CELP CODING BASED ON SIGNAL CLASSIFICATION USING THE DYADIC WAVELET TRANSFORM

Joachim Stegmann, Gerhard Schröder, Kyrill A. Fischer

Deutsche Telekom AG, Technologiezentrum
Am Kavalleriesand 3, 64295 Darmstadt, Germany

e-mail: stegmann@fz.telekom.de

ABSTRACT

This paper describes a CELP speech-coding algorithm which makes use of a specific signal classifier especially designed for this purpose. The classification method is based on the Dyadic Wavelet Transform (D_yWT) and has proved to be superior to common classifiers that use the open-loop long-term prediction gain for mode selection. The classifier's output is used for the control of several coder parameters, such as the choice of the subframe length and the selection of the synthesis model and the corresponding codebooks. We designed a fully quantised coder operating at a fixed bit rate of 4 kbit/s with a 20-ms frame. The proposed coder improves the weighted segmental signal-to-noise ratio (WSegSNR) by 2.3 dB on the average in comparison with a conventional CELP coder, thereby achieving high speech quality.

1 INTRODUCTION

The ITU-T SG 15 is planning to standardise a 4-kbit/s speech coding algorithm that should perform not worse than the reference codec G.726 (ADPCM) at 32 kbit/s under a variety of conditions, e.g. clean speech quality, tandemings, presence of background noise and channel errors [1]. The main applications for this codec will be very low-rate PSTN visual telephony, personal communications, and mobile-telephony satellite systems.

Code-Excited Linear Prediction (CELP) based coding schemes with a fixed combination of codebooks have been successfully applied to provide wireline-quality speech at bit rates down to 8 kbit/s [2]. However, it is difficult to maintain high speech quality at 4 kbit/s with a relatively short frame length of 20 ms. Therefore, some authors [3,4,5] have proposed time-variant signal-dependent codebook combinations based on classification to match the temporal input signal characteristics.

In many speech coding schemes, classification is based on the open-loop long-term prediction gain because it is a simple estimate of the periodicity in a speech frame. More sophisticated classification algorithms [6] additionally use different parameters, like reflexion coefficients or zero-crossing rate, to get a rough estimate of the distribution of signal energy in the frequency domain. However, all these parameters are based on temporal averages over a window of fixed length. Thus, the time-frequency resolution depends on the choice of the window length and cannot be matched to the temporal input signal characteristics.

For example, during transition segments like voicing onsets both the ear's sensitivity to spectral errors and the prediction gain are small. Exact timing of excitation pulses is, however, crucial to maintain high-quality coded speech. As these transients mostly cover only a short time interval, the classifier should provide high temporal resolution in this case. On the other hand, during stationary and periodic frames, a longer analysis window — with a higher frequency resolution — can be more efficient to extract the important signal features. Further disadvantages of common classification methods are either degradation in the presence of background noise or a large computational complexity.

The Wavelet Transform is an analysis method that offers higher flexibility in adapting time- and frequency resolution to the input signal [7]. This flexibility is achieved by correlating the input signal with basis functions that are scaled (dilated or contracted) and shifted versions of a so-called *mother wavelet* which itself is a bandpass function. As there are only small constraints on the mother wavelet, it is possible to choose wavelets with similar time shapes as typical events in the input signal. Based on the theory described in [8], the use of cubic spline wavelets may be very well suited to detect the locations of the pitch pulses in a speech signal. Since it is often sufficient in speech analysis to consider the signal over a few dyadic scales only, the computationally efficient filter-bank implementations of the D_yWT can be applied [8].

2 OVERVIEW OF THE CODING SCHEME

This paper describes a 4-kbit/s CELP coder that is controlled by signal classification based on the D_yWT. The algorithm uses a frame length of 20 ms and a lookahead of 5 ms as allowed by ITU-T [1]. Each frame is further subdivided into a variable number of subframes. A general overview of the encoder is given in Figure 1.

First, an input speech frame is classified into one of the three typical modes (0) *background noise / unvoiced*, (1) *transients / voicing onsets*, (2) *periodic / voiced*. Then, linear predictive coding (LPC) analysis and quantisation are performed on each speech frame to obtain the parameters for the synthesis filter and the perceptual weighting filter needed for the analysis-by-synthesis codebook search. The parameters are interpolated dependent on the chosen subframe length.

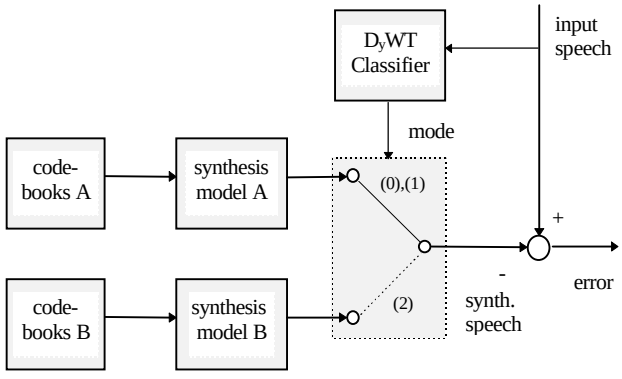

Figure 1: Principle of coding scheme.

The decision on the excitation coding model and the subframe length is based on the classification result. Modes (0) and (1) represent rather instationary speech segments and use synthesis model A with a short subframe length of 5 ms to produce the locally synthesised speech. For either mode of model A, a fixed excitation codebook was designed to match the mode-specific characteristics. Mode (2) respresents more stationary frames and uses synthesis model B with a double-size subframe length of 10 ms. Here, we use a fixed excitation codebook in combination with pitch prediction to utilise the signal periodicity.

The important parts of the coder are described in detail in the following sections. Block diagrams for synthesis models A and B are given in Figures 2a and 2b, respectively.

3 CLASSIFICATION ALGORITHM

The classifier [9] operates on input speech frames 20 ms in length. To avoid edge effects due to noncausal wavelets in the following transformation, we add a lookahead and a history of 5 ms, respectively, which results in an input buffer of 30 ms with an algorithmic delay of 25 ms.

For each buffer, the D_yWT is calculated using a cubic spline wavelet. This transformation can be efficiently implemented by means of cascaded filter banks because the splines chosen yield simple 4-tap lowpass and highpass FIR filters. The impulse responses of the filters are upsampled at every consecutive dyadic scale, as this proves to be better for the estimate of the signal periodicity than downsampling of the filter outputs. This filter bank is performed on the input buffer and iterated until scale 5 which is sufficient for telephone-bandwidth speech. The outputs of the highpass filters at all scales constitute the D_yWT coefficients of the current frame.

In order to achieve a higher time resolution, we subdivide the frame into 4 subframes so that the subframe length of 5 ms is matched to the smallest subframe length of our coding scheme. To classify each subframe into one of the three modes defined above, we calculate a set of three parameters (P_0, P_1, P_2) from the D_yWT coefficients, thereby defining decision scores in the interval (0,1) that indicate the corresponding modes.

Parameter P_0 is based on the variance of energy of the current frame across all scales and is an indicator of background noise or unvoiced frames. Parameter P_2 uses the distances between neighbouring local maxima of two consecutive *smooth* scales with low temporal resolution to derive a periodicity measure for the current frame. Both parameters are relatively rough indicators of the positions of the voiced and unvoiced parts of the speech segment. However, parameter P_1 is based on the energy difference per subframe of *fine*-scale coefficients with a high temporal resolution, so that the instants of significant changes in the input signal and, therefore, the correct boundaries of the voiced and unvoiced parts are well-localised.

Finally, the parameter set (P_0, P_1, P_2) controls a finite-state model that calculates the classifier's decision for each subframe. In this way, the history of past decisions can be taken into consideration and unvalid transitions, e.g. from (0) directly to (2) without intermediate step to (1), can be excluded. In order to save bits for transmission, we transform the four subframe results into one frame result according to the following rule: If all subframes are (0), then the frame is (0), else if all subframes are (2), then the frame is (2), otherwise the frame is (1).

The performance of the classifier for telephone-bandwidth speech is described in [9]. The algorithm is robust to various types of background noise, such as office noise and vehicular noise, for signal-to-noise ratios (SNR) down to 10 dB. In comparison with a common classifier based on the long-term prediction gain, the D_yWT classifier proved to be superior.

4 CODING OF LPC COEFFICIENTS

Linear predictive analysis is performed on every input speech frame using the recursive method as proposed in [10]. The input speech is weighted with the reversed impulse response of an IIR filter with a double pole at 0.983. Matching the framing of our classifier, we introduced a lookahead of 5 ms, so that the asymmetric window has its maximum in the last unit subframe of the current frame.

The LPC coefficients are transformed into line-spectral-frequency (LSF) parameters for quantisation. We utilise multi-stage vector quantisation (VQ) with interframe moving-average (MA) prediction as described in [2,11]. Since the frame length is 20 ms, we found that a 2nd order MA predictor is sufficient for exploiting the interframe redundancy of the LSF parameters. The prediction residue is quantised by a 2-stage VQ with a split structure at the 2nd stage. The codebook uses 7 bits for the first stage, 8 (4+4) bits for the second stage and 1 bit for the choice of the prediction matrix, resulting in a total of 16 bits.

The quantised LPC coefficients are used by the synthesis filter in the last subframe of the current frame. To improve the time resolution of the short-term spectrum, the intermediate one (mode 2) or three (mode 0,1) subframes are linearly interpolated in the LSF domain with last frame's coefficients. The coefficients of the perceptual weighting filter for excitation codebook search are updated in a similar manner using the unquantised LPC coefficients and bandwidth expansion coefficients of 0.94 for the numerator and 0.6 for the denominator, respectively.

5 EXCITATION CODING

Excitation coding depends on the mode chosen by the D_yWT classifier. Mode (0) and (1) are represented by synthesis model A with 4 subframes at 5 ms per frame whereas mode (2) is

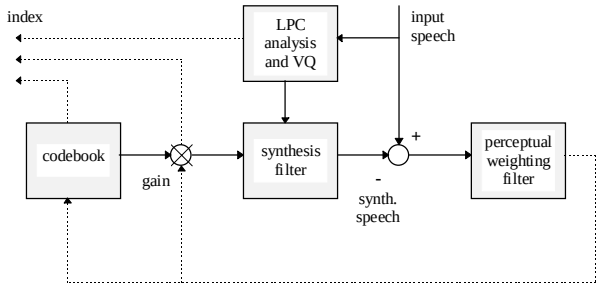

Figure 2a: Block diagram for synthesis model A

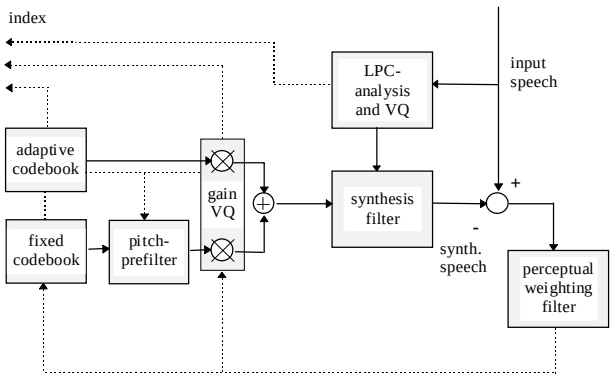

Figure 2b: Block diagram for synthesis model B

represented by model B with 2 subframes at 10 ms per frame.

5.1 Model A

In synthesis model A (Figure 2a) each codevector of the excitation codebook is scaled with the optimum gain and filtered by the synthesis filter to produce the synthesised speech vector. The codevector that minimises the perceptually weighted mean square error is selected for transmission.

We designed different codebooks for mode (0) and mode (1) to match the mode-specific characteristics of the signal. For mode (1) we use a sparse ternary codebook containing only M=3 non-zero pulses. Therefore, the codevector $c(i)$ is given by

$$c(i) = \sum_{k=1}^{M} s_k \, \delta(i - m_k) \, , \, i = 0, \ldots, N_{sf} - 1, \qquad (1)$$

where N_{sf} is the subframe length, δ(i) is a unit pulse and m_k, s_k, k=1,2,3, are the pulse positions and signs, respectively. In addition, the following conditions must hold:

$$m_1 \in [0, N_{sf} - 1]$$
$$m_2 - m_1 \geq \Delta_{\min}$$
$$m_3 - m_2 \geq \Delta_{\min}$$
$$s_3 = s_1 \, .$$

Setting Δ_{min} , the minimum difference between two successive pulses, to 16, we obtain an 11-bit codebook. For mode (0), an 11-bit Gaussian codebook is used.

At a second step the prediction residue γ of the gain G is quantised with 4 bits in the logarithmic domain utilising a simple gain predictor according to the following equation

$$\gamma^{(n)} = \log G^{(n)} - \alpha^{(n)} \log G_p^{(n)} \, , \text{with}$$
$$G_p^{(n)} = \sqrt{\frac{1}{N_{sf}^{(n-1)}} \sum_{i=0}^{N_{sf}^{(n-1)}-1} \left(r^{(n-1)}(i) \right)^2} \, , \qquad (2)$$

where n is the subframe index, G_p is the predicted gain, N_{sf} is the subframe length of the last subframe, $r(i)$ is the quantised excitation signal of the synthesis filter of the last subframe and α is the recursion factor. The predicted gain is initialised with 23 dB and is bounded between 10 and 60 dB. The recursion factor depends on the chosen mode and is set to 0.9865 for mode (0) or to 0.9332 for mode (1). The gain codebooks were designed using the Max-Lloyd algorithm with reordering of codebook entries in order to be more robust against single bit errors.

Two bits are reserved for parity check of the mode bits and the 2 most significant bits of each gain value, respectively. The overall bit allocation for synthesis model A is shown in Table 1a.

Parameter	Bits/subframe	Bits/frame
Mode	-	2
LSF	-	16
Gain	4+4+4+4	16
Codebook	11+11+11+11	44
Parity	-	2
Total	-	**80**

Table 1a: Bit allocation for synthesis model A

Parameter	Bits/subframe	Bits/frame
Mode	-	2
LSF	-	16
Gain	6+6	12
Pitch lag	8+5	13
Fixed codebook	18+18	36
Parity	-	1
Total	-	**80**

Table 1b: Bit allocation for synthesis model B

5.2 Model B

Synthesis model B is used in mode (2) and combines the output of two codebooks to compute the excitation signal of the synthesis filter for the double-size subframes. Excitation coding in model B is based on the CS-ACELP coder described in [2].

The adaptive codebook contains the past excitation signal and exploits the pitch periodicity during voiced segments. The optimum pitch lag T is determined by closed-loop search of the adaptive codebook with 3-times precision for lags in the interval (20,85) and with integer precision in the interval (86,143). In the first subframe the lag is encoded with 8 bits and in the second subframe it is differentially encoded with 5 bits. The codevector from the adaptive codebook is then scaled by the optimum gain, filtered and subtracted from the target speech vector to obtain the new target vector for the fixed codebook search.

The fixed codebook is based on an algebraic codebook structure. Each codevector $c(i)$ contains M=4 non-zero pulses and is defined by equation (1). However, the pulse positions are now given by an interleaved single-pulse permutation design according to Table 2. The bracketed values for m_k indicate that the pulses may be missing which allows a variable number of pulses per subframe. Permitting all possible sign combinations, 18 bits are used for the fixed codebook excitation.

To improve the periodicity in the reconstructed speech, an adaptive pitch prefilter is incorporated in the fixed codebook search which enhances the harmonic components when the pitch lag is shorter than the subframe length. The z-transform $P(z)$ of this filter is given by

$$P^{(n)}(z) = \frac{1}{1 - \beta^{(n-1)} z^{-T^{(n)}}} \, , \qquad (3)$$

where T is the integer part of the pitch lag of the current subframe and β is the adaptive codebook gain of the previous subframe bounded by (0.2,0.8).

The adaptive codebook gain β and the logarithmic prediction residue γ of the fixed codebook gain, as defined in equation (2), are vector-quantised by closed-loop search of the 6-bit gain codebook. The recursion factor of the gain predictor for mode (2) is set to α=0.9944. The gain codebook for mode (2) was designed using the generalised Lloyd algorithm and reordering of codebook indices.

One parity bit is left to protect both the mode and the three most significant bits of each gain vector. The overall bit allocation of synthesis model B is given in Table 1b.

k	s_k	m_k
1	±1	0,8,16,24,32,40,48,56,64,72,(80)
2	±1	2,10,18,26,34,42,50,58,66,74,(82)
3	±1	4,12,20,28,36,44,52,60,68,76,(84)
4	±1	6,14,22,30,38,46,54,62,70,78,(86)

Table 2: 18-bit ternary sparse fixed codebook for mode (2)

6 DECODER

For each speech frame, the decoder synthesises the speech according to synthesis model A or B dependent on the detected mode flag. To improve the subjective speech quality in the tandemed and non-tandemed cases, an adaptive postfilter [2] is applied to the synthesised speech on a subframe-by-subframe basis.

An error concealment procedure has been incorporated in the decoder to reduce the degradation in the reconstructed speech because of frame erasures or single bit errors in the mode bits. In this case, the LPC filter coefficients of the previous frame are repeated. The missing excitation signal for the current frame will only be replaced by the adaptive codebook signal, while gradually attenuating its energy. This is done by using the detected mode of the previous frame. If last subframe's mode was (2), the pitch lag of the last subframe in the last frame will be repeated to form the excitation signal. Otherwise, the pitch lag will be set equal to the frame length.

7 RESULTS

To evaluate the performance of the proposed coding scheme, we consider a speech data base consisting of 16 German sentences spoken by 4 male and 4 female speakers. All test sentences were filtered by a modified IRS filter to simulate the frequency response of a telephone handset and scaled to an average speech level of -30 dB with respect to the system overload. After discarding of silent frames with an energy of 40 dB below the maximum frame energy of the sentences, the following distribution of modes for active speech was found: mode (0): 22.5%, mode (1): 9.1%, mode (2): 68.4%.

At a first step, we analysed the performance of the LPC coding scheme. Table 3 shows the logarithmic spectral distortion for non-silent frames dependent on the chosen mode. It can be seen that the 16-bit quantisation scheme performs well, although the 1-dB limit for transparent quantisation is not reached. The distortion slightly increases for mode (1) but remains inaudible in most cases.

For comparison of the overall speech quality of the proposed coding scheme with that of a conventional CELP coder, we define a coder that always operates in mode (2) according to synthesis model B (see Figure 2b and Table 1b). We use the average perceptually-weighted segmental SNR (WSegSNR) as an objective measure for the speech quality. WSegSNR is a modification of the commonly used segmental SNR where the variance of the quantisation error signal is replaced by the variance of the weighted error signal as obtained at the output of the perceptual weighting filter in the CELP model.

Table 4 shows the WSegSNR in dB, evaluated over non-silent frames, for the proposed and the conventional coders, respectively. The proposed coder is superior to the conventional coder for all three modes. It improves the WSegSNR by 2.3 dB on the average. Note that there is even a significant improvement for mode (2), although the same synthesis model is used in this case. This is due to a better startup phase during voicing onsets according to synthesis model A in mode (1). In informal listening tests the subjective quality of the proposed coder was assessed to be significantly better than that of the conventional coder and slightly below that of G.726 operating at 32 kbit/s.

8 CONCLUSION

We designed a 4 kbit/s CELP coding algorithm with a 20-ms frame that is based on signal classification using the D_yWT. The algorithm utilises specific synthesis models for each of three modes to match the temporal input signal characteristics. The proposed coder provides a significantly better speech quality than a conventional CELP coder operating at the same bit rate.

Mode	**average SD** [dB]	**SD > 2 dB** [%]
(0)	1.297	4.89
(1)	1.499	21.27
(2)	1.354	8.13
Total	**1.356**	**8.60**

Table 3: Spectral distortion (SD) of 16-bit LPC quantisation

	Conventional coder			**Proposed coder**		
Mode	Male	Female	Total	Male	Female	Total
(0)	4.15	4.13	4.14	5.74	5.69	5.71
(1)	3.77	4.08	3.94	5.89	6.17	6.05
(2)	9.14	10.14	9.67	11.72	12.94	12.36
Total	7.65	8.11	**7.91**	10.00	10.55	**10.30**

Table 4: WSegSNR (dB) for conventional and proposed coder

REFERENCES

[1] ITU-T, SG 15: *Terms of Reference for 4-kbit/s speech coding*, 1995.

[2] ITU-T, COM 15-152*: Draft Recommendation G.729 - Coding Of Speech at 8 kbit/s using Conjugate-Structure Algebraic-Code-Excited-Linear-Prediction (CS-ACELP)*, 1995.

[3] Gerson, I. et al.: *Speech and Channel Coding for the Half-Rate GSM Channel*, Proc. of ITG-Conference „Codierung für Quelle, Kanal und Übertragung", Munich, 1994.

[4] Wang, S., Gersho, A.*: Phonetically-Based Vector Excitation Coding of Speech at 3.6 kbit/s*, Proc. ICASSP, 1989.

[5] Nomura, T., Ozawa, K., Serizawa, M.: *Efficient Excitation Model and LPC Coefficients Coding in 4 kbps CELP with 20ms Frame*, Proc. of IEEE Workshop on Speech Coding for Telecommunications, 1995.

[6] Campbell, J., Tremain, T.: *Voiced/Unvoiced Classification of Speech with Application to the U.S. Gouvernment LPC-10e Algorithm*, Proc. ICASSP, 1986.

[7] Rioul, O., Vetterli, M.: *Wavelets and Signal Processing*, IEEE Signal Processing Magazine, Oct. 1991.

[8] Mallat, S., Zhong, S.: *Characterization of Signals from Multiscale Edges*, IEEE Transactions on Pattern Analysis and Machine Intelligence, July 1992.

[9] Stegmann, J., Schröder, G., Fischer, K.A.: *Robust Classification of Speech Based on the Dyadic Wavelet Transform with Application to CELP Coding*, Proc. ICASSP, 1996.

[10] Barnwell, T.: *Recursive Windowing for Generating Autocorrelation Coefficients for LPC Analysis*, IEEE Transactions on Acoustics, Speech, and Signal Processing, Oct. 1981.

[11] Ohmuro, H., Moriya, T., Mano, K., Miki, S.: *Coding of LSP Parameters Using Interframe Moving Average Prediction and Multi-Stage Vector Quantization*, Proc. of IEEE Workshop on Speech Coding for Telecommunications, 1993.