QUANTIZATION OF SPECTRAL SEQUENCES USING VARIABLE LENGTH SPECTRAL SEGMENTS FOR SPEECH CODING AT VERY LOW BIT RATE

Geneviève Baudoin(1), Jan Cernocký(1,2), Gérard Chollet(3)

1 ESIEE, Dpt Signaux-télécommunications, BP 99, Noisy Le Grand, 93162 CEDEX France, baudoin@esiee.fr
2 FEIVUT Brno, Purkynova 118, 61200 Brno, Czech Republic, cernocky@urel.fee.vutbr.cz
3 ENST, Dpt Signal, 46 rue Barrault, 75013 Paris, France, chollet@sig.enst.fr

ABSTRACT

This paper deals with the coding of spectral envelope parameters for very low bit rate speech coding (inferior to 500 bps). In order to obtain a sufficient intelligibility, segmental techniques are necessary. Variable dimension vector quantization is one of these.

We propose a new interpretation of already published research from Chou-Loockabaugh [2] and Cernocky-Baudoin-Chollet [4,6] on the quantization of variable length sequences of spectral vectors, named respectively Variable to Variable length Vector Quantization (VVVQ) and Multigrams Quantization (MGQ). This interpretation gives a meaning to the Lagrange multiplier used in the optimization criterion of the VVVQ, and should allow new developments as, for example, new modelization of the probability density of the source.

We have also studied the influence of the limitation of the delay introduced by the method. It was found that a maximal delay of 400 ms is generally sufficient.

Finally, we propose the introduction of long sequences in the segmental codebook by linear interpolation of shorter ones.

1. INTRODUCTION

For very low bit rates (below 500 bps) speech coding, it is useful to take into account the interframe dependencies, by using segmental quantization techniques for the coding of spectral parameters.

Chou and Loockabaugh [2] have proposed a method for the quantization of spectral vector sequences with variable length segments, under the name VVVQ (Variable to variable Vector Quantization). The results are satisfactory for spectral bit rates as low as 50 bps, for a single speaker, despite some limitations as long delay and complexity.

Another such method has been proposed by Cernocký, Baudoin and Chollet [4,6] with the name Quantization of spectral sequences with Multigrams (noted MGQ).

This paper develops the following topics: new interpretation and comparison of the 2 approaches, study of the introduced delay and proposition of an algorithm to optimize the performances when a maximal delay is imposed, introduction of long sequences in the dictionary by linear interpolation of shorter ones.

2. DESCRIPTION AND COMPARISON OF THE VVVQ AND MULTIGRAMS METHODS

2.1. The VVVQ method

This method segments and quantizes a temporal sequence of spectral vectors with variable dimension vector quantization using a codebook of variable length spectral vectors sequences. The length of these segments can vary from 1 to n spectral vectors.

The codebook sequences are entropy coded. So, they are represented by a variable number of bits depending on their probabilities. Therefore, both the length of the codebook sequences and the number of coding bits for one sequence are variable, which explains the name Variable to Variable length Vector Quantization.

The codebook is obtained by minimizing, on a training database, the average spectral distortion for a limited average bit rate. A Lagrange multiplier technique is applied and the optimization criterion can be written:

$$\min_{S,s} \{d_S + \lambda r_S\} = \min_{S,s} \left\{ \sum_{i,j} d_{i,j} + \lambda \sum_{i,j} n_{i,j} \right\}$$ (1)

where $S$ is the set of all possible segmentations of the database in segments of length inferior to $n$, $S_i$ is one such segmentation, $d_S$ is the corresponding distortion, $r_S$ the associated bit rate and $\lambda$ the Lagrange multiplier.

More precisely:

$$d_S + \lambda r_S = \sum_{s_i \in S} \left( d_{i,j} + \lambda n_{i,j} \right)$$ (2)

where $s_i$ is the $j^{th}$ segment of $S_i$, $n_{i,j}$ the number of coding bits for $s_i$ and $d_{i,j}$ the distortion on this segment (sum of the distortions on all the vectors of the segment). In this paper it is always supposed that:

$$n_{i,j} = -\log_2(\text{proba}(M_{i,j}))$$ (3)

$M_{i,j}$ being the codebook sequence coding $s_i$.

The codebook is initialized with Z vectors sequences and their probabilities. Then an iterative 2 steps EM (Expectation Maximization) algorithm [1] is used to construct the codebook. At the $q^{th}$ iteration, the codebook $C_q$ contains $Z$ sequences $M_{q,i}$ with their probabilities $p(M_{q,i})$, the new codebook $C_{q+1}$ is calculated in 2 steps:

- step 1: Segmentation of the database in $N$ segments using Viterbi algorithm to optimize the criterion (1). To each $M_{q,i}$ corresponds a class of $N_{q,i}$ sequences of the training database coded by $M_{q,i}$.
2.2. The Multigrams Quantization method

As for VVVQ, the basic idea is to segment and quantize the spectral vectors sequences using a codebook of variable length segments called multigrams. In a first approach, the spectral vectors were vector quantized and the multigrams Mk were sequences of n or less quantization indices.

The codebook was obtained by maximizing the joint likelihood L of the optimal segmentation S_{opt} and of the training observation (sequence of quantization indexes) [3]. The segments were supposed to be independent and the optimization criterion was to maximize L:

\[ L(\text{observation}, S_{opt}) = \max_{S \in S} \prod_{k \in S} p(M_k) \]  
(4)

The codebook was initialized with the sequences present in the training database, the probabilities of the sequences being initialized by counting their number of occurrences. The EM algorithm was used to calculate the codebook (probabilities only) and the multigrams were entropy coded. The results were insufficient for VQ size above 128, due to the large variability of index sequences. The results were good on the training database but unsatisfactory on the test database.

This led to a second approach. Spectral vectors are no longer vector quantized. A multigram Mk is a sequence of n or less spectral vectors. The observation sequence of spectral vectors is segmented in segments Uk which are quantized by multigrams Mk in order to maximize the new criterion L' :

\[ L'(\text{observation}, S_{opt}) = \max_{S \in S} \prod_{k \in S} p'(M_k) \]  
(5)

Where \( p'(M_k) \) is the penalized probability of Mk, defined as the product of Mk probability with a penalization factor Q depending on the distance \( d_k \) between the observed segment Uk and its coding multigram Mk.

\[ p'(M_k) = p(M_k)Q(d_k) \]  
\[ d_k = d(U_k, M_k) \]  
(6)

\[ Q(d) = \begin{cases} 
1 - \frac{d}{d_{\text{max}}} & \text{for } d \leq d_{\text{max}} \\
0 & \text{for } d > d_{\text{max}} 
\end{cases} \]  
(7)

Where \( d_{\text{max}} \) is an arbitrary constant. The number of multigrams of each size in the initial codebook is limited a-priori. The multigrams codebook of fixed dimension is initialized, then the EM algorithm is used to calculate the codebook iteratively optimizing the criterion (5). In each iteration, the 2 steps EM algorithm proceeds in the same way as it was described for VVVQ.

2.3. New interpretation and comparison of the methods

While independently developed, these 2 techniques are very similar. The VVVQ is mathematically better expressed and is locally optimal for a given codebook structure.

The MGQ approach brings a different perspective. It will be first reformulated, and with this new interpretation, the 2 approaches will be compared.

To reformulate the MGQ method, we consider that a spectral sequence is generated by a source emitting variable length independent multigrams and that the vectors (size p) of the multigrams are gaussian with a covariance matrix \( \sigma^2 I \), I is the pxp identity matrix. The parameters \( \theta \) (multigrams and probabilities) of the source are identified by maximizing the joint likelihood of the optimal segmentation \( S_{opt} \) and of the observation:

\[ \max_{\theta} L(\text{obs}, S_{opt}) \Leftrightarrow \max_{\theta} L(S_{opt})L(\text{obs} / S_{opt}) \]  
(8)

\[ L(S) = \prod_{k} p(M_k) \]  
(9)

\[ L(\text{obs} / S) = \prod_{k} p(U_k / M_k) \]  
(10)

\( U_k \) being a length h segment from the training database and \( M_k \) the multigram that quantifies \( U_k \) in the segmentation S. With the proposed gaussian model and using a logarithm, the criterion is equivalent to:

\[ \max_{\theta} \sum_{k} \left( \log(p(M_k)) - \sum_{j=1}^{n} \sum_{m=1}^{n} \frac{(c_{k,j,m} - m_{k,j,m})^2}{2\sigma^2} \right) \]  
(11)

\[ \Leftrightarrow \min_{\theta} \sum_{k} \left( \sum_{j=1}^{n} d(c_{k,j,m}, m_{k,j,m}) - 2\sigma^2 \log(p(M_k)) \right) \]  
(12)

\( c_{k,j,m} \) and \( m_{k,j,m} \) are the mth coefficients of jth vectors of segment \( U_k \) and multigram \( M_k \), d(c_{k,j,m}, m_{k,j,m}) is a quadratic distance between the jth vectors of \( U_k \) and \( M_k \).

Equation (12) is the VVVQ criterion with \( \lambda = 2\log(2)\sigma^2 \) and a quadratic distance on the spectral vectors.

On the other hand, it is possible to interpret the arbitrary criterion of the MG method by observing that, for \( d < d_{\text{max}} \):

\[ \log(p) + \log(1 - \frac{d}{d_{\text{max}}}) = \log(p) - \frac{d}{d_{\text{max}}} \]  
(13)

with \( d_{\text{max}} = 2\log(2)\sigma^2 \). We have used here, a triangular probability, which is closed to a gaussian, for d small compared to \( d_{\text{max}} \), and guarantees that d is always limited by \( d_{\text{max}} \).

Another possible interpretation can be obtained by considering that the source emits variable length independent constant multigrams to which is added a centered gaussian noise of variance \( \sigma^2 \).

A further difference between the VVVQ and the MGQ approach is the spectral distortion used. Chou & al worked with a modified Itakura distance while we used a quadratic distance on cepstral coefficients. With the modified Itakura measure, the preceding interpretations must be applied on the residual signal of the linear prediction which is supposed to be white and gaussian.

3. LIMITATION OF THE DELAY

The theoretical delay introduced by these methods is equal to the length of the signal. When the delay is limited to a value of \( k_{\text{max}} \) frames, the performances are
degraded. To limit the delay, the classical technique uses a buffer of \( k_{\text{max}} \) frames and imposes segmentation points at the buffer extremities.

We have developed a new algorithm. At each new input of a spectral vector in the buffer, we examine if the \( n \) best possible segmentations from the origin of the buffer to the last received vector coincide until a certain position in the buffer. If yes, the buffer is cleared until there. As long as the buffer does not saturate, the performances are not degraded despite the limitation of the buffer size to \( k_{\text{max}} \).

We have studied the statistical characteristics of the buffer filling for different values of \( n \) and \( \lambda \). Results are given in section 6.

4. CONSTRUCTION OF LONG MULTIGRAMS BY INTERPOLATION

When the maximum length of multigrams is increased, the number of spectral vectors to train augments rapidly. For example, for \( n=16 \), 64 multigrams per length, there are 8704 vectors to train and 35088 if \( n=32 \).

In order to increase the maximum length of multigrams without being obliged to increase the size of the training database, we have constructed a codebook with multigrams of length 1 to \( n \) from a codebook of maximum length \( n/2 \), by stretching with linear interpolation the multigrams of length \( n/2 \) to obtain the long multigrams from \( n/2+1 \) to \( n \), taking into account the fact that the same acoustic sequences can be uttered at different speeds.

During the training, the multigrams of size \( n/2 \) are actualized from associated segments of size \( n/2 \) and from longer segments associated to the stretched multigrams. In the last case, the actualization is done by linearly contracting the long segments. At each iteration, the Multigrams of size 1 to \( n/2 \) and all the probabilities are saved.

The obtained results (distortion - rate curves), for \( n=16 \) with stretching of length 8 multigrams, are better than those obtained with a non stretched codebook of maximal length 12 containing the same number of spectral vectors to train (figure 2). But the complexity is greatly increased.

5. EXPERIMENTAL RESULTS

Distortion-Rate definitions:

The spectral distortion was defined as the average over all frames of the logarithm of the spectral distance:

\[
D_{\text{log}} = \sqrt{\int \left[ 10\log S(f) - 10\log \tilde{S}(f) \right]^2 df}
\]

(in dB) where \( S(f) \) and \( \tilde{S}(f) \) are the power LPC-spectra with original and quantized coefficients respectively. It was approximated by a scaled euclidian distance on the cepstral coefficients.

The bit rate is defined as the average number of bits for the coding of one spectral vector, it is a number of bits per frame. The average bit rate per frame \( R \), corresponding to a multigrams codebook with entropy coding, is given by the ratio of the codebook entropy \( H \) to the average Multigram length \( \bar{l} \):

\[
R = \frac{H}{\bar{l}} = \frac{\sum_{i=1}^{Z} p(M_i) \log_2 p(M_i)}{\sum_{i=1}^{Z} p(M_i) p(M_i)}
\]

(15)

where \( l(M_i) \) and \( p(M_i) \) are the length and the probability of \( M_i \), and \( Z \) the number of multigrams in the codebook.

Database :

We have used a single speaker of the Swiss-French database PolyVar. It contains telephone calls recorded over 6 months, consisting of read sentences, spelled words, digits, some control words and spontaneous speech.

The signal is digitized at 8 kHz with a A-law 8 bits quantization. The spectral vectors are 10 LPCC calculated with preemphasis on 20 ms Hamming windows with 10ms overlapping. The first cepstral coefficient is not used.

The corpus was divided into 213270 vectors for the training and 122903 vectors for the test.

Codebook initialization :

Different codebook initialization have been compared.

- Initialization with the most frequent vector-quantized multigrams: after vector-quantizing the training database with a VQ codebook of \( L \) vectors, we used for each multigram of length 1 the most frequent quantized sequences of length 1.

- Matrix quantization initialization: for each multigram length, the multigram codebook is initialized with a matrix quantization codebook. [2]

- Natural random codebook: the multigram codebook is initialized with natural spectral vector sequences chosen randomly. [5].

After a few iterations of the EM algorithms, the 3 different initializations gave similar results, so we used the last one.

Test configurations :

The following results have been obtained for different Multigram or Matrix quantization (only one possible sequence length in the codebook) codebook topologies:

- MG16 Multigram Quantizations with \( n=16 \) and 64 multigrams per length, \( 0 \leq A \leq 1 \). There are 8704 cepstral vectors in the codebook.

- MQ8704, Matrix Quantization (\( \lambda=0 \)) with different codebooks each corresponding to a unique sequence length 1 between 2 and 20 vectors, all codebooks containing 8704 cepstral vectors. For example, for \( l=8 \), there are 1088 sequences of 8 vectors in the codebook and for \( l=16 \), 544 sequences of 16 vectors.

- MQ1, MQ2, MQ4, 3 entropy constrained matrix quantizations with codebooks containing 8704 sequences of respective lengths 1, 2, 4 corresponding to 8704, 17408 and 34816 cepstral vectors, \( 0 \leq A \leq 1 \).

- MG8, MG12, MG8 stretch, Multigrams quantization with respectively \( n = 8, 12, 16 \), and 64, 64, 113 multigrams per length, \( 0 \leq A \leq 1 \). The long multigrams
Limitation of the delay, buffer filling
Figure 1 gives, for 0≤λ≤0.5, the probability distributions of the buffer filling (number of frames in the buffer), for the Multigram quantization MG16, when the new algorithm for limitation of the delay is used.
In case of limitation of the delay to k_max frames, the performances are clearly improved with this algorithm compared to the classical clearing of the buffer every k_max frames. For example, for MG16, a buffer length of k_max=40 frames (400 ms delay) gives results equivalent to the unlimited delay ones, when 0<λ<0.1. For 0.1<λ<0.2, a buffer length of 70 frames is sufficient.

Comparison of Variable dimension Vector Quantization and Matrix Quantization (MQ)
Chou & al compared VVVQ and entropy constrained MQ for the same complexity, but they could not train the big MQ codebooks for sequence lengths superior to 4, because of the limited size of the training database. So we have also compared the 2 approaches for the same total number of spectral vectors in the codebooks (configuration MG16 vs MQ8704). Figure 3 gives distortion-rate curves obtained on the test database for the different configurations described previously.
For small spectral bit rate (inferior to 2 bits/frame, 200 bits/s), the Multigram Quantization is superior to Matrix Quantization. But, when the comparison is done for the same number of cepstral vectors in the MG or MQ codebooks, the performances improvement is rather small for a large increase in complexity.

6. CONCLUSION
In this paper, we have given a new interpretation of Variable to Variable length Vector Quantization VVVQ and a unified view of VVVQ and Multigram Quantization.
The variable length vector quantization or multigram quantization has only been tested for a single speaker, but the new proposed interpretation should allow to use the adaptation techniques of speech recognition.
When compared with entropy constrained Matrix Quantization, the Multigram Quantization gives better distortion rate performances but at the price of an increased complexity and delay.
One shortcoming of the method is that it does not explicitly take into account the variability of the speaking rate. An attempt to linearly stretch the multigrams slightly improved the results but not sufficiently. In our future work, we will apply the Multigram Quantization on the spectral targets of a temporal decomposition.

7. REFERENCES