A 13.0 KBIT/S WIDEBAND SPEECH CODEC BASED ON SB-ACELP

Jürgen Schnitzler

RWTH Aachen, University of Technology
Institute of Communication Systems and Data Processing (IND), D-52056 Aachen, Germany
Juergen.Schnitzler@ind.rwth-aachen.de http://www.ind.rwth-aachen.de/~juergen

ABSTRACT

This paper describes a wideband (7 kHz) speech compression scheme operating at a bit rate of 13.0 kbit/s, i.e. 0.8 bit per sample. We apply a split-band (SB) technique, where the 0-6 kHz band is critically subsampled and coded by an ACELP approach. The high frequency signal components (6-7 kHz) are generated by an improved High-Frequency-Resynthesis (HFR) at the decoder such that no additional information has to be transmitted. In informal listening tests, the subjective speech quality was rated to be comparable to the CCITT G.722 wideband codec at 48 kbit/s.

1. INTRODUCTION

The interest in using wideband (50 ... 7000 Hz) speech and audio signals has grown within the last years. Compared to 'narrowband', i.e. telephone band limited signals, the larger signal bandwidth provides much more naturalness and intelligibility, and thus promises a significant quality improvement for telecommunication services. As a first wideband speech compression standard released in 1988, the CCITT G.722 [1] subband ADPCM scheme operates at bit rates of 48, 56 or 64 kbit/s (i.e. at effictive rates of 3-4 bit per sample).
Recently, ITU-T study group 16 has started a new standardization of a coding algorithm which is required to exhibit, at bit rates of 16, 24 and 32 kbit/s (1-2 bit per sample), a similar performance as the G.722 codec at its respective rates under most operating conditions [2]. The new codec aims at wireline applications such as ISDN wideband telephony, videoconferencing, and also at packet transmission applications as B-ISDN and 'multimedia' transmissions in the internet. In [3] we have proposed a split band coding scheme that fulfilled most of the requirements for speech at 16 kbit/s.
New applications for wideband speech will arise in the domain of mobile communications, which experienced a tremendous development during the last decade. Future interconnections between fixed and mobile networks and the increasing competition between their operators, e.g. in the Wireless Local Loop, will certainly excite a need for high quality services. Low rate wideband speech coding schemes (i.e. at effective rates of 0.5-1 bit per sample) may play an important role in this context. In ETSI SMG 11 the introduction of a wideband mode is currently being discussed for the forthcoming AMR (Adaptive Multi-Rate) codec standard [4], which shall replace the existing GSM codecs. In a previous proposal [5] we have introduced an algorithm that provided, at a rate well below 13 kbit/s, a similar clean speech quality as our original algorithm [3] for *clean speech* at 16 kbit/s.
In this paper, we present a modified scheme that shows an improved performance under both *clean speech and acoustic background noise conditions*. In the sequel, section 2 gives an overview of the general codec structure, whereas section 3 focusses on the core codec, an ACELP algorithm designed for the main 0-6 kHz subband signal. In section 4 we propose an improved high-frequency resynthesis of the 6-7 kHz band that does not require the transmission of any side information.

2. GENERAL CODEC STRUCTURE

Similarly to CCITT G.722, our basic approach is to split the input signal into two subbands, in order to allocate the available bit rate according to both the spectral distribution and the subjective importance of the subband components. An important difference is that we found an unequal splitting at a cutoff frequency of 6 kHz to be a more suitable solution [3]. This conclusion was motivated by an inspection of the instantaneous bandwidth of speech signals and by the spectral resolution of human perception: the 6-7 kHz band corresponds to about one critical band only.
In our configuration, those spectral portions of the upper subband (6-7 kHz) which are sufficient to convey a correct subjective impression of wideband speech can be represented either by coding them at a very low bit rate or even, as described in this paper, by extrapolation at the decoder side. Furthermore, this band splitting allows the lower subband (0-6 kHz) to be more efficiently quantized: at an overall target bit rate of 13 kbit/s, the effective bit rate increases from $\bar{R} \approx 0.8$ bit per sample at a sampling rate of $f_s = 16$ kHz to $\bar{R} \approx 1.1$ bit per sample at $f_s = 12$ kHz.
This suggests the use of state-of-the-art ACELP (Algebraic Code-Excited Linear Prediction) techniques for coding the lower subband. Currently the domain of toll-quality, medium rate *narrowband* speech codecs is dominated by algorithms based on ACELP, as they best fulfill the performance requirements in terms of subjective quality, complexity, robustness and delay. Examples of ACELP codecs are the GSM Enhanced Full Rate (EFR) codec [6] (12.2 kbit/s, i.e. $\bar{R} \approx 1.53$ bit per sample), the ITU-T G.729 universal 8 kbit/s codec [7] ($\bar{R} = 1$ bit per sample) and its extensions, or the IS-641 standard [8] (7.4 kbit/s or $\bar{R} \approx 0.93$ bit per sample) for the US-TDMA system.
Figure 1 a) shows the encoder structure of our proposal. A rate conversion module extracts the 0-6 kHz lower subband from the input wideband (7 kHz) signal and reduces the

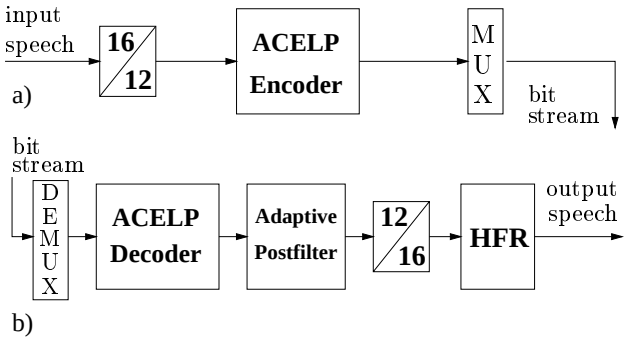

Figure 1: a) Structure of wideband encoder
b) Structure of wideband decoder

sampling rate from 16 kHz to 12 kHz using a linear phase analysis filter.
The ACELP core codec operates on speech frames of 20 ms (240 samples at $f_s = 12$ kHz). For every frame, a total of 260 bits is transmitted over the channel, including 8 bits of protection information that can be used for error concealment in the decoder. The resulting bit allocation is shown in Table 1. Details of the ACELP configuration are given in the next section.
The decoder structure is shown in Figure 1 b) and will be refined in section 4. The received bits are used in the ACELP decoder to synthesize the lower band signal. A postfilter is applied to the signal in order to enhance the perceptual quality. The receiver rate conversion module interpolates the postfilter output to the original sampling rate of 16 kHz. Both the decimation (transmitter) and the interpolation (receiver) filter contribute a delay of 2.5 ms each. In conjunction with the framing, the overall algorithmic delay therefore amounts to 25 ms.
Finally a High-Frequency-Resynthesis (HFR) module generates an upper band (6-7 kHz) signal portion. As it will be described in section 4, the regenerated upper band signal consists of a gain-amplified and filtered bandpass noise. All necessary parameters are solely adapted based on the received lower band parameters and the use of a priori knowlegde of the input speech.

Parameter	Bit Allocation	Bits/Frame	Bit Rate
LPC		32 bit	1.6 kbit/s
ACB-Index	2×(8+6) bit	28 bit	2.2 kbit/s
ACB-Gain	4× 4 bit	16 bit	
FCB-Index	8×18 bit	144 bit	8.8 kbit/s
FCB-Gain	8× 4 bit	32 bit	
Parity		8 bit	0.4 kbit/s
Σ			13.0 kbit/s

Table 1: Bit allocation of the proposed codec

3. ACELP CODING OF 0-6 KHZ BAND

3.1. Short-term LP analysis

The linear prediction (LP) analysis uses a modified Split-Levinson approach as described in [9] to compute the Line Spectral Frequencies (LSF) from the windowed speech signal. The analysis window covers 300 samples and is right aligned with the current 20 ms frame, i.e. no lookahead is used. The order of the LP filter is $p_{lb} = 14$. Before computing the LSF coefficients, the autocorrelation matrix is weighted using a binomial window, providing an additional amount of bandwidth expansion to the LP filter.
The 14 LSF parameters are quantized by a Predictive Multistage Split Vector Quantizer scheme. For the prediction of the LSF vector, a Moving Average (MA) model of order 4 is used. The closed-loop residual quantizer consists of two stages of split vector quantizers, using 2 segments for the first and 3 segments for the second stage, respectively. One of two fixed predictor sets can be chosen. This approach resulting into an overall bit rate of 32 bit per frame is similar to the one used for the ITU-T G.729 codec [7].
In addition, a linear interpolation of the LP filter coefficients is performed in the LSF domain every 5 ms.

3.2. Long-term prediction analysis

Every 5 ms, the long-term prediction (LTP) is carried out in a combination of open-loop and closed-loop LT-analysis based on an adaptive codebook (ACB) representation (see [5]). The ACB delays in the four LTP subframes are coded by 8+6+8+6=28 bits. In the lower delay range a fractional pitch approach is used. The ACB gains are nonuniformly quantized with 4 bits each.

3.3. Fixed codebook (FCB) excitation

Every 2.5 ms (30 samples), an excitation shape vector is selected from a sparse algebraic pulse codebook. An innovation vector contains 4 nonzero pulses, as shown in Table 2. The pulses 1 and 2 can take one of 16 possible positions, the pulses 3 and 4 one of 8 positions. Since each pulse can have an individual sign, 18 bits are necessary to encode the shape vector. Note that the pulses 1 and 3 as well as pulses 2 and 4 may share the same position, and that all pulses can fall outside the valid range of positions 0 ... 29. This allows a variable number of pulses and pulse amplitudes of $0, \pm 1, \pm 2$.
The codebook structure and the efficient focussed search method are based on [7]. The FCB gain is quantized using a fixed autoregressive predictor in order to reduce the dynamic range [10]. The residual of the gain predictor is nonuniformly scalar quantized with 4 bits.

3.4. Perceptual weighting

The perceptual weighting filter $W(z)$ applied during the optimization processes of the ACB and FCB search has a

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

Table 2: 18-bit sparse algebraic pulse codebook

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 computed from the unquantized, interpolated LSF parameters. Different sets of weighting factors $\{\gamma_1, \gamma_2\}$ are used for the adaptive and fixed codebook search. For the fixed codebook search, the parameters are adapted with respect to the tilt and the strength of resonances of the LP synthesis filter [7].

3.5. Adaptive postfilter

As described in [11], the adaptive postfilter consists of a cascade of a formant postfilter, an harmonic postfilter and a tilt compensation filter. The postfilter is updated every LTP subframe (5 ms). The formant postfilter uses the transmitted LPC filter coefficients. After postfiltering, an adaptive gain control is performed.

4. HIGH FREQUENCY RESYNTHESIS

For speech, spectral components above 6 kHz are almost always due to unvoiced, i.e. fricative, sounds. Informal listening tests showed that the presence of this spectral band is still very well perceivable. However, a sufficient subjective quality does not require an exact reproduction of the noise-like signal waveform. In [5] we have demonstrated a very simple and efficient spectral folding technique to regenerate an upper band signal. This approach exploited the observation that the spectral distributions in the 5-6 kHz and 6-7 kHz bands are very similar.
On the other hand, many operating conditions typically include the presence of acoustic background noise. In such situations of non-speech signal components, our previous proposal sometimes revealed perceivable degradations. Techniques as proposed in [12] do not yield the intended quality, either.
In this paper, we describe a more elaborate scheme based on High-Frequency Resynthesis (HFR) techniques that have been initially studied for the extension from telephone-band to wideband speech [13].

4.1. Resynthesis of the 6-7 kHz band

Similarly to [3], we model the upper band signal by a bandpass (6-7 kHz) noise excitation whose magnitude spectrum has to be shaped properly (see Figure 2). The basic idea is to separately extrapolate the spectral envelope and the residual of the signal. Typically there exists a good correlation between the lower band spectral envelope and the spectrum of the upper band. Since no side information shall be transmitted, the task is to predict the spectral shape and the energy of this excitation from the received lower band parameters.
The overall spectral shape of the output wideband signal is determined by selecting an appropriate LPC synthesis filter $1/A_{Hfr}(z)$. Provided that this filter matches the synthesized lower band spectrum, it is expected that its behaviour above 6 kHz will reflect the original upper band speech components. $1/A_{Hfr}(z)$ is determined from a codebook $\mathcal{C}_{wb}$ describing N_{Hfr} LPC filters that are computed at $f_s = 16$ kHz and stored in their LSF representation [15]. The decoded and interpolated lower band signal is first filtered by $A_{Hfr}(z)$, such that the gain-adapted noise excitation for the upper band is added in the residual domain, before the sum is again filtered using $1/A_{Hfr}(z)$. This implies that, regardless of the choice of $A_{Hfr}(z)$, the HFR module does not introduce any additional degradation to the lower band. The spectral fit of the filter to the actual lower band signal does not have to be very exact and the number of stored filter parameters can be limited. We have found $N_{Hfr} = 48$ different LSF sets of order $p_{wb} = 16$ to be sufficient.
For the selection of $A_{Hfr}(z)$, a second codebook $\mathcal{C}_{lb}$ is necessary: it contains N_{Hfr} LSF vectors describing the spectral envelope of the lower band at a sampling frequency of 12 kHz. For each LSF vector $\underline{\omega}_{wb,\nu} \in \mathcal{C}_{wb}$ ($\nu = 0 \ldots N_{Hfr} - 1$), the associate vector $\underline{\omega}_{lb,\nu} \in \mathcal{C}_{lb}$ approximates the lower band part of the spectrum given by $\underline{\omega}_{wb,\nu}$. Thus the selection process can be understood as re-quantizing the lower band LPC filter, $A(z)$, in $\mathcal{C}_{lb}$ and looking up the LSF parameters for $1/A_{Hfr}(z)$ in the 'shadow' codebook $\mathcal{C}_{wb}$. The found HFR filter defined by $1/A_{Hfr}(z)$ is linearly interpolated every 5 ms.
The adaptation of the HFR gain g_{Hfr} is performed in the residual domain. Assuming that the inverse HFR filter $A_{Hfr}(z)$ yields a rather flat spectrum in the 0-6 kHz frequency range, the bandpass noise d_{ub} is adjusted in order to adopt the same power spectral density level in the 6-7 kHz range. Since d_{lb} is not completely decorrelated, better results are obtained when using a high-pass (5 kHz) filtered portion for the gain adaptation. This scaling is updated every 5 ms, and the resulting gains are smoothed on a sample-by-sample basis.
The described HFR method serves to achieve a more transparent subjective quality than our previous spectral folding approach. In particular, the performance in background noise conditions has been improved.

4.2. Design of HFR codebooks

To obtain the HFR codebooks $\mathcal{C}_{lb}$ and $\mathcal{C}_{wb}$, an approach close to the Linde-Buzo-Gray (LBG) algorithm [14] is chosen [15]. Prior to the training phase, an initial codebook $\mathcal{C}^0_{lb}$ is required for the partioning of the p_{lb}-dimensional vector space filled by all possible lower band LSF parameter sets. $\mathcal{C}^0_{lb}$ contains N_{Hfr} LSF vectors $\underline{\omega}^0_{lb,\nu}$ ($\nu = 0 \ldots N_{Hfr} - 1$) and is obtained by applying the LBG algorithm to the lower band portion of the training speech data.
During the training process, for each 20 ms frame λ an LPC analysis (order p_{wb}) is performed on the wideband input speech ($f_s = 16$ kHz); thus, an LSF vector $\underline{\omega}_{wb}(\lambda)$ is computed. In parallel, the lower band signal portion ($f_s = 12$ kHz) of frame λ is subject to a second LPC analysis (order p_{lb}), yielding an LSF vector $\underline{\omega}_{lb}(\lambda)$. Using $\mathcal{C}^0_{lb}$, the current frame's parameters $\underline{\omega}_{lb}(\lambda)$ and $\underline{\omega}_{wb}(\lambda)$ are assigned to the sets $\mathcal{P}_{lb}(\nu)$ and $\mathcal{P}_{wb}(\nu)$, respectively. $\mathcal{P}_{lb}(\nu)$ and $\mathcal{P}_{wb}(\nu)$, $\nu = 0 \ldots N_{Hfr} - 1$, define the partioning of the vector spaces containing the LSF parameters $\underline{\omega}_{lb}(\lambda)$ and $\underline{\omega}_{wb}(\lambda)$. This assignment is achieved by searching $\mathcal{C}^0_{lb}$ and selecting ν such that an inverse LPC filter, built from $\underline{\omega}^0_{lb,\nu} \in \mathcal{C}^0_{lb}$ and applied to the lower band speech, yields the minimum mean squared prediction error.
After processing all frames of training data, the final code-

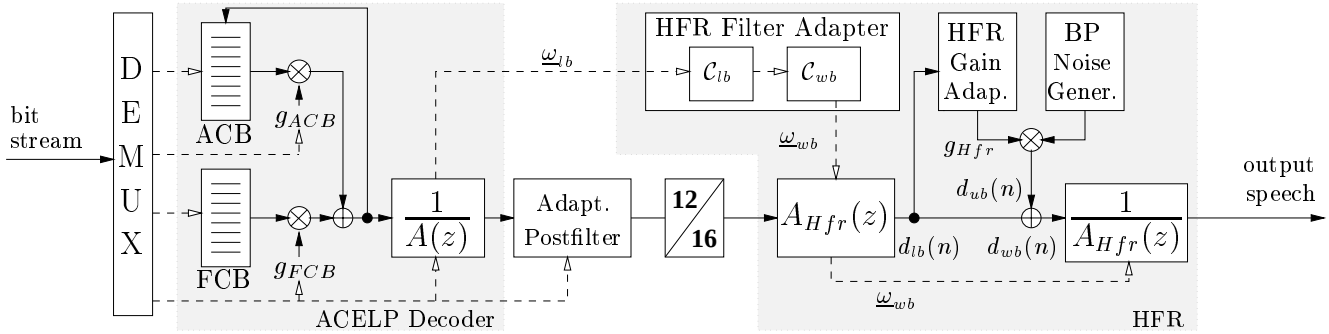

Figure 2: Wideband speech decoder details: ACELP decoder and High-Frequency Resynthesis (HFR)

vectors of $\mathcal{C}_{lb}$ and $\mathcal{C}_{wb}$ are found as the centroids of partions $\mathcal{P}_{lb}(\nu)$ and $\mathcal{P}_{wb}(\nu)$, $\nu = 0 \ldots N_{Hfr} - 1$, respectively. This procedure ensures to produce pairs of lower band and wideband LSF parameters having a good spectral fit in the 0-6 kHz range. Furthermore, the stability of the resulting filters is guaranteed [15].
It can be noted that, in order to save memory in a practical implementation, the codebook $\mathcal{C}_{wb}$ may be directly linked to the high-resolution LSF quantizer of the lower band ACELP decoder, instead of explicitely storing $\mathcal{C}_{lb}$.

5. CONCLUSION

In this paper an SB-ACELP encoding scheme for 13.0 kbit/s wideband speech encoding has been presented. The algorithm is based on a split band (SB) structure. A state-of-the-art ACELP codec operating at a 12 kHz sampling frequency is used to transmit the 0-6 kHz subband signal. An LPC-based High-Frequency-Resynthesis technique has been successfully applied to fill the perceptually significant upper 6-7 kHz band on the decoder side, without the need to transmit any side information. By informal listening tests the speech quality was judged to be comparable to the CCITT G.722 wideband codec operating at 48 kbit/s.

6. REFERENCES

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

[2] ITU-T SG 16 Q.20, "Terms of Reference for the ITU-T Wideband (7 kHz) Speech Coding Algorithm," April 1997.

[3] J. Paulus und J. Schnitzler, "16 kbit/s Wideband Speech Coding Based on Unequal Subbands" in *Proc. Int. Conf. Acoust., Speech, Signal Processing, ICASSP*, (Atlanta, Georgia, USA), pp. 651–654, 1996.

[4] ETSI SMG11, "Draft Adaptive Multi-Rate (AMR) Study Phase Report." Version 0.4, Tdoc SMG11-AMR 128/97, August 1997.

[5] J. Paulus und J. Schnitzler, "Wideband Speechcoding for the GSM Fullrate Channel?" in *Proceedings ITG-Fachtagung "Sprachkommunikation"*, (Frankfurt am Main), pp. 11–14, 1996.

[6] ETSI/TC SMG, "Recommendation GSM 06.60: Enhanced Full Rate Rate Speech Transcoding" European Telecommunications Standards Institute, Januar 1996.

[7] CCITT/ITU-T, "Rec. G.729: Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP)" in *General Aspects of Digital Transmission Systems; Terminal Equipments, Series G Recommendations*, International Telecommunication Union, 1996.

[8] T. Honkanen, J. Vainio, K. Järvinen, P. Haavisto, R. Salami, C. Laflamme, und J.-P. Adoul, "Enhanced Full Rate Speech Codec for IS-136 Digital Cellular System" in *Proc. Int. Conf. Acoust., Speech, Signal Processing, ICASSP*, (Munich, Germany), pp. 731–734, IEEE, 1997.

[9] S. Saoudi und J. Boucher, "A New Efficient Algorithm to Compute the LSP Parameters for Speech Coding" *Signal Processing*, vol. 28, pp. 201–212, 1992.

[10] R. Salami, C. Laflamme, J. Adoul, und D. Massaloux, "A Toll Quality 8 Kb/s Speech Codec for the Personal Communications System (PCS)" *IEEE Transactions on Vehicular Technology*, vol. 43, pp. 808–816, August 1994.

[11] J. Chen und A. Gersho, "Adaptive Postfiltering for Quality Enhancement of Speech" *IEEE Trans. Speech and Audio Processing*, vol. 3, pp. 59–71, January 1995.

[12] J. Makhoul und M. Berouti, "High-Frequency Regeneration in Speech Coding Systems," in *Proc. Int. Conf. Acoust., Speech, Signal Processing, ICASSP*, (Washington, DC), pp. 428–431, IEEE, 1979.

[13] H. Carl, *Untersuchung verschiedener Methoden der Sprachcodierung und eine Anwendung zur Bandbreitenvergrößerung von Schmalband-Sprachsignalen*. PhD thesis, Ruhr-Universität Bochum, 1994.

[14] Y. Linde, A. Buzo, und R. Gray, "An Algorithm for Vector Quantizer Design" *IEEE Transactions on Communications*, vol. 28, pp. 84–95, January 1980.

[15] J. Kenkenberg, *"Untersuchungen zur künstlichen Bandbreitenvergrößerung von Sprachsignalen"*. Diploma thesis D25/96, Institut für Nachrichtengeräte und Datenverarbeitung, IND, RWTH Aachen, 1996.