16 KBIT/S WIDEBAND SPEECH CODING BASED ON UNEQUAL SUBBANDS

Jürgen W. Paulus and *Jürgen Schnitzler*
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, juergen.paulus@ind.rwth-aachen.de

ABSTRACT

In this paper we propose a split-band encoding scheme for 16 kbit/s wideband speech coding (50-7000 Hz), using 2 unequal subbands from 0-6 kHz and from 6-7 kHz. This approach was motivated by experimental evaluation of the signal bandwidth of speech frames. The higher subband is simply represented by white noise with adjustment of the short term energy. For the lower subband code-excited linear prediction (CELP) is used. By informal listening tests the speech quality was rated higher than the speech quality of the CCITT G.722 wideband codec operating at 48kbit/s.

1. INTRODUCTION

During the last few years there has been an increasing effort in wideband speech coding at lower bit rates. This not only arises from high quality videophone and digital mobile telephone applications, but also from the increasing market for multimedia systems where high quality speech and audio is demanded. Compared to narrowband telephone speech, the reduction of the lower cut off frequency from 300 Hz to 50 Hz contributes to increased naturalness and fullness. The high frequency extension from 3400 Hz to 7000 Hz provides better fricative differentiation and therefore higher intelligibility. In 1986 the International Telegraph and Telephone Consultative Committee (CCITT, now ITU-T) recommended the G.722 standard for wideband speech and audio coding. This wideband speech codec provides high speech quality at 64 kbit/s with a bandwidth of 50 Hz to 7000 Hz [1]. Slightly reduced qualities are achieved at 56 and 48 kbit/s. Since September 1993, the International Telecommunications Union Study Group 15 (ITU-T SG 15) studies in Question 6 ("Audio and Wideband Coding for Public Telecommunication Networks") new coding schemes for low-rate wideband speech coding at 16, 24, and 32 kbit/s [2]. The G.722 standard will serve as a reference for the development of this alternative coding scheme.

In the past, linear prediction models have been used very successfully for the coding of telephone speech. Recently, a new 8 kbit/s narrowband speech coder has been selected by the ITU-T SG 15 which provides telephone quality at 1 bit/sample [3, 4]. This indicates, that very good coding quality might be possible for wideband speech signals with 1 bit/sample, too. However, for audio signals the desired quality has not yet been achieved using LPC techniques with long term prediction (LTP) which are based on a model of speech production. For those signals, subband coding, transform coding and various forms of entropy coding have been used for efficient coding with 2-3 bits per sample, if no oversampling is applied.

In the following sections an encoding scheme for speech will be presented which consists of a 2-band splitband scheme with unequal bandwidths of the subbands. This approach is motivated by the experimental evaluation of the instantaneous signal bandwidth. First, in Section 2 a classification scheme is explained which leads to the unequal splitting of the subbands. Afterwards the analysis filter bank is described which performs the unequal band splitting combined with critical subsampling of the sub-bands. In Section 3 and Section 4 the encoding techniques for both bands are explained. In Section 5 a bit error concealment technique is described and in Section 6 the final bit allocation is given. In Section 7 we discuss the extension of the coding scheme towards variable bitrate.

2. ANALYSIS FILTERBANK

The use of unequal subbands 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 present in the lower frequency region. Therefore it is not necessary to encode the higher part of the frequency range. Transform coding techniques behave in a similar way in that they allocate in voiced frames more bits to code lower frequency components than 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 10 ms we found that almost 40% of the frames could be encoded using a bandwidth of 6 kHz without loss of perceptual quality. The full bandwidth was selected mainly during unvoiced parts of the speech signals. The voice activity of the speech material used was 95%. It was extracted from the European Broadcasting Union database [5]. The speech material consists of various languages (English, German, and French), each with male and female speakers, and was bandlimited to a frequency range of 50-7000 Hz, according to the specifications in the G.722 recommendations [1]. As a result of the classification, a 2-band encoding scheme is proposed which consists of subbands with unequal bandwidth. The lower subband has a frequency range from 0-6 kHz and the upper subband

0-7803-3192-3/96 $5.00©1996 IEEE

covers a frequency range from 6-7 kHz, i.e. we obtain a subband coder with 2 bands having a bandwidth of 6 kHz and 1 kHz respectively. Figure 1 shows the analysis filterbank for unequal subband splitting and critical subsampling of the subbands.

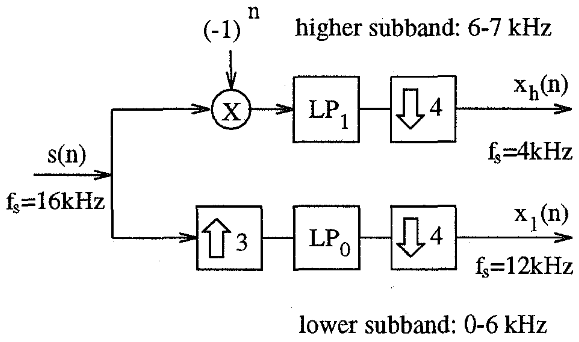

Figure 1. Analysis filterbank for subband splitting and critical subsampling of the subband signals.

The analysis filterbank is implemented using the efficient structure for sampling rate conversion with a fractional ratio of the sampling rate, as described for example by Crochiere *et al.* [6].

3. ENCODING OF THE 0-6 KHZ BAND

For encoding the decimated lower subband code-excited-linear-prediction (CELP, Atal *et al.* [7]) is performed. The coder operates on speech frames of 10 ms (120 samples). In the following, the main parts of the CELP-codec will be described: LP-analysis, pitch analysis, fixed codebook structure and perceptual weighting filter.

The subframe lengths used for the different parts of the codec are indicated in Figure 2, being 5 ms for the pitch analysis and 2.5 ms for the fixed codebook.

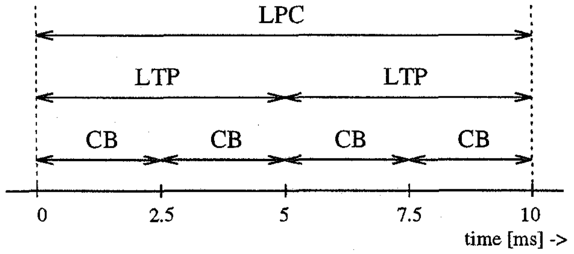

Figure 2. Update of the codec parameters.

3.1. LP-analysis

The Linear-Prediction (LP) analysis uses a covariance-lattice approach as described by Cumani [8]. The analysis frame length is 15 ms, centered around the middle of the second LTP-subframe, resulting in a look-ahead of 5 ms. In our realization the order of the LP-filter is 14. The prediction coefficients are updated every 10 ms. Prior to solving the equations for the coefficients, the covariance matrix is modified by weighting it with a binomial window having an effective bandwidth of 80 Hz [9]. This provides a small amount of bandwidth expansion to the final LP-filter coefficients. This is advantageous for the following conversion of the LP filter parameters to line spectral frequencies (LSF) [10], as well as for the quantization of the LSF's.

The LSFs are encoded using 44 bits by interframe moving average prediction and split vector quantization of the line spectral frequencies resulting in an average spectral distortion of 1 dB.

A linear interpolation of the LP-filter coefficients is performed for the first LTP-subframe. This is done in the LSF-domain between the quantized actual coefficient set and the quantized coefficient set of the previous frame. For the second subframe, no interpolation is performed.

3.2. Pitch analysis

Every 5 ms, a long-term-prediction (LTP) is carried out in a combination of open-loop and closed-loop LT-analysis. For each 10 ms speech frame, 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. In the first subframe a focussed closed-loop adaptive codebook search is performed around the open-loop estimate τ_{ol}, and in the second subframe a restricted search is performed around the pitch lag of the closed-loop analysis of the first subframe $\tau_{cl,1}$, as depicted in Figure 3.

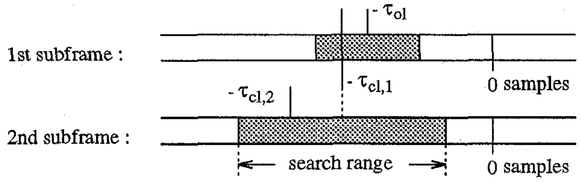

Figure 3. Long-Term analysis using combined open-loop and closed-loop analysis and a focussed search strategy.

This procedure results in a delta encoding scheme leading to 8+6=14 bits for coding the 2 pitch lags.

The closed-loop search is performed using an adaptive codebook filled with previously computed excitation samples. The minimum pitch lag is half of the subframe length, i.e. $\tau_{\min} = 30$ samples. Additionally, in the lower delay range a fractional pitch approach is used [11], as shown in Figure 4.

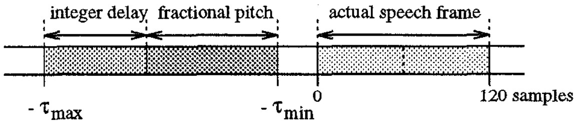

Figure 4. Combined integer and fractional pitch search ranges during closed-loop adaptive codebook search ($\tau_{\max}$=193 samples).

Informal listening tests indicate, that a resolution of 1/2 sample is sufficient for an improvement in speech quality.

The pitch gain is nonuniformly scalar quantized with 4 bits.

3.3. Codebook

Every 2.5 ms (30 samples), an excitation vector is selected from a modified 16-bit ternary sparse codebook, as described by Salami *et al.* [12]. An innovation vector contains 4 nonzero pulses, as shown in Table 1.

Amplitude	Position
±1	0, 4, 8, 12, 16, 20, 24, 28
±1	1, 5, 9, 13, 17, 21, 25, 29
±1	2, 6, 10, 14, 18, 22, 26, (30)
±1	3, 7, 11, 15, 19, 23, 27, (31)

Table 1. 16-bit ternary sparse codebook [12].

Note that the last position of the 3rd and 4th pulse falls outside the subframe boundary. This gives the possibility of a variable number of pulses per frame.

Each pulse has 8 possible positions. Therefore the pulse positions are encoded for each pulse with 3 bits. Furthermore, each pulse amplitude is encoded with 1 bit, resulting in a total of 16 bits for the 4 pulses.

Due to the structured nature of the codebook, a fast search procedure is ensured. Additionally, a focussed search approach is used to further reduce the computational load of the codebook search [12].

To reduce the dynamik range of the fixed codebook gain, a fixed gain predictor is used. The gain predictor is predicting the log. energy of the current fixed codebook vector based on the log. energy of the previously selected scaled fixed codebook vector. This is done in a similar way as in a preliminary version of ITU-T G.729 [13]. The residual of the gain predictor is nonuniformly scalar quantized with 4 bits.

3.4. Perceptual weighting filter

The perceptual weighting filter $W(z)$ used during the minimization process has a transfer function of the form

$$W(z) = \frac{A(z/\gamma_1)}{A(z/\gamma_2)}, \qquad 0 \leq \gamma_2 \leq \gamma_1 \leq 1 \qquad (1)$$

with $A(z)$ being the LP-analysis filter, using unquantized LP-filter coefficients. Different sets of weighting factors $\{\gamma_1, \gamma_2\}$ are used for the adaptive and fixed codebook search. During the adaptive codebook search, weighting factors $\{1.0, 0.4\}$ are used, and during the fixed codebook search $\{0.9, 0.8\}$ is used. This was found to give better results compared to a fixed weighting filter.

The perceptual weighting filter is updated every 5 ms, using in the first subframe a linear interpolation between the actual unquantized filter coefficients and the unquantized filter coefficients of the previous frame. In the second subframe the actual unquantized coefficients are used.

4. ENCODING OF THE 6-7 KHZ BAND

The classification experiment shows, that the full bandwidth is selected mainly during the unvoiced parts of the speech signal. This indicates that the higher subband has a noise like character. Furthermore, it turned out by experiment that during unvoiced parts it is sufficient to add some noise like spectral components above 6 kHz to obtain the perceptual speech quality of a 7 kHz speech signal. Therefore, the higher subband (6-7 kHz) is simply represented by white noise with adjustment of the short term energy. At the output of the analysis filterbank of Section 2 the subband signal $x_h(n)$ has a sampling rate of 4 kHz, i.e. a bandwidth of 2 kHz (see Figure 1). Since the input signal is bandlimited to 7 kHz, a further reduction of the sampling rate by a factor of 2 could be done without use of an aliasing filter. The input frame length of 10 ms (20 samples) is split up into 4 subframes of 2.5 ms, each consisting of 5 samples. For each subframe the short term energy is logarithmically quantized with 3 bits using MA-prediction with a fixed set of coefficients. This results in a bitrate of 1.2 kbit/s for the higher subband.

An informal listening test was performed using a high quality loudspeaker. The higher subband was processed using the encoding scheme as described above. The lower band remained uncoded, however the sampling rate conversion of the lower sub-band was carried out. As a result, it was difficult to distinguish between the original and the processed speech signal. Thus, this very simple encoder can be used to encode the subband from 6-7 kHz.

5. BIT ERROR CONCEALMENT

For the previously described scheme, the overall bit-rate sums up to 15.8 kbit/s. This gives the possibility of using 2 parity-bits per frame for reducing the sensitivity of the codec to random bit errors up to BER=10^{-3}. After performing informal listening tests, it was concluded, that the LP-coefficients are most sensitive against bit errors.

Therefore, the first parity-bit is computed from the 44 bits of the LP-coefficients. This bit is transmitted, and at the decoder the parity-bit is recomputed from the received LP-filter cofficients. If a parity-error occurs, the LP-coefficient set is replaced by the values of the previous frame.

The second parity-bit is computed from the 8 bits of the LTP-index of the first subframe. If a parity-error occurs, the value of the LTP-index is set to the integer delay value of the previous subframe.

6. BIT ALLOCATION

In the previous sections, the main components of the wideband codec were presented. According to Table 2, a final bit-rate of 16 kbit/s is achieved.

6-7 kHz	Energy	4*3 Bit	12 bits	1.2 kbit/s
0-6 kHz	LPC		44 bits	4.4 kbit/s
	LTP-Index	8+6 Bit	14 bits	
	LTP-Gain	2*4 Bit	8 bits	2.2 kbit/s
	CB-Index	4*16 Bit	64 bits	
	CB-Gain	4*4 Bit	16 bits	8.0 kbit/s
	Parity bits		2 bits	0.2 kbit/s
Σ				16.0 kbit/s

Table 2. Bit allocation for a 10 ms frame of the proposed 16 kbit/s splitband wideband codec

7. EXTENSION TO VARIABLE BITRATE

One of the results of Section 2 has been, that 40% of the speech signal with a voice activity of 95% could be encoded using just the subband from 0-6 kHz. This means, during 40% of the active talk time it is not necessary to encode the higher subband. This encourages us to consider different coding schemes.

The first alternative is to neglect the bits necessary to encode the higher subband. This leads to a coder with a variable bitrate. Transmission of 1 Bit/frame is necessary in this case to indicate the encoding mode of the higher subband.

The second possibility is to use these bits to encode the lower subband more precisely, resulting in an encoder with an overall constant bitrate, but variable bitrate in the two bands. Since this happens most of the time during voiced parts of a speech signal this is advantageous with respect to speech quality. Again one additional Bit/frame is necessary to indicate the encoding mode of the higher subband.

Another possibility was recently presented by the author in [14], based on a similar approach in [15] in the context of wideband ADPCM. The wideband speech signal is encoded using only the spectral bandwidth from 0-6 kHz and the higher subband is neglected. The missing components above 6 kHz are replaced at the receiver by interpolating the lower subband signal from 12 kHz to 16 kHz using an interpolation filter with cut-off frequency 7 kHz which violates the interpolation rules. This is possible due to the fact, that the signals within the frequency ranges 5-6 kHz and 6-7 kHz exhibit a similar distribution of energy along the time axis for a given speech sound. In this case a fixed bitrate of 14.8 kbit/s is achieved, with only very small degradations compared to the fixed bit-rate version of the previous sections.

8. CONCLUSION

In this paper a split-band encoding scheme for 16 kbit/s wideband speech coding has been presented. It is based on two unequal subbands from 0-6 kHz and 6-7 kHz. This approach was motivated by experimental evaluation of the instantaneous signal bandwidth of the speech frames. The coder operates on speech frames of 10 ms, using a look-ahead of 5 ms for LP-analysis. Together with the 10 ms delay introduced by the analysis-synthesis filterbank, this results in an overall algorithmic delay of 25 ms. By informal listening tests the speech quality was judged to be better than the CCITT G.722 wideband codec operating at 48 kbit/s.

ACKNOWLEDGEMENTS

This work has been supported by the Research Center of Deutsche Telekom AG. The author would like to thank especially Mr. G. Schröder. Acknowledgements are made to Prof. P. Vary and the colleagues of the speech coding group for inspiring discussions, especially T. Fingscheidt.

REFERENCES

[1] CCITT, "7 kHz Audio Coding within 64kbit/s", in *Recommendation G.722*, vol. Fascile III.4 of *Blue Book*, pp. 269–341. Melbourne 1988.

[2] Study Group 15 ITU-T, "Report February 1995 Meeting Working Party 2/15", February 1995, Geneva, Switzerland.

[3] S. Dimolitsas, "ITU Voice Coding Standards: Standardization of Voice Coding Milestones Reached", comp.speech Newsgroup, February 1995.

[4] ITU-T SG15 COM 15-152, "G.729 - Coding of Speech at 8kbps using conjugate-structure algebraic-code-excited linear-predictoin (CS-ACELP)".

[5] European Broadcasting Union (EBU), *Sound Quality Assesment Material (Recordings for Subjective Test)*, no. 422 204–2 edition.

[6] R.E. Crochiere and L.R. Rabiner, *Multirate Digital Signal Processing*, Signal Processing. Prentice-Hall, 1983.

[7] B.S. Atal and M.R. Schroeder, "Stochastic Coding of Speech Signals at Very Low Bit Rates", in *Proc. Int. Conf. Communication (ICC)*, May 1984, pp. 1610–1613.

[8] A. Cumani, "On a Covariance-Lattice Algorithm for Linear Prediction", in *Proc. Int. Conf. Acoust., Speech, Signal Processing, ICASSP*, Paris, France, 1982, pp. 651–654.

[9] Y. Tohkura and F. Itakura nad S. Hashimoto, "Spectral Smoothing Technique in PARCOR Speech Analysis-Synthesis", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 26, no. 6, pp. 587–596, December 1978.

[10] P. Kabal and R.P. Ramachandran, "The Computation of Line Spectral Frequencies Using Chebyshef Polynomials", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, no. 6, pp. 1419–1426, December 1986.

[11] J. S. Marques, J. M. Tribolet, I. M. Trancoso, and L. B. Almeida, "Pitch Prediction with Fractional Delays in CELP Coding", in *Proc. EUROSPEECH*, Genua, Italien, 1989, pp. 509–513.

[12] R. Salami, C. Laflamme, J-P. Adoul, A. Kataoka, S. Hayashi, C. Lamblin, D. Massaloux, S. Proust, P. Kroon, and Y. Shoham, "Description of the Proposed ITU-T 8kb/s Speech Coding Standard", in *Proc. IEEE Workshop on Speech Coding*, Annapolis, Maryland, USA, September 1995, pp. 3–4.

[13] R. Salami, C. Laflamme, J.-P. Adoul, and D. Massaloux, "A Toll Quality 8 kb/s Speech Codec for the Personal Communications System (PCS)", *IEEE Trans. Vehicular Technology*, vol. 43, no. 3, pp. 808–816, August 1994.

[14] J. Paulus, "Variable Bitrate Wideband Speech Coding Using Perceptually Motivated Thresholds", in *Proc. IEEE Workshop on Speech Coding for Telecommunications*, Annapolis, Maryland, USA, September 1995, pp. 35–36.

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