AN ALGORITHM FOR COMPRESSION OF WIDEBAND DIVERSE SPEECH AND AUDIO SIGNALS

Trevor R. Trinkaus and Mark A. Clements

Center for Signal and Image Processing
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0250

ABSTRACT

A compression scheme for diverse speech and audio signals is proposed. In this scheme, signals are analyzed with a 2-band QMF filter bank followed by the application of a Modulated Lapped Biorthogonal Transform (MLBT) to each of the filter bank channels. Subsequent encoding of transform coefficients is performed using Laplacian optimized scalar and vector quantizers, whose rates are determined by an estimated noise threshold, i.e., masking threshold. Listening tests show that the coder achieves a quality at 32 Kbits/s that is preferred over the ITU G.722 coder at 64 Kbits/s, for speech, music, and more diverse signals consisting of speech in the presence of eventful background sounds. Both the delay of the coder, at 40 ms, and the level of complexity are moderate.

1. INTRODUCTION

Compression algorithms designed specifically for speech or audio signals, such as music, have been successfully utilized in application areas such as telecommunications, digital broadcasting, and storage. In many instances, however, the algorithms were designed for a particular input signal or application, or have not met quality expectations when applied to a broader class of input signals. Until recently [1], [2], algorithms designed for both speech and other, more diverse, audio signals have not received considerable attention. Recent progress in this area, however, has shown that increased quality levels at low bit rates (16 Kbits/s) could only be achieved at the expense of higher algorithmic delay, or complexity, or a compromised quality for more diverse signals.

In this paper, a coding scheme is presented which produces satisfactory quality over a broad class of input signals, with moderate delay and complexity. This scheme makes use of subband and transform coding techniques, and uses both scalar and vector quantization methods. In addition, a masking threshold is used for effective spectral shaping of quantization noise. The algorithm operates on 16 kHz sampled signals, and uses a frame size of 32 ms, or 512 samples. The method of signal analysis combines a 2-band filter bank with a lapped transform. These analysis stages produce respective delays of approximately 8 ms and 32 ms, which gives an overall delay of 40 ms. The complexity of the algorithm, which is largely due to vector quantization and codebook searching routines, is kept down by restricting the maximum vector dimension and codebook sizes to within reasonable limits. It is shown that the reconstructed signal quality produced with this algorithm at a bit rate of 32 Kbits/s, or 2 bits/sample, is significantly preferred over the quality produced with the G.722 algorithm at 64 Kbits/s, for a variety of input signals.

2. CODING ALGORITHM

The algorithm may be described as a hybrid subband-transform coding scheme which employs perceptual masking properties to achieve an efficient reduction in bit rate. A block diagram of the coder is shown in Figure 1. As seen in the

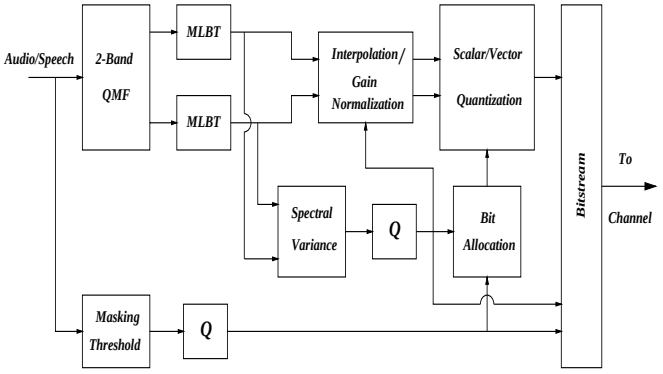

Figure 1: Block Diagram of Encoder

figure, each input frame is first split into its low and high subbands using a Quadrature Mirror Filter (QMF) bank. Each of the subband signals is then converted to a set of frequency domain coefficients, or transform coefficients, by applying a Modulated Lapped Transform (MLT) which contains an analysis window which differs from its synthesis window. In such a case, the transform is referred to as a Modulated Lapped Biorthogonal Transform (MLBT), and is described in more detail in the next section. The transform coefficients are encoded using combined scalar-vector quantization techniques similar to the techniques used in [3] and [4]. Here, however, the quantizers are optimized for Laplacian distributed data. A rate-distortion characteristic for each quantizer is determined and used in sub-

sequent bit allocation operations. Prior to quantization, transform coefficients are partitioned into non-uniform frequency bands which correspond to the ear's critical bands, and are then normalized by their estimated standard deviation. The standard deviation estimates are computed by geometrically interpolating the transform coefficient variance values in each critical band [5], [6]. These spectral variance values are quantized and transmitted to the decoder as side information.

The resolution of each critical band quantizer is determined by best matching its corresponding distortion, scaled by the spectral variance, to the minimum of an estimated allowable noise threshold, commonly known as the masking threshold. The masking threshold determines, as a function of frequency, the maximum value that quantization distortion levels may reach before becoming audible. In this algorithm, the masking threshold is computed according to the procedure described in [7]. In situations where the number of quantization bits needed to meet the masking threshold requirement exceeds the number of available bits, a bit pruning procedure is applied. This procedure basically involves locating the minimum absolute difference between a quantizer's scaled distortion level and the masking threshold in each critical band, and reducing that critical band's current bit assignment by one bit. This operation is repeated until the bit total reaches the acceptable limit. The final bit allocation among critical bands must be reproduced at the decoder, so it is, therefore, necessary to quantize and transmit the minimum levels of the masking threshold in each critical band. The overall bitstream consists of quantizer or codebook indices representing the transform coefficients, in addition to the indices used to represent quantized spectral variance values and masking threshold levels.

3. SIGNAL ANALYSIS AND REPRESENTATION

The method of signal analysis consists of two stages. The input 512-sample data frame is first split into its low and high frequency subbands using a 2-band QMF bank. Each subband signal is then converted to a set of frequency domain coefficients by applying an MLBT. The MLBT offers, for some signals, a slight improvement in perceptual quality, possibly due to the increased frequency selectivity of the synthesis basis functions [8].

3.1. Filter Bank

Since it is assumed that the input signal may contain speech, music, or possibly speech in the presence of important background sounds, a filter bank is applied to split the signal into its two primary frequency bands. A 2-band quadrature mirror filter bank is used for this purpose. The impulse response of each filter contains 128 coefficients, which, as mentioned previously, adds approximately 8 ms to the overall delay. A plot of the magnitude of the frequency response of the filter bank is shown in Figure 2. As mentioned previously, each input frame contains 512 samples, therefore, each subband signal will contain 256 samples after filtering. A transform of size 256 is, therefore, used in the subsequent analysis of each subband signal.

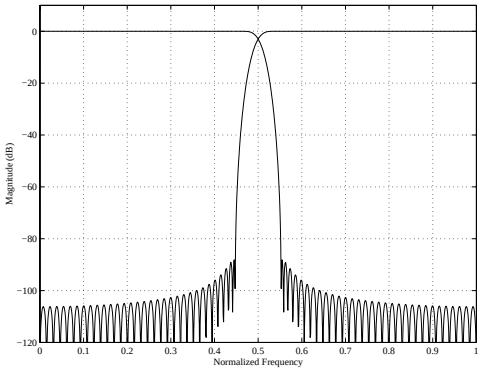

Figure 2: QMF Bank Frequency Response Magnitude

3.2. Lapped Transform

The lapped transform used in this coder is very similar to the Modified Discrete Cosine Transform (MDCT) [9], or Modulated Lapped Transform (MLT) [10]; however its analysis basis functions of the forward transform differ from its synthesis basis functions of the inverse transform. The transform, therefore, may be referred to as the Modulated Lapped Biorthogonal Transform (MLBT), as described by Malvar in [8]. The development of the MLBT was initiated by Smart and Bradley in [11] where the conditions for perfect reconstruction for an oddly stacked Time Domain Aliasing Cancellation (TDAC) filter bank with non-identical analysis and synthesis filters were derived. These conditions were later rederived in [12] in the context of lapped transforms, and were formulated in terms of the analysis and synthesis windows used in the forward and inverse transforms, respectively. As stated in [12], the length M forward transform may be expressed as the $2M \times M$ matrix, $\mathbf{P}$, whose elements are given by

$$p_{nk} = h_a(n)\sqrt{\frac{2}{M}}\cos\left[\left(n+\frac{M+1}{2}\right)\left(k+\frac{1}{2}\right)\frac{\pi}{M}\right], \quad (1)$$

where $h_a(n)$ is the analysis window of length $2M$. Similarly, the inverse transform matrix, denoted $\mathbf{Q}$, is given by

$$q_{nk} = h_s(n)\sqrt{\frac{2}{M}}\cos\left[\left(n+\frac{M+1}{2}\right)\left(k+\frac{1}{2}\right)\frac{\pi}{M}\right], \quad (2)$$

where $h_s(n)$ is the synthesis window. Since the window length is exactly twice the transform length, only adjacent frames overlap. Assuming the synthesis window is symmetric, the conditions which guarantee perfect reconstruction, as described in [12] and [8] are given by

$$h_a(n) = \frac{h_s(n)}{h_s^2(n)+h_s^2(n+M)}, \quad (3)$$

$$h_a(n) = h_a(2M-1-n), \quad n = 0, 1, \ldots, M-1. \quad (4)$$

A number of choices for the synthesis window, $h_s(n)$, are specified in [8] as a single parameterized equation given by

$$h_s(n) = \frac{1-\cos\left[\left(\frac{n+1}{M}\right)^{\alpha}\pi\right]+\beta}{2+\beta}. \quad (5)$$

In this coder, the choices $\alpha = 1$ and $\beta = 0$ were sufficient. In this case, the synthesis window reduces to

$$h_s(n) = \frac{1}{2} - \frac{1}{2}\cos\left[(n+1)\,\frac{\pi}{M}\right]. \qquad (6)$$

Plots of the analysis and synthesis windows for $M = 256$ are shown in Figure 3.

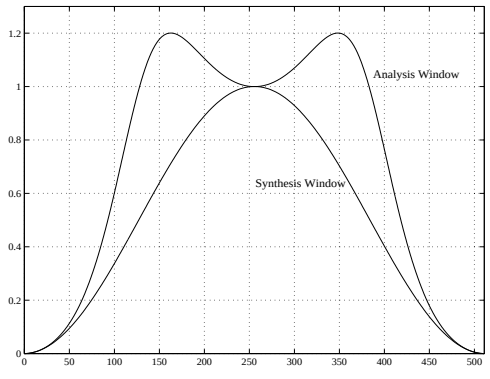

Figure 3: Analysis and Synthesis Windows

4. QUANTIZATION AND BIT ALLOCATION

A combined scalar-vector quantization technique is used to encode the MLBT coefficients. Scalar quantization is used for coefficients lying in the first five critical bands (0-510 Hz), whereas vector quantization is used for coefficients which reside in the remaining sixteen critical bands (510-7700 Hz). The design of the quantizers is based on the assumption that the distribution of the normalized MLBT coefficients in each critical band is approximately Laplacian. Analysis of the long-term critical band distributions for some speech and music inputs has verified this assumption.

Prior to quantization, the MLBT coefficients are partitioned into frequency bands corresponding to the 21 critical bands which span the 8 kHz bandwidth. Since there are 256 samples in each subband, and therefore 256 MLBT coefficients after transformation, there are a total of 512 coefficients that need to be partitioned. The coefficients in the frequency ranges below 50 Hz and above 7700 Hz could be discarded without significantly affecting the overall subjective quality of the output. Following critical band partitioning, the coefficients in the upper sixteen critical bands are blocked into vectors, with dimensions ranging from 2 to 4. Table 1 summarizes the coefficient partitioning. The final choice of quantization mode for each critical band was based on the observed subjective quality of the coded output.

Critical Bands	No. Coefficients	Vector Dim.	No. Vectors
1 - 5	31	1	31
6 - 14	114	2	57
15 - 18	135	3	45
19 - 21	208	4	52

Table 1: Critical Band Coefficient Partitioning

4.1. Scalar and Vector Quantization

The scalar quantizers used in this coder are optimized for Laplacian distributed data. The quantizers were designed using the Lloyd II algorithm described in [13], with a mean-squared error criterion and sizes ranging from 2 output levels to a maximum of 256 output levels (1-8 bits). The resulting distortion for each quantizer was recorded for later use in the bit allocation scheme.

The vector quantizers were designed using the LBG algorithm with the splitting initialization technique, and a long, unit variance Laplacian training set. Codebooks were designed for vectors of 2, 3, and 4 dimensions, using a mean-squared error criterion. The maximum size of the codebooks for 2 and 3-dimensional vectors was restricted to 128, and for 4-dimensional vectors it was restricted to 64. A randomly generated training set of approximately 3.8 million samples was used for the designs. A training set of such size allowed approximately 10,000 training vectors per codevector to be used for the design of the largest sized codebook. Distortion values for each codebook were, again, recorded for use in the bit allocation scheme. Since the codebooks are moderately sized, a simple nearest neighbor encoding method was used to quantize the MLBT coefficient vectors.

4.2. Side Information

The side information, which consisted of critical band masking threshold levels and spectral variance values, was also quantized using vector quantization (VQ). The masking threshold levels in each of the 21 critical bands were quantized as a single vector, whereas the spectral variance values were first split into subvectors, and each subvector was quantized separately in a log domain. This so-called Split VQ technique reduces complexity and allows for more accurate quantization. This technique was previously applied in [14] to quantize LPC vectors consisting of line spectral frequencies. VQ codebooks were, again, designed using the LBG algorithm for training data computed over approximately 75,000 frames, or 40 minutes of source data at 16 kHz. The source data consisted of speech, and vocal and non-vocal music. It was observed that the masking threshold vector could be quantized with an MSE of 8.3 dB averaged over 14,000 frames of test data. This amount of error, however, did not significantly affect critical band bit assignments reproduced by the decoder. The error due to quantization of the spectral variance subvectors, for the same set of test frames, resulted in an average MSE of -0.6 dB for the set of 21 spectral variance values. In this case, the quantization error was more severe and did produce perceptible differences when compared to results using unquantized spectral variance values.

4.3. Bit Allocation

The method of bit allocation is basically a bit pruning procedure. Critical band bit assignments are initially determined by locating the minimum of the masking threshold in each band and matching this value, on a linear scale, to a quantizer's scaled distortion. The distortion value used depends, of course, on the quantization mode of the critical band. The scaling factor is the variance of the MLBT coeffi-

Signal	Preference
Male speech	37/48
Female speech	34/48
Orchestra (horns, violins)	14/16
Acoustic guitar	13/16
Violin	11/16
Organ, triangle	12/16
Female opera	13/16
Orchestra (clarinets)	14/16
Piano	13/16
Pop vocal	15/16
Male speech + background	25/32
Female speech + background	17/32

Table 2: Preference of Coder Over G.722 Coder

cients in that band. Bits are pruned when a situation arises where the number of bits needed to meet this requirement exceeds the number of available bits. The rule for pruning critical band bit assignments is relatively straightforward. The absolute difference between the minimum masking threshold level and the scaled quantization distortion in each critical band is first computed. The critical band whose difference is smallest has its initial bit assignment reduced by one bit. This process is repeated until the bit total reaches, or falls below, a fixed limit. It has been observed that, on average, less than 50 iterations are required to accomplish this bit pruning task. It is also worth noting that the overall subjective quality, for most inputs, could be improved by restricting bit pruning to take place only in the lower 17 critical bands (0-3700 Hz). This restriction was applied for the encoding of all items used in the listening tests.

5. PERFORMANCE

Informal listening tests were used to assess the performance of the coder. The format of the tests was an A/B comparison between the output of the coder at 32 Kbits/s and the output of the G.722 coder at 64 Kbits/s. Subjects were presented the test segments in a random order and were asked to indicate their preference. The test segments consisted of clean male and female speech, music, and speech in the presence of a background cocktail party. There were a total of 18 test segments used. All 18 test segments were presented to 16 different subjects. Table 2 indicates the preference of the coder over the G.722 coder for all 16 subjects. Results for more general inputs were combined.

6. CONCLUSION

A coder for diverse speech and audio signals was presented. The coder is capable of achieving 8:1 compression, while maintaining good quality for speech, music, and other diverse speech signals. Listening test results show that operating at a bit rate of 32 Kbits/s, the coder outperforms the G.722 coder at 64 Kbits/s over a broad class of test signals. The performance is achievable at a moderate delay (40 ms) and complexity level.

7. REFERENCES

[1] T. Moriya, N. Iwakami, A. Jin, K. Ikeda, and S. Miki, "A design of transform coder for both speech and audio signals at 1 bit/sample," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, (Munich, Germany), pp. 1371–1374, April 1997.

[2] S. A. Ramprashad, "A two stage hybrid embedded speech/audio coding structure," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, (Seattle, WA), pp. 337–340, May 1998.

[3] W.-Y. Chan and A. Gersho, "High fidelity audio transform coding with vector quantization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, (Albuquerque, NM), pp. 1109–1112, April 1990.

[4] S. Boland and M. Deriche, "Audio coding using the wavelet packet transform and a combined scalar-vector quantization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, (Atlanta, GA), pp. 1041–1044, May 1996.

[5] R. Zelinski and P. Noll, "Adaptive transform coding of speech signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 299–309, August 1977.

[6] J. M. Tribolet and R. E. Crochiere, "Frequency domain coding of speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 512–530, October 1979.

[7] ISO/IEC JTC1/SC29/WG11 MPEG, IS11172-3, *Information Technology - Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1.5 Mbit/s, Part 3: Audio*, 1992.

[8] H. S. Malvar, "Biorthogonal and nonuniform lapped transforms for transform coding with reduced blocking and ringing artifacts," *IEEE Trans. Signal Processing*, vol. 46, pp. 1043–1053, April 1998.

[9] J. P. Princen and A. B. Bradley, "Analysis/synthesis filter bank design based on time domain aliasing cancellation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 1153–1161, October 1986.

[10] H. S. Malvar, *Signal Processing with Lapped Transforms*. Norwood, MA: Artech House, 1992.

[11] G. Smart and A. B. Bradley, "Filter bank design based on time domain aliasing cancellation with non-identical windows," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, (Adelaide, South Australia), pp. 185–188, April 1994.

[12] S. Cheung and J. S. Lim, "Incorporation of biorthogonality into lapped transforms for audio compression," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, (Detroit, MI), pp. 3079–3082, May 1995.

[13] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Norwell, MA: Kluwer Academic, 1992.

[14] K. K. Paliwal and B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," *IEEE Trans. Speech, Audio Processing*, vol. 1, pp. 3–14, January 1993.