Very Low Bit Rate Speech Coding: Comparison of Data-Driven Units with Syllable Segments*

Jan Černocký[1], Ivan Kopeček[2], Geneviève Baudoin[3], and Gérard Chollet[4]

[1] Brno Univ. of Technology, Inst. of Radioelectronics, cernocky@urel.fee.vutbr.cz
[2] Masaryk University Brno, Faculty of Informatics, kopecek@fi.muni.cz
[3] ESIEE Paris, Dpt. Signal et Télécommunications, baudoing@esiee.fr
[4] ENST Paris, Dpt. Signal et Images, chollet@sig.enst.fr

Abstract. Very low bit-rate (VLBR) coding of speech offers the opportunity to test methods of automatic generation of sub-word units. This paper describes two approaches to VLBR coding: the first based on ALISP (Automatic Language Independent Speech Processing) techniques, the second based on syllable segments. Experimental results are reported on a database of one Czech professional speaker. The obtained rates for unit encoding were approximately 135 bps for the former approach and 62 bps for the latter. The quality was evaluated by measuring the logarithmic spectral distortion (computed on LPC-spectra), and in informal listening tests. Possible mutual profits of each technique to the other are discussed.

1 Introduction

Current systems of automatic speech processing (ASP), including the recognition, synthesis, very low bit-rate (VLBR) coding 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 know 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 dictionary. The same holds for the use of such units in identification/verification or coding.

It is however possible to investigate the methods of *automatic determination* of such units based uniquely on raw speech data. In our previous works [10–12], 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". On the other hand, the use of syllable segments for speech processing was investigated in [4, 5, 3, 6]. The syllable

* The research has been partially supported by the Ministry of Education, Youth and Sports of the Czech Republic, project Nbs. VS97060, VS97028, and by the Grant Agency of the Czech Republic under the Grant 201/99/1248.

based approach is motivated by the fact that syllables create perceptually and acoustically coherent units and that they also represent the basic prosodic units. Automatic segmentation into syllable segments [4, 9] is achieved using estimation of syllable time duration (by estimation of speech rate and articulation rate), followed by the estimation of the syllable boundary by means of functions of sonority decrease, acoustical intensity and acoustical changes.

Although the main target of proposed methods are speech recognition and synthesis, the *VLBR coding* offers an excellent possibility to test those methods without a passage 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 is appropriate also in other domains. Moreover (in contrast with classical approach, 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*. This paper is organized as follows: section 2 describes the ALISP framework for VLBR coding while the following section 3 concentrates on syllable segments. Section 4 covers the experimental part of the work and section 5 details on the possible mutual enhancements of ALISP and syllable techniques. Section 6 concludes the paper.

2 Use of ALISP for VLBR coding

The use of ALISP units for coding at very low bit rates was already described in [10–12]. Therefore, 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] 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 was trained with one vector per segment (that in gravity center of TD interpolation function), but the VQ-coding worked with entire segments using cumulated distances. *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. Once the units are found (and the models trained), the *coding* of previously unseen signals can be done following Fig. 1.

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

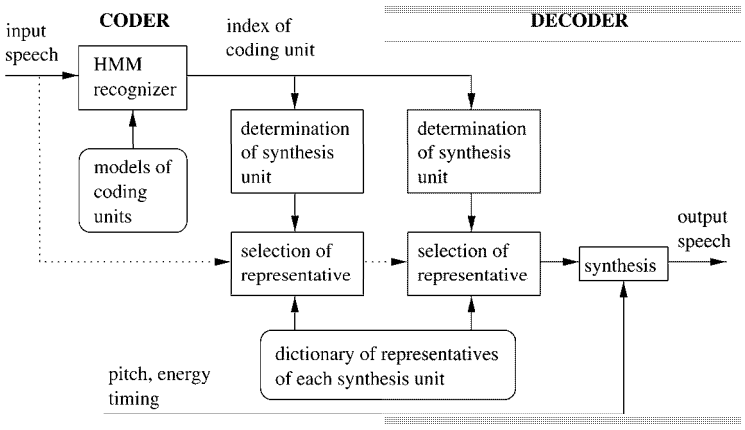

Fig. 1. Coding units, synthesis units and representatives in the coder and decoder. The information about chosen representative can be either transmitted (dotted line) or re-created in the decoder.

3 Syllables and syllable segments

We use the notion *syllable segments* instead of syllables because we have no reliable specification of syllables. Although the feeling of syllables and syllable boundaries is usually very natural and strong, it is subjective and in many cases also not unique. A problem complicating the determination of an appropriate set of syllable segments is the necessity to respect the coarticulation between the adjacent syllables and to keep the number of segments reasonable. To solve this problem we use the fact that the coarticulation is in most cases reduced on the boundaries of the adjacent syllables in comparison to the inner parts of syllables. Also, syllabic onsets are often realized in a standard form [4, 8] and acoustical intensity typically decreases at the syllable boundaries. For this purpose, some classes of syllable segments were proposed and analyzed (L-syllables, CVS-syllables, and other types [7, 6]).

4 Experiments

4.1 Database

The experimental work has been conducted in speaker-dependent mode with a Czech database of one professional speaker. Two tapes with read texts of the well known actor Martin Růžek[2] 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.

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

4.2 ALISP experiments

The data were parameterized using 12 LPCC coefficients computed on 220-sample frames with 110-sample overlap (pre-distortion with $\alpha = 0.95$ and Hamming window were applied). The pitch was computed using FFT-cepstral method on larger frames (500 samples). 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.

In the coding step, 8 longest representatives per coding unit were selected in the training corpus. The synthesis units from Fig. 1 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 information about representative was transmitted (dotted line in Fig. 1) thus adding 3 bits per segment. The overall rates with 5-th generation models for coding unit transmission (assuming uniform encoding of their indices) are 115.46 bps and 116.96 bps for the training and test sets respectively. The average unit rate per second is 19.26 and 19.49, so that the total coding rates (not taking the prosody information into account) are **173.26 bps** and **175.44 bps**.

The *objective* measure of the quality was the mean of logarithmic spectral distortions:

$$D = \sqrt{\int_{-1/2}^{1/2} \left[10 \log \hat{S}(f) - 10 \log S(f)\right]^2 df} \quad \text{in dB}, \tag{1}$$

evaluated using the power frequency responses of prediction filters ($\hat{S}(f)$ stands for the response with coded coefficients and $S(f)$ for the same function with the original ones). For the part of test file `df232000` (the same was tested with syllable segments), $D = 4.39$. *Subjectively,* the resulting speech is intelligible, but quite unnatural and with strongly audible artifacts. One must however take into account, that no smoothing was performed on the borders of segments, and that the synthesis (pure LPC) is quite primitive. Smoothing techniques and better synthesis (HNM, PSOLA) should improve the quality of the coder. Another issue is the encoding of prosody, which was not resolved in our work, and should be done as well on the segmental basis.

4.3 Syllable segments experiments

The procedure for segmentation to the syllable segments, used in the experiment, was based on the method which was used for semi-automatic segmentation procedure for syllable speech synthesizer *Demosthenes* [7, 6]. It detects local

maxima of the *function of sonority decrease*. Provided $v_i = (v_{i,1}, v_{i,2}, ..., v_{i,n})$ is the acoustic vector in the time i, the function of sonority decrease is based on the value

$$D(i) = \sum_{j=k}^{l} c_j(i),$$

where $c_j(i) = max\{v_{i,j} - v_{i-\delta,j}, 0\}$. The indices k, l correspond to the vowel formants frequency region, and δ is the length of the used difference. The coding consists in comparing a previously unseen syllable segment to all segments in the training corpus and in choosing the optimal one. For this comparison, 16-element acoustic vectors are obtained by modeling the Corti organ. Segments are compared using a metric based on a sum of differences of vector elements, weighted by ratios of those elements. The time is warped using linear transformation.

The resulting rate on the test corpus is 62.3 bps. The logarithmic spectral distortion for part of the test file `df232000`, evaluated using Eq. 1, was 8.90 dB. Subjectively, the quality of coded speech is poor and the utterances are hard to understand.

One of the goal of the experiment was to test whether the inaccuracy of the method will have substantial negative consequences for coding. Rather poor quality of the output shows, that such a simple type of the segmentation should be substituted by a more precise method. Also, use of a metrics developed specially for syllable segments could increase the quality of the output. On the other hand, low transmission rate and possibilities of enhancing the used method give a chance for further development in this direction. Statistical evidence of the syllable segments frequencies can be used to further decrease the bit rate (using entropy encoding).

5 ALISP and syllables helping each other

It is obvious (though we have not yet managed to prove it experimentally), that ALISP techniques and syllable segments can be mutually profitable. So the later approach can take advantage from its ALISP counterpart in the following way:

- passage from semi-automatic to fully automatic determination of syllable segments in the corpus by application of clustering techniques.
- automatic segmentation of the corpus into syllable segments by training an HMM for each segment, and by Viterbi decoding.

while ALISP techniques can be enhanced by the works in syllable segments:

- studies of allowed transitions and coarticulation can be used as rules to drive the sequencing of units by multigrams [2], so that the intra-unit coarticulation is maximized, while the inter-unit one is minimized.
- prosody patterns determination techniques can be used for prosody encoding in an ALISP-based VLBR coder.

6 Conclusion

The coding using ALISP units gives fair and promising results; with the improvements mentioned in par. 4.2, this scheme should be suitable for very low bit rate transmission over the Internet, storage of huge amount of speech data, and other applications.

As for the possible use of the syllable segments, we would like to use more precise methods for segmentation, apply special metrics for comparison of the syllable segments and apply statistical method for evaluation of the optimal amount of the training data as well as for the optimization of the bit rate.

Our future works in the VLBR coding will aim at the mutual improvement of the approaches (as described in section 5), and at using those techniques in the continuous speech recognition.

References

1. F. Bimbot. An evaluation of temporal decomposition. Technical report, Acoustic research department AT&T Bell Labs, 1990.
2. S. Deligne. *Modèles de séquences de longueurs variables: Application au traitement du langage écrit et de la parole.* PhD thesis, École nationale supérieure des télécommunications (ENST), Paris, 1996.
3. G. Doddington. Syllable based speech processing. Technical report, J. Hopkins University, 1997. WS97 Project Report, Research Notes No. 30.
4. S. Greenberg. Speaking in shorthand - a syllable-centric perspective for understanding pronunciation variation. In *Proc. Workshop Modeling Pronunciation Variation for Automatic Speech Recognition*, pages 47–56, 1998.
5. L. Josifovski, D. Mihajlov, and D. Gorgevik. Speech synthesizer based on time domain syllable concatenation. In *Proc. SPECOM'97*, pages 165–170, Cluj-Napoca, 1997.
6. I. Kopeček. Syllable based speech synthesis. In *Proc. 2nd International Workshop SPECOM'97*, pages 161–165, Cluj-Napoca, 1997.
7. I. Kopeček. Speech synthesis based on the composed syllable segments. In *Proc. of Workshop on Text Speech and Dialogue (TSD'98)*, pages 259–262, Brno, Czech Republic, September 1998.
8. I. Kopeček. Syllable segments in czech. In *Proc. XXVII. Mezhvuzovskoy naucznoy konferencii*, pages 60–64, St. Petersburg, March 1998. Vypusk 10.
9. I. Kopeček and K. Pala. Prosody modeling for syllable-based speech synthesis. In *Proc. IASTED Conference on AI and Soft Computing*, pages 134–137, 1998.
10. J. Černocký. *Speech Processing Using Automatically Derived Segmental Units: Applications to Very Low Rate Coding and Speaker Verification.* PhD thesis, Université Paris XI Orsay, 1998.
11. 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.
12. J. Černocký, G. Baudoin, D. Petrovska-Delacrétaz, J. Hennebert, and G. Chollet. Automatically derived speech units: applications to very low rate coding and speaker verification. In *Proc. of Workshop on Text Speech and Dialogue (TSD'98)*, pages 183–188, Brno, Czech Republic, September 1998.