FREQUENCY SELECTIVITY VIA THE SpEnt METHODOLOGY FOR WIDEBAND SPEECH COMPRESSION[†]

Mark G. Kokes and Jerry D. Gibson

Multimedia Communications and Networking Laboratory
Department of Electrical Engineering
Southern Methodist University
P.O. Box 750338, Dallas, TX 75275-0338
{kokes, gibson}@seas.smu.edu

ABSTRACT

In speech and audio coding, frequency selectivity of the basis functions is an important property of the codec. The more precise the frequency selectivity, the less chance there is for audible coding effects due to uncanceled aliasing. In this work, we use Campbell's coefficient rate and the spectral entropy (SpEnt) of the source random process as a guide to formulate adaptive nonuniform modulated lapped biorthogonal transforms (NMLBT). The use of the NMLBT allows for efficient implementation of a time-varying transform which possesses both good frequency and time resolution at all instances, without the need for transitional filters. By coupling the SpEnt methodology with modulated lapped biorthogonal transforms (MLBT), we develop band combining strategies to produce an adaptive NMLBT. Due to the nature of the SpEnt methodology, the new frequency selection process comprises a non-linear approximation method to determine the best n basis functions to represent the current speech frame. We implement a wideband speech compression scheme based on this strategy and verify its imprived performance in coding speech and audio signals at 16 and 24 kbps.

1. INTRODUCTION

In speech and audio coding applications, transform domain signal processing represents a popular approach to source compression. In particular, the most recent wideband speech compression standard, G.722.1 [1], is a transform-based algorithm. In typical block-based transform coding scenarios, a frame of speech samples is represented as a linear combination of a particular fixed set of transform basis functions, and the coefficients representing the contribution of each basis function to the speech frame are efficiently quantized and entropy coded in order to achieve data compression.

An important property of transform codecs is the frequency selectivity of the basis functions [2]. The more precise the frequency selectivity, the less chance there is for audible coding effects due to uncanceled aliasing. Unfortunately, as stated previously, typical block-based transform coding schemes often use a fixed set of basis functions and thus the frequency selectivity is non-adaptive. By adapting the transform basis functions, we can enhance the reproduced signal fidelity by tuning the frequency selectivity of the transform.

At present, the most successful methods for lossy source compression are sample-function adaptive codecs. Prominent examples of such compression schemes are variable rate speech and audio coders that allocate bits to parameters within a frame based upon the classification of the frame [3, 4]. These techniques can be classified as nonlinear approximation methods. Building on these ideas and related theories in computational harmonic analysis [5], we use Campbell's coefficient rate [6] and the spectral entropy (SpEnt) [7, 8, 9] of the source random process, as a mechanism to formulate adaptive nonuniform modulated lapped biorthogonal transforms.

In previous related work, Coifman and Wickerhauser [10] proposed the following best basis selection algorithm. For transform coefficients x_n, the theoretical dimension of a signal is defined as $d = \exp\left(-\sum_n p_n \log p_n\right)$ where $p_n = \frac{|x_n|^2}{\|x\|^2}$, and the exponent is the spectral entropy. They used d as a measure of the number of coefficients to be coded in a wavelet transform, and by minimizing d, the best wavelet packet basis could be chosen as the best transform basis since it produces the minimum number of coefficients. A similar method was extended to other transformations for speech processing in [11] using an adaptive windowing procedure. Such an algorithm causes increased overhead in that it switches between multiple pre-determined window sizes and thus adapts the transform's frequency selectivity by changing the blocksize over which it is calculated. In [2] and [12], Malvar explores band combining strategies using either quarter resolution or half resolution time domain basis functions within the NMLBT without reference to the theoretical dimension of a signal, the minimum coefficient rate, or the spectral entropy of the source random process.

In this paper, we couple the SpEnt methodology with modulated lapped biorthogonal transforms, for the purpose of developing an automated mechanism for determining frequency band combinations to produce a frame adaptive NMLBT. We use the SpEnt methodology as a means of suggesting both the number and location of which bands to

[†]This research was supported, in part, by NSF Grants NCR-9796255 and CCR-0093859.

combine. By adjusting the bandwidth of some of the original basis functions, we generate a procedure by which the frequency selectivity of the transform basis functions are dictated by the SpEnt result. In this way, we use a frame adaptive NMLBT to tune the frequency selectivity of the transform by shortening the duration of basis functions located at particular frequencies. Because of the nature of the SpEnt methodology, the resulting frequency selection process represents a non-linear approximation method for the purpose of selecting the best n basis functions for the current speech frame.

2. THE MLT, MLBT, AND NMLBT

Two types of artifacts are often observable in transform coding of speech signals at low bitrates: blocking (i.e. clicks) and ringing (i.e. reverberation and pre-echo). Blocking artifacts are generated by signal discontinuities at speech frame boundaries due to quantization effects. Ringing artifacts are generated when quantization errors in the transform coefficients cause reconstruction errors that last the duration of a reconstructed speech frame. Because of its inherent properties, the lapped transform (LT) can significantly reduce blocking artifacts in reconstructed speech segments [2]. Modulated lapped transforms (MLT) can reduce blocking effects even further [2]. In low bitrate speech and audio coding, ringing artifacts often occur during sounds which contain transient-like signals, such as plosive sounds or transitional speech segments. It is in the encoding of these signal types that most low bitrate transform codecs suffer the most severe performance degradations.

In developing the NMLBT, Malvar relaxes the constraint of identical analysis and synthesis windows. In [2], he suggests that if one uses a symmetric synthesis window and applies biorthogonality conditions, then a MLBT can be generated by using an analysis window, $h_a(n)$, which satisfies the generalized Princen-Bradley conditions [2], i.e.

$$h_a(n) = \frac{h_s(n)}{h_s^2(n) + h_s^2(n+M)}, \quad n = 0, 1, ..., M-1,$$
$$\text{and} \quad h_a(n) = h_a(2M-1-n). \tag{1}$$

By incorporating biorthogonality, Malvar suggests that the MLBT can be used to improve the frequency selectivity of the synthesis basis functions. In [2], as in this work, a synthesis window, $h_s(n)$, is chosen such that

$$h_s(n) = \frac{1 - cos[(\frac{n+1}{M})^{\alpha}\ \pi] + \beta}{2+\beta}, \tag{2}$$
$$n = 0, 1, ..., M-1,$$

where the parameter α controls the width of the window and β controls the end points of the window.

Since the MLBT is essentially free of blocking artifacts, we focus on the its use in reducing ringing artifacts in reconstructed speech. It is hypothesized that by generating shorter basis functions, we can increase the ability of a wideband speech compression algorithm to capture the transient nature of certain speech segments. One approach is to generate shorter high frequency basis functions by merging frequency subbands. In this way, we can equivalently create a non-uniform filter bank (i.e. NMLBT) in which some frequency subbands have larger bandwidths. It is suggested in [13, 14] that only the high frequency subbands be merged. In the following, we provide a motivation for combining adjacent subbands (including by not limited to the high frequency subbands). We show that by using the NMLBT as a basis for our algorithm, we can achieve minimal ringing artifacts in the reproduced speech frame.

3. CAMPBELL'S COEFFICIENT RATE

Our motivation for using spectral entropy as a basis for transform coder design is the work of Campbell on the coefficient rate of a random process [6]. The coefficient rate of a random process was first derived and defined by Campbell in 1960. Campbell considered the product of N sample functions of a random process, and showed, using an AEP[1]–like argument, that a Karhunen–Loeve expansion of this product could be separated into two sets – one set with average power very close to that of the product and the other set having very low average power. Asymptotically in the number of sample functions forming the product and in the support interval of the process, he showed that the average number of terms in the high power set approached a quantity that he interpreted as a coefficient rate given by

$$Q = \exp\left[-\int_{-\infty}^{\infty} S(f) \log S(f) df \right], \tag{3}$$

where we denote the quantity in the exponent as the spectral entropy. The implications of coefficient rate and spectral entropy for source compression were not explored by Campbell and no coding theorems were presented.

At this point, it is important to point out the relationship of coefficient rate and classical rate distortion theory results. For continuous time, bandlimited sources, classical rate distortion theory assumes Nyquist sampling and then proceeds to specify the minimum number of bits/sample (or bits/coefficient here) required to represent the source with a desired fidelity. On the other hand, the coefficient rate is not a source coding result at all. It is a statement about the minimum number of coefficients/second required to represent a random process in the sense of the number of terms needed to approximate the energy in the product of sample functions, as described earlier in this section. Interestingly, coefficient rate as specified by Campbell gives us an analytical indicator of the required minimum rate that was previously unavailable.

Yang and Gibson have recently provided two alternative derivations of Campbell's coefficient rate [7, 8]. One derivation tightens the connection between coefficient rate and the source bandwidth. This theorem is known as the equivalent bandwidth explanation. In this theorem, Yang and Gibson do not necessarily imply that the coefficient rate is the number of samples one should use to represent a random process. Instead, they propose that the sampling rate may still be the Nyquist rate, but the importance of the individual samples to the entirety of the source data frame may be different.

[1] AEP: Asymptotic Equipartition Property [15].

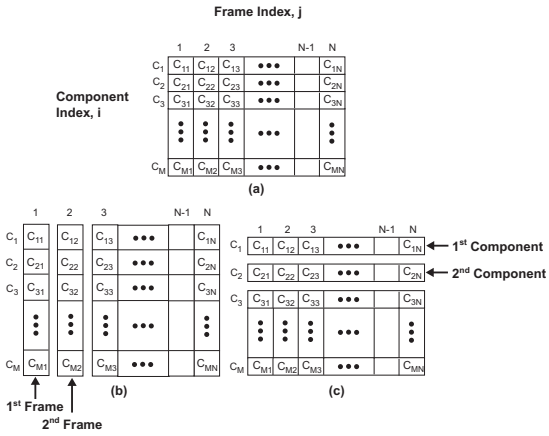

Fig. 1. Encoding Transform Coefficients. (a) Coefficients of N Sample Functions, (b) Encoding Sample Function by Sample Function, (c) Encoding Component by Component.

A second derivation by Yang and Gibson implies an explicit technique for adaptively coding a source sequence. In this derivation, known as the dominant terms expansion, the relationship

$$n_i = \frac{\lambda_i}{T} N, \qquad i = 1, 2, \ldots, M. \tag{4}$$

is derived. A graphical representation of this idea is illustrated in Fig. 1. In particular, this statement tells us that if we perform an orthogonal decomposition of a series of frames (i.e. 1 to N) of source data (Fig. 1a), and consider the coefficients corresponding to the i^{th} basis function in the series of frames as a coefficient sequence (Fig. 1c), then the number of coefficients, n_i, to be coded in this sequence (i.e. c_{x1} to c_{xN}) is proportional to the energy, λ_i, in the i^{th} component. The implied coder here "looks ahead" at N coefficient samples and thus employes a delay of N frames. Note that the SpEnt methodology is not a coding method, but it is a procedure for selecting which coefficients in the sequence need to be coded. Notice also that this is very different from the usual approach (Fig. 1b) of considering the orthogonal decomposition on a single frame and allocating bits to coefficients according to their relative energy in the current frame.

In this work, we demonstrate that Campbell's minimum coefficient rate can be used to suggest the number of basis functions as well as the bandwidth used to comprise these basis functions. The selection of bandwidths for each basis function in the decomposition will dictate the selection of the frequencies at which the best n basis functions are located.

4. ADAPTIVE NMLBT VIA SpEnt

In [2], $+1/-1$ butterflies are used in combination with the MLBT in order to achieve a NMLBT with perfect reconstruction capabilities (Fig. 2). In [12], an automatic switching mechanism based on frame energy levels and classification methods is used to turn on and off a subband combination matrix. In either experiment, the idea is to combine subbands in such a way that the reproduced speech or audio signal has reduced error spreading in frames which contain segments of transient signals. In this section, we will suggest another procedure for selecting how subbands can be combined in order to achieve the same goal.

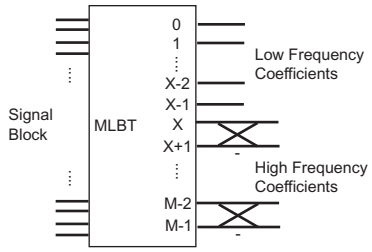

Fig. 2. NMLBT block diagram (Assuming R=2).

Assume that we take a similar approach to that in [2]. In this approach the first X basis functions are not modified, and each R-length set of the remaining $M - X$ functions are combined to generate a new set. At this time, let us define K as the sum of the number of unmodified basis functions (X) plus the number of R-length basis function sets. Assuming $R = 2$, if the original two basis functions were centered at frequencies $(f + 1/2)\pi/M$ and $(f + 3/2)\pi/M$, with bandwidths π/M, then when combined via $+1/-1$ butterflies, the result is two new basis functions that are both centered at $(f + 1)\pi/M$ with bandwidths $2\pi/M$, but have different time localization.

For the purpose of adaptive speech and audio coding, it can be inferred from above that the value of K can change on a frame-by-frame basis. During stationary frames, we could make $K = M$ which would turn the NMLBT into a length-M MLBT (as suggested in [2]). During transient sounds, we can choose K among a number of possible values for best reproduction of the block. The question becomes, is there a theoretical value of K for which the best reproduction can be obtained. In the following discussion, we suggest a possible way to obtain the best value of K based on the minimum coefficient rate.

Using Campbell's minimum coefficient rate result and the theorem regarding the property of dominant terms, we demonstrate that the SpEnt methodology can be used to adaptively determine a theoretical value for K. This K value will depend on the signal bandwidth as well as the overall significance of each basis function to the representation of the source data frame. The basic approach to the selection of K is implied by the spectral entropy result illustrated in Fig. 1. In Fig. 1a, we show M MLBT coefficients or coefficients of the M basis functions for N frames of source data, denoted c_{ij}, $i = 1, ..., M, j = 1, ...N$. The coefficients in the first frame are c_{i1}, $i = 1, 2, ..., M$, while the coefficients for the second frame are c_{i2}, $i = 1, 2, ..., M$, and so on. Therefore, the frame index is indicated by the second subscript (j) and the coefficient index is indicated by the first subscript (i). The spectral entropy result implies that each transform coefficient should be considered as a separate sequence, C_{ij}, $j = 1, ..., N$, as shown in Fig. 1c, and the significant values in that coefficient sequence can be determined by comparing to a threshold derived from the energy within that same confficient sequence.

After deriving the relevent number of coefficients for a

particular frame of speech, we now know the total number of singular (i.e. uncombined) basis functions (X) plus the number of basis function sets to be used in the reproduction of the speech data frame (i.e. $K = X + \frac{M-X}{R}$). Note that in using this methodology, a basis function set can be either a grouping of two (R=2) or four (R=4) combined functions. Remember that in either case, the functions in the set will differ only in time localization. Given a value of K, and assuming bands are combined using pairs of basis functions (i.e. $R = 2$), X is equal to $2K - M$. Consequently, given a value of K, and assuming that frequency bands are combined using the combinational matrix in [12] (i.e. $R = 4$), X must equal $\frac{4K-M}{3}$. These results imply that there can be no less than $\frac{M}{4}$ basis functions and no more than M basis functions used in representing a given speech data frame. Given these rules, the process of updating K can then be done on a frame-by-frame basis by dropping the oldest frame from the N-frame block and adding a new M-coefficient frame in its place. In this way, we have a sliding window approach for determining the value of K for each data frame.

Using this adaptive selection process, we are forced to send 7–9 bits of side information per coded speech frame to represent K. This is a relatively insignificant amount of overhead per frame considering that we typically use 256–512 bits/frame when coding at 16–24 kbits/second. By including an additional single bit of overhead, we can select between using half resolution (i.e. $+1/-1$ butterflies) or quarter resolution (i.e. combinational matrix [12]) for band combining in order to achieve the SpEnt-based value for K.

5. RESULTS

For our wideband speech coding trials, we developed a codec similar to that in [9, 16] operating at 16 and 24 kbits/second which uses the NMLBT and its SpEnt-based adaptation algorithm. As seen in Table I, for coded samples of clean speech files using this wideband speech codec, we achieved average segmental SNR values between 21–22 dB for framesizes (M) of 320 samples. For the same samples and codec configuration, we achieved peak segmental SNR values between 28 and 38 dB. These results represented approximately a 1 dB gain in the average segmental SNR and roughly a 1.5–2.0 dB gain in the peak segmental SNR values. Currently, this gain in average and peak segmental SNR is coming at a cost of approximately 400 bps. Work is underway to develop seamless methods of transmitting K so that no additional rate is used in achieving this better quality.

Table 1. Wideband Speech Coding Results (Average Segmental SNR [ASSNR], Peak Segmental SNR [PSSNR])

Sequence (256 kbits/s)	SpEnt (16 kbits/s) ASSNR/PSSNR [dB]	SpEnt/NMLBT (16 kbits/s) ASSNR/PSSNR [dB]
Male #1	21.234 / 35.834	21.787 / 37.666
Male #2	21.117 / 26.255	21.974 / 29.738
Male #3	21.473 / 32.826	22.160 / 34.122
Female #1	19.328 / 31.289	20.304 / 33.239
Female #2	20.570 / 26.709	21.865 / 28.011
Female #3	21.571 / 33.679	22.812 / 35.891

6. REFERENCES

[1] ITU-T G.722.1, "7 KHz audio coding at 24 and 32 Kbps for hands free operation in systems with low frame loss," Sept. 1999.

[2] H. Malvar, "Biorthogonal and nonuniform lapped transforms for transform coding with reduced blocking and ringing artifacts," IEEE Trans. on Signal Proc., vol. 46, pp. 1043–1053, Apr. 1993.

[3] S. McClellan, Variable rate speech coding based on subband measures of spectral flatness, Ph.D. thesis, Texas A&M University, Aug. 1995.

[4] E. Paksoy, K. Srinivasan, and A. Gersho, "Variable rate speech coding with phonetic segmentation," Proc. ICASSP, vol. 2, pp. 155–158, 1993.

[5] D. Donoho and et al, "Data compression and harmonic analysis," IEEE Trans. on Info. Theory, vol. 44, pp. 2435–2476, Oct. 1998.

[6] L. Campbell, "Minimum coefficient rate for stationary random processes," Information and Control, vol. 3, pp. 360–371, 1960.

[7] W. Yang and J. Gibson, "Spectral entropy, equivalent bandwidth and minimum coefficient rate," Proc. IEEE Int. Symp. on Info. Theory, p. 181, July 1997.

[8] W. Yang, Coefficient rate and adaptive coding of side information, Ph.D. thesis, Texas A&M University, May 1998.

[9] J. Gibson, M. Kokes, and J. Ridge, "The SpEnt methodology for lossy source coding," Proc. IEEE DSP Workshop, Oct. 2000.

[10] R. Coifman and M. Wickerhauser, "Entropy-based algorithms for best basis selection," IEEE Trans. on Info. Theory, vol. 38, pp. 713–718, Mar. 1992.

[11] E. Wesfreid and M. Wickerhauser, "Adapted local trigonometric transforms and speech processing," IEEE Trans. on Info. Theory, vol. 41, pp. 3596–3600, Mar. 1993.

[12] H. Malvar, "Enhancing the performance of subband audio coders for speech signals," Proc. IEEE Int. Symp. on Cir. and Sys., June 1998.

[13] R. Cox, "The design of uniformly and nonuniformly spaced pseudo-quadrature mirror filters," IEEE Trans. on Acoust., Speech, and Signal Proc., vol. 34, pp. 1090–1096, Oct. 1986.

[14] J. Lee and B. Lee, "A design of nonuniform modulated filterbanks," IEEE Trans. Circuits Syst. II, vol. 42, pp. 732–737, Nov. 1995.

[15] T. Cover and J. Thomas, Elements of information theory, Wiley and Sons, New York, NY, 1991.

[16] M. Kokes and J. Gibson, "SpEnt-based wideband speech compression," Proc. of the Asilomar Conf. on Sig., Sys. and Comp., Nov. 2000.