# TOWARDS A VERY LOW BIT RATE SEGMENTAL SPEECH CODER

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

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

## ABSTRACT

In this paper we are trying to define a novel scheme for speech coding on segmental basis. The goal is 100–200 bit/s coding rate in multi-speaker and multi-lingual environment. The main part of the algorithm is the research and modelization of typical spectral sequences. We have performed this search using temporal decomposition (TD), vector quantization (VQ) and multigram (MG) techniques, on a mono-speaker database. We report the results in terms of lengths and numbers of typical spectral sequences, and we are discussing their phonetical relevance. The following steps will be a modelization of sequences using Hidden Markov Models (HMM), and a study on resynthesis and speaker adaptation.

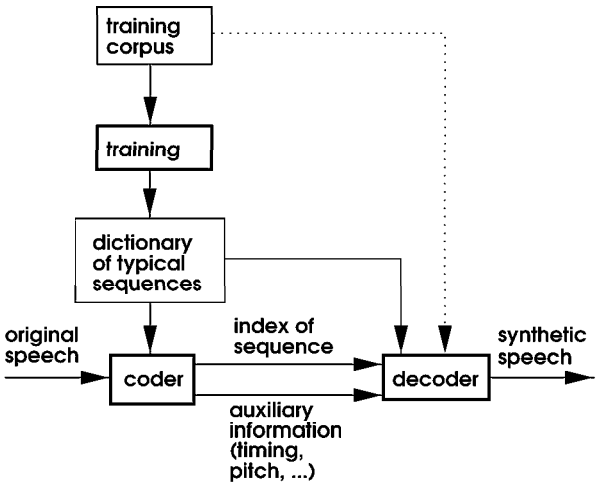


Figure 1: Scheme of the segmental coder.

## 1. INTRODUCTION

Standard low-bit rate techniques used in present applications (FS DoD at 4.8 kbit/s, LPC10 at 2.4 kbit/s) need, in some situations, to be replaced by a coding scheme reaching lower bit rates (hundreds of bit/s). "Black boxes" for airplanes, tapeless answering machines, speech storage, and military applications can be mentioned rather than standard telecommunication ones, where high quality and limited delay are requested. At those rates, one must leave the classical coding schemes working frame-by-frame for the algorithms using larger speech segments. The scheme of such coder can be seen on Figure 1. For the spectrum representation, which is the biggest "consumer" of bit rate, Chou et al. [4] suggested the Variable to Variable length Vector Quantization (VVVQ). We worked on a similar approach which we called Modified Multigrams (MMG) [6, 7]. Section 2 gives a very brief overview of these experiences. However, we found the multigrams themselves to be insufficient for a significant decrease in the bit rate; therefore we are looking for a method using multigrams for the research of typical spectral sequences, but not for the speech spectrum modelling itself. In section 3 we present a general scheme of the novel algorithm, where the search of sequences and their modelization are divided. Section 4 deals in detail with the method of search of typical spectral sequences, section 5 presents the experiences and results. The following section 6 discusses the application of HMM to segmental coding, and the synthesis and adaptation issues. We conclude in section 7.

This work is supported by the French Government scholarship No. 94/4516 and by the grant No. VS97060 of the Ministry of Education, Youth and Sports of the Czech Republic.

## 2. SPECTRUM CODING BY MODIFIED MULTIGRAMS

The *multigram segmentation* originally proposed by Bimbot et al. in [3], is a method for the division of a string of symbols $W$ into variable length sequences (1 to $n$) using a decision oriented likelihood maximization:

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

where $p(S_k)$ are the probabilities of sequences and $\{B\}$ is the set of all possible segmentations. To be able to represent speech spectra sequences, we modified the method by adding a distance notion to the evaluation of segmentation likelihood. A detailed description of experiences and their results can be found in [6, 7], but we can summarize, that for a significant decrease in bit rate, the spectral distortion deterioration is too important. We have qualified the modified multigrams themselves to be unsuitable for the low bit rate speech coding. The main drawback of multigrams is the time rigidity (a spectral sequence can not be represented by a multigram of different length). These disadvantages conducted us to define a new scheme of segmental coder, which we describe in the following section.

## 3. NEW SCHEME FOR SEGMENTAL CODING

During our previous experiences with the segmental coding we found, that the etaps of characteristical segments search and spectrum representation must be separated. We suggest the following five steps to build an algorithm for very low bite-rate coding:

1. **Non-supervised search of characteristical segments.** We propose the using of TD for the timing normalization, and VQ and MG of the target vectors to find the typical sequences. Following two sections deal with this point in detail.
2. **Clustering and modelling of segments.** The same technique as that used in 1. (TD+MG) could be employed, but we must note, that a badly quantified spectral target can "destroy" a sequence, which will not be able to be represented by a multigram. If we return from targets and interpolation functions of TD to spectral vectors, those can be represented by a set of HMMs.
3. **Segment recognition.** The segmentation and segment recognition can be done using techniques known from continuous speech recognition. Only the index of HMM and a timing information must be transmitted from the coder to the decoder.
4. **Segment reconstruction.** While the points 1.–3. are common for speech recognition and segmental coding, there is no need to reconstruct the original speech in the recognition. To obtain the complete information, the pitch and the energy of speech must be reconstructed as well as the spectrum. Another problem is the smoothing in concatenations of segments.
5. **Adaptation.** The resulting set of typical segments will be strongly dependent on the database used for the training. Several approachs can be considered to overcome the inter-speaker variability (normalization of voices to a generic one, voice modification).

## 4. SEARCH OF TYPICAL SPECTRAL SEQUENCES

The first division of the speech signal is into active and passive parts using a voice avtivity detector (VAD). For the search of sequences, only active parts are taken into account. The signal is parametrized of frame basis by a set of spectral coefficients, which form the $P \times N$ matrix $Y$, where $P$ is the number of coefficients and $N$ the number of frames.

As next step, this matrix is separated into limited amount of spectral *events*, each consisting of a *target* and an *interpolation function (IF)* using Temporal Decomposition (TD), introduced by Atal in [1] and refined by Bimbot in [2]. The spectral parameters are approximated by a product of two matrices:

$$Y = G\Phi \tag{2}$$

where $G$ is a $P \times M$ target matrix and $\Phi$ is a $M \times N$ matrix of interpolation functions, concentrated in time (the function is non-zero only on the interval $[begin_i, end_i]$. The number $M$ of events is inferior to $N$. The method used for this decomposition is a short-time SVD with an adaptive windowing, with post-processing of interpolation functions (smoothing, decorrelation) and with iterative refinement of $G$ and $\Phi$. It is described in detail in [2]. We can not determine the exact localization of an event in the time, but we can approximate it as a gravity center of the corresponding interpolation function.

Then, the target vectors are quantified using a simple VQ with low-size codebook to obtain a set of symbols. These are the input into the "classical" (without the notion of distance) multigram method, looking for characteristical repeating patterns of variable length in the training string. The process consists of an initialization of dictionary using all occurences of $1-$ to $n-$symbol sequences, and of iterations of segmentation (Eq. 1) and probabilities reestimation. Two modifications were added to the original method [3]:

- introduction of **forced segmentation** on the borders of parts. Those are determined by VAD, and the resulting training string is created by their concatenation, so no mutligram should cross their borders.
- introduction of **minimum occurence number** for the multigram dictionary entries. In the original work, a penalized probability evaluation was used to prune the dictionary, but as this pruning does not control directly the number of representants of each MG in the training string, we used thresholds for the numbers of occurences.

Using this procedure, we obtain a set of variable length charactersitical spectral sequences. The time variability is introduced by two factors: the different length of interpolation functions of TD and by the variable length of multigrams.

## 5. TYPICAL SEQUENCES – EXPERIENCES AND RESULTS

We used one speaker data from French Swiss DB *Polyvar* created at IDIAP. It is recorded over telephone, with $F_s = 8000$ Hz and 16 bit quantization. The set of 218 calls was divided into training ($\frac{4}{5}$) and test ($\frac{1}{5}$) sets. Only the training set was used for the search. The signal was parametrized using 10 LAR coefficients in frames of 20 ms, with overlapping of 10 ms. In the same time, the pitch (using FFT-cepstrum on 400 ms frames) and energy were computed. The voice activity was detected using one absolute and one relative energy thresholds, and the raw decisions were smoothed using a 11-tap OR-filter ("all around must be passive to consider a frame passive"). We obtained 5.2 hours of active speech containing 15813 active parts and $1.8 \times 10^6$ frames.

The TD was done using the `td95` package of Frederic Bimbot[1]. The parameter controlling the number of spectral targets was empirically set to have approximately the same numer of events per second as the phonetical rate (15 events/sec). The mean length of one interpolation function is 87 ms. The total number of events in the training

[1]Thank you very much, Fred !

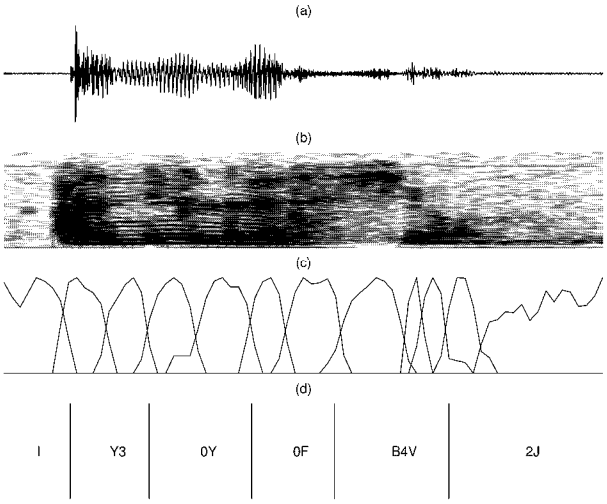


Figure 2: Example for the French word "annulation". a) signal, b) spectrogram, c) TD interpolation functions, d) MG segmentation.

corpus is 280273. An example of TD can be seen on Figure 2c.

The TD target vectors were quantified using VQ with 32 code-vectors. For the codebook training, we used an LBG algorithm with $2 \rightarrow 4 \rightarrow \ldots \rightarrow 32$ splitting. For simplicity, the code-vectors are marked by letters A...Z, 0...5. The MG dictionary training and segmentation was performed on the symbol string resulting from the VQ quantization. The maximal length of sequence was $n = 5$. We performed 10 iterations of the segmentation–reestimation cycle. The borders of parts gave us the imposed segmentation. The thresholds for minimum number of occurences of one sequence were two: the first was applied right after the initialization of dictionary, the second in the iterations. Both thresholds were set to min. 20 representants of one sequence in the training string. The numbers of $1-$ to $5-$grams in the resulting dictionary are given in Table 1. An example of multigram segmentation can be seen on Figure 2d.

### 5.1. Phonetical relevance of sequences

One dictionary entry represents several ($\geq 20$) speech segments in the training corpus. We tried to find, if the segments represented by the same sequence are phonetically coherent. An example of such comparison is shown on Figure 3: we took the most frequent sequence from $3-$grams, "FQ3" and we were looking for the speech segments represented by this sequence. Phonetically, the signals contain a fricative "s" and a nasalized "a". In some representations, we observe the substitution of "s" for an unvoiced plosive "t". For some signals, the final "ã" does not end properly, but we can hear artefacts from following interpolation functions ("sãẽ" in FQ3.3). Speech signals for this and other examples can be found on Web page `http://www.fee.vutbr.cz/~cernocky/English.html` as `wav`-files.

| | no. of sequences |
|---|---|
| 1-grams | 32 |
| 2-grams | 627 |
| 3-grams | 478 |
| 4-grams | 32 |
| 5-grams | 2 |
| Total in the dictionary | 1171 |
| Mean length [events] | 1.936 |

Table 1: Numbers of characteristical sequences in the resulting MG dictionary

Generally, the sequences are phonetically coherent, sometimes with the above mentioned problems: substitutions of sounds with similar character, and not clear beginnings and ends. In our opinion, the former problem is caused by the low dimension of VQ used. Also, we used only the LPC spectra for the quantization, without an energy or voicing criterion. The later problem comes from the nature of TD, where we had to determine the point separating two events in time. For two neighbouring interpolation functions $p$ and $q$ (where $q = p + 1$) we place the border to the mean value of $end_p$ and $begin_q$.

It was also found that the human evaluation of coherence of sequences is not objective – a strong event on the beginning of sequence (a plosive for example) is attracting attention and we are less able to evaluate the middle and end of sequence, especially in case of very short ones.

## 6. HMM, SYNTHESIS, AND ADAPTATION

The combination of TD+VQ+MG is used only for the first segmentation and labelling of training corpus. As the next etap, we are going to train one HMM for each dictionary sequence, and the coding itself will be done on similar base as connected word recognition. Next, we need to resynthesize the sequences in the decoder and to pay attention to adaptation issues. We are only beginning with the experiences, but this section contains some reflections about these problems.

### 6.1. HMM and "How to obtain desired bit rate ?"

Having a labelled training corpus, we can train a set of models using standard training methods. However, we must think about number of states per sequence and about the nature of state distributions. For an $i-$gram, we suggested either $i$ states (one per original TD event) or $2i + 1$ states, where event transitions should be better modelized. For the number of possible state distributions, we will certainly not able to have an independent one for each state. We suggest the tying of distributions for the states represented originally by the same VQ code-vector. For $2i + 1$ states per model, we should have 32 distributions for constant parts of IF plus one distribution per possible event to event transition.

Each HMM will have its probability a-priori $p(M_i)$ defining a simple "language model" and the segmentation will be performed by maximization of well known likelihood:

$$L = \prod p(M_i)p(O_i|M_i) \qquad (3)$$

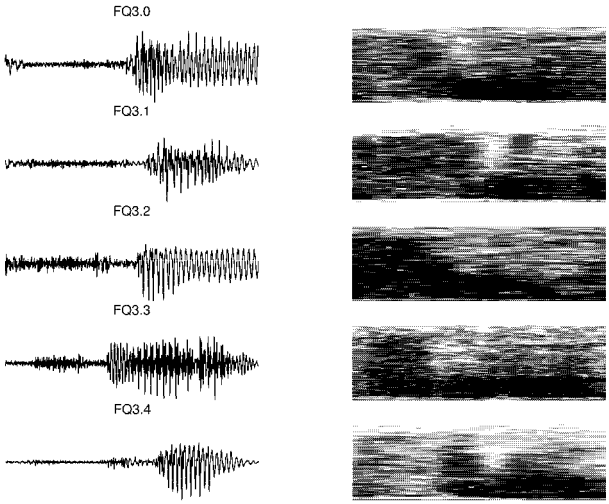


Figure 3: Five speech segments from the training string represented by the same sequence "FQ3". Signals and spectrograms.

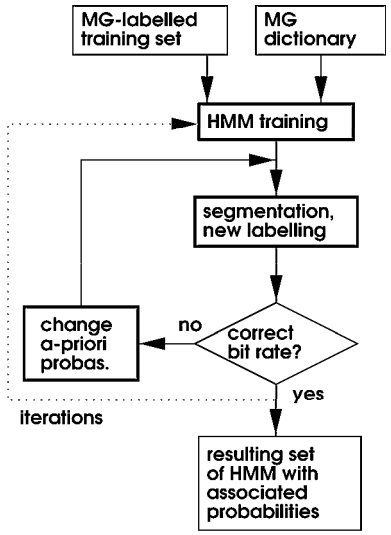


Figure 4: Algorithm for training of HMMs for sequences.

where $O_i$ is a sequence of observations, over the set of all possible segmentations and models. As the lengths of sequences representing the spectrum are directly linked to the resulting bit rate, we proposed an iterative algorithm to adjust the probabilities a-priori. These probabilities are initially given by the MG dictionary, then readjusted using the scheme on Figure 4.

### 6.2. Resynthesis of segments

For the reconstruction, we must transmit not only the spectrum information but also the pitch/voicing and energy. Having the typical spectral sequences, we hope to be able to fing characteristical patterns also for those parameters. The synthesis itself can be done using standard LPC synthesizer (impulsions or noise excited filter) or by a PSOLA based method. In this case, the decoder must dispose of the training speech (or at least of several examples for each sequence) which is marked by dotted line on Figure 1. Another important issue is the smoothing on segment to segment transitions.

### 6.3. Speaker adaptation

The spectral sequences are strongly dependent on the speaker(s) who created the training database. To be able to code any speaker, we are considering the methods known from recognition: the normalization of voice to a generic one, with an adaptation on the decoder side [5].

## 7. CONCLUSION

Our work was aimed to the search of typical spectral sequences of variable length for the very low bit rate coding. We have found a set of sequences using the combination of TD, VQ and MG, and we observed a sufficient phonetical coherence of speech segments, represented by the same sequence. As next step, we are going to modelize these sequences by HMM and use them for the segmentation and coding. As the experiences are just on the beginning, this article contains only several reflections on these topics; we are going to present the results in our future publications.

## 8. REFERENCES

[1] B. S. Atal. Efficient coding of LPC parameters by temporal decomposition. In *Proc. IEEE ICASSP 83*, pages 81–84, 1983.

[2] F. Bimbot. An evaluation of temporal decomposition. Technical report, Acoustic research departement AT&T Bell Labs, 1990.

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

[4] 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.

[5] K. Choukri. *Quelques approches pour l'adaptation aux locuteurs en reconnaissance automatique de la parole.* PhD thesis, École nationale supérieure des télécommunications (ENST), Paris, November 1987.

[6] 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.

[7] J. Černocký, G. Baudoin, and G. Chollet. Speech spectrum representation and coding using multigrams with distance. In *Proc. IEEE ICASSP 97*, pages 1343–1346, Munich, Germany, April 1997.