

EFFICIENT QUANTIZATION OF SPEECH SPECTRAL INFORMATION

Torbjørn Svendsen

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

ABSTRACT

The transmission of the spectral information requires a major part of the total bit rate in today's medium-to-low bit rate speech coders. The speech spectrum is relatively smooth for a much longer period than the update rate of the spectral information for many speech sounds. A method for utilizing this is by variable frame rate segment quantization which as a first step identifies steady state portions of the speech signal and then represents each steady state segment by a simple approximation. In the present paper we show that segment quantization can be applied to reduce the bit rate necessary for transmitting the speech spectral information by a factor of two without compromising the total spectral distortion. As an example, using a simple scalar quantizer with 40 bits/segment, an average bit rate of 22.6 bits/frame resulted in a average spectral distortion of 1.06 dB. Using a more sophisticated quantizer allow for reducing the bit rate without increasing the spectral distortion.

INTRODUCTION

In most current medium-to-low bit rate speech coding systems, the transmission of the spectral information consumes a major proportion of the overall bit rate. Several studies have addressed the issue of efficient quantization of the spectral information, achieving a reduction from the previously accepted standard of 40 bits/frame down to 24 bits/frame whilst maintaining the quantizer induced spectral distortion at a level at or below 1 dB (see e.g. [1]).

The bit rate necessary for transmitting the spectral information is however also dependent upon the frequency at which the spectral vectors are transmitted which has traditionally equalled the rate of performing the spectral analysis. This rate has generally been determined by the variability over time of the speech spectrum. The speech spectrum is mostly slowly varying. However, the spectral transitions between phonemes (and in some cases also the transitions within phonemes, e.g., in plosives) can be rapid, in some cases on the order of 3-5 ms. The spectral transitions are important to the perceived speech quality, and thus one has to reach a compromise between the desire for a low frame rate to reduce the bit rate and the need for a higher frame rate in order to obtain accurate time-frequency representations of the spectral transitions.

Since the resulting frame rate is higher than the rate of spectral change for many speech sounds, neighboring frames exhibit quite strong correlations. This is exploited in inter-frame differential coding schemes to reduce the required bit rate (see e.g. [2]).

A different approach to data rate compression is taken in segment (or matrix) quantization schemes [3]. In this case the speech is segmented into variable or fixed length segments. Each speech segment is quantized as a single entity by codebook look-up. The codebook entries are matrices consisting of a spectral vector sequence. This is an approach well suited for very low bit rate applications. For applications where the requirements to the accuracy of the spectral vector is higher, the codebook size will be prohibitive.

A different approach to segment quantization is when the length of the segments used is dependent upon the rapidity of the spectral variation. Steady state vowels will in this case produce longer segments than e.g. plosives. Thus, the rate of emitted spectral frames will be variable. 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 [4]. 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 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 [5]. Recently, a similar method searching for anchor points at points of maximal spectral change has been proposed [6].

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.

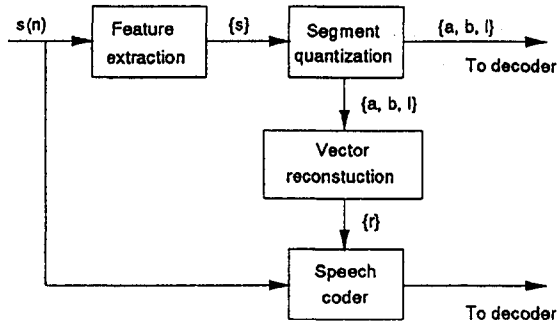


Fig. 1. Segment quantization in a speech coder.

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

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

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

This measure will however be too complex to use 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. If a suitable parameter representation for the speech spectrum is chosen (e.g. LPC cepstrum), there is a direct relation between the squared parameter error and the spectral distortion. 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.

In the following three methods for obtaining a parametric representation of a speech segment will be discussed.

Sequential segmentation

In a sequential segmentation scheme, the input spectral vectors are processed as they are received by the segmentation algorithm. Denoting the first speech frame of a segment s_1 , subsequent speech spectral vectors are determined to belong to the segment if the resulting mean parametric distortion is below some predetermined threshold, Θ , i.e., if

$$\frac{1}{k} \sum_{n=1}^k (s_n - r_n)^T (s_n - r_n) < \Theta \quad ; \quad 1 \leq k \leq N \quad (3)$$

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

The parametric representation of the speech segment 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,

$$r_n = \frac{1}{k} \sum_{i=1}^k s_i \quad (4)$$

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

$$r_n = na + b \quad (5)$$

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 slowly varying segments, thereby resulting in a longer average segment length. However, as can be seen the segment representation now consists of two vectors that must be transmitted, thus doubling the required bit rate.

Another method for sequential segmentation is based on linear interpolation [4],[6]. Here, the first frame of the segment is selected as an anchor point. For each new frame, the segment representation is taken as the linear interpolation of vectors between frame 1 and the last received frame. The representation and resulting segmentation criterion can be computed from (5) and (3) using

$$\begin{aligned} a &= (s_k - s_1)/(k-1) \\ b &= (ks_1 - s_k)/(k-1) \end{aligned} \quad (6)$$

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

Locally global segmentation

Due to their sequential nature, the previously described segmentation methods are sensitive to single frame variations. A possible way of overcoming this problem is by the use of delayed decision. In [4], a procedure which is aimed at reducing the influence of spurious frames by the use of delayed decision is outlined.

A different approach is taken here. The basic idea is that in order to obtain a good segment representation, the segmentation should be performed over a speech window large enough to contain at least one, possibly several speech segments. Within this window a globally optimal segmentation with respect to the chosen distortion measure is to be performed. This scheme will have two major advantages over the sequential segmentation: It will be less sensitive to spurious speech frames and it will ensure that the distortion is minimized.

The segmentation scheme is based on constrained clustering segmentation [7],[8]. Assuming that the sequence of speech vectors within the present window is denoted by $S = \{s_1, s_2, \dots, s_w\}$ where w is the window length. Given the number of segments, Θ , the aim of constrained clustering segmentation is to obtain the $\Theta-1$ segment boundaries, $B = \{b_1, b_2, \dots, b_{\Theta-1}\}$ that minimizes the overall distortion,

$$D = D(1, w) = \sum_{i=1}^{\Theta} D(b_{i-1}, b_i - 1) \quad (7)$$

where $D(i,j)$ denotes the distortion due to the segment with start frame i and end frame j and b_0 and b_∞ are defined as 1 and $\Theta+1$ respectively. The distortion within a segment can be calculated using (2) and summing for all frames within a segment.

As for sequential segmentation, the representation vector can be chosen as a zeroth order or first order approximation, defined by (4) and (5) respectively provided that the summation boundaries and normalization factors are suitably adjusted. Other choices are of course also possible.

Straightforward minimization of (7) would include a search over all possible combinations of segment boundaries. However, as shown in [7], the minimization can be performed using a dynamic programming approach resulting in an efficient solution.

The dynamic programming approach to the segmentation will start by finding the best 2 segment solution. It will proceed by using the $l-1$ segment solution to find the best l segment solution. Increasing the number of segments will obviously reduce the mean distortion. Thus, an obvious stopping criterion will be to terminate the segmentation when the average per frame distortion goes below a predefined threshold,

$$D' = D/w < \epsilon \quad (8)$$

When using constrained clustering segmentation one has to consider how long the window should be shifted between one analysis and the next. If the entire speech sequence to be encoded could be encompassed within the window an optimal segmentation would be result. For practical purposes this is not possible, both due to computation and to delay considerations. Thus, a sub-optimal solution must be found. Possible solutions are:

- Shift by the length of the window. This will ensure that the locally optimal solution is preserved. However, the segmentation can yield different results if speech events in the immediate future is taken into account. This will primarily effect the final segment although it may also have some effect on the segment boundaries within the window. An intuitively preferable solution would be to shift by the window length minus the final segment.
- Shift by a fixed number of segments. The size of the window can be on the order of 20 frames. This will imply that a large delay is incurred when the shift is by the length of the window. A smaller delay can be obtained by shifting by a fixed (small) number of segments.
- Shift by a variable number of segments. In this scheme we would employ a system with memory. The minimum number of segments to shift would be pre-specified and is typically set to one. The system remembers the segment boundaries obtained by the previous segmentation. In addition to the minimum number of segments, the shift is made over the number of segments that are identical from one segmentation to the next.

In fig. 2, reconstructed speech spectra from linear segmentation and constrained clustering segmentation with a first order segment representation are shown.

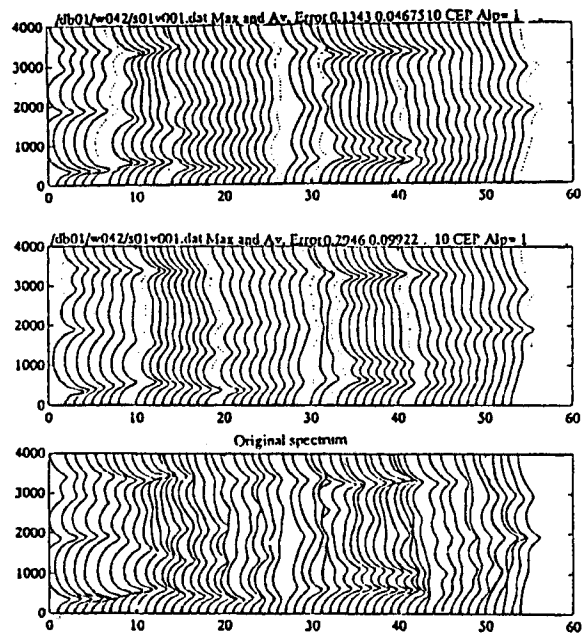


Fig. 2. Reconstructed speech spectra from first order constrained clustering segmentation (top) and sequential segmentation using linear interpolation representation (middle) (from [9])

Quantization

As the system is described above, the step subsequent to the segmentation and segment representation would be the quantization of the segment representation vector(s). For this any scalar or vector quantization scheme could be employed. The quantization of the segment representation will lead to an increase in distortion, but the increase will not be of the same magnitude as the quantizer distortion seen isolated. This is due to the requirement that the segmentation in itself should not introduce distortion that is audible. Because of this, the variation of the parameters within a segment will be minor, often within one quantization interval. Thus, the subsequent quantization of the representation vectors will not introduce an increase in spectral distortion that is on the same order of magnitude as the spectral distortion introduced by using the same quantizer to quantize each frame individually.

However, performing the segmentation and the quantization in two separate processes will be suboptimal and will give less control over the end result than if the quantization and segmentation were to be performed simultaneously. This has been relatively easily obtained for very low bit rate coders employing vector quantizers by computing the segment distortion as the minimum average distortion between the segment vectors and the codebook vectors. For larger codebooks this approach is not viable due to computational considerations and alternative solutions must be found.

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 were then pre-emphasized using a first order filter with coefficient 0.95. A 10th order LPC analysis were 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.

Initial experiments were designed to investigate the performance of the segment representation schemes without quantization of the representation vectors. These experiments showed that all the described methods were capable of representing the speech spectral information adequately (i.e. an average spectral distortion of ~1 dB) whilst reducing the average number of parameters that need to be transmitted by a factor of approximately 2. Only minor differences were detected between the methods in terms of reduction factor and spectral distortion. As an example, the results using LPC derived cepstral parameters for the spectral representation are shown in Table 1. In the table, the CCS denotes constrained clustering segmentation and SEQ denotes sequential segmentation with a segment representation defined by (4) or (5). The digits 0 and 1 denote zeroth order and first order approximations respectively. LIN denotes sequential segmentation with linear interpolation representation.

Method	Spectral distortion	Reduction factor	Variance
CCS-0	0.99 dB	2.20	0.59
CCS-1	1.02 dB	2.17	0.55
SEQ-0	0.95 dB	2.15	0.63
SEQ-1	0.95 dB	2.09	0.55
LIN	1.07 dB	2.43	0.99

Table 1. Performance of the methods using LPC derived cepstral parameters.

It should however be noted that the average spectral distortion does not tell the whole story. The linear segmentation method seems to perform very well compared to the other methods. However, if we look at the proportion of the speech frames with spectral distortion greater than 1 dB, this figure is much higher for the linear segmentation method than for the other methods. This is also reflected in the variance of the spectral distortion as shown in the table. Thus, a larger number of the speech frames will have audible distortion as a result of the segment representation prior to quantization.

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 frame rate of 1.95 yielded an average spectral distortion of 1.06 dB. Including 4 bits/segment for the segment length this 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.

For the constrained clustering segmentation based experiments a number of observations were of interest. The amount of segments used to shift the window had only minor effect on the performance. This means that the shift can be determined by the system specifications allowing for a trade off between delay and computational cost (a large shift will result in a longer delay, but will be less computationally costly than a smaller shift provided that the window length is kept constant).

Some forms of parametric representation of the LP based speech spectra are considered to be better for intraframe coding due to their smoothness in time (e.g. Log Area Ratios, Line Spectral Frequencies). In this study we have tried various representations of the LP spectrum (LPC coefficients, reflection coefficients, Log Area Ratios, LPC cepstra and Line Spectral Frequencies). Although the LAR and LSF representations had a slight edge over the other representations, the differences were not large enough to be conclusive.

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] 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
- [2] E.Erzin, A.Enis Cetin: "Interframe Differential Vector Coding of Line Spectrum Frequencies", Proc. ICASSP'93, pp. II25-II28, Minneapolis, April 1993
- [3] M.Honda, Y.Shikari: "Very Low-Bit-Rate Speech Coding", in *Furui and Sondhi: Advances in Speech Coding*, M. Dekker, 1992
- [4] 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
- [5] M.Copperi: "Rule-Based Speech Analysis and Application to CELP Coding", Proc. ICASSP'88, pp. 143-146, April 1988
- [6] 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
- [7] T.Svendsen, F.K.Soong: "On the Automatic Segmentation of Speech Signals", Proc. ICASSP'87, pp. 77-80, Dallas, April 1988
- [8] 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
- [9] P.O.Husøy: "Forward Connected Artificial Neural Networks Applied to Automatic Speech Recognition", Dr.ing. thesis, The Norwegian Institute of Technology, January 1991.