SPEECH SPECTRUM REPRESENTATION AND CODING USING MULTIGRAMS WITH DISTANCE

Jan Černocký [1,2], *Geneviève Baudoin* [1] *and Gérard Chollet* [3]

[1]ESIEE, Département Signal et Télécommunications, Noisy-le-Grand, France, {cernockj,baudoing}@esiee.fr
[2]Technical University of Brno, Institute of Radioelectronics, Brno, Czech Republic, cernocky@urel.fee.vutbr.cz
[3]ENST, Département Signal, Paris, France, chollet@sig.enst.fr

ABSTRACT

The multigrams allow us to split a string of symbols into a stream of variable length sequences. The direct application of this method to vector-quantized speech spectra fails, we develop an extension of the method called modified multigrams or multigrams with distance. The algorithm for modified multigram dictionary training as well as experimental results are presented. We found a significant improvement of rate/distortion ratio in comparison to vector quantization with small codebooks. For precise spectrum representation, this method is less suitable and we see its application rather in speech segmentation or in very low bit rate coding.

1. INTRODUCTION

In standard speech coding systems working on low bit rates, the signal is modelized by source and filter. The filter parameters called "spectral vector" (determining the synthesized speech envelope or shortly "spectrum") are usually transmitted once per frame, without taking into account the inter-frame dependencies. In our work, we investigated the representation of spectral vectors on segmental basis. The vector sequences are represented using a pre-trained set of variable length sequences called *multigrams*. We used two methods for the matching of input vectors with the sequences in the dictionary and for dictionary creation: *classical multigrams* [1, 4, 2], where the sequences of vector quantized vectors are matched by symbol strings equivalence, and *modified multigrams*, where the choice is done on minimum-distance basis.

2. CLASSICAL MULTIGRAMS

In this case, the string of vectors $\boldsymbol{X} = \boldsymbol{x}_1 \dots \boldsymbol{x}_N$ is converted to string of symbols $W = w_1 \dots w_N$ by means of Vector Quantization (VQ). The multigram method is described in the articles of Bimbot et al. [1, 4, 2] and we limit us to a brief description: The string of symbols W is segmented into variable length (1 to n) sequences $S_1, S_2, \dots S_q$ using decision oriented likelihood maximization:

$$L(W) = \max_{\{B\}} \prod_k p(S_k) \qquad (1)$$

This work is supported by the French Government scholarship No. 94/4516 and by the Elite programme of Texas Instruments.

where $p(S_k)$ are the probabilities of sequences and $\{B\}$ is the set of all possible segmentations. The number of all these segmentations is large and we use a Viterbi-based algorithm to find the optimal one. The probabilities of sequences are not known a-priori and we must create a *dictionary* of sequences. After an initialization step, an iterative procedure is run, consisting of segmentation (using Eq. 1) and of probabilities reestimation. As a result, we dispose of a dictionary of Z characteristic sequences M_i with associated probabilities $p(M_i)$.

In the representation (coding) phase, the string is segmented using Eq. 1 and the sequences are represented by multigrams of the dictionary. In this case, the representation introduces the same spectral distortion as the VQ, because the symbols in the input string and in the coded one are equal.

3. MODIFIED MULTIGRAMS

Here, no conversion of input vectors to symbols is performed, and the method uses a distance measure between spectral vectors rather than symbol strings equivalence to match the input vectors with the dictionary sequences. We define the distance of two sequences of vectors $\boldsymbol{A} = \boldsymbol{a}_1 \dots \boldsymbol{a}_l$ and $\boldsymbol{B} = \boldsymbol{b}_1 \dots \boldsymbol{b}_l$ of equal length l as the mean distance of corresponding vectors:

$$D(\boldsymbol{A}, \boldsymbol{B}) = \frac{1}{l} \sum_{i=1}^{l} d(\boldsymbol{a}_i, \boldsymbol{b}_i) \qquad (2)$$

In our case, $d(\boldsymbol{a}_i, \boldsymbol{b}_i)$ is the Euclidian distance.

Similarly as in the previous case, we are looking for the segmentation of a string of vectors and for a set of representative sequences of variable length, consisting here of unquantized vectors: $\boldsymbol{M}_i = \boldsymbol{m}_{i,1} \dots \boldsymbol{m}_{i,l_i}$. The optimal segmentation of $\boldsymbol{X}$ into sequences $\boldsymbol{U}_1 \dots \boldsymbol{U}_q$ of length 1 to n is given by maximization of the quantity:

$$L'(\boldsymbol{X}) = \max_{\{B\}} \prod_k p'(\boldsymbol{M}_k) \qquad (3)$$

over the set of all possible segmentations $\{B\}$. The main difference from the previous likelihood definition (Eq. 1) is the use of *penalized probability* p'. It is no more a probability of a sequence but a modified probability of the code-multigram $\boldsymbol{M}_k$ representing the sequence $\boldsymbol{U}_k$. The multigram $\boldsymbol{M}_k$ is chosen among all multigrams of length $l(\boldsymbol{U}_k)$

to minimize the distance $D(M_k, U_k)$. The penalized probability is derived from the multigram probability (found in the dictionary) by:

$$p'(M_k) = Q[D(U_k, M_k)]\, p(M_k) \tag{4}$$

where $p(M_k)$ is the probability of code-multigram M_k and $D(U_k, M_k)$ is the distance between this multigram and the sequence U_k. The function $Q[.]$ must penalize the probability of multigram in function of its distance from the represented sequence. We used a simple partially linear function defined by:

$$Q[D] = \begin{cases} 1 - \frac{D}{D_{max}} & \text{for} \quad D \leq D_{max} \\ 0 & \text{for} \quad D > D_{max} \end{cases} \tag{5}$$

where D_{max} is a constant giving the maximal distance for which p' may be nonzero. Similarly as for "classical" multigrams, we need a dictionary of code-multigrams. After initialization (see Subsection 5.3) the following iterative procedure consists of 3 steps:

- *Segmentation* which is found by maximizing the quantity $L'(\boldsymbol{X})$ (see Eq. 3).
- *Reestimation of probabilities*:

$$p(M) = \frac{c(M)}{C} \tag{6}$$

  where $c(M)$ is the number of sequences represented by multigram M and C is the total number of sequences in the optimal segmentation.
- *Recalculation of code-multigrams* - new multigram M is calculated as a centroid of all sequences represented by M in the optimal segmentation.

In the representation (coding) phase, the string is segmented using Eq. 3 and the sequences of vectors are represented by the multigrams in the dictionary. In this case, the distortion introduced by the representation must be evaluated.

4. EVALUATION

The *distortion* SD introduced by the spectrum representation is evaluated as the mean over all frames of the logarithmic spectral distance:

$$D_{log} = \sqrt{\int \left[10 \log S(f) - 10 \log \hat{S}(f)\right]^2 df} \tag{7}$$

(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:

$$\hat{D}_{log} = \mu \sqrt{2 \sum_{i=1}^{10} (c_i - \hat{c}_i)^2} \tag{8}$$

where $\mu = 10/ln(10)$ is the conversion constant, and c_i and $\hat{c}_i$ the original and quantized LPCC coefficients.

In the *rate* evaluations, we compare the average number of bits necessary for the transmission of one spectral vector for different methods. For both VQ and multigrams, it is the index pointing to a code-vector or code-multigram which needs to be transmitted. We assume the ideal entropy coding of those indices, so that the number of bits necessary for the transmission of index i is $\log_2 p(o_i)$, where o_i stands for code-vector $\boldsymbol{y}_i$ or code-multigram M_i. For VQ, the average rate for the training string is given by:

$$H(V) = -\sum_{i=1}^{L} p(\boldsymbol{y}_i) \log_2 p(\boldsymbol{y}_i) \tag{9}$$

and for the test one by:

$$R(V) = -\frac{\sum_{i=1}^{L} c_{test}(\boldsymbol{y}_i) \log_2 p(\boldsymbol{y}_i)}{N_{test}} \tag{10}$$

where $c_{test}(\boldsymbol{y}_i)$ is the number of vectors of the test string represented by code vector $\boldsymbol{y}_i$, and N_{test} is the length of the test string. For the multigrams (classical or modified), the average rate for the training string is given by:

$$H'(M) = -\frac{\sum_{i=1}^{Z} p(M_i) \log_2 p(M_i)}{\sum_{i=1}^{Z} l(M_i) p(M_i)} \tag{11}$$

where $l(M_i)$ is the length of multigram M_i, and for the test string by:

$$R(M) = -\frac{\sum_{i=1}^{Z} c_{test}(M_i) \log_2 p(M_i)}{N_{test}} \tag{12}$$

where in this case, $c_{test}(M_i)$ is the number of sequences represented by multigram M_i.

5. EXPERIENCES

5.1. Database, Vector Quantization

Approximately 1 hour of mono-speaker telephone speech with deleted pauses was used for the experiments. The spectrum was parametrized by 10 LPCC coefficients calculated on 20 ms frames with 10 ms overlapping. The corpus was divided into 213270 vectors for training and 122903 vectors for tests. The VQ codebook training was performed by a simple LBG algorithm for codebook lengths $L = 2, 4, 8, 16, 32, 64, 128, 256, 512$. The training and test set of vectors were quantized using those codebooks and we obtained 9 training and 9 test strings of symbols. The rate-distortion curve of VQ (see Fig 1) was compared with those obtained by multigram methods.

5.2. Classical multigrams

Experimental conditions and results for classical multigrams applied to cepstral vectors were previously published in [5]. The summary of the results is the following: we obtained strong decrease of $H'(M)$ in comparison to $H(V)$ for the training strings, but catastrophical results ($R(M) > R(V)$) for the test ones. This fact together with an enormous size of the dictionaries proved a strong overlearning and incapability of classical multigrams to represent sequences of vector quantized spectral vectors due to their high variability - we can say, that for larger codebooks of VQ, almost

each long sequence is unique. The main problem of classical multigrams is caused by the constraint of equivalence between a sequence in the dictionary and on the input: the modified multigram method with distance notion is the result of efforts to bypass this constraint.

5.3. Modified multigrams

The crucial problem of the modified multigram method is the initialization of the dictionary. Not only we have to choose the numbers of multigrams of different lengths and their initial values, but we must also initialize their probabilities. In our experiences, we used the vector quantized training string and classical multigrams without the iterative refinement for this step. For the VQ of dimension L, the number of initial 1-grams is L. The sections of 2- to $n-$grams were initialized each with $2L$ classical multigrams with the highest probability (the symbols were converted back to vectors).

The modified multigrams were trained for maximal length $n = 5$ and for initializations $L = 2, 4, 8 \ldots 512$. Maximal distance D_{max} was varied in the experiences from 0 to 1.0 with 0.1 step. The resulting rate-distortion curves for $L = 2, 8, 32, 128, 512$ initializations are presented on Fig. 1. For small initialization codebooks, we observe a "u"-like curve indicating the decrease of distortion for a given rate. For larger ones, this form does not appear, but the rate-distortion curves are all situated below that of VQ. A detailed example is given in Table 1: the comparison of distortions for approximately equal rate of 3.5-3.9 bits/vector. We see, that for modified multigrams with $L = 512$ initialization, we obtain a 0.664 dB improvement of the distortion (on the test string) in comparison to VQ with 16 codevectors. If limited to VQ, we would need 6.8 bits/vector to reach the same distortion.

The changing of proportion of of 1- to 5-grams in the resulting dictionary is illustrated on Fig. 2. When "relaxing" the maximal distance, more longer multigrams appear, but this "relaxation" has a negative effect on the overall distortion.

Despite the results obtained with small dictionaries, for a precise spectrum representation (for which we need on contrary a large dictionary), the method is not completely without problems:

1. the choice of maximal distance D_{max} is critical for the resulting ratio spectral distortion/entropy.
2. the computational load for large multigram dictionaries is significantly higher in comparison to VQ.
3. the overlearning causes that the results obtained on the learning string are not validated on the test one (see Fig. 1, initialization $L = 512$).

5.4. Split vector quantization

In order to improve the precision of spectrum description and decrease the computational load, we tested the combination of modified multigrams and split vector quantization (SVQ). The vector of cepstrum coefficients was divided into 3 subvectors (with 3, 3 and 4 coefficients) and the modified multigrams were applied to each of them independently. After vector quantization with $L_1 = 256, L_2 = 64, L_3 = 64$ and classical multigram initialization, we performed the dictionary training for maximal distances $D_{max} = 0.0 \ldots 1.0$. For $D_{max} = 0.2$ for all three subvectors, we obtained the sum of rates per symbol 7.359 bit. When reassembling the quantized vectors, we could evaluate the overall spectral distortion which was 2.232 dB. When comparing this point with Fig 1, we see, that a simple VQ performs better than this sophisticated method. Another problem was the instability of synthesis filters with cepstral coefficient quantized on the sub-vector basis.

VQ		$H(V)$	SD_{train}	$R(V)$	SD_{test}
L=16		3.881	2.892	3.838	2.841
MMG	D_{max}	$H'(M)$	SD_{train}	$R(M)$	SD_{test}
L=16	0.3	3.797	2.826	3.776	2.785
L=32	0.4	3.563	2.514	3.523	2.498
L=64	0.4	3.647	2.349	3.613	2.356
L=128	0.4	3.602	2.235	3.595	2.270
L=256	0.4	3.574	2.143	3.647	2.211
L=512	0.4	3.544	2.055	3.701	2.177

Table 1: Comparison of VQ and modified multigrammes with different initializations for approxiamtely equal rate of 3.5-3.9 bits/vector.

6. COMPARISON WITH VVVQ

The Variable to Variable Length Vector Quantization (VVVQ) introduced by Chou and Lookabaugh in [3] can be compared to modified multigrams approach. The entropy constrained quantization of variable length sequences of spectral vectors by symbols of variable length uses the same tools as our method: Viterbi-based segmentation of the string of vectors and entropy coding of transmitted symbols. However, the criterion of the optimal segmentation differs: in our case, the multigram $\boldsymbol{M}_k$ representing the sequence $\boldsymbol{U}_k$ is chosen in order to minimize the distance and only then, its probability is penalized by this distance (see Eqs. 4 and 3). In [3], the distortion and rate are minimized jointly, using a Lagrange multiplicator in the segmentation formula:

$$B_{opt} = \arg\min_{\{B\}} \sum_k [d(\boldsymbol{U}_k, \boldsymbol{M}_k) - \lambda \log_2 p(\boldsymbol{M}_k)] \qquad (13)$$

(when using the same notations as in Section 3). The comparison of performances of those two methods was unfortunately not possible due to using of different speech data and of different distance evaluation.

7. CONCLUSION

In our work we tested the performances of different multigram methods when applied to speech spectra representation. The ideal result would be an improved rate/distortion ratio for arbitrary precision of spectrum description.

First, classical multigrams were used to represent the sequences of vector quantized vectors. The results, described in detail in [5], are not satisfactory: only for small VQ codebooks we are able to find the typical sequences, for larger ones, this method performs well only for training strings which is a proof of an overlearning.

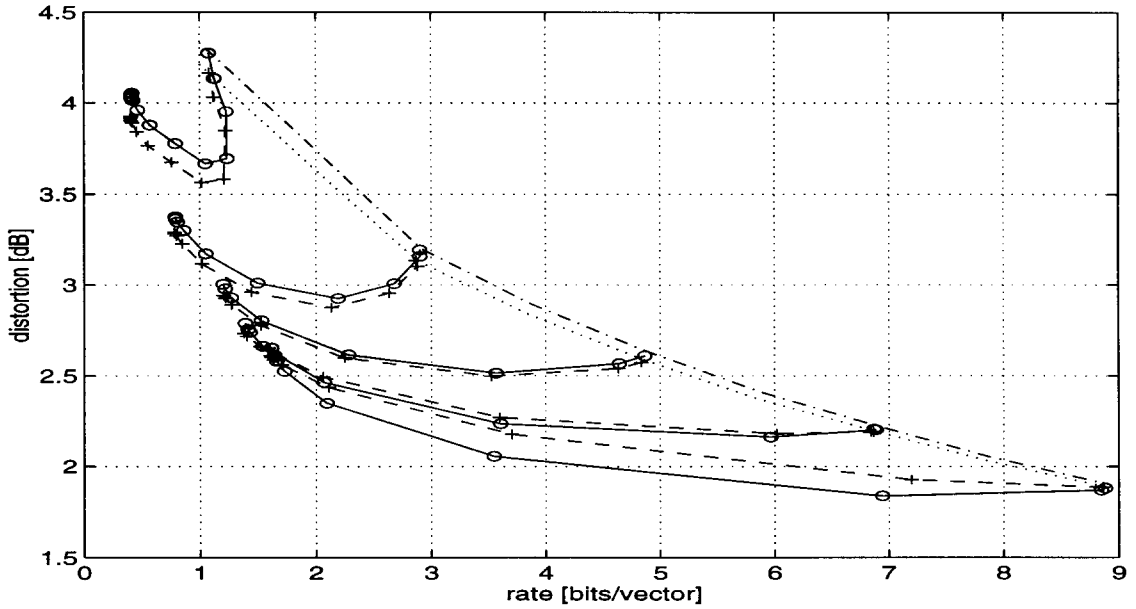

Figure 1: Rate-distortion curves for VQ and modified multigrams (MMG). Lines: dashdot - VQ learning, dotted - VQ test, solid - MMG learning, dashed - MMG test. The couples of MMG curves stand from top to bottom for initializations: $L = 2, 8, 32, 128, 512$.

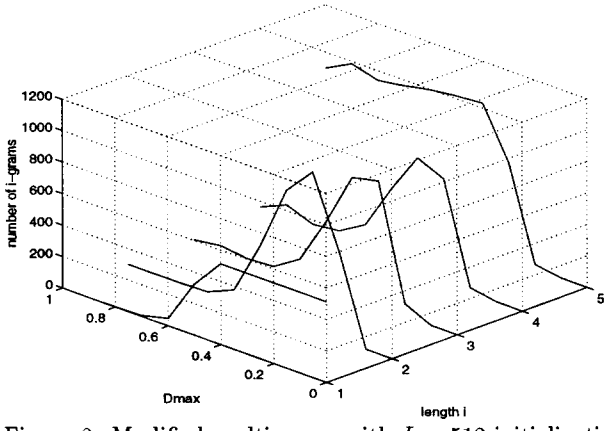

Figure 2: Modified multigrams with $L = 512$ initialization: 1- to 5-grams proportions as a function of D_{max}.

The method of modified multigrams performs better in terms of generalization of typical spectral sequences. By the introduction of distance, we attain a certain flexibility in comparison to classical multigrams. However, for precise spectrum representation with large multigram dictionaries, the problems persist and are mentioned in Subsection 5.3: in these cases, the method is also disposed to overlearning and the computational load is significant.

Multigrams were tested with SVQ, too, but without a great succes. The instability of synthesis filter could be improved by using other coefficient than LPCC or by polynom stabilization, but it is not probable that modified multigrams used with SVQ would perform better than SVQ itself.

We can conclude, that the performance of multigrams in speech spectrum representation is not fully satisfactory and that their use in direct coding of spectral parameters is arguable. However, these methods are an excellent tool for detecting the typical patterns of variable length. In our opinion, the major domain of use of multigrams is in conjunction with time alignment methods (DTW, temporal decomposition) as a preprocessing for another method of representation. Our future work will be conducted in this direction.

8. REFERENCES

[1] F. Bimbot, R. Pieraccini, E. Levin, and B. Atal. Modèles de séquences à horizon variable: Multigrams. In *Proc. XX-èmes Journées d'Etude sur la Parole*, pages 467–472, Trégastel, France, June 1994.

[2] F. Bimbot, R. Pieraccini, E. Levin, and B. Atal. Variable length sequence modelling: Multigrams. *IEEE Signal Processing Letters*, 2(6):111–113, June 1995.

[3] 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.

[4] 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.

[5] 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.