VERY LOW BIT RATE SEGMENTAL SPEECH CODING USING AUTOMATICALLY DERIVED UNITS

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

[1] FEI VUT Brno, Inst. of Radioelectronics, Czech Republic, cernocky@urel.fee.vutbr.cz
[2] ESIEE Paris, Dpt. Signal et Télécom., France, {cernockj,baudoing}@esiee.fr
[3] ENST Paris, Dpt. Signal, France, chollet@sig.enst.fr

ABSTRACT

In our paper, the problem of very low bit rate segmental speech coding is addressed. The basic units are found automatically in the training database using temporal decomposition, vector quantization and multigrams. They are modelled by HMMs. The coding is based on recognition and synthesis. In single speaker tests with the PolyVar database, we obtained mostly intelligible and naturally sounding speech at mean rate of 211.2 b/s. First results on the Boston University Radio Corpus are reported and future extensions of our scheme (synthesis and speaker adaptation) are discussed.

1. INTRODUCTION

The very low bit rate speech coders (see [7] for overview) operating at hundreds of bits per second, mostly do not use *frames* of fixed length, but *variable length segments* as basic coding units. The *phonetic* vocoders [8, 4, 6] reaching the lowest rates, make use of a phoneme-based recognition in the coder, followed by synthesis in the decoder. The main drawback of these schemes is the necessity of a transcribed database (DB) for the training of phone-models, so that their use in languages lacking standard speech databases is discutable. Our approach is based on deriving the basic units directly from the speech data, without the need of transcriptions, using **ALISP** (automatic language independent speech processing [3]) tools. The paper is organized as follows: section 2 describes the global structure of our coder, section 3 gives a brief overview of tools used in the search, modelling and recognition of units, section 4 summarizes the experiences performed on the monospeaker part of PolyVar database and in section 5 we describe first experiences on the BU database. Section 6 contains the discussion and conclusions.

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

2. SEGMENTAL CODING

The overall scheme of our coder is given on Fig. 1. The algorithm consists of five steps:

Non-supervised search of characteristic segments. The temporal decomposition (TD), vector quantization (VQ) and eventually multigram segmentation (MG) were used to find the initial set. It was further refined by hidden Markov model (HMM) training, and resegmentation.

Clustering and modelling of segments. Left-right HMMs were used to model the sequences.

Segment recognition. The segment recognition is done using techniques known from continuous speech recognition. The index of recognized sequence is transmitted to the decoder.

Segment reconstruction. In this etap, a simple synthesis using examples from the training corpus was applied. Only the choice of example and energy correction is transmitted.

Adaptation. The resulting set of typical segments is strongly dependent on the training database. Several approaches can be considered to overcome the inter-speaker variability (normalization of voices to a generic one, voice modification). The adaptation has not yet been tested experimentally.

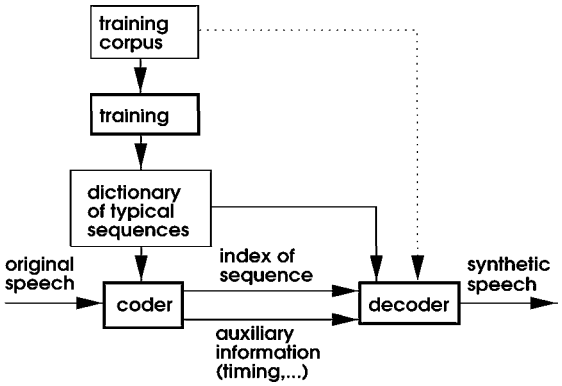

Figure 1: Scheme of the segmental coder.

3. TOOLS

This section gives a brief overview of the tools used for initialization and refinement of the segment dictionary, for segment modelling and recognition.

Temporal decomposition [1] was used to initially segment the speech into quasi-stationary parts. A spectral parameter matrix is decomposed into a limited amount of *events*, each represented by a *target* and an *interpolation function* (IF). We used the `td95` package[1] for this decomposition, where a short-time SVD with adaptive windowing takes place for the initial search of IFs, followed by post-processing of IFs (smoothing, decorrelation) and iterative refinement of targets and IFs.

Vector quantization was used to initially cluster the events found by the TD. It can be performed either with one vector per segments, or on whole segments using cumulated distances. Another possibility is a creation of 2 codebooks of lower dimension, one for voiced, another for unvoiced segments.

Multigrams [2] provide a segmentation of a symbol string (resulting from VQ) into characteristical sequences, and a dictionary of such sequences using maximization of the likelihood: $L(W) = \max_{\{B\}} \prod_k p(S_k)$, where W is the symbol string and B is the set of all possible segmentations into sequences S_k of length 1 to n.

Hidden Markov Models [10] are the standard framework for the speech recognition. Here, the HMMs are used to refine segments in the dictionary, to model them and to detect these segments in the input speech. For the recognition, we were experimenting with a simple "unigram" language model (LM) based on a-priori probabilities of sequences. With this model, the likelihood to maximize is: $L = \prod p(M_i)^\gamma p(O|M_i)$, where O are the observations, $p(M_i)$ is the a-priori probability of model M_i and γ is the LM scale factor. For all HMM experiences, the HTK software [10] was used.

An example of processing by above mentioned tools is shown on Fig. 2.

[1]Written by Frédéric Bimbot, we are grateful to the author for the permission to use it.

4. EXPERIENCES – POLYVAR

First experiences were done with one speaker from the Swiss French PolyVar database [5] recorded at IDIAP. The set of 218 calls was divided into training ($\frac{4}{5}$) and test ($\frac{1}{5}$) sets. The signal was parametrized by 10 LPCC coefficients in 20 ms frames, with a shift of 10 ms and a cepstral mean substraction for each call. The voice activity decisions were taken and only active parts were used in the rest of the work: 5.2 hours of speech containing 15813 parts and 1.8×10^6 frames. The **TD** was adjusted so that the average number of events per second is closed to the phonetic rate (15 events/sec approximately). The total number of events in the training corpus is 271078. The **VQ** was performed on the original cepstral vectors situated in gravity centers of IFs. A codebook with $L = 64$ code-vectors was trained using a standard LBG algorithm; the quantized vectors form a symbol string. The **MG** dictionary training and segmentation were performed with a maximum length of multigram m=5 and with 10 iterations of the segmentation-reestimation loop. With respect to the following HMM training, a threshold for minimal number of occurencies of one sequence in the corpus was set to 20. During the iterations, only 1-, 2-, and 3-grams were

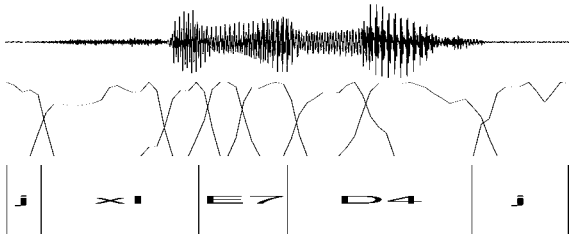
Figure 2: Example for the French word "cinema". signal, TD interpolation functions and MG segmentation.

retained, with the numbers 64, 1514 and 88 respectively (total of 1666 sequences).

In the **HMM** phase, a prototype left-right model with $2i + 1$ states was constructed for each i-gram in the dictionary. These HMMs were then estimated using the initial MG transcriptions of the training corpus and HTK tools `HInit`, `HRest` and `HERest` (5 iterations). These "1st generation" models were used to resegment the corpus with 3 different LM factors γ=0.0,5.0,10.0 leading to 3 transcriptions for each file (modified `HVite` tool was used for this task). Using these transcriptions, the original models were refined, with the check for minimal number of occurences in the corpus: this is why the 3 resulting dictionaries contain less than 1666 models. With those "2nd generation" models, the training as well as test corpora were segmented. The results of this segmentation, giving actually the coding rates, are reviewed in Table 1.

The **synthesis** used in these experiences was very simple, based on concatenation of units without any pitch or duration modification. A set of 8 *examples* was found in the training set for each sequence. When coding, for each labelled input sequence, the example with the best duration and DTW-distance match was chosen (coding of this choice by 3 bits). To avoid energy discrepancies between the example and original, an energy correction constant (5 bits) was transmitted, forcing the mean energy of example to match with the original.

4.1. PolyVar – Results

The informal listening tests showed, that for all three γ factors, the output speech sounds naturally, on contrary to LPC-synthesis based schemes. However, for γ=5.0 and 10.0 it is not fully intelligible. For γ=0.0, the intelligibility was excellent for command word, names, short phrases and digits well represented in the DB, worse for longer phrases. The resulting bit rate including the additional information is thus: 117+11.07×8 = 205.6 b/s for training and 120+11.40×8 = 211.2 b/s for test corpus. The corresponding audio files can be downloaded from:

`www.fee.vutbr.cz/~cernocky/Icassp98.html`.

Several problems emerged during the experimental work: the *acoustical quality* of found units is determined by the DB used. With a telephone DB and a non-profesional speaker, it is not excellent. The use of *MG before HMM* makes the basic coding units longer, thus alleviating some coarticulation problems and making the speech more natural, but the number of free parameters in HMM training is considerable and makes the system very complex. Finally, the using of simple *synthesis* was realized only to test the segmentation and must not be retained in the final coder. Surprisingly, the effects of this synthesis were not so disastrous as we would expect (compare to subsec. 5.1).

5. EXPERIENCES – BU CORPUS

The problems listed above led us to use a high quality DB. We choose 2 speakers from the Boston University Radio Speech Corpus, one female (F2B) and one male (M2B). The radio news portions were taken as training corpus (56 and 71 min. of speech respectively), the laboratory news portion served as test corpus (22 and 12 min.). This DB is sampled at 16 kHz, so the parametrization was done by 16 LPCC coefficients, with cepstral mean substraction on "story" basis. The **TD** was set up to produce slightly more events per second than for PolyVar, the average was 18.3 events/sec for female and 17.8 for male. We made this change in order to prevent more distinct sounds to be represented by one event, as it happenned occassionally for PolyVar. The **VQ** codebooks were trained on the vectors in gravity centers of IFs, but the quantization was performed using whole segments.

No MG "lenghtening" of sequences was performed before the HMM phase to limit

		train. set			test set		
γ	N	R_e	R_u	R_s	R_e	R_u	R_s
0.0	1514	113	117	11.07	116	120	11.40
5.0	1201	68	74	7.19	69	74	7.29
10.0	894	51	58	5.95	50	58	5.95

Table 1: Resulting rates [b/s] for sequence indices coding. N – size of sequence dictionary, R_e – average rate for entropy coding ($-\log_2 p(M_i)$ bits for sequence M_i), R_u – avg. rate for uniform coding ($\log_2 N$ bits per sequence), R_s – avg. number of sequences per second.

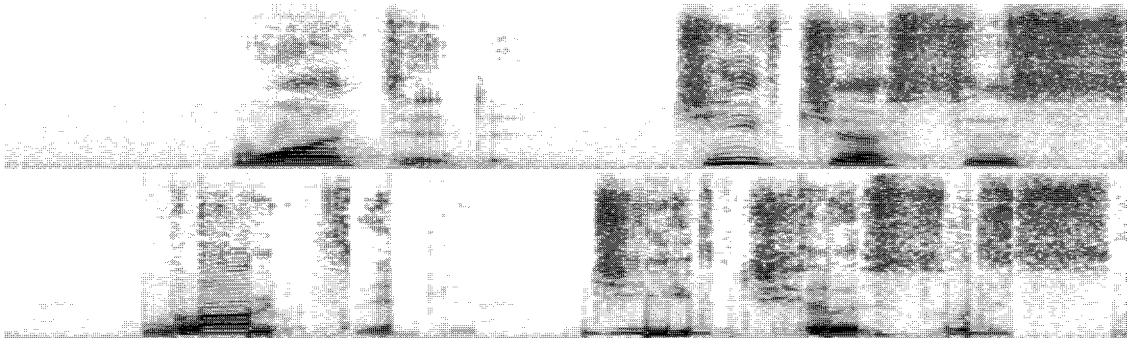

Figure 3: Example from the BU database: "wanted. chief justice" – original and synthesized signal spectrograms. The segment transitions are clearly visible as vertical bars.

the number of free parameters. The models were trained directly on the TD+VQ transcriptions, so that their number is 64, each with 3 states. Using these initial models, three successive HMM sets were trained using resegmentations (no LM) and HMM reestimations. We found that the acoustical coherence of sequences improved from iteration to iteration . With "last-generation" models, we segmented the training and test corpus and tested the simple concatenation synthesis.

5.1. BU – Results

The resulting bit rates (R_u) were 120.7 and 121.5 b/s on the training and test corpus for F2B, and 118.5 and 122.0 b/s for M2B. The quality of the synthetic speech is not very good, due mainly to transition effects on segment borders (one can hear a "machine-gun" at the segment rate), see Fig. 3. Currently we are experimenting with a more sophisticated synthesis using Harmonic and Noise Model (HNM), which is capable to modify the duration and pitch of segments [9] and should improve considerably the quality of synthetic speech. Another improvement should be obtained using the MG postprocessing of HMM segmentation.

6. DISCUSSION, CONCLUSIONS

We have proved that very low rate speech coding with automatically derived units is a promissing topic. However, many problems remain to be solved: for the HNM synthesis, the pitch and energy contours must be efficiently coded, and the speaker adaptation must be taken into accout. In our opinion, using this approach, good quality coding at rates of hundreds of b/s is possible and can be used in multimedia technology, archivation and Internet telephony applications.

7. REFERENCES

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

[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] G. Chollet, J. Černocký, A. Constantinescu, S. Deligne, and F. Bimbot. *NATO ASI: Computational models of speech pattern processing*, chapter Towards ALISP: a proposal for automatic language independent speech processing. Springer Verlag, in preparation.

[4] C.M.Ribeiro and I.M.Trancoso. Phonetic vocoding with speaker adaptation. In *Proc. EUROSPEECH 97*, pages 1291–1294, Rhodes, Greece, 1997.

[5] A. Constantinescu and G. Chollet. Swiss PolyPhone and PolyVar: building databases for speech recognition and speaker verification. In *Speech and image understanding, Proc. of 3rd Slovenian-German and 2nd SDRV Workshop*, pages 27–36, Ljubljana, Slovenia, 1996.

[6] M. Ismail and K. Ponting. Between recognition and synthesis – 300 bits/second speech coding. In *Proc. EUROSPEECH 97*, pages 441–444, Rhodes, Greece, 1997.

[7] C. Jaskie and B. Fette. A survey of low bit rate vocoders. *DSP & Multimedia technology*, pages 26–40, April 1994.

[8] J. Picone and G. R. Doddington. A phonetic vocoder. In *Proc. IEEE ICASSP 89*, pages 580–583, Glasgow, 1989.

[9] I. Stylianou. *Modèles harmoniques plus bruit combinés avec des méthodes statistiques, pour la modification de la parole et du locuteur.* PhD thesis, École nationale supérieure des télécommunications (ENST), Paris, January 1996.

[10] S. Young, J. Jansen, J. Odell, D. Ollason, and P. Woodland. *The HTK book.* Entropics Cambridge Research Lab., Cambridge, UK, 1996.