TRANSFORM CODING OF LSF PARAMETERS USING WAVELETS

Davor Petrinović

Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia

ABSTRACT

A method of inter-frame transform coding of Line Spectrum Frequencies (LSF) using the Discrete Wavelet Transform is presented in this paper. Each component of the LSFs (or of their linear transform) is treated separately and is decomposed into a set of subband signals using the nonuniform filter bank. Subband signals are quantized and coded independently. By the appropriate choice of the mother Wavelet, subband signal with the lowest rate comprises most of the LSF waveform energy. Filter bank effectively decorrelates the input signal, enabling more efficient quantization of the subband signals. A suitable weighted Euclidean distance measure in the Wavelet domain is proposed, defining optimal static or dynamic bit allocation of the subband signals. It is shown that the average bit rate for coding of the DCT transformed LSFs can be reduced by 0.9 bits per vector component by using a very simple Wavelet. The total delay due to inter-frame coding is only 90 ms that is acceptable even for a medium bit rate speech coders.

1. INTRODUCTION

In speech coding, speech signal is usually decomposed into two parts: the envelope of the short time speech spectrum, and the residual excitation signal. Decomposition is performed using the linear prediction method, resulting in a set of coefficients of an all-pole filter representing the short-time spectrum envelope. One of the most popular parameter sets for filter representation are the line spectrum frequencies (LSF), first introduced by Itakura and Sugamura [1].

Due to the non-stationary behavior of the speech signal, linear prediction can be employed accurately only over a short time period known as a frame, 10 - 35 ms long. The parameters of the spectrum envelope must be sampled at least once during this period, to capture perceptually important spectral transitions. On the other hand, the speech spectrum is slowly varying most of the time and may exhibit stationarity extending over as much as a few hundred ms. In a typical speech coder, spectrum envelope is estimated and transmitted every 10 to 25 ms, irrespectively of the current spectral variations, so the parameters of successive spectrum estimates may be highly correlated.

In applications where analysis time (delay) can be extended to several frames, the inter-frame correlation can be exploited to reduce the average bit-rate for spectrum envelope coding. Various techniques have been developed to remove the inter-frame redundancy in smoothly varying segments of speech. The vector predictive coding method, originally developed for coding of a blocked scalar process [2], was later applied to the switched adaptive inter-frame vector prediction for coding of the LPC parameter vectors [3]. Other approaches such as segment quantization and matrix quantization [4] encode several successive parameter vectors as a single entity. They represent a generalization of the vector quantization technique, where the codebook entries are matrices consisting of a spectral vector sequence. Although the compression ratios offered by these methods are significant, the codebook size and complexity are prohibitive for the real-time implementations.

Another possibility of exploiting the inter-frame correlation is by employing transform coding. The sequence of the P-dimensional parameter vectors (e.g. LSFs) is treated as a set of P time-varying signals (one signal for each vector component), and transformed to the parameter 'frequency' domain using any suitable linear transform. Quantization of the decorrelated transform coefficients results in a lower average spectral distortion for a given bit rate, due to transform gain. Application of the two dimensional Discrete Cosine Transform (2D-DCT) performed on a fixed block of successive LSF vectors was investigated in [5]. It was found that even better results could be obtained if the transform is performed after the segmentation process, so the transform size (across time) depends on the actual segment length. A segment is a block of frames with similar spectrum pattern. In [6], the performance of two linear transforms, DCT and the Karhunen-Loeve Transform (KLT) for segmented (adaptive size) parametric speech coding was investigated. Such transform size adaptation is only mimicking the natural behavior of the Wavelets: better time resolution for shorter events and better frequency resolution for longer events. It has been confirmed by several studies that Wavelet decomposition is much closer to the human speech production and perception. Therefore, the approach of using Wavelets as a time-frequency analysis/synthesis technique for inter-frame decorrelation of spectrum parameter vectors seems very appealing. The method of LSF coding in Wavelet domain is presented in this paper, along with some preliminary results verifying its effectiveness.

2. DISTORTION MEASURE AND LSF INTRA-FRAME TRANSFORMATION

Before any discussion of the LSF coding itself, the appropriate distortion measure must firstly be introduced, defining the optimal bit allocation and enabling objective evaluation of the results. Spectral distortion measure is commonly used as a measure of system performance defined as in (1):

$$D_S = \frac{1}{N_f}\sum_{n=1}^{N_f}\left(\frac{100}{\pi}\int_0^{\pi}\left(\log S_n(\omega) - \log \hat{S}_n(\omega)\right)^2 d\omega\right), \quad \left[dB^2\right] \qquad (1)$$

where $S_n(\omega)$ is the unquantized and $\hat{S}_n(\omega)$ is the quantized power spectrum of the n-th speech frame. N_f is the total number of frames. It has been shown in several studies, that whenever D_S drops below 1dB^2, the introduced distortion is perceptually almost negligible. In practice, for the quantizer design, an approximation of D_S is used, based on weighted squared Euclidean distance (WED), d_n^2, between the unqauntized and the quantized LSF vectors, $\mathbf{x}_n$ and $\hat{\mathbf{x}}_n$:

$$D_S \cong \frac{k}{N_f}\sum_{n=1}^{N_f} d_n^2, \quad d_n^2 = \left(\mathbf{x}_n - \hat{\mathbf{x}}_n\right)^T \mathbf{W}_n \left(\mathbf{x}_n - \hat{\mathbf{x}}_n\right). \tag{2}$$

$\mathbf{W}_n$ is the diagonal weighting matrix, which may depend on $\mathbf{x}_n$, while k is the factor of proportionality. In this study, the Inverse Harmonic Mean Weights were used due to their good approximation of the D_S measure.

Due to the ordering property of the LSF vector components, the neighboring LSF parameters within a frame (vector) are highly correlated [5]. By removing this intra-frame correlation, the quantization of the LSF parameters can be improved. This can be done by transforming the P-dimensional LSF vector $\mathbf{x}_n$ into a new P-dimensional vector $\mathbf{y}_n$ using any suitable PxP invertable transform matrix $\mathbf{A}$ (e.g. difference LSF frequencies, DCT or KLT), according to the expression (3):

$$\mathbf{y}_n = \mathbf{A}\mathbf{x}_n, \quad \hat{\mathbf{x}}_n = \mathbf{A}^{-1}\hat{\mathbf{y}}_n . \tag{3}$$

It can be shown that the WED between $\mathbf{x}_n$ and $\hat{\mathbf{x}}_n$ based on matrix $\mathbf{W}_n$ is identical to WED between transformed vectors $\mathbf{y}_n$ and $\hat{\mathbf{y}}_n$ based on the transformed weighting matrix $\mathbf{V}_n$ as in:

$$d_n^2 = \left(\mathbf{y}_n - \hat{\mathbf{y}}_n\right)^T \mathbf{V}_n \left(\mathbf{y}_n - \hat{\mathbf{y}}_n\right), \quad \mathbf{V}_n = \left(\mathbf{A}^{-1}\right)^T \mathbf{W}_n \mathbf{A}^{-1} . \tag{4}$$

Matrix $\mathbf{V}_n$ is symmetrical but generally not diagonal. However, if the quantization errors of different components of $\mathbf{y}_n$ are not correlated, only the diagonal elements of $\mathbf{V}_n$ need to be considered in evaluation of the mean value of d_n^2.

3. WAVELET CODING OF THE TRANSFORMED LSF PARAMETERS

Each component of the transformed LSF vector is treated separately as a time-varying waveform and is decomposed into Wavelet coefficients using the Discrete Wavelet Transform (DWT). If the consecutive transformed LSF vectors $\mathbf{y}_n$ are placed together into a matrix $\mathbf{Y}$ with P rows and N_f columns, then the DWT transform and its inverse can be expressed by the conventional linear operator notation as in (5):

$$\mathbf{Z} = \left(\mathbf{B}\mathbf{Y}^T\right)^T, \quad \hat{\mathbf{Y}} = \left(\mathbf{B}^{-1}\hat{\mathbf{Z}}^T\right)^T . \tag{5}$$

Matrices $\mathbf{Z}$ and $\hat{\mathbf{Z}}$ hold unquantized and quantized DWT coefficients, respectively. The rows of the transform matrix $\mathbf{B}$, as well as the columns of $\mathbf{B}^{-1}$, hold basis vectors, which are actually dilated and translated Wavelet and Scaling functions. Although this definition is mathematically complete, it does not imply the possible of the real-time implementation required for coding applications.

A very simple and efficient realization of the DWT is based on a nonuniform multirate filter bank, as shown in Figure 1. For the proposed method, DWT of the sequence of the transformed LSF vectors $\mathbf{y}_n$ is performed using P independent identical filter banks, each decomposing one of the vector components: $y_1(n)$ up to $y_P(n)$. Analysis filter bank corresponding to the DWT transform of the component i is shown on the left side of the Figure 1., while the synthesis bank corresponding to the inverse DWT, is shown on the right. For a J level DWT, the input signal is decomposed into J+1 subband signals: one approximation signal a_J and J detail signals, d_J, d_{J-1}, ... d_1. These signals correspond to the DWT coefficients in matrix $\mathbf{Z}$. All the signals are critically sampled, i.e. the total number of samples on all of the J+1 outputs is identical to the total number of samples at the input of the analysis bank in a given period of time. In the synthesis bank, the subband signals are combined back together.

If the filters $H_0(z)$ and $F_0(z)$ are chosen to satisfy equation (6):

$$F_0(z)H_0(z) - F_0(-z)H_0(-z) = z^{-L} \tag{6}$$

and if $H_1(z)$ and $F_1(z)$ are derived from $H_0(z)$ and $F_0(z)$ as in :

$$H_1(z) = F_0(-z), \quad F_1(z) = -H_0(-z) \tag{7}$$

then the perfect reconstruction condition is satisfied, and the resulting synthesized signal is only the delayed version of the signal at the input. The total delay equals to (2J-1)L samples, where L is the delay of a single level analysis/synthesis bank. Causal and stable FIR lowpass/highpass filter pairs satisfying (6) and (7) are known as biorthogonal filters and Wavelets corresponding to these filters are known as biorthogonal Wavelets. These Wavelets were employed in this study due to their simplicity and efficiency.

The basic idea of the proposed LSF Wavelet coding method is in the fact that instead of the direct quantization of the vector $\mathbf{y}_n$, J+1 subband signals of each of the P analysis banks are

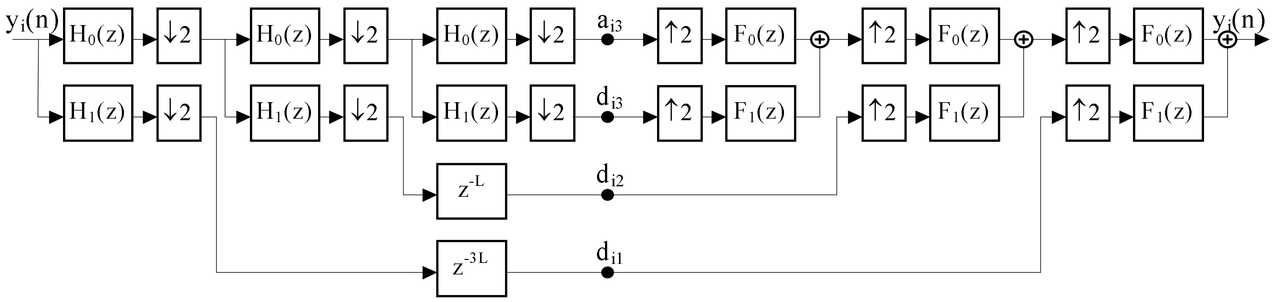

Figure 1: Multirate filter bank realization of a 3-level Discrete Wavelet Transform and its inverse (order=L)

quantized, coded and transmitted. It was expected that due to the transform gain, the same average spectral distortion could be achieved using the lower bit rate. The purpose of the study was to find out the suitable biorthogonal Wavelets that result with the maximum compression of the transformed LSF vector, with the minimum introduced delay.

4. DISTORTION MEASURE AND BIT ALLOCATION IN WAVELET DOMAIN

Although the biorthogonal Wavelets exhibit many good features, they also have one significant drawback: DWT using the biorthogonal Wavelets is neither orthonormal, nor orthogonal transformation. It is only invertable. To fulfill the spectral distortion requirement, the WED distortion measure, d_n^2, between $\mathbf{y}_n$ and $\hat{\mathbf{y}}_n$ must be translated into the Wavelet domain, in order to define how accurately each of the subband signals for each of the vector components must be quantized. DWT transforms the rows of matrix **Y**, $\mathbf{Y}_1$ to $\mathbf{Y}_P$, into the rows of the matrix **Z**, $\mathbf{Z}_1$ to $\mathbf{Z}_p$, where each row $\mathbf{Y}_i$ corresponds to the one of the P components of the vector sequence $\mathbf{y}_n$ in time. A new Wavelet domain WED, e_i^2, between the unquantized and the quantized DWT coefficients in rows i of the matrices **Z** and $\hat{\mathbf{Z}}$ is introduced and defined as follows:

$$e_i^2 = \left(\mathbf{Z}_i - \hat{\mathbf{Z}}_i\right)\mathbf{U}_i\left(\mathbf{Z}_i - \hat{\mathbf{Z}}_i\right)^T, \qquad \sum_{n=1}^{Nf} d_n^2 = \sum_{i=1}^{P} e_i^2. \tag{8}$$

It is obvious that the order of summation is changed. WED e_i^2 is the distortion through time that is finally summed up for all of the P vector components, while d_n^2 is the distortion across the vector that is summed up for all of the N_f vectors. Weighting matrix $\mathbf{U}_i$ can be found as in (9):

$$\mathbf{U}_i = \left(\mathbf{B}^{-1}\right)^T \begin{bmatrix} v_{1ii} & 0 & 0 \\ & \ddots & \\ 0 & 0 & v_{Nfii} \end{bmatrix} \mathbf{B}^{-1}, \tag{9}$$

where v_{1ii} up to v_{Nfii} are the diagonal elements located on the position (i,i) of the weighting matrices $\mathbf{V}_1$ up to $\mathbf{V}_{Nf}$. $\mathbf{B}^{-1}$ is the inverse DWT matrix. As it was stressed before, if the quantization errors of different DWT coefficients in matrix **Z** are not correlated, only the diagonal elements of $\mathbf{U}_i$ need to be considered. The final step in the procedure is to recognize which diagonal elements correspond to each of the subband signal weights through time.

By evaluating (9), a simpler formulation of the subband weights can be given, that is much closer to the concept of the filter bank realization. Firstly, two row vectors must be defined: the input vector $\tilde{\mathbf{V}}_i = \left[v_{1ii} \; \cdots \; v_{Nfii}\right]$, where v_{1ii} up to v_{Nfii} are the weights described as above, and the output vector $\tilde{\mathbf{U}}_i = \left[u_{i11} \; \cdots \; u_{iNfNf}\right]$, where u_{i11} up to u_{iNfNf} are the diagonal elements of $\mathbf{U}_i$. Output vector can be split into J+1 sub-vectors :

$$\tilde{\mathbf{U}}_i = \left[\left[\tilde{\mathbf{U}}_{iaJ}\right]\left[\tilde{\mathbf{U}}_{idJ}\right]\left[\tilde{\mathbf{U}}_{idJ-1}\right]\cdots\left[\tilde{\mathbf{U}}_{id1}\right]\right] \tag{10}$$

where each sub-vector corresponds to one of the subband signals: a_{iJ}, d_{iJ}, d_{iJ-1},...,d_{i1}. These weighting vectors are of different lengths that are directly proportional to the rates of the corresponding subband signals ($\tilde{\mathbf{U}}_{iaJ}$ and $\tilde{\mathbf{U}}_{idJ}$ are the shortest and $\tilde{\mathbf{U}}_{id1}$ is the longest). FIR weighting filters $H_{Wsb}(z)$ are defined next. They are derived from the impulse responses of the filter bank synthesis filters $H_{Ssb}(z)$ by squaring each of the impulse response coefficients:

$$\left.\begin{aligned} H_{Wsb}(z) &= \sum_{k=0}^{Nsb}\left(h_{Ssb}(k)\right)^2 z^{-k} \\ H_{Ssb}(z) &= \sum_{k=0}^{Nsb} h_{Ssb}(k) z^{-k} \end{aligned}\right\} sb = a_J, d_J, d_{J\text{-}1}, \ldots, d_1, \tag{11}$$

where N_{sb} is the synthesis filter order for the subband sb. Subband signal weights $\tilde{\mathbf{U}}_{iaJ}$, $\tilde{\mathbf{U}}_{idJ}$ up to $\tilde{\mathbf{U}}_{id1}$ are found by filtering the weighting signal $\tilde{\mathbf{V}}_i$ with the corresponding weighting filters H_{WaJ}, H_{WdJ}, up to H_{Wd1} and finally decimating the filter output by the same factor that is used for subband signal decimation.

A special case of weighting is the static weighting for which all the elements of $\tilde{\mathbf{V}}_i$ are identical, $v_{1ii} = v_{2ii} = \ldots = v_{Nfii} = \overline{V}_i$. Since the input vector is constant, the filter outputs determining the subband weights will also be constant and can be found by multiplying the component weight $\overline{V}_i$ with each of the subband weighting factors $\overline{U}_{sb}$. Subband weighting factors are equal to the DC responses of the weighting filters:

$$\overline{\tilde{U}}_{isb} = \overline{V}_i \cdot \overline{U}_{sb}, \qquad \overline{U}_{sb} = \sum_{k=0}^{Nsb}\left(h_{Ssb}(k)\right)^2, \tag{12}$$

where sb= a_J, d_J, d_{J-1},...,d_1. Factors $\overline{U}_{sb}$ are different for each of the subband signals and depend on the chosen Wavelet type. For biorthogonal Wavelets used in this study these weighting factors varied from 0.551 to 2.5 and are direct consequence of their nonorthogonality. Once the weighting in the Wavelet domain is established, the optimal bit allocation can be determined using any conventional technique. For the static bit assignment, the average distortion D_W should be minimized :

$$D_W = \sum_{i=1}^{P} \overline{V}_i \sum_{sb} \overline{\overline{U}}_{sb} \lambda_{i,sb} \sigma_{i,sb}(b_{i,sb}), \tag{13}$$

where $\lambda_{i,sb}$ is the variance of the subband signal sb and $\sigma_{i,sb}(b_{i,sb})$ is the variance-normalized minimum mean squared error incurred in quantizing the subband signal sb with $b_{i,sb}$ bits, all for the filter bank i. $\overline{\overline{U}}_{sb}$ is the normalized weighting factor given by:

$$\overline{\overline{U}}_{aJ} = \frac{\overline{U}_{aJ}}{2^J}, \; \overline{\overline{U}}_{dJ} = \frac{\overline{U}_{dJ}}{2^J}, \; \overline{\overline{U}}_{dJ-1} = \frac{\overline{U}_{dJ-1}}{2^{J-1}}, \ldots, \overline{\overline{U}}_{d1} = \frac{\overline{U}_{d1}}{2^1} \tag{14}$$

The minimization is performed by varying $b_{i,sb}$ with the average number of bits per LSF vector component, $\overline{b}$, as the constraint:

$$\frac{1}{P}\sum_{i=1}^{P}\left(\frac{b_{i,aJ}}{2^J} + \frac{b_{i,dJ}}{2^J} + \frac{b_{i,dJ-1}}{2^{J-1}} + \cdots + \frac{b_{i,d1}}{2}\right) \le \overline{b} \tag{15}$$

5. PERFORMANCE RESULTS

The proposed LSF Wavelet coding technique was evaluated on a limited database with the analysis parameters given in Table 1. LSF vectors were intra-frame decorrelated using DCT. Optimal scalar Lloyd-Max quantization was performed on these transformed vectors and the average spectral distortion D_s was calculated as a function of average bit rate per vector component (denoted as *no inter-frame coding* on Figure 2.). Distortion of 1dB was obtained with the average bit rate of 3.27 bits. Unquantized DCT transformed vectors were inter-frame decorrelated using DWT with 15 different biorthogonal Wavelets (currently supported by the *Matlab Wavelet toolbox*). All Wavelets were tested for decomposition levels from J=1 to J=4. The transform coefficients were scalar quantized using the optimal Lloyd Max algorithm, with static bit assignment (same for all frames) according to (13). The example of one cluster assignment matrix is given in Table 2.

Speakers :	1 male speaker	Sampl. fr.:	11025 Hz
Sentences:	16 nonsense	Frame rate:	100.2 Hz
# of frames:	6012	Window:	23.2ms
Preemphas.	0.9375	LPC order:	12

Table 1. Experimental conditions

sb\i	1	2	3	4	5	6	7	8	9	10	11	12
a_2	46	24	22	17	17	16	11	10	11	10	10	10
d_2	14	9	7	5	5	4	4	4	4	4	4	4
d_1	7	5	4	3	2	2	2	2	2	2	3	3

Table 2. Example of the cluster allocation matrix for 'bior2.2', for J = 2, $\overline{b} = 2.32$ and $\sqrt{D_S} = 1.05$dB

J	best		'bior2.2'	worst	
1	0.71	'bior3.7'	0.66	0.44	'bior1.5'
2	0.88	'bior6.8'	0.86	0.53	'bior1.5'
3	0.95	'bior6.8'	0.89	0.49	'bior3.1'
4	0.97	'bior6.8'	0.90	0.45	'bior3.1'

Table 3. Average bit rate reduction per vector component, $\Delta\overline{b}$

D_s was calculated for each of the 60 different Wavelet combinations and for 14 different average bit rates. Three of these 60 combinations are shown in Figure 2., for J=3. The best and the worst results are summarized in the Table 3, for each decomposition level. It is obvious that the best results are obtained with the most sophisticated Wavelet ('bior6.8'), but the results achieved with a very simple Wavelet, 'bior2.2', are only slightly inferior. Two level decomposition using 'bior2.2' decreases bit rate from 3.27 to 2.41, while the total introduced delay is only 9 frames (it includes both analysis and synthesis delays). It is also worthwhile to mention that the two level DWT with 'bior2.2' requires in average only 4.5 additions, 2.25 shifts and 0 multiplication per input vector component, so it represents computationally excellent solution. The results also demonstrate that for static bit assignment using more then 3 decomposition levels results in only minor bit rate reduction.

6. CONCLUSION

A method of LSF inter-frame coding using the Discrete Wavelet Transform was proposed in this paper. The method can be applied either to LSF vector directly or to any version of linearly transformed LSF vector, therefore the intra-frame coding can also be utilized. An efficient real-time realization is suggested, based on the set of the nonuniform filter banks. A transformation procedure for the weighted Euclidean distance measure between LSFs is described, enabling optimal bit allocation in the Wavelet domain. It was shown that even with the static bit assignment and scalar quantization, the average bit rate for LSF coding can be reduced by approximately 0.9 bits per LSF component, compared to the case without inter-frame coding. Since the subbands are quantized independently, the quantization errors are localized within the band, so any perceptually significant subband can be emphasized. The proposed method is also very computationally efficient.

7. REFERENCES

1. Itakura,F. "Line spectrum representation of linear predictor coefficients of speech signals," *J. Acoust. Soc. Am. Vol.57, Supplement No.1, 1975, p S35.*
2. Cuperman,V., Gersho,A. "Vector Predictive Coding of Speech at 16 kbit/s", *IEEE Trans.COM-33, No.7: 685-696, 1985.*
3. Yong,M., Davidson,G., Gersho,A. "Encoding of LPC Spectral Parameters Using Switched-Adaptive Interframe Vector Prediction ", *Proc. IEEE ICASSP, Vol.1 402-405, 1988.*
4. Honda,M., Shikari,Y. "Very Low-Bit-Rate Speech Coding" in *Furui and Sondhi: Advances in Speech Coding,* M.Dekker, 1992.
5. Farvardin,N., Laroia,R. "Efficient Encoding of Speech LSP Parameters Using the Discrete Cosine Transformation", *Proc. IEEE ICASSP, 168-171, 1989.*
6. Mudugamuwa,D.J., Bradley,A.B. "Optimal Transform for Segmented Parametric Speech Coding", *Proc. IEEE ICASSP, 53-56, 1998.*

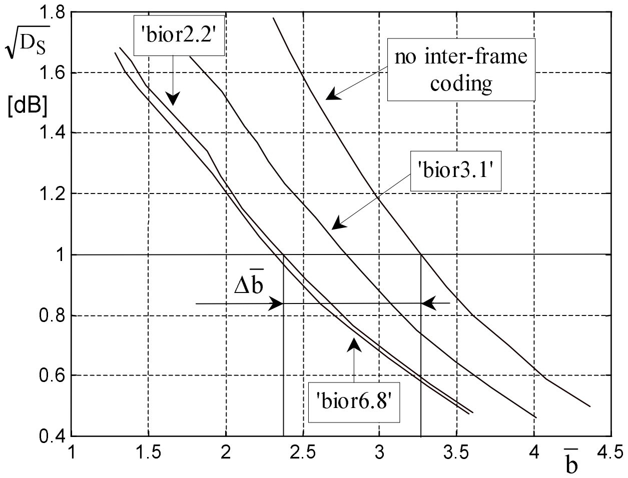

Figure 2: Average spectral distortion as function of average bit rate, with and without Wavelet inter-frame coding, J=3