CORPUS BASED VERY LOW BIT RATE SPEECH CODING

G. Baudoin, F.El Chami *

ESIEE
Telecommunications systems laboratory
g.baudoin@esiee.fr, BP 99, 93162 Noisy Le Grand, CEDEX, France

ABSTRACT

This paper presents a new Very Low Bit Rate segmental speech coding approach applying speech recognition in the coder and corpus based speech synthesis in the decoder. The system uses a large corpus of speech signal that is searched to find a speech segment similar to the segment to be coded. The elementary acoustical units for recognition and synthesis are determined automatically by an unsupervised training method. This approach is an alternative to using phoneme-derived linguistic units.

Very good results are obtained at an average bit rate of 400 bits/second for a corpus of about 1 hour of speech. We present an efficient method for finding the best synthesis unit taking into account the good concatenation of successive segments. The proposed organization of the speech segments in the corpus allows a very efficient search of the best unit.

1. INTRODUCTION

Several standards define speech coders with a bit rate between 1200 and 4800 bps (bits/second) that give an acceptable quality for communication applications. These coders split the speech signal into frames of 20 to 30 ms. Then they extract and code pertinent parameters from each frame and transmit these parameters to the decoder.

It is possible to decrease the bit rate by coding the set of parameters of a few successive frames (typically 3 frames). The NATO STANAG 4479 defines a 800 bps vocoder using this principle.

But in order to achieve Very Low Bit Rates (VLBR) coding below 500 bps, while keeping a sufficient speech quality, this approach is no longer sufficient. One cannot work with frames of fixed length. A segmental approach using segments of variable length is necessary [1, 2, 3, 4, 5, 6, 7, 8].

VLBR coders implement recognition of acoustic segments in the analysis phase and speech synthesis from a set of segment indices in the decoder. The coder generates a symbolic transcription of the speech signal from a dictionary of elementary units. These units can be linguistic units (e.g. phonemes, subword units) in "phonetic vocoders" or they can be acoustical units automatically obtained from a speech training corpus. Phonetic vocoders require a phonetic transcription of the training corpus, which is time-consuming, prone to errors and has to be carried out for each language. The automatic determination of acoustic units from an unlabeled speech corpus is therefore an interesting alternative.

* This work was realized in the RNRT project SYMPATEX with funding from the French ministry of research.

In this paper, we present a new VLBR speech coder applying speech recognition and synthesis techniques and working with a large corpus of speech signal. The recognition is based on acoustical units obtained automatically in an unsupervised manner and named Recognition Acoustical Units (RAU). The decoder is a corpus based speech synthesizer. The synthesis units are named Synthesis Acoustical Units (SAU) and the speech segments representative of a synthesis unit are named Synthesis Speech Representatives (SSR). The quality of the coder is very good for an average bit rate of 400 bps. This paper focusses on the the dynamic selection of the synthesis representatives in the speech corpus.

Section 2 describes the principles of the VLBR speech coder. Section 3 presents the techniques used to determine the recognition units. Section 4 explains the proposed method for the dynamic selection of the synthesis representaives. Section 5 presents the realized experiments and the obtained results.

2. PRINCIPLES OF THE VLBR SPEECH CODER

The proposed VLBR speech coding system comprises a coder that acts as a speech recognition system and a decoder that works as a speech synthesizer. The system uses a large corpus of about one hour of speech signal.

2.1. Principle of the Coder

The Fig. 1 illustrates the tasks of the coder that can be summarized in: recognition of RAU, prosodic analysis, determination of the synthesis speech representatives.

The coder uses a set of N_R recognition units RAU modelized by Hidden Markov Models (HMM). These models are trained during an off-line training phase described in section 3.

A spectral analysis extracting Linear Prediction Cepstral Coefficients (LPCC) and an energy calculation are applied frame by frame on the signal. The most likely sequence of RAU is determined by a Viterbi algorithm working on the sequence of parameter vectors: LPCC, derivative of LPCC and derivative of the energy. Therefore, the signal is segmented and labeled into a sequence of RAU.

Each recognized segment is analyzed to determine its prosodic parameters such as fundamental frequency, energy contours, segment length. These parameters are coded segment by segment [9].

Then the coder searches in its speech corpus, the speech segment that best matches each recognized segment and that will be used by the decoder as a representative for the synthesis. The organization of the speech corpus and the search for the best match are described in section 4

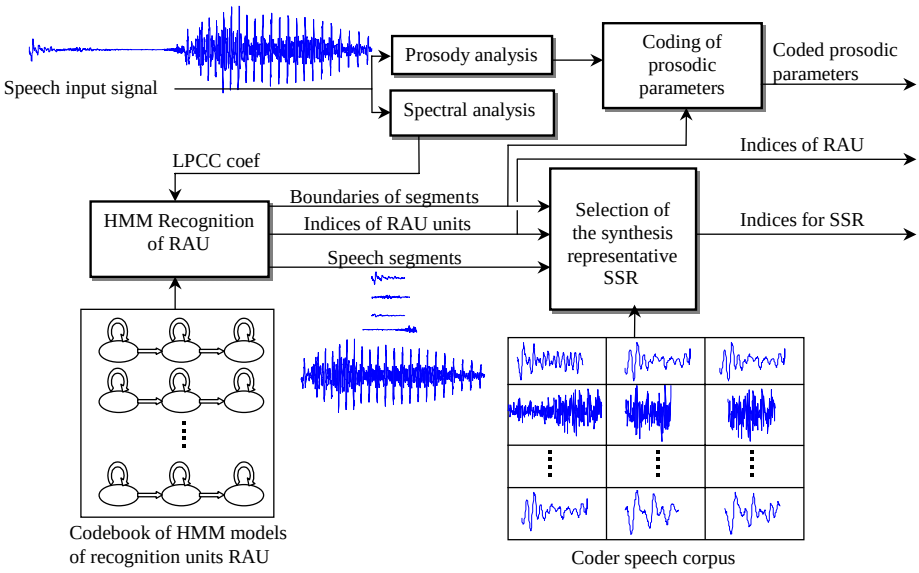

Fig. 1. Principle of the coder.

Finally, for each recognized segment, the coder transmits the index of the recognized RAU, a complementary index for the selected SSR, and the coded prosodic parameters.

2.2. Principle of the decoder

The Fig. 2 represents the decoder.

The decoder receives the indices of the RAU and the complementary indices for the SSR plus the coded prosodic parameters.

It generates the decoded speech by concatenation of SSR segments chosen in its speech corpus. This concatenation is done using an Harmonic plus Noise Model (HNM) of speech that makes prosodic transformation easy to realize.

It uses the 2 indices to select the SSR. Then it modifies the prosody of the selected segment according to the received prosodic parameters.

3. TRAINING OF THE RECOGNITION UNITS

The recognition units are obtained automatically in an unsupervised manner. The different steps of this training is described in [7].

The first step uses a Temporal Decomposition (TD) to segment the training corpus into spectral targets connected by interpolation functions. The average rate of segments after the TD is about 18 segments per second. The spectral targets vectors are clustered in N_R classes by vector quantization (VQ) of the spectral parameters. A first segmentation and labeling of the speech corpus is realized by segmenting the signal at the intersections of interpolation functions of the temporal decomposition and classifing the segments according to the VQ class of their spectral target. The elementary speech segments are more or less spectrally stable, at least in their central part. The N_R classes are named H_j with $j \in [0, N_{R-1}]$.

A left-to-right HMM model with 3 emitting states is trained on each of the N_R classes H_j. Then the training corpus is re-segmented and labeled using these HMM models with a Vieterbi algorithm. The process of training of HMM models and segmentation of the training corpus is iterated a few times until the likelihood of the models do not significantly increase any more. At the end of the training, we obtain N_R HMM and the training corpus is segmented and labeled with the labels H_j. Each class H_j is modelized by a HMM and contains all the speech segments of the training corpus that were labeled by H_j.

4. DYNAMIC SELECTION OF SYNTHESIS REPRESENTATIVES

4.1. Possible approaches for the speech synthesis.

The decoder is a speech synthesizer operating with the information received from the coder: RAU and SSR indices and prosodic parameters. The speech synthesis is based on concatenation of synthesis representatives SSR. Two types of SAU can be used:

- *Many SAU units with spectrally stable extremities and only one speech representative SSR per unit.* This approach is similar to diphone concatenative synthesis techniques. The concatenation of SAU is easy. Each SAU is respresented by a single speech segment chosen in the training corpus. We have experimented this approach [10] in constructing some kind of diphone-equivalent SAU H_iH_j by resegmenting in the spectrally stable parts of the RAU segments. The results were not completely satisfactory, concatenation noise was still clearly audible and due to the limitation of the training corpus, some synthesis units were missing and difficult to replace.
- *Few short SAU units with many representatives for each SAU and dynamic selection of representatives*. This method is comparable to corpus based text to speech synthesis [11]. This approach provides clearly better results than the first one. It is the method that we have retained for our VLBR coder.

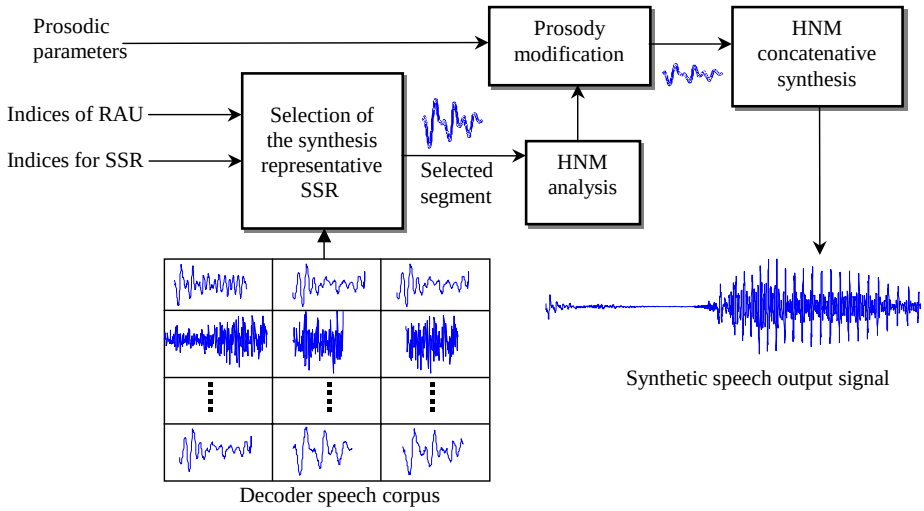

Fig. 2. Principle of the decoder.

4.2. Synthesis with dynamic selection of short representatives.

4.2.1. Constitution of the coder-decoder speech corpus

The speech corpus used in the coder and in the decoder for the synthesis is constituted and organized in the following way.

The SAU and RAU classes are identical. The speech segments of the H_j class are called synthesis representatives for H_j. The coder-decoder speech corpus is made of these classes of variable length speech segments labeled by their recognition class RAU.

The classes H_j are furthermore organized in subclasses according to the left context of speech segments. Each H_j class is partitioned in N_R sub-classes called H_iH_j containing all the speech segments of class H_j that were preceded by a segment belonging to the class H_i in the training corpus.The Fig. 3 illustrates this organization of the speech corpus into classes and subclasses of synthesis representatives for an example with $N_R = 3$.

4.2.2. Selection of the synthesis unit

During the coding phase the determination of the SSR is done in order to fulfill criteria of good representation of a given segment to be coded and criteria of good concatenation of successive segments.

A possible solution consists in defining representation and concatenation distances (D_R and D_C) and in choosing the synthesis representatives in order to minimize a criterion of the form $aD_I + bD_C$, where a and b are 2 weighting factors. The representation distance D_R is a distance between the segment to be coded and the synthesis representatives. The concatenation distance D_C is a measure of the quality of concatenation between successive synthesis representatives. The main drawback of this solution is that it requires to adjust the parameters a and b leading to heavy experimental trials.

Therefore, we developed a different method where we have avoided the use of a concatenation criterion D_C by the organization of each class H_j of the speech corpus in subclasses H_iH_j according to the left context of segments. During the coding phase, the speech signal is first segmented and labelled with RAU units H_j. Then, if a segment is recognized as belongig to the class H_j and is preceded by a segment in the class H_i, the synthesis representative is searched in the subclass H_iH_j of the class H_j

It is possible to keep all the segments of the training corpus as synthesis representatives or to limit the size of each sub-class to a maximal value of K segments. This limitation allows to control the maximum bit rate for the coding of the SSR indices. When the size of a suclass is smaller than K segments, we keep all the segments of the subclass.

The selection of the best SSR representative in a sub-class H_iH_j is done in order to minimize a criterion D_C of good representation of the segment. The D_C criterion is based on a spectral comparison by dynamic time warping DTW between the segment to code and the possible SSR. We used spectral vectors resulting of a concatenation of parameters representing the spectral envelope of both the harmonic part and the noisy part of the HNM model. The criterion D_C can also include a distance on prosodic parameters.

The index of the RAU is transmitted on $\log_2(N_R)$ bits and the index for the SSR on $\log_2(K)$ bits or $\log_2(Nmax)$ bits where N_{max} is the maximum number of segments in a sub-class. It is not necessary to transmit the index of the sub-class, since the decoder has the same information as the coder concerning the preceding unit.

If the training corpus is not large enough, some of the subclasses H_iH_j may be empty. We developed an algorithm of substitution of the missing classes. We have calculated an average spectral distance between classes H_j. When a class H_iH_j is empty, the algorithm searches in the non empty subclass H_kH_j where H_k is the closest class to H_i.

When the number of representative segments is not limited the coder does an exhaustive search in the training corpus, but this is done efficiently: because of pre-classification by preceding segments the calculation is divided by 64.

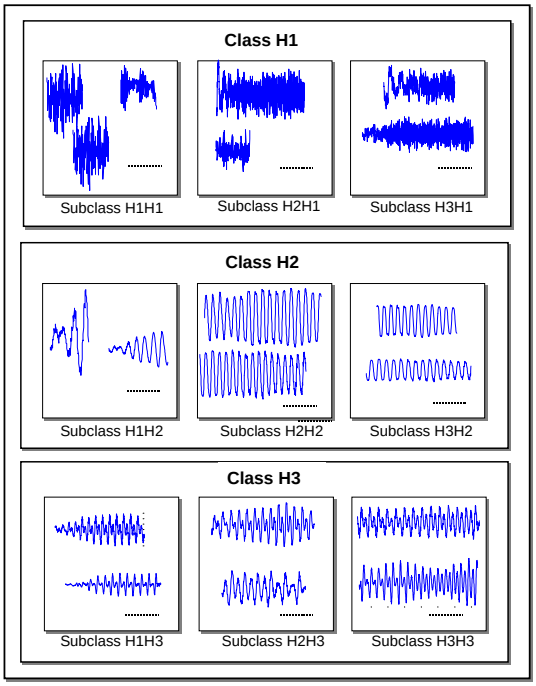

Fig. 3. Organization of the speech corpus in classes and subclasses of synthesis representatives, example with $N_3 = 3$.

5. EXPERIMENTS AND RESULTS

The proposed VLBR coder was tested on a French corpus of wideband speech sampled at 16 KHz, with 83 min and 15 min of signal respectively for the training and the test corpus.

The number of RAU classes N_R was set to 64 (same order of magnitude as the number of phonemes).

We obtained the following results for the training corpus. The average length of RAU segments is 62 ms. The number of segments per class is between 470 and 1267. There are 69% of non-empty subclasses (2827 out of 4096). The average number of segments per subclass is 29.

This approach gives very good results in term of subjective quality of speech, but it requires a large memory size in the system for storing the speech corpus and the delay introduced by the coder is important (typically 150 ms).

If no limitation is done on the number of segments in a subclass, the complete training corpus is present in the system (83 min in our experiments). If the number of segments per subclass is limited to $K = 16$, the quality is only very sligthly degraded and the size of the speech corpus is limited 45 minutes.

There is a compromize between the size of the stored corpus and the speech quality. The average bit rate, measured on the test corpus, for the coding of the RAU and SSR indices is equal to 220 bps when the complete speech corpus is used and to 180 bps when $K = 16$ (i.e. an average of 18 segments per second with 10 bits to code both indices). The average bit rate for the coding of the prosodic parameters is 200 bps. So the total average bit rate is between 380 bps and 420 bps for $K \geq 16$.

6. CONCLUSIONS

We have proposed a new corpus based VLBR speech coder. The speech corpus used by the system contains variable length speech segments organized in acoustical RAU classes and sub-classes. The acoustical RAU classes contains segments with spectral similarities. Each class is divided in sub-classes according to the class of their preceding segment. Each recognition class is modelized by an HMM.

Incoming speech is segmented and labeled by a Viterbi algorithm. For each resulting segment a synthesis representative is chosen in the speech corpus according to the label of the segment and to the label of the preceding segment.

At a bit rate between 350 bps and 400 bps, the obtained speech quality is very good for a corpus corresponding to about 1 hour of speech. Demonstration coded signals can be listened to at http://www.esiee.fr/~baudoing/sympatex/demo.

Work is ongoing on the compression of the coder corpus, on the extension to the speaker dependent case and on the robustness to noisy environments.

7. REFERENCES

[1] K.-S. Lee and R. Cox, "A very low bit rate speech coder based on a recognition/synthesis paradigm," *IEEE Trans. SAP*, vol. 9, no. 5, pp. 482–491, July 2001.

[2] K.-S. Lee and R. Cox, "A segmental speech coder based on a concatenative tts," *Speech communication*, vol. 38, pp. 89–100, 2002.

[3] Picone and G. R. Doddington., "A phonetic vocoder.," in *Proc. ICASSP-89*, 1989, pp. 580–583.

[4] S. Roucos, R. Schwarz, and J. Makhoul., "A segment vocoder at 150 b/s.," in *Proc. ICASSP-83*, 1983, pp. 61–64.

[5] C. Ribeiro and M. Trancoso., "Phonetic vocoding with speaker adaptation.," in *Proc. Eurospeech-97*, Rhodes, 1997, pp. 1291–1294.

[6] M. Ismail and K. Ponting., "Between recognition and synthesis 300 bit/s speech coding," in *Proc. Eurospeech-97*, Rhodos, 1997, pp. 441–444.

[7] J. Černocký, G. Baudoin, and G. Chollet, "Segmental vocoder - going beyond the phonetic approach," in *Proc. ICASSP 98*, Seattle, USA, 1998, pp. 605–608.

[8] G. baudoin, J. Černocký, P. Gournay, and G. Chollet, "Codage de la parole à bas et très bas débit," *Annales des télécom.*, vol. 55, no. 10, pp. 462–482, Nov. 2000.

[9] Y.-P. Nakache, P. Gournay, and G. Baudoin, "Codage de la prosodie pour un codeur de parole trs bas dbit par indexation d'units de taille variable," in *Proc. CORESA' 2000*, France, Oct. 2000.

[10] P. Motlíček, G. Baudoin, and J. Cernocky, "Diphone-like units without phonemes - option for vey low bit rate speech coding," in *Proc. conf. IEEE - EUROCON'2001*, Slovakia, July 2001, pp. 463–466.

[11] M. Balestri, A. Pacchiotti, S. Quazza, P.-L. Salza, and S. Sandri, "Choose the best to modify the least: a new generation concatenative synthesis system," in *in Proc. Eurospeech*, Hungaria, 1999, pp. 2291–2294.