UNSUPERVISED LEARNING FOR VERY LOW BIT-RATE SPEECH CODING

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

[1] Institute of Radioelectronics, Brno University of Technology, Czech Republic
[2] Département Signal et Télécommunications, ESIEE Paris, France
[3] Département TSI, ENST Paris, France

ABSTRACT

Very low bit-rate (VLBR) coding of speech offers the opportunity to test methods of automatic discovry and modeling of sub-word units, namely the ALISP (Automatic Language Independent Speech Processing) techniques. Experimental results are reported on American English and Czech databases. The obtained rates for unit encoding vary from 120 to 195 bps. Informal listening has convinced us, that itelligibility and speaker identity can be preserved in this coding.

Keywords: speech coding, very low bit rate, ALISP, Hidden Markov models, temporal decomposition.

1. INTRODUCTION

Current systems of automatic speech processing (ASP), including the recognition, synthesis, very low bit-rate (VLBR) coding, spoken language identification and text-independent speaker verification, rely on sub-word units determined using phonetic knowledge (phonemes, diphones, etc). To make an automatic system use those units, one needs to find their position in the speech signal: in the recognition, phonetically labelled or at least transcribed databases are widely used, while in the synthesis, an important amount of human efforts must be devoted to the creation of (usually) diphone dictionaries. The same holds for the use of such units in identification/verification or coding.

It is however possible to investigate methods for the *automatic determination* of such units based uniquely on raw speech data. In our previous work [3, 9, 10], ALISP (Automatic Language Independent Speech Processing) techniques were used for unsupervised, data-driven determination of basic speech units. Temporal decomposition (TD), vector quantization (VQ) and hidden Markov models (HMM) are the main ALISP "tools".

Although the primary target of the proposed methods are speech recognition and synthesis, *VLBR coding* offers an excellent possibility to test those methods without a link to the lexical domain (a step, which is not straightforward for data-driven techniques). The efficiency of algorithms is evaluated by re-synthesizing the speech and by comparing it to the original. If the output is intelligible, one must admit, that this representation is capable of capturing acoustic-phonetic structure of the message and that it could also be used in other domains. Moreover (in contrast with classical approaches, where the unit set is fixed a-priori and can not be altered), the coding rate in bps and the dictionary size carry information about the *efficiency* of the representation, while the output speech quality is related to its *precision*.

2. USE OF ALISP FOR VLBR CODING

As the use of ALISP units for coding at very low bit rates was already described in [3, 9, 10], we include only a brief description of different "tools".

In all experiments, the speech parameterization is done by classical LPC-cepstral (LPCC) coefficients calculated on fixed-length frames.

Temporal decomposition [1, 2] 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, de-correlation) and iterative refinement of targets and IFs.

Vector quantization was used to initially cluster the events found by the TD. It was trained with one vector

This work is supported by the Ministry of Education, Youth and Sports of the Czech Republic – project No. VS97060, and by the Research programme of Brno University of Technology "Research of electronic communication systems and technologies", No. CEZ: J22/98:262200011.

[1] Thanks to Frédéric Bimbot (IRISA Rennes, France) for the permission to use it.

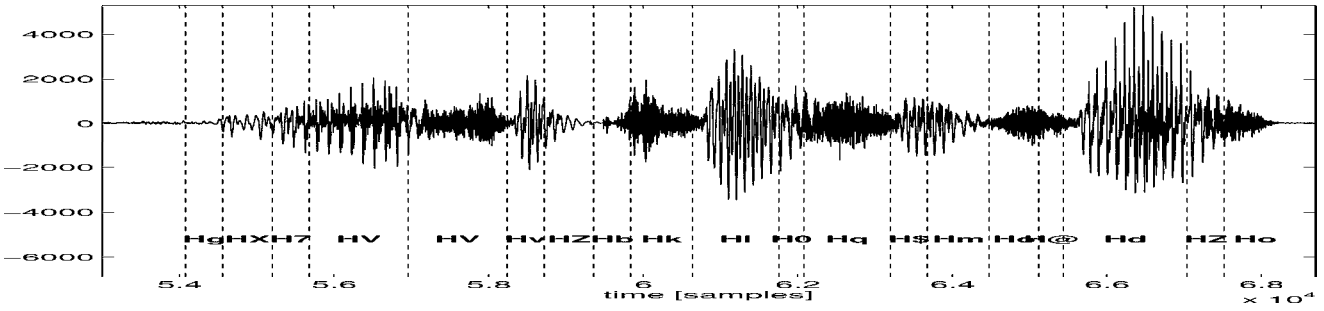

Figure 1: Example of initial segmentation and labelling obtained by temporal decomposition and vector quantization. Words "Massachusetts has" from BU corpus.

per segment (that in gravity center of TD interpolation function), but the VQ-coding worked with entire segments using cumulated distances. Figure 1 shows an example of such automatic segmentation obtained by TD and VQ.

Hidden Markov Models 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. The refinement consists of iterative steps of recognition with previously trained models, and of re-estimation of those models using the new segmentations and labellings.

Trained HMMs determine *coding units – CU* (see Fig. 2). To complete the coder, we need to define *synthesis units – SU*, which will be used in the decoder to re-synthesize the speech. Here, SU were equivalent to CU, so that the "determination of synthesis unit" box in Fig. 2 actually does not exist. This is not the optimal solution and subsection 4.2 contains a discussion on that topic. For each SU, we must dispose of several *representatives* (short matrices of spectral envelope parameters). Those are selected in the training corpus and serve as concatenation units in the synthesizer.

When coding a previously unseen speech, first the coding units are detected using the HMM recognizer. Then, synthesis units are determined. For each synthesis unit, the best representative is selected in the dictionary (currently, we use the representative which minimizes the DTW distance of input segment to all representatives for a given unit). This information is sent to the decoder. In this work, the coding of pitch trajectories and energy contour was not done, and the decoder disposes of the original information.

Based on the information received, the decoder looks up in the dictionary of representatives, and selects the appropriate one. The parameter matrix is modified using the DTW warp for each unit. Then, all those modified matrices are concatenated, the pitch and energy information is re-introduced, and the output speech is produced by the synthesizer. Currently, the synthesis is a simple LPC one.

3. EXPERIMENTS

3.1. American English

The experimental work has been conducted in speaker-dependent mode with professional speaker databases. The first set of experiments was conducted with 2 speakers from the Boston University Radio Speech Corpus [5], 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 parameterization was done by 16 LPCC coefficients, with cepstral mean subtraction on "story" (e.g. paragraph) basis. The TD was set up to produce 17-18 events per second in average. The VQ codebooks of size 64 were trained on the vectors in gravity centers of IFs, while the quantization was performed using whole segments.

The models were trained on the TD+VQ transcriptions, therefore their number is 64, each with 3 states. Using these initial models, five successive HMM sets were trained using re-segmentations and HMM re-estimations. We found that the subjectively evaluated acoustical coherence[2] of sequences improved from iteration to iteration . With "last-generation" models, we segmented the training corpus and selected the 8 longest representatives per coding unit. The synthesis units from Fig. 2 were equal to coding ones. The original pitch and energy contours, as well as optimal DTW time-warps between the original segment and the coded one were used. The index of the representative was transmitted (dotted line in Fig. 2) thus adding 3 bits per segment.

[2]Speech segments carrying the same label should be similar from a perceptual point of view.

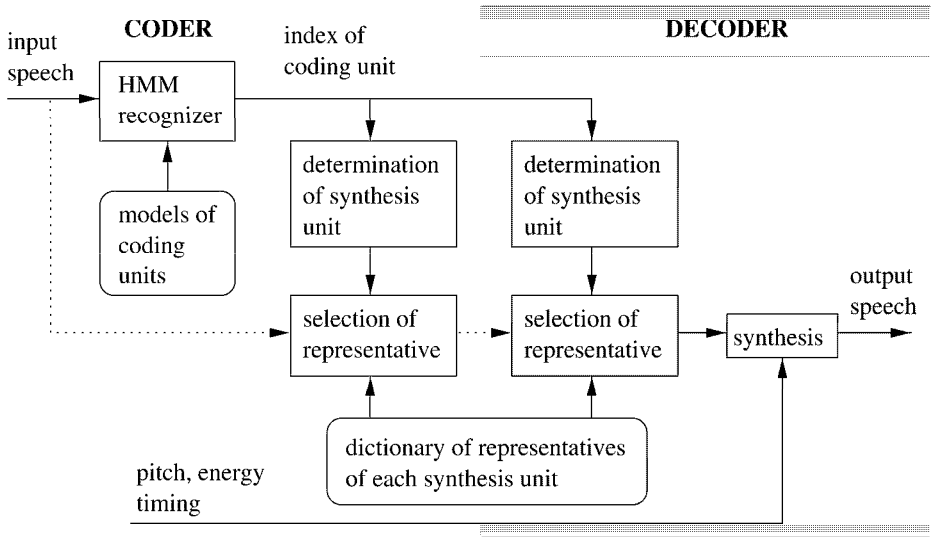

Figure 2: Coding units, synthesis units and representatives in the coder and decoder.

The resulting bit rates for unit encoding, assuming uniform encoding of their indices, are given in

Table 1 summarizes the bit rates achieved on the training and test corpora of the two speakers. The unit rate is evaluated assuming uniform encoding of their indices (rates which may be obtained using entropy encoding have been evaluated too, but they are not significantly lower than the "uniform" ones). The encoding of representatives increases the rate by 3 additional bits per unit.

3.2. Czech

Second set of experiments was conducted with a Czech database of one professional speaker. Two tapes with read texts of the known actor Martin Růžek[3] were sampled at 11025 Hz (16 bit linear resolution), and split using energy minima detection into parts of 6 to 18 seconds. Parts with foreground and background music were discarded. The total amount of data is 126.5 minutes. This DB was split into training (7/8) and test (1/8) portions.

The data were parameterized using 12 LPCC coefficients computed on 220-sample frames with 110-sample overlap. The pitch was computed using FFT-cepstral method on larger frames (500 samples). The setups for temporal decomposition, vector quantization, and HMMs were approximately the same as for BU corpus.

The temporal decomposition was set to produce 15 events per second in average. The VQ codebook size for the initial clustering was 64. Prototype HMMs with 3 data-streams (LPCC, ΔLPCC, E and ΔE), each with 3 emitting states carrying a single Gaussian component, were initialized and trained using HTK tools `HInit`, `HRest` and `HERest` (5 iterations). The first HMM decoding on the training set was done using the obtained models, and then, 5 iterations of model re-estimation (5 $\times$ `HERest`) and Viterbi decoding were run (creating 1st to 5th "generation" of HMMs). The whole refinement took approximately 8 hours on a Pentium 233 MMX under Linux.

The bit rates are presented in similar way as for BU Corpus in Table 2.

[3]Thanks to Czech Radio Brno for having granted us the access to those recordings for research and education purposes.

4. DISCUSSION AND POSSIBLE IMPROVEMENTS

Very low bit rate for unit encoding was achieved in both experimental setups. Subjectively evaluated, the speech is intelligible, but its quality is far from being optimal; the examples can be found at:
`http://www.fee.vutbr.cz/~cernocky/Samples.html`
The following subsections mention the major improvements of our scheme we are currently working on:

4.1. Synthesis

The synthesis currently used (pure LPC-one) is itself responsible for a lot of artifacts and unnatural sound of the output speech. Moreover, the link between coefficients used for unit determination and recognition, and parameterization for the speech synthesis (both are currently LPCC) can be broken – the recognition parameterization may be different from the synthesis one.

speaker	F2B		M2B	
corpus	training	test	training	test
units [bps]	189.27	190.28	189.75	195.51
unit rate [sec^{-1}]	21.03	21.14	21.08	21.72
units+representatives [bps]	**126.18**	**126.86**	**126.51**	**130.35**

Table 1: Bit rates obtained on Boston University corpus for F2B and M2B speakers. The first line stands for encoding unit indices only. The second one does not give a bit rate, but the average rate of units per second. The third one gives the bit rate necessary for units and representatives.

corpus	training	test
units [bps]	115.46	116.96
unit rate [sec^{-1}]	19.26	19.49
units+representatives [bps]	**173.26**	**175.44**

Table 2: Bit rates obtained on the corpus of Martin Růžek. The meaning of lines is the same as in Table 1

.

A more sophisticated synthesis method producing better quality speech can therefore be used, without increasing the needed bit-rate. Harmonic Noise Model [8], PSOLA or HSX are among the candidates. An issue here is the availability of the code for research purposes.

4.2. Re-definition of coding and/or synthesis units, smoothing

Coding units are trained on segments determined using the Temporal Decomposition. The TD "events" span usually a transition, a stable portion (only longer events) and another transition. The coding units therefore begin with a transition and end with another one. Obviously, we obtain distortions when concatenating such units. The idea is therefore to re-define them so that a unit would begin and end by a stable part (Figure 3). This can be done on two levels:

- prior to the model training. HMMs can be already trained as longer units.
- using the original coding units based on "short" models, but re-define synthesis units as longer ones.

Further improvement should be obtained by a more sophisticated selection of representatives (they should not only match well the input speech segment, but also concatenate well with the previous and following representative) and by properly smoothing the parametric representations on the transitions.

4.3. Coding of pitch, energy, and timing

The coding of pitch (and voicing) trajectory, energy contour and timing (DTW path), not completed in this work, should be done on segmental basis as well. All three functions can be easily parameterized, for example by linear sections. Moreover, the pitch and energy contour present regularities, so that a dictionary of typical patterns can be trained. Transmitting only indices of such patterns would further lower the bit rate. DTW alignment path can eventually be replaced by a simple linear warp.

We have estimated the bit rate necessary for pitch, energy and timing encoding to be about 200 bps, so that the total bit rate of the coder should not exceed 400 bps.

4.4. Speaker adaptation and voice modification

All experiments conducted so far in the coding were speaker-dependent. There are therefore all the issues of *speaker adaptation* (which can be resolved by the normalization of voices to a generic speaker, or by the adaptation of models) which have to be investigated. Fortunately, here we can rely on many references, as those issues have been extensively studied in the recognition [4]. It is however to be verified, how the standard approaches (maximum-likelihood linear regression (MLLR) and others) do perform in the coding with ALISP units, where the objectives are different from the recognition.

Linked to the previous problem is the *speaker modification*. If not used, the voice of decoder will always be that of the speaker(s), who created the training set. There is a possibility to transmit modification parameters, allowing to transform the voice of a generic

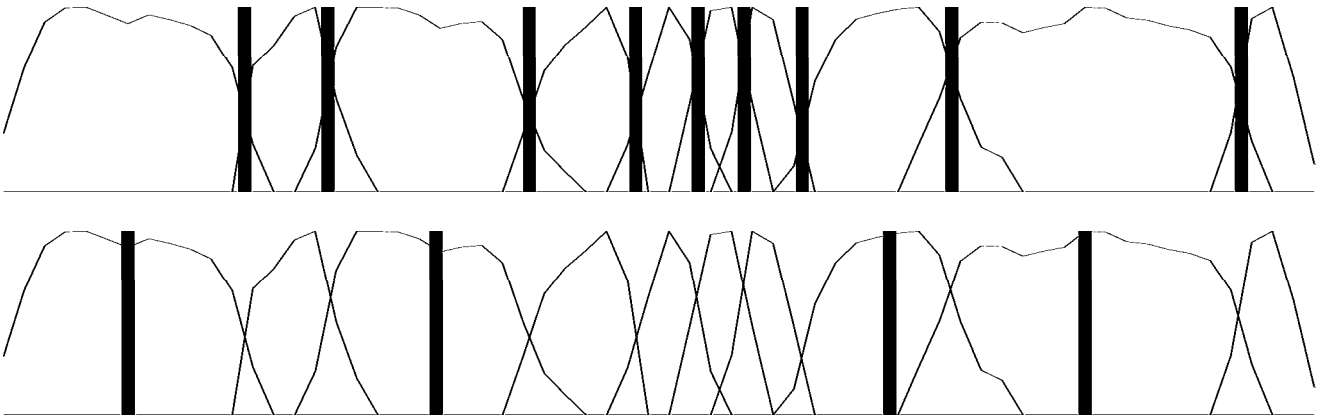

Figure 3: Two ways of unit determination based on temporal decomposition events (bold lines stand for segment boundaries). On the upper panel the current one, making use of intersections of interpolation functions. On the lower panel the proposed one, spanning a unit from one stable part to another. Short (unstable) TD events should not be cut during the unit creation.

speaker to the desired one [8]. It has however been shown by Ribeiro and Trancoso [6, 7], that this information increases considerably the bit-rate. In some applications (military for example), we would accept the alteration of voice at the price of very low rate.

5. CONCLUSION

The coding using ALISP units gives fair and promising results; with the improvements mentioned in the previous section, this scheme should be suitable for very low bit rate transmission over the Internet, storage of huge amount of speech data, and other applications.

6. REFERENCES

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

[2] F. Bimbot, G. Chollet, P. Deleglise, and C. Montacié. Temporal decomposition and acoustic-phonetic decoding of speech. In *Proc. IEEE ICASSP 88*, pages 445–448, New York, 1988.

[3] G. Chollet, J. Černocký, A. Constantinescu, S. Deligne, and F. Bimbot. *Computational models of speech pattern processing*, chapter Towards ALISP: a proposal for Automatic Language Independent Speech Processing, pages 375–388. NATO ASI Series. Springer Verlag, 1999.

[4] C.J. Leggetter and P.C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, (9):171–185, 1995.

[5] M. Ostendorf, P.J. Price, and S. Shattuck-Hufnagel. The Boston University radio news corpus. Technical report, Boston University, February 1995.

[6] C.M. Ribeiro and I.M. Trancoso. Application of speaker modification techniques to phonetic vocoding. In *Proc. ICSLP 96*, pages 306–309, Philadelphia, 1996.

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

[8] 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.

[9] J. Černocký. *Speech Processing Using Automatically Derived Segmental Units: Applications to Very Low Rate Coding and Speaker Verification.* PhD thesis, Université Paris XI Orsay, December 1998.

[10] J. Černocký, G. Baudoin, and G. Chollet. Segmental vocoder - going beyond the phonetic approach. In *Proc. IEEE ICASSP 98*, pages 605–608, Seattle, WA, May 1998. http://www.fee.vutbr.cz/~cernocky/Icassp98.html.