TOLL QUALITY VARIABLE-RATE SPEECH CODEC

Pasi Ojala

Speech and Audio Systems Laboratory,
Nokia Research Center, Tampere, Finland
pasi.ojala@research.nokia.fi

ABSTRACT

This paper presents a source controlled variable-rate CELP type speech codec. First, a voice activity detection block distinguishes active speech frames from silence and background noise. The active speech is further classified into voiced and unvoiced frames. The voiced frames have variable bit-rate pitch-lag quantization based on the characteristics of the speech, whereas the unvoiced frames are coded without pitch information. A variable bit-rate fixed codebook excitation with a variable number of excitation pulses is determined for each speech frame. The performance of the linear analysis part of the codec as well as the input speech characteristics determine the excitation bit-rate. The average bit-rate of the codec is around 7.0 kbit/s for active speech, and the overall bit-rate ranges from 0 to 7.85 kbit/s. The described variable-rate codec produces toll quality speech equal to that of the 32 kbit/s ADPCM (G.726) standard.

1. INTRODUCTION

Speech signals have bursty and time varying characteristics consisting of clearly voiced, unvoiced and silent parts. Considering these characteristics, the number of parameters and bits needed to optimally encode the speech signal varies in consecutive analysis frames. Hence, the number of coding parameters, as well as the quantization, vary while the quality of decoded speech remains constant. Therefore, the average bit-rate is lower than that of a constant bit-rate codec while sustaining the same subjective speech quality.

The most common method to introduce variable-rate operation into a coding algorithm is the Voice Activity Detection (VAD) function classifying the speech into active and inactive parts [1]. When VAD detects inactive speech, the encoder produces comfort noise parameters or an empty frame. In addition to the VAD-function, a widely used method is to classify the speech signal into voiced, unvoiced, and silence parts [2]. Voiced frames are coded with a higher bit-rate than unvoiced and silence frames. Another method is a closed-loop quality evaluation to select the bit-rate [3]. If the quality of the synthesised speech is not good enough, coding functions will be repeated with a higher bit-rate. Variable-rate codecs in IS-96 CDMA system make the rate decision based on sub-band energies of the input signal [4]. Adaptive thresholds on the energy select the bit-rate of the codec. This paper presents source controlled criteria for a variable-rate codec to determine the bit-rate and the number of coding parameters. Rate criteria select the modes so that the subjective quality of the decoded speech is the same as the quality produced at the highest bit-rate. Hence, the codec maintains the high quality despite the lower average bit-rate.

The presented variable-rate speech encoder applies a VAD function to detect the active speech from the input signal. The encoder generates comfort noise parameters for inactive speech frames. After the Linear Predictive Coding (LPC) and spectral content analysis a voicing criterion classifies the speech into voiced and unvoiced frames. Consequently, in the case of unvoiced speech, the Long Term Prediction (LTP) parameters are unnecessary. On the other hand, in the case of voiced frames the bit-rate of LTP parameters is determined based on the source signal. Since the characteristics of speech are partly voiced, unvoiced, and noisy, the variable-rate encoder determines the accuracy of the non-integer closed-loop lag calculation based on the knowledge about the spectral content of the speech frame. Voiced frames with multiple formants in the spectral envelope need an accurate fractional closed-loop lag search algorithm and quantization to maintain good speech quality. On the other hand, speech frames with a flat spectrum do not contain any formants. Hence, a less accurate lag quantization is good enough to maintain the high quality of speech. With a reliable lag search algorithm selection the quality of the decoded speech will remain high despite the lowering of the average bit-rate.

Finally, the excitation criterion determines the bit-rate of the pulse excitation based on the performance of the linear prediction analysis filters, as well as on the input signal characteristics. The selection of the number of pulses for consecutive speech frames is based on adaptive thresholds of the performance of the short-term and long-term analysis filters. In addition, the spectral content of the input signal controls the excitation bit-rate determination.

2. VARIABLE-RATE ENCODER

The LPC analysis gives indication about the speech characteristics. The spectral content is analysed using the two first reflection coefficients. These coefficients indicate the characteristics of the spectrum of the analysis frame [5]. We can distinguish between a low-pass type of spectrum and a flat spectrum with uniform frequency distribution.

Voiced/unvoiced classification

The speech frame is divided into voiced and unvoiced frames according to the speech characteristics and the coding

parameters. Indication for this is obtained from the LTP gain, the zero-crossing rate, and the smoothness of the LTP lag track.

The classifier sets the voiced flag when LTP gain is greater than an adaptive threshold. The use of an adaptive threshold is quite speaker independent and it makes the classification more robust against background noise. The adaptive threshold b_{th} is computed by low-pass filtering the scaled LTP gain b

$$b_{th,i} = (1-\alpha)K_b b + \alpha b_{th,i-1}$$

The scaling factor is $K_b = 0.15$. The threshold is limited into proper intervals. The considerably large decaying factor is chosen to slowly adapt the thresholds over voiced and unvoiced periods, because the background noise is typically relatively stationary. The analysis frame with LTP gain below the adaptive threshold is considered as unvoiced. As an additional criterion, a high zero-crossing rate indicates unvoiced speech.

The LTP lag varies about 1 % / ms during voiced speech [1]. Therefore, we can adjust a threshold for the LTP lag variations to classify speech frames with dramatic lag changes as unvoiced. First, the pitch lag values of each frame are low-pass filtered to get a smooth track d_{th} of LTP lag.

$$d_{th,i} = (1-\alpha)d_{opt} + \alpha d_{th,i-1}$$

The open-loop LTP lag of the current frame is compared with the smoothed track. If the lag value differs significantly from the tracked value, the frame is considered as unvoiced.

Figure 1 presents an example of voiced/unvoiced classification results. Typically, around 12 % of active speech frames are classified as unvoiced with the criteria described above.

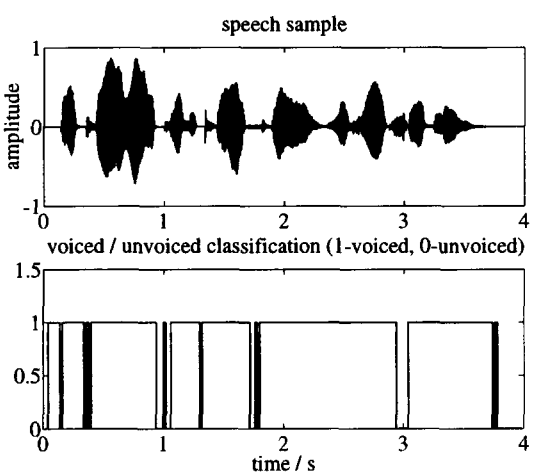

Figure 1. Voiced unvoiced classification based on zero crossing rate, LTP-gain, and smoothed LTP lag.

Closed-loop LTP

The closed-loop pitch-lag search is performed on a subframe basis. Typically, the performance of the LTP is improved when the pitch-lag is searched with a high resolution, i.e., with subsample accuracy [6]. The high resolution LTP lag consists of performing a fractional pitch lag search and computing the adaptive codevector by interpolating the past excitation at the selected fractional lag with the selected fractional accuracy. The accuracy of the search algorithm has three possibilities: 1/3, 1/2 and 1/1 fraction of the lag. Each LTP lag with different fraction accuracy is coded with a different number of bits. Hence, the LTP parameters have a variable bit-rate. To reduce the bit-rate, the LTP lag of the first subframe is scalar quantized, and the following three subframes have differential scalar coding with respect to the previous lag.

The decision of the LTP lag accuracy is based on the spectral content of the frame, as well as on the open-loop LTP gain. As the frames containing several formants in the spectrum typically contain voiced speech, the codec selects the most accurate 1/3 fractional search for those frames. Consequently, the 1/2 and 1/1 fractions are chosen for the frames with less clear formant structure. The two first reflection coefficients give an indication to select the highest resolution when a low-pass type of spectrum is detected. In addition, a high open-loop LTP gain indicates voiced speech, for which a high resolution LTP lag is required.

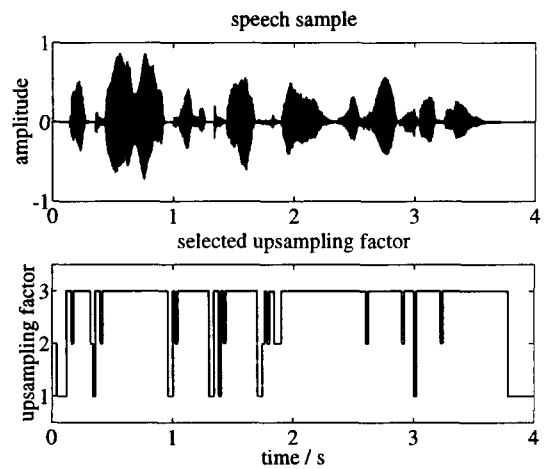

Figure 2. Upsampling factor selected based on the LPC model formant structure, the spectral content, and the open-loop LTP gain.

Variable excitation

An algebraic codebook excited codec consumes a lot of bits for quantization of the excitation signal. A set of excitation pulses is fitted to the perceptually weighted target signal. A fixed-rate speech codec calculates a fixed number of pulses for each subframe regardless of the source signal itself. However, since the characteristics of speech are partly voiced, unvoiced, and noisy, the number of pulses of the codebooks may differ as a function of the input signal. In this case, the number of pulses and the quantization accuracy of the position of the pulses is changed according to the input signal. The variable-rate excitation is implemented by employing different bit-rate excitation codebooks. The excitation criterion selects the codebook.

The excitation generation is done on a subframe basis. Each codebook covers the subframe with a different number of non-zero pulses. Currently, the variable-rate codec employs three different bit-rate codebooks: a codebook with five, four, and three pulses.

The variable-rate excitation is realized by choosing different codebooks for consecutive frames based on an adaptive criterion. The method is to set an adaptive threshold according to the LTP residual energy G and change the codebook when the energy crosses the threshold. The adaptive threshold is calculated as follows

$$G_{th,i} = (1-\alpha)(G+\Delta G) + \alpha G_{th,i-1},$$

where $G_{th,i}$ is the threshold value and the scale factor is $\Delta G = -1.0\,dB$. When the calculated LTP residual energy is above the adaptive threshold the encoder applies a codebook with more pulses and more accurate quantization. A low energy residual signal results in less accurate excitation with fewer pulses. When more than two different codebooks are available another adaptive threshold is formed by changing the scale factor of the threshold equation.

Figure 3 presents the adaptive thresholds (dash dotted). The most accurate codebook is selected when the LTP residual signal energy exceeds the highest threshold. On the other hand, the three-pulse codebook is selected when the residual signal energy falls below the lowest threshold level.

The energy distribution of the signal on the frequency axis is an important factor when selecting the excitation codebooks. When the signal energy is concentrated at the lower end of the frequency band the signal can be considered containing voiced speech. A voiced signal needs accurate excitation for the best possible quality. Thus, the five-pulse codebook is preferred based on perceptual selection. The bit-rate of the codec increases slightly because of the perceptual codebook selection, but on the other hand, the quality of the coded speech is improved.

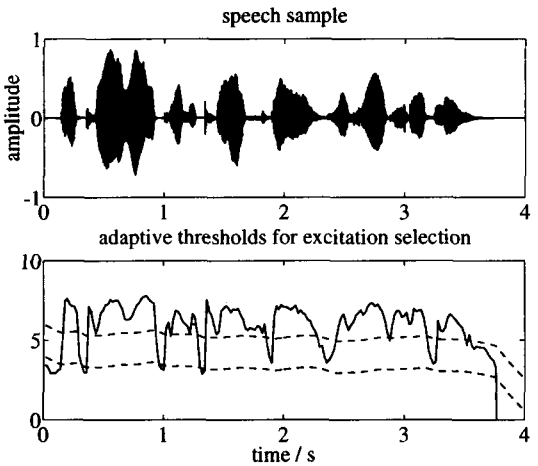

Figure 3. Thresholds for codebook selection.

Gain quantization

Typically, the LTP gain and the excitation gain are vector quantized. However, the variable-rate codec processes unvoiced frames without using LTP parameters. Therefore, only the excitation gain needs to be quantized for unvoiced frames. In the case of voiced speech, the LTP and excitation gains are vector quantized. In addition, to obtain high speech quality, the gains are vector quantized with the high bit-rate codebook when the highest bit-rate excitation codebook is selected.

3. BIT-RATE

Figure 4 presents the schematic diagram of the variable-rate encoder containing the bit-rate control blocks. Since every control block operates independently of each other, the codec contains as many as 12 modes and an additional function generating comfort noise parameters in every tenth silence frame.

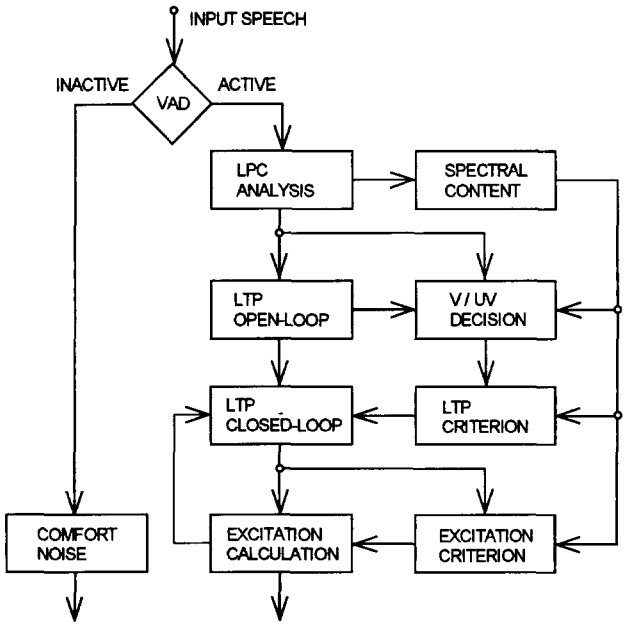

Figure 4. Source controlled variable-rate speech encoder

In a typical telephone conversation with speech activity around 50%, the VAD function brings the average bit-rate near 3.0 kbit/s. Figure 5 presents the bit-rate of the codec for a short speech sample. The average bit-rate is around 7.0 kbit/s for active speech.

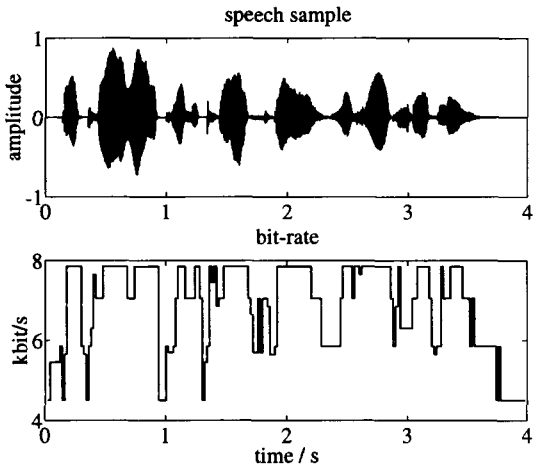

Figure 5. Bit-rate of the variable-rate speech codec

Table 1 presents the bit-allocation of the variable-rate codec for each coding parameter. Naturally, the LTP, excitation, and gain parameters are coded on a subframe basis.

Table 1. *Bit allocation of the variable-rate speech codec. The first two modes are comfort noise frames. Modes 3-5 are unvoiced, and modes 6-14 are voiced frames.*

Mode	LPC	LTP	Exc.	Gain	Total bits	Total kbit/s
1	–	–	–	–	0	0.00
2	26	–	–	5	31	1.55
3	26	–	44	20	90	4.50
4	26	–	68	20	114	5.70
5	26	–	80	20	126	6.30
6	26	15	44	24	109	5.45
7	26	19	44	24	113	5.65
8	26	23	44	24	117	5.85
9	26	15	68	24	133	6.65
10	26	19	68	24	137	6.85
11	26	23	68	24	141	7.05
12	26	15	80	28	149	7.45
13	26	19	80	28	153	7.65
14	26	23	80	28	157	7.85

Figure 6 presents the bit-rate distribution of a ten-minute-long speech sample with 40% speech activity. Only the percentage of active speech frames is presented in the chart. The average bit-rate for this sample is around 2.8 kbit/s.

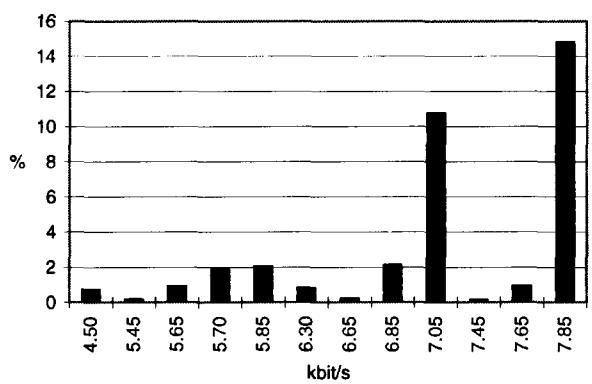

Figure 6. The percentage of selected modes for clean speech.

Mode information

The independent source controlled rate selection criteria result in a high number of different rates. The variable-rate codec has 14 different modes. Naturally, the decoder needs to receive the information about the modes. Therefore, the encoder transmits mode bits as side information. The options for the LTP calculation are three different fractional pitch-lag search algorithms and the decision whether the analysis frame is unvoiced. Hence, two bits are needed for it. The excitation has three options. In addition, the decision about the activity of the frame could be connected to excitation selection. Thus, two mode bits are needed for it. In total, the mode bits increase the average bit-rate by 0.2 kbit/s.

4. SPEECH QUALITY

A subjective Mean Opinion Score (MOS) listening test was performed to evaluate the speech quality of the variable-rate codec. The test contained modified IRS filtered male and female speech samples. The test was carried out by ten experts in speech coding. Figure 7 presents the MOS test results of GSM enhanced full rate (12.2 kbit/s), variable-rate (7.85-0.0 kbit/s), G.726 (32.0 kbit/s ADPCM), CDMA IS-96 (8.0-0.8 kbit/s), GSM half rate (5.6 kbit/s), and GSM full rate (13.0 kbit/s) codecs.

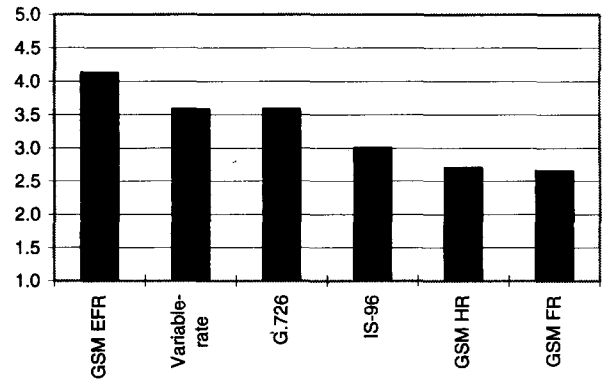

Figure 7. Results of the MOS test.

5. CONCLUSION

A high quality source controlled variable-rate speech codec with an average bit-rate 7.0 kbit/s for active speech was presented in this paper. The subjective tests showed that the codec provides quality equal to the G.726 32 kbit/s standard.

REFERENCES

[1] Kondoz A. M. Digital Speech (Coding for Low Bit Rate Communication Systems). John Wiley & Sons, New York. 1994.

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

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

[4] Cox, R. V. Speech Coding Standards. In Kleijn, W. B. and Paliwal, K. K., editors, *Speech Coding and Synthesis*, pp. 49-78. Elsevier Science B.V., Amsterdam, 1995.

[5] Massaloux, D. and Proust, S. *Spectral Shaping in the Proposed ITU-T 8 kb/s Speech Coding Standard.* Proc. of IEEE Workshop on Speech Coding for Telecommunications. Annapolis, MD, Sept. 1995. pp. 9-10.

[6] Kroon, P. and Atal, B. S. *Pitch Predictors with High Temporal Resolution.* Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing. Alberquque, NM, April. 1990. pp. 661-664.