Variable Bitrate Wideband Speech Coding Using Perceptually Motivated Thresholds

Jürgen W. Paulus

Institute of Communication Systems and Data Processing (IND)
RWTH Aachen, University of Technology, D-52056 Aachen, Germany
phone: +49.241.806961, fax: +49.241.8888186, joschi@ind.rwth-aachen.de

Abstract

This paper proposes a variable bitrate wideband CELP coding scheme (50-7000 Hz) with a one-way coder-decoder delay of 25 ms only. The bitrate varies between 6.4 kbit/s and 14.8 kbit/s with a mean of 11.2 kbit/s at a voice activity of 95%. The variation of the bitrate is based on a voiced-unvoiced-silence classification of the speech frame to be encoded. For voiced frames a perceptually based focussed LTP analysis is used, and for unvoiced or silent frames the LTP filter is omitted. Additionally, in adjacent frames with small spectral changes LPC parameters are repeated using an adaptive LPC codebook. For the classification of the different modes a cepstral distance measure, and a new adaptive detector is used. By informal listening tests the speech quality was rated higher than the CCITT G.722 wideband codec operating at 48 kbit/s.

1. Introduction

The average bitrates of analysis-by-synthesis CELP coders can be reduced significantly if they are not restricted to a fixed bitrate (see for example [1, 2]). Starting with a CELP speech codec operating at a fixed bitrate of 14.4 kbit/s, we will describe in the following our methods to obtain a variable bitrate codec.

2. Fixed Bitrate Codec

Recently, we presented a split-band encoding scheme using 2 unequal subbands, i.e. 0-6kHz and 6-7kHz [6]. This approach was motivated by the experimental evaluation of the instantaneous signal bandwidth of speech frames. During voiced parts of a speech signal, most of the signal energy is placed in the lower subband. Therefore it is not necessary to encode the higher subband of voiced frames. Transform coding techniques behave in a similar way in that they allocate more bits to lower frequency components than to higher frequency components. For that reason simulations were performed to find out the actual cut-off frequency necessary to encode the current frame without loss of perceptual speech quality. By applying a frame size of 10ms we found that almost 40% of the frames could be encoded using a bandwidth of 6kHz only. The full bandwidth was selected only during unvoiced parts of the speech signal. Furthermore, it turned out by experiment that during unvoiced parts it is sufficient to add some noise like spectral components above 6kHz to obtain the perceptual speech quality of a 7kHz speech signal. With the mechanism of speech production in mind, spectral components above approximately 4kHz are almost exclusively due to fricative speech sounds [8]. Visual inspection of the waveforms of speech components within the frequency ranges 5-6kHz and 6-7kHz shows that both of these signals exhibit a similar distribution of energy along the time axis for a given speech sound. This observation can be used successfully to approximate the missing speech components in the frequency range of 6-7kHz very efficiently without transmitting any side-information as shown below. The wideband speech signal is encoded using only the spectral bandwidth from 0-6kHz and the missing components above 6kHz are replaced at the receiver by interpolating the lower subband signal from 12kHz to 16kHz using an interpolation filter with cut-off frequency 7kHz which violates the interpolation rules. The enhancement of the perceptual speech quality due to this is surprisingly high. A similar approach was already proposed in [3] in the context of wideband ADPCM.

For encoding the decimated signal a CELP scheme is applied. In our realization we use a 14th order LPC predictor, updated every 10ms (120 samples). The spectral parameters are encoded using 44 bits by interframe moving average prediction and split vector quantization of the line spectral frequencies (LSF) parameters. Every 5ms, a long-term-prediction (LTP) is carried out in a closed loop pitch search, i.e. using an adaptive codebook filled with previous computed excitation signals. The minimum pitch lag searched is half of the subframe length, i.e. 30 samples. Additionally, a fractional pitch is used with a resolution of 1/2 sample, resulting in 8 bits for indexing the adaptive codebook. The pitch gain is nonuniformly scalar quantized with 4 bits. Each 2.5ms (30 samples), a 15 bit algebraic ternary sparse codebook is used as described for example by Laflamme *et al.* [7]. An innovation vector contains 4 nonzero pulses (+1,-1,+1,+/-1). The global sign bit to change all pulse signs simultaneously is included in the codebook size. A fixed gain predictor comparable to the one in G.728 standard is used for quantizing the codebook gain. The residue of the gain predictor is nonuniformly scalar quantized with 4 bits. Different perceptual weighting filters are used for the adaptive and algebraic codebook search. During the adaptive codebook search, a bandwidth expansion

This work has been supported by the Research Center of Deutsche Telekom. The author would like to thank Prof. P. Vary, H.J. Mundt and T. Krug for inspiring discussions.

factor of γ=0.4 is used, and during the algebraic codebook search γ=0.9 is used. This was found to give better results compared with a fixed weighting filter.

3. Variable Bitrate Codec

To realize the advantage of a variable bitrate coder, which selects the bitrate depending on the type of the speech frame, an adaptive voiced-unvoiced-silence detector was developed. This detector applies adaptive thresholds for the energy, zerocrossing rate, first reflection coefficient, and a new criterion d_N especially for voiced-unvoiced decisions:

$$d_N = \frac{\sum_{n=2}^{N-1} \Delta\Psi(n)\,|s(n)-s(n-1)|}{\sqrt{E_N}}$$

with

$$\begin{aligned} \Delta\Psi(n) &= \frac{|\Psi(n)-\Psi(n-1)|}{2}, \ \in \{0,1\} \\ \Psi(n) &= \frac{s(n)-s(n-1)}{|s(n)-s(n-1)|}, \ \in \{-1,1\}. \end{aligned}$$

and $s(n)$ the speech samples. All different thresholds depend on the history of the classification.

In the following, the mode dependent strategy for bitrate reduction is presented.

(1) Previous quantized LPC parameter sets are repeated if spectral changes are small. To measure these changes, a cepstral distance measure is used [4]. Informal subjective listening tests were performed to obtain a threshold for the cepstral distance measure, such that a repetition of previous sets cannot be noticed. A scheme presented in [4] is extended to the evaluation of the distance measure of the last P frames. This can be interpreted as an adaptive codebook for the LPC parameter quantization, similar to the one recently proposed in [5]. Using a codebook with P=3 parameter set entries, 30% of the sets can be replaced by previously quantized sets without loss of subjective speech quality.

(2) In silent frames the prediction gain of the LPC analysis filter is relatively low compared to voiced frames. Also the detailed structure of the spectral shape is not as important as in voiced frames. In this case the order of the LPC predictor can be reduced. Informal listening tests indicated, that a filter order of 6 is sufficient.

(3) Unvoiced or silent frames do not contain any pitch periodicity. In these frames the LTP analysis and the transmission of the corresponding parameters can be omitted. At the receiver the LTP gain is set to zero during these frames.

(4) During voiced frames, a perceptually motivated combination of open-loop and closed-loop LTP analysis is performed. Each 10 ms speech frame is divided into 2 subframes of 5 ms duration. For each subframe an open-loop pitch estimate is calculated using a weighted correlation measure to avoid multiples of the pitch period. Thus a smoothed estimate of the pitch contour is obtained. If there are only small variations within these pitch estimates, in the first subframe a focussed closed-loop adaptive codebook search is performed around the first open-loop estimate and in the second subframe a restricted search is performed around the pitch lag of the closed-loop analysis of the first subframe. This procedure results in a delta coding scheme leading to 8+3=11 bits for coding the pitch lags of a 10 ms frame.

Otherwise in case of strong variations a full closed-loop search is performed in each subframe resulting in 8+8=16 bits for coding the pitch lags. The decision whether a restricted search is performed or not is based on a perceptually motivated threshold. During low-frequency voiced speech a larger variation of the pitch contour is perceptually acceptable than for high-frequency voiced speech. Therefore, the threshold depends on the relative change of the open-loop pitch estimate in the actual subframe in comparison to the estimate of the previous subframe. Informal listening tests indicated a threshold for the difference of the open-loop pitch estimates of about 5%. Using this threshold 80% of the voiced frames can be encoded using a combined open-loop/closed-loop LTP applying a restricted search.

4. Conclusion

Introducing the above presented features a variable bitrate 6.4 kbit/s $\leq B \leq$ 14.8 kbit/s is obtained with an average bitrate of $\bar{B}$=11.2 kbit/s for a voice activity of 95%. The speech quality of the variable bitrate codec was rated in informal subjective listening tests to be better than the CCITT G.722 wideband codec operating at 48 kbit/s.

5. References

[1] E. Paksoy, K. Srinivasan, and A. Gersho, "Variable Bit-Rate CELP Coding of Speech with Phonetic Classification," *European Transactions on Telecommunications*, Vol.5, No.5, pp. 591-601, October 1994.

[2] L. Cellario, and D. Sereno, "CELP Coding at Variable Rate," *European Transactions on Telecommunications*, Vol.5, No.5, pp. 603-613, October 1994

[3] M. Dietrich, "Performance and Implementation of a Robust ADPCM Algorithm for Wideband Speech Coding with 64 kbit/s," *Proc. Int. Zürich Seminar on Digital Communications*, Zürich, Switzerland, March 6-8, 1984.

[4] P. Meyer, W. Peters, and J. Paulus, "Variable Rate Speech Coding Using Perceptive Thresholds and Adaptive VUS Detection," Proc. EUROSPEECH, Genua, Italy, pp.809-812, 1991

[5] C.S. Xydeas and K.K.M. So, "A Long History Quantization Approach to Low Bit Rate Speech Coding," *Proc. EUSIPCO*, Brussels, Belgium, pp. 479-482, 1992.

[6] J.W. Paulus, "Wideband Speech Coding with 1-2 bit per Sample," in *Proc. Int. Conf. Digital Signal Processing*, Limassol, Cyprus, June 26-28, 1995.

[7] C. Laflamme, R. Salami, and J.P. Adoul, "16 kbps Wideband Speech Coding Technique Based on Algebraic CELP," in *Proc. Int. Conf. Acoust., Speech, Signal Processing, ICASSP*, Toronto, Kanada, pp.13-16, 1991.

[8] J.L. Flanagan, *Speech Analysis, Synthesis, and Perception*, Springer Verlag Berlin Heidelberg New York, 2nd ed.,1972.