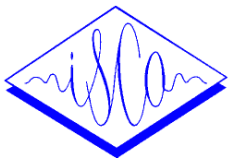

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

VECTOR QUANTIZATION OF GLOTTAL PULSES

Thomas Eriksson, Jan Lindén, and Jan Skoglund

Department of Information Theory
Chalmers University of Technology
S-412 96 Göteborg, SWEDEN.

E-mail:thomas.eriksson@it.chalmers.se,
jan.linden@it.chalmers.se,
jan.skoglund@it.chalmers.se

ABSTRACT

An efficient codebook driven voiced excitation coding method producing natural sounding speech is proposed. It can be incorporated as an essential part in a complete speech coder working at low bit rates. The interpulse correlation of such a coding scheme is investigated and exploited using linear predictive vector quantization and finite state vector quantization (FSVQ). A new and more robust FSVQ method that is able to more efficiently exploit this correlation is proposed. The new method dynamically combines two memoryless vector quantizers. This dynamic combination method is not restricted to pulse codebooks but can be employed for any coding scheme with intervector correlation.

1. INTRODUCTION

Most modern speech coders are based on the source-filter concept which tries to separate the excitation source waveform from the spectral characteristics of the vocal tract. For voiced speech, the excitation waveform essentially describes the airflow through the vocal folds. For coders operating at high and medium rates, the majority of the available bits are used for excitation coding. Hence, that part can contribute most to an efficient bit reduction. One approach to reduce the number of bits in the excitation coding is to use schemes relying on a parametric description of the source signal (vocoders). The problems associated with a parametric description concern the limitations of the model which can result in un-natural sounding speech. Utilization of various codebooks is another possibility, as in Code-Excited Linear Prediction (CELP) coders. CELP coders can achieve good performance if the codebooks are populated with a sufficient number of different excitation shapes, but the performance deteriorates at bit rates lower than say 4 kb/s.

In recent years the research community has reached a point where one major goal is high quality speech coding at lower rates than traditional CELP coders have achieved. This motivates interest in coding techniques combining the speech quality of codebook driven waveform coders and the compression efficiency of parametric coders. In this paper, a novel method for coding voiced parts of the excitation by using a prototype pulse codebook, is proposed. The codebook is populated by typical glottal flow derivative waveforms which originate from real speech. Given such a pulse codebook we present methods for exploiting interpulse correlation to improve performance, alternatively to decrease coding rate. The concept of introducing a pulse codebook has previously been restricted to CELP variations, still having relatively high rates [1,2]. Our work can be seen as an extension of a fully parameterized vocoder [3] by replacing excitation pulse parameters by a codebook.

2. VECTOR QUANTIZATION OF GLOTTAL PULSES

It is widely known that it is crucial to accurately represent the periodic behavior of speech [4]. Therefore we (and others with us, e.g. [5]) have focused on what happens during one pitch cycle regarding the shape of the signal. Pitch and amplitude information are assumed to be handled separately from this work, although it is clear that a combined analysis of all three parameters pitch, amplitude and shape could yield a more efficient coding procedure.

An inverse filter $A(z)$ is obtained by performing an LPC analysis on pre-emphasized speech $S_p(z) = S(z)P(z)$. We have utilized an adaptive pre-emphasis filter $P(z) = 1 - \mu z^{-1}$. The inverse-filtered speech signal, $R(z) = S(z)A(z)$, is now a simple approximation of the glottal flow derivative waveform, although no attempts have been made to model the "true" glottal signal. The waveform is divided into separate pulses by finding the instants of glottal closure and each of these pulses is quantized separately. In order to compare pulses of different lengths and amplitudes, a transformation procedure is performed. The original pulse having a period of N_P samples is transformed into a vector of normalized length N_{CB} and normalized amplitude as depicted in Figure 1. The time-scaled pulse is also positioned in time by fractions of samples. An N_{CB}-dimensional vector quantizer (VQ) produces a quantized pulse which then is transformed back to the inverse-filtered speech signal domain.

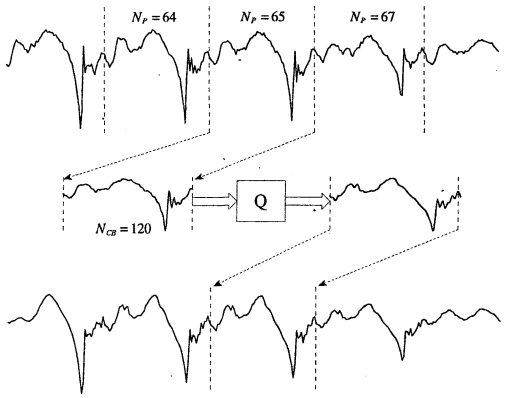

Figure 1. The vector quantization process. Pulses having different lengths N_P are each transformed to normalized length N_{CB}, quantized, and then retransformed to original length.

We have trained pulse VQs by utilizing the generalized Lloyd algorithm [6]. The performance of memoryless VQ quantization of excitation pulses for different sizes is depicted in Figure 7.

3. EXPLOITING INTERPULSE CORRELATIONS

In order to improve the performance of the pulse quantizer, we here analyze what can be gained by exploiting interpulse correlations.

If the pulses are quantized in a VQ with M entries, the required rate R (bits per entry) to encode the stream of indices I_n has a lower bound, determined by the entropy of the index source

$$R \geq H(I_n) = -\sum_{i=0}^{M-1} P(I_n = i) \log_2 P(I_n = i)$$

To estimate the required rate when knowledge of the previous index is exploited, we compare the entropy above with the conditional entropy, computed as

$$H(I_n|I_{n-1}) = -\sum_{i=0}^{M-1} p_i \sum_{j=0}^{M-1} p_{j|i} \log_2 p_{j|i}$$

where $p_{j|i} = P(I_n = j | I_{n-1} = i)$ and $p_i = P(I_{n-1} = i)$. The mutual information is defined as the entropy difference $\mathcal{J}(I_n, I_{n-1}) = H(I_n) - H(I_n|I_{n-1})$ and this is a measure of the possible gains achievable by fully exploiting the knowledge of the previously coded vector. We note that estimates of the entropies above are straightforward to compute when the probabilities have been estimated.

The entropy for a 7-bit memoryless vector quantizer was found to be $H(I_n) = 6.5$ bits, while the conditional entropy was $H(I_n|I_{n-1}) = 4.4$ bits which gives the mutual information $\mathcal{J}(I_n, I_{n-1}) = 2.1$ bits. The entropy measurements indicate that large gains can be achieved if interpulse correlations are exploited. In the following sub-sections, we try to improve the coder performance by including memory of previous pulses.

3.1. Linear predictive VQ

A straightforward method to exploit interpulse correlation is to design a linear (vector) predictor for the pulses, and train a VQ to quantize the prediction error [7]. This method, very similar to a DPCM system, is illustrated in Figure 2.

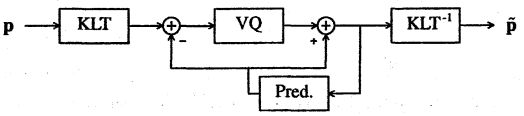

Figure 2. The linear predictive coding scheme. A pulse vector $\mathbf{p}$ *is input, and a quantized pulse* $\tilde{\mathbf{p}}$ *is output. Compare with a DPCM system.*

As illustrated in Figure 2, we have chosen to let the predictor in the system work with Karhunen-Loeve transformed (KLT) vectors. This approach has two advantages:

I) The vector components are uncorrelated after the KLT, which means that each component can be predicted separately[1], and

II) Only a few strong components of the KLT-vector need to be predicted, typically 10-20% of the full vector size. The rest of the vector has low energy and is thus of less importance. Also, our measurements show that the low energy components gain very little from prediction.

The results from the linear prediction experiments are listed in Table 2 and 3. The conditions for all simulations are discussed in Section 4. The results confirm that the interpulse correlation can be efficiently exploited as expected. In the next sub-section we investigate other methods to exploit the correlation.

[1]This conclusion holds for stationary Gaussian processes. The Gaussian assumption does not hold for this application but is an acceptable approximation.

3.2 Finite-state VQ

A finite-state VQ (FSVQ) is a vector quantizer with memory, where the output depends not only on the current input vector, but also on the past history of quantized vectors [6]. An FSVQ can be viewed as a collection of memoryless vector quantizers, together with a selection rule (next-state function) that determines which of the VQs to use for the encoding of the current input vector, c.f. Figure 3.

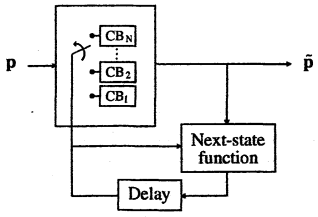

Figure 3. A finite-state VQ.

Several different FSVQ methods have been proposed in the literature. Also linear predictive VQ can be viewed as a special case of FSVQ. Here we study a successful FSVQ technique, *omniscient* FSVQ, which has been found to yield the best codes with reasonable complexity in most applications [6]. We also propose a new FSVQ technique, denoted *dynamic combination VQ (DCVQ)*, which is shown to outperform the omniscient technique for this application.

3.2.1 Omniscient FSVQ

We have experimented with an FSVQ technique called omniscient FSVQ [8]. The omniscient FSVQ technique has shown good performance in many other applications, including speech coding [9].

The next-state selection rule consists of classifying the current input vector in a memoryless vector quantizer. The index of the classifying quantizer determines which of the state quantizers to use for quantization of the *next* input vector. For each of the states (classes) in the classifying VQ, a state codebook is trained, using only the training data belonging to the actual state.

The decoder operates in a slightly different way. The decoder cannot track the next-state rule defined above, since it depends on the input rather than on the encoded input. However, if we replace the actual input with the encoded input, we get an approximation of the next-state rule used in the design.

The results from omniscient FSVQ of glottal pulses, with 8 and 16 classes, are presented in Table 1. The FSVQ method does not succeed in exploiting the interpulse correlation to the same extent as the linear prediction method, yet it outperforms a memoryless VQ.

The omniscient FSVQ technique requires very large databases for training purposes, especially if the number of states is large. Even for the relatively small number of states we have experimented with, the training is very complex and requires a large training database. Another problem is the robustness, both against changes in the input signal and against channel errors. Most FSVQ systems also suffer from a problem that is referred to as "derailment" [8], similar to the slope-overload phenomenon well known from delta modulator quantizers.

The problems with the omniscient technique motivated us to propose a new FSVQ scheme in the next sub-section.

3.2.2 Dynamic combination of VQs

In Figure 4, two histograms, measured over the training database, are plotted. The dotted histogram shows the distribution of squared distances from the mean pulse, $D_{\bar{\mathbf{p}}} = \|\mathbf{p}_n - \bar{\mathbf{p}}\|^2$, and the solid histogram shows the distribution of squared distances from the previous pulse, $D_{\mathbf{p}_{n-1}} = \|\mathbf{p}_n - \mathbf{p}_{n-1}\|^2$.

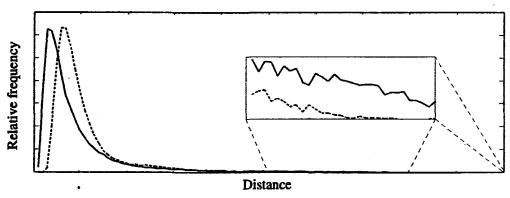

Figure 4. Histogram of $\|\mathbf{p}_n - \bar{\mathbf{p}}\|^2$ (dotted line) and $\|\mathbf{p}_n - \mathbf{p}_{n-1}\|^2$ (solid line). As we see in the magnified section, the solid histogram has a much longer tail of "outliers". The mean is approximately the same for both variables.

We see that the most probable value for $D_{\mathbf{p}_{n-1}}$ is about half the most probable value of $D_{\bar{\mathbf{p}}}$, which indicates that a pulse $\mathbf{p}_n$ with high probability is closer to the previous pulse $\mathbf{p}_{n-1}$ than to the mean pulse $\bar{\mathbf{p}}$. However, the histogram of $D_{\mathbf{p}_{n-1}}$ has a much longer tail of "outliers", and both variables have about the same mean value. This motivated us to try to encode the outliers separately from the typical, slowly-changing pulses. We write the probability density of a pulse $\mathbf{p}_n$, given the history of pulses $\mathcal{H} = \{\mathbf{p}_{n-1}, \mathbf{p}_{n-2}, \ldots\}$, as

$$f(\mathbf{p}_n|\mathcal{H}) = f(\mathbf{p}_n|\mathcal{H}, \mathcal{S}) \cdot P(\mathcal{S}) + f(\mathbf{p}_n|\mathcal{H}, \mathcal{T}) \cdot P(\mathcal{T})$$

where $\mathcal{S}$ is the event "steady state segment" and $\mathcal{T}$ is the event "transition". We define a steady state segments as a segment where a pulse is similar to the previous pulse(s), so the density $f(\mathbf{p}_n|\mathcal{H}, \mathcal{S})$ has a maximum for pulses close to the previous. On the other hand, in a transition segment the pulse can be assumed to be independent of the history, $f(\mathbf{p}_n|\mathcal{H}, \mathcal{T}) \approx f(\mathbf{p}_n|\mathcal{T})$. The method presented below is based on the above probability density.

In the new approach, called dynamic combination VQ (DCVQ), two VQs with different resolutions are dynamically combined into one (see Figure 5).

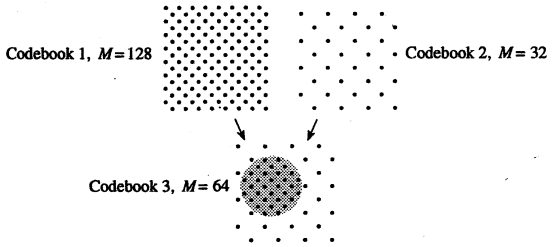

Figure 5. The combination principle: Construct a new codebook 3, by adding vectors from a dense codebook 1 to a sparse codebook 2 in a specified region

We start with a sparse codebook and add vectors from a more dense codebook (called super codebook) to the region where the previous codebook entry was chosen. The dense codebook part (called state codebook) extends the sparse codebook to improve performance for speech segments $\mathcal{S}$ where the pulse form is stable. The underlying sparse codebook takes care of the outliers in segments $\mathcal{T}$ where the pulseform undergoes rapid changes. The current state, which is the same as the last chosen codeword, determines which part of the dense codebook that will be used. Figure 4 shows that successive vectors lie fairly close to each other. The subset of the super codebook should hence be chosen as the codevectors that have minimum distance to the last chosen codevector.

The introduction of a sparse codebook introduces some ambiguity of what is the current state. Hence, when a code vector is chosen from the sparse codebook the state is chosen as the closest codevector of the super codebook. Note that even if the state is chosen in the super codebook the actual index produced by the encoder is referring to a vector in the sparse codebook. This is the main reason why this method is less sensitive to channel errors and causes less derailment problems than common FSVQ schemes.

An obvious problem that arises when trying to combine two different codebooks is how many vectors should be chosen from each of the codebooks. In Figure 6 we show results for different mixes given that the super codebook is 13 bits and the combined codebook is 9 bits. The performance is presented in SNR_Q (average Signal-to-quantizing Noise Ratio for the pulse VQ).

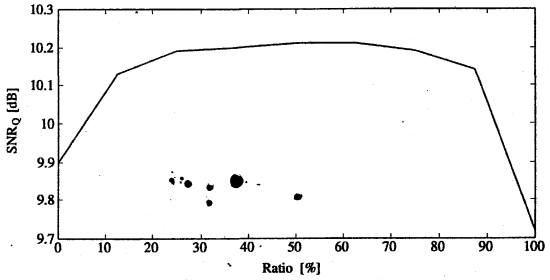

Figure 6. SNR_Q for a 9 bit DCVQ versus the ratio of the sparse and the combined codebook size. (A ratio of 100% is equivalent to a 9 bit memoryless VQ.)

These results suggest that the choice of mixing is not critical in terms of performance but that the sparse codebook and the state codebooks should be roughly equal in size. In all simulations the size of the state codebooks and the sparse codebook have been chosen to be equal. The problem of finding the best size of the super codebook given the sizes of the sparse and the state codebooks is more intricate. Figure 7, which shows the results of a number of simulations for different sizes of the super codebook, indicates that the size of the super codebook should be chosen as large as possible in order to get the lowest average distortion. Yet, it is important to note that large savings in storage requirements and complexity can be achieved by decreasing the size of the super codebook.

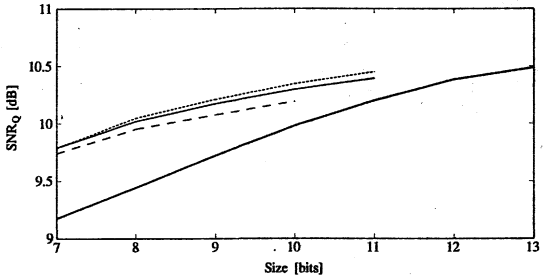

Figure 7. SNR_Q for DCVQ with different sizes of super codebook as a function of number of bits in the combined quantizer. 11 bits (dashed), 12 bits (solid) and 13 bits (dotted). The thick line corresponds to the memoryless VQs.

In Table 1, the dynamic combination method is compared to the omniscient FSVQ described in Section 3.2.1. Beside the performance in SNR_Q, we have also computed the resulting segmental SNR, SNR_{SEG}, of the coded speech when the

quantized pulses are transformed back to synthetic speech. The segmental SNR was computed using 5 ms segments.

We see that the DCVQ outperforms the omniscient FSVQ in terms of both SNR_Q and SNR_{SEG}.

TABLE 1. COMPARISON OF FSVQS (UNITS IN dB)

VQ	SNR_Q 8 bits	SNR_{SEG} 8 bits
Memoryless	9.44	9.53
Omniscient, 8 states	9.59	9.66
Omniscient, 16 states	9.69	9.74
Dynamic combination	10.05	10.22

A major problem with coding schemes that employ some type of feedback, such as FSVQ, is that channel errors propagate, which can significantly degrade the performance. There exists a number of standard methods that try to decrease the effect of channel errors. An example is to periodically perform a full search that forces the encoder into the best possible state, and transmit the new state to the decoder. This can only be done quite in-frequently otherwise the cost of sending extra information gets too high. Another possibility is to account for the noisy channel already in the design of the VQ. This approach is called Channel Optimized VQ (COVQ) [10]. The dynamic combination method does itself provide some robustness against propagating channel errors. This is due to the fact that every time an index is chosen from the sparse codebook the current state is unambiguously determined, which removes the need for full search. The sparse codebook is utilized 20-30% of the time which means that the state will be frequently updated. COVQ methods can be incorporated in the DCVQ design procedure, and hence these gains are also achievable for the dynamic combination method.

The dynamic combination method also decreases the impact of the derailment problem of FSVQ systems, again due to the fact that each time a codevector is chosen from the sparse codebook the current state is unambiguously determined.

In this section we have shown that the proposed DCVQ method outperforms the omniscient FSVQ technique in terms of average distortion and at the same time is more robust against derailment and channel error propagation.

4. RESULTS AND DISCUSSION

The vector quantizers were trained on a database of over 200 000 pulses originating from several speakers, both male and female. Coder performance is evaluated by use of another database with 15 000 pulses. The evaluation database contains speech from other speakers than those in the training set. The pulses had a normalized length of 120 samples.

Table 2 and 3 shows the performance in SNR_Q and SNR_{SEG} for the presented VQs of different sizes.

For the linear predictive VQ, 20 scalar predictors of order 3 for the 20 strongest KLT components are used. For the rest of the components in the vector, no prediction is performed.

For the dynamic combination VQ, the sparse codebooks and the state codebooks are equal in size and 13 bits super codebooks are used.

TABLE 2. COMPARISON OF VQS - SNR_Q (UNITS IN dB)

VQ	7 bits	8 bits	9 bits	10 bits
Memoryless	9.17	9.44	9.72	9.98
Linear predictive	9.62	10.00	10.30	10.57
Dynamic combination	9.79	10.05	10.21	10.35

TABLE 3. COMPARISON OF VQS - SNR_{SEG} (UNITS IN dB)

VQ	7 bits	8 bits	9 bits	10 bits
Memoryless	9.33	9.53	9.78	10.10
Linear predictive	9.76	10.14	10.46	10.75
Dynamic combination	9.97	10.22	10.39	10.52

From these tables we can see that the use of vector quantizers with memory implies large gains in performance compared to memoryless, typically over 2 bits. For low rates DCVQ is the best of the investigated methods but for higher rates the linear predictive scheme is the most efficient.

When incorporating this coding method in a speech coder operating at low bit rates it will not be possible to encode each individual pulse. Therefore a straightforward procedure would be to periodically quantize isolated pulses and the intermediate pulses would then be interpolated [5]. This sampling would decrease the correlation between encoded pulses but for typical sampling periods of 20-30 ms the correlation would still be relatively high.

Informal listening tests reveal that the proposed coder sounds more natural and non-buzzy than a fully parameterized vocoder. Compared to a CELP coder, the coded speech suffers less from the noisy character evident in most CELP coders, but is perceived as more unstable.

5. SUMMARY

We have presented an efficient codebook driven voiced excitation coding method producing natural sounding speech. It can be incorporated as an essential part in a complete speech coder working at low bit rates. We have also investigated the interpulse correlation of such a coding scheme and further improved the efficiency by exploiting this correlation. A new FSVQ method is proposed, dynamic combination VQ. The new method is not restricted to pulse codebooks but can be employed for any coding scheme.

6. REFERENCES

[1] A. Bergström and P. Hedelin, "Code-Book Driven Glottal Pulse Analysis", *ICASSP-89* (Glasgow, Scotland), pp. 53-56, 1989.

[2] C. McElroy, B.P. Murray, and A.D. Fagan, "Wideband Speech Coding Using Multiple Codebooks and Glottal Pulses", *ICASSP-95* (Detroit, USA), pp. 253-256, 1995.

[3] P. Hedelin, "A Glottal LPC-vocoder", *ICASSP-84* (San Diego, USA), pp. 161-164, 1984.

[4] G. Fant, *Speech sounds and features*. MIT Press, 1973.

[5] W.B. Kleijn, "Encoding speech using prototype waveforms", *IEEE Trans. on Speech and Audio Processing* , vol. 1, no. 4, pp. 386-399, 1993.

[6] A. Gersho and R.M. Gray, *Vector quantization and signal compression*. Kluwer Academic Publishers, 1992.

[7] V. Cuperman and A. Gersho, "Vector predictive coding of speech at 16 kbits/s", *IEEE Trans. on Communications* , vol. 33, no. 7, pp. 685-696, 1985.

[8] J. Foster, R.M. Gray, and M.O. Dunham, "Finite-state vector quantization for waveform coding", *IEEE Trans. on Information Theory* , vol. 31, no. 3, pp. 348-359, 1985.

[9] M.O. Dunham and R.M. Gray, "An algorithm for the design of labeled-transition finite-state vector quantizers", *IEEE Trans. on Comm.* , vol. 33, pp. 83-89, 1985.

[10] Y. Hussain and N. Farvardin, "Finite-state vector quantization over noisy channels and its application to LSP parameters", *ICASSP-92* (San Francisco, USA) pp. 133-136, 1992.