APPLICATION OF RE-SEGMENTATION IN VERY LOW BIT RATE SPEECH CODING

Petr Motlíček, PhD student
Institute of Radio Electronics, Faculty of Electrical Engineering, TU Brno
Supervisor: Dr. Ing. Jan Černocký

Abstract:

The aim of our effort is to reach higher quality of resulting speech coded by very low bit rate (VLBR) segmental coder. The basic units are found automatically in a training database using temporal decomposition and vector quantization. They are modeled by HMMs. Then two methods of re-segmentation were used in order to find new longer units. In the first approach borders are set to the centers of previous units. In the second, borders are fixed to the centers of middle HMM states of previous units. Number of frames in new units is conditioned to be bigger than a fixed constant. Hence, new units can consist of a several previous segments. Decreasing transition noise of resultant speech was obtained using these techniques.

1 Introduction

When we speak of very low bit rate coders, segmental or phonetic vocoders are meant. Only those vocoders based on recognition and synthesis are able to efficiently limit bit rate. The coder and the decoder share the database of speech units (segments) that are considered to be representatives of any speech uttered by any speaker. Only the indices of representatives and some prosodic informations are transmitted by this coder. Hence, the bit rate of these types of coders can be less than 300 bps. The quality of this speech coding approach depends on a lot of factors. Among the most important is the quality of recognition of speech units. But speech analysis and synthesis are not less significant.
In our approach defining of speech units influences resulting quality of coder. Using phonetically labeled or transcribed speech database would discharge us these troubles. Unfortunately, current phone recognizers provide sufficient performances when the recognized speech is similar to the training database only. Hence, in our approach, speech units are supposed to be found automatically (by Automatic Language Independent Speech Processing (ALISP) tool) before training of recognizer. The fact that we do not need transcribed and labeled speech database can be great benefit of this method. The coder can be easily used in languages lacking standard speech databases.

2 ALISP tools

Using of ALISP units for very low bit rate speech coding is in more details described in [1], [2]. All our techniques of the re-segmentation are based on ALISP. Hence, only brief description is included.

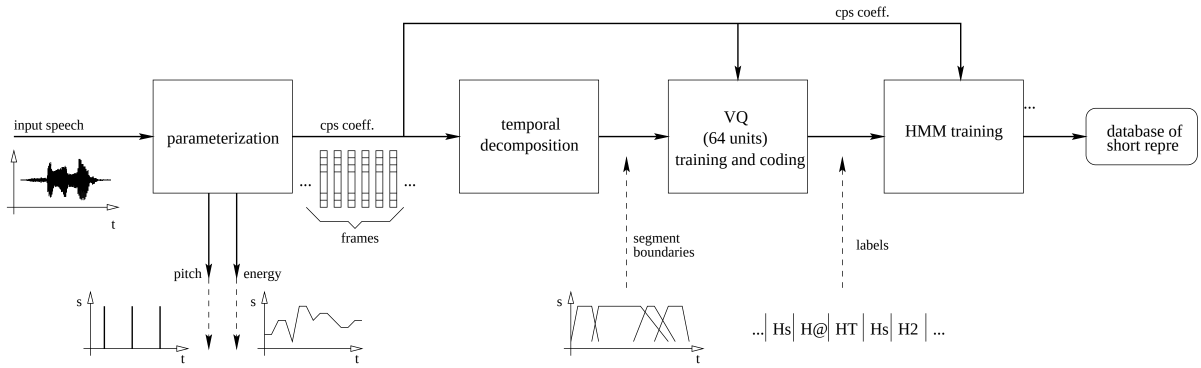

Figure 1: Scheme of training process.

2.1 Temporal decomposition

Temporal decomposition (TD) [3] is used for the initial segmentation of the speech into quasi-stationary parts. The matrix of spectral parameters is decomposed into a limited number of events. Each event is represented by a target vector and an interpolation function. The td95 package is used for the TD[1]. Only on training corpus, the temporal decomposition was applied. In addition, segments are generated in spectrally stable parts of speech. For each segment, the gravity center frame is computed.

2.2 VQ clustering

Vector Quantization (VQ) [4] is used for the clustering of the segments created by TD. It is obvious, that VQ is applied on training data, only. VQ consist of training, quantization and post-processing.

2.3 HMMs

Hidden Markov Models (HMMs) are widely used in speech recognition because of their acoustic modeling capabilities.
HMMs are related to original VQ symbols, so that their number is 64. The number of emitting states per model is fixed to 3. The models are initialized as left-right without state skipping. Not only one set of HMMs is generated. We have found that an iterative approach can improve the acoustical quality of units. Hence, several generations of models are created.
The whole training process described above is shown in Fig. 1.

3 Re-segmentation

At this point, the re-segmentation of the original units recognized by HMMs is used. This technique is applied in order to decrease the influence of transition noise on the synthesized resulting speech. It is obvious that the transition noise will be always presented in our resulting decoded speech, because of the technique on which our approach is based. But applying appropriate segmentation, its influence can be notably decreased.
Original TD segments, on which HMMs are afterwards trained, are created, so that they have contained stable parts of processed speech. Therefore, the boundaries of these segments are

[1] Thanks to Frédéric Bimbot (France) for permission to use it.

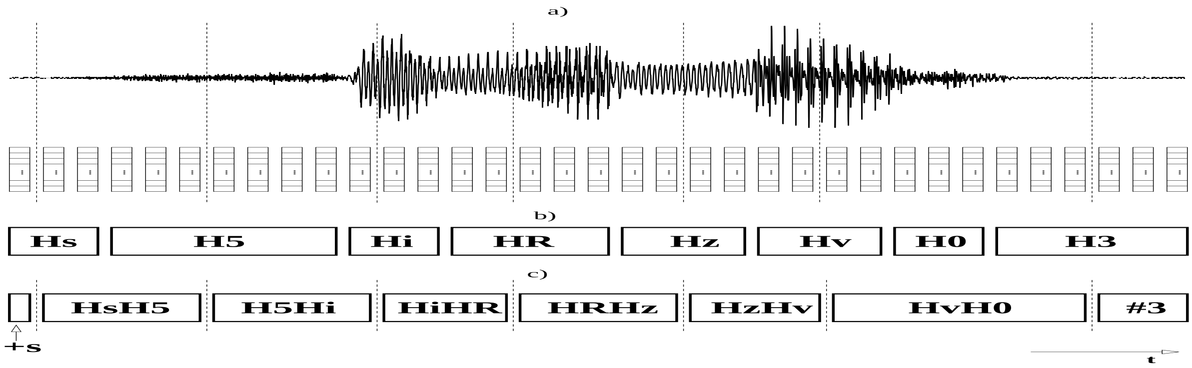

Figure 2: Example of re-segmentation according to middle frames of old units. Minimal length of new units is 4 frames. a) speech signal with its splitting into the frames, b) original segmentation recognized by HMMs, c) new re-segmentation.

set to non-stable parts of speech that mostly contains small energy of signal. In decoder, where chosen appropriate representatives are concatenated to create resulting speech; these representatives are concatenated mostly in parts with small energy, as well. It is obvious that representatives do not come from original coded speech. They are chosen from training data set, so that they the best substituted coded units. Hence, transition noise appears in resulting speech.

The quality of signal can be described by ratio of signal to noise. In localities of segment boundaries, the ratio is small because of small energy of signal and quite big value of transition noise. But, if representatives in synthesis are concatenated in localities of stable parts of signal, where the energy of signal is mostly much higher, the transition noise would not come out so much in resulting speech and the ratio of signal to noise would be higher, as well. This is the general principle of our approach.

One can say that instead of the re-segmentation of original units, new alternative segmentation could have been done at the beginning of our job. However the aim is not only creating units, so that their boundaries are set to the stable parts of speech signal. A new longer units that cover more non-stable parts of signal are required. It would be difficult to create them by TD and to train HMMs afterwards. Hence, the re-segmentation of original units is done after HMM recognition. The method is applied on recognized label sequence that comes from last HMM iteration. Obviously, it must be applied on training data set, as well as on the test one. One of the difficulties can be where to put the boundaries of new created segments in original ones. Hence, two experiments were done.

In first experiment, the new segment boundaries are set to the centers of previous segments. These centers were determined according to their middle frames. In the second experiment, the middle frames of middle HMM states in original segments interpret new segment boundaries. As mentioned before, new units are supposed to be longer than original ones. Hence, not each original segment will contain new segment boundary.

3.1 Re-segmentation according to middle frames of old units

In this approach, the boundaries of new units are put to the centers of old ones. Several experiments were done with this method that were different in minimal length of new units, as mentioned before. The minimal length represents the minimal number of frames in created

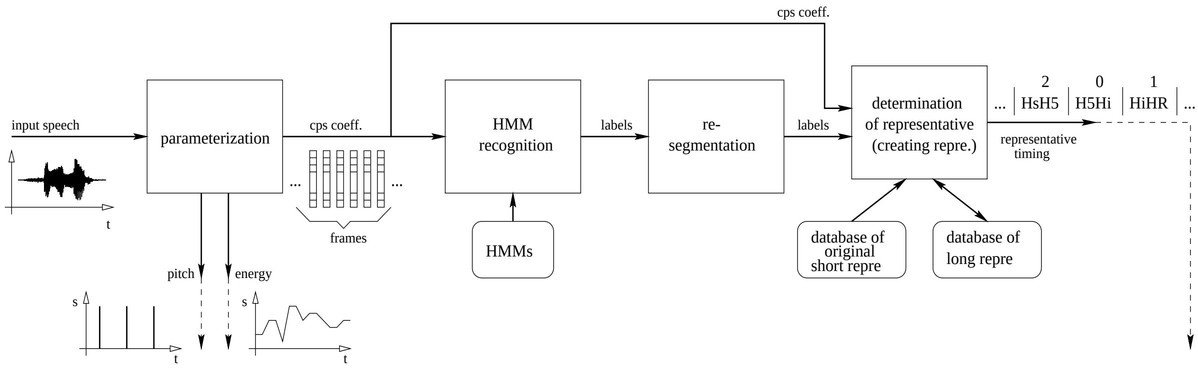

Figure 3: Scheme of coder used in our experiments. Only indices of coding units and numbers of chosen representatives are transmitted to the decoder.

new units. The algorithm of the re-segmentation is: First, the centers of old units are found. Then, we move from one center to another and remember the number of frames we went over. If number of frames between two neighboring centers is less than required, the second center is not declared as new segment boundary and we move to another old unit's center. This advance is still repeated unless we go over required minimal number of frames. Graphically the re-segmentation method can be seen in Fig. 2. The names of the whole new units consist of the names of old units that are covered by new one, as again seen in Fig. 2c.

3.2 Re-segmentation according to middle frames of middle states of HMMs

As mentioned before, in this approach, the new segment boundaries are represented by center frames of the middle HMM states of old units. The number of emitting states per one HMM is fixed to 3, as described in 2.3. Each state must contain one frame, at least. Hence, the minimal number of frames in an original unit is 3. If the number is higher, the frames are split into states according to score of recognition. It is obvious that the resulting segmentation based on this approach will be different to the first one.

4 Representatives and synthesis

In our experiments, parametric LPC synthesis was applied. To complete the coder we need to define define the synthesis units that will be used in the decoder to synthesize the resulting speech. Hence, for each unique dictionary unit, the three longest units from the training data set are kept, so that they are mostly down-sampled when being converted to shorter segments. It is obvious that the attention is already payed to the training units after the re-segmentation. When coding a previously unseen speech, first the coding units are detected using the HMM recognizer. Then, the stream of recognized units is re-segmented by one of the method described in 3. For each coding unit, the best synthesis unit (from 3 representatives) is chosen. The choice is done using minimum Dynamic Time Warping (DTW) distance between a representative and an input speech segment (coding unit). Important part of the synthesis is the prosody. The original prosody was used in our experiments.

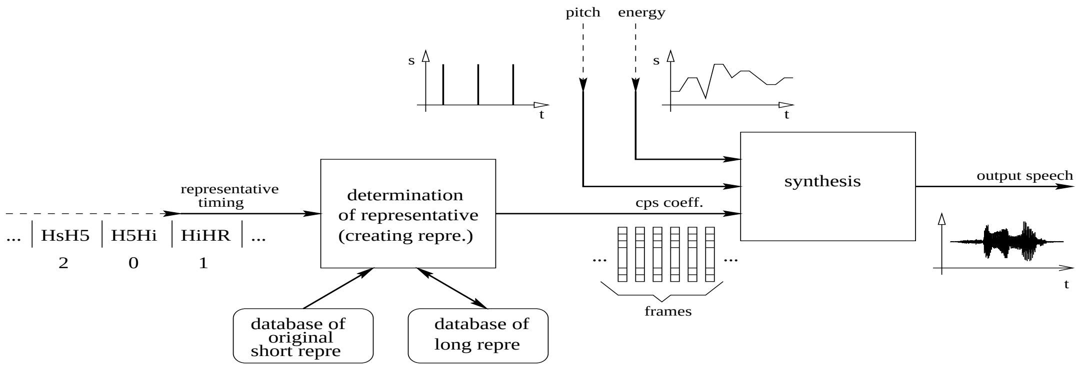

Figure 4: Scheme of decoder.

5 Results

The scheme of coder is given in Fig. 3. Only indices of coding re-segmented units with numbers of chosen representatives, as well as the DTW paths are transmitted to the decoder. The scheme of decoder can be seen in Fig. 4. Several experiments based on re-segmentation were done. They distinguish from each other in method used for re-segmentation, as described in 3.1 and 3.2. Furthermore, they are different in the length of created units, as mentioned in section 3. When applying the re-segmentation methods on original units, the number of created units in coding sentence is always less than without re-segmentation. In spite of this fact, the bit-rate does not necessarily decrease. The re-segmentation greatly increases the number of re-segmented unique units. Hence, more bits are needed when transmitting the indices of coding units. Therefore, resulting bit-rate depends on the lengths of new re-segmented units.

6 Conclusion

The purpose of applying the re-segmentation techniques was to reach higher quality of resulting speech coded by VLBR coders. This aim was achieved with all our experiments that were built on the re-segmentation. Some examples of resulting speech can be found on:
http://www.fee.vutbr.cz/∼motlicek/speech.html.
The speech coded only using original units (re-segmentation not used) and the resulting average bit rates of speech coming from all our experiments are given there, as well.

References

[1] J. Černocký. Speech Processing Using Automatically Derived Segmental Units, *PhD Thesis, ESIEE*, France, 1998.

[2] J. Černocký, G. Baudoin, and G. Chollet. Segmental Vocoder-going beyond the phonetic approach. *Proc. ICASSP Seattle* pp. 605-608, May 1998.

[3] B. S. Atal. Efficient Coding of LPC Parameters by Temporal Decomposition. In *Proc. IEEE ICASSP 83*, pp. 81-84, 1983.

[4] J. Makhoul, S. Roucos, and H. Gish. Vector quantization in speech coding. In *Proc. IEEE*, 73(11):1551-1589, November 1995.