VARIABLE FRAME SIZE FOR VECTOR QUANTIZATION

Carlos Moreno and Fabrice Labeau

McGill University
Department of Electrical & Computer Engineering
Montreal, Canada

ABSTRACT

Vector Quantization (VQ) is a lossy data compression technique that is often applied in the field of speech communications. When applying VQ to speech signals, usually combined with some for of Linear Prediction Coding (LPC), the input signal is divided into *frames* of a given length. Each frame is then processed and quantized using VQ. A typical assumption is the fact that the frame size is fixed. In this study, we propose a modification to this technique that allows for variable size frames, providing an additional degree of freedom for the optimization of the encoding process. The results show that this technique effectively improves the quality of the encoded signal at a given bit rate, even if the improvement is not dramatic.

1. INTRODUCTION

Vector Quantization (VQ) is a lossy data compression technique that is often applied in the field of speech communications [1]. In VQ encoding, a group of values or *vector* is replaced with a codeword that is chosen following some optimality criterion — usually the closest vector from a list of possible values, the *codebook*. In the case of speech signals, VQ is usually combined with some form of Linear Predictive Coding (LPC) technique. The signal is divided into frames of a given length, and each frame is processed and encoded using VQ.

Even though VQ has proved extremely effective in many practical applications, some opportunities for potential improvement may have been neglected. In particular, a typical assumption is the fact that the frame size is fixed. Little attention has been paid to the possible benefits that could derive from encoding with variable frame size, and it has been often considered an unnecessary complexity [1].

A technique using variable analysis frame sizes has been proposed for the coding of multiband excitation model parameters [2]. In that study, however, the encoding allowed for variable frame size under certain conditions, to ensure stationary spectral parameters of the signal within frames.

Other studies have focused on variable rate VQ. One such study reports success in using variable bit allocation for Line-Spectrum Pairs and generalized spectral distributions, taking advantage of the relative entropy in these parameters [3]. Variable dimension spectral vectors has also received some attention — vectors of harmonic spectral peaks have variable dimension to optimize the bit rate according to the current spectral characteristics [4].

Another important class of related techniques where the use and optimization of VQ has been sought is the Analysis-by-Synthesis (ABS), in particular, Code-Excited Linear Prediction (CELP) techniques [5]. In the context of CELP, variable rate quantization has been applied with success [6, 7].

All of these ideas, however related to what we propose in this study, still do not consider the possibility of an additional level of optimization that could result from using variable frame size VQ.

In this study, we propose a modification to the VQ technique that allows for variable size frames, providing an additional degree of freedom for the optimization of the data compression process. The *quantization error* goes through a first level of minimization by choosing the closest point in the codebook for the given frame. We now minimize this by choosing the frame size that yields the lowest quantization error — notice that the quantization error is a function of the given frame and the codebook; by considering different frame sizes, we get different actual frames that yield different quantization errors, allowing us to choose the optimal size. This effectively provides a second level of optimization.

This idea raises an important uncertainty: additional bits are required on a per-frame basis, to encode the selected frame size. It is not immediately obvious whether these bits would be better spent in increasing the codebook size, which in turn decreases the average quantization error.

2. VARIABLE FRAME SIZE

There are several aspects that lead us to think that there may be important benefits resulting from variable frame size in VQ. First, we have the fact that interaction between consecutive frames is fairly complex. This means that the artifacts introduced in the transition from one distorted frame to the next (also distorted) frame could be dramatically reduced by an optimal choice of the frame size, which in turn determines the exact position for the transition between the frames.

Also, extreme values in the quantization errors can have a severe effect in the case of VQ, since these affect the entire frame. In the case of speech signals, this has an interesting implication: the points where the short-term spectral characteristics change (such as transitions between a vowel and a consonant) are critical for speech intelligibility; if one of these transitions falls in the middle of a frame, then the quantization error will be large, and the distortion for that *critical* frame will be extremely high.

An optimal choice of the frames sizes — and thus positions — could have a noticeable impact in the quality and intelligibility of the reconstructed speech signal, as it would naturally tend to avoid *critical* frames with large quantization errors.

3. EXPERIMENTAL SETUP

We implemented an LPC + VQ setup with the standard fixed-size frames technique, and with our modification, allowing for variable-size frames. The optimality criterion used to determine the codewords for the various components was based on an Analysis-by-Synthesis (ABS) strategy, using the square error as the objective function to be minimized.

In the case of fixed-size frames, we proceed as follows: for each frame, choose the filter codeword that minimizes the distance in the p-dimensional space corresponding to the Line-Spectral Frequencies (LSF) of the filter; we then compute the residual (using the quantized filter), split it into a fixed number of chunks (we used 8 chunks per frame), and for each chunk, we choose the codeword that minimizes the square error when reconstructing the signal for the segment corresponding to the given chunk of the residual.

The encoding of each frame involves applying VQ to two components of the speech signal: the LSF representing the prediction filter, and the residual. In the case of the residual, we use 8 codewords, encoding each of the residual chunks.

For the case of the variable-size frames, we proceed exactly as described above for the fixed-size case, except that we repeat the procedure for each possible frame size, and we choose the size that minimizes the square error *per sample* after reconstruction.

The main objective of the tests is to compare the modified technique vs. the conventional fixed-size frame technique at the same bit rate. This addresses the issue that when implementing the variable-size technique that we propose, additional bits are required for the encoding process on a per-frame basis. One of the uncertainties that we tried to address is the issue of whether or not those bits are better spent in encoding the frame size, or spent on increasing the size of the codebook, which in turn leads to reducing the average quantization error.

When applying the variable-size frame technique, we considered the cases where we have 8 and 16 possible sizes, requiring 3 and 4 additional bits, respectively, to encode the frame size. The codebook sizes were between 10 and 14 bits. Table 1 summarizes the tests and comparisons done, indicating number of bits assigned to each part of the setup, and the actual bit rate, taking into account the various parameters used in each case.

Table 1. Number of Bits for Codebooks and for the Encoding of the Frame Size. *Cbk.* is the number of bits for the codewords; *Fr. size* is the number of bits to encode the frame size — 0 indicates fixed-size frames.

Bits (Cbk. / Fr. size)	Chunks (Residual)	Bits per Frame	Bit Rate (bits/sec)
14 / 0	8	126	6.300
13 / 0	8	117	5850
12 / 0	8	108	5400
11 / 0	8	99	4950
13 / 4	8	121	5900
12 / 4	8	112	5460
11 / 4	8	103	5025
10 / 4	8	94	4585
13 / 3	8	120	6000
12 / 3	8	111	5550
11 / 3	8	102	5100
10 / 3	8	93	4650

In the table, Bits per Frame accounts for the eight codewords of the residual chunks plus one for the prediction filter. Bit rate is obtained as the number of bits per frame times the average frame size (see table 2).

Table 2 shows the values for the frame sizes, including the size for the fixed-size case, and the choices of frame sizes for the cases of 3 and 4 additional bits to encode the sizes.

Table 2. Frame Sizes for the Experimental Setup.

Bits	Sizes	Avg. Size
0	160	160
3	104, 120, 136, 152, 168, 184, 200, 216	160
4	104, 112, 120, 128, 136, 144, 152, 160, 168, 176, 184, 192, 200, 208, 216, 224	164

The list of available sizes for the variable-size case were chosen in arithmetic progression, centered around the size used for the fixed-size case (if possible), and such that all the values are divisible by 8 (so that the residual chunks fit exactly).

For each of the configurations, a test audio file was run through the encoding and decoding process, and the Signal-to-Noise Ratio (SNR) of the output was computed (the noise was computed as the difference between the output signal and

the input signal). Also, the Segmental SNR was computed, using 20 ms segments.

4. RESULTS

The results consistently show an improvement in the quality of the processed speech when using our proposed technique. The comparisons take into account the bit rate — we consider the quality of the processed speech by the two methods as a function of the bit rate.

The results and comparisons are presented in two parts: the measurable and objective parameters for both techniques, and the subjective comparison that we performed by listening to the processed audio files for both techniques.

4.1. Objective Measurable Parameters

Table 3 shows the results with the details of the corresponding configurations and bit rates.

Table 3. Signal-to-Noise Ratio (SNR) After Reconstruction.

Bits (Cbk. / Fr. Sz.)	SNR	Seg. SNR
14 / 0	6.60	4.47
13 / 0	6.32	4.16
12 / 0	6.12	3.99
11 / 0	5.81	3.64
13 / 4	7.24	5.03
12 / 4	7.06	4.41
11 / 4	6.31	4.19
10 / 4	6.04	3.39
13 / 3	7.04	4.96
12 / 3	6.80	4.51
11 / 3	6.20	4.18
10 / 3	5.73	3.66

Figure 1 shows a graphical display comparing the Signal-to-Noise Ratio (SNR) for fixed size against the results for variable size.

Figure 2 shows the Segmental SNR (with 20 ms segments) for the same configurations. The fact that the curves cross at approximately 5.4 kbps is perhaps a little surprising. However, we have to keep in mind that, even though the technique being used is the same, using different configuration parameters (in this case, different amounts of choices for the frame size) effectively leads to two separate classes of encoding systems. Each of the two configurations may exhibit benefits that have a more noticeable impact at different operating parameters.

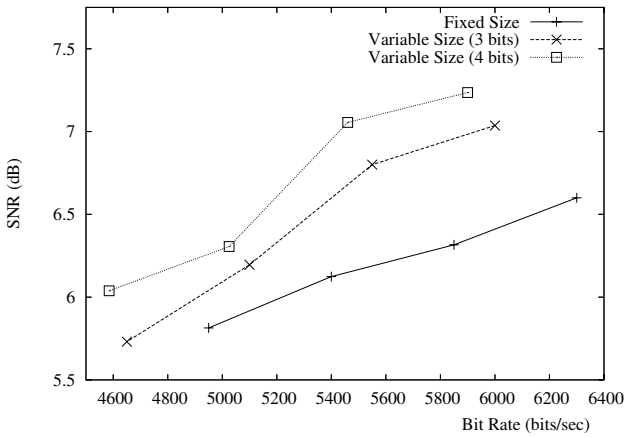

Fig. 1. Signal-to-Noise Ratio (SNR) vs. Bit Rate.

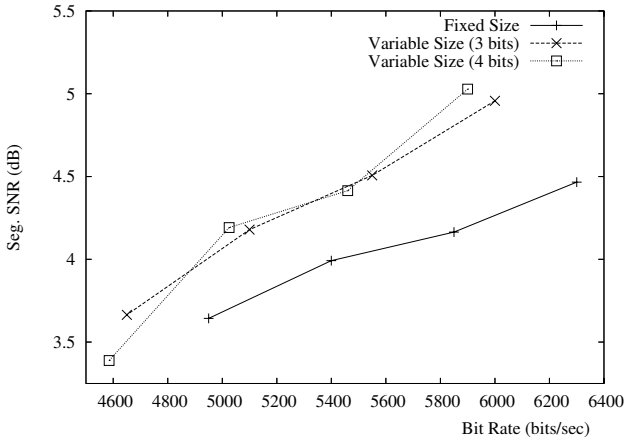

Fig. 2. Segmental SNR (20 ms Segments) vs. Bit Rate.

4.2. Perceptual Evaluation

In addition to the measurable parameters presented in section 4.1, we listened to the processed speech samples to try to confirm the conclusions that were drawn from the measurable results. Our perceptual comparison does match the results, in that in most cases, we do hear a slightly higher quality for the processed audio files when using the variable-size frames technique proposed in this study.

All of the processed audio files do suffer from a slight "scratchy" quality, which is the result of an encoding technique that is not optimized using perceptual criteria and the relatively low bit rate that we used.

However, we did perceive an increased level of "scratchiness" for the fixed-size frames audio files, and an increased number of artifacts. These artifacts are particularly noticeable in the form of "pops" that follow certain explosive phonemes such as stop consonants. This is consistent with our previous analysis, since those explosive sounds correspond to sharp

transitions in the short-term spectral characteristics. These transitions are not well handled by a technique that is based on processing entire frames of speech under the assumption that the spectral properties are quasi-stationary. The encoding errors tend to have large values in these transition frames.

Our subjective evaluation was confirmed by an audio analysis software designed to measure the perceptual degradation of the signal, following the ITU-R BS.1387 (PEAQ) standard [8].

Figure 3 shows a plot of the grades reported by the Perceptual Evaluation of Audio Quality (PEAQ) software.

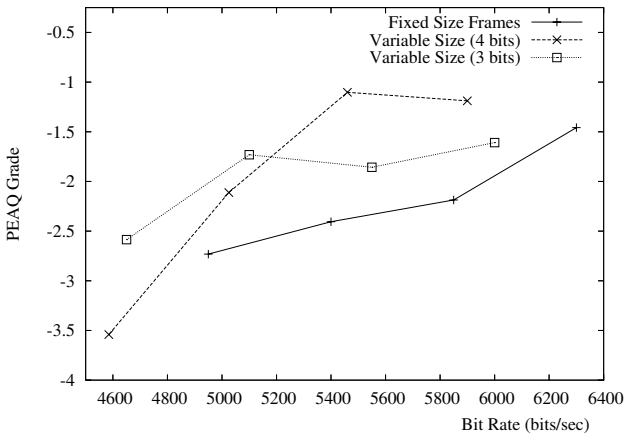

Fig. 3. Perceptual Evaluation of Audio Quality (PEAQ) Grades for the Processed Speech Samples.

5. CONCLUSIONS

In this study, we proposed and evaluated a modified VQ technique in which we allow variable size frames for the encoding of speech signals. The idea can be applied in many contexts where VQ is applicable, but this study focused on using VQ to encode speech signals in the context of an LPC setup. The results from our tests clearly show an improvement in the quality of the encoding method when compared to the standard technique at the same bit rate, even if the difference was not dramatic. The quantitative measures that we used to compare both methods were the Signal-to-Noise (SNR) ratio and the Segmental SNR for the reconstructed signal.

In addition to these objective parameters, we listened to the processed speech segments to establish — in an informal manner — a perceptual, subjective evaluation of the quality of both methods. The results from our subjective evaluation match the quantitative, objective results that we measured; the reconstructed speech using the standard, fixed-size frames technique exhibited an increased level of noise and an increased number of artifacts, mostly noticeable in the form of "pops," often synchronized with the explosive portions of consonants like P or T.

As a final remark in terms of conclusions, it is important to highlight that the improvement that we observed was not dramatic. One of the potential criticisms to this proposed technique is the considerable increase in complexity — at least the way that it was implemented — with respect to the standard technique. It could be argued that such increase in complexity and computing-power requirements for a practical implementation might be justified only in exchange for a dramatic increase in performance.

6. REFERENCES

[1] Douglas O'Shaughnessy, *Speech Communications, Human and Machine*, IEEE Press, second edition, 2000.

[2] Eric W.M. Yu and Cheung-Fat Chan, "Robust multiband excitation coding of speech based on variable analysis frame sizes," *EUSIPCO-96*, 1996.

[3] Stan A. McClellan and Jerry D. Gibson, "Variable rate vector quantization of the speech spectral envelope," *Proceedings of the IEEE Southeastcon '96*, pp. 208 – 215, Apr. 1996.

[4] Philip A. Chou and Tom Lookabaugh, "Variable dimension vector quantization of linear predictive coefficients of speech," *ICASSP-94*, vol. I, pp. I505 – I508, Apr. 1994.

[5] Manfred R. Schroeder and Bishnu S. Atal, "Speech coding using efficient block codes," *ICASSP-82*, vol. 7, pp. 1668–1671, May 1982.

[6] Jerry D. Gibson, Myron L. Moodie, and Stan A. McClellan, "Variable rate techniques for CELP speech coding," *29th Asilomar Conference on Signals, Systems and Computers*, vol. 2, pp. 1219 – 1224, Nov. 1995.

[7] Lei Zhang, Tian Wang, and Vladimir Cuperman, "A CELP variable rate speech coded with low average rate," *ICASSP-97*, vol. 2, pp. 735 – 738, Apr. 1997.

[8] Peter Kabal, "An examination and interpretation of ITU-R BS.1387: Perceptual Evaluation of Audio Quality," Tech. Rep., TSP Lab Technical Report, Dept. Electrical & Computer Engineering, McGill University, May 2002.