

# SEGMENTAL QUANTIZATION OF SPEECH SPECTRAL INFORMATION

*Torbjørn Svendsen*

Department of Telecommunications  
The Norwegian Institute of Technology  
N-7034 Trondheim  
NORWAY

## ABSTRACT

The majority of current speech coding algorithms for medium-to-low bit rates transmit two information components, a short-time spectrum estimate and an excitation signal. Even though advanced intraframe quantization schemes have been proposed, the spectral information still consumes a large proportion of the available bit rate. For many speech sounds, the speech spectrum is relatively smooth for time intervals much longer than the sampling rate of the spectrum estimates. Thus, compression can be obtained by identifying smoothly varying segments of the speech spectrum and only transmit the spectral information once for each segment. The segment spectral information is then an approximation to the true spectrum, but if the segmentation criterion is properly chosen, the induced distortion can be controlled to be within the acceptable 1dB mean spectral distortion limit. In the present paper we show that segment quantization can be applied to reduce the required bit rate for the spectral information by a factor of approximately two without compromising the total spectral distortion.

## INTRODUCTION

The transmission of the speech spectral information consumes a major part of the available bit rate in current medium-to-low bit rate speech coders. Several researchers have made important contributions to the design of efficient quantization schemes for the spectral information, thereby reducing the previously lower bound for acceptable quantization (i.e. achieving a mean spectral distortion of 1 dB) from 40 bits per 10-coefficient spectral vector first to 32 bits[1] and then to 24 bits/vector[2].

Although the above achievements have resulted in a significant bit rate reduction, they have only exploited intraframe properties of the speech spectrum. The bit rate required for proper transmission of the spectral information is dependent upon the frame rate as well as the number of bits used to quantize each individual frame. The frame rate has traditionally been determined by the frequency of performing the spectral analysis, i.e., the sampling rate (in time) of the time-frequency pattern. This sampling rate has been chosen such as to obtain a compromise between conflicting interests. The rate should be high enough to capture the spectral transitions that are important to the perceptual quality of the coded speech and at the same time be low enough to give a reasonable bit rate. The speech spectrum is mostly slowly varying, in some cases it can be considered as being stationary over as much as a few hundred

ms. However, the spectral transitions between phonemes (and in some cases also within phonemes, e.g. in plosives) can be rapid, in some cases on the order of 3-5 ms. The typical compromise taken in speech coders is to estimate and transmit the speech spectrum every 10-25 ms irrespective of the current spectral variation.

Because of the necessary compromise when selecting a fixed sampling rate of the spectral information, there is a significant correlation between successive spectral estimates. This can be exploited in inter-frame differential coding schemes (see e.g. [3]) to reduce the bit rate.

Another approach to exploiting the inter-frame correlation is taken in segment (or matrix) quantization schemes [4]. In this case the speech is segmented into variable or fixed length segments. If the segment length is variable, some segmentation criterion needs to be applied. Each speech segment is quantized as a single entity by codebook look-up. The codebook entries are matrices consisting of a spectral vector sequence. For fixed length segments, this is a straightforward extension to vector quantization. If variable length segments are used, an interpolation scheme is necessary to align the input with the codebook entries. The matrix quantization approach is well suited for very low bit rate applications. For applications where the requirements to the spectral distortion caused by the quantizer is higher, the codebook size will be prohibitive for real time realizations.

The use of variable length segments is appealing as this makes feasible the utilization of the varying stationarity duration of the various speech sounds. When the length of the segments used is dependent upon the rapidity of the spectral variation, steady state vowels will produce longer segments than e.g. plosives. These quasi-stationary segments can be efficiently represented by some simple mathematical approximation, which is much more computationally efficient than the matrix quantization approach. A number of different strategies for obtaining a segmentation proper for low bit rate transmission of the spectral information have been proposed in the literature.

A linear interpolation approach is taken in [5]. Here, starting with a segment length of 2 frames, the first and last frame in a segment is taken as reference vectors. An approximation of the LAR spectral vectors

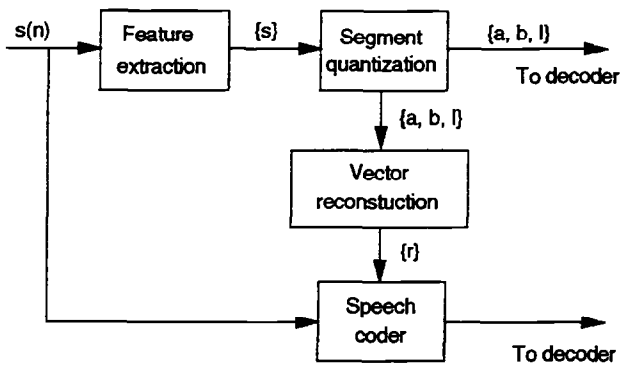


Fig. 1. Segment quantization in a speech coder.

is then made by linearly interpolating between these two vectors. If the mean segment distortion is below a threshold, all frames within the segment are considered to belong to that segment. The segment length is then increased by 1 and the procedure is repeated. This method was originally developed for LPC vocoders. It has later been applied to CELP based speech coders [6]. Recently, a similar method searching for anchor points at instances of maximal spectral change has been proposed [7].

Most of the above methods have been targeted for very low bit rate coding schemes. In the present paper we investigate methods for segment quantization aimed at producing low spectral distortion (i.e. ~1dB). A segment quantization scheme can be embedded in a speech coder as shown in fig. 1. In addition to the linear interpolation method, a new method for segment quantization is described. Experimental data show that the methods could reduce the required bit rate relative to standard intraframe coders by a factor of approximately two for speech passages, possibly by more for silent portions of a conversation. The schemes can be applied to any speech coder using explicit representation of the spectral information in a forward adaptive manner. The drawbacks of the method are increased coding delay and variable bit rate.

### PARAMETRIC REPRESENTATION OF SEGMENTS

Assume that a speech segment consists of a sequence of contiguous parameter vectors,  $S = \{s_1, s_2, \dots, s_k\}$ . For the present purposes, the parameter vectors can be any suitable parametric representation of the speech spectrum, e.g., reflection coefficients, log area ratios, Line Spectral Pairs etc. We wish to represent the speech vectors by a vector sequence,  $R = \{r_1, r_2, \dots, r_k\}$  in such a way that the distortion introduced by the new representation is minimized. Ideally we would like to minimize the mean spectral distortion which on a frame basis is defined as

$$d_2(n) [dB] = \quad (1)$$

$$4.34 \left( \frac{1}{2\pi} \int_{-\pi}^{\pi} |\log S_n(e^{j\omega}) - \log S_{r_n}(e^{j\omega})|^2 d\omega \right)^{1/2}$$

This measure will however be impractical when seeking the best approximation to the sequence  $S$ , and we have to resort to a simpler measure for estimating the representation vector. A simple and efficient measure is the squared parameter error, defined as

$$d(n) = (s_n - r_n)^T (s_n - r_n) \quad (2)$$

where the T denotes vector transpose. Minimization of (2) will be used to obtain the best approximation of the input sequences. However, in order to evaluate the performance of the representation, the spectral distortion will be employed.

The parametric representation of the speech segment,  $S = \{s_1, s_2, \dots, s_k\}$ , can be calculated in a number of ways. One possibility is to represent each segment by its centroid, i.e. the single vector that minimizes the intrasegment squared parameter error. Then,

$$r_n = \frac{1}{k} \sum_{i=1}^k s_i \quad (3)$$

If we wish to exploit that the speech spectrum can be slowly varying, we can represent the speech segment using a first order approximation:

$$r_n = na + b \quad (4)$$

Selecting  $b$  to be the segment centroid and  $a$  as the first order orthogonal polynomial coefficient vector will minimize the mean squared parameter error. The first order approximation will give the opportunity of representing approximately linearly varying segments, and results in a longer average segment length. However, the segment representation now consists of two vectors that must be transmitted.

### Sequential segmentation

In a sequential segmentation scheme, an instant decision is made as to if the input spectral vector belongs to the current segment. Denoting the first speech frame of a segment  $s_1$ , subsequent speech spectral vectors are classified as belonging to the segment if the resulting mean parametric distortion is below some predetermined threshold,  $\Theta$ , i.e., if

$$\frac{1}{k} \sum_{n=1}^k d(n) < \Theta \quad ; \quad 1 \leq k \leq N \quad (5)$$

then  $\{s_1, s_2, \dots, s_N\}$  belong to the same segment.

Another method for sequential segmentation is based on linear interpolation [5],[7]. Here, the first frame of the segment is selected as an anchor point. For each new frame,  $s_k$ , the segment representation is taken as the linear interpolation of vectors between frame 1 and frame  $k$ . The representation and resulting segmentation criterion can be computed from (4) and (5) using

$$a = (s_k - s_1)/(k - 1)$$

$$b = (ks_1 - s_k)/(k - 1) \quad (6)$$

Here, the segment length is  $k-1$  and  $k$  is the index of the next anchor point.

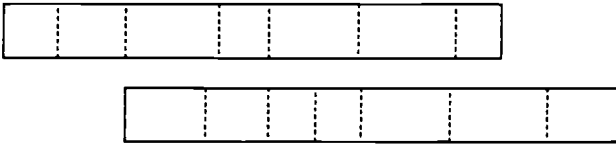


Fig.2. Segmentation variation when the buffer position is changed

### Locally optimal segmentation

The sequential methods described above are inherently sensitive to a single spectral estimate deviating from its neighbors. Spurious frames will induce the sequential segmentation to regard the segment as completed, thus reducing the obtainable compression rate. In [5] it has been suggested to overcome this problem by delaying the decision of segment termination until the immediate future has been examined. We will adopt a different segmentation strategy that alleviates the problem of spurious frames.

If we, instead of making the decision of whether or not to place a new segment boundary at the instance of receiving a new spectral estimate, buffer a sequence of frames sufficiently long to contain at least one segment, we can obtain a segmentation that is optimal within the scope of the buffer size. This can be compared to the advantages of vector quantization over scalar (instantaneous) quantization, and should be able to obtain a better segmentation, i.e., the average segment length should be at least as long as for the sequential strategy, at the same level of spectral distortion.

Assume that our buffer contains the input vectors  $\{s_1, s_2, \dots, s_W\}$ . Our segmentation criterion is now

$$D(m) = \min_{\{b_1, b_2, \dots, b_m\}} \left\{ \frac{1}{W} \sum_{k=1}^m \sum_{n=b_k}^{b_{k+1}-1} d(n) \right\} \leq \Theta \quad (7)$$

Here,  $m$  is the number of segments,  $B = \{b_1, b_2, \dots, b_{m+1}\}$  are the segment boundaries where we have defined  $b_1 = 1$  and  $b_{m+1} = W + 1$  and  $\Theta$  is a predetermined threshold. Segment  $i$  will then consist of the sequence  $\{s_{b_i}, s_{b_i+1}, \dots, s_{b_{i+1}-1}\}$ . The optimal segmentation (and the corresponding number of segments) is found by evaluating  $D(m)$  for  $m = 1, 2, \dots, m_{max}$  until  $D(m) \leq \Theta$ .

Straightforward minimization of (7) will mean having to try out  $\binom{W-1}{m-1}$  boundary combinations for each  $m$ . An efficient method for finding the optimal segmentation based on dynamic programming level building, Constrained Clustering Segmentation (CCS), is described in [8],[9].

If it were possible to fit the entire speech sequence to encoded into the buffer, the CCS algorithm would yield the optimal segmentation. This is generally not possible, both due computation and to delay considerations. Thus, we have to deal with a situation where the locally optimal boundaries may change if the buffer position is changed. This is illustrated in fig. 2.

Because of this, a strategy for adjusting the buffer position after each local segmentation must be devised. Several possibilities are open. One can encode the entire buffer and shift by the buffer length. This has the disadvantage that the last segment always will be terminated at the buffer end, which only rarely is the optimal point.

Strategies like shifting by a fixed or variable number of segments and make arrangements such that the segment boundaries from succeeding segmentations coincide are also possible. We have however found that the simple method of transmitting all segments but the last will give a slightly better performance in terms of compression at a given distortion level.

### Quantization

The compression schemes outlined above will yield a coding gain that is dependent upon the ratio of the number of original speech spectral vectors versus the required number of segment representation vectors: the reduction factor. In the outline of the system it is assumed that the quantization of the segment representation vectors is performed after the actual segmentation, thus any suitable quantization scheme can be applied. In order to get a fair evaluation of the coding gain it is reasonable use the same type of quantizer in the compression scheme and in a system without segment compression.

The quantization of the segment representation will lead to an increase in the spectral distortion at a given compression ratio. The total distortion resulting from the compression and the quantization will however be much smaller than the sum of the distortions from the two contributing sources. This is due to the requirement of a high quality segment representation. When the thresholds are set to yield an unquantized segment representation with an average spectral distortion of less than 1 dB, the spectral variation within a segment will be minor. This means that if a suitable spectral representation is used, the parameter variations within the segment will also be minor, often smaller than the size of a quantization interval. Thus, the subsequent quantization of the segment representation vectors will lead to only a small increase in spectral distortion. In our experience, using a quantizer (vector or scalar) designed to yield an average spectral distortion on the order of 1 dB will be sufficient to yield a mean spectral distortion on the quantized and segment compressed data of approximately 1 dB if the unquantized compression results in a spectral distortion of 0.9 dB.

It should be noted that performing the segmentation and the quantization in separate steps is suboptimal, and that some increase in performance should be expected by using a segmentation scheme where the quantization is embedded in the segmentation process.

### EXPERIMENTS

The various methods for compression of the data rate of the speech spectral information have been tested using the TIMIT speech database. The TIMIT data have a 8 kHz bandwidth and are sampled at 16 kHz. The present study is aimed at telephone bandwidth speech and thus the original TIMIT speech data were lowpass filtered using a 128 order Blackman FIR filter with cut-off frequency of 3650 Hz and decimated to 8 kHz.

The decimated speech was then pre-emphasized using a first order filter with coefficient 0.95. A 10th order LPC analysis was then performed using a 200 sample (25 ms) Hamming window with a frame shift of 80 samples (10 ms). The resulting LP parameters were stored as reflection coefficients and duly converted to the parametric representation desired in the experiments.

For the experiments, some 11500 speech frames from the SX portion of the TIMIT database were used. The speakers were chosen to cover the various dialect regions and only one sentence per speaker was used.

The segmentation methods can be used for any suitable parametric representation of the speech spectrum. Initial experiments were carried out using reflection coefficients, LPC cepstra, Log Area Ratios (LAR), area ratios and Line Spectral Pairs (LSP). It is well known that the LAR and LSP representations in general have relatively smooth time evolutions and thus should be well suited for a segmentation based on a linear (or even zeroth order) approximation. However, the results showed no significant differences in terms of compression ratio for a given distortion level.

At the next stage, the various segmentation approaches were compared. In the previous and in this experiment the segment representation was not quantized as the purpose of the exercise was to investigate the properties of the segmentation methods. The experiments showed that all the described methods were capable of representing the speech spectral information adequately (i.e. at an average spectral distortion of ~1 dB) whilst reducing the average number of parameters for transmission by a factor of approximately two. Only minor differences were found between the methods in terms of reduction factor and average spectral distortion. As an example, in table 1. we have given the results obtained for the LSP representation of the spectral information. In the table, CCS denotes constrained clustering segmentation, SEQ denotes sequential segmentation and LIN denotes sequential segmentation with linear interpolation representation. The digits 0 and 1 denote zeroth and first order approximations respectively (i.e. segment representations according to eq. (3) or eq. (4)).

It should however be noted that the average spectral distortion does not give the full picture of the performance of the methods. The distribution of the frame distortions are of major importance to the perceived quality. In table 1 we have also indicated the proportion of frames within the 0-2dB and 2-4dB ranges as well as the percentage of frames with a distortion greater than 4 dB. As can be seen from the table, the sequential segmentation with linear interpolation between anchor frames gave a significantly greater proportion of frames in the 2-4dB range than the other methods. This means that a larger proportion of frames will have an audible distortion when using this method.

Further experiments were performed using quantization of the segment representation vectors. Simple scalar quantizers for ~1 dB spectral distortion performance were designed. As expected, the introduction of quantization did not result in a large increase in spectral distortion. As an example, a 40 bit/frame quantizer for LP-cepstral parameters at an average reduction factor of 1.95 yielded an average spectral distortion of 1.06 dB. Including 4 bits/segment for the segment length this

Method	Spectral distortion	Reduction factor	0-2dB	2-4dB	>4 dB
CCS-0	1.00 dB	2.05	96.7%	3.3%	0%
CCS-1	1.05 dB	2.06	97.2%	2.8%	0%
SEQ-0	0.99 dB	2.01	96.9%	3.1%	0%
SEQ-1	1.04 dB	2.04	95.7%	2.4%	0%
LIN	1.00 dB	2.19	79.1%	20.9%	0%

Table 1. Performance of the methods using LSP parameters.

means that 22.6 bits/frame is the mean transmission rate for the spectral information. A better quantizer should give the possibility of reducing the necessary bit rate by another 25-45%.

The introduction of quantizers led the overall distortion to increase by 0.01-0.1 dB depending on what type of LP-derived spectral representation and segment representation were used. The variation in performance were however not consistent enough to lead to conclusions as to which combinations were preferable.

## CONCLUSIONS

In this paper several methods for compression of the speech spectral information have been presented and investigated. Segment quantization schemes capable of representing the LP-based spectra at distortions of approximately 1 dB have been shown to be able to reduce the necessary bit rate for transmission by a factor of two.

## REFERENCES

- [1] F.K. Soong, B.H. Juang: "Optimal Quantization of LSP Parameters", IEEE Trans. on Speech and Audio Processing, pp. 15-24, Jan. 1993
- [2] K.K. Paliwal, B.S. Atal: "Efficient Vector Quantization of LPC Parameters at 24 Bits/Frame", IEEE Trans. on Speech and Audio Processing, pp. 3-14, Jan. 1993
- [3] E. Erzin, A. Enis Cetin: "Interframe Differential Vector Coding of Line Spectrum Frequencies", Proc. ICASSP'93, pp. II25-II28, Minneapolis, April 1993
- [4] M. Honda, Y. Shikari: "Very Low-Bit-Rate Speech Coding", in *Furui and Sondhi: Advances in Speech Coding*, M. Dekker, 1992
- [5] V. Viswanathan et al.: "Variable Frame Rate Transmission: A Review of Methodology and Application to narrow-Band Speech Coding", IEEE Trans. on Communications, pp. 674-686, April 1982
- [6] M. Copperi: "Rule-Based Speech Analysis and Application to CELP Coding", Proc. ICASSP'88, pp. 143-146, April 1988
- [7] J.M. Lopez-Soler, N. Farvardin: "A Combined Quantization-Interpolation Scheme for Very Low Bit Rate Coding of Speech LSP Parameters", Proc. ICASSP'93, pp. II21-II24, Minneapolis, April 1993
- [8] T. Svendsen, F.K. Soong: "On the Automatic Segmentation of Speech Signals", Proc. ICASSP'87, pp. 77-80, Dallas, April 1988
- [9] J.S. Bridle, N.C. Sedgwick: "A method for segmenting acoustic patterns, with applications to automatic speech recognition", Proc. ICASSP'77, pp. 656-659, 1977