Dynamic unit selection for Very Low Bit Rate coding at 500 bits/sec

Marc Padellini[1] and Francois Capman[1] and Geneviève Baudoin[2]

[1] Thales Communications, 160, Bd de Valmy , BP 82,
92704 Colombes, CEDEX, France
{marc.padellini, francois.capman}@fr.thalesgroup.com
[2] ESIEE,Telecommunication Systems Laboratory,
BP 99, 93162 Noisy-Le-Grand, Cedex, France
baudoing@esiee.fr

Abstract. This paper presents a new unit selection process for Very Low Bit Rate speech encoding around 500 bits/sec. The encoding is based on speech recognition and speech synthesis technologies. The aim of this approach is to use at best the speech corpus of the speaker. The proposed solution uses HMM modelling for the recognition of elementary speech units. The HMM are first trained in an unsupervised phase and then are used to build the synthesis unit corpus. The coding process relies on the synthesis unit selection. The speech is decoded by concatenating the selected units through HNM-like decomposition of speech. The new unit selection aims at finding the unit that best match the prosody constraints to models its evolution. It enables the size of the synthesis unit corpus to be independant of the targeted bit rate. A complete quantisation scheme of the overall set of encoded parameters is given.

1 Introduction

Classical frame-by-frame coding can't model speech with sufficient quality at Very Low Bit Rate (VLBR), below 600 bits/sec. Even if bit rate reduction can be achieved through optimised quantisation of successive frames like in the NATO STANAG 4479 at 800 bits/sec and the newly standardised NATO STANAG 4591 at 1200 bits/sec, the spectral envelope is coarse and can't reflect the evolution of speech with good naturalness. An other approach must be taken to cope with the bit rate reduction. A solution was proposed in [1],[2]: using a codebook of speech segments, it is possible to synthesise speech with a set of indice segments which best fit the original speech signal. The spectral envelope can be accurate and full correlation between frames is used. Inspired by speech recognition and speech synthesis, the speech unit can be linguistic like phonemes in [5]. But to have a fully unsupervised coding scheme (without phonetic transcription of the speech corpus), automatically derived units must be used [3], [4]. Using Hidden Markov Models variable length units can be automatically derived in [6], [7], [8], [9]. This paper starts from [9], Section 2 presents VLBR basis of speech coding, the training, the coding, and the 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, an evaluation of the speech quality is presented as well as the estimated average bit rate.

2 Principles of VLBR speech coding

The current system uses about one hour of speech from the speaker for training. It is fully unsupervised. The coding scheme is compound of three phases.

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 [6].

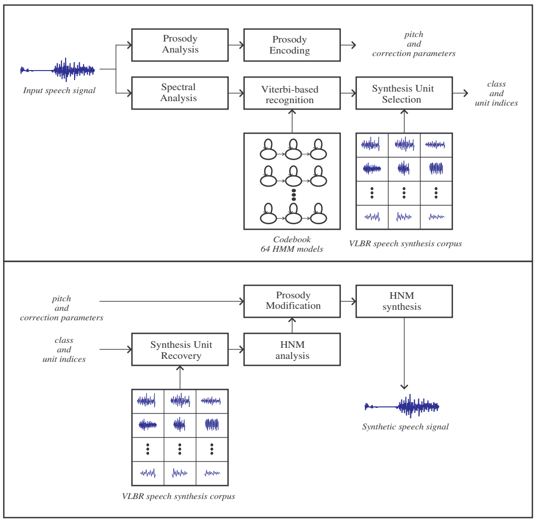

Fig. 1. VLBR coding (upper) and decoding principle (lower)

Encoding phase: During the encoding phase (Figure 1, upper), 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.

Decoding phase: During the decoding phase (Figure 1, lower), 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 most VLBR structure, [1], [2], [3], [4] and [9], 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 [10] it is suggested for TTS systems, that a large number of units should be used, in order to select the best and modify the least the synthesis units. We propose to performs a pre-selection of the 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. In effect, the number of allocated bits to the selected unit indices can be chose independantly, whatever the number of available units in the sub-class. We fixed this number to Nu = 16 units, (4 bits) in the dynamic pre-selection. On Figure 2, the occurences of the units in the pre-selection process are plotted, for one class of the synthesis unit corpus and for the coding of 15 minutes of speech. A broad range of units are pre-selected (more than 80%). The pre-selection process can be viewed as a window taking the 16 closest units to the target unit, in the pitch domain.

3.2 Final unit selection

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

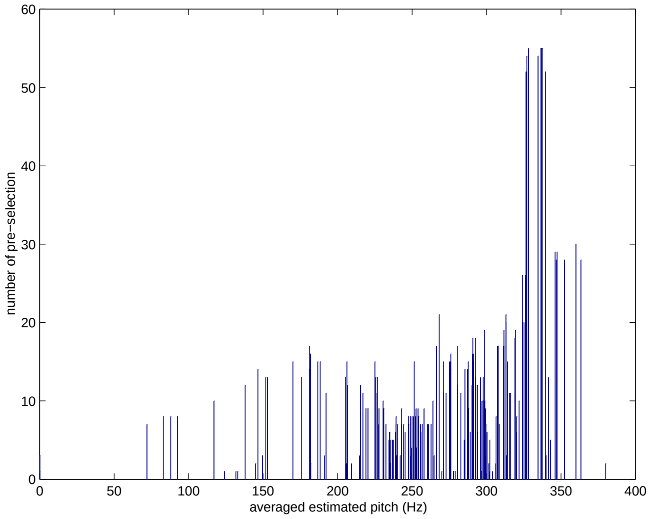

Fig. 2. number of occurencies in the pre-selection of the 231 units of the subclass H44/H12, for the coding of 15 minutes of speech (this subclass was found 152 times)

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.

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.

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.

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.

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

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).

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.

Table 1. VLBR frame bit allocation

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

5 Experiments and results

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, [11]. 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.

Experiments: Figure 3 is an illustration of the proposed unit selection process. Upper-left hand corner shows the sequence of log-spectrum interpolated at harmonic frequencies for the segment to be encoded, and the equivalent sequence of log-spectrum for the selected synthesis unit after correction. Upper-right hand corner shows the interpolated mean harmonic profiles. A comparison of the different energy profiles is given in the Lower-left hand corner, showing the effectiveness of the selection process. Similarly, the Lower-right hand corner illustrates the selection process regarding the pitch profile.

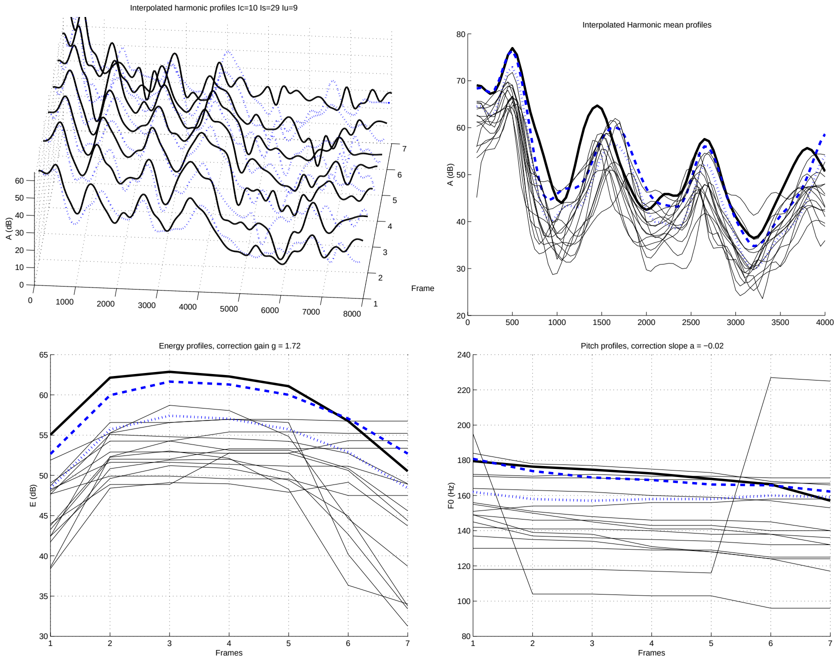

Fig. 3. Unit selection process: target parameters are in bold solid black line, selected unit in bold dotted line, selected unit after correction in bold dashed line, pre-selected units in solid lines.

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.

6 Conclusion

A new dynamic selection of units has been proposed for VLBR coding. An averaged bit rate around 500 bits/sec is obtained through quantisation of 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 the French database BREF, [11]. Recent developments on concatenation on spectrally stable zones should improve the quality of speech synthesis. Moreover, for the special case of plosive sounds, the HNM-like model should better model transient sounds and the recognition core should perform dedicated classification. 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, [12]. Finally, compression of the speech synthesis units for low-cost memory storage will have also to be further investigated.

Table 2. Intelligibility scores

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

References

1. Roucos, S., Schwartz, R.M., Makhoul, J.: A segment vocoder at 150b/s. Proc. ICASSP'83(1983) 61–64
2. Roucos, S., Wilgus, A.M., The waveform segment vocoder: a new approach for very-low-bit-rate speech coding. Proc. ICASSP'85(1985) 236–239
3. Lee, K.S., Cox, R.: A very low bit rate speech coder based on a recognition/synthesis paradigm. IEEE Trans. SAP. **9** (2001) 482–491
4. Lee, K.S., Cox, R.: A segmental speech coder based on a concatenative TTS. Speech Communication **38** (2002) 89–100
5. Ribeiro, C.M., Trancoso, I.M.: Phonetic vocoding with speaker adaptation. Proc. Eurospeech'97 (1997) 1291–1294
6. Cernocky, J., Baudoin G., Chollet, G.: Segmental vocoder - going beyond the phonetic approach. Proc. ICASSP'98 (1998) 605–608
7. Motlicek, P., Baudoin G., Cernocky J.: Diphone-like units without phonemes - option for very low bit rate speech coding. Proc. Conf. IEE- EUROCON-2001 (2001) 463–466
8. Baudoin, G., Capman, F., Cernocky, J., El Chami, F., Charbit, M., Chollet, G., Petrovska-Delacrtaz, D.: Advances in very low bit rate speech coding using recognition and synthesis techniques. TSD'02, (2002) 269–276.
9. Baudoin, G., Chami, F.El: Corpus based very low bit rate speech coding. Proc. ICASSP'03 (2003) 792–795.
10. Balestri, M., Pacchiotti, A., Salza, P.L., Sandri, S.: Choose the best to modify the least: a new generation concatenative synthesis system. Proc. Eurospeech'99 (1999) 2291–2294
11. Lamel, L.F., Gauvain, J.L., Eskenazi, M.: BREF, a large vocabulary spoken corpus for French. Proc. Eurospeech'91 (1991)
12. document: ES202212. Distributed speech recognition; Extended advanced front-end feature extraction algorithm; Compression algorithms; Back-end speech reconstruction algorithm. ETSI (2003)