Minimization of transition noise and HNM synthesis in very low bit rate speech coding

Petr Motlíček[1], Geneviève Baudoin[2], Jan Černocký[1] and Gèrard Chollet[3]

[1] Brno Univ. of Technology, Inst. of Radioelectronics,
Purkyňova 118, 612 00, Brno, Czech Republic
{motlicek,cernocky}@urel.fee.vutbr.cz
http://www.fee.vutbr.cz/~motlicek

[2] ESIEE Departement Signaux et Télécommunication, BP 99, cedex,
Noisy-le-Grand, 93162, France
baudoing@esiee.fr

[3] CNRS URA-820, ENST, Dept. TSI,
46 rue Barrault, 75634, PARIS-cedex 13, France
chollet@tsi.enst.fr

Abstract. The aim of our effort is to reach higher quality of resulting speech coded by very low bit rate (VLBR) segmental coder. In already existing VLBR coder [1], we want to improve the determination of acoustical units. Furthermore, better analysis-synthesis technique for the synthesis part (Harmonic-Noise Model) instead of LPCC is going to be used. The VLBR coder consists of a recognition system followed by a speech synthesizer. The recognizer identifies recognition acoustic units (RU). On the other hand, the synthesizer concatenates synthesis acoustic units (SU). However, the two kinds of acoustic unit can be identical or different and then can be modeled in different ways such Hidden Markov Model for the RU and Harmonic-Noise model for the SU. Both kinds of units are obtained automatically from a training database of raw speech that does not contain any transcription. In the original version of the coder [1], the quality of the synthetic speech was not sufficient for these two main reasons: the SU units were too short and difficult to concatenate and the synthesis was done using basic LPCC analysis-synthesis. In order to remove first drawback, three methods of re-segmentation were used. Afterwards, the basic LPCC analysis-synthesis was replaced by HNM.

1 Introduction

When we speak of very low bit rate coders, segmental or phonetic vocoders are meant [4]. Only those vocoders based on recognition and synthesis are able to efficiently limit the 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 information are transmitted by this coder. Hence, the bit rate of these types of coders can be less than 350 bps. The quality of this speech coding approach depends on many factors. Among the most important is the quality of recognition of speech units. But speech analysis and synthesis are not less significant. The definition of speech units influences resulting quality of the coder. In our experiments, speech units are found automatically (by Automatic Language Independent Speech Processing (ALISP) tools) before training of recognizer. The fact that we do not need transcribed and labelled speech database is a great benefit of this method. The coder can be easily used in languages lacking standard speech databases. When a set of speech units is obtained, they can be used for coding. The coder consists of recognizer acoustically labelling the speech and of additional information encoder. In the decoder, synthesis built on concatenation of examples from the training corpus is applied in order to obtain output speech. A technique based on applying automatically derived speech units was developed at ENST, ESIEE and VUT-Brno [1], [2]. However, the quality of synthesized speech is not sufficient. This paper reports experiments based on re-segmentation of original units. The aim of the re-segmentation is removing transition noise from the synthesized speech, which is caused by concatenation of chosen representatives in the decoder.

2 Basic structure of the coder

All our experiments are built on Boston University Radio Speech Corpus [5] database (DB) collected in 1995. Use of ALISP units for very low bit rate speech coding is in more details described in [1], [2]. For the initial segmentation Temporal decomposition (TD) [3] is applied. Created segments are clustered by Vector Quantization (VQ).

Hidden Markov Models (HMMs) [1] are widely used in speech recognition because of their acoustic modeling capabilities. HMMs were applied only in our first two experiments of the re-segmentation. 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. We have found that an iterative approach can improve the acoustical quality of units. Hence, several generations of models are created.

Units found using HMM or TD+VQ segmentation are referred as "original" or "short" units. In the baseline version of the coder, a limited number of representatives is found in the training data for each unit. In the coding of unseen speech, the input signal is labelled by HMM or TD+VQ, and the optimal representative is selected for each detected unit. The information about units as well as about representatives is transmitted to the decoder, where a concatenation synthesis takes place. During this synthesis, a transition noise can appear in points where representatives were concatenated.

3 New units

As mentioned before, the re-segmentation of the original units recognized by HMMs or VQ is applied in order to decrease the influence of transition noise on

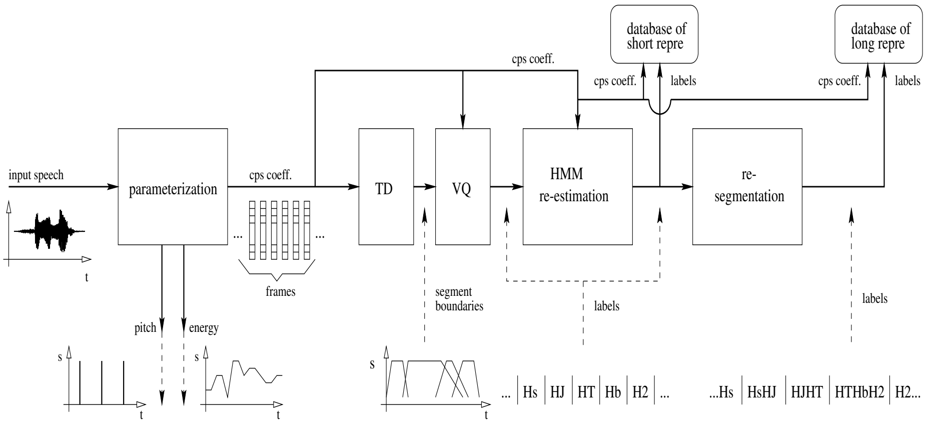

Fig. 1. Scheme of the whole training process based on ALISP.

the resulting synthesized speech. Original TD segments, on which VQ or HMMs are afterwards trained, contain stable parts of speech in their centers. Therefore, the boundaries of these segments are set to non-stable parts of speech that mostly contain small energy of signal. Hence, in the decoder (where chosen appropriate representatives are concatenated to create resulting speech) these representatives are concatenated mostly in parts with small energy, so that the signal-to-noise ratio is low. 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, whose boundaries are set to the stable parts of speech signal. 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 VQ codebook or HMMs afterwards. Hence, the re-segmentation of original units is done after VQ or HMM recognition.

3.1 Re-segmentation according to middle frames of original units

In this approach, the boundaries of new units are placed to the centers of original ones, as can be seen in Fig. 2. Several experiments were done that differ in minimal length of new units. The minimal length represents the minimal number of frames in created 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 process is iterated unless we go over required minimal number of frames. It is obvious that the re-segmentation starts from the first center of first original unit. The "prefix" part of first original unit is declared as an independent new unit. The same problem appears in the last processed old unit. The names of the whole new units consist of the names of original units that are covered by new one.

Let suppose that the original label sequence is: Hs HF H7 Hr Hi ... and a new segment boundary is going to be fixed into center of H7 segment. After the re-segmentation according to this approach, the label sequence will be: HsHFH7 H7HrHi ...

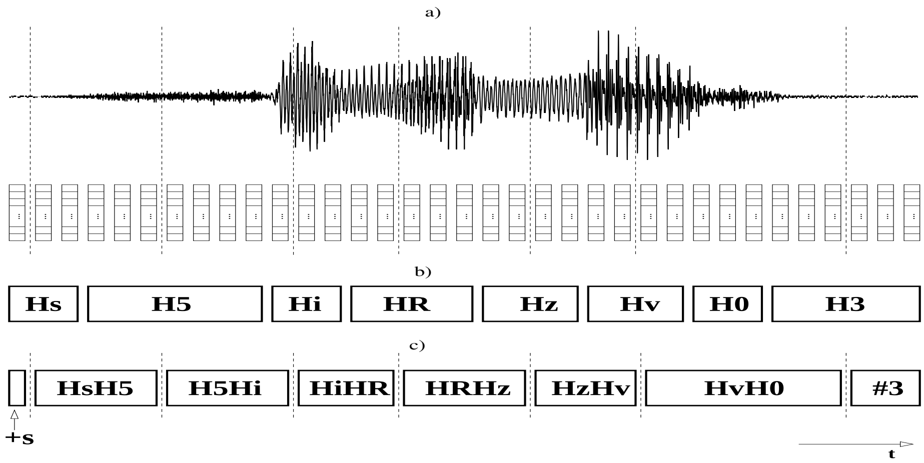

Fig. 2. Example of re-segmentation according to middle frames of original units. Minimal length of new units is 4 frames: a) speech signal with its splitting into the frames. b) original segmentation recognized by HMMs. c) new re-segmentation.

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

In this approach, new segment boundaries are represented by center frames of the middle HMM states of old original units. The number of emitting states per one HMM is fixed to 3. Each state must contain one frame, at least. Hence, the minimal number of frames in an original unit is 3, as well. If the number is higher, the frames are assigned to states according to likelihood scores. It is obvious that the resulting segmentation based on this approach will be different from the first one.

3.3 Re-segmentation according to gravity centers of original TD-based units

In the last experiment, the segment boundaries are supposed to stand in gravity centers of original segments, derived by TD. The goal is that gravity centers are one of the most suitable positions in segments, where the spectrally stable parts of speech can be expected. The re-segmentation can not be built on label sequence recognized by HMM, because this sequence does not match with label sequence obtained by TD. Hence, the re-estimation by HMMs is not used. However, as before, not each gravity center of original segment will represent the new segment boundary. In the sequence of interpolation functions (IFs), their width in frames is determined. Only a gravity center which lies in IF that is wide enough can represent a segment boundary. The sufficient width of IFs is evaluated according to an a-priori chosen constant. In sequence of IFs, several consecutive narrower IFs than required can appear. Afterwards, when applying the previous condition only, the distances in frames between new segment boundaries would be too important, and new units would be too long. Therefore, each new unit is constrained in length so that it can cover less original units than another a-priori set constant. Due to, we can easily control lengths of new units. The example is given in Fig. 3.

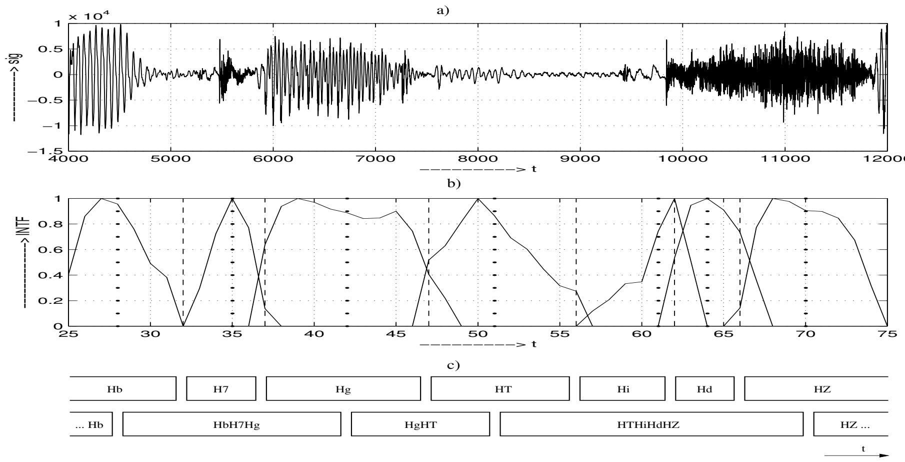

Fig. 3. Illustration of Temporal decomposition and re-segmentation process according to gravity centers of original units on chosen part of speech: a) speech signal. b) interpolation functions (solid lines) with positions of gravity centers of each segment (dotted lines) and segment boundaries of original units (dashed lines). c) original segmentation and new re-segmentation of speech.

4 Representatives for synthesis

In our experiments, LPCC and Harmonic-Noise model (HNM) analyzes-synthesis algorithms were applied. To complete the coder, we need to define the synthesis units that will be used in the decoder to synthesize the resulting speech. 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. Obviously, the attention is already paid to the training units after the re-segmentation. When coding a previously unseen speech, first the coding units are detected using the HMM recognizer (in first two methods of re-segmentation) or by TD+VQ (in the third method of re-segmentation). Then, the stream of recognized units is re-segmented. 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. When selecting the representatives to synthesize a previously unseen speech, we can easily find out that in the coded speech, there are some coding units that do not have equivalent representative stored in DB of representatives (based on training data set). It is caused by the re-segmentation of original units, because the theoretical number of unique units after the re-segmentation is infinite. A new long unit created by some of the re-segmentation method can consist of two, three, or more original units, depending on the minimal required length. Hence, many re-segmented coding units can appear that have not been seen in training data set and for which we do not have any appropriate representative. Therefore, two approaches were developed in order to solve this problem.

4.1 Seeking the best synthesis unit from existing ones

Instead of non-existing synthesis unit, some existing one will be used. Seeking the appropriate existing synthesis unit by DTW or another method, based on searching minimum distance between two segments, would result in very long search time. Hence, in our experiments, replacing unit will be sought using Euclid distances between original short units from which the longer ones consist of.

Unfortunately it can happen that any appropriate representative is found applying this method. In this case, the representative will be created.

4.2 Creating the representative for non-existing coding unit

The representative for non-existing coding unit can be made from original short representatives. These short representatives were created from tge original units before the process of re-segmentation. Only the longest synthesis unit from each class (from training data set, of course) was chosen. Then, this unit was split into two halves, according to its middle frame. Both halves and the entire unit are going to be later used in creating long representative.

5 HNM synthesis

In the previous experiments [1], [2], [3], LPCC synthesis was used to produce the output speech. Despite all re-segmentation methods, LPCC synthesis was highly responsible for the low quality of the resulting speech (that can be proved by a copy LPC analysis-synthesis). Therefore, the Harmonic-Noise Model (HNM) which brings much higher quality of the synthesized speech, was applied in our experiments. The principle of HNM is described in [7]. First, the pitch is detected for all the frames of analyzed speech. According to score of pitch detection, the frames are marked voiced or unvoiced. For all the frames, the parameters of the noise model are calculated. Furthermore, the parameters of the harmonic model are calculated only for voiced frames. The LPCC parameterization is still being used for TD, VQ and HMM recognition, because HNM features are not suitable for it. The representatives are modeled only by HNM parameterization.

6 Results

a. Quality of resulting speech: If new re-segmented units are short, the probability of not-existing representative for coding unit is small. Hence, an appropriate representative will be used almost every time. However, the influence of transition noise on resulting speech will not be anywise decreased. In case of too long re-segmented units, a small number of transitions appear in synthetic speech obtained by the concatenation of resulting units' sequence. However, using not large enough DB in our experiments, the probability of non-existing representative is large. Hence, the non-existing coding unit has to be replaced by the most suitable existing one (or created from original representatives (more transition parts will appear there)). The quality of resulting speech is then lower, of course. Therefore, the optimal lengths of re-segmented units should be determined according to the desired quality of resulting speech and the availability of data.
b. Bit rates: When applying the re-segmentation methods on original short units, the number of units in coding sentence is always less than without the 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. Whence it follows that resulting bit-rate depends on the lengths of new re-segmented units. In all our experiments, the original prosody as well as timing are used in the decoder, so that the bit rates refer only to encoding of unit and representative selection. Summary of all our experiments with average bit rates is given in Tab. 1.

7 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. Subjectively we can say that the best quality of resulting speech was obtained with "nse1" version of the re-segmentation (objectively "nse" version, according to Tab. 1). 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 are given there, as well. In our experiments, the prosody and timing (DTW) path were not coded. However according to [6], the sufficient

Table 1. Table of the results.

version	const1	const2	bit rate [bps]	N	l [%]	sp. d. [dB]
rec5	-	-	129	64	100	6.46
nse	4	-	219	7363	81	**5.51**
nse1	7	-	157	13656	53	5.73
nse2	9	-	137	17144	42	5.81
nseNV	4	-	205	9748	65	5.61
nse1NV	7	-	155	16955	48	6.01
sps	8	3	162	19817	50	5.65
sps1	10	4	105	20541	32	6.01
sps2	7	4	139	19229	45	5.80

bit rate [bps]: average bit rates, (without prosody, for one representative).
N: Number of unique segments in training data set.
l [%]: Average relative number of segments in re-segmented label files to rec5 version.
sp. d. [dB]: Spectral distortion between coded version and original speech (only for LPC synthesis).
rec5: Old segmentation, *nse*: Re-segmentation according to middle frames of original units (const1. = min. length of new segments in frames), *nseNV*: Re-segmentation according to middle states of HMMs (const1. = min. length of new segments in frames), *sps*: Re-segmentation according to gravity centers of original units (const1. = min. width of old units in frames, const2. = max. number of original units that can be covered by new unit).

average bit rate for coding prosody is about 200bps. With this value, the difference between original and coded prosody is almost indistinguishable. We can therefore expect the total bit rate for a speaker dependent coder to be of about 370 bps.

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. Picone, and G. R. Doddington. A phonetic vocoder. In *Proc. IEEE ICASSP 89*, pp. 580-583, Glasgow, 1989.
5. M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel. The Boston University radio news corpus. *Technical report*, Boston University, 1995.
6. Y. P. Nakache, P. Gournay, G. Baudoin. Codage de la prosodie pour un codeur de prole a tres bas debit par indexation d'unites de taille variable, CORESA 2000.
7. M. C. Oudot. Etude du modele "Sinusoides and bruit" pour la traitement des signaux de parole, estimation robuste de l'enveloppe spectrale. These de doctorat de l'ENST France, 1996.