PITCH QUANTIZATION IN LOW BIT-RATE SPEECH CODING

Thomas Eriksson[*] *and Hong-Goo Kang*

AT&T Labs-Research, SIPS
180 Park Avenue, Florham Park, NJ 07932
eriksson@speech.kth.se goo@research.att.com

ABSTRACT

This paper describes a new pitch quantization method for low bit-rate speech coding systems. The logarithm of the pitch period is quantized in a combination of two uniform quantizers, one working directly on logarithmic pitch values and the other working on the difference between current and previous logarithmic pitch. The best of the two output values is transmitted to the receiver. This scheme can exploit both redundancy in the signal and properties of the ear to achieve an efficient quantization. Listening tests show that the proposed scheme allows the pitch parameter to be quantized using 4 bits, with no degradation in audible quality.

1. INTRODUCTION

In typical speech coding applications, the quantization of *pitch period* or *pitch frequency* requires the use of 7 or 8 bits per pitch sample for accurate representation. In many low bit-rate algorithms, such as waveform interpolation coders [1] and the MELP coder [2], the quantization of pitch uses a large proportion of the overall bit-rate (up to as much as 25 %).

Even though a good pitch quantization scheme is crucial for low bit-rate speech coding, the subject has not been studied previously in as much as detail as other parameters (i.e. line spectral frequencies). Most previous research efforts were concentrated to quantization of the long term prediction lag (LTP lag, or LTP delay) in a CELP coding context.[1] For example, in [3] *restrictive pitch deviation coding* is proposed, in which an average LTP lag for an entire frame is computed, and then an optimal pitch lag is found for each subframe within some predefined offset limits (typically a few samples from the pitch of the entire frame). This method is refined in [4], where the restrictive coding is applied only on voiced frames. In [5] the differential LTP lag in a CELP coder is Huffman coded, and the saved bits are used to enhance the innovation codebook. This has the advantage that there is no restriction on the range of LTP delays. The authors of [6] determine a subset of the most probable LTP lags, and the optimal LTP search is then restricted to this subset. In contrast to [3], the subset is not a simple neighborhood of the previous pitch value, and no average frame pitch is used.

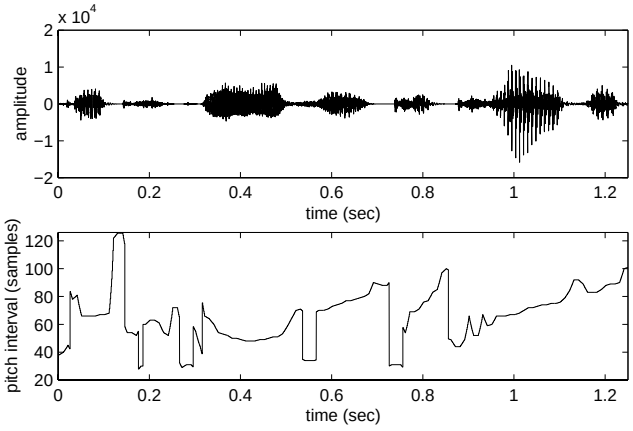

Figure 1: Waveform and pitch contour for a male speaker saying 'it's easy to tell the'. Note the slow evolution of the pitch in voiced segments, interupted by sudden jumps in unvoiced segments.

In this paper, we propose a differential and logarithmic quantization scheme for pitch period or pitch frequency. The logarithm of the pitch is quantized both in a differential and a memoryless quantizer, and the best of the two output values is transmitted to the receiver. The major advantage of this scheme is that the high correlation of consecutive pitch values during voiced speech segments can be exploited, without losing performance for unvoiced speech, where consecutive pitch values has low correlation. Another advantage is that the proposed scheme quantizes rapidly and slowly changing pitch in separate quantizers, which allows for different resolution for these two cases. It is shown in [7] that the human ear is much more sensitive for pitch changes in stable segments, and the proposed scheme can exploit this by a high resolution in the differential quantizer, which typically quantizes slowly changing pitch. Also, the logarithmic quantization is advantageous since relative pitch change is more important than absolute pitch change. With the proposed scheme, the pitch can be quantized with only 4 bits without showing audible artifacts in the reconstructed speech signal. Results from listening tests confirming this conclusion are provided.

In Section 2, properties of pitch are discussed, and in Sec-

[*]Thomas Eriksson is currently at the Department of Speech, Music and Hearing, KTH (Royal Institute of Technology), 100 44 Stockholm, Sweden.

[1]Note that there is a major difference between quantization of pitch in parametric coders (such as sinusoidal or waveform coders) at one hand, and quantization of pitch in a waveform coder (i.e. pitch lag in a CELP coder) at the other hand. In parametric coders, the quantization of pitch is fairly orthogonal to quantization of other parameters, while this is generally not true in a CELP coder; the quantization of the LTP lag affects the quantization of the innovation. The scheme proposed in this paper is designed for coders with orthogonal pitch and waveform quantization.

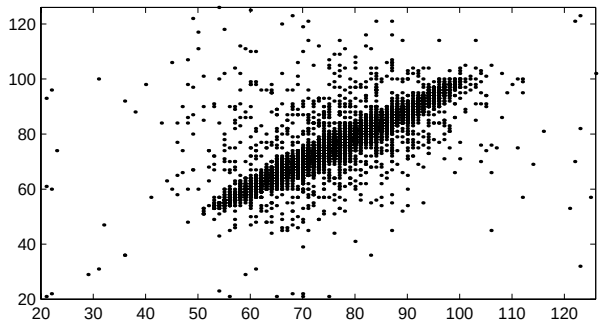

Figure 2: Scatter plot of two consecutive pitch periods for a male speaker.

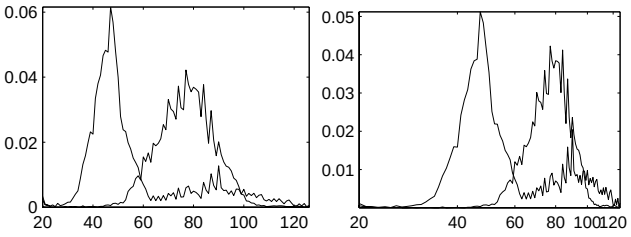

Figure 3: Histogram of pitch periods for female (max value at around 50 samples) and male (max at 80 samples) speakers. Left: linear pitch scale. Right: logarithmic pitch scale.

tion 3 preliminary listening tests to determine the sensitivity of the human ear to pitchare presented. Section 4 introduces a new scheme for pitch quantization with 4 bits per sample, and Section 5 presents a listening test to verify the quality of the new scheme.

2. PROPERTIES OF THE PITCH

In this section, we investigate temporal and statistical properties of pitch intervals. In Figure 1, a pitch period contour is depicted.[2] From the figure, it is clear that the pitch trace has several interesting characteristics. We see that consecutive pitch period values are highly correlated most of the time. This is also illustrated in Figure 2, where a scatter plot of consecutive pitch period values is depicted.

Histograms can also reveal a lot about the relevant features of the pitch. In Figure 3, histograms of both a male and a female speaker are given. The left peak in the histogram (shorter pitch period) corresponds to female speakers, and the right peak to male speakers. Note that the male peak is much wider in the plot with linear pitch scale. However, the relative variation is almost the same for male and female speaker. This can be seen in the plot with logarithmic pitch scale, where the peaks are approximately of equal width. The histograms in Figure 3 are derived from only one male and one female speaker, but they are typical over a wide range of speech; histograms for other speakers are similar.

[2] All the pitch estimations in this paper are based on the pitch estimator proposed in [8]. This algorithm gives fairly stable pitch values, and it has built-in functionality to reduce pitch doubling and halving.

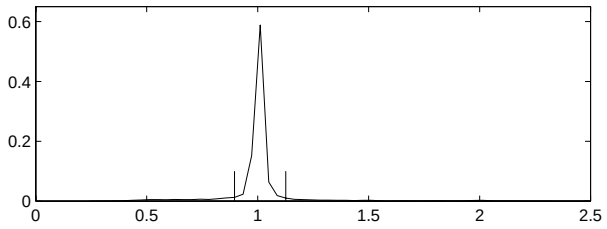

Figure 4: Histogram of the ratio between two consecutive pitch periods. A ratio between 0.9 and 1.11 (the lines) is classified as belonging to the high-correlation group.

3. PRELIMINARY LISTENING TESTS

In this section we report on listening tests to determine how sensitive the human ear is to pitch quantization. The listening tests were performed using synthetic speech from a waveform interpolation (WI) speech coder [9], with all parameters except pitch quantized. The encoder determines a new pitch value every 20 ms, rounds it off to the desired resolution, and sends it to the decoder. The resulting synthetic speech is presented to the listeners. The WI coder with unquantized pitch gave a close to transparent quality of the synthesized speech signal, which allowed distortion due to pitch quantization to be easily detected. Six experienced listeners were presented sentences from four speakers. The subjects listened to several versions of the same sentence, first the original WI output with unquantized pitch (1 sample resolution), and then versions with lower pitch resolutions. The subjects were then asked to indicate at what resolution they start to hear artifacts. The procedure was repeated for four different sentences.

3.1. Male and female pitch resolution

In this experiment, we tried to estimate the necessary pitch resolution for two male voices with average pitch period of 80 samples, and for two female voices with average pitch period of 50 samples (8 kHz sampling frequency). The pitch estimation algorithm in the coder gives the pitch period with a resolution of 1 sample, and this value is then rounded off to give a resolution of 2,3,4,5,6 and 8 samples. Results from this experiment indicated that a pitch resolution of 3 samples is enough for the female speakers. For the male speakers, a pitch resolution of 5 samples is required. This result indicates that voices with long pitch period needs lower absolute resolution for transparent quantization. The conclusion we draw from this listening test, and from the histograms in Figure 3, is that logarithmic pitch quantization may be preferable to standard uniform quantization of the pitch period.

3.2. Resolution for pitch samples with high and low correlation

In the second test, we subdivide the pitch samples into two groups: those with high temporal correlation (with a value close to the previous value) are sorted into one group, and the rest, i.e. samples with low temporal correlation, are sorted into another group. To determine which group a pitch sample belongs to, the ratio between present and previous pitch sample is studied. If this ratio is between 0.9-1.11, the pitch is classified into the high correlation

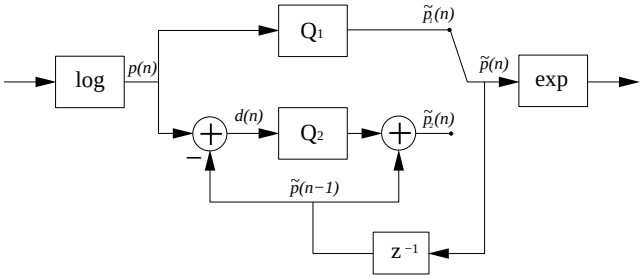

Figure 5: Block diagram of the proposed pitch quantization scheme. A uniform memoryless quantizer, Q_1 is used in parallel with a differential quantizer (the lower branch, including the uniform quantizer Q_2).

group. Otherwise the pitch is classified into the low correlation group (see Figure 4).[3]

As in the experiments with male and female speakers in the previous subsection, the results indicated that for the high correlation group, the necessary pitch resolutions for male and female speakers are 5 and 3 samples respectively. For the low-correlation group, much lower resolution is tolerable. For females, a pitch resolution of 10-15 samples was enough and for the male voices, with longer pitch period, 20 samples resolution still gave transparent quality. We conclude that the low-correlation and the high-correlation group should be encoded with different resolution, and in the next section we suggest a combination of a differential and a memoryless quantizer that separately quantizes rapidly and slowly changing pitch.

4. PITCH QUANTIZATION

The comparison between the necessary resolution for a male and female voice in the previous section, and the histogram in Figure 3, motivate us to propose a logarithmic quantization scheme. Furthermore, the long runs of highly correlated pitch samples interrupted by sequences with low correlation, as is obvious from Figure 1 and 2, suggests that a combination of a predictive quantizer and a memoryless quantizer might be a good solution.[4]
In Figure 5, a block diagram of the proposed quantization scheme is depicted. The function of the algorithm is as follows: First the logarithm of the pitch is computed. The logarithmic pitch value is used in two branches of the algorithm. In the first branch, the pitch is directly quantized in a uniform quantizer. In the second branch, the previous value of the quantized pitch is subtracted before quantization, and added back after quantization, to form a differential quantization scheme. The output of the two branches are compared to the unquantized pitch, and the best is selected for transmission. To design the quantizer for the predictive branch, we recall that an absolute resolution of 5 samples (see Section 3) is desirable for a voice with average pitch period 80 samples. This corresponds to a logarithmic pitch resolution of $\log 85 - \log 80 = \log 85/80 = \log 1.0625$. For the females in the test, the necessary resolution is $\log 53/50 = \log 1.06$, which is very close to the resolution for male speakers. Therefore we chose the step size for the quantizer for the high-correlation group to be $\log 1.06$. A 5-step uniform quantizer, with levels $\log 1.06 \cdot \{-2, -1, 0, 1, 2\}$, is enough to cover the interval for the high-correlation group $\log 0.9 ... \log 1.11$, as discussed in Section 3.2. With a 4 bits/sample pitch quantizer, we have 11 levels left to use for the low-correlation group, which is enough if a standard pitch range of 20-147 samples is used. The full algorithm is given in Table 1.

[3]The reason for using pitch ratios instead of differences for the classification is the same as the reason for logarithmic quantization; relative changes are more relevant than absolute.

[4]The combination of a predictive and a memoryless quantizer has been previously studied in e.g. [11]. It was proposed for quantization of pitch in [10].

Table 1: The pitch quantization algorithm

1. Initialization (this is only done the first call)
   $p_{\min}$ and $p_{\max}$ are the minimum and maximum pitch period, respectively.

   $p_{\text{range}} = \log p_{\max} - \log p_{\min}$
   (the range of logarithmic pitch values)
   $\mathcal{E}_1 = p_{\text{range}}/10 \cdot \{0, 1, 2, ..., 10\} + \log p_{\min}$
   (11 entries, index 0-10)
   $\mathcal{E}_2 = \log 1.06 \cdot \{-2, -1, 0, 1, 2\}$
   (5 entries, index 11-15)
2. Get an estimated pitch value $P(n)$, and compute the logarithmic pitch, $p(n) = \log P(n)$
3. Find the closest value to $p(n)$ in $\mathcal{E}_1$, $\tilde{p}_1(n) = \text{argmin}_{c \in \mathcal{E}_1} |p(n) - c|^2$
4. Find the closest value to $d(n) = p(n) - \tilde{p}(n-1)$ in $\mathcal{E}_2$, $\tilde{p}_2(n) = \text{argmin}_{c \in \mathcal{E}_2} |d(n) - c|^2 + \tilde{p}(n-1)$
5. Compare $\tilde{p}_1(n)$ and $\tilde{p}_2(n)$ to $p(n)$, and select the best, $\tilde{p}(n) = \min(\tilde{p}_1(n), \tilde{p}_2(n))$
   The index to the selected codebook entry (see definition of $\mathcal{E}_1$ and $\mathcal{E}_2$ for index assignment) is output.

It can be argued that the pitch estimation algorithm during a fairly stable segment of speech might give a pitch period that jumps up and down with small amplitude, and that the amplitude of this oscillation might be amplified to several samples with the new scheme. This rapidly oscillating pitch would probably be clearly noticable. If problems like those above are encountered, we propose the use of hysteresis in the algorithm, so that oscillotary behaviour is prohibited. A simple extra step to the pitch quantization algorithm is given in Table 2.

Table 2: An extra step in the pitch quantization algorithm, to prohibit oscillation.

6. Prohibit oscillation
   (see Table 1, step 1, for index assignment)
   IF (previous index was 12 and current index is 14) OR
   (previous index was 14 and current index is 12) THEN
   current index is set to 13
   END IF

However, as will be clear in Section 5, the algorithm worked fine with the test material we used, and no problems of this kind were noticed. Note that even though the listening test were performed using only 40 sentences from 20 speakers, our listening experience is much wider than that.

Table 3: A-B test; the numbers indicate listener preference

	female	male
unquantized	52%	51%
quantized	48%	49%

An important consideration when differential quantization schemes are used is the performance for transmission over noisy channels. Since the differential quantization scheme has memory, a single channel error leads to a sequence of faulty indices at the decoder (error propagation). However, the proposed combination of a differential and a memoryless quantizer suffer much less from error propagation than a purely differential scheme, and it has proven reliable in other applications [12]. Even though no experiments with noisy channels have been performed, we believe that also for noisy channels, the subjective performance for the proposed scheme is comparable to the performance of a standard memoryless pitch quantizer. Note that the quantized pitch is only used in the decoder, not in the encoder. The encoder still relies on the unquantized pitch, since high pitch resolution is generally necessary for the analysis stage.

5. LISTENING TEST

To verify the performance of the pitch quantization scheme proposed in Section 4, we present a listening test in this section. The test setup is the same as in the preliminary listening tests in Section 3, using synthetic speech from an unquantized waveform interpolation speech coder. 15 test persons (10 experienced, 5 inexperienced) listened (using headphones) to 40 short sentences from 10 male and 10 female speakers, encoded by the (unquantized) WI coder with and without the proposed pitch quantization scheme. The listeners were presented pairs of sentences, with random order of the unquantized and quantized version, and the test persons were asked to indicate a preference for either the first or the second sentence.

The listening tests revealed that there is no significant difference in perceived quality between unquantized sentences (7 bit pitch resolution) and sentences quantized with the proposed 4-bit pitch quantization scheme. The general opinion among the test subjects was that it was very difficult to hear any difference between the quantized and unquantized versions, and in the cases when the test persons did hear a difference, the quantized sentence was not always perceived as worse. Table 3 depicts the results from the A-B listening test. The absolute pitch period resolution for typical women (high pitch frequency) and men (low pitch frequency) is different because of the logarithmic quantization scheme, and therefore the results for male and female speakers are separately reported.

6. SUMMARY

We propose a pitch quantization algorithm based on logarithmic pitch values, quantized in a combination of two uniform quantizers, one working directly on the logarithmic pitch values and the other working on the difference between current and previous logarithmic pitch. The complexity of the proposed scheme is very small, since only two uniform quantizers are involved. Preliminary listening tests to investigate the perception of pitch and the necessary resolution for transparent pitch quantization are presented. These preliminary tests illustrate the advantages with logarithmic and differential quantization, and thus motivate the algorithm. In a subsequent listening test, the performance of the proposed quantization scheme is evaluated, and the results show that transparent pitch quantization can be achieved with 4 bits/sample.

7. REFERENCES

[1] W. B. Kleijn and J. Haagen, "Waveform interpolation for speech coding and synthesis," W. B. Kleijn and K. K. Paliwal, editors, *Speech Coding and Synthesis*, pp. 175-208, Elsevier Science Publishers, Amsterdam, 1995.

[2] A. V. McCree, K. Truong, E. B. George, T. P. Barnwell, III, and V. Viswanathan, "A 2.4 kbit/s MELP coder candidate for the new U.S. federal standard," *Proc. Int. Conf. Acoust. Speech Sign. Process.*, Atlanta, USA, 1996.

[3] M. Yong and A. Gersho, "efficient encoding of the long-term predictor in vector excitation coders," in *Advances in speech coding*, B. S. Atal, V. Cuperman, and A. Gersho, Eds.: Kluwer Academic Publishers, 1991, pp. 329-338.

[4] K. Ozawa, M. Serizawa, T. Miyano, T. Nomura, M. Ikekawa, and S.-I. Taumi, "M-LCELP speech coding at 4 kb/s with multi-mode and multi-codebook," *IEEE Transactions on Communications*, vol. 77, pp. 1114-1121, 1994.

[5] T. Eriksson and J. Sjoberg, "Dynamic bit allocation in CELP excitation coding,",*Proc. Int. Conf. Acoust. Speech Sign. Process.*, Minneapolis, 1993.

[6] A. Popescu, N. Moreau, and C. Lamblin, "A differential encoding method for the LTP delay in CELP coders," *Eurospeech '95*, Madrid, Spain, 1995.

[7] E. Zwicker and H. Fastl, "Pitch and pitch strength", in *Psychoacoustics*, pp. 103-132, Springer-Verlag, New York, 1990.

[8] W. B. Kleijn, P. Kroon, L. Cellario and D. Sereno, "A 5.85 kb/s CELP algorithm for cellular applications," *Proc. Int. Conf. Acoust. Speech Sign. Process.*, Minneapolis, vol. II, pp. 596-599, 1993.

[9] W. B. Kleijn, Y. Shoham, D. Sen and R. Hagen, "A low-complexity waveform interpolation speech coder," *Proc. Int. Conf. Acoust. Speech Sign. Process.*, Atlanta, vol. I, pp. 212-215, 1996.

[10] T. Eriksson, J. Linden, and J. Skoglund, "A safety-net approach for improved exploitation of speech correlations," *Proc. Int. Conf. on Digital Signal Processing*, Cyprus, 1995.

[11] T. Eriksson, J. Linden, and J. Skoglund, "Exploiting interframe correlation in spectral quantization - A study of different memory VQ schemes,"*Proc. Int. Conf. Acoust. Speech Sign. Process.*, Atlanta, USA, 1996.

[12] T. Eriksson, J. Linden, and J. Skoglund, "Interframe LSF quantization for noisy channels," To appear in *IEEE Transactions on Speech and Audio Processing*.