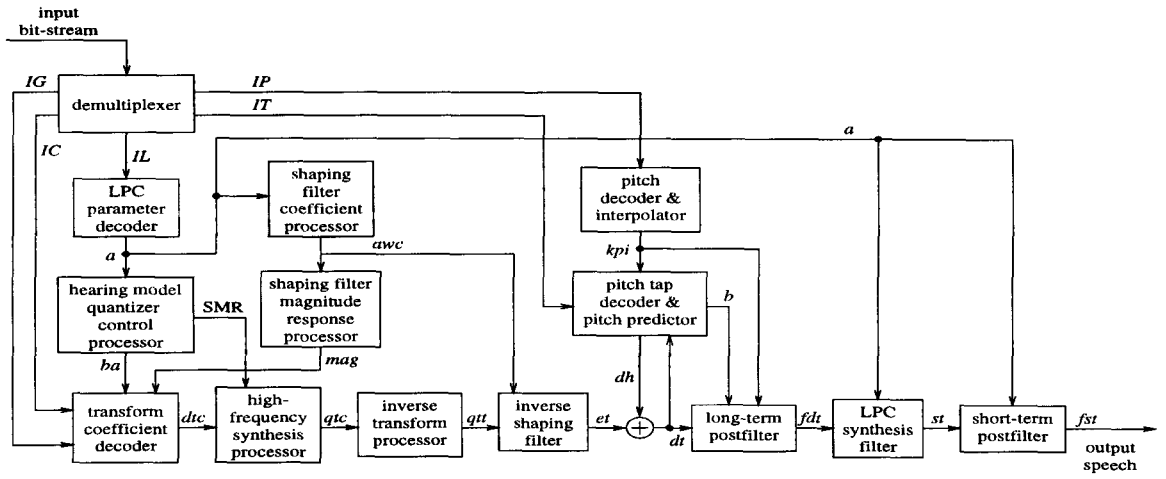

Figure 2 DFT-Based TPC Decoder Block Diagram

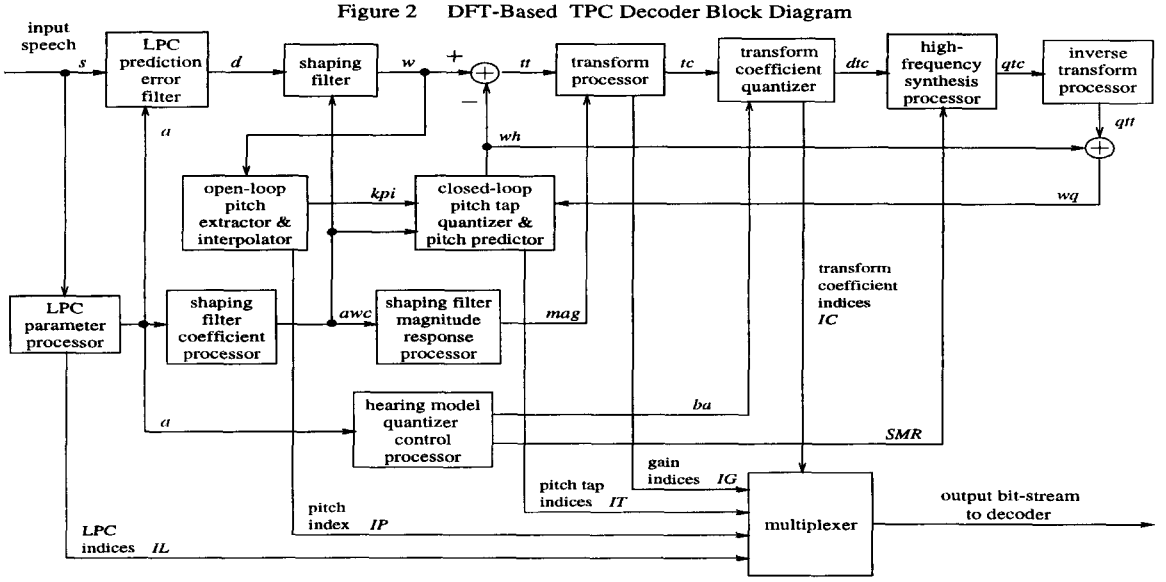

Figure 3 MDCT-Based TPC Encoder Block Diagram

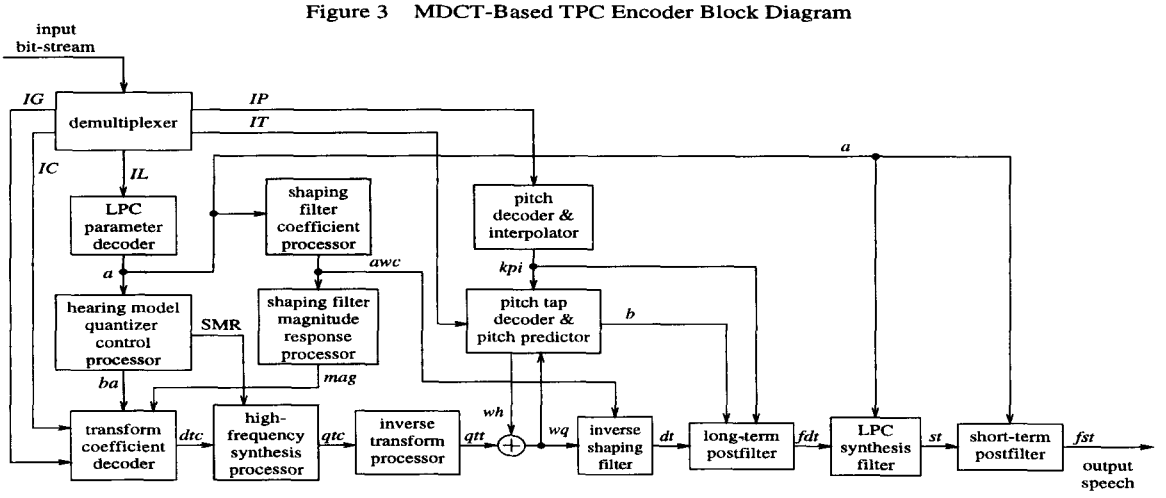

Figure 4 MDCT-Based TPC Decoder Block Diagram