A SEGMENT VOCODER AT 150 B/S

S. Roucos, R.M. Schwartz, J. Makhoul

Bolt Beranek and Newman Inc.
Cambridge, MA 02238

ABSTRACT

In this paper we investigate several methods for reducing the bit rate of a segment vocoder [1] by 35% to 150 b/s. In the original vocoder we used a random sample of vectors as a set of templates for vector quantization. We demonstrate in this paper that this random quantizer is near-optimal by comparing it with quantizers that use clustering algorithms for quantizing speech segments. The reduction of the bit rate of the segment vocoder was achieved primarily by using a segment network, i.e., not all segment templates are allowed to follow a given segment template. The spectral continuity of speech is used to determine the subset of templates, that can be used to quantize an input segment. To achieve the low rate of 150 b/s, we also reduced the bit rate for coding pitch, gain, and segment duration.

Finally, we present the bit allocation used for transmitting speech at 150 b/s as a single speaker segment vocoder.

1. INTRODUCTION

In an earlier paper [1], we presented the segment vocoder as a new method for transmitting speech for a single speaker at the low bit rate of 230 b/s. In the segment vocoder, we represent the output of LPC analysis at 100 frames/s as a sequence of segments. Each segment consists of a variable number of consecutive frames and has an average duration of 90 ms. To achieve the low rate of the segment vocoder, vector quantization is used for quantizing all the LPC spectra in a segment as a single unit. The segment templates of the vector quantizer were obtained without using clustering techniques due to the large computational load of segment clustering. Instead, a random set of segments was used as the set of templates. We hypothesized then, that the performance of such a random quantizer is near-optimal. In this paper, we confirm this hypothesis by comparing the above random quantizer to quantizers that use clustering algorithms to determine locally optimal sets of templates.

We also describe in this paper the quantization techniques that were used to reduce the bit rate of the segment vocoder from 230 b/s to 150 b/s, a reduction of 35% in the bit rate. The major decrease in the bit rate was achieved by using a segment network that restricts the number of segment templates that can follow a given template. The segment network is used in the following manner: if the current input segment is quantized to a given template, then only those segment templates that follow this template in the network can be used to quantize the following input segment. We describe below the method used to define the segment network that was used to reduce the segment bit rate from 13 bits to 10 bits (per segment) with minimal increase in the segment quantization error.

Finally, we will describe the bit allocation used for the segment vocoder operating at 150 b/s as a single speaker system. In the following section, a brief description of the original segment vocoder is presented. A more detailed discussion of the design issues of the segment vocoder can be found in [1, 2].

2. DESCRIPTION OF THE SEGMENT VOCODER

The segment vocoder uses a 14th order, unquantized LPC analysis at 100 frames/s to represent the input speech. The LPC input sequence is automatically segmented with an average segmentation rate of 11 segments/s. The segmentation algorithm is a heuristic algorithm that uses a set of thresholds on two spectral derivatives to determine the spectral steady-state regions in the input speech. The segments are defined to begin and end in the middle of consecutive steady-states.

In the original segment vocoder, each segment is quantized as single unit independently from other segments. The spectral trajectory, gain track, pitch track and total duration of a segment are also separately quantized.

The vector quantizer used to quantize the spectral trajectory of a segment uses a distance measure that incorporates the required time alignment of two segments. We describe this distance measure [1] because it will also be used for segment clustering which is described in the following section.

The sequence of LPC spectra in a segment represent a piece-wise linear trajectory in the 14 log-area-ratios (LARs) space. We use LARs to represent an LPC spectrum. The total length (using a Euclidean norm on LARs) of a segment is computed. Then, the spectral trajectory is resampled at M equispaced points in LARs space (usually M=10). The space-sampling representation of a segment consists of M 14-dimensional LARs vector (in our case, this corresponds to a 140 dimensional vector for M=10).

The distance measure used for segment quantization defines the time warping between two segments by the correspondence of the consecutive space-samples of the two segments. The distance measure between two segments X and Y is:

$$d = \sum_{i=1}^{M} w_i \; || \; \underline{x}_i - \underline{y}_i \; ||^2$$

2.1

CH1841-6/83/0000-0061 $1.00 © 1983 IEEE

where $\underline{x}_i$, $\underline{y}_i$ are 14 LARs vectors that represent the i-th space-samples of the two segments X and Y, $\| \; \|$ is the Euclidean norm, and w_i is a weight.

The above distance measure is used to quantize an input segment to the nearest segment template. In the original segment vocoder we use 8000 segment templates which corresponds to 13 bits. In the segment vocoder, voicing is not transmitted. The receiver uses the voicing of the template. Similarly, the gain track of the template is used at the receiver but a level adjustment is transmitted using 2 bits. This adjustment equalizes the average gain (in dB) of the input and template segments. The total duration of a segment is quantized and transmitted with 3 bits. Finally, pitch is modelled as a piece-wise linear trajectory and one pitch value is transmitted per segment using 3 bits. Therefore, we transmit 21 bits/segment which yields an average bit rate of 230 b/s for a segment rate of 11 segments/s. We will describe in Section 5 the bit allocation used for the new segment vocoder operating at 150 b/s.

3. SEGMENT CLUSTERING

The above segment vocoder uses 8000 segment templates (~13 bits) obtained by automatically segmenting 15 minutes of speech at an average segment rate of 11 segments/s. The resulting set of segment templates is considered as a random sample of speech segments. We postulated in [1] that the random quantizer, that uses the above set of segment templates, is near optimal. This was based on the result that for a Gaussian random vector with a large dimensionality and with independent and identically distributed components, a random quantizer is optimal. Since we do not expect segments to be Gaussian, we used clustering techniques to determine if the performance of the random quantizer used in segment quantization is far from optimal.

The database used for the clustering experiments consisted of one hour of speech from a single male speaker reading text. The data was collected in four 15 minute sessions spanning a 6 month interval. The database was automatically segmented with an average segment rate of 11 segments/s. The total number of segments was 34,872 (~15 bits). Clustering was used to determine several codebooks of different sizes which were compared to random quantization. We describe below the clustering algorithm used to determine locally optimal sets of templates.

3.1 Clustering Algorithm

The distance measure used for clustering is the same measure used in segment quantization with unit weights and incorporates the time alignment of two segments with different total durations. Each segment is represented by 10 equally spaced space-samples, where each space-sample is a vector of 14 LARs.

The clustering algorithm is a two-pass procedure: a binary clustering to determine a set of clusters, followed by a template selection process for each cluster. The binary clustering phase uses an 8-th order LAR vector instead of 14 to minimize the storage requirement. Therefore, each segment was an 80-dimensional vector and the Euclidean distance on this vector was used in the binary clustering phase. The template selection phase uses the 14 LARs representation to get a set of segment templates.

The binary clustering algorithm is described in [2, 3]. This algorithm is applied sequentially in the following manner on the training data set of segments. Initially, the binary clustering algorithm divides the training data set into two clusters using the K-means algorithm (with K=2). Each cluster is represented by its mean vector. Then, the quantization error for each cluster is computed (using Euclidean distance on 80 dimensional vector). The cluster that has the largest total quantization error (sum of the quantization errors of all the segments in the cluster) is selected for further subdivision into two clusters. This process is repeated until the desired number of clusters is obtained. The K-means algorithm (with K=2) is always used in dividing a given cluster.

The binary clustering algorithm defines a non-uniform binary tree that can be used to efficiently find the nearest cluster representative to an input vector. The large savings in the computational load of the binary clustering over the K-means algorithm (2 $\log_2 M$ distances instead of M distances for M clusters) is needed for segment clustering. We have also found that the quantization error of binary clustering is not much higher than that of the K-means algorithm both for quantizing 14 dimensional LARs vectors and 140 dimensional segments [2].

Given the M clusters obtained by the non-uniform binary clustering algorithm, the second pass determines a set of segment templates. We have used two methods for selecting these segment templates.

1. i) Mean Template: For each cluster, we recompute the mean segment of all the segments in it using the 14 LARs representation. Since the space-sampling representation specifies a time alignment of the segments, the detailed timing is also averaged to specify the timing of the template. The voicing decision for each space-sample is obtained by a majority vote from the corresponding space-sample of all segments in the cluster.

2. ii) Nearest to the Mean Template: To avoid the smearing due to the averaging process, the template of a cluster was chosen as that segment in the cluster nearest to the mean. This segment was represented using 14 LARs.

Using the training data set of 34,872 segments, we compared two sets of codebooks: random codebooks and clustering based codebooks. The random codebooks were obtained by uniformly sampling the training set, i.e., a 10 bit random codebook is obtained by including the segments at multiples of 32 from the 34,872 (~15 bits) segments.

Figure 1 shows the mean square error in quantizing an independent test set of 6 sentences for both the random quantizer and cluster mean quantizer. The random quantizer requires an additional 2 bits to get the same quantization error as the cluster mean quantizer. This loss in performance is rather small and confirms our earlier assumption. Further, for the same bit rate the synthesized test sentences of the random quantizer are crisper and more intelligible than the cluster-mean quantizer. This effect is probably due to the smearing of the averaging process.

To reduce the smearing of the spectral trajectory of a segment template, we used the nearest to the mean template selection method. The quantization error of this quantizer increases to be almost equal to the quantization error of the random quantizer for the same bit rate. Therefore, due to the small amount of data available for each cluster (for the dimensionality of a segment) the nearest to the mean template selection is equivalent to random template selection, i.e., all segments in a cluster are equally far from the mean of the cluster. Also, in listening tests the quality of the random quantizer and the nearest to the mean quantizer is the same. Therefore, random quantization is an effective method for segment quantization which does not use the computationally expensive clustering algorithm to determine a set of segment templates.

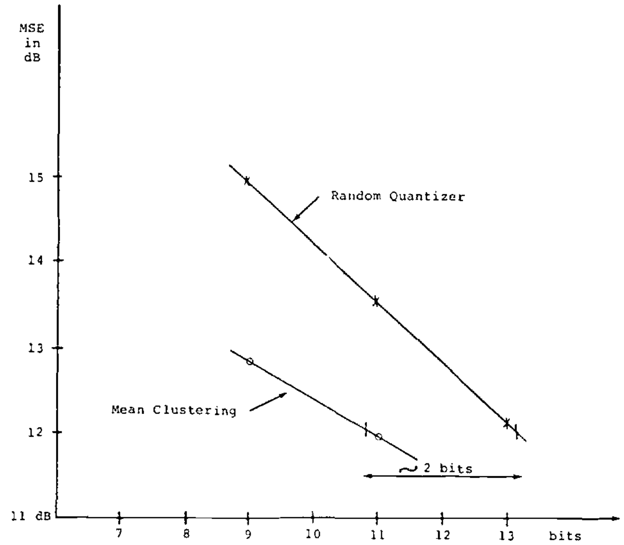

Fig. 1 Comparison of the mean quantization error of a random quantizer with a cluster mean quantizer.

4. SEGMENT NETWORK

In the original segment vocoder the sequence of segment templates that can be used to quantize the sequence of input segments is unconstrained, i.e., any segment template can follow any template. To reduce the bit rate of the segment vocoder we use a segment network to constrain the sequence of segment templates. That is, if the current input segment is quantized to a given segment template, then the following input segment can be quantized only to one of a subset of all available templates. Thus reducing the number of bits used to code the segment template.

A general method for determining the segment network would be to determine statistically which segment templates are most likely to follow a given segment template, i.e., to estimate a Markov chain model for the sequence of speech segment templates. However, this approach would require a prohibitive amount of data to estimate the model. A practical alternative approach is based on an assumption of the spectral continuity of the input speech at the boundary between successive segment templates. If a given segment template matches the current input segment, we assume that the segment template that best matches the following input segment must begin with a spectrum that is close to the last spectrum of the previous segment template. Since the segment templates are generally defined as beginning and ending in relative steady states the assumption stated above is reasonable.

We implemented the above model of the segment network in the following manner. Suppose that the current input segment is quantized to a given template. The last spectrum of this best segment template is used to determine a subset of templates that are allowed in quantizing the following input segment. The templates allowed are those whose first spectrum is nearest to the last spectrum of the template used in quantizing the current input segment. The distance measure used to select the subset of templates is the Euclidean distance using the first 8 LARs. We found that using 8 LARs was slightly better than using all 14, since the high-order LARs vary in time more rapidly than the low-order LARs.

We have tested the segment network idea for a wide range of initial bit rates and bit savings. The effect of using the segment network on the performance of the segment vocoder is shown in Fig. 2. In this figure, we compare the quantization error for the case where all templates were allowed with a segment vocoder that used the segment network. For the unconstrained case, in this figure, we used a total of 1024 segment templates (10 bits). Then for each lower bit rate for the segment network, we restricted the number of templates that could follow a given template to satisfy the bit rate requirement. For example, for the reduced rate of 8 bits/segment, we restricted the choice to the 256 segment templates whose first spectrum was nearest the last spectrum of the previous segment template. The solid line indicates the quantization error for a random quantizer at the same bit rate. When 10 bits of templates are allowed, the segment network allows all templates to follow any template and therefore will have the same quantization error as a 10-bit random quantizer. We can draw two conclusions from the curves in Fig. 2.

1. The segment network achieves a saving of 2.5 to 3 bits over the random quantizer with a negligible increase in quantization error.

2. If the bit rate is reduced by more than 3 bits, the quantization error with the segment network increases much faster, with decreased bit rate, than that of the random quantizer. This indicates that the method used to subset the templates does something consistently worse than a random selection of the templates.

We have found the above two conclusions to hold for several cases. In particular, for the 150 b/s segment vocoder we use a 10-bit network derived from a 13-bit random codebook. During careful listening experiments we were generally not able to distinguish the case of the 10-bit segment network from that of the unconstrained 13-bit segment templates codebook. Careful examination of the templates used in quantization verified that indeed for more than 90% of the input segments, the segment template chosen was the same in both cases.

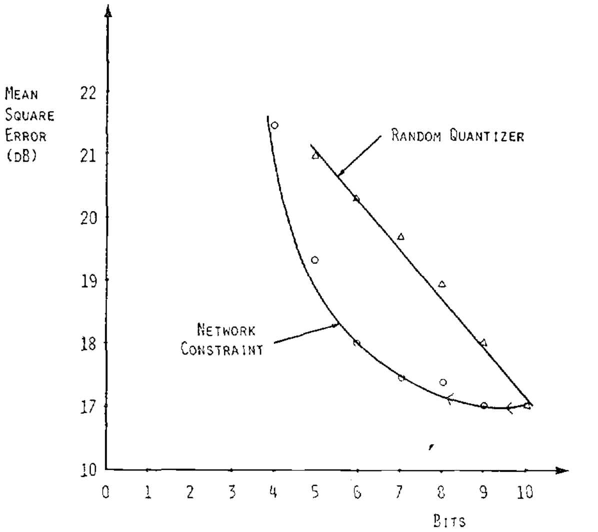

Fig. 2 Mean square quantization error of a random quantizer and a segment network that uses 1024 segment templates.

5. SEGMENT VOCODER AT 150 B/S

We present in this section the quantization methods and bit allocation used for vocoding speech at 150 b/s with the segment vocoder.

The segment templates were obtained by segmenting 15 minutes of speech into 8000 segments (13 bits). A segment network was used to reduce the spectral bit rate to 10 bits/segment as described above.

Pitch was transmitted using 1 bit only for each segment. The input pitch was modelled by a piecewise linear model (linear over a segment) similarly to the original segment vocoder. The change in pitch from the last transmitted value (quantized value) is quantized using an adaptive 2-level quantizer. The two levels were $\pm 3\sqrt{t}$, where t was the segment duration. The resulting pitch sounded quite similar to the original. However, this coding scheme did occasionally change the intended intonation of the input speech.

The gain adjustment and segment duration were quantized using 3 levels each. Therefore we used 14 bits for each segment. At an average segment rate of 11 segments/s, the average bit rate of the segment vocoder is 154 b/s. We used the above segment vocoder to transmit the speech of a single male speaker. In informal listening tests, we have fond that the output speech is quite intelligible (9 words in 10 are intelligible in context).

6. CONCLUSION

We demonstrated in this paper that a random quantizer is effective for segment quantization. The computationally expensive clustering algorithms are not needed to determine a set of segment templates; a random set of segments is equally effective.

We also demonstrated that a segment network based on a spectral continuity model can be used to save 3 bits per segment in quantizing the spectral trajectory of a segment. The segment network was used with quantization techniques for pitch, gain and segment duration to implement a segment vocoder with an average bit rate of 150 b/s. The single speaker vocoder has a good intelligibility to be a useful communication system.

ACKNOWLEDGEMENT

This work was supported by the Advanced Research Projects Agency and monitored by RADC under Contract NO. F19628-80-C-0165.

REFERENCES

1. S. Roucos, R. Schwartz, and J. Makhoul, "Segment Quantization for Very-Low-Rate Speech Coding," ICASSP, Paris, France, May 1982, pp. 1565-1568.

2. S. Roucos, R.M. Schwarts, and J. Makhoul, "Research in Narrowband Communications," Final Report, Contract No. F19628-80-C-0165, Bolt Beranek and Newman Inc., BBN Report No. 5231, November 1982.

3. S. Roucos, R. Schwartz, and J. Makhoul, "Vector Quantization for Very-Low-Rate Coding of Speech," GLOBECOM, Miami, FL, November 1982, pp. 1074-1078.