A 16, 24, 32 KBIT/S WIDEBAND SPEECH CODEC BASED ON ATCELP

Pierre Combescure[‡], *Jürgen Schnitzler*[*], *Kyrill Fischer*[†], *Ralf Kirchherr*[†],
Claude Lamblin[‡], *Alain Le Guyader*[‡], *Dominique Massaloux*[‡],
Catherine Quinquis[‡], *Joachim Stegmann*[†], *Peter Vary*[*]

[*]Aachen University of Technology (RWTH), IND, D-52056 Aachen, Germany
[†]Deutsche Telekom Berkom GmbH, D-64276 Darmstadt, Germany
[‡]France Telecom CNET, DIH/DIPS, F-22307 Lannion Cedex, France

ABSTRACT

This paper describes a combined Adaptive Transform Codec (ATC) and Code-Excited Linear Prediction (CELP) algorithm, called ATCELP, for the compression of wideband (7 kHz) signals. The CELP algorithm applies mainly to speech, whereas the ATC mode is selected for music and noise signals. We propose a switching scheme between CELP and ATC mode and describe a frame erasure concealment technique. Subjective listening tests have shown that the ATCELP codec at bit rates of 16, 24 and 32 kbit/s achieved performances close to those of the CCITT G.722 at 48, 56 and 64 kbit/s, respectively, at most operating conditions.

1. INTRODUCTION

The ATCELP coder [1] is a combination of an ATC and a CELP coding algorithm, where the CELP mode is used mainly for speech signals and the ATC mode for music and stationary background noise signals. The codec operates at bit rates of 16, 24 and 32 kbit/s. For the bit rate of 16 kbit/s, the transcoded signal bandwidth is reduced to 5 kHz only. For 32 kbit/s, only the ATC mode of the ATCELP coder is selected.

The general structure of the ATCELP codec is shown in Figure 1. The switch between the two coding techniques (ATC and CELP) is controlled by a signal classifier which works exclusively on the input signal. The chosen mode of the signal classifier is transmitted as a side information to the decoder. In section 2 and 3, we describe the particular CELP and ATC coding algorithms, respectively, before we focus on the ATCELP switching scheme in section 4. Furthermore, attention has been put on the error concealment of erased frames (section 5).

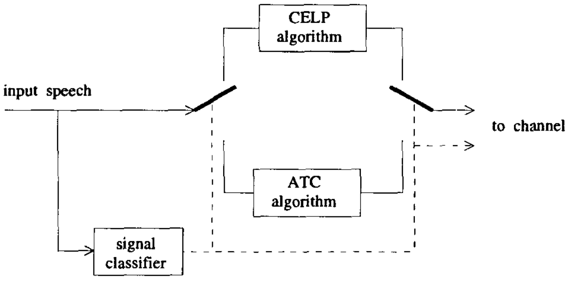

Figure 1: Overview of ATCELP encoder

The ATCELP codec has been developed in accordance to the ITU-T Q.20/16 requirements [2] for a new wideband coding standard for ISDN, video-conferencing and multimedia applications. Section 6 reports on the subjective listening tests performed on the ATCELP codec with reference to the CCITT G.722 codec.

2. SB-CELP CODING FOR SPEECH SIGNALS

For speech signals at 16 and 24 kbit/s, the ATCELP codec mainly operates in CELP mode. The Subband-CELP (SB-CELP) algorithm is based on a split-band scheme [3] using two unequal subbands from 0-5 kHz and 5-7 kHz. A block diagram of the encoder and decoder is shown in Figure 2. The analysis filter bank performs unequal subband splitting of the wideband input signal and critical subsampling of the two subbands [4]. At the decoder, the synthesis filter bank interpolates and superposes the decoded lower and upper band signals, yielding the wideband output signal. The delay of the analysis/synthesis filter bank amounts to 10 ms. The bit allocation for the CELP mode is shown in Table 1.

For encoding the decimated **lower band** (LB, 0-5 kHz), Algebraic Code-Excited Linear Prediction (ACELP) is used. The short term (LP) synthesis filter coefficients are updated every 20 ms frame (200 samples at f_s=10 kHz). A look-ahead of 5 ms is used within the autocorrelation analysis. The quantization of the 12 LP parameters is performed in the LSF (Line Spectral Frequencies) domain and based on predictive multistage split vector quantization using 33 bits. Every 10 ms, an open-loop pitch estimate is calculated. Using this estimate, a voicing decision is taken and coded by 1 bit. Provided the 10 ms subframe is declared voiced, a constrained closed-loop adaptive codebook (ACB) search with fractional delays is performed every 5 ms [5]. This procedure requires 8+6=14 bit per 10 ms for coding the pitch lags. Every 5 ms ACB-subframe, the pitch gain is nonuniformly quantized with 4 bit.

At 16 kbit/s, an excitation shape vector is selected from a

0-7803-5041-3/99 $10.00 © 1999 IEEE

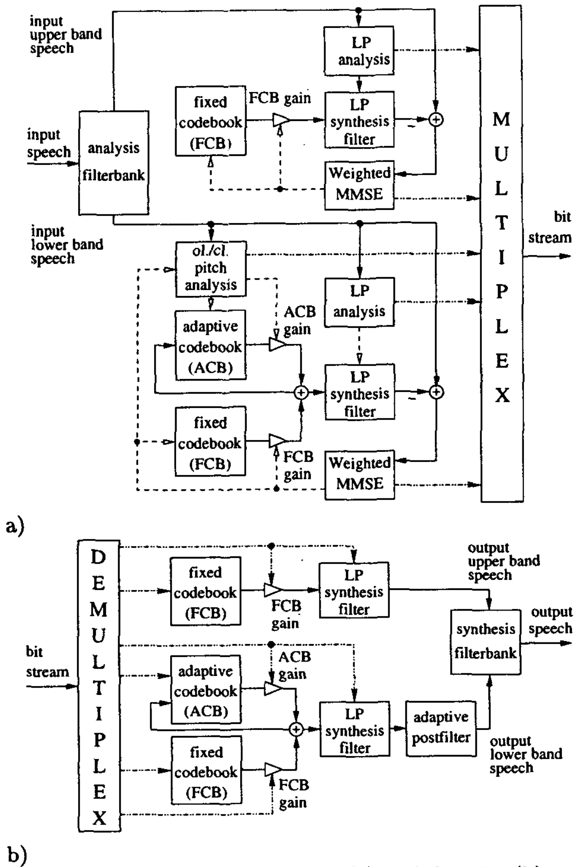

Figure 2: SB-CELP encoder (a) and decoder (b)

ternary sparse codebook („pulse codebook") [5] every 2.5 ms (25 samples). Depending on the voicing mode, i.e. the available bit rate, an innovation vector contains 4 or 5 tracks with a total maximum of 10 nonzero pulses, resulting in 25 or 30 bits to encode a shape vector. The FCB gain is encoded using fixed interframe recursive prediction with the prediction residual being nonuniformly quantized with 4 or 5 bits. At 24 kbit/s, every 1 ms (10 samples) an excitation shape vector is selected from either a pulse codebook similar to the 16 kbit/s mode or a binary algebraic VSELP codebook [6]. Depending on the voicing mode, an innovation shape vector requires 12 or 14 bits for encoding. A shape vector of the pulse codebook contains 2 tracks with a total maximum of 2 or 3 nonzero pulses. Both codebooks are searched for the optimum innovation, and that codebook is selected which minimizes the reconstruction error. For each FCB subframe, the FCB mode is transmitted by 1 bit. Every FCB gain is allocated 3 or 4 bits for quantization.

In the ACB and FCB search processes, a perceptual weighting filter is used [3]. At the decoder, adaptive postfiltering is applied to the synthesized lower band speech.

The **upper band** (UB, 5-7 kHz) is not transmitted at 16 kbit/s. For the 24 kbit/s mode, a frame of the decimated upper subband (80 samples at f_s=4 kHz) is encoded by a CELP algorithm technique. Based on a Burg covariance approach, a short term (LP) synthesis filter of order 8 is computed and vector quantized with 10 bits. In subframes of length 16, a stochastic Gaussian codebook with 256 entries is searched for the innovation shape vectors according to a WMSE criterion. The codebook gain is encoded by fixed interframe recursive prediction and quantized with 3 bits.

	16 kbit/s	**24 kbit/s**
ATC/CELP	2	2
LB LPC	33	33
LB Voicing	2	2
LB ACB	22/0+22/0	22/0+22/0
LB FCB	116/136+116/136	162/185+162/185
UB LPC	-	10
UB FCB	-	55
Error prot.	7 ... 11	8 ... 10
Total # bits	320	480

Table 1: Bit allocation for a frame (20 ms): CELP mode at 16 and 24 kbit/s

3. ATC CODING FOR MUSIC SIGNALS

For music at 16 and 24 kbit/s, and for any signal at 32 kbit/s, the ATCELP codec operates in ATC mode. The ATC coder presented here is based on a MDCT transform, that exploits psychoacoustical results by the use of masking curves calculated in the transform domain and expressed on the Bark scale. The principle of the ATC encoder is depicted by Figure 3 a) and follows the scheme described in [7]. For each 20 ms frame, the 320 MDCT transform coefficients are calculated, with a window overlapping two successive frames. A tonality detector and a voiced/unvoiced detector transmit to the decoder information on the tonal/non tonal and voiced/unvoiced nature of the input signal. A first masking curve is calculated and coefficients below the mask minus a given threshold are cleared. The spectrum envelope of the current frame is estimated, divided into 32 bands nonuniformly spaced along the frequency axis. The energies of the bands are quantized and encoded using entropy coding, the quantizers and Huffman codes depending on the voiced/unvoiced and tonality indices. Then for the not fully masked bands a dynamic allocation of the bits for the coefficients encoding is performed both by the encoder and the decoder. This avoids transmitting any information on the bit allocation. This procedure (see [7]) uses a second masking curve and relies on the decoded spectrum envelope.

The transform coefficients are then quantized using the decoded spectrum envelope to reduce the quantizers dynamic range. Two types of quantizers (scalar and vector quantizers), have been designed: In the scalar case, masked coefficients receive the null value, which is allowed by the use of symmetric quantizers including zero as reconstruction level. The quantizers having an odd number of levels, a packing procedure encodes jointly the indices corresponding to the scalar quantized coefficients. The VQ employs embedded algebraic codebooks designed for various sizes and dimensions. For a given dimension, the codebooks are composed of the union of permutation codes, all signs combinations

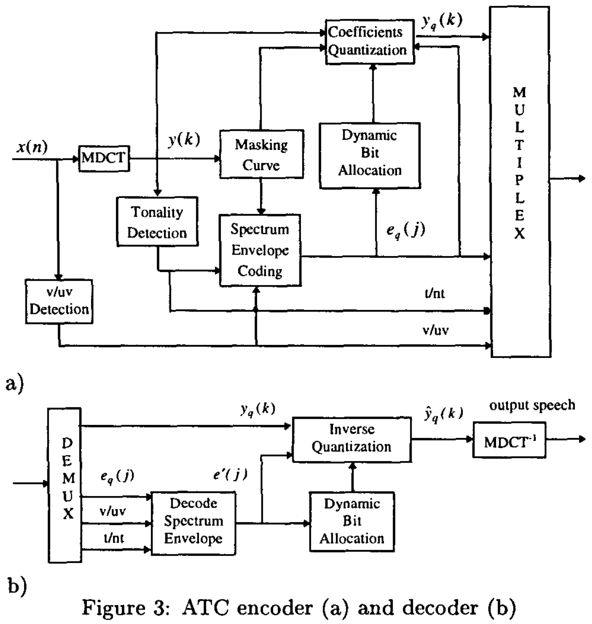

Figure 3: ATC encoder (a) and decoder (b)

being possible. An optimal fast search (cf. [8]) that takes advantage of the codes' structure has been designed. For the ATCELP purpose, a local decoding is included, similar to the distant decoding (see Figure 3 b)). The decoded coefficients are inverse transformed to produce the output signal.

4. ATC/CELP SWITCH

The ATC/CELP-switch, which is used in this coder for bit rates of 16 and 24 kbit/s, is designed in such a way that the coder favours the CELP mode for non stationary signals such as speech and the ATC mode for stationary signals (as is most often the case for music). The switch between the two coding techniques (CELP and ATC) is controlled by a signal classifier, which works exclusively on the input signal. The chosen mode of the signal classifier (i.e. „continue CELP", „transition ATC to CELP", „transition CELP to ATC", „continue ATC") is transmitted as a side information to the decoder. With respect to the bad frame handling, two mode bits are necessary for this purpose.
The **classifier**, which is inspired by an approach used in switched forward/backward adaptive LPC [9], consists of three functional blocks:

1. calculation of input parameters,
2. calculation of the stationarity measure and
3. checking procedure.

The input parameters for the stationarity measure are two prediction gains, based on a lower order LPC analysis of the current speech frame and on a higher order LPC analysis of the previous input frames. An additional input parameter for the stationarity measure is the difference between the LSF coefficients of the previous and the current speech frame. The following calculation of the stationarity measure is based on the short-term and long-term difference of the prediction gains and the difference of the LSF coefficients. Special conditions are checked for noisy speech. The calculated stationarity measure is then used as an indicator for the current frame being either speech or music. Finally a checking procedure is performed to examine if a transition of one codec mode to another leads to a smooth output signal at the decoder. If it is likely that switching will lead to an audible degradation, the switching decision will be delayed to the next frame.
For switching between the two coding techniques a special transition frame is needed to avoid discontinuities in the synthesized output signal: If the classifier decides to perform a **transition from ATC to CELP** at frame n, the n-th frame is the last frame to be computed by the ATC algorithm using a modified window function (see Figure 4. This modified window function, used for frames n and $(n+1)$, is set to zero for the last 5 ms. This enables the coder to decode the leading 5 ms of frame $(n+1)$, which would otherwise cause time domain aliasing effects without the ATC coefficients of the next frame. In the $(n+1)$th frame, where the CELP mode is performed for the first time, only the last 5 ms can be decoded by the CELP coder (caused by filterbank delay). Therefore, 10 ms of the speech signal in this frame (see the shaded section in Figure 4) have to be extrapolated by extending the residual signal of the previous synthesized output frames periodically based on the pitch lag, followed by LPC synthesis filtering. This is also done in the checking procedure of the classifier described above, but using the input signal; i.e. if the extrapolated input signal is very similar to the original input signal, the probability of a smooth transition is high and the switch can be performed. The **transition from CELP to ATC** is done using a similar window function like the one used at the ATC to CELP transition, but reversed in time.

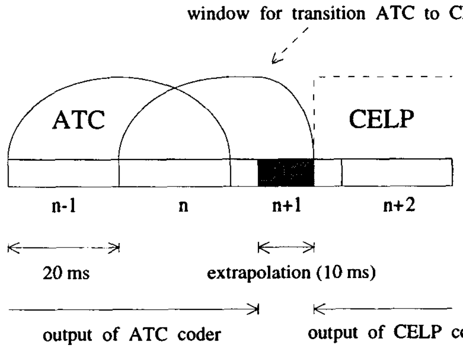

Figure 4: ATCELP coder: transitions from ATC to CELP

5. BAD FRAME HANDLING

If a frame erasure occurs and the last frame was processed in mode „continue CELP" , then the CELP-mode will be kept for this frame. Otherwise, the erased frame will be handled like an erased ATC frame. If a frame indicating a transition from ATC to CELP is erased, the ATC bad frame handling (ATC-BFH) will be used. However, since the following non-erased frame is already a CELP frame, a signal extrapolation covering 15 ms has to be performed. On the other hand, if a frame indicating a transition from CELP to ATC is erased, the CELP-BFH will be used. Upon the detection of the following non-erased frame, which is in ATC mode, an extra ATC-BFH has to be performed in order to enable the decoding of the non-erased ATC frame. In case of a detected **frame erasure during CELP mode**, the LP synthesis filter of the previous frame is re-used. Based on a voiced/unvoiced decision of the previous frame, either a pitch-synchronous or an asynchronous extrapolation of the previous excitation is constructed and used for synthesizing the signal in the current, lost frame. For subsequent lost frames, an attenuation of the excitation is performed. When a **frame erasure during ATC mode** is detected, the output signal is extrapolated. The treatment differs for the first erased frame and the following successive frames. For the first erased frame, a 14th order LPC analysis is performed on the past synthesized decoded speech, and if the past frame was tonal or voiced, a simplified LTP analysis yields a pitch period. Those estimated parameters are kept for the next successive erased frames. The extrapolated signal samples are obtained by LPC filtering an excitation signal generated from the calculated past residual signal, using pitch periodicity in the voiced and tonal cases or a simple copy else. The energy of the synthetic excitation is controlled by an adaptive gain control procedure and tends towards an estimation of the residual energy that depends on the local stationarity of the past residual.

6. CODEC PERFORMANCE

For the ITU-T Q.20/16 qualification phase, subjective listening tests according to [2] have been performed. In [1], the results are described in detail. In an ACR experiment, the 16 and 24 kbit/s modes were tested for various input levels, single and multiple transcodings and frame erasures with clean speech and music as input signals. A further DCR experiment tested the codec for speech with office, babble and frequency harmonic noise and interfering talkers at all bit rates. For all conditions, the CCITT G.722 codec at 48, 56, and 64 kbit/s was used as a reference. The ATCELP codec achieved its goal to provide a good overall quality for both speech and music signals. The main problems were related to the tandem conditions (2 and 3 tandems at 16 and 24 kbit/s). Apart from some isolated points such as high input levels and office noise at 16 kbit/s, the codec met all requirements for the 32 kbit/s rate and for the single transcoding of speech (with and without background noise) and music at 16 and 24 kbit/s.

7. CONCLUSION

A new codec structure called ATCELP, which is based on a combination of Subband-CELP and ATC techniques, has been proposed for the coding of wideband speech and music at 16, 24, and 32 kbit/s. In this paper, we have reported on a switching scheme between the ATC and CELP modes and described a bad frame handling procedure. Subjective listening tests have shown that the ATCELP approach is appropriate to cope with both speech and music signals at all bit rates and provides a high quality close to the ITU-T Q.20/16 requirements. The remaining problems, especially for multiple transcodings, are expected to be overcome by improvements in the ATC/CELP switching procedure and re-optimizations of the quantization procedures.

8. ACKNOWLEDGEMENTS

The authors wish to thank the DT Berkom and FT CNET speech quality groups, in particular Dominique Pascal and Wolfgang Koch, for conducting the subjective listening tests.

9. REFERENCES

[1] Deutsche Telekom and France Telecom/CNET, "High-level description / Subjective qualification test results for the ITU-T wideband (7kHz) ATCELP speech coding algorithm of Deutsche Telekom, Aachen University of Technology (RWTH) and France Telecom (CNET)" Delayed contributions COM16-129/130, ITU-T Study Group 16, Q.20, Geneva, January 1998.

[2] ITU-T SG 16 Q.20, "Terms of Reference for the ITU-T Wideband (7 kHz) Speech Coding Algorithm," September 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] J. Schnitzler und M. Kappelan, "On Nonuniform Filter Banks for Subband Speech Coding and Their Efficient Implementation" in *Proc. ITG-Fachtagung "Sprachkommunikation"*, (Dresden), pp. 73–76, September 1998.

[5] R. Salami, C. Laflamme, J. Adoul, A. Kataoka, S. Hayashi, T. Moriya, C. Lamblin, D. Massaloux, S. Proust, P. Kroon, und Y. Shoham, "Design and description of CS-ACELP: A Toll Quality 8 kb/s Speech Coder" *IEEE Trans. Speech and Audio Processing*, vol. 6, pp. 116–130, March 1998.

[6] A. Le Guyader, C. Lamblin, und E. Boursicaut, "Embedded Algebraic CELP/VSELP Coders for Wideband Speech Coding" *Speech Communication*, vol. 16, pp. 319–328, 1995.

[7] Y. Mahieux und J. Petit, "High Quality Transform Coding of Speech at 64 kbps" *IEEE Trans. Communications*, vol. 42, pp. 3010–3019, November 1994.

[8] C. Lamblin, *Quantification vectorielle algébrique sphérique par le réseau de Barnes-Wall. Application au codage de la Parole.* PhD thesis, University of Sherbrooke, March 1988.

[9] S. Proust, C. Lamblin, und D. Massaloux, "Dual Rate Low Delay CELP Coding (8 kbit/s 16 kbit/s) using a Mixed Backward/Forward Adaptive LPC Prediction" in *Proc. IEEE Workshop on Speech Coding for Telecommunications*, (Annapolis, Maryland, USA), pp. 37–38, September 1995.