A 6.1 TO 13.3-KB/S VARIABLE RATE CELP CODEC (VR-CELP) FOR AMR SPEECH CODING

S. Heinen, M. Adrat, O. Steil, P. Vary

Inst. of Comm. Systems and Data Processing
Aachen University of Technology
heinen@ind.rwth-aachen.de
marc@ind.rwth-aachen.de

W. Xu

Dept. of Mobile Phone Development, Siemens AG
Hofmannstr. 51, 81359 Munich
wen.xu@pn.siemens.de

ABSTRACT

We propose a new 6.1 to 13.3-kb/s speech codec called variable rate code-excited linear prediction (VR-CELP) for Adaptive Multi-Rate (AMR) transmission over mobile radio channels such as GSM or UMTS.

The AMR concept allows to operate with almost wireline speech quality for poor channel conditions and better quality for good channel conditions. This is achieved by dynamically splitting the gross bit rate of the transmission system between source and channel coding according to the current channel conditions. Thus the source coding scheme must be designed for seamless switching between rates without annoying artifacts.

To enhance the transmission quality under very poor channel conditions, a new powerful error concealment strategy based on estimation theory is applied.

1. INTRODUCTION

In digital mobile radio systems, due to the powerful combination of equalization, interleaving and channel coding, an almost error-free transmission is achieved down to a certain threshold of the carrier to interferer ratio (C/I). If this threshold is threspassed, the error controlling code will fail with the result that residual errors possibly cause very annoying artifacts in the reconstructed speech signal.

Therefore, in current systems a worst case design is applied where the channel coding is powerful enough to remove most transmission errors as long as the system operates within a reasonable C/I-range. Yet, the drawback of this solution is a lower speech quality than achievable for good channel conditions, since a large amount of the gross bit rate is consumed by the (in this case) over-scaled channel coding.

The AMR concept solves the problem in a more intelligent, i.e. *adaptive* way. The ratio between net bit rate and error protecting redundancy is adaptively chosen according to the current channel conditions. While the channel is bad, the speech encoder operates at low bit rates thus allowing powerful forward error control. In turn, for good channel conditions the speech encoder may use its highest net rate implying high speech quality, as in this case weak error protection is sufficient.

Despite the inherent robustness of the AMR concept itself there is still the necessity to conceal residual errors that are not corrected by the forward error control code. By means of error concealment the speech quality under very bad channel conditions can significantly be increased since annoying sound effects are removed or muted. For error concealment we employ an approach recently proposed in [1], which is strictly based on estimation theory. We advance this approach by adding explicit redundancy to the transmitted parameters. In this way the performance of the applied error concealment technique can drastically be increased.

The paper is organized as follows: First we give a general description of the speech codec modules before focussing on the rate adaptation capabilities. In section 3 we describe the concealment techniques. Finally we shall state results of the performance tests and draw the possible conclusions.

2. VR-CELP SPEECH CODING

A simplified block diagram of the VR-CELP speech encoder is depicted in Figure 1. The basic CELP [2] encoder operation is as follows: The sampled input speech signal $s(n)$ is partitioned into segments of 20 ms (160 speech samples) duration and a linear predictor is computed for each speech segment. The coefficients of this predictor are used to build an LPC synthesis filter $\frac{1}{1-A(z)}$ describing the spectral envelope of the signal. An analysis-by-synthesis procedure is employed to find the excitation that minimizes the weighted Minimum Mean Square Error (MMSE) between the synthesized and the original signal. The applied weighting filter $W(z)$ is derived from the LPC synthesis filter and takes into account the psychoacoustic effect, that quantization noise in the spectral neighbourhood of the formants is less perceptible.

For complexity reasons adaptive and fixed codebook are sequentially searched for the best entry, i.e. first the best adaptive contribution is selected and then the fixed one. The adaptive codebook consists of time-shifted versions of past excitation sequences. In this way, the long-term characteristic (periodicity) of the speech signal is considered. As described below, the fixed codebook can be exchanged during operation to realize the rate adaptation. In order to track the varying channel conditions we developed four different codec modes, i.e. 6.1, 8.1, 9.5 and 13.3 kb/s. All four codebooks are sparse excitation algebraic codebooks [3].

Short-Term Prediction

We perform a 10th order LPC-analysis on the actual speech frame using a split-Levinson approach as proposed in [4]. This algorithm represents a very efficient means to compute the direct prediction coefficients $a_i, i = 1 \ldots 10$, the reflection coefficients k_i as well as pairs of line spectral frequencies (LSF). The LSFs are specifically appropriate for quantized transmission. In order to avoid sharp transitions between successive coefficient sets we use a 5 ms

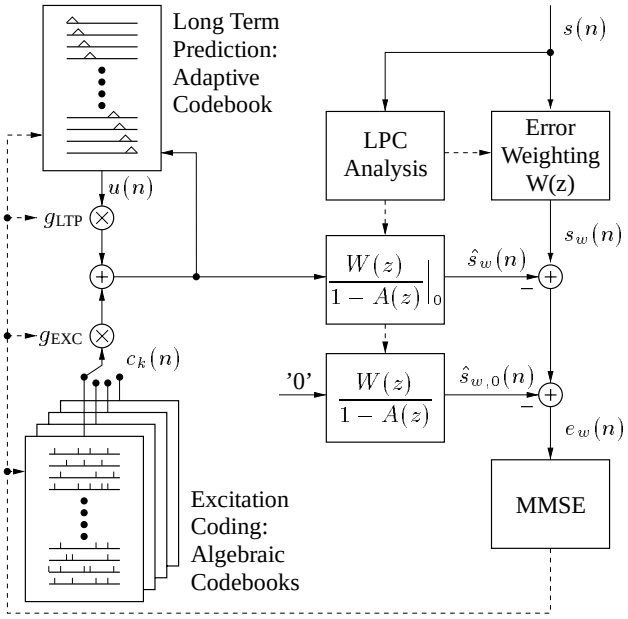

Figure 1: Block diagram of the VR-CELP speech encoder

lookahead for the LPC analysis in combination with an interpolation of the LSF coefficient set for subframes of 40 samples.

LSF quantization

For the quantization of the 10 LSF parameters we use a memory based 2-stage split vector quantization (VQ). This scheme allows the transmission of high quality spectral information at very low bit rates [5, 6]. Figure 2 depicts the basic structure of a memory based (predictive) VQ. Note that the prediction is carried out on each component ω_i of the LSF vector independently.

As predictor order we choose $K = 4$ making a tradeoff between complexity and prediction gain. In order to increase the flexibility of the prediction we trained two different coefficient sets $g_{ij}^{(1)}$, $g_{ij}^{(2)}$, one optimized for stationary speech segments, the other for transient ones. For each LSF vector the optimum of these coefficient sets is selected in the MMSE sense. One extra bit has to be transmitted to specify the selected predictor. The predictor structure allows to exploit the interframe correlation of the LSF vectors without being too sensitive to transmission errors as a possible error does only propagate up to K subsequent frames.

Figure 3 shows the VQ structure. The residue $\mathbf{c}$ of the LSF prediction is subject to a two stage vector quantization maintaining the highest possible speech quality at a reasonable complexity. The first stage VQ_1 yields a centroid vector $\hat{\mathbf{c}}_1$. The quantization error of the first stage is split into two component sets, each containing 5 of the 10 vector components. The two subsets are quantized by

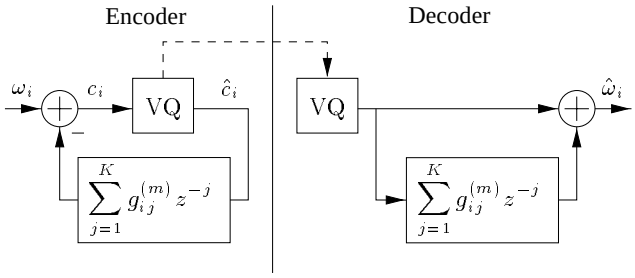

Figure 2: Memory based LSF vector quantization

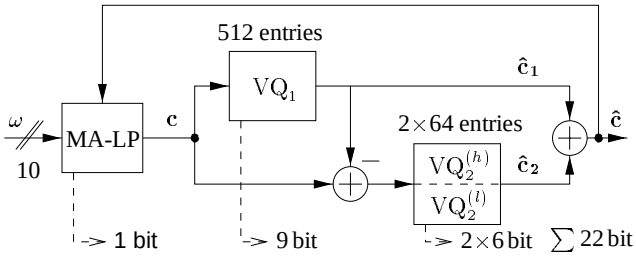

Figure 3: Block diagram and bit allocation of the 2-stage vector codebook for the LSF quantization

$\text{VQ}_2^{(l)}$ and $\text{VQ}_2^{(h)}$ respectively. Finally both centroid vectors $\hat{\mathbf{c}}_1$ and $\hat{\mathbf{c}}_2$ are combined yielding $\hat{\mathbf{c}}$.

As can be seen from this figure only 22 bits are required to quantize the LSF vector, which is a remarkably low rate for the given frame length of 20 ms. In our simulations we achieved average spectral distortions of 1.2 to 1.4 dB compared to the unquantized spectral envelope. Moreover, informal listening tests confirmed that the distortion due to the LSF quantization is not perceptible.

For the quantizer training, an important issue is the dependence between prediction coefficients and codebook entries which requires a joint optimization of predictor and codebook. We solved this problem by an iterative algorithm alternately optimizing predictor and vector codebook [5].

Long Term Prediction

For the long term prediction we utilize a 1 tap adaptive codebook. As the integer pitch resolution is rather low for short pitch periods which can often be found for female speech, we introduced a fractional pitch in the range of pitch periods from 20 up to 84 samples by steps of $1/3$. The encoding of the complete pitch range requires 8 bits per 5 ms subframe.

The LTP gain factor shows a limited dynamic range [7]. Therefore we found it sufficient to employ an eight level scalar Lloyd/Max quantizer [8, 9] with three bits. Thus we require in total $4 \times (8 + 3) = 44$ bits to encode the complete LTP information.

Excitation Coding

Right from the beginning of the codec design it appeared most promising for the realization of various bit rates to provide merely different fixed excitation codebooks while leaving the coding schemes for all other speech parameters invariant. This is due to the following reasons:

- The different codec modes share most of the software and tables which results in an overall codec with a program memory and ROM size comparable to a single mode codec.
- Seamless mode switching can very easily be implemented by simply changing the vector codebook for excitation coding.
- In the case of mode misdetection the VR-CELP exhibits a very robust behaviour since only the fixed excitation is misinterpreted, which usually leads only to minor distortions of the reconstructed speech.

The excitation codebooks are realized as sparse algebraic codebooks [7]. Table 1 comprises the codebook properties of the various codec modes.

Rate [kb/s]	Length	Tracks	Pulses	Pulse Type
13.3	20	7	7	ternary
9.5	40	3	6	ternary
8.1	40	5	5	binary
6.1	40	2	2	binary

Table 1: Excitation codebook properties of codec modes

For the excitation of the highest rate we bisect the excitation vector length of 40 into two sub-subframes of 20 samples. This is mainly done to achieve a complexity reduction, but on the other hand this improves the reconstruction of the excitation signal.

Due to the high dynamic range of the fixed excitation gain factor we employ a memory based scalar quantization which exploits the fact that the energy of subsequent excitation vectors changes rather slowly [7]. We choose a similar predictor structure as for the LSF quantization (see Figure 2) to avoid infinite error propagation.

For the training of the memory based quantizer in this case we use a two step procedure. First we compute an optimum predictor for the unquantized gain factor sequence via Durbin's method [10], then we train a Lloyd/Max quantizer subject to the prediction residual.

Post-Processing

To enhance the subjective quality of the synthesized speech we utilize a noise shaping adaptive postfilter [11]. The postfilter consists of a long term part, short term part, tilt compensation and an automatic gain control.

The principle of noise shaping is to attenuate all those frequencies that are less relevant for the speech signal as such. In this way the quantization noise, which is assumed to be almost white, is partly suppressed below the masking threshold.

By informal listening tests we found that it is advantageous to have two different postfilter adjustments, one for rates 13.3 and 9.5 kb/s the other for rates 8.1 and 6.1 kb/s. This is a tradeoff between speech naturalness and quantization noise.

Bit Allocation

The overall bit allocation of the various codec modes is given in Table 2.

Rate [kb/s]	LPC	LTP	g_{LTP}	g_{EXC}	EXC	total
13.3	22	32	12	32	168	266
9.5				16	108	190
8.1					80	162
6.1					40	122

Table 2: Overall bit allocation of codec modes

The table shows that all rates share the same parameter encoding part. Only the highest rate requires twice the number of gain factors due to the halved excitation vector length. To improve the reconstruction quality in case of mode misdetection when the highest rate is involved, we position the odd numbered gain factors of the highest rate adequately within the coded frame such that they can serve at least as rough estimates for the gain values of other rates and vice versa.

3. VR-CELP ERROR CONCEALMENT

The aim of error concealment is to make residual errors (i.e. errors that could not be corrected by the channel decoder) inaudible or at least less annoying for the listener. The concealment technique presented here is based on two informations: Reliability information about the received bits and a priori information about the sent speech codec parameters [1].

Figure 4 depicts the simplified block diagram of the speech transmission system. In principle the speech encoder performs two functions. First it analyzes the speech signal s and computes real valued parameters v which describe the signal (note that v may also be a vector in case of a multidimensional parameter). In a second step the parameters are quantized and encoded by sequences of bits that are composed to bit vectors $\mathbf{x}$. The bit stream is then transmitted over an equivalent channel which shall comprise the components channel (de)coding, (de)interleaving, (de)modulation, equalization and, of course, the physical noisy channel itself.

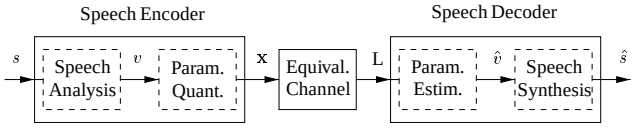

Figure 4: Block diagram of the speech transmission system

The output symbols of the equivalent channel shall be so called log likelihood ratios (shortly called L-values) [12], which are defined as $\mathrm{L}(x_j) = \log\left(\frac{\Pr\{x_j=0\}}{\Pr\{x_j=1\}}\right)$ for a component x_j of the bit vector $\mathbf{x}$. The absolute of the L-values corresponds to the reliability of the received channel values while their signs represent the hard decoded bits ($x_j = \begin{smallmatrix}1 & \mathrm{L}<0\\ 0 & \mathrm{L}>0\end{smallmatrix}$).

The error concealment unit exploits reliability information (L-values) and a priori knowledge (parameter pdf $p_v(v)$) to compute *estimates* $\hat{v}$ of the actually sent parameters. Finally, the parameter estimates are fed into the speech synthesis block which yields the reconstructed speech signal $\hat{s}$.

For a given parameter quantizer and a known channel an optimal estimator with respect to a specific fidelity criterion can be designed. Actually, this criterion should be related to the subjective speech perception. Mostly the subjective speech perception is well correlated with the speech *parameter* SNR. Therefore, we propose the MMSE as fidelity criterion for the speech parameters, which, as known from estimation theory, leads to the Mean Square (MS) estimator.

Let $\hat{c}_i$, $i = 1..N$ be the N possible quantization region centroids of a parameter v, then the MS estimate is

$$\hat{v}_{\text{MS}} = \sum\nolimits_{i=1}^{N} \hat{c}_i \cdot \Pr\{\hat{c}_i\}. \tag{1}$$

The evaluation of this formula becomes rather complex for parameters that are quantized employing large codebooks with high dimensionality. For these parameters we use a much simpler Maximum A Posteriori (MAP) detector with only small performance loss.

$$\hat{v}_{\text{MAP}} = \arg\max_{\hat{c}_i} \Pr\{\hat{c}_i\} \tag{2}$$

Assuming a memoryless channel the a posteriori probabilities $\Pr\{\hat{c}_i\}$ can be computed by applying Bayes' rule as

$$\Pr\{\hat{c}_i\} = C \cdot p_{\hat{c}_i}(\hat{c}_i) \cdot \exp\left(\sum\nolimits_{j\in\{j|x_j^{(i)}=0\}} \mathrm{L}(x_j^{(i)})\right), \tag{3}$$

where $x_j^{(i)}$ is the j-th component of the bit vector that is assigned to the quantization centroid $\hat{c}_i$ for transmission and $p_{\hat{c}_i}(\hat{c}_i)$ is the discrete pdf of the quantizer output. The normalization constant C is determined by the condition $\sum_i \Pr\{\hat{c}_i\} = 1$. Equation (3)

demonstrates the combination of a priori knowledge and reliability information.

The performance of our error concealment concept can be further improved by adding *explicit* redundancy to the transmitted bit vectors $\mathbf{x}$. This redundancy may be generated e.g. by a linear block code. A simple example is the Single Parity Code (SPC), which produces one bit redundancy by using the parity equation $x_{M+1}^{(i)} = \sum_{j=1}^{M}{}^{\oplus} x_j^{(i)}$ ($\oplus$ denotes modulo 2 addition). The M-dimensional bit vector is now simply lengthened by one additional bit and equation (3) can be evaluated as before. The basic idea of extending our error concealment approach by using additional parity bits has been successfully applied by us in an AMR codec proposal submitted to ETSI (European Telecommunications Standardisation Institute) [13]. A similar approach has also been proposed recently in [14]. Further improvement of this concept by using more complex codes like Hamming or BCH codes is straightforward.

The reconstruction quality under bad channel conditions is dramatically increased by just protecting the first LPC index in the described way by a shortened (13,9) Hamming code, although the channel related error control code was weaker[1].

4. CONCLUSIONS

We presented a new variable rate CELP speech codec operating at rates from 6.1 up to 13.3 kb/s. The codec was designed as homogeneous as possible, requiring the changing of only few modules when switching between the various rates. Additionally, our rate switching mechanism exhibits an inherent robustness against mode mismatch.

A major aspect in the codec design was the robustness against transmission errors. We proposed a new error concealment technique which is supported by explicit redundancy added specifically to single speech codec parameters. In simulations we showed that with respect to subjective speech quality this approach is able to outperform systems with conventional error control coding and parameter estimation.

The VR-CELP speech coding and the error concealment algorithms described in this study were combined with the channel coding and codec rate adaptation schemes developed at TU Munich, Germany to build an AMR codec for GSM speech transmission [13]. Subjective tests were carried out to evaluate the AMR codec for the conditions defined at ETSI SMG11 AMR standardisation meetings. The test results showed that the AMR codec met most of the qualification requirements and constraints. No audible effects were identified when switching the bit rate from one to another. For clean speeches under error- free condition, the presented VR-CELP codec at the rate of 9.5 kb/s achieved performance close to that of the enhanced full rate speech codec, the latest GSM standard with a speech coding rate of 12.2 kb/s, and at the rate of 13.3 kb/s it performed superior to the enhanced full rate codec. In addition, the codec at the rate 8.1 kb/s outperformed the 16 kb/s ITU-T G. 728 standard speech codec. And at its lowest bit rate 6.1 kb/s, it provided better performance than the GSM full rate speech codec which operates at a bit rate of 13.0 kb/s. All the GSM full rate, enhanced full rate and G.728 speech codecs were reference codecs used for the AMR codec standardisation in ETSI.

[1] In this case the bit rate for the channel related error control coding is reduced to keep the overall gross bit rate constant.

ACKNOWLEDGEMENTS

We want to thank Tim Fingscheidt, with RWTH Aachen, Prof. Joachim Hagenauer, Thomas Hindelang and Max Schmautz, with TU Munich, and Dr. Stefan Oestreich, Dr. Jürgen Paulus, with Siemens AG, for many fruitful discussions. Siemens AG, Munich, Germany, has supported this work.

REFERENCES

[1] T. Fingscheidt and P. Vary, "Robust Speech Decoding: A Universal Approach to Bit Error Concealment," in *ICASSP'97*, vol. 3, pp. 1667–1670, Apr. 1997.

[2] M. R. Schroeder and B. S. Atal, "Code-Excited Linear Prediction (CELP): High-Quality Speech at Very Low Bit Rates," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, (Tampa, Florida), pp. 937–940, 1985.

[3] J. Adoul, P. Mabilleau, M. Delprat, and S. Morissette, "Fast CELP Coding Based on Algebraic Codes," in *Proc. Int. Conf. Acoust., Speech, Signal Processing, ICASSP*, pp. 1957–1960, IEEE, 1987.

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

[5] A. Kataoka, T. Moriya, and S. Hayashi, "An 8-kb/s Conjugate Structure CELP (CS-CELP) Speech Coder," *IEEE Trans. Speech and Audio Comm.*, vol. 4, pp. 401–411, Nov. 1996.

[6] J. Lindén, "Interframe LSF Quantization for Noisy Channels," *IEEE Transactions on Speech and Audio Processing*, revised Jan. 1998.

[7] R. Salami, C. Laflamme, J. Adoul, and 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, Aug. 1994.

[8] J. Max, "Quantizing for Minimum Distortion," *IRE Trans. Information Theory*, vol. 6, pp. 7–12, Mar. 1960.

[9] S. P. Lloyd, "Least Squares Quantization in PCM," *IEEE Trans. Information Theory*, vol. 28, pp. 129–137, Mar. 1982.

[10] J. Durbin, "Efficient Estimation of Parameters in Moving-Average Models," *Biometrica*, vol. 46, pp. 306–316, 1959.

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

[12] J. Hagenauer, "Source-Controlled Channel Decoding," *IEEE Transactions on Communications*, vol. 43, pp. 2449–2457, Sept. 1995.

[13] ETSI SMG11, "Proposal of an Adaptive Multi-Rate Codec," AMR #10 Tdoc AMR /98, Siemens, Stockholm, Sweden, June 1998.

[14] N. Görtz, "Joint Source Channel Decoding Using Bit-Reliability Information and Source Statistics," in *Proc. Int. Symp. on Information Theory, ISIT*, p. 9, IEEE, 1998.