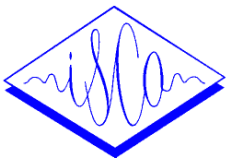


4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

WAVELETS FOR LOW BIT RATE SPEECH CODING APPLICATIONS

F. J. Ancín, M.L. Larreategui, B.L. Burrows and R.A. Carrasco
e-mail:franj@bss10a.staffs.ac.uk
School of Engineering, Staffordshire University
Beaconside, PO 333,
Stafford ST18 0DF
United Kingdom

ABSTRACT

Adaptive transform coding (ATC) and Analysis-by-Synthesis (AbS) have been shown to provide high quality speech above 10 kb/s and below 8 kb/s respectively. The application of the discrete wavelet packet transform (DWPT) and the finite interscale basis coefficient sequences are investigated and evaluated for speech coding applications. A fully quantised DWPT based ATC is presented. Objective and subjective measurements prove the superior WT performance in the adaptive quantisation and bit allocation processes in comparison with other transform based ATC algorithms. Moreover, a hybrid Wavelet-Binary pulse excitation (WBPE) model for CELP speech coding is also presented. Results of a WBPE targeted at 6.3 kb/s are also compared with the conventional CELP coder at 8 kb/s using Gaussian, sparse and ternary codebook population.

1. INTRODUCTION

Wavelets are a new family of basis functions for the space of square integrable functions. A signal $f(t) \in L^2(R)$ can be expanded in terms of translated and dilated versions of a single function called the *mother* wavelet $w(t)$ as

$$f(t)=\sum_{j}\sum_{k=-\infty}^{\infty}2^{j/2}d(j,k)w(2^j t-k) \qquad (1)$$

where the wavelet coefficients $d(j,k)$ are obtained from the inner product of $f(t)$ and $2^{j/2}w(2^j t-k)$. This wavelet expansion yields to a multiresolution analysis of the signal $f(t)$. The mother wavelet function $w(t)$ is, by definition, a band-pass signal and has a centre frequency ω_0. Moreover, the coefficients $d(j,k)$ carry information about the analysed signal $f(t)$ near the frequency $2^j\omega_0$ and the time instant $2^j k$. These coefficients are called the *detail* coefficients. In practical applications, the expansion series of Eq.(1) is truncated and $f(t)$ is approximated by the function $f^J(t)$ defined as

$$f^{(J)}(t)=\sum_{k=-\infty}^{\infty}a(k)g(t-k)+\sum_{j=0}^{J-1}\sum_{k=-\infty}^{\infty}2^{j/2}d(j,k)w(2^j t-k) \qquad (2)$$

where $a(k)$ are referred to as the *approximation* coefficients at scale 2^J and $g(t)$ is the scaling function, which plays an important role in the wavelet expansion. In this paper, only discrete orthogonal compactly supported wavelets of support size equal to, or less than N-1 (where N is an integer) are addressed. These wavelets are completely specified by the constrained sequences $\{c_k\}$ [1][2] of length N which have to satisfy three conditions assuring that the decomposition of any function is numerically stable, the translations and dilation of the wavelet analysis are mutually orthogonal and the regularity of the wavelet decomposition. This sequences $\{c_k\}$ of finite length N are referred to as the interscale basis sequences.

2. A DWPT BASED ADAPTIVE TRANSFORM CODER

In practice, the wavelet coefficients for each scale, i.e. $d(j,k)$ and $a(j,k)$, are recursively computed from $a(j+1,k)$ using a fast efficient pyramidal algorithm. Furthermore, the shapes of $g(t)$ and $w(t)$ are not computed and $a(J,k)$ is approximated by the sampled signal $f(k)$. If a finite set of data of length N is assumed, $2^J=N$ is taken as the finest scale and the discrete signal $f(k)$ is set equal to $a(J,k)$. Therefore, the finest scale is given by the block size N and the sampling rate applied to the discrete wavelet analysis.

Assuming that the incoming data has a period of N samples, the DWT can be seen as an orthonormal linear transform DWT:$\boldsymbol{R}^N \rightarrow \boldsymbol{R}^N$ [3]. The DWT matrix $\mathbf{W}$ is fully specified by the coefficients c_k. The signal processing interpretation of this matrix is passing the periodised input data through a cascade of perfect reconstruction two-channel filter banks, with $L(\omega)$ and $H(\omega)$ the low and high-pass filters, respectively, and followed by a decimation by 2. The practical implementation of the DWT is a tree structure as shown in Fig. 1, where the nodes correspond to the wavelet decomposition subbands in an octave-band manner.

The matrix $\mathbf{W}$ consists of circular shifted blocks, one block for each subband, of a size $M \times N$, where $M=(N/2^m)$ for a band at depth m in the decomposition tree. Within each block, each row is a circular shifted version of the previous row by 2^m samples. Wavelet packet representations [4] are an extension of the DWT. In the DWT discussed above, only the approximation of the signal at a given scale is further decomposed. The DWPT decomposition can be represented as a binary graph in which the subspace components can be freely chosen. A DWPT with a maximum frequency resolution has been chosen for the implementation of the wavelet based ATC and wavelets with different number of vanishing moments will be evaluated. Non-overlapped blocks of speech data $s(n)$ are grouped every N samples (N being a power of 2). In absence of quantisers, the invertibility of the WT makes the output equal to the input, $s(n)=\hat{s}(n)$, having a perfect reconstruction system. However, the aim of any TC system is to reduce the number of bits and a quantisation process has to be performed.

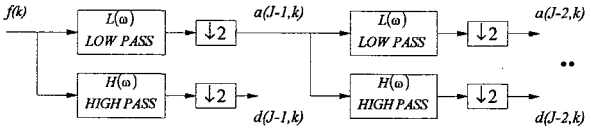


Figure 1: Discrete wavelet decomposition tree structure.

Let us define a DWPT $w_{ik}\{i=1,...,I;\ k=1,...,K\}$ with uniform frequency resolution and having enough subbands, i.e. I, to take advantage of the dynamic spectral characteristics of the speech and transforming it into I wavelet subspaces with unequal variance patterns. On the other hand, the time resolution is higher than in traditional ATC, i.e. K. This time-frequency distribution will characterise the final linear transform **W**.

The properties of the wavelet matrix permit us to assume that the variance of the input is obtained by adding the variance of the different wavelet subbands. Similarly, the variance of the reconstruction error for each wavelet subband σ_i^2 will add up to yield to the variance signal reconstruction error σ_r^2. Hence, the variance of the error introduced by the i quantiser is given as [5] (assuming a constant quantiser performance factor, $\varepsilon_i^2 = \varepsilon^2$, for all w_i) by

$$\sigma_i^2 = \varepsilon^2 2^{-2R} \sigma_{wi}^2 \tag{3}$$

Being R the average number of available bits and R_i the number of bits allocated for each wavelet band; $IR=\sum_i R_i$, the optimal bit allocation is given by [5]

$$R_i = R + \frac{1}{2} log_2 \left[\sigma_{wi}^2 \Big/ \left(\prod_{i=1}^{I} \sigma_{wi}^2 \right)^{1/I} \right] \tag{4}$$

The block diagram of the DWPT based ATC speech coder is shown in Fig. 2. The speech signal is transformed by the DWPT decomposition matrix **W**, obtaining the w_{ik} samples. The variance for each of the wavelet subbands, σ_{wi}^2, is calculated and its square root is quantised using an uniform logarithmic quantiser. The quantised square root of the variance is deployed to obtain the optimal bit allocation and the normalised wavelet coefficients $(w_{ik})_n$. The final quantisation of the normalised wavelet coefficients is realised by means of an optimised non-uniform quantiser [6] which assumes that within each of the wavelet subbands there is a Normal distribution. The stream bit to be sent to the decoder is formed by the side information composed by the quantised square root of the variances, $\tilde{\sigma}_{wi}$, and the bits obtained from the respective quantisers, b_{wi}. In the decoder, the algorithm is performed by first decoding $\tilde{\sigma}_{wi}$ and calculating the proper bit allocation in order to obtain $(\tilde{w}_{ik})_n$ from b_{wi}.

After multiplying the quantised wavelet coefficients by $\tilde{\sigma}_{wi}$ the synthesised speech is obtained by applying the inverse DWPT. The expression of Eq.(4), besides its meaningful theoretical result, is computationally complex for practical applications. Since it provides only an approximation of the bits to be adaptively allocated, eventually, one has to heuristically adjust an integer bit assignment. A very good approximation for practical realisations and which gives better results in practice is obtained by simply taking logarithms in base 2 of all the variances and assigning the bits in a sequential manner to each wavelet subband until all the available bits are assigned. This can be very easily implemented by assigning one bit to the biggest variance, dividing it by 2, and continuing the process until all the bits are allocated. It is important to notice that a similar procedure has already been successfully deployed in traditional subband coding schemes [7].

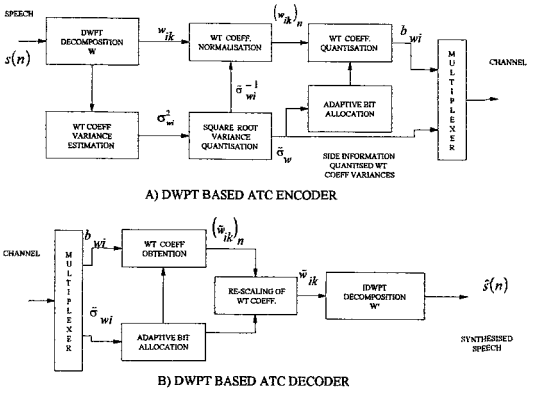


Figure 2: DWPT based ATC coder block diagram.

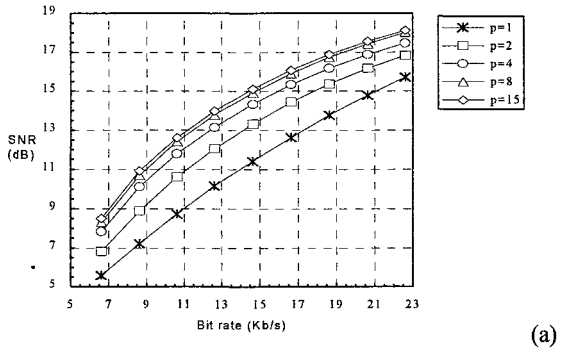
2. 1. Results and evaluation

The DWPT based ATC speech coder has been evaluated for different configuration parameters. Wavelets basis with different numbers of vanishing moments are compared. The number of vanishing moments is varied for p = 1, 2, 4, 8 and 15. This corresponds to interscale basis coefficients of length K = 2, 4, 8, 16 and 30 respectively. The scale of maximum resolution is given by $2^J=N$, being N=128. The number of levels in the wavelet decomposition is also changed and two different number of wavelet subbands are evaluated, I being the number of wavelet subbands.

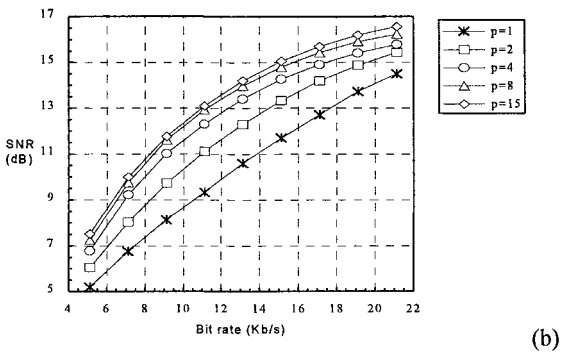
Fig. 3-*a* and Fig. 3-*b* show the objective performance of the DWPT algorithm for N=128 and with I equal to 16 and 8 respectively. In both systems the two last wavelet subbands are discarded without checking their coefficient variances. The bit rates of the side information in the systems evaluated in Fig. 3-*a* and Fig. 3-*b* are 2625 and 1125 bits/second, respectively.

From these figures can be observed that except for the wavelet with p=1, the other wavelet bases perform in a similar way. This can be explained because of the poor frequency resolution given by wavelets with few vanishing moments. Given that a wavelet has more vanishing moments, the sequence $\{c_k\}$ is a longer filter with a sharper stop-band attenuation. Therefore, subbands or wavelet subspaces are better isolated from each other. In other words, the aliasing is reduced and the quantisation error in one subband gets more independent from the rest of the subbands.

Wavelet decompositions can be also performed in a noncausal manner. This is the normal way of implementing a traditional subband coder. In this case there is an improvement over the performance presented in the blockwise mode for high bit rates because of its assumption of signal periodicity. However, the noncausal DWPT introduces an additional delay to the system which is proportional to the length of the sequence $\{c_k\}$ which is doubled with the number of scales in the decomposition.



(a)



(b)

Figure 3: Objective performance of the blockwise DWPT. N = 128, wavelets with different number of vanishing moments versus bit rate. (a) I = 16. (b) I = 8.

3. WAVELET-BINARY PULSE EXCITATION (WBPE)

After the experiments dealing with the wavelet modelling process in [2], the best performance was achieved by locating the wavelet function correlating with the glottal closure and using an optimum distance of approximately the pitch period. This concept is introduced into AbS coder systems. An elegant way of doing this was also proposed in [2] using the pitch adaptive codebook (PAC) approach [8]. This codebook is formed by unit samples, one every α samples starting from the first position. The value α is given by the LTP analysis. The different codewords are then obtained by shifting the codeword one position and repeating α-1 times; or if α is greater than the frame excitation frame N, N-1 times. To avoid pitch doubling errors from the LTP search, if the $\alpha/2$ distance is greater than a given α_{min}, the same process is applied using the $\alpha/2$ distance.

The signal to be modelled by the wavelet functions is the reference or target vector $t(n)$. It is the weighted speech input after subtracting the zero-input response of the pitch synthesis and weighted synthesis filters. The minimisation of the squared weighted error gives the optimum gain g_m for the wavelet function excitations by setting $\partial E_M / \partial g_m = 0$ as [2]

$$g_m = \frac{\sum_{n=0}^{N-1} t(n)\left[c_m(n) * \left(h(n) * w_m^2(n)\right)\right]}{\sum_{n=0}^{N-1}\left[c_m(n) * h(n)\right]^2} \qquad (6)$$

Substituting g_m into the squared error term, E_M, the optimum location into the PAC is obtained by minimising [2]

$$E_M = \sum_{n=0}^{N-1} t(n)^2 - \frac{\left[\sum_{n=0}^{N-1} t(n)\left[c_m(n) * \left(h(n) * w_m^2(n)\right)\right]\right]^2}{\sum_{n=0}^{N-1}\left[c_m(n) * h(n)\right]^2} \qquad (5)$$

where * denotes convolution, $h(n)$ is the impulse response of the weighting filter $W(z)$, $w_m^2(n)$ is the function to project the signal into the correspondent wavelet subspace and $c_m(n)$ is one of the possible PAC codewords. Wavelet parameters are defined by the set of locations L_m from Eq. 6 and then, by applying this location into Eq. 5, the gains g_m are also obtained. As shown is Fig. 4, after estimating these parameters, the pulse amplitudes are projected onto the corresponding wavelet subspaces. The weighted contribution of this signal is finally subtracted from the reference vector $t(n)$. The remaining reference vector $t(n)$ is modelled with a binary pulse excitation which yields a very efficient stochastic contribution. The optimum binary vector $b(n)$ is obtained by minimising

$$E = \mathbf{t}^T\mathbf{t} - \frac{\left(\mathbf{z}^T\mathbf{b}\right)^2}{\mathbf{b}^T\Theta\mathbf{b}} \qquad (7)$$

where $\mathbf{b}$ is the binary pulse vector as defined in [9]. For a frame excitation length of N, there are M binary pulse of value (1,-1) regularly spaced at N/M different phase values. Hence, the optimum stochastic contribution is defined by $\log_2(N/M)$ bits indicating the phase, and M bits indicating the binary vector. Fig. 5 shows the block diagram of the complete WBPE coder. In [2], a WBPE coding system implementation at the bit rate of 6.3 kb/s was presented. The synthesis filter is a direct form 10-th all-pole filter. The LPC coefficients are coded once per 30 ms frame and updated in each 7.5 ms excitation frame through interpolation. Table 1 shows the bit allocation for this specific WBPE coder. The 10 LPC coefficients are coded using nonuniform scalar quantisation of the line spectrum pair LSP coefficients.

A wavelet model of one level of resolution and with functions generated from a 10-sample-length interscale coefficient sequence [1] is employed. Hence, two gains $\{g_g, g_w\}$ have to be obtained. A suboptimal version of the model is employed by using the same optimum location for both functions. The location is coded using 7 bits to cover the PAC. A 12-bit binary pulse codeword is used with four possible phase locations. The optimum binary codeword is efficiently obtained searching only in a M+1=13 codebook. The adaptive codebook approach using delta coding of the delay in odd frames is deployed for the LTP analysis.

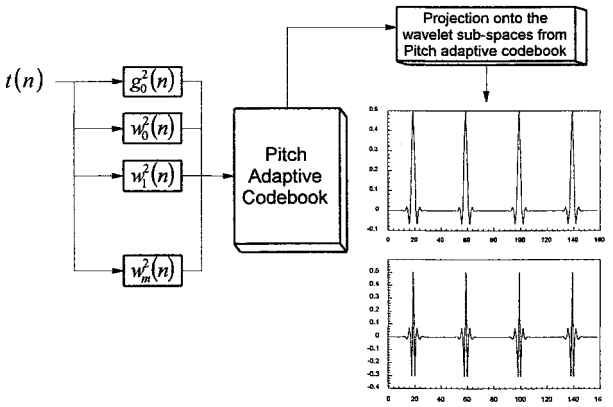


Figure 4: Block diagram of the Wavelet modelling applied as the PAC with the following projection of the optimum amplitudes onto the wavelet subspaces

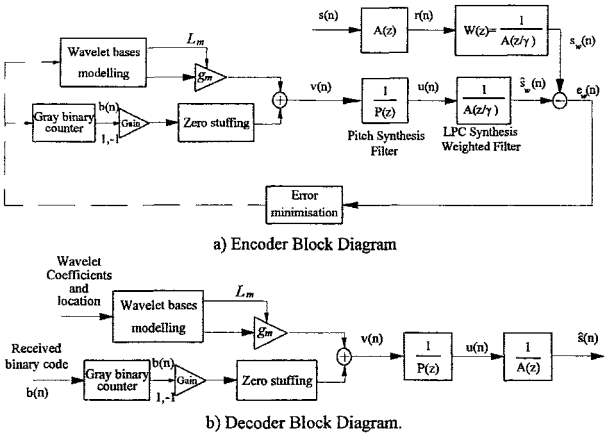


Figure 5: WBPE coding system block diagram.

The autocorrelation formulation is employed for obtaining all the parameters. In order to efficiently code the LTP gain G, the excitation gain β and the wavelet function gains g_w and g_{g_s} the parameters G_s-P_0-P_1-P_2 are created and vector quantised to 7 bits by using a splitting LBG algorithm [10]. These new parameters appear as a function of the average speech energy, which is encoded once per LPC analysis frame, and an estimation of the excitation subframe energy, which is based on the normalised prediction gain of the LPC filter as specified [11].

Parameters	Number of bits
LSF's	36 [3,3,4,4,4,4,4,43,3]
energy	5
LTP delay	24 [7,5,7,5]]
g-w1-s1 signs	12 [3 x 4]
G_s-P_0-P_1-P_2	28 [7 x 4]
wavelet location	28 [7 x 4]
binary pulses	48 [12 x 4]
pulse position	8 [2 x 4]
Total	189 bits per 30 ms

Table 1 : Bit allocation for the 6.3 kb/s WBPE coding

Table 2 shows the similar segmental signal-to-noise ratio, about 1.5 dB less, of our WBPE system in comparison with a standard 8 kb/s CELP for different stochastic excitations even though a lower bit rate was employed. It is also worthy to mention that the complexity of the WBPE is much less than the other algorithms considered, due to the constrained binary pulse excitation used.

Coders	Excitation	SNR-SEG (dB)
8 kb/s CELP	Gaussian	11.63
	Sparse	11.60
	Ternary	11.59
6.3 kb/s WBPE	Wavelets + Binary pulses	10.11

Table 2 : SNRSEG comparison for the different coders.

The subjective quality is indistinguishable and in some cases is even better in the WBPE system. Another WBPE coder targeted at the bit rate of 2.4 kb/s has also been simulated. In this model the excitation is either wavelet function based or binary pulse. This coder structure will be detailed at the conference.

4. CONCLUSIONS

In this paper, the application of the WT to speech coding has been investigated. A waveform coder using one of the possible available DWPT matrices has been implemented in a blockwise and noncausal manner. In the former case, it is a DWPT based ATC system whereas the latter is a subband coding scheme. In both cases, the discrete coefficients specifying the transformation are the interscale basis sequences yielding orthogonal wavelet bases of compact support and with a number of vanishing moments. In the second approach, these sequences are introduced into the excitation of an AbS coding system yielding a hybrid excitation optimised in a closed loop fashion. After some simplifications the coder algorithm has been fully quantised for the bit rate of 6.3 kb/s achieving high speech quality with a competitive complexity.

5. REFERENCES

[1] Daubechies I., ' Orthonormal bases of compactly supported wavelets', Comm. on Pure and Applied Math., Vol.XLI, Pages 909-96, 1988.

[2] Ancín F.J. PhD. thesis, 'Application of the wavelet transform for pitch estimation and compression of speech signals', Staffordshire University, Stafford, UK, May 1995.

[3] Strang G.,'Wavelets and dilation equations: A brief introduction', SIAM Rev., Vol. 31, No. 4. pp. 614-627, Dec 1989.

[4] Coifman R. R., Meyer Y.,Quake S. and Wickerhauser, 'Signal processing and compression with wave packets', Dept. of Mathematics, Yale Univ., 1990.

[5] Jayant, N.S. and Noll, P., ' Digital Coding of Waveforms: Principles and Applications to Speech and Video', Prentice-Hall Signal Processing Series, 1984.

[6] Max J., 'Quantising for minimum distortion', IRE Trans. Inform. Theory, pp. 7-12, March 1960.

[7] Ramstad T.A., 'Sub-and coder with a simple adaptive bit allocation algorithm', Proc. ICASSP'82, pp. 203-207, 1982.

[8] Kondoz A.M., Evans B.G. and Suddle M.R., 'A Speech Coder for TV Programme Description', Eurospeech 1993, pp 257-260.

[9] Salami R.A., 'Binary code excited linear prediction (BCELP): New Approach to CELP coding of speech without codebooks', Electronic Letters, Vol. 25, No. 6, pp. 401-403, 16th. Mar 1989.

[10]Linde Y., Buzo A. and Gray R., 'An algorithm for vector quantizer design', IEEE Trans. Comm., vol. COM-28, pp. 84-95, Jan. 1980.

[11] I.R. Gerson and M.A. Jasiuk, 'Vector sum excited linear prediction (VSELP)', in B.S. Atal et al, ed., Advances in speech coding, Kluwer, Amsterdam, 1991, pp. 69-79.

ACKNOWLEDGEMENTS:

Support was provided under the "Programa de Becas Predoctorales; Formacion de personal investigador", Diputacion Foral de Navarra, Spain.