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]

[1]ESIEE, BP 99, Noisy Le Grand, 93162 CEDEX, France, baudoing@esiee.fr
[2]FEIVUT Brnö, Purkynova 118, 61200 Brnö, Czech Republic, cernocky@urel.fee.vutbr.cz
[3]ENST, 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 (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-Lockabaugh [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 usefull 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 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 interpretaion 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 dictionnary 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 COMPARIZON OF THE VVVQ AND MULTIGRAMS METHODS

2.1. The VVVQ method

This method quantizes the spectral envelop 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:

$$\min_{S_i \in S} d_{S_i} + \lambda r_{S_i} \tag{1}$$

where S is the set of all possibles 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 λ the Lagrange multiplier. More precisely:

$$d_{S_i} + \lambda r_{S_i} = \sum_{s_j \in S_i} d_{i,j} + \lambda n_{i,j} \tag{2}$$

where s_j is the jth segment of S_i, $n_{i,j}$ the number of coding bits for s_j 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:

$$n_{i,j} = -\log(proba(M_{i,j})) \tag{3}$$

$M_{i,j}$ being the codebook sequence coding s_j.
An iterative 2 steps EM (Expextation 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 likehood of the optimal segmentation and of the observation (sequence of quantization indexes) [3]. The segments were supposed to be independant and the optimization criterium was:

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

The codebook was initialized with the sequences present in the training database, the probabilities of the sequences being intialized by counting the number of occurences of each sequence. The EM algorithm was used to calculate the codebook and the mutigrams 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 6.8 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 vectors is segmented in segments U_k and quantized by multigrams M_k in order to minimize the new criterium :

$$\max L(S) = \max_{S_i \in S} \prod_k p'(M_k) \tag{5}$$

Where $p'(M_k)$ 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.

$$p'(M_k) = p(M_k)Q(d_k) \tag{6}$$

$$d_k = d(U_k, M_k) \tag{7}$$

$$Q[d] = \begin{cases} 1 - \frac{d}{d_{max}} & \text{pour} \quad d \leq d_{max} \\ 0 & \text{pour} \quad d > d_{max} \end{cases} \tag{8}$$

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 comparizon of the methods

While independantly developed, these 2 techniques are very closed. 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 independant multigrams. and that the spectral vectors (size p) of the MGs are gaussian with a covariance matrix $\sigma^2 I$, I is the pxp identity matrix. The parameters θ (probabilities and MGs) of the source are identified by maximizing the likehood of the optimal segmentation S_{opt} for the observation:

$$\max_{\theta} L(S_{opt}/obs) \Leftrightarrow \max L(S_{opt}) L(obs/S_{opt}) \tag{9}$$

$$L(S) = \prod_k p(M_k) \tag{10}$$

$$L(obs/S) = \prod_k p(U_k/M_k) \tag{11}$$

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:

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

$$\Leftrightarrow \min \sum_k \sum_{j=1}^{l_k} d(c_{k,j}, m_{k,j}) - 2\sigma^2 log(p(M_k)) \tag{13}$$

$c_{k,j,m}$ and $m_{k,j,m}$ are the m^{th} coefficients of j^{th} vectors of segment U_k and multigram M_k. $d(c_{k,j}, m_{k,j})$

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 $d << dmax$:

$$log(p) + log(1 - \frac{d}{dmax}) \simeq log(p) - \frac{d}{dmax} \qquad (14)$$

with $dmax = 2\sigma^2$. In we have used here a triangular probability, which is closed a gaussian for d small compare to dmax and guaranties that d is always limited by dmax.
Another possible interpretation can be obtained by considering that the source emits variable length independant constant multigrams to which is added a centered gaussian noise of variance σ^2.
A further difference between the VVVQ and the Multigrams approach is the used spectral distotion. Chou & 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 theorical delay introduced by these methods is equal to the length of the signal.
When the delay is limited to a value of kmax frames, the performances are degraded. To limit the delay, the classical technique uses a buffer of kmax 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 n 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 kmax.
We have studied the statistical characteristics of the buffer filling for different values of n and λ. Results re 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 streching 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 streched 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 unefficient.

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:

$$D_{log} = \sqrt{\int \left[10 \log S(f) - 10 \log \hat{S}(f)\right]^2 df} \qquad (15)$$

(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-cepstral vectors, this distance was approximated by:

$$\hat{D}_{log} = \mu \sqrt{2 \sum_{i=1}^{10} (c_i - \hat{c}_i)^2} \qquad (16)$$

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 $\bar{l}$:

$$R = \frac{H}{\bar{l}} = -\frac{\sum_{i=1}^{Z} p(M_i) \log_2 p(M_i)}{\sum_{i=1}^{Z} l(M_i) p(M_i)} \quad (17)$$

where $l(M_i)$ and $p(M_i)$ are the length and the probability of multigram M_i, and Z the number of multigrams in the codebook.
For the test string the average bit rate is obtained by:

$$R(M) = -\frac{\sum_{i=1}^{Z} c_{test}(M_i) \log_2 p(M_i)}{N_{test}} \quad (18)$$

where $c_{test}(M_i)$ is the number of sequences represented by multigram M_i and N_{test} is the length of the test string..
In the case of Matricial quantization on seqences 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 distortion and rate in the following figures.
The complexity by frame of the VVVQ or Multigrams method is proportional to $\sum_{i=1}^{Z} l(M_i)$, while the complexity by frame of a Matricial Quantization with sequences of n vectors is only proportionnal 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 telehone calls made of read sentences, spelled words, digits, some control words and spontaneous speech.
The signal is digized with a sampling frequency of 8 Khz and a linear 16 bits quantization. The spectral vectors are 10 LPCC calculated after preemphasis on 20 ms Hamming windows with 10ms overlapping. The first cepstral coeeficient is not used.
The approximately one hour long corpus was divided into 213270 vectors for the training and 122903 vectors for the test.

6.2. Multigrams Codebook initialization

Different codebook initialization techniques have been compared.

- More frequent vector-quantized multigrams initialization, where after vector-quantizing the training database with a VQ codebook of L vectors, we used for each multigram length l the more frequent quantized sequences of length l.
- Matricial quantization initialization, where we initialized the multigrams codebook for each multigram length with a matricial quantization codebook. [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 respectively 2176 and 544 sequences in the matricial codebook.
- MQ1, MQ2, MQ4 are 3 different matricial quantization with codebooks containig 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 $lambdabetween0and1$.
For kmax=40 (delay=400 ms) the performances are clearly improved with this algorithm compared to the classical clearing of the buffer every kmax 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 trames is sufficient.

6.3.2. Comparizon 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 compared the 2 approaches for the same total number of spectral vectors in the MG or MQ codebooks. Figure 3 gives distortion-rate curves obtained on the test base with Multigrams Quantization and for Matricial Quantization for configurations:

For small bit rate (¡2 bits/frame, 200 bits/s), the Multigrams Quantization is superior to Matricial quantization. But , when the comparizon is done with a fixed number of cepstral vectors in the dictionary the performances improvement for distortion-rate is rather small for a a great increase in complexity.

fig 1:cumulated distributions of the buffer filling for λ =0,0.01,0.02,0.05,0.1,0.2, kmax=512 n=16

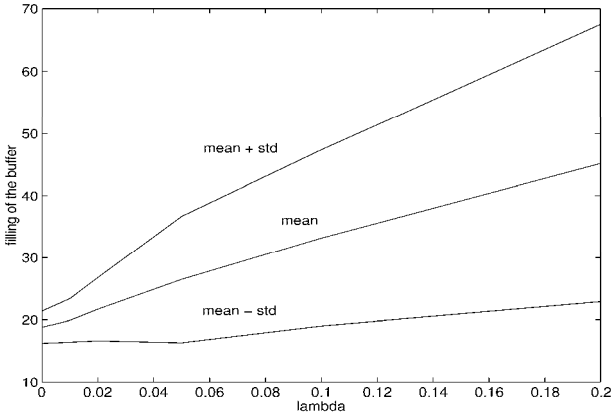

fig 2: Average and standard deviation values of the buffer filling for MG16 and MG8

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 comparizon of 2 approaches of variable length vector quantization, the VVVQ and the Multigram quantiztion 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 streching the small ones, but the obtained results were not satisfactory. The Multigrams quantization has only be applied here in the monolocutor 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 spectarl decomposition.

8. REFERENCES

[1] N. M. Laird A. P. Dempster and D. B. Rubin. Maximum likelihood from incomplete data with the em algorithm. *J. Roy. Stat.*, 39(1):1–38, 1977.

[2] P. A. Chou and T. Lookabaugh. Variable dimension vector quantization of linear predictive coefficients of speech. In *Proc. IEEE ICASSP 94*, pages I–505–508, Adelaide, June 1994.

[3] S. Deligne and F. Bimbot. Language modelling by variable length sequences: Theoretical formulation and evaluation of multigrams. In *Proc. IEEE ICASSP 95*, pages 169–172, Detroit, USA, 1995.

[4] G. Baudoin J. Černocký and G. Chollet. speech spectrum representation and coding using multigrams with distance. In *Proc. IEEE ICASSP 97*, pages –, Munich, Germany, April 1997.

[5] R. Schwarts S. Roucos and J. Makhoul. Segment quantization for very low bit rate speech coding. In *Proc. IEEE ICASSP 82*, pages 1565–1568, Paris, France, April 1982.

[6] R. Schwarts S. Roucos and J. Makhoul. A segment vocoder at 150bits/s. In *Proc. IEEE ICASSP 83*, pages 61–64, 1983.

[7] J. Černocký and G. Baudoin. Représentation du spectre de parole par les multigrammes. In *Proc. XXI-es Journées d'Etude sur la Parole*, pages 239–242, Avignon, France, June 1996.