VARIABLE RATE SPECTRAL QUANTIZATION FOR PHONETICALLY CLASSIFIED CELP CODING

Roar Hagen†, Erdal Paksoy†† **, and Allen Gersho††

†Dept. of Information Theory
Chalmers University of Technology
S-412 96 Goteborg, Sweden

††Dept. of Electrical and Computer Engineering
University of California, Santa Barbara, CA 93106 U.S.A.

ABSTRACT

Variable rate quantization of the linear predictive coding (LPC) parameters based on phonetic classification of the speech frame results in substantial performance gain. Speech frames are classified as unvoiced or voiced and are separately quantized with VQ codebooks designed for each class. Performance results, including listening tests, show that for transparent quality roughly 9 bits is sufficient for unvoiced frames and 24 bits for voiced frames. Test results of LPC quantization are described for a variable rate phonetically segmented CELP coder and for the synthesis of speech from the prediction residual.

1. INTRODUCTION

Efficient quantization of the LPC parameters is a critical issue for low-rate speech coding. In almost all coding schemes studied today, a fixed allocation of bits for each frame is assigned to the LPC parameters. In this paper, we examine variable rate coding of the LPC parameters based on phonetic classification of each frame and we report on the resulting performance advantages that are attained.

"Transparent" quality spectral quantization is said to be achieved if the coding of the LPC parameters does not introduce any audible distortion into coded speech. Based on the *spectral distortion* (SD) between the spectra obtained from quantized and unquantized LPC coefficients, Paliwal and Atal [1] have given the following requirements for transparent quality LPC quantization: 1) average SD below 1 dB, 2) no outlier frames with SD larger than 4 dB, and 3) less than 2% of outliers with SD in the range 2-4 dB. In the usual fixed-rate coding of speech, the transparency criteria are evaluated by averaging across all speech frames, regardless of their varying phonetic character.

Several recent coders employ *multi-mode* coding, where some form of phonetic classification is performed on each frame and the coding scheme and bit allocation for the frame is class-dependent. In the simplest cases, the classification distinguishes only between voiced and unvoiced speech. In *variable-rate* coders, the total bit quota for the frame also varies with time. All such multi-mode or variable-rate schemes can readily be modified by including class-dependent LPC quantization schemes.

The quantization of the LPC parameters for a phonetically specific class of speech has received very little attention. Liu and Hoege [2] have reported on experiments with wideband speech synthesized with an unquantized excitation. Wang and Gersho introduced a multi-mode coder based on phonetic classification with different CELP coding schemes [3]. They noted that considerably fewer bits are required for unvoiced spectra compared to voiced spectra and they allocated 10 bits for unvoiced speech and 24 bits for voiced speech when each frame is coded in isolation, (i.e., without interframe coding). Here we specifically study variable rate quantization of class-specific LPC coefficients for a CELP coder that employs a phonetic classifier and we examine the coding advantage that can be obtained by designing and training class-specific codebooks for voiced and unvoiced classes of speech, the bit-rate for LPC parameters needed to achieve the Paliwal-Atal criteria for "transparent quality", and the SD performance needed to achieve subjectively distortion-free spectral quantization for different classes.

2. PHONETICALLY CLASSIFIED SPECTRAL QUANTIZATION

As a specific framework for assessing performance of phonetically classified spectral quantization, we use the VRPS coder [4] , [5] which is a variable-rate, CELP-based coder where frames of the input signal are

This work was supported in part by the National Science Foundation under grant number 8914741, Fujitsu Laboratories Ltd., the UC Micro program, Rockwell International Corporation, Qualcomm, Inc., Echo Speech Corporation, Signal Technology, Inc., Hughes Aircraft Company, and Lockheed Missile and Space Company.

** Erdal Paksoy is currently with Echo Speech Corporation, 6460 Via Real, Carpinteria CA 93013.

classified into one of four main categories: *non-speech* (background noise), *unvoiced speech*, *voiced speech*, and *onsets*. We do not consider non-speech segments in this study. We focus only on spectral quantization of the speech signal itself. Also, for this study, we treat the perceptually important onset frames as members of the voiced class. Thus, we include them in the class with the most demanding requirements on accurate reproduction. Therefore, we assign each LPC spectrum of an active speech frame to one of two classes: unvoiced or voiced.

Voicing is the most relevant phonetic feature for LPC quantization in CELP coders. Voiced sounds, notably vowels, have most of their energy in the lower frequencies and are strongly characterized by their formants corresponding to peaks in the spectral envelope. In contrast, the short-term spectrum of unvoiced sounds is typically much flatter than that of voiced sounds. The formants are usually fewer and less distinct.

With a ''universal'' vector quantization (VQ) coding scheme, the set of codebooks designed and used for all types of speech may not adequately cover the spectral variability. In a phonetically classified coder such as VRPS, quantization can be tailored to the particular types of spectra that occur.

3. OBJECTIVE PERFORMANCE ANALYSIS

It is intuitively palatable that the statistical character of the LPC spectra of unvoiced speech is distinctly different from that of voiced speech. Rate-distortion theory [6] can be used to determine the fundamental distinction in coding performance of these two types of LPC spectra. We have computed the distortion-rate function (DRF) of tenth-order LPC spectra for SD versus rate for both voiced and unvoiced cases by representing the spectra with cepstral coefficients and assuming a Gaussian distribution for these coefficients. The results indicate that, as expected, the unvoiced spectrum is easier to quantize than the voiced spectrum. For example, the DRFs show that a rate of 11 bits for voiced LPC spectra gives the same SD value as a rate of 8.36 bits for unvoiced spectra, a difference of 2.64 bits. This study will be reported in detail elsewhere.

We confirmed these theoretical values by separately designing unconstrained VQ codebooks for voiced and unvoiced speech *line spectrum frequency* (LSF) parameter sets and testing their performance on test sets independent of the data used from training. The experimental results also demonstrate that, in the range 5 to 11 bits, the unvoiced spectra require 2 bits less than the voiced spectra for the same performance. Class-specific VQ exhibited an advantage of 1 bit over universal VQ in the unvoiced case but essentially no advantage was found with voiced spectra. We infer that universal codebooks are dominated by the character of voiced spectra.

Unconstrained VQ with rates up to 11 bits for unvoiced spectra gives reasonably good perceptual quality. The degradation in the voiced class is, however, highly objectionable and higher rates are needed. We therefore use a 2-split VQ of the LSFs in the voiced class. With this scheme an average SD of less than 1 dB requires 25 bits. The outliers with SD between 2 and 4 dB are then less than 1%.

4. PERCEPTUAL EVALUATION

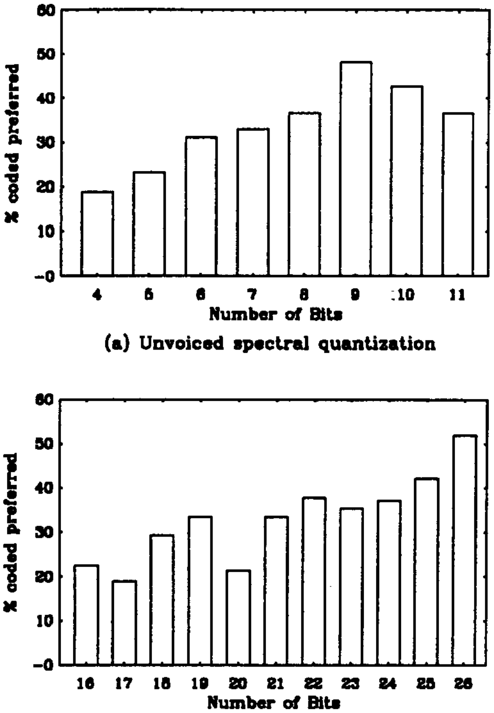
One of the main features of phonetically-based multi-mode coding is that, for each class, we need to allocate to the spectrum only the bits necessary to maintain transparent quality. To determine the rate needed for each class, we performed two separate listening tests. The first test was intended to assess the audible degradation due to spectral quantization for a particular class in the context of the VRPS coder. The second test measured audible degradation in synthesizing speech by applying the true prediction residual signal to a synthesis filter that has been quantized only for frames of a particular class.

A forced-choice A-B preference listening test was conducted. In this test, the listeners were presented with pairs of sentences and asked to indicate which of the two versions in the pair they preferred, or perceived as most natural-sounding. The ordering of the sentences in each pair and the ordering of the sequence of pairs in the test were randomized.

4.1 Listening Tests for Phonetically-Classified CELP

The first test was designed to investigate the role of class-specific spectral quantization on the perceived quality of speech coded with a phonetically classified CELP coder. As references, we generated sentences processed with VRPS where all parameters of the coder are fully quantized except for the LPC synthesis filter which was left unquantized. These were compared with test sentences that were processed with VRPS where the synthesis filter was quantized only for frames belonging to the particular class under test. Again, all other parameters of the coder were fully quantized. The results are compiled from a test consisting of 19 listeners each evaluating 8 sentence-pairs at each bit rate in each class. The reference and test sentences were obtained from 8 different English sentences spoken by 2 male and 2 female speakers.

The test was separated into two parts. In the first part of this test, only the unvoiced spectra were coded with unstructured VQ for the LSFs at a resolution ranging from 4 to 11 bits/frame. In the second part only the voiced spectra were quantized with a 2-split VQ of the LSFs. The rates ranged from 16 to 26 bits/frame. The test results given in Figure 1 show the percentage of time that the sentence with quantized spectra was preferred. Ideally, this number should be 50% if the sentences were of equal quality.



(a) Unvoiced spectral quantization

(b) Voiced spectral quantization

Figure 1. Listening test results for spectral quantization in a VRPS coder.

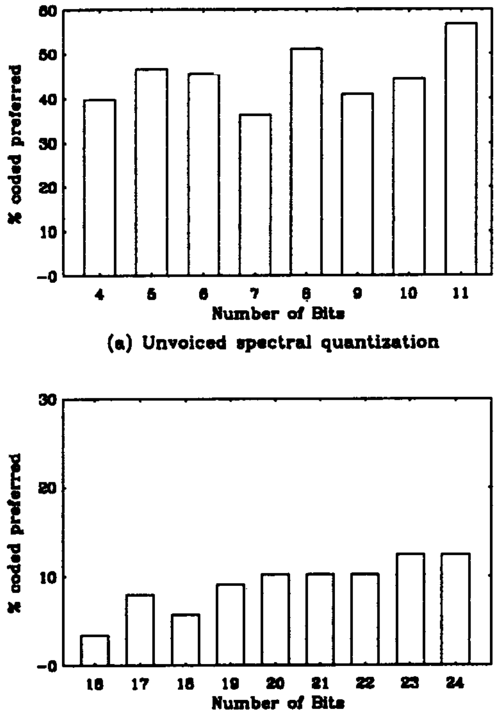
Since preference for spectrally quantized speech over unquantized should theoretically saturate at 50% as the number of bits approaches infinity, we would expect a knee in the curve of percent preference versus bits. While recognizing the statistical variability of a subjective test with limited sample size, we infer from Figure 1(a), that there is a knee in the preference versus bits curve for unvoiced spectral quantization at around 9 bits, where a preference of more than 40% for the coded version takes place. At 11 bits per frame, the percentage of time the coded version is selected is somewhat lower. However, for the range of 9 to 11 bits, careful listening does not demonstrate any significant difference and no audible distortion is observed in the coded versions. Thus, this behavior is attributed to statistical variability. Below 9 bits, the results indicate that the listeners identify a graceful increase in the degree of distortion. We observed this distortion to have the character of a "blow of air," somewhat like breath noise, and it is especially prominent in high-pass sounds such as fricatives. These sounds appear to be the primary source of annoying artifacts. We found that a resolution of 9 bits per frame corresponds to an average SD of 2.1 dB, thus showing that the "1 dB rule" for transparent quality is not valid for unvoiced LPC quantization. The requirements on the number of outliers mentioned in the "transparent quality" rules of [1] do not apply for unvoiced spectra either as the average distortion level is higher. Limiting the percentage of frames with large SD, however, may still be perceptually important. We observed that using 9 bits for unvoiced frames results in less than 1% of the coding frames having SD greater than 4 dB.

Fig. 1(b), for the voiced class, shows a different behavior. The percentage of time that coded frames were preferred is almost monotonically increasing with the bit-rate. This suggests that we have not quite reached the knee of the curve at 24 bits. In other words, "transparent" quality is obtained above the 24-bit level, which, from our experimental measurements, corresponds to less than 1 dB average SD, thus confirming the "1 dB rule" for the voiced class. The type of distortion heard when the voiced spectra is quantized is also very different from that of the unvoiced class: it is typically manifested as clicks, bells, tonal noise, reverberation, or similar effects. These artifacts can be very annoying and confirm the fact that the percentage of outliers is an important measure. For 25 bits, we found that the percentage of spectra with SD between 2 dB 4 dB is less than 1%, suggesting an even stronger requirement than the "2% rule" of [1] for transparent quality. A reason for this may be that for a universal LPC quantization scheme, many of the outliers come from the unvoiced class.

4.2 Tests With "True" Residual Excitation

For the second test, we wished to isolate the role of class-specific quantization of the LPC synthesis filter when all other sources of CELP coder degradations are eliminated. Thus, the "true" linear prediction residual was first obtained by applying the original speech to the unquantized linear prediction error filter. The test speech sentences were then generated by exciting an LPC synthesis filter with the true residual. If this synthesis filter were left unquantized, the original speech would be reproduced. The test sentences were obtained by quantizing the LPC filter parameters of the synthesis filter only for those frames corresponding to the class under study. The sentence pairs consisted of the original and a version where the LPC parameters for the class being tested were quantized and the synthesis filter was driven with the true LPC residual. The format of the tests were the same as for the previous test involving the the VRPS coder, except that only four of the sentences (from one male and one female speaker) were used. Hence, 4 sentence pairs for each bit rate in each class. were presented to each of the 19 listeners The results are shown in Figure 2.

For the unvoiced class, we observe in Figure 2(a) a very different behavior from the results of the first test shown in Figure 1(a). Here, the distortion incurred in quantization of the unvoiced class is barely audible for any of the bit rates. We believe that the reason for the difference lies in the *analysis-by-synthesis* (A-b-S) structure of CELP, which is not present in the current test. When a very small number of bits are assigned to



(a) Unvoiced spectral quantization

(b) Voiced spectral quantization

Figure 2. Listening test results for spectral quantization with true residual excitation.

the coding of the LPC parameters, the high frequency portion of the spectral envelope is not well represented. In the A-b-S coding structure, the algorithm minimizes the *perceptually weighted squared error* (WSE) between the coded and original speech waveforms. However, at low bit rates, such as those used for the unvoiced class in VRPS, the WSE is not so perceptually meaningful for speech coding. Hence, the coder tries to compensate for the high frequency spectral error in the search for an excitation vector, but in the process, introduces low frequency noise. This suggests that it might be preferable to abandon the A-b-S coding method for unvoiced speech and focus on accurately tracking the rapid time-variation in the energy of the excitation. This observation also agrees to some extent with Kubin et al. [7] , where the authors found that a noise excitation can be adequate for representing unvoiced speech.

For the voiced class, the situation is the opposite. Specifically, the A-b-S structure and perceptual weighting may result in an excitation vector that helps to compensate for spectral errors in the synthesis filter. On the other hand, the results for the true residual as excitation, seen in Figure 2(b), show that the spectral quantization introduces audible distortion at all bit rates. This again emphasizes the fact that for voiced sounds good spectral reproduction in the synthesized speech is perceptually crucial whether it is achieved by the synthesized filter alone or, as in an A-b-S structure, by a combination of reasonable spectral quantization augmented by the excitation codebook search process.

5. CONCLUDING REMARKS

Class-dependent VQ of the speech LPC parameters offers a significant coding gain for multi-mode or variable-rate coders when a simple voiced/unvoiced classification is used. A given spectral distortion can be obtained with fewer bits for unvoiced speech.

The rate required to achieve the so-called "transparent quantization" criteria outlined in [1] depends on the phonetic class. Informal subjective listening tests have shown that, while an average SD below 1 dB must be obtained for "transparent" coding of voiced frames, a distortion level of approximately 2.1 dB is sufficient for unvoiced speech. Furthermore, a reasonable requirement for unvoiced speech is that less than 1% of the unvoiced frames should have SD above 4 dB. The listening tests indicate that 9 bits/frame are sufficient for unvoiced spectra whereas 25 bits/frame are needed for voiced spectra. Finally, our experimental results with the true LPC residual as excitation add some strength to the viewpoint that the A-b-S mechanism should be abandoned from the coding of unvoiced speech.

References

[1] K.K. Paliwal and B.S. Atal, ''Efficient vector quantization of LPC parameters at 24 bits/frame,'' *IEEE Trans. Speech and Audio Processing*, vol. 1, no. 1, p. 314, 1993.

[2] T.M. Liu and H. Hoege, ''Phonetically-based LPC vector quantization of high quality speech,'' *Proc. European Conf. Speech Technology*, September 1989.

[3] S. Wang and A. Gersho, ''Phonetically-based vector excitation coding of speech at 3.6 kbit/s,'' *Proc. of the IEEE Int. Conf. on Acoust., Speech, and Sig. Proc.*, pp. I-49-52, Glasgow, May 1989.

[4] E. Paksoy, K. Srinivasan, and A. Gersho, ''Variable rate speech coding with phonetic segmentation,'' *Proc. of the IEEE Int. Conf. on Acoust., Speech, and Sig. Proc.*, pp. II-155-158, Minneapolis, April 1993.

[5] E. Paksoy, K. Srinivasan, and A. Gersho, ''Variable bit-rate CELP coding of speech with phonetic classification,'' *to appear in Europen Transactions on Telecommunications*, no. 5, 1994.

[6] T. Berger , *Rate distortion theory*, Prentice Hall, Englewood Cliffs, 1971.

[7] G. Kubin, B.S. Atal, and W.B. Kleijn, ''Performance of noise excitation for unvoiced speech,'' *Proc. IEEE Workshop Speech Coding for Telecom.*, pp. 35-36. Saint Adele, Quebec, October 1993.