VERY LOW BIT RATE (VLBR) SPEECH CODING AROUND 500 BITS/SEC

Marc Padellini[(1)], *François Capman*[(1)], *and Geneviève Baudoin*[(2)]

(1), Thales Communications,
160, Bd de Valmy , BP 82, 92704 Colombes, CEDEX, (France)
(2), ESIEE, Telecommunications Systems Laboratory
BP 99 , 93162 Noisy-Le-Grand, CEDEX, (France)

ABSTRACT

New solutions to Very Low Bit Rate speech coding have been recently proposed based on speech recognition and speech synthesis technologies, [1,2,3,4,5,7,8]. In the continuation of the work described in [8], this paper presents a complete encoding scheme around 500 bits/sec. The proposed solution is based on automatic recognition of elementary acoustical units using HMM modelling. An unsupervised training phase is used to build the HMM models and the codebook of synthesis units. The decoded speech is then obtained by concatenating the corresponding synthesis units based on a HNM-like decomposition of speech.

A new unit selection process is proposed integrating some prosody constraints. Through this approach, the size of the synthesis codebook is independent of the targeted bit rate. A complete description of the unit selection process and of the associated prosody modelling is given, together with the quantisation scheme of the overall set of encoded parameters.

1. INTRODUCTION

Classical frame-based encoding of speech is insufficient for addressing Very Low Bit Rate (VLBR) below 600 bits/sec while keeping a sufficient quality. Some already existing schemes achieve bit rate reduction through optimised quantisation of successive frames, as for instance the NATO STANAG 4479 at 800 bits/sec and the newly standardised NATO STANAG 4591 at 1200 bits/sec working on super-frame merging three successive elementary frames.

Improved solutions for a targeted bit rate below 600 bits/sec have been proposed based on variable length segmentation of speech, [4,5,7,8]. Starting from the description of [8], Section 2 presents the basis for VLBR coding of speech, including training, encoding and decoding phases. In Section 3, the proposed solution for unit selection is presented. Section 4 gives a description of the complete VLBR quantisation scheme. In Section 5, some results are given together with the estimated averaged bit rate.

2. PRINCIPLES OF VLBR SPEECH CODING

2.1 Training phase

An unsupervised training phase is used to build the HMM models and the codebook of synthesis units. During the initial step, spectral target vectors and corresponding segmentation are obtained through Temporal Decomposition (TD) of the training speech corpus. Vector Quantisation (VQ) is then used to cluster the different segments in a limited number of classes (64). Finally, for each class of segments, 3-states left-to-right HMM (Hidden Markov Model) models are trained using an iterative process refining both the segmentation and the estimation of the HMM models. The final segmentation is obtained with the final set of HMM models, and is used to build the reference codebook of synthesis units. More details on the training process can be found in [4].

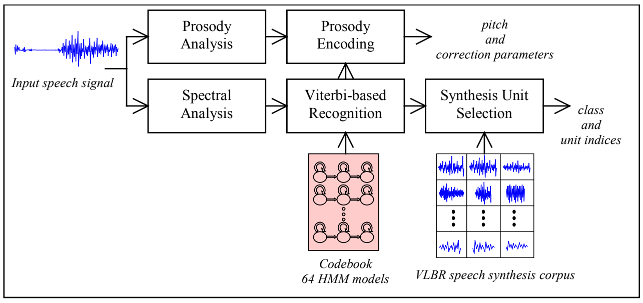

Figure 1: VLBR encoding principle.

2.2 Encoding phase

During the encoding phase, a Viterbi algorithm provides the on-line segmentation of speech using the previously trained HMM models, together with the corresponding labelling as a sequence of class (or HMM) indices. Each segment is then further analysed in terms of prosody profile: frame-based evolution of pitch and energy values. The unit selection process is finally used to find an optimal synthesis unit in the reference codebook. In order to take into account the backward context information, each class of the synthesis codebook is further organised in sub-classes, depending on the previous identified class. The selection process is described in details in Section 3.

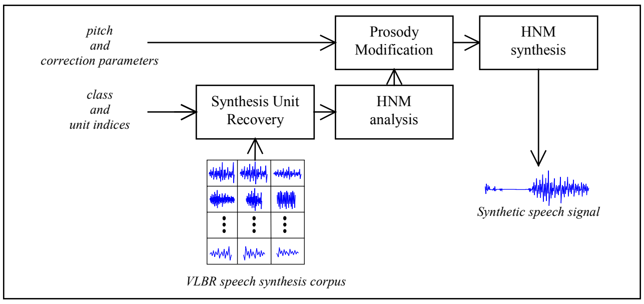

Figure 2: VLBR decoding principle.

2.3 Decoding phase

During the decoding phase, the synthesis units are recovered from the class and unit indices and concatenated with a HNM-like algorithm (Harmonic plus Noise Model). Additional parameters characterising the prosody information are also incorporated to match the original speech signal.

3. UNIT SELECTION PROCESS

3.1 Pre-selection of units according to F0

In the previous VLBR structure, [8], the bit allocation for indexing the synthesis units depends on the size of the stored corpus. An improved quality will then be obtained by both increasing the size of the corpus and the corresponding bit rate. In the new structure, we propose to performs a pre-selection of synthesis units according to the averaged estimated pitch of the segment to be encoded. It is then possible to keep the original training corpus with no limitation regarding its duration, and to choose independently the number of allocated bits to the selected unit indices: whatever the number of available units in the sub-class, therefore a fixed number (Nu = 16 units, 4 bits) of units is dynamically pre-selected.

Once the Nu synthesis units have been pre-selected, the final selection process is performed by incorporating both prosodic and spectral information. For this purpose, time-alignment between the segment to be encoded and the pre-selected synthesis units has been investigated. During our experiments, it was found that a precise alignment at the frame level through Dynamic Time Warping was not essential, and therefore a simple linear correction of the unit's length was sufficient. In order to avoid transmitting additional alignment information, we have used this linear length correction with parameter interpolation to calculate the different selection criteria. The calculation of these criteria is given in the following.

3.2 Correlation measure on pitch profile

For each pre-selected synthesis unit, the pitch profile is compared to the one of the segment to be encoded, using a normalised cross-correlation coefficient. For unvoiced frames, the estimated pitch value is arbitrarily set to zero, therefore introducing a penalty for voicing mismatch.

3.3 Correlation measure on energy profile

Similarly to the pitch profile, a normalised cross-correlation coefficient on the energy profiles is also estimated between each pre-selected synthesis unit and the segment to be encoded.

3.4 Correlation measure on harmonic spectrum

Spectral information can easily be incorporated using various kind of spectral parameters (LPCC, MFCC, LSF) with adequate distances. We suggest to compute an averaged cross-correlation measure between harmonic log-spectrum sequences of pre-selected synthesis unit and segment to be encoded, both being re-sampled either at the F0 profile of the segment to be encoded, or at a fixed predefined F0 (typically equal or less than 100 Hz). Pre-defined F0 reduces the overall complexity since the re-sampling of the synthesis units could then be done at the end of the training phase. A low-complexity alternative scheme consists in first time-averaging the sequences of harmonic log-spectrum and computing the normalised cross-correlation measure on the averaged harmonic log-spectrum.

3.5 Combined selection criteria

The final selection of the synthesis unit is based on a combined criteria of the three previously defined normalised cross-correlation measures. In the current experiments, a linear combination with equal weights has been used.

4. QUANTISATION OF VLBR PARAMETERS

4.1 Quantisation of spectral information

The spectral information is completely represented by the selected synthesis unit. The necessary information for retrieving the corresponding synthesis unit at the decoder is composed of the class index and the unit index in the associated sub-class. The class index is coded with 6 bits (64 classes/64 HMM models), and the unit index is coded with 4 bits (16 closest units according to the averaged pitch).

4.2 Quantisation of prosody

The averaged pitch time lag is quantified in the log-domain using a uniform 5-bit quantifier. A linearly varying gain is determined to match the pitch profile of the segment to be encoded from the one of the selected synthesis unit. This model requires an additional pitch profile correction parameter, which is encoded using a non-uniform 5-bit quantifier. The energy profile is fully determined from the profile of the synthesis unit, with average energy correction. The resulting energy profile correction parameter is also encoded using a non-uniform 5-bit quantifier. Finally, the segment length is coded with 4 bits, in the range of 3 to 18 frames. The corresponding VLBR bit allocation is summarised in *Table 1*. The proposed scheme leads to a bit allocation of 29 bits/segment.

VLBR parameters	Bit Allocation
Class / HMM index (64)	6 bits
Unit index (16)	4 bits
Spectral Information	***10 bits per frame***
Segment length (3-18)	4 bits
Averaged pitch	5 bits
Pitch profile correction	5 bits
Energy profile correction	5 bits
Prosody Information	***19 bits/frame***
Frame bit allocation	**29 bits/frame**

Table 1: VLBR frame bit allocation.

5. EXPERIMENTS AND RESULTS

5.1 Estimated averaged bit rate

For bit-rate evaluation, the coder has been trained on ten speakers individually (5 males/5 females), taken from the French read corpus BREF, [9]. 70 test utterances from each speaker have been coded yielding a global averaged bit rate of ***481 bits/sec***. The maximum and minimum averaged bit-rate per speaker are 512 and 456 bits/sec respectively.

5.2 Experiments

The following figures give an illustration of the proposed unit selection process. *Figure 3* shows the sequence of log-spectrum interpolated at harmonic frequencies for the segment to be encoded. *Figure 4* shows the equivalent sequence of log-spectrum for the selected synthesis unit. A comparison of the different energy profiles is given in *Figure 5*, showing the effectiveness of the selection process. Similarly, the *Figure 6* illustrates the selection process regarding the pitch profile.

5.3 Intelligibility test

The Diagnostic Rhyme Test (DRT) is a common assessment for very low bit rate coders. It uses monosyllabic words that are constructed from a consonant-vowel-consonant sound sequence. In our test, 55 French words are arranged in 224 pairs which differ only in their initial consonants. A word pair is shown to the listener, then he is asked to identify which word from the pair has been played on his headphone. The DRT is based on a number of distinctive features of speech and reveals errors in discrimination of initial consonant sounds.

The test was performed on 10 listeners using the voice of a female speaker coded with three different coders: the MELP (Stanag 4591), the HSX (Stanag 4479), and the VLBR.

The results gathered in Table 2 are the mean recognition score per coder. The VLBR is ranked before the Stanag 4479 but does not reach Stanag 4591 performances. Indeed, the training speech corpus was continuous speech and was not adapted to isolated word coding. Yet, it points out the lack of accuracy of the VLBR coder in recognising and synthesising transient sounds like plosives. Further works will be done in this direction since plosives play an important role in speech intelligibility.

Coder	Recognition score (%)
Stanag 4591 2400 bit/s	88
VLBR 500 bit/s	**80**
Stanag 4479 800 bit/s	77

Table 2: Intelligibility scores.

6. CONCLUSIONS

A complete VLBR encoding system has been proposed based on recognition and synthesis techniques. An averaged bit rate around 500 bits/sec is obtained thanks to a joint process for unit selection and prosody modelling. For illustration purpose, some speech audio files are available at the following address:

http://www.esiee.fr/~baudoing/sympatex/demo

from both the French database BREF, [9], and the Boston University radio news corpus, [10]. The intrinsic quality of the speech synthesis core module (HNM) should be improved through a better incorporation of phase information, and concatenation on spectrally stable zones. If the joint process should help the adaptation of this VLBR scheme to a speaker-independent mode, some work still have to be done in this area. Some studies on robustness to noisy environments are also on-going, in particular with the integration of an AURORA-like front-end, [11]. Finally, compression of the speech synthesis units for low-cost memory storage will have also to be further investigated.

7. ACKNOWLEDGEMENTS

This work has been carried out within the framework of the French RNRT project SYMPATEX, and partly under the grant n° 449/2002 from the ANRT (CIFRE).

REFERENCES

[1] K.S. Lee, R. Cox, “A very low bit rate speech coder based on a recognition/synthesis paradigm,” *IEEE Trans. SAP, Vol.9, N°5*, pp. 482-491, July 2001.

[2] K.S. Lee, R. Cox, “A segmental speech coder based on a concatenative TTS,” *Speech Communication, Vol.38,* pp. 89-100, 2002.

[3] C.M.Ribeiro, I.M.Trancoso, “Phonetic vocoding with speaker adaptation,” *Proc. Eurospeech-97,* pp. 1291-1294, 1997.

[4] J.Cernocky, G.Baudoin, G.Chollet, “Segmental vocoder – going beyond the phonetic approach,” *Proc. ICASSP-98,* pp. 605-608, 1998.

[5] P.Motlicek, G.Baudoin, J.Cernocky, “Diphone-like units without phonemes – option for very low bit rate speech coding,” *Proc.Conf. IEEE - EUROCON-2001,* pp. 463-466, July 2001.

[6] M.Balestri, A.Pacchiotti, S.Quazza, P-L.Salza, S.Sandri, “Choose the best to modify the least: a new generation concatenative synthesis system,” *Proc. Eurospeech-99,* pp. 2291-2294, 1999.

[7] G.Baudoin, F.Capman, J.Cernocky, F.El Chami, M.Charbit, G.Chollet, D.Petrovska-Delacrétaz, “Advances in very low bit rate speech coding using recognition and synthesis techniques,” *TSD-02,* pp. 269-276, Brno, Czech Republic, September 2002.

[8] G.Baudoin, F.El Chami, “Corpus based very low bit rate speech coding,” *Proc. ICASSP-03,* pp. 792-795, 2003.

[9] L.F.Lamel, J.L.Gauvain, M.Eskenazi, “BREF, a large vocabulary spoken corpus for French,” *Proc. EUROSPEECH-91,* Genoa, Italy, 1991.

[10] M.Ostendorf, P.J.Price, S.Shattuck-Hufnagel, “The Boston University radio news corpus.”, Technical Report, Boston University, February 1995.

[11] ETSI document: ES202212, “Distributed speech recognition; Extended advanced front-end feature extraction algorithm; Compression algorithms; Back-end speech reconstruction algorithm”, August 2003.

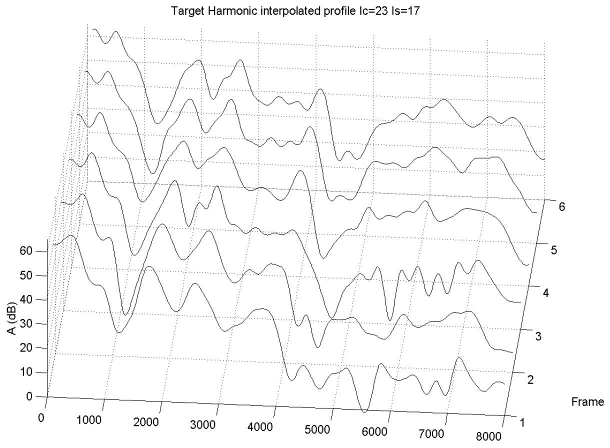

Figure 3:
Interpolated harmonic spectrum of the segment to be encoded.

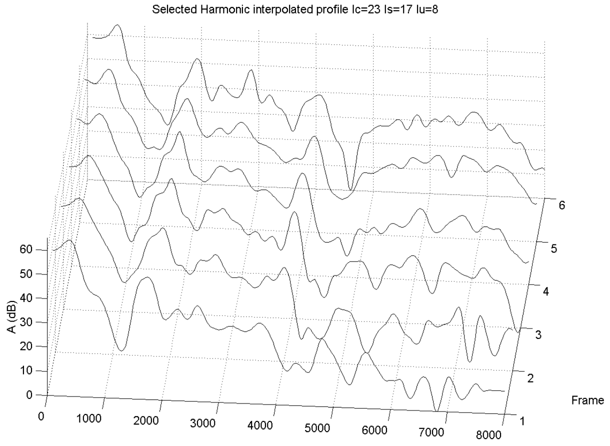

Figure 4:
Interpolated harmonic spectrum of the selected unit.

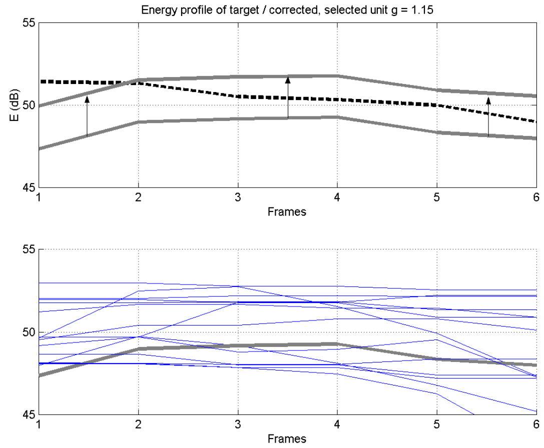

Figure 5:
Energy profile of the segment to be encoded (upper-dashed), and the selected unit before/after correction (upper-bold). Energy profile of the pre-selected units and the selected unit (lower-bold).

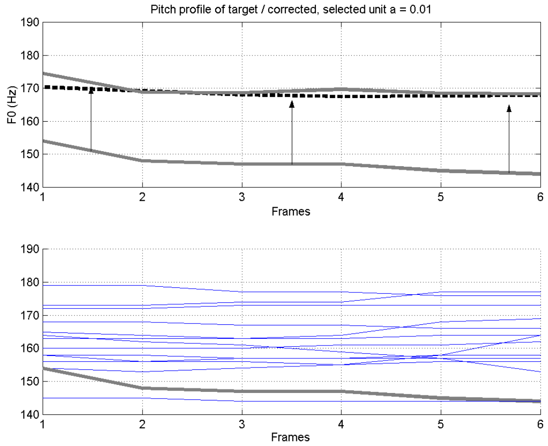

Figure 6 :
Pitch profile of the segment to be encoded (upper-dashed), and the selected unit before/after correction (upper-bold). Pitch profile of the pre-selected units and the selected unit (lower-bold).