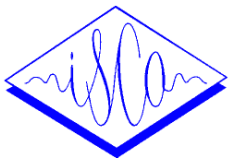

2nd European Conference on
Speech Communication and Technology
EUROSPEECH '91
Genova, Italy, September 24-26, 1991

VARIABLE RATE SPEECH CODING USING PERCEPTIVE THRESHOLDS AND ADAPTIVE VUS DETECTION

[1]P. Meyer, [1]W. Peters, [2]J. Paulus

[1]Philips Kommunikations Industrie AG, Thurn-und-Taxis-Str. 14, D-8500 Nuernberg
[2]Technical University of Aachen, Templergraben 55, D-5100 Aachen

Abstract

This paper describes a 8.4 kBit/s RPE-coder which is modified to form a variable bit rate coder with an average bit rate of 5 kBit/s.

To realize such variable bit rates, the coding scheme must depend on the type of the speech segment to be encoded. Reduction has been achieved for segments that are
- spectrally stationary (repetition of an LPC parameter set),
- non-periodic (omission of an LTP parameter set),
- unvoiced (modeling of the residual as a noise source)
- or contain background noise (simple coding of stationary background noise).

To classify these states we used distance measures, perceptive thresholds and a robust adaptive Voiced-Unvoiced-Silence (VUS) detector.

1. Introduction

With the increasing capacity of digital signal processors, a variety of residual excited linear predictive (RELP) coders, with a bit rate between 8 and 16 kbit/s can be realized for speech transmission tasks like mobile telephony. Obviously these coding techniques should be also usable for speech storage in applications like voice mail or for an answering machine. However in some of these application even a bit rate of 8 kBit/s is to high if only few memory space is available.

The bit rate of fixed rate speech coders depends on the speech parts that are most difficult to code. These parts are within voiced speech segments. In these segments, especially the coding of the residual signal consumes most of the bits. However, in unvoiced parts fewer bits are necessary to code the residual as well as the LPC parameters. Parameters of a long term prediction are not necessary to code at all in these segments. Even more drastically is the bit rate reduction in silence or background segments where only few bits are necessary to code spectral shape and energy of a stationary background.

To realize the advantages of a variable bit rate coder, which selects the bit rate dependent on the type of the speech (or nonspeech) segment, good and robust criteria to classify these types are necessary. Ideally these criteria should be chosen, in a way that no, or only little degration of the speech quality with respect to a fixed rate coder is audible.

2. Description of the RPE-coder

For our investigation we used a regular-pulse excited linear predictive coder as described in [1]. Frame size and the coding of the LPC parameter is similar to the coder chosen for the European mobile radio system [2] called the GSM-Codec in the following. The incoming speech is segmented using a frame length of 20 ms. The sampling frequency is 8 kHz. There is no overlap in successive frames. For an 8th order LPC analysis, the Schur recursion is used to obtain the parcor coefficients. These parcor coefficients are then transformed to log. area ratios (LAR). The LARs are coded and retransformed to parcor coefficients to control the LPC filter.

Other than for the GSM-codec, an open loop, long-term prediction, with an analysis length of 80 samples (10 ms) is used. The delay can vary between 8 and 135 samples. The delay and the gain factor calculated by the LTP analysis is coded and used to control the LTP filter.

For the RPE coding the residual is segmented in 5 ms subframes. These sub-frames are then weighted with a fixed shape, perceptive filter. After the search for the optimal grids, using a grid spacing of 4, the residual is coded using a block maxima coded by 6 bits and 2 bits per sample for the normalized grids. The final bitrate of this coder is 8.4 kbit/s.

3. Repetition of LPC Parameters

3.1 Definition of the spectral distance

A large portion of speech, especially long vowels have steady-state behaviour. In such portions it is useful, not to transmit (or store) LPC parameters but to repeat the old set. To decide if such a set can be repeated, a spectral distance measure and a threshold are necessary. Such a distance measure is the cepstral distance. If c_i and c_j denote the vectors of cepstral coefficients of two different spectra this distance is given by:

$$d_{ij} = \sum_{n=1}^{N} (c_{i,n}-c_{j,n})^2 \qquad (1)$$

For N it is useful to chose the order of the LPC analysis. The cepstral coefficients can be calculated recursively from the LPC parameter as described in [3]. The cepstral distance measure is easy to implement and is known to give better results in speech recognition tasks than the likelihood distance [4].

3.2 Determination of a spectral threshold

To find an appropriate threshold for the cepstral distance we used a procedure known in psychoacoustics as the 2 IFC (Interval Forced Choice) method. We twice coded and decoded whole sentences using the RPE-Coder. In one of the examples we repeated the old LPC-parameter sets if the distance of two successive sets was below a given threshold. The examples were synthesized with random succession. Subjects were asked to listen to the sentences and to decide, which of the tho examples was worse. We started with a large threshold of 1. In this case the subject could allways hear large distortions and find the worse example. If, like in the beginning of the procedure, the answer of the subject was correct, the threshold was lowered and new examples were synthesized and the subjects were asked again to give the position of the worse example. During this first phase, the threshold is lowered until it reaches a point, where no differences are audible. Now the subject is forced to guess the position. If he gives a wrong answer the threshold is rised again. If, from now on, a correct answer is given, the threshold is hold constant. If there is a second correct answer the threshold is lowered again. Each wrong answer rises the threshold. 40 points after the first wrong answer, the session is stopped.

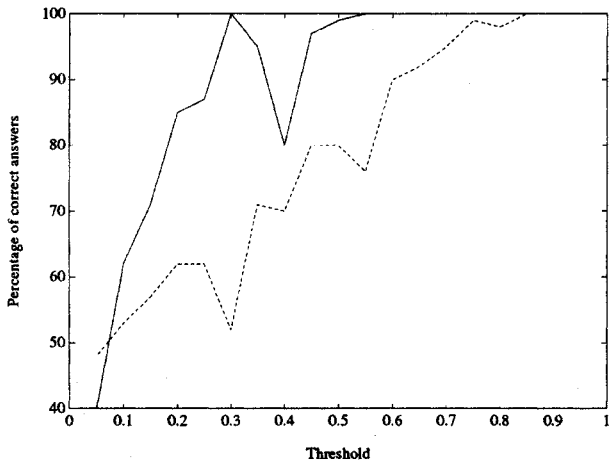

Fig. 1: Percentage of correct answers for the 2IFC method dependent on the spectral threshold. Solid: male speech, dashed: female speech.

For our investigation 16 phonetically balanced German sentences spoken by two female and two male speakers were used. 10 subjects participated the described listening tests. From each session, the number of correct answers for the different thresholds where stored and averaged. Fig. 1 shows the result for the female and male speech. It can be seen, that for high thresholds the modified speech can allways be identified. For a threshold near 0 the correct answer rate is about 50% which shows, that in this case the subjects where only guessing. If we define a correct answer rate of 75% as the switching point between guessing and correct answers we find a reasonable threshold of 0.2 for male and 0.4 for female speaker. This implies that changes below the threshold of 0.2 are not audible. With such a threshold, it was found, that the average frame rate of the coder was about 25 frames per second which is twice as low as for the unchanged RPE-Codec.

4. Omission of Long Term Prediction Parameter

As demonstrated e.g. in [1], a long term predictor or pitch predictor is very useful to reduce the SNR values of coded-speech significantly. However, this advantage is only valid for voiced sounds where a pitch can be detected. In unvoiced parts of the speech and in pauses, it is useless to transmit (or store) LTP parameter sets. To examine, if the LTP is useful or not, we have to look at the attenuation that is achieved by the LTP inverse filter. If E_{in} describes the energy of the residual after the inverse LPC filter and E_{out} the energy of the residual after the LTP filter, this attenuation is described by

$$d_{LTP} = \frac{E_{out}}{E_{in}} \quad . \qquad (2)$$

Using this attenuation we can find a threshold where the LTP parameters can be omitted.

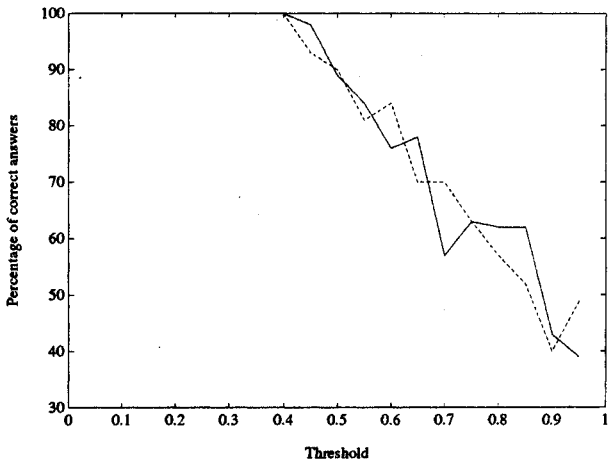

Fig. 2: Percentage of correct answers for the 2 IFC method dependent on the threshold for LTP-attenuation. Solid: male speech, dashed: female speech.

To get a perceptive relevant threshold we used again the 2 IFC method described in Section 3. All tests were made using the same speech corpus. Fig. 2 shows the result. It can be seen that above a threshold of 0.8 it was impossible to hear any differences between synthesized examples with and without LTP filter. For a threshold of 0.8 about half of the LTP parameter sets or about 500 bits/s can be saved without perceptive significant changes.

5. Voiced-Unvoiced-Silence Detection

5.1 Basic Detector

A prerequisite for a robust VUS detector is a good parameter vector. Good elements for such a vector are described in [5,6]. We decided to select the following elements: the logarithm of the energy of the lowpass filtered signal, the Zero crossing rate, the first parcor coefficient and the logarithm of the LTP attenuation (equ. 2). All these elements have different probability density functions for the different classes (voiced unvoiced and silence).

To estimate these density functions we hand-labeled the speech corpus described above. This labeling was done interactively using time-signal plots, spectra and spectrograms. Furthermore, it was possible to listen to labeled speech segments. From the speech corpus we obtained 3290 Frames, 1988 voiced, 628 unvoiced and 674 silence.

If we assume the propability density functions of the elements to be statistically independent, we get the following joined propability density function:

$$p(x|\omega_k) = \prod_{i=1}^{4} p_i(x_i, \omega_k) \qquad (3)$$

where $x = (x_1, x_2, x_3, x_4)^t$ is the parameter vector and ω_k (k =1,2,3) the class.

A vector x should be assigned to a class j if

$$p(x|\omega_j) \geq p(x|\omega_k) \quad \text{for all } k \neq j \qquad (4)$$

The propability function $p_i(x_i,\omega_k)$ can be approximated by a histogram $\hat{p}_i(n,\omega_k)$ with the bin number n. The histograms were calculated using the labeled speech corpus. For each histogram approximation, we used 24 bins.

To test the described VUS detector, we used a test set which consisted on 30 phonetically balanced German sentences, 15 spoken by a male, 15 spoken by a female speaker. This test set was hand-labeled in the same way as the training set and consisted on 3561 voiced, 1326 unvoiced and 1224 silence vectors.

The overall recognition rate was 8.42 % for the training set and 10.31 % for the test set. This high rate is mainly due to the misclassification of unvoiced and silence patterns. Most of these errors are in transition areas like voiced to silence and unvoiced to silence. Such regions are called *uncertain* regions by Rabiner et al [6]. In these regions it is allways hard to classify, even for a human labeler looking at spectra and time plots. The error rate for the classification of voiced frames was 4.8% and 3.8% respectively. This lower rate is more important for our task, because a misclassification of a voiced frame as unvoiced would have strong consequences for a coder due to lower speech quality. Misclassification of unvoiced or silence frames will only rise the bit rate.

5.2 Background Adaptation

A serious disadvantage of a VUS Detector as described above is its sensitivity to different background noise. If background noise occurs, the error rate rises dependent on the signal to noise ratio.

If we assume that only the background noise is changing, it is necessary to retrain the silence propability density functions. However, in most applications it is impossible to collect about 600 silence patterns, as it was done in our training procedure, before starting the coding. So we decided to take the trained silence density functions and to adapt them recursively to new background patterns.

If x is a new background vector whose element i lies in the bin j of the histogram approximation of the background propability density function, this adaptation can be done in the following way:

$$\hat{p}_{i,new}(n, \omega_{BG}) = \alpha \hat{p}_{i,old}(n, \omega_{BG}) \text{ for } n \neq j \qquad (5)$$

and

$$\hat{p}_{i,new}(j, \omega_{BG}) = \alpha \hat{p}_{i,old}(j, \omega_{BG}) + 1-\alpha \qquad (6)$$

with a constant α = 0.96.

We tested the background adaption using white noise and car noise that we added to the test set. The errors of the VUS detector were less than 5% for voiced segments even for segmental signal to noise ratios of 10 dB.

To get background patterns for an adaptation, the distance calculation described in Section 3 can be used. If the distances of successive frames remain below a threshhold of 0.2 for more than 200 ms we can assume a non speech interval that can be used for background adaptation.

6. Type dependent residual signal coding and perceptive evaluation

The described VUS detector is used to classify incoming speech frames of 10 ms length. If the segment is voiced the residual is coded like in the 8.4 kBit/s RPE coder, using 10 bits for the LTP coding (3 bits gain, 7 bits delay) and 58 bits for the RPE coding (2* 2 bits for the grid position, 6 bits for the block maxima, 20 bits for the grids).

If the speech frame is classified as unvoiced, the energies of two 5 ms subframes are calculated and coded logarithmically using 12 bits. The LTP parameters are omitted.

If the speech frame is classified as silence (background), the energy of the 10 ms frame is calculated and coded logarithmically using 6 bits. Also here the LTP parameters are omitted.
This type dependent residual coding saves 46 and 52 bits for each non voiced frame. For our used speech corpus this results in a 2137 saved bits per second.

For synthesis of the speech, the unvoiced and silence (background) frames are produced by decoding the energy values and synthesizing a residual signal of the same energy using a simple noise generator. However, it might be asked if such a crude coding and decoding of the residual does not produce severe distortions for the synthesized speech.

To answer this question we made the following listening test. Subjects listened three times to a sentence, coded and decoded by the original 8.4 kBit version of the coder. Then the subjects heard in random succession another unmodified resynthesized example and an example were all nonvoiced segments were coded and decoded as described above. The subject was forced to decide which of the two examples was worse (modified). After the decision the subject was asked if the modified example was clearly different, slightly different or if he could hear no difference at all.

37 subjects participated the test. 18 different sentences of the speech corpus were used which had different percentage of nonvoiced segments. The percentage was between 25 and 60. Each person listened to 15 different sentences.

The results depend on the percentage of nonvoiced speech segments. For a percentage between 25 and 30 the misclassification of the modified speech examples was 50%. This indicates that no differences were audible in these examples. 70% of the subjects claimed to hear no differences and only a small percentage heared small or large differences. For examples with a larger percentage of nonvoiced segments the misclassification rate was between 40% and 30%. This indicates, that for 60% - 80% of these examples there were no audible differences. For percentages of nonvoiced speech of over 50%, which is normally rare in fluent speech, already 30% of the subjects claimed to hear clear differences while only 30% heared no differences. However for normal speech, which has a percentage of nonvoiced speech segments of about 40, the modified coding scheme is only slightly worse than the unmodified one.

6. Summary and Discussion

In this paper we have presented a 8.4 kBit/s RPE coder which has been modified to form a variable bit rate coder of an average bit rate of 5-6 kBit/s. Three methods have been used to reduce the bit rate.

- Repetition of LPC parameter sets can be done without audible differences if successive LPC parameter sets have a cepstral distance of less than 0.2. Such repetition of LPC parameters can save approximately 800 bits per second.

- Omission of LTP parameter sets without audible differences can be done if the attenuation of the LTP analysis filter is higher than 0.8. Approximately 500 bits per second can be saved by omission of the LTP parameters.

- Synthesis of the residual signal using a noise generator, if the speech segment is unvoiced. This has only small effects to the quality of the synthesized speech. More than 2000 bits per second can be saved using such type dependent coding of the residual. Further reduction can be achieved if the segment is background noise and constant over several frames.

For the classification of voiced, unvoiced and silence (background) intervals we developed a robust VUS detector. This detector can easily be adapted to the actual background noise. Eventhought the detector has an error rate of more than 4% for the detection of voiced segments, there is no severe distortion audible if it is used for the varible bit rate coder.

The advantage of the shown coder is its low complexity. It can easily be implemented on a standard DSP and runs in real time. Even though we have used a RPE coder for our varible bit rate coder, the same reduction methods should be valid for other residual coders which depend on LPC, an LTP and a residual coding part, like a code excited or multipulse excited coder.

7. References

1 P. Kroon, E.F. Deprettere and R.J. Sluyter, "Regular-Pulse Excitation - A Novel Approach to Effective and Efficient Multipulse Coding of Speech," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-34, pp.1054-1063, Oct.1986.

2 P. Vary, K. Hellwig, R. Hofmann, R.J. Sluyter, C. Galand and M. Rosso, "Speech Codec for the European Mobile Radio System," in *Proc. IEEE int. Conf. Axoust., Speech, Signal Processing,* Mar. 1988, pp. 227-230.

3 J.D. Markel and A.H. Gray, Jr. *Linear Prediction of Speech* Springer Verlag, 1976.

4 F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-23, pp. 67-72 1975.

5 B.S. Atal and L.R. Rabiner, "A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition," in *Proc. IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-24, No.3, June 1976.

6 L.R. Rabiner, C.E. Schmidt and B.S. Atal, "Evaluation of a Statistical Approach to Voice-Unvoiced-Silence Analysis for Telphone-Quality Speech," in *The Bell System Technical Journal* Vol.56, No.3, March 1977.