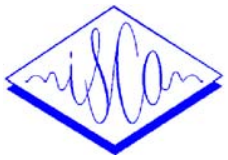

EUROSPEECH '97
5th European Conference on Speech
Communication and Technology
Rhodes, Greece, September 22-25, 1997

LPC QUANTIZATION USING WAVELET BASED TEMPORAL DECOMPOSITION OF THE LSF

Aweke N. Lemma[⋆], *W. Bastiaan Kleijn*[†], *and Ed. F. Deprettere*[⋆]

[⋆] Department of Electrical Engineering
Delft University of Technology
2628 CD Delft, The Netherlands

[†] Department of Speech, Music and Hearing
KTH (Royal Institute of Technology)
Box 700 14, 100 44 Stockholm, Sweden

ABSTRACT

The quantization of linear prediction coefficients (LPC) is an important aspect in low bit rate speech coding. In this work, we introduce a new approach, which exploits the temporal dependencies in the line spectral frequencies (LSF). We approximate each LSF track using expansion into wavelet basis functions. As the LSF vary fairly smoothly as functions of time, they perform very well when interpolated. By vector quantizing the resulting wavelet expansion coefficients, the interpolated LSF tracks could be quantized with a distortion of 0.91 dB using only 15.6 bits per 20 ms update (780 bits per second). This is about 4 bits per update less than the results obtained with previously described procedures.

1 INTRODUCTION

The amount of information in a speech signal can be measured from different perspectives. For instance, one can try to estimate the information rate transmitted from the brain to the speech-production apparatus (or vice-versa), or one can also estimate the information rate of the speech signal itself. The former approach generally results in values that are much lower than the latter. In fact, estimates that are based on the former approach are considered to give the lower bound for the bit rate required for a speech coder [1]. Moreover, the low information rate associated with these estimates provides the motivation to study the human speech-production apparatus. One commonly used speech coding technique that is based on the understanding of the human speech production system is the linear prediction (LP) approach.

In the LP based speech coding, the speech signal is characterized in terms of an LP residual signal and a set of LP parameters. The LP parameter set is generally interpreted as a description of the short-time speech power spectrum. When reconstructing a speech signal, accuracy of this spectrum is essential both for good speech quality and high intelligibility. As a result, the search for more efficient quantization of the LP parameters has become a very active research area. In this paper, we present a method for transparent quantization of the LP parameters at a very low bit rate. It exploits the temporal structure in the parameters.

To put our LP quantization approach in proper perspective, we will first provide some comments on the state-of-the-art. In this discussion, it is useful to distinguish between memoryless quantization and quantization with memory of the LP parameters. In memoryless quantization, only dependencies within a set of LP parameters can be exploited. An example of memoryless quantization is the split vector quantizer of Paliwal and Atal [2]. Quantizers with memory can also exploit temporal dependencies. Examples of such quantizers are the temporal decomposition method of Atal [3] and the long-history quantizer of Xydeas and Ko [4].

To allow quantizer comparisons, most authors have adopted the same distortion measure: the mean of the rms log spectral distortion. A distortion of 1 dB or less, with no outliers beyond 4 dB and less than 2% above 2 dB is commonly used as a definition for "transparent coding" [2]. Practical memoryless vector quantization (VQ) schemes seem to require bit rates of 24 bits per LP set or more (e.g., [2]). However, memoryless quantization at bit rates of down to 20 bits per LP set have been reported [5]. At the common update rate of 50 Hz, these correspond to 1200 and 1000 b/s.

Several authors provide estimates of the bit rate reduction made possible by exploiting temporal dependencies. Svendsen [6] reports that for scalar quantizers, the bit rate can be reduced by 50%. Svendsen estimates that his method, in combination with a practical VQ, could result in 14.5 to 17 bits per LP set for transparent coding. (For a 50 Hz sampling rate his estimates would be lower.) Based on information-theoretical arguments, Eriksson [7] estimates savings of 5 to 6 bits per set for exploiting temporal dependencies. This would lead to a bit rate of between 14 and 18 bits per LP set depending on the memoryless quantizer used as starting point. Neither Svendsen nor Eriksson actually created coders of rates below 20 bits per set. Several other authors do report transparent coding below 20 bits per LP set by using temporal dependencies. Bruhn [8], reports 19.5 bits per set using a noiseless coding scheme which reduced the bit rate down from 24 bits per LP set. Xydeas [4] reports 19 bits per set using a codebook containing the history of the LP parameters.

Thus, previous works suggest that with the exploitation of temporal dependencies transparent coding of the LP parameters at 15 bits per set is possible. We will confirm this hypothesis with a transparent coding scheme requiring only 15.6 bits per LP parameter set (corresponding to 780 b/s).

Atal [3] showed that it is advantageous to use a temporal decomposition into interpolating functions prior to the quantization of the LP parameters. We use a set of predefined wavelet basis functions whereas Atal used adaptive basis functions which had to be transmitted. We convert the LP parameters first to a line-spectral frequencies (LSF) representation and then expand the individual LSF track into the basis functions. The coefficients of the wavelet basis functions are vector quantized. Our quantization procedure is computationally undemanding, but temporal decompositions generally require a high delay.

2 SIGNAL EXPANSION USING WAVELET BASIS

The motivation (from coding view point) for signal expansion lies in the fact that it may provide a lower coding cost, and that the resulting representations can be more robust to disturbances, such as, noise and quantization. If our objective is low coding cost, for instance, the expansion should not affect the coding error significantly. In this work, we

show that describing the LSF using wavelet expansion provides this advantage.

Given the function $f(n)$ satisfying some boundary conditions, in wavelet based signal expansion, we want to find the representation

$$f(n) = \sum_j \sum_k c_{j,k}\psi_{j,k}(n), \tag{1}$$

where the wavelet basis functions $\psi_{j,k}(n)$ are generated by frequency scaling and time shifting the so-called mother wavelet $\psi(n)$ [9–13], that is,

$$\psi_{j,k}(n) = 2^{j/2}\psi(2^j n - k). \tag{2}$$

2.1 Generating the Wavelet Basis functions

A discrete-time wavelet basis can be generated by iterating a two channel filter bank in a tree-structure [10, 11]. The equivalent of a J-level tree structured filter bank is shown in Fig. 1. The left and right halves of the filter bank are

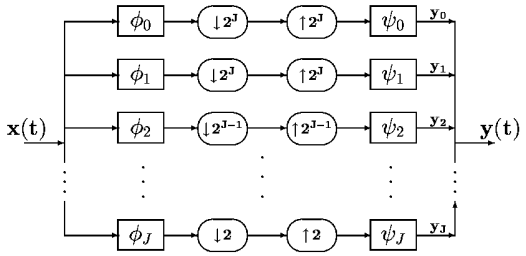

Fig. 1. Filter bank interpretation of wavelets

called the decomposition and the reconstruction filters, respectively. The basis functions of the discrete time wavelet expansion are the impulse responses of the reconstruction filter. The decomposition filter generates the weight coefficients for the expansion. It is important to note that, except for the initial two channels, all adjacent channels have relative sampling rates differing by a factor 2. This is known as the *dyadic sampling grid* [10]. In the dyadic sampling grid the relative sampling rate differences translate to shifts in the time domain giving rise to the so called multi-resolution pyramid. This is demonstrated in Fig. 2 for $J = 3$. In the figure, $\psi_{j,k}$ represents the k-th translation of the j-th scale wavelet basis function ψ_j. The set $\{\psi_{j,k}\}$ for $j \in \{0, 1, \ldots, J\}$ and $k \in \{0, 1, \ldots, n_j\}$ forms the wavelet basis. The two basis functions $\psi_{0,0}$ and $\psi_{1,0}$ are called the scaling function and the mother wavelet, respectively. They formulate the platform for generating the rest

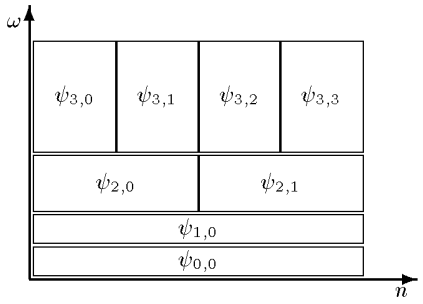

Fig. 2. The time-frequency grid for $J = 3$

of the basis functions (See [11]). Similar representation of the decomposition filter generates the corresponding analysis wavelet family $\{\phi_{j,k}\}$. If we start with a two channel biorthogonal filter bank, the resulting wavelet pair (ϕ, ψ) is said to be biorthogonal.

2.2 The wavelet expansion

Let ϕ and ψ represent the analysis and synthesis biorthogonal wavelet pair. Given the function $f(n)$, we want to find the expansion in (1) where the weighting coefficients $c_{j,k}$, $j \in [0, J]$, $k \in [0, n_j]$ are determined by the inner product[1]

$$c_{j,k} = \langle \phi_{j,k}(n), f(n) \rangle.$$

For simplicity of notation, we replace the double subscripts with a single one. The ordering of the basis functions is retained if the single subscript is defined as $m = \lfloor 2^{J-1} \rfloor + k$. Then, we write

$$f(n) = \sum_m c_m \psi_m(n), \tag{3}$$

$$c_m = \langle \phi_m(n), f(n) \rangle, \tag{4}$$

$m \in [0, \ldots, M-1]$, where M is the total number of basis functions involved in the expansion. Let the function $f(n)$ be partitioned into time segments of N samples, where N is the maximum of the supports of the wavelet basis functions[2]. Let this partitioned time segment be defined as

$$\mathbf{f} = [\ f(n_o) \quad f(n_o+1) \quad \ldots \quad f(n_o+N-1)\ ],$$

with $n_o = 0, N, 2N, 3N, \ldots$
Also, define the matrices Φ and Ψ as

$$\Phi = \begin{bmatrix} \phi_0(0) & \ldots & \phi_0(N-1) \\ \vdots & & \vdots \\ \phi_{M-1}(0) & \ldots & \phi_{M-1}(N-1) \end{bmatrix}, \tag{5}$$

$$\Psi = \begin{bmatrix} \psi_0(0) & \ldots & \psi_0(N-1) \\ \vdots & & \vdots \\ \psi_{M-1}(0) & \ldots & \psi_{M-1}(N-1) \end{bmatrix}. \tag{6}$$

Then we can write

$$\mathbf{f} = \mathbf{c}^T \Psi, \tag{7}$$

where the $M \times 1$ weight vector $\mathbf{c}$ is determined by

$$\mathbf{c} = \Phi \mathbf{f}^T. \tag{8}$$

3 TEMPORAL DECOMPOSITION OF THE LSF USING WAVELET BASIS

It is often convenient to use a transform prior to quantization of a set of parameters. Such a transform can facilitate quantization by lowering the interdependency of the quantized parameters and/or by simplifying the error criterion (or an approximation thereof). Thus, the transformation to line-spectral frequencies (LSF) is advantageous, particularly for memoryless quantization of the LP parameters. Our transformation to wavelet coefficients is a further step towards efficient quantization.

In general, the LSF vary fairly smoothly as functions of time, as can be noted from the fact that these representations perform very well when interpolated [14]. To describe these parameters efficiently, we use expansions into

[1] The inner product $\langle .,. \rangle$ is defined by $\langle \phi_{j,k}(n), f(n) \rangle = \sum_n \phi_{j,k}(n) f(n)$ where $\phi_{j,k}(n)$ and $f(n)$ are both real valued functions.

[2] This corresponds to the support of the mother wavelet.

biorthogonal wavelet basis functions. Thus, given an LSF track[3] $f_k(n), \quad k = 1, \ldots, p$, we want to find the expansion described in (3) for each $f_k(n)$, where the basis functions are generated by iterating a two channel filter bank as discussed in section 2.1.

Prior to the expansion, the means of the LSF are removed. The original LSF are sampled at 50 Hz which means the dyadic sampling intervals for the wavelets of different scales are 40 ms, 80 ms, 160 ms, 320 ms, etc. We used an expansion into wavelets at three scales and the scaling function. We found that the coefficients for the wavelets on the higher scales nearly vanish and they were not used as basis functions. Thus, for each 320 ms segment each LSF track is described by $M = 8$ coefficients (four for the finest scale, two for the intermediate scale, and one for the coarse scale wavelets; and one for the scaling function). Let the partitioned time segment of the LSF be represented by the $p \times N$-matrix F, i.e.

$$\mathsf{F} = \begin{bmatrix} f_1(n_o) & f_1(n_o+1) & \ldots & f_1(n_o+N-1) \\ \vdots & \vdots & & \vdots \\ f_p(n_o) & f_p(n_o+1) & \ldots & f_p(n_o+N-1) \end{bmatrix}. \quad (9)$$

Then, the $M \times p$ wavelet coefficient matrix containing the p weight vectors corresponding to the p LSF tracks is given by (vis. (8))

$$\mathsf{C} = \Phi \mathsf{F}^T, \quad (10)$$

and the approximate LSF tracks by

$$\mathsf{F} \approx \mathsf{C}^T \Psi, \quad (11)$$

where Φ and Ψ are as given in (5) and (6). We use the prediction gains of the auto-regressive filters generated by the original LSF and by the wavelet interpolated LSF to validate the approximation in (11). As an example. the original and the wavelet approximated LSF corresponding to the speech segment "*don't ask me to carry an oily rag like that*" are shown in Fig. 3*a* and Fig. 3*b*, respectively.

4 QUANTIZATION OF THE WAVELET COEFFICIENTS

Each LSF track is quantized independently. As mentioned in the previous section, for each 320 ms interval there are eight coefficients to be quantized per LSF track. The vector of eight wavelet coefficients is vector quantized using a split VQ method: the coefficient that corresponds to the scaling function is scalar quantized with 6 bits, the three coefficient representing the two coarser scales are vector quantized with 11 bits, and the four coefficients representing the finest scale are vector quantized with 8 bits. Using a conventional 10-th order LP representation, 250 bits are required to describe the LSF for 320 ms. This corresponds to an average bit rate of 15.6 bits per 20 ms update interval, or 780 b/s. The quantized LSF that corresponds to the utterance "*don't ask me to carry an oily rag like that*" is shown in ig4.

The quantizer is trained using the standard LBG algorithm. We used a weighted Euclidean distortion measure. Let the M-vector $\mathbf{c}_k$, $k = 1, \ldots, p$ consist of the wavelet coefficients corresponding to the k-th LSF track[4] and $\hat{\mathbf{c}}_k$ the corresponding codebook value. The error criterion is then given by

$$D(\mathbf{c}_k, \hat{\mathbf{c}}_k) = \sum_{m=1}^{M} w_{m,k}(c_{m,k} - \hat{c}_{m,k})^2, \quad (12)$$

[3] p is the order of the LSF coder. In most cases $p = 10$

[4] This corresponds to the k-th column in (10).

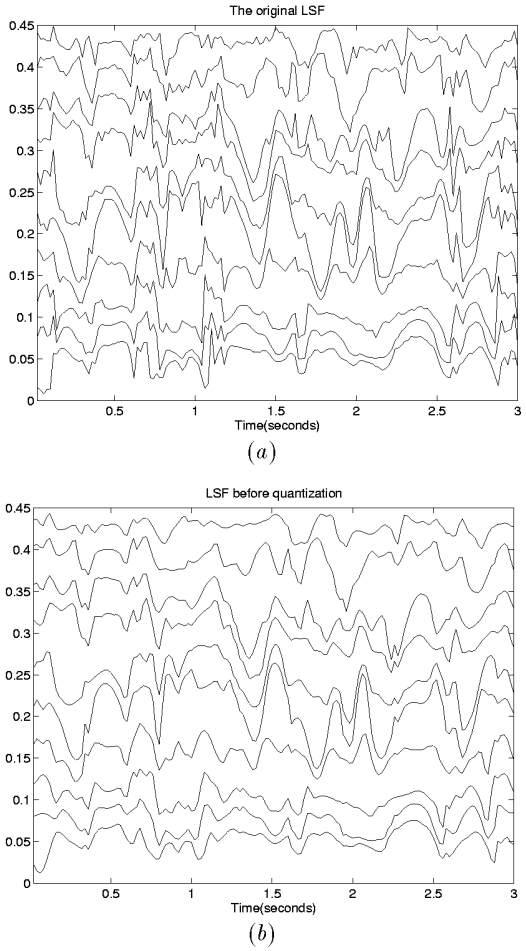

Fig. 3. The LSF corresponding to the speech segment "*don't ask me to carry an oily rag like that*" spoken by a female speaker a) the original LSF b) the interpolated LSF

where the $w_{m,k}$ are weighting factors. Good and even asymptotically optimal weighting procedures exist for the LSF [14–16]. For our purposes, these weighting procedures must be extended to the wavelet coefficient criterion, taking into account the support of the corresponding wavelet. Let $v_{k,j}$ denote a conventional criterion weighing for k-th LSF track at time sample j. As a simple approximation we used as weighing in equation (12)

$$w_{m,k} = \sum_{j \in S_m} v_{k,j} \quad (13)$$

where S_m is the support of the wavelet basis function with the index m. Again for simplicity, we used for $v_{k,j}$

$$v_{k,j} = \frac{1}{f_k(j) - f_{k-1}(j)} + \frac{1}{f_{k+1}(j) - f_k(j)}. \quad (14)$$

5 RESULTS

As we described above, our coding procedure involves two steps; the wavelet expansion (interpolation) step and the quantization step. The interpolation is evaluated using prediction gain and the quantization using spectral distortion.

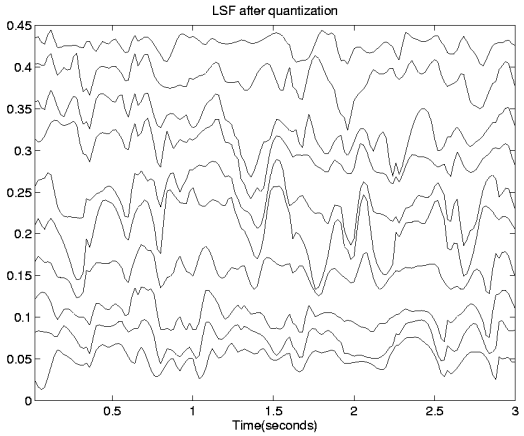

Fig. 4. The quantized LSF corresponding to the speech segment "*don't ask me to carry an oily rag like that*" spoken by a female speaker. Compare it with Fig. 3*b*.

5.1 Interpolation

We found that the prediction gains associated with the interpolated and the original LSF are comparable (the mean prediction gain goes from 9.1 to 8.8 dB, averaged over one minute of speech).

5.2 Quantization

Table 1 provides the main experimental results for our wavelet based quantization procedure. We used data from the TIMIT data base. The training and testing data base were separate and consisted of 200,000 and 6000 sets of LP coefficients, respectively. Our results indicate that a mean spectral distortion of 1 dB is possible at a bit rate of 15 bits per LP parameter set, at an update rate of 50 Hz.

Table 1. Results of the wavelet based LP parameter quantizer.

update interval (ms)	bits per LP set	mean SD (dB)	2-4 dB (%)	>4 dB (%)
20	15.6	0.91	0.85	0

6 CONCLUSION

We found that, a biorthogonal wavelet expansion of the LSF results in an LP parameter representation which facilitates very efficient quantization. We were able to 1) interpolate the LSF with good accuracy and 2) obtain a transparent quantization of the interpolated LSF at a rate of only 15.6 bits per 20 ms update. This result is obtained without exploiting dependencies from between the LSF and with an arbitrarily chosen wavelet basis. We expect that it is possible to obtain further improvement by exploiting the dependencies between the LSF, by choosing optimized wavelet basis functions, and also by using more accurate weighting procedures. We also expect that, with an appropriate choice of the wavelet basis, the slight deficit in the prediction gain will disappear.

In general, our results suggest that temporal dependencies are of paramount importance when designing an efficient quantizer for the LP parameters. Our particular method for exploiting these dependencies resulted in a quantizer having a frame size of 320 ms, making it practical only for nonreal-time applications such as voice storage and speech synthesis. The method can be made robust against channel errors since the wavelets have finite support, and channel errors therefore affect only finite time intervals.

REFERENCES

[1] W. B. Kleijn and K. K. Paliwal, "An introduction to speech coding," in *Speech Coding and Synthesis* (W. B. Kleijn and K. K. Paliwal, eds.), pp. 1–47, Elsevier Science B.V, 1995.

[2] K. K. Paliwal and B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," *IEEE Trans. Speech Audio Process.*, vol. 1, no. 1, pp. 3–14, 1993.

[3] B. S. Atal, "Efficient coding for LPC parameters by temporal decomposition," *Proc. Int. Conf. Acoust., Speech and Signal Processing*, vol. ICASSP'83, pp. 81–84, 1983.

[4] C. S. Xydeas and K. K. M. So, "A long history quantization approach to scalar and vector quantization of LSP coefficients," in *Proc. IEEE Int. Conf. Acoust. Speech Sign. Process.*, pp. II1–II4, 1993.

[5] P. Hedelin, "Single stage spectral quantization at 20 bits," in *Proc. Int. Conf. Acoust. Speech Sign. Process.*, (San Francisco), pp. 57–60, 1992.

[6] T. Svendsen, "Segmental quantization of speech spectral information," in *Proc. IEEE Int. Conf. Acoust. Speech Sign. Process.*, pp. I517–I520, 1994.

[7] T. Eriksson, *Vector quantization in speech coding.* PhD thesis, Chalmers University, Goteborg, Sweden, 1996.

[8] S. Bruhn, "Efficient interblock noiseless coding of speech lpc parameters," in *Proc. IEEE Int. Conf. Acoust. Speech Sign. Process.*, (Adelaide), pp. I501–I504, 1994.

[9] I. Daubechies, *Ten Lectures on Wavelets.* SIAM, Philadelphia, 1992.

[10] M. Vetterli and J. Kovačević, *Wavelets and Subband Coding.* Prentice Hall, 1995.

[11] P. Vaidyanathan, *Multirate Systems and Filter Banks.* Prentice Hall, 1993.

[12] A. N. Lemma and E. F. Deprettere, "Multi-scale nonlinear system modeling using wavelet networks," *The SPIE Proceedings 1996, Denver Colorado*, vol. SPIE-96, August 4-9 1996.

[13] Y. Yu, S. Tan, J.Vandewalle, and E. Deprettere, "Near-optimal construction of wavelet networks for nonlinear system modeling," *Proceedings of 1996 ISCAS Int. Symp. on CS*, vol. ISCAS'96, May 1996.

[14] K. K. Paliwal and W. B. Kleijn, "Quantization of LPC parameters," in *Speech Coding and Synthesis* (W. B. Kleijn and K. K. Paliwal, eds.), pp. 433–466, Amsterdam: Elsevier Science Publishers, 1995.

[15] J. Erkelens and P. Broersen, "On the statistical properties of line spectrum pairs," in *Proc. IEEE Int. Conf. Acoust. Speech Sign. Process.*, (Detroit), pp. 768–771, 1995.

[16] W. R. Gardner and B. D. Rao, "Optimal distortion measures for the high rate vector quantization of LPC parameters," in *Proc. IEEE Int. Conf. Acoust. Speech Sign. Process.*, (Detroit), pp. 752–755, 1995.