SPEECH COMPRESSION WITH COSINE AND WAVELET PACKET NEAR-BEST BASES

Carl Taswell[*]

ABSTRACT

Compression of speech from the TIMIT corpus was investigated for several transform domain methods coding near-best and best bases from cosine and wavelet packet transforms. Satisficing (suboptimizing) search algorithms for selecting near-best bases were compared with optimizing algorithms for best bases in these adaptive tree-structured transforms. Experiments were performed on several hundred seconds of speech spoken by both male and female speakers from all dialect regions of the TIMIT corpus. Near-best bases provided rate-distortion performance effectively as good as that of best bases but without the additional computational penalty. Cosine packet bases outperformed wavelet packet bases.

1. INTRODUCTION

Satisficing search algorithms have been developed [1, 2, 3, 4] for adaptively selecting near-best basis and near-best frame decompositions in redundant tree-structured wavelet transforms. Any of a variety of additive or non-additive information cost functions can be used as the decision criterion for comparing and selecting nodes when searching through the tree. The algorithms are applicable to tree-structured transforms generated by any kind of wavelet whether orthogonal, biorthogonal, or non-orthogonal. These satisficing search algorithms implement sub-optimizing rather than optimizing principles, and acquire the important advantage of reduced computational complexity with significant savings in memory, flops, and time. Despite the sub-optimal approach, top-down tree-search algorithms with additive or non-additive costs that yield near-best bases can be considered in many practical situations better than bottom-up tree-search algorithms with additive costs that yield best bases. Here "better than" means that effectively the same level of performance can be attained for a fraction of the computational work. These principles and methods are demonstrated here in this report on their application to the compression of speech from the TIMIT corpus [5].

2. METHODS

A complete exposition of definitions, notation, and algorithms for adaptive tree-structured wavelet transforms, information cost functions, and basis selection methods for near-best and best bases can be found elsewhere [1, 2, 3, 4]. Here in this report, only methods specific to experiments investigating lossy compression of speech will be detailed. We wish to minimize the distortion D resulting between the reconstructed estimate $\hat{\mathbf{x}}$ and the original signal $\mathbf{x}$ following compression and coding of the transform domain packet coefficients. Compression can be achieved by thresholding the N packets of the transform to the fixed compression rate of $M < N$ largest absolute-value packets and then quantizing and coding the remaining M packets. As explained in Section 2.2., other experimental paradigms can also be investigated.

2.1. Quantization Coder

A uniform mid-tread quantizer with adaptive feed-forward gain control [6] was modified to include some of the features of the wavelet scalar quantizer (WSQ) characteristic of Bradley and Brislawn [7]. Using their notation, let Z_k be the bin width of the zero bin for the k^{th} subband, and Q_k be the uniform bin width of all other bins for the k^{th} subband. Their WSQ characteristic was simplified to the case of Q_k and Z_k constant for all subbands k, thus enabling use of the same values of Q and Z for all N transform coefficients considered together as one collection instead of as separate subbands. These bin widths Q and Z were adaptively computed as a function of the maximum absolute value amplitude A, the bit rate parameter β in bits per quantized coefficient, and a thresholding parameter α as a fractional multiplier. Thus,

$$\begin{aligned} A &= \max_n |a(n)| \\ Z &= 2\alpha A \\ Q &= (1-\alpha)A/(2^{\beta-1}-1+C) \end{aligned}$$

provided a quantizer characteristic that thresheld values smaller than the fraction α of the maximum A and appropriately "centered" the maximum A in its bin so as to minimize its error. In this specification of the quantizer, the important parameters are α and β, of which α determines the resulting number M of surviving non-zero transform coefficients in each segment of length N, and β determines the precision of the M surviving coefficients. Alternatively, M can be used as the parameter which determines the zero bin width Z either directly as

$$Z = 2|a(n_M)| - \epsilon$$

or indirectly via the fractional multiplier

$$\alpha = |a(n_M)/a(n_1)| - \epsilon = |a(n_M)|/A - \epsilon$$

where the index n_i identifies the i^{th} largest of the coefficients $\{|a(n)| : 1 \le n \le N\}$ sorted in decreasing absolute value order, and ϵ is a tolerance taken as a small multiple of machine precision.

[*]C. Taswell is with Scientific Computing and Computational Mathematics, Stanford University Department of Computer Science, Stanford, CA 94305-2140. Internet: taswell@sccm.stanford.edu; Tel: 415-723-4101; Fax: 415-723-2411.

2.2. Rate-Distortion Curves

Speech signals (each an entire spoken sentence of several seconds duration sampled at 16 kHz as 12 bit integers) were scaled to zero-mean unit-variance signals in floating point format and then segmented with a frame length of $N = 512$ samples. There were approximately $O(10^2)$ segments $\mathbf{x}$ per spoken sentence, with each segment containing sampled data from a time interval of 32 milliseconds of speech. Defining the peak signal data value X in each segment as $X = \max_n |x(n)|$, the peak signal-to-noise ratio (PSNR) distortion measure was computed in each segment as

$$\mathrm{PSNR} = 10 \log_{10} \frac{NX^2}{\|\mathbf{x} - \hat{\mathbf{x}}\|^2}$$

in decibels. These distortion measures were computed for each segment of all signals in an experiment, averaged over all segments from all signals, and reported as the mean segmental values with standard errors of the means (SEM). Rate-distortion curves were then plotted as the mean segmental PSNR versus the mean segmental number M of transform coefficients that were quantized in the non-zero bins of width Q.

The number M of surviving non-zero coefficients was determined by the quantization coder as a function of the thresholding parameter α. For the comparison of different transforms, several kinds of experiments were performed. In the first kind called a "fixed α" experiment, the value of α was held fixed at the same constant value for all transforms and all segments with resulting variable M different for each transform and each segment. In the second kind called a "fixed M" experiment, the value of M was held fixed at the same constant value for all transforms and all segments with resulting variable α different for each transform and each segment. In the third kind called a "fixed f" experiment, the value of M was used to determine α as in the fixed M experiment, but instead of holding M fixed, M itself was determined by the data compression number function $\mathcal{N}_f^p$ [1] with $p = 2$ and the fraction f held fixed for all transforms and all segments. Thus, in the fixed f experiment, the energy of the surviving non-zero transform coefficients was constant for all transforms and segments. Experiments were performed comparing operational rate-distortion curves for adaptive pulse code modulation (APCM), the discrete wavelet transform (DWT), the wavelet packet transform (WPT), and the cosine packet transform (CPT).

3. RESULTS

A fixed M experiment was performed on a total of 6968 segments from 80 sentences of type `sx` from 16 different TIMIT speakers consisting of one male and one female speaker from each of the eight dialect regions. Figure 1 displays results from this experiment with $\beta = 12$ and $M = 57$, 161, 266, 357, 425. Both CPDD and WPDD were computed as the near-best basis Decompositions selected by top-Down search through the CPT and WPT, respectively. They both outperformed the DWT, with the CPDD also surpassing the WPDD, by several dB higher PSNR with less distortion for given rates of compression measured by the number M of surviving coefficients. Analogous results were obtained from fixed α and fixed f experiments.

Table 1 lists interpolated values of M for given PSNR in an experiment comparing top-Down near-best and bottom-Up best basis decompositions. Using the selection criterion $\mathcal{C} = \mathcal{G}$, WPDU outperformed WPDD at all compression rates; however, CPDD actually outperformed CPDU at high compression rates. This superior performance for a top-down near-best basis over a bottom-up best basis occurred at high rates of compression and distortion (low values of M and PSNR) with the cost $\mathcal{G}$ known to perform better at low rates (high values of M and PSNR). In contrast, CPDU did outperform CPDD at all compression rates with the cost $\mathcal{E}^1$ known to perform well at all compression rates.

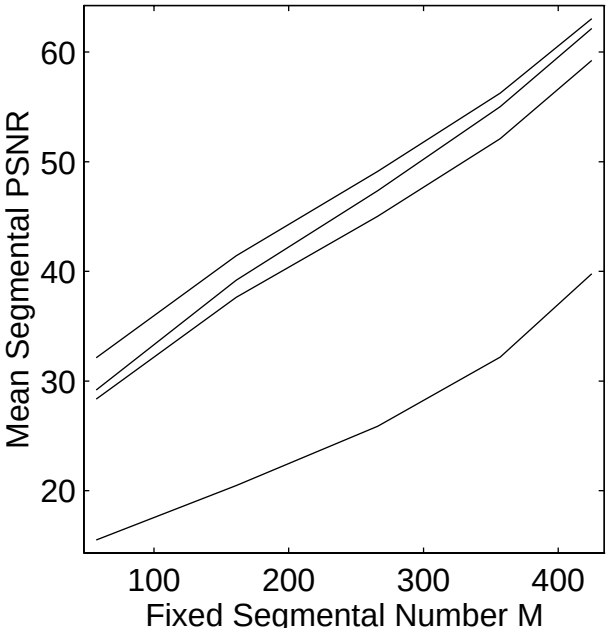

Figure 1. Mean segmental PSNR versus M for a fixed M experiment. Curves, all with $\beta = 12$, are ordered from upper left to lower right as CPDD, WPDD, DWT, and APCM. Points on curves have the same abscissa: α was determined from fixed M with values $M = 57, 161, 266, 357, 425$.

Since SEM values were typically 0.1 dB for the PSNR values, the differences in results between search methods $\mathcal{S}$ and cost functions $\mathcal{C}$ were not as dramatic as the differences between transforms. For example, at $M = 64$, PSNR was 33.9 and 34.2 for CPD($\mathcal{D}, \mathcal{E}^1$) and CPD($\mathcal{U}, \mathcal{E}^1$) compared to 29.1 for DWT; while at PSNR = 30, M was 35.5 and 33.9 for CPD($\mathcal{D}, \mathcal{E}^1$) and CPD($\mathcal{U}, \mathcal{E}^1$) compared to 70.8 for DWT. Thus the additional improvement provided by the CPD with bottom-up best basis relative to the CPD with top-down near-best basis was negligible in comparison to the improve-

Table 1. Interpolated M for Given PSNR in a Fixed M Experiment with Search $\mathcal{S} = \mathcal{D}$ Top-Down or $\mathcal{S} = \mathcal{U}$ Bottom-Up.

	WPD($\mathcal{S}, \mathcal{G}$)		CPD($\mathcal{S}, \mathcal{G}$)		CPD($\mathcal{S}, \mathcal{E}^1$)	
PSNR	$\mathcal{D}$	$\mathcal{U}$	$\mathcal{D}$	$\mathcal{U}$	$\mathcal{D}$	$\mathcal{U}$
30	63.4	53.0	40.8	43.8	35.5	33.9
35	111.5	97.8	84.2	88.0	74.1	70.8
40	171.1	154.6	142.5	144.3	135.4	129.5
45	235.7	217.4	209.2	207.5	209.1	200.7
50	299.1	280.3	277.7	272.3	285.2	275.2
55	356.9	338.7	342.1	333.8	354.8	344.4
60	406.6	390.5	397.5	388.5	410.3	401.9

ment already attained by the CPD over the DWT. Moreover, the mean segmental value of the mean tree depth level $\hat{L} = 1.94$ for $\mathrm{CPD}(\mathcal{D}, \mathcal{E}^1)$ was significantly smaller than the fixed value of $L = 6$ required for $\mathrm{CPD}(\mathcal{U}, \mathcal{E}^1)$. Thus, the small increase in compression rates provided by the bottom-up best basis relative to the top-down near-best basis did not justify the large increase in computational complexity required to obtain that improvement in performance.

4. DISCUSSION

To select a basis adaptively within a redundant tree-structured wavelet transform, it is necessary to specify a search path through the tree and a decision criterion by which to compare and select branches of the tree. It is then possible to implement appropriate algorithms incorporating the desired search path and selection criterion [4]. Experiments described in this report demonstrated that the choice of information cost function used as selection criterion for finding a basis decomposition may have a significant impact on data compression especially when considering different design constraints (such as high or low rates of compression with corresponding high or low levels of distortion). Comparing the various cost functions investigated, the ℓ^p and $\ln \ell^2$ functionals (both additive costs) and the data compression area (a non-additive cost) provided the best performance in general. However, the choice of search path had an even greater impact on performance than did the choice of cost function. Thus, the sub-optimal top-down tree search, instead of optimal bottom-up tree search, significantly increased the efficiency of computation without proportionately decreasing the efficiency of compression of signals from a speech database.

These results support the validity of the use of satisficing principles and methods [4]. This general conclusion derived from the speech compression experiments was based on comparisons of the mean segmental PSNR value as distortion measure and either the fixed or mean segmental number M of quantized non-zero coefficients as rate measure. Since DeVore et al [8] demonstrated an empirical relationship between the number of coefficients and the number of bytes of compressed data in the context of image compression, use of this number has become a common rate measure for compression studies in the wavelet literature. However, for the development of an actual speech coder, entropy coding in addition to quantization coding, actual bit rates rather than coefficient counts, and psychoacoustic distortion measures rather than PSNR should all be investigated [9, 10, 11].

Nevertheless, the use of the rate-distortion measures studied in this report does not invalidate the following key results and conclusions obtained with regard to data compression. First, for speech compression in particular, the CPT (which incorporates a DCT) performed significantly better than the WPT and the DWT. This result reconfirms from another perspective the long established superiority of the discrete cosine transform (relative to other transforms) due to its close fit to the optimal Karhunen-Loeve transform [12]. Second, for data compression in general (as inferred from the results on speech compression reported herein), the choice of a decision criterion (information cost function) and basis selection method (tree search path) may very well impact performance and should be considered when designing a data compression method intended for application to a particular class of data and type of compression. For example, a decision criterion and tree search path appropriate for best performance at high rates of compression and distortion may not be appropriate at low rates of compression and distortion, and vice versa. Third, despite the caveat of the second conclusion, sub-optimizing near-best bases can be considered "good enough" relative to optimizing best bases in most practical situations wherein computational complexity must also be considered, especially in real-time signal processing applications.

REFERENCES

[1] C. Taswell, "Near-best basis selection algorithms with non-additive information cost functions," in *Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis* (M. G. Amin, ed.), (Philadelphia, PA), pp. 13–16, IEEE Press 94TH8007, Oct. 1994.

[2] C. Taswell, "Top-down and bottom-up tree search algorithms for selecting bases in wavelet packet transforms," in *Wavelets and Statistics* (A. Antoniadis and G. Oppenheim, eds.), Lecture Notes in Statistics, Springer Verlag, 1995. Proceedings of the Villard de Lans Conference November 1994.

[3] C. Taswell, "Image compression by parameterized-model coding of wavelet packet near-best bases," in *SPIE Conference on Wavelet Applications* (H. Szu, ed.), pp. 153–161, SPIE Press, Apr. 1995.

[4] C. Taswell, "Satisficing search algorithms for selecting near-best bases in adaptive tree-structured wavelet transforms," tech. rep., Scientific Computing and Computational Mathematics, Stanford University, Stanford, CA, June 1995. submitted for publication.

[5] DARPA, *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Gaithersburg, MD: National Institute of Standards and Technology, Oct. 1990. NIST Speech Disc 1-1.1.

[6] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Prentice Hall Signal Processing Series, Englewood Cliffs, NJ: P T R Prentice-Hall, Inc., 1978.

[7] J. N. Bradley and C. M. Brislawn, "The FBI wavelet/scalar quantization standard for gray-scale fingerprint image compression," Technical Report LA-UR-93-1659, Los Alamos National Laboratory, Los Alamos, NM, 1993.

[8] R. A. DeVore, B. Jawerth, and B. J. Lucier, "Image compression through wavelet transform coding," *IEEE Transactions on Information Theory*, vol. 38, pp. 719–746, Mar. 1992.

[9] P. Noll, "Wideband speech and audio coding," *IEEE Communications Magazine*, pp. 34–44, Nov. 1993.

[10] L. R. Rabiner, "Applications of voice processing to telecommunications," *Proceedings of the IEEE*, vol. 82, pp. 199–228, Feb. 1994.

[11] A. Gersho, "Advances in speech and audio compression," *Proceedings of the IEEE*, vol. 82, pp. 900–918, June 1994.

[12] J. L. Flanagan, M. R. Schroeder, B. S. Atal, R. E. Crochiere, N. S. Jayant, and J. M. Tribolet, "Speech coding," *IEEE Transactions on Communications*, vol. 27, pp. 710–737, Apr. 1979.