LOW RATE SPEECH CODING USING CONTOUR QUANTIZATION

Joseph Picone[1] and George R. Doddington

Texas Instruments Inc.
P.O. Box 225016 MS 238
Dallas, Texas 75266

ABSTRACT

Vector quantization-based approaches to speech coding have generated new interest in very low bit rate speech coding, that is, speech coded to bit rates below 1200 bits/sec. To achieve such low bit rates, it is necessary to quantize the pitch and energy parameters at rates below 100 bits/sec. Contour quantization is introduced as a technique in which the contour of a given parameter is normalized by a nominal value and vector quantized. Contour quantization is shown to be extremely robust and efficient in encoding the pitch and energy parameters of the LPC vocoder.

In this paper, a low rate speech coding system which uses contour quantization to encode the LPC excitation is presented. The system is a fixed bit rate system which is intended to operate at bit rates ranging from 400 bits/s to 800 bits/s. The overall system delay varies from 300 ms at 800 bits/s to 400 ms at 400 bits/s. At 800 bits/s, the system achieved a score of 89 on a three male speaker DRT, and a score of 81 on a three female speaker DRT.

INTRODUCTION

Recent developments in vector quantization [1,2] and recent advances in VLSI technology have made low rate speech coding feasible. Low rate coding systems have typically suffered from a number of deficiences including speaker sensitivity, reduced intelligibility, and large system delay. Further, a majority of these systems postprocess the output of an LPC vocoder, thus falling prey to many fundamental vocoder problems such as robust pitch detection. The scope of this study was, nevertheless, limited to the problem of postprocessing the output of an LPC vocoder.

Since the introduction of vector quantization as a speech coding technique, much interest has been focused upon the task of improving computational efficiency via sub-optimal search algorithms. Less interest has been focused on the more fundamental problems of developing codebooks which are truly speaker independent, do not require adaptive training, and are robust to variations in the recording environment. Ultimately, to achieve all of these requirements, a perceptually-based distance metric must be constructed. Borrowing from experience in speech recognition [3], a principal spectral components (PSC) distance metric is proposed as a means to improve spectral distance computations.

Traditionally, much less emphasis has been placed on excitation coding, mainly because excitation parameters typically require much fewer bits than the spectral parameters. We introduce the notion of contour quantization of the fundamental frequency and root-mean-square (RMS) energy as an efficient and reliable means of encoding excitation information. This technique has the advantages of being inherently speaker independent and of being a fixed bit rate quantization scheme. Contour quantization of the fundamental frequency is similar to approximating the fundamental frequency as a piecewise linear function, and encoding the frequency as a series of line segments. The major difference is that a vector quantization approach is used to determine the optimal set of line segments with which a contour is encoded.

Most low rate coding systems rely on some type of frame repeat strategy or frame interpolation strategy to achieve a low bit rate with acceptable speech intelligibility. We introduce a frame interpolation algorithm which replaces selected frames of LPC spectral information with interpolated values derived from adjacent frames, using a synthesis-based distortion criterion. Side information which contains the positions of the interpolated frames must be transmitted to the synthesizer. This algorithm is capable of dropping a maximum of 40 percent of the incoming speech data with minimal degradation in speech quality.

A LOW RATE SEGMENT VOCODER

The segment vocoder in [2] exploits the variable rate nature of speech by segmenting the incoming speech into variable length blocks of data. Our approach to achieving a low data rate, yet maintain a fixed data rate, is to postprocess the output of an LPC vocoder, and segment the incoming speech into N frame blocks of LPC data. The number of frames per block is usually fixed at 10. The blocking length is chosen as a compromise between what is required to achieve a low bit rate (which dictates as long a block length as possible), and a small enough segment such that the speech signal is relatively stationary through the segment. The spectral coefficients, energy, pitch, and voicing within this block are each quantized separately. The frame interpolation algorithm also replaces a predetermined number of frames per block with interpolated values. Each quantization step is described below in detail.

Spectral Quantization. Each LPC frame of spectral information is vector quantized using a perceptually-motivated distance metric. A PSC-based distance metric is chosen mainly because of its superiority [3] to other LPC-based distance measures for tasks such as isolated word speech recognition. The incoming LPC filter parameters are converted to filter bank samples by sampling the LPC filter frequency response at a predetermined number of points. The filter bank is a critical band filter bank [4] consisting of 14 bins. Each bin contains 5 equally spaced frequency samples. The output of each bin in the filter bank is the average of the samples within the bin.

1. J. Picone is currently with AT&T Bell Laboratories, Room IH 6C-336, Naperville-Wheaton Road, Naperville, Illinois, 60566.

CH2396-0/87/0000-1653 $1.00

The filter bank samples are converted to a PSC vector using a whitening transformation [4], a process which statistically decorrelates the PSC vectors. Though one might suspect that the transformation matrix needs to be adapted to the particular recording conditions, experimental results indicate that performance is relatively insensitive to variations in recording conditions. In fact, the transformation matrix used in our system was generated using a database different from both the vector quantization training database and the DRT test database.

Though the PSC vector is actually a 14th order vector corresponding to the 14 filter bank outputs, the PSC vector is truncated to 10 components, removing the 4 components which have the smallest eigenvalues. This is essentially a noise reduction technique [4] which attempts to maximize the discriminating capability of the distance measure. Distances in the PSC domain are computed as the Euclidean distance between two PSC vectors.

The clustering algorithm used to generate the codebook is essentially the K-Means algorithm [5]. We have incorporated one major modification to this procedure. The initial guesses for cluster centroids are chosen through an iterative scheme, such that the initial guesses for cluster centers represent vectors which span the entire vector space. This is accomplished by searching for a set of initial centers which are separated by some predetermined distance threshold. This threshold iteratively decreased from some nominally large value until the desired number of initial cluster centroids are found. In subsequent iterations of the K-MEANS procedure, cluster centroids are represented as an average PSC vector, arithmetically averaged over all elements in the cluster.

Since the PSC representation is not amenable to LPC synthesis (a PSC vector to LPC reflection coefficient transformation is not unique), the autocorrelation vectors corresponding to each PSC vector are retained. As the PSC cluster centroids are computed through averaging, a corresponding autocorrelation vector, also computed through averaging, is maintained. The averaged autocorrelation vector is used for LPC synthesis. This technique of autocorrelation averaging has been found to be acceptable in applications such as template generation in speech recognition [3].

Energy Contour Quantization. Contour quantization is an extremely efficient method of encoding the energy and pitch contours. The energy term typically transmitted in the LPC vocoder is the root-mean-square of the hamming windowed and pre-emphasized speech signal, or RMS energy. The N frame block of RMS values are normalized by an average, or nominal RMS value. The particular form of the normalization chosen is:

$$\hat{E}_n = 1 + \beta \ln (E_n/E_{avg}) , \qquad (1)$$

where,

$$E_{avg} = (1/N) \sum_{n=1}^{N} \hat{E}_n . \qquad (2)$$

β, the normalization weighting factor, controls the amount of emphasis placed on quantizing low energy sections of the speech signal. A typical value for β is 4.0.

In this quantization scheme, the normalized energy vectors, E, are vector quantized. In addition, the nominal value of energy is transmitted once per block. Borrowing from the notion of predictive coding, the energy vector is normalized by the quantized value of the nominal RMS value of Eq. 1. Thus, the quantization error is only via the vector quantization process. The energy codebook is trained using the K-MEANS clustering algorithm mentioned above, with the clustering algorithm operating on the normalized energy vectors.

Pitch Frequency Contour Quantization

The pitch contour can also be quantized using contour quantization. However, the unvoiced values of the pitch must be replaced with some average value of the pitch, such that the contour appears smooth. The nominal value of the pitch is the average pitch, where averaging is performed in the log domain:

$$P_{nom} = \exp [(1/M) \sum_{n=1}^{N} \ln (\max(P_n,1))] , \qquad (3)$$

where M represents the number of non-zero pitch values in the block, and P_n represents the pitch frequency in Hz for the nth frame.

All unvoiced frames are filled with the nominal value of the pitch. Note that unvoiced frames are not interpolated, since this has been found to produce inferior speech quality, due to the tendency of the vector quantized pitch contours to be excessively smooth. The filled contour is then normalized by the quantized value of the nominal pitch, as follows:

$$\hat{P}_n = \ln(P_n/P_{nom}) . \qquad (4)$$

As before, the contour is vector quantized, and the nominal value of the pitch, along with the pitch codebook index, is transmitted. Note that this quantization scheme requires explicit quantization of the voicing.

Voicing Quantization. Voicing information requires at most one bit per frame to transmit, since it is a binary parameter. Yet, the entropy of the voicing parameter is very low, typically less than 0.5 bits for an LPC vocoder operating at 20 ms frame period. Our approach to quantizing voicing is to vector quantize the N frame binary voicing vector using a hamming distance measure. This is equivalent to accumulating the most frequently occurring voicing patterns.

A table of the likelihoods of the most common voicing patterns for a 10 frame block is shown in Table 1(a). This table was constructed from 20 ms LPC data collected over a large database described in the next section. The entropy of the voicing vector is 3.9 bits/vector. Observe that a majority of the patterns are merely locating a voicing boundary. In Table 1(b), a similar table of voicing vector likelihoods has been constructed for 30 ms LPC data. The entropy of the voicing vector at 30 ms is 4.5 bits/vector. The entropy is slightly higher, indicating that the 10 frame vector, which spans 300 ms for the analysis in Table 1(b), represents a more complicated voicing structure than the analysis in Table 1(a) (which spans 200 ms).

Frame Interpolation. Further bit rate reduction can be achieved through some frame repeat or frame interpolation strategy. Experience with speech recognition template generation [4] indicates that replacing LPC frames with frames which have been interpolated in the autocorrelation domain from their adjacent neighbors will maintain an acceptable level of speech quality. With this in mind, we impose the constraint that no two adjacent frames can be replaced by interpolated values. Thus, we will decimate only isolated frames.

The choice of frames to be decimated is guided by an analysis-synthesis strategy which considers what combination of frames, if interpolated, will produce the least distortion at the synthesizer. Our distortion measure is defined as:

$$d_n = || (1/2) (\hat{S}_{n-1} + \hat{S}_{n+1}) - S_n ||^2 , \qquad (5)$$

where S_n denotes the PSC vector for the nth frame, and $\hat{S}_n$

denotes the quantized PSC vector for the nth frame. Eq. 5 represents the error if the nth frame is replaced by its interpolated value. The set of frames to be dropped is chosen to minimize the total decimation error over the N frame block, which is the sum of Eq. 5 over the set of frames under consideration. For example, if two frames per block are desired to be dropped, all combinations of two frames which are non-adjacent must be tested. For each combination of two frames, Eq. 5 is computed for both frames, and summed. The combination that minimizes this sum is chosen as the set of frames to be replaced. Experimental results indicate that a maximum of 40 percent of the incoming speech frames can be replaced by their interpolated values with minimal degradations in speech quality.

EVALUATIONS AT 400 BITS/S AND 800 BITS/S

DRT [6] results were collected at two different bit rates. In each case the system was trained over a database consisting of 16 male and 16 female speakers, using a total of 30 minutes of speech data. The speakers contained in the DRT test data were not included in the training database. The training database was collected under studio conditions recording conditions. The speech material was excised from recordings of non-rehearsed conversational speech.

Two variations of this low rate coding system were evaluated. The first system postprocessed the output of an LPC vocoder which used a pre-emphasis of 1.0, an LPC analysis window length of 30 ms, an LPC model order of 10, and a frame period of 20 ms. The sizes of the various codebooks are shown in Table 2(a). Note that only one frame per block was decimated. Further bit rate reduction can be achieved at virtually no reduction in synthetic speech quality by increasing the number of frames decimated per block to three, which lowers the overall bit rate of the system to 670 bits/s.

A second system which operates at 400 bits/s was also evaluated. One major tradeoff that is necessary to achieve a 400 bits/s bit rate is that the frame period must be increased to 30 ms. The sizes of the various codebooks are given in Table 2(b). The number of spectral vectors has been decreased to 1024, or 10 bits. Quality degraeds significantly if the number of spectral vectors is decreased even further, while slight improvements in quality can be achieved by increasing the number of spectral vectors to 2048. A DRT was conducted using a three male (Tape E-2-A) and three female (Tape E-5-A) test. The DRT results are compared to the baseline 2400 bits/s system in Fig. 1. The combined DRT score is an unweighted average of the male and female scores. Observe that the DRT scores are significantly lower for females than males, and decline faster for females as the bit rate decreases. It is no surprise that females vocode more poorly at these low rates, since that phenomena has been observed for 2400 bits/s vocoders.

The performance of the 800 bits/s system for the three male speakers approaches that of the 2400 bits/s LPC vocoder. The majority of the degradation is a result of the spectral vector quantization process. Even though the spectral vector quantization uses 13 bits for spectral information, or 8192 codebook vectors, there is still significant degradation in speech quality over the 2400 bits/s scalar quantization system. Informal listening tests which tested only the spectrum quantization portion of the algorithm have verified that the majority of distortion is introduced by the spectral quantization. Overall, the speech quality at 800 bits/s can be best described as intelligible. The synthetic speech at 400 bits/s, though intelligible, is somewhat muffled, and significanly more buzzy.

CONCLUSIONS

Contour encoding of the LPC excitation has been introduced as an effective coding technique for low rate coding applications. A system based upon contour quantization has been shown to produce synthetic speech at 800 bits/s which is highly intelligible. This low rate coding system requires approximately 6 MFLOPS, and currently runs in real-time on two Floating Point Systems AP120B array processors. Though the effects of fast codebook searches and sub-optimal search strategies have not been studied, there is good reason to believe that the number of MFLOPS can be easily reduced. Over 75 percent of the available real-time is consumed by the spectral vector quantization.

The intelligibility for female speakers was demonstrated to be significantly lower than the intelligibilty for male speakers. A robust frame decimation criterion was introduced as an effective means to reduce the bit rate without compromising speech quality. The reduced synthetic speech quality of the low rate coder presented in this paper is mainly attributable to the specral vector quantization. A PSC-based distance metric has been proposed as a way to improve the vector quantization process. To further improve speech quality, a better perceptually-based distance measure must be developed. One possible representation which might accomplish this is a formant-based description of the spectrum. Our future research will be focused towards developing a more efficient representation of the spectral information.

REFERENCES

[1] R. M. Schwartz and S. E. Roucos, "A Comparison Of Methods For 300-400 B/S Vocoders," Proc. 1983 IEEE Int. Conf. Acoustics, Speech, Signal Processing, pp. 65-68, April 1983.

[2] S. E. Roucos, R. M. Schwartz, and J. Makhoul, "A Segment Vocoder At 150 B/S," in Proc. 1983 IEEE Int. Conf. Acoustics, Speech, Signal Processing, pp. 61-64, May 1982.

[3] P. K. Rajasekaran and G. R. Doddington, "Speech Recognition In The F-16 Cockpit Using Principal Spectral Components," in Proc. 1983 IEEE Int. Conf. Acoustics, Speech, Signal Processing, pp. 882-885, March 1985.

[4] E. L. Bocchieri and G. R. Doddington, "Frame-Specific Statistical Features For Speaker Independent Speech Recognition," IEEE Trans. Acoustics, Speech, and Signal Processing, Vol. ASSP-34, No. 4, pp. 755-764.

[5] M. R. Anderberg, Cluster Analysis For Applications, Academic Press, New York, 1973.

[6] W. D. Voiers, "Diagnostic Evaluation Of Speech Intelligibility," Benchmark Papers on Acoustics: Speech Intelligibility and Recognition, ed. by M. E. Hawley, Dowden, Hutchinson, and Ross, Stroudsburg, PA, Vol. 11, pp. 250-275, 1977.

Voicing Pattern	Likelihood
1111111111	0.319359
0000000000	0.237641
0011111111	0.024094
1111111100	0.023891
0001111111	0.023531
1111111000	0.023516
0111111111	0.022906
1111111110	0.022719
0000111111	0.021906
1111110000	0.021688
0000011111	0.019422
1111100000	0.019047
0000001111	0.016297
1111000000	0.015797
0000000111	0.013875
1110000000	0.013547
0000000011	0.011750
1100000000	0.011359
0000000001	0.010594
1000000000	0.010266

Table 1(a). Voicing Pattern Likelihoods For 20 ms LPC Voicing Data

Voicing Pattern	Likelihood
1111111111	0.260875
0000000000	0.203766
1111111100	0.026578
0111111111	0.026531
0011111111	0.026516
1111111110	0.026375
1111111000	0.024500
0001111111	0.024172
1111110000	0.020469
0000111111	0.020453
1111100000	0.017328
0000011111	0.017172
0000001111	0.015750
1111000000	0.015422
0000000111	0.014313
1110000000	0.014000
0000000011	0.013266
1100000000	0.012906
0000000001	0.012375
1000000000	0.011984

Table 1(b). Voicing Pattern Likelihoods For 30 ms LPC Voicing Data

```
Spectral Codebook     : 13 bits
Energy   Codebook     : 12 bits
Nominal  Energy       :  6 bits
Pitch    Codebook     : 10 bits
Nominal  Pitch        :  5 bits
Voicing  Codebook     :  7 bits
Frame    Interpolation:  3 bits (1 Frame Per Block)
```

Table 2(a). Bit Allocations For An 800 Bits/s System

```
Spectral Codebook     : 10 bits
Energy   Codebook     : 11 bits
Nominal  Energy       :  4 bits
Pitch    Codebook     :  9 bits
Nominal  Pitch        :  4 bits
Voicing  Codebook     :  8 bits
Frame    Interpolation:  4 bits (2 Frames Per Block)
```

Table 2(b). Bit Allocations For A 400 Bits/s System

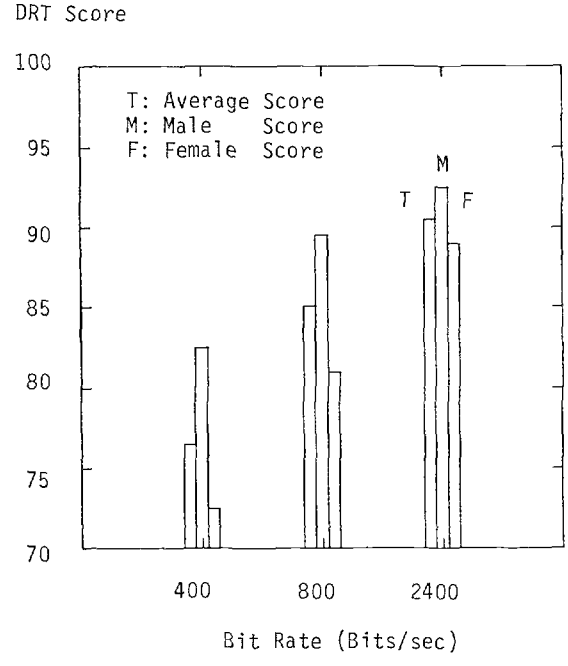

Fig. 1. A Comparison Of Speech Intelligibility As Measured By The Diagnostic Rhyme Test