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

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

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

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

In this paper, we propose a new interpretation of already published research from Chou-Lockabbaugh [2] and Cernocky-Baudoin [7] on Variable to Variable Vector Quantization (VVVQ). This interpretation gives a meaning to the Lagrange multiplier used in the optimization criterium of the VVVQ, and should allow new developments as, on example, new modelization for 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 delay of 400 ms is generally sufficient.

Finally, the interest of a new codebook type was studied.

1. INTRODUCTION

For very low bit rates (< 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 Lockabbaugh [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 100 bps, in monolocutor, despite some limitations as long delay and complexity.

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

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

In this paper, VQ represents Vector Quantization, MQ Matrix Quantization (quantization of fixed length sequence of vectors), and MGQ or VVVQ Quantization of Variable length sequences of vectors (named multigrams or MG).

2. DESCRIPTION AND COMPARISON OF THE VVVQ AND MULTIGRAMS METHODS

2.1. The VVVQ method

This method quantizes the spectral envelope with variable vector 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 VVVQ : Variable to Variable 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 criterium can be written

$$\min_{S_i \in S} d_{S_i} + \lambda r_{S_i}$$

(1)

where S is the set of all possible segmentations of the database with segments of length inferior to n, $S_i$ is one such segmentation, $d_{S_i}$ is the corresponding distortion, $r_{S_i}$ the associated bit rate and $\lambda$ the Lagrange multiplier. More precisely

$$d_{S_i} + \lambda r_{S_i} = \sum_{s_j \in S_i} d_{i,j} + \lambda n_{i,j}$$

(2)

where $s_j$ is the jth 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:

$$n_{i,j} = -\log(proba(M_{i,j}))$$

(3)
$M_{i,j}$ being the codebook sequence coding $s_j$.
An iterative 2 steps EM ('Expection Maximization' algorithm [1]) is used to construct the codebook:

- step 1: Segmentation of the database using Viterbi algorithm.
- step 2: Codebook actualisation (probabilities and values of the segments).

This is a locally optimal algorithm for a given topology of the codebook.

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 $M_k$ were sequences of n or less quantization indexes. The codebook was obtained by maximizing the joint likelihood of the optimal segmenting and of the observation (sequence of quantization indexes) [3]. The segments were supposed to be independent and the optimization criterium was:

$$\max_{S_i} \prod_{M_k \in S_i} p(M_k)$$ (4)

The codebook was initialized with the sequences present in the training database, the probabilities of the sequences being initialized by counting the number of occurrences of each sequence. The EM algorithm was used to calculate the codebook and the multigrams were entropy coded. The results were insufficient for VQ size above 128, due to the great variability of indexes sequences. The results are good on the training database but not generalized on the test database.

On example, while for a simple entropy constrained VQ on 128 vectors, the spectral distortion and the average bit rate are equal respectively to 2.2 dB and 8.8 bits/vector, for multigrams of maximum length equal to 10 with a vector quantization codebook of 128 vectors, the average bit rate is 1.4 bits/vector on the training database but 12.7 bits/vector on the test string.

So a second approach was developed. Spectral vectors are no longer vector quantized. A multigram $M_k$ is a sequence of n or less spectral vectors. The observation sequence of spectral sequences is segmented in segments $U_k$ and quantized by multigrams $M_k$ in order to minimize the new criterium:

$$\max L(S) = \max \prod_{S_i \in S} L' \left( M_k \right) \left( 5 \right)$$

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

$$p'(M_k) = \frac{p(M_k)}{Q(d_k)} \left( 6 \right)$$

$$d_k = d(U_k, M_k) \left( 7 \right)$$

$$Q[d] = \begin{cases} \frac{d}{d_{max}} \text{ for } d \leq d_{max} \\ 0 \text{ otherwise} \end{cases} \left( 8 \right)$$

Where $d_{max}$ is an arbitrary constant. In this new approach the number of multigrams of each size in the initial multigrams codebook was arbitrarily a-priori limited. The multigrams codebook of fixed dimension is initialized then the EM algorithm is used to calculate the codebook iteratively. In each iteration the training spectral vectors string is segmented and quantized with the existing multigram codebook, then this codebook is modified by calculating the new centroid of each class of multigrams and the associated estimated probability by counting the number of sequences in the class.

2.3. New interpretation and comparison of the methods

While independently developed, these 2 techniques are very close. The VVVQ is mathematically better expressed and optimizes the distortion for a given rate and a given codebook structure.

The MG approach brings a different lighting. It will be here more cleanly reformulated. And with this new interpretation, the 2 approaches will be compared.

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

$$\max_{\theta} L(S_{OPT} / obs) = \max_{S} L(S_{OPT}) L(obs / S_{OPT}) \left( 9 \right)$$

$$L(S) = \prod_k p(M_k) \left( 10 \right)$$

$$L(obs / S) = \prod_k p(U_k / M_k) \left( 11 \right)$$

$U_k$ being a length $l_k$ 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 monotonic function, the criterium is equivalent to:

$$\max_k \sum_{j=1}^{l_k} \sum_{m=1}^{p} \frac{1}{2\sigma^2} (c_{k,j,m} - m_{k,j,m})^2 + \log(p(M_k)) \left( 12 \right)$$

$$\Leftrightarrow \min_k \sum_{i=1}^{l_k} d(c_{k,j,m_k,j}, m_{k,j,m_k,j}) - 2\sigma^2 \log(p(M_k)) \left( 13 \right)$$

c_{k,i,m} and $m_{k,i,m}$ are the $m^{th}$ coefficients of $j^{th}$ vectors of segment $U_k$ and multigram $M_k$. $d(c_{k,j,m_k,j}, m_{k,j,m_k,j})$
is a quadratic distance between the \( j \)th vectors of \( U_k \) and \( M_k \).
In the last equation, can be recognised the VVVQ criterium with \( \lambda = 2\sigma^2 \) and a quadratic distance on the spectral vectors.
On another hand, it is possible to interpret the arbitrary criterium of the MG method by observing that, for \( d << d_{\text{max}} \):
\[
\log(p) + \log(1 - \frac{d}{d_{\text{max}}}) \approx \log(p) - \frac{d}{d_{\text{max}}}
\]  
with \( d_{\text{max}} = 2\sigma^2 \). In we have used here a triangular probability, which is closed 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 Multigrams approach is the used spectral distortion. Chou et al. worked with a modified Itakura distortion measure while we used an euclidian distance on cepstral coefficients. With the modified Itakura measure, the precedings 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. For each new spectral vector, we search for a common point of segmentation in the buffer. That is to say that we examine if the \( k \) best possible segmentations from the origin of the buffer to the last received vector are formed by the same segments until a certain point called common segmentation point. 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
In order to increase the maximum length of multigrams, it is necessary to augment the size of the training database. On example, for \( n = 16 \), 64 multigrams by 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 MGs of length \( n/2 \) to obtain the long MGs 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 MGs of size \( n/2 \) are actualized from associated segments of size \( n/2 \) and from longer segments associated to the stretched MGs. In the last case, the actualization is done by linearly contracting the long MGs. At each iteration, the MGs of size 1 to \( n/2 \) as well as all the probabilities are saved. The obtained results (distortion - rate curves) are inferior to those obtained with a normal codebook of maximum length \( n \) and slightly better than with a maximum length \( n/2 \), but the complexity is greatly increased. In fact, after a few iterations of the EM algorithm, most of the long MGs disappear from the dictionary. So this approach has revealed to be inefficient.

5. EVALUATION
Three criteria have been considered for the evaluation: the average spectral distortion, the average bit rate and the complexity.
The spectral distortion was defined as the average over all frames of the logarithm of the spectral distance:
\[
D_{\log} = \sqrt{\int \left[ 10 \log S(f) - 10 \log \hat{S}(f) \right]^2 df}
\]  
(in dB) where \( S(f) \) and \( \hat{S}(f) \) are the power LPC-spectra with original and quantized coefficients respectively. As we were working with LPC-cestral vectors, this distance was approximated by:
\[
\hat{D}_{\log} = \mu \sqrt{\frac{1}{2} \sum_{i=1}^{10} (c_i - \hat{c}_i)^2}
\]  
where \( \mu = 10/\ln(10) \) is the conversion constant, and \( c_i \) and \( \hat{c}_i \) the original and quantized LPCC coefficients.
The bit rate is defined as the average number of bits for the coding of one spectral vector, it is in fact the number of bits by frame for the transmission of the spectrum. We have supposed that the indexes of multigrams in the Multigram codebooks were entropy coded and that the number of bits for the transmission of index \( i \) of multigram \( M_i \) was \(-\log_2(p(M_i))\), where \( p(M_i) \) is the probability of \( M_i \). The same hypothesis was applied for Vector or Matrix Quantization.
The average bit rate by frame $R$ corresponding to a
Multigrams codebook (classical or modified), is given
by the ratio of the multigram codebook entropy $H$
with the average Multigram length $\bar{I}$:

$$ R = \frac{H}{\bar{I}} = -\sum_{i=1}^{Z} p(M_i) \log_2 p(M_i) \sum_{i=1}^{Z} I(M_i) p(M_i) $$

(17)

where $l(M_i)$ and $p(M_i)$ are the length and the proba-

bility of multigram $M_i$, and $Z$ the number of multi-
grams in the codebook.

For the test string the average bit rate is obtained by:

$$ R(M) = -\frac{\sum_{i=1}^{Z} c_{test}(M_i) \log_2 p(M_i)}{N_{test}} $$

(18)

where $c_{test}(M_i)$ is the number of sequences repre-
sented by multigram $M_i$ and $N_{test}$ is the length of
the test string.

In the case of Matricial quantization on sequences of $n$

vectors (VQ is the case with $n=1$), the same formulas

are used and the average length is equal to $n$.

The spectral distortion and the bit rate are noted dis-
tortion and rate in the following figures.

The complexity by frame of the VVVQ or Multigrams

method is proportional to $\sum_{i=1}^{Z} I(M_i)$, while the com-
plexity by frame of a Matricial Quantization with se-
quences of $n$ vectors is only proportional to $Z/n$.

6. EXPERIMENTS AND QUANTITATIVE RESULTS

6.1. Database

We have used a Swiss-French monolocutor database
which is a part of PolyVar from IDIAP. It contains
telephone calls made of read sentences, spelled words,
digits, some control words and spontaneous speech.
The signal is digitized with a sampling frequency of
8 Khz and a linear 16 bits quantization. The spec-
tral vectors are 10 LPCC calculated after preemphasis
on 20 ms Hamming windows with 10ms overlapping.
The first cepstral coefficient is not used.

The approximately one hour long corpus was divided
into 213270 vectors for the training and 122903 vec-
tors for the test.

6.2. Multigrams Codebook Initialization

Different codebook initialization techniques have been
compared.

- More frequent vector-quantized multigrams ini-
tialization, where after vector-quantizing the train-
ing database with a VQ codebook of L vectors,
we used for each multigram length 1 the more fre-
quent quantized sequences of length 1.
- Matricial quantization initialization, where we ini-
itialized the multigrams codebook for each multi-
gram length with a matricial quantization code-
book. [2]

- Natural random codebook, where we initialized
  the multigrams codebook with natural spectral
  vectors sequences choosen randomly in one of the
  training speech files [5] [6]

After very few iterations of the EM algorithms, the
3 different initialization gave similar results, so we
decide to use the last one.

6.3. Results

The following results have been obtained for differents
Multigrams or Matricial codebooks topology:

- MG16, Multigram Quantization with $n$=16, 64
  Multigrams by length. There are 8704 cepstral
  vectors in the codebook.
- MQ8704, Matricial Quantization with codebooks
  containing 8704 cepstral vectors and sequences
  lengths $n$ from 2 to 20 vectors. On example,
  for $n$=4 and $n$=16 there are sequences respec-
  tively 2176 and 544 sequences in the matricial
  codebook.
- MQ1, MQ2, MQ4 are 3 different matricial quan-
tization with codebooks containing 8704 sequences
  of respective lengths 1, 2, 4 corresponding to
  8704, 17408 and 34816 cepstral vectors.
- MG5 with $n$=5 . 64 MG for length 1, 128 MG
  for length 2 to 5.

6.3.1. Limitation of the delay, buffer filling

Figure 1 gives examples of cumulated distributions
for the buffer filling, for the Multigram quantization
MG16. Figure 2 gives the average values and the
standard deviations of the buffer filling, for MG16
with $\text{lambda} \leq \text{between0and1}$.

For $k_{max}$=40 (delay=400 ms) the performances
are clearly improved with this algorithm compared to
the classical clearing of the buffer every $k_{max}$ frames.

In conclusion, a 400 ms delay gives results equivalent
to the unlimited delay ones, for $n$ <= 16 with $0 <
\lambda < 0.05$. For $n$=16 et 0.05 < $\lambda < 0.2$ a buffer length
of 60 frames is sufficient.

6.3.2. Comparison of Multigrams Quantization and
Matricial Quantization

Chou & al compared the two approaches for the same
complexity, but very rapidly it was no more possible
to train the big MQ codebook because of the limited
size of the training database. So we have also com-
pared the 2 approaches for the same total number of
spectral vectors in the MG or MQ codebooks. Fig-
ure 3 gives distortion-rate curves obtained on the test
base with Multigrams Quantization and for Matricial
Quantization for configurations.
For small bit rate (12 bits/frame, 200 bits/s), the Multigrams Quantization is superior to Matricial quantization. But, when the comparison is done with a fixed number of cepstral vectors in the dictionary the performances improvement for distortion-rate is rather small for a great increase in complexity.

\[ \lambda = 0.0, 0.01, 0.02, 0.05, 0.1, 0.2, \text{ } \kappa_{\text{max}} = 512 \text{ } n = 16 \]

---

fig 1: Cumulated distributions of the buffer filling for \( \lambda = 0.0, 0.01, 0.02, 0.05, 0.1, 0.2 \), \( \kappa_{\text{max}} = 512 \) \( n = 16 \)

---

fig 2: Average and standard deviation values of the buffer filling for MG16 and MG8

---

fig 3: Distortion-rate curves for config MG5, MG16 and some MQ configs

---

7. CONCLUSION

In this paper we have given a new interpretation and a comparison of 2 approaches of variable length vector quantization, the VVVQ and the Multigram quantization method. We have proposed a new algorithm to improve the performances when the delay is limited and constated that a delay of 400 ms is sufficient for low bit rate situations. We have compared the Multigrams quantization with Matricial quantization and obtained better results for The Multigrams quantization for low bit rates to the price of a great complexity increase. We have tried to construct long sequences in the Multigrams codebook by linearly stretching the small ones, but the obtained results were not satisfactory. The Multigrams quantization has only be applied here in the monolocator case, but the new formulation should allow to use adaptation techniques of speech recognition. One default of the method is that it does not explicitly take into account the fact that the same acoustic sequences can be uttered at different speed. In our new work we use first a temporal decomposition and then apply a Multigram segmentation and quantization on the target of the spectral decomposition.

8. REFERENCES


