WIDEBAND SPEECH CODING FOR THE GSM FULLRATE CHANNEL?

*Jürgen Paulus** *Jürgen Schnitzler*

Institute of Communication Systems and Data Processing (IND)
RWTH Aachen, University of Technology, D-52056 Aachen, Germany
Phone: +49.241.806982, Fax: +49.241.8888186, E-mail: Juergen.Schnitzler@ind.rwth-aachen.de

ABSTRACT

In this paper we propose a wideband speech encoding scheme (50-7000 Hz) having a bit rate of 12.3 kbit/s which could be used for the GSM Fullrate channel. The coding scheme is based on 2 unequal subbands from 0-6 kHz and from 6-7 kHz. This approach was motivated by experimental evaluation of the instantaneous signal bandwidth of speech frames. The lower subband is encoded using code-excited linear prediction (CELP). The higher subband is replaced at the receiver by aliased components of the lower band using an interpolation filter with a cut-off frequency of 7 kHz. By informal listening tests the speech quality was rated higher than the speech quality of the CCITT G.722 wideband codec operating at 48 kbit/s. In comparison to the GSM Fullrate codec with 13 kbit/s, naturalness and intelligibility are improved significantly.

1. INTRODUCTION

Since several years the so called 'full rate codec' has been used in the GSM system for mobile communication [1]. This codec which is operating at 13 kbit/s was designed almost one decade ago for the encoding of narrowband speech signals (0.3-3.4 kHz) which are sampled at 8 kHz. The effective bit rate is therefore $13/8 = 1.625$ bit per sample.

Recently, a new 8 kbit/s narrowband speech coder has been standardized by the Study Group 15 of ITU-T which provides telephone quality at 1 bit per sample only [2]. In comparison to these codecs, a wideband coding scheme would require a sampling rate of 16 kHz and the effective target bit rate would be $13/16 = 0.8125$ bit per sample.

During the last few years there has been an increasing effort in wideband speech coding at lower bit rates. This arises not only from applications such as high quality videophone and digital mobile telephone, 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.

* Now with Siemens AG München,
E-mail: juergen.paulus@hl.siemens.de

2. ENCODER STRUCTURE

Recently, we presented a split-band encoding scheme using 2 unequal subbands, i.e. 0-6 kHz and 6-7 kHz [3]. This approach was motivated by the experimental evaluation of the instantaneous signal bandwidth of speech frames, i.e. the cut-off frequency necessary to encode the current frame without loss of perceptual speech quality. In an experiment, a bank of linear phase FIR filters with cut-off frequencies decreasing in steps of 1 kHz (starting at 7 kHz) were applied to the speech data. After filtering a frame of 10 ms length using all the lowpass filters, an energy ratio $\kappa_E = E/\triangle E$ was calculated from the full-band signal energy E and the energy $\triangle E$ of the difference signal between the filtered and original signal. If the value of this energy ratio is greater than a certain threshold $\kappa_{E_{thr}}$, the cut-off frequency of the actual signal could be reduced to the cut-off frequency of the lowpass. The threshold $\kappa_{E_{thr}}$ was obtained by informal listening tests. In Figure 1 a histogram of the final cut-

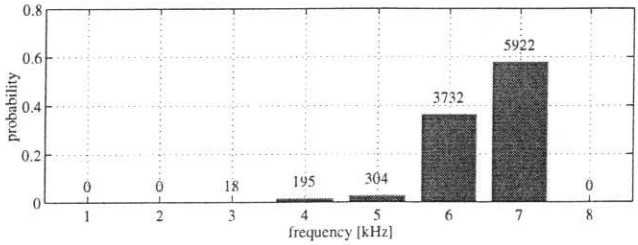

Figure 1. Histogram of allowable cut-off frequencies

off frequencies is presented. A simulation was performed using 100 s multilingual speech (english, german, french) each with male and female speakers. The speech material having an speech activity of 95 % was extracted from the European Broadcasting Union database [4]. This material was bandlimited to a frequency range of 50-7000 Hz, according to the CCITT G.722 recommendation [5]. By applying a frame size of 160 samples (10 ms) and choosing an energy threshold of $\kappa_{E_{thr}} = 800000$, it resulted that almost 40% of the frames (4249 out of 10171) could be encoded using a bandwidth of only 6 kHz, without loss of perceptual

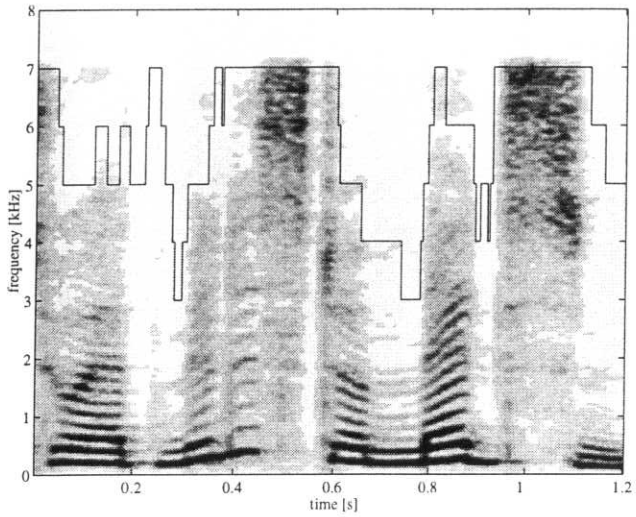

Figure 2. Spectrogram and quantized cut-off frequency $f'_{g_q}(t)$ for the speech segment *'To administer medicine'*.

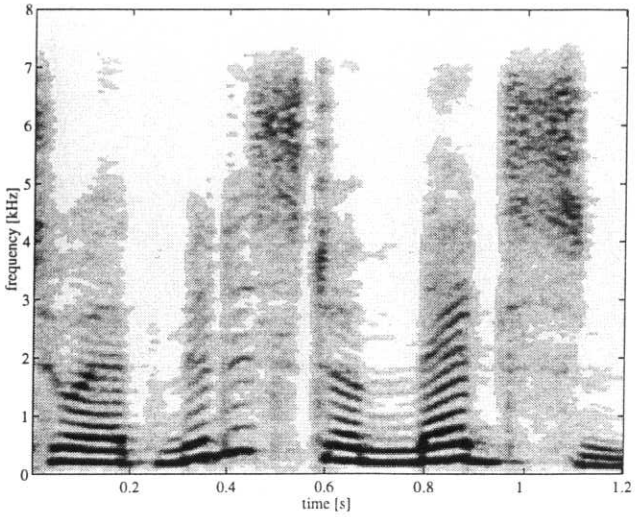

Figure 3. Spectrogram of the previous speech segment, with the original signal decimated from 16 kHz to 12 kHz and subsequently re-interpolated using an 7 kHz interpolation filter.

speech quality.

The full bandwidth was selected only during unvoiced parts of the speech signal. Due to the mechanism of speech production, spectral components above approximately 4 kHz are almost exclusively referring to fricative speech. In contrast to that, during voiced parts of speech, most of the signal energy is present 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 the low frequency components than to the high frequency components of speech. Figure 2 gives an example of a speech segment and its quantized instantenous bandwidth.

Visual inspection of the waveforms of speech components within the frequency ranges 5-6 kHz and 6-7 kHz shows that both of these signals exhibit a similar distribution of energy along the time and frequency axis for a given speech sound. Furthermore, it turned out by informal listening experiments 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. This effect has been used in our previous proposal [3].

This observation can be alternatively used to substitute the speech components in the frequency range of 6-7 kHz by something else without transmitting any side-information as shown below. The wideband speech signal is encoded using only the spectral bandwidth from 0-6 kHz and the missing components above 6 kHz are produced 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 requirements.

A further spectrogram for the same speech example is given in Figure 3. The decimation/interpolation chain described above was applied to the original speech signal of Figure 2, but without the decimated signal being encoded yet. Both spectrograms exhibit a considerable similarity, although, especially near $t = 0.5$ ms, the aliasing effect is visible. However, the enhancement of the perceptual speech quality is surprisingly high. The basic approach was already proposed by Dietrich [6] in the context of wideband ADPCM.

In Figure 4 and Figure 5 the encoder and decoder structures are given.

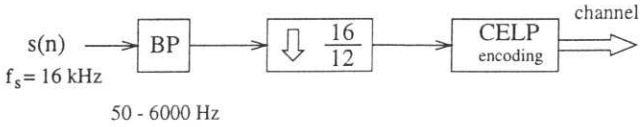

Figure 4. Encoder structure of the proposed codec

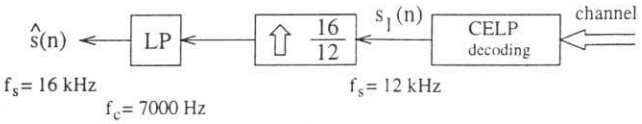

Figure 5. Decoder structure of the proposed codec

3. CELP ENCODING SCHEME

For encoding the decimated signal, code-excited linear prediction (CELP, Atal *et al.* [7]) is performed. The coder operates on speech frames of 20 ms (240 samples).

The subframe lengths used for the different codec parts are illustrated in Figure 6, being 5 ms for the pitch analysis and 2.5 ms for the fixed codebook search.

3.1. LP analysis

The linear prediction (LP) analysis uses a covariance-lattice approach as described by Cumani [8]. The analysis win-

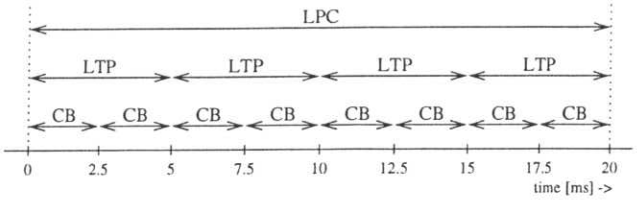

Figure 6. Update of the codec parameters

dow length is 260 samples (≈ 21.7 ms), centered around the middle of the frame. The order of the LP-filter of our realization is 14. The prediction coefficients are updated every 20 ms and converted to line spectral frequencies (LSF) [9]. 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 [10].

The LP coefficients are encoded using 42 bit by an hybrid trained predictive vector and lattice vector quantization (PVQ-LVQ) scheme for the line spectral frequencies (LSF) [11]. The computationally efficient quantization scheme leads to an average spectral distortion [12] of 1 dB only.

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

3.2. Pitch analysis

Every 5 ms, the long-term prediction (LTP) is carried out in a combination of open-loop and closed-loop LT-analysis. Every 10 ms, 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 and third LTP subframe, a focussed closed-loop adaptive codebook search is performed around the open-loop estimate τ_{ol}, and in the second and fourth subframe a restricted search is performed around the pitch lag of the closed-loop analysis of the first (third) subframe $\tau_{cl,1}$, as depicted in Figure 7.

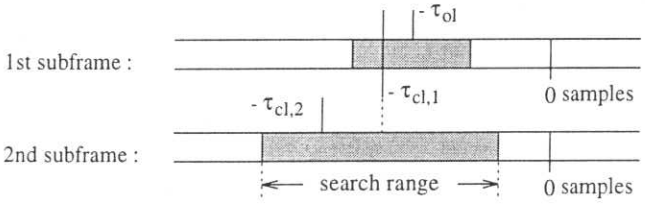

Figure 7. Long-term analysis using combined open-loop and closed-loop analysis and a focussed search strategy.

This procedure results in a delta encoding scheme requiring 2×(8+6)=28 bits for coding the 4 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

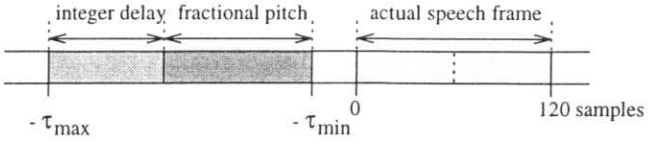

Figure 8. Combined integer and fractional pitch search ranges during closed-loop adaptive codebook search.

fractional pitch approach is used [13], as shown in Figure 8.

The pitch gain is quantized nonuniformly quantized with 4 bits.

3.3. Fixed 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.* [14]. 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 [14].

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 (i.e. ±1), resulting in a total of 16 bits for the 4 pulses.

Due to the structure of the codebook, a fast search procedure is possible. Additionally, a focussed search approach is used for further reductions of the computational load of the codebook search [14].

To reduce the dynamic 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 [15]. The residual of the gain predictor is quantized nonuniformly 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

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

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.

The coefficients of the weighting filters are updated similarly to the LP synthesis filter, but using the unquantized LSF.

4. BIT ALLOCATION

According to Table 2, a total bit-rate of 12.3 kbit/s is achieved. The codec will be demonstrated at the conference by audio tape.

Parameter	Bit Allocation	Bits/Frame	Bit Rate
LPC		42 bit	2.1 kbit/s
LT-Index	2×(8+6) bit	28 bit	2.2 kbit/s
LT-Gain	4× 4 bit	16 bit	
CB-Index	8×16 bit	128 bit	8.0 kbit/s
CB-Gain	8× 4 bit	32 bit	
Σ			12.3 kbit/s

Table 2. Bit allocation for a 20 ms frame of the proposed 12.3 kbit/s wideband codec

5. CONCLUSION

In this paper a split-band encoding scheme for wideband speech coding at 12.3 kbit/s 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 instantenous signal bandwidth. The lower band codec operates on speech frames of 20 ms using code-excited linear prediction. Taking advantage of a similar energy distribution in the 5-6 kHz and 6-7 kHz bands, no information is transmitted for the upper band; the missing components above 6 kHz are generated at the decoder by extrapolation from the 5-6 kHz band. With respect to informal listening tests, this experimental encoding scheme exhibits a speech quality rated higher than the CCITT G.722 wideband codec operating at 48 kbit/s. Comparing the speech quality to the GSM Fullrate codec at 13 kbit/s, a significant improvement of naturalness and intelligibility has been achieved.

ACKNOWLEDGEMENTS

This work has been supported by the Technologiezentrum of Deutsche Telekom AG. The authors 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 to T. Fingscheidt.

REFERENCES

[1] P. Vary, K. Hellwig, R. Hofmann, R.J. Sluyter, C. Galand, and M. Rosso, "Speech Codec for the European Mobile Radio System,", in *Proc. Int. Conf. Acoust., Speech, Signal Processing, ICASSP*, New York, USA, 1988, pp. 227–230.

[2] ITU-T Recommendation G.729, "Coding of Speech at 8kbps using conjugate-structure algebraic-code-excited linear-prediction (CS-ACELP)".

[3] J. Paulus and J. Schnitzler, "16 kbit/s Wideband Speech Coding Based on Unequal Subbands,", in *Proc. Int. Conf. Acoust., Speech, Signal Processing, ICASSP*, Atlanta, Georgia, USA, 1996, pp. 255–258.

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

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

[6] 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.

[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] 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.

[10] 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.

[11] J. Schnitzler, "Lattice-Quantisierung der LP-Filterkoeffizienten auf der Basis der Line Spectral Frequencies (LSF),", *ITG Fachtagung Sprachkommunikation*, Frankfurt am Main, 1996.

[12] K.H. Paliwal and B.S. Atal, "Efficient Vector Quantization of LPC Parameters at 24 Bits/Frame", *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 1, pp. 3–13, January 1993.

[13] 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.

[14] 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.

[15] 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.