A LOW-COMPLEXITY TOLL-QUALITY VARIABLE BIT RATE CODER FOR CDMA CELLULAR SYSTEMS

Peter Kroon

AT&T Bell Laboratories
Murray Hill, NJ 07974
email: kroon@research.att.com

Michael Recchione

AT&T Bell Laboratories
Whippany, NJ 07981
email: mcr@research.att.com

ABSTRACT

With the deployment of digital cellular systems such as 13 kb/s GSM RPE-LTP, 8 kb/s fixed rate IS54 VSELP and 8 kb/s variable rate IS96 QCELP it is becoming clear that these coders are not as robust as originaly expected. As a result there has been renewed interest in higher bit rate coders that provide toll quality performance with regard to various input signals and multiple encodings. This paper describes the implementation of a toll quality 13 kb/s CELP coder. This coder can also operate at a variable bit rate mode (13, 6.5 and 0.8 kb/s), and has potential applications for a revised CDMA cellular system and PCS systems based on GSM.

1. INTRODUCTION

With the increased use of cellular communication systems (e.g. GSM [4],IS54 VSELP[3] and IS96 QCELP[2]), it is becoming clear that customer expectations with regard to speech quality are not fully met. One of the main reasons for this failure is that the speech coder, which is one of the key components for achieving efficient use of the available bandwidth, is not providing high enough speech quality at its operating rate around 8 kb/s. Examples are the performance for noisy backgrounds such as car noise, babble noise or nonspeech inputs such as music. Transcoding situations, in which the output of one speech encoder/decoder is used as input for another speech encoder/decoder have also led to unacceptable speech quality. Even coders that incorporate the latest technology, such as the new ITU-T 8 kb/s candidate [5][9] still have problems with robustness against speech with background noise. Obviously increasing the bit rate alleviates the aforementioned problems at the expense of capacity. However, since for most applications capacity is not a serious problem, it makes increasing the bit rate a viable option. In the US CDMA IS96 system, this option can be implemented by a relative simple modification of the error correction layer. As a result the modified channel can support 13 kb/s for the full rate coder, 6.5 kb/s for the half rate coder and 0.8 kb/s for the 1/8 rate.

This paper discusses in detail a speech coder for such an application Both the 13 kb/s and 6.5 kb/s coders are based on a CELP coder and differ only in the bit allocation used for its various parameters.

2. SPEECH CODER DESCRIPTION

The coder design objectives are tabulated in table 1. To meet these requirements is was decided to design a speech coder based on a CELP analysis-by-synthesis structure (see e.g.[7]). CELP has been used extensively over the previous years, which means that a lot of experience is available regarding optimizing its performance for the cellular environment.

Table 1: Coder requirements for 13 kb/s operation. G.728 is the 16 kb/s LDCELP ITU-T coder.

Attribute	Requirement
speech quality	equivalent or better than G.728
background noises	equivalent or better than G.728
2 encodings	equivalent or better than G.728
complexity	no more than IS96 QCELP
frame size	20 ms or submultiple
delay	$\leq$ 25 ms

The coder computes LPC information every 20 ms. By using an asymmetrical window only 5 ms of look-ahead is needed without degrading performance. The 12 LPC coefficients are quantized and interpolated as Line Spectral Frequencies (LSF). The increased order makes the coder more robust for nonspeech sounds. The LSF quantizer is a scalar intra-frame quantizer [10] using 48 bits per frame (4 bits/coefficient). The quantizer not only accurately describes the corresponding spectra, but also minimizes the quantization "jitter" from frame to frame which is a necessary requirement for high quality coding.

To efficiently represent the periodic components in the input signal, the coder uses an adaptive codebook with noninteger delays. To reduce complexity an initial estimate of the delay is obtained from the residual signal. To obtain better delay estimates, the correlation sequence is smoothed, which is equivalent to low-pass filtering the residual signal. This estimate is used to find the optimal fractional pitch every 5 ms. The nonuniform delay table is encoded with 8 bits, and the gain is encoded with 3 bits.

To increase robustness and to maintain low-complexity, the codebooks are based on a 7 bit multi-pulse structure, using either 2 or three unit amplitude pulses within a 10 dimensional vector. Table 2 shows the codebook structure. The first 2 rows define the 50 2-pulse vectors, the next three

0-7803-2431-5/95 $4.00

Table 2: Codebook structure for 10-dim vectors.

Amplitude	Positions				
+1.225	0	2	4	6	8
±1.225	1	3	5	7	9
+1	0	3	6	9	
-1	1	4	7		
±1	2	5	8		

rows define the 72 3-pulse vectors. The last vector is used to define an all-zero excitation. This leaves 5 vectors unused. Some experiments were done to see if populating these unused vectors with single pulse vectors could improve the performance for plosives. It was found that without an a-priori decision about the inclusion of the single-pulse vector, the performance actually degraded. Incorporating logic to make an a-priori decision was found not to be a good trade-off between performance and increased complexity. Hence it was decided to leave these vectors unused.

A pole-zero weighting filter is used to increase the quality of the signal. By using a Toeplitz weighting matrix, many of the search components can be pre-computed resulting in a very efficient search algorithm. To reduce the dynamic range of the fixed codebook gain, a fixed 1st order 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 previous scaled fixed code book vector, in a similar way as done in [1]. The codebook search procedure determines the scale factor needed to optimally scale the fixed codebook vector. Due to the short length of the codebook vectors it would be ineffecient to use scalar quantizers for the gain. Instead, two gain values are jointly quantized using a 2-dimensional VQ. It was found that it was sufficient to do the gain quantization in an open-loop mode using an Euclidian distortion measure. In other words, the optimal scale factors were computed and once 2 scaling factors are available, the optimal representation is found by doing an exhaustive codebook search using a 5 bit codebook. For optimum performance it is necessary to take the gain predictor into account. Once the coefficients are quantized, the filter conditions are recomputed before the next codebook search is done.

In the current version the decoder incorporates a postfilter, similar to the one used in G.728[1]. This postfilter uses an adaptive pole-zero postfilter with spectral tilt compensation and a harmonic enhancement filter. The coefficients of these filters are computed from the decoded signal. Preliminary tests with a simplified version using the received spectral and long-term correlation coefficients have shown that similar performance can be obtained, but no formal subjective test data is available yet.

2.1. Operation at 6.5 kb/s

To operate the coder in half rate mode, the frame sizes remain the same. The main difference is that the LPC order is reduced to 10, and that the LSF coefficients are encoded with 34 bits. The adaptive codebook information is the same as for the 13 kb/s mode (3 + 8 bits) every 5 ms. The fixed codebook gain is now encoded with a scalar quantizer

Table 3: Codebook structure for 40 dimensional vectors.

Amplitude	Positions				
+1	0	8	16	24	32
±1	3	11	19	27	35
±1	6	14	22	30	38

using 3 bits/ 5 ms, and the fixed code book contains 40 sample vectors (5 ms), with a codebook size of 9 bits and 1 bit for the sign. The fixed codebook contains 500 3-pulse vectors as is shown in table 3. The remaining 12 vectors are unused and identify a zero excitation. Due to the simple structure of the codebook it is easy to derive fast search strategies. We found that satisfactory performance was obtained by searching one pulse at a time, and recomputing the overall gain once all pulse positions have been determined. In addition, the spectral weighting and postfilter parameters are changed.

3. CODING AT 1/8 RATE

For the 1/8 rate (800 b/s) it is assumed that the rate selector has properly identified that no speech is present, and that only background noise has to be encoded. Both GSM and IS96 have some mechanism for encoding silence segments. In GSM this mechanism is used for discontinuous transmission [11], and is only used when relatively long periods of background noise are encountered. It updates spectral and energy parameters at a relativey low rate (e.g. > 100 ms). IS96 QCELP[2] encodes the background noise at 800 b/s using a very coarse (10 bit) quantizer for the 10 LSF coefficients and using a noise excitation modulated by a 2 bit gain parameter. Both gain and spectral information are updated every 20 ms. In addition, a 6 bit seed is transmitted to initialize the random generator. This ensures that both encoder and decoder use the same noise sequence. Both approaches work reasonably well with stationary noises. However, many noises are nonstationary (e.g. office noises), and will get distorted when encoded with aforementioned approaches. It was found that most noises (except music) can be encoded properly with amplitude modulated noise with some overall spectral shaping [8]. To accommodate this an envelope detector has been developed[6], which samples the envelope at 10 ms intervals to allow accurate reconstruction of the envelope countour. The corresponding spectra are modeled by a 4-th order predictor and quantized with a (4+4) bit split VQ for each pair of LSF's. A total of 8 bits every 20 ms is used for the spectral information and 4 bits are used to encode each envelope amplitude sample (100 Hz sampling rate). The normalized (spectrally flat excitation signal) is represented by Gaussian noise which does not require any side information. Ideally, the noise generators in both encoder and decoder should by synchronized, but it was found that if they loose synchronization due to transmission it only has a small impact on the quality of the reconstructed signals. Informal testing for a large variety of background noises showed that this approach is superior to either the GSM or IS96 approaches.

When the coder is switching between the higher rates (1 and 1/2 rate) and the 1/8 rate, a mechanism is needed

for smooth transitions. Since the filter orders are different, it is not trivial to switch simply the excitation sequences. Moreover, this mechanism would make the coder also more sensitive to a loss of synchronization between encoder and decoder when operating in 1/8 rate. It was found that by using the parallel filter approach[13] smooth transitions between the different rates can be obtained. Since the amount of overlap required is no more than 5 ms, very little complexity is added.

4. RATE SELECTION MECHANISM

Within a CDMA system, the normal operating mode is determined by a rate selection mechanism or is imposed by the system operator as a function of capacity or other considerations. In this system, the system operator selects the highest permissible rate, i.e. either 13 kbps or 6.5 kbps. The rate selection mechanism then switches between this maximum rate when speech is present, and the lowest rate when speech is not present. It is important for the rate selector to make a reliable decision even when background noise is present because speech cannot be represented well at the 1/8 rate. It is especially important for the rate selector to correctly identify speech onsets, as a delay in switching from 1/8 to full rate at a speech onset greatly impairs speech quality. Conversely, the rate selector must not be so biased in favor of declaring speech present that system capacity is degraded.

The rate selector is driven by a speech/nonspeech decision made by a voice activity detector (VAD). The input to the VAD is the average power in a number of frequency bands, calculated in 5 non-overlapping 5 ms subframes that span the current 20 ms speech frame and the 5 ms lookahead interval. The VAD uses the average power measurements to generate two test statistics: a speech signal power estimate that acts primarily as an onset detector, and a short-term variability estimate that is used to detect steady-state speech. An estimate of the average background noise power in each band is computed for every subframe by filtering the average power measurements with a nonlinear first-order exponential smoother, and is subtracted from the total average power to produce an estimate of the subframe speech power. Speech onset is indicated by a sudden increase in subframe speech power.

While speech onsets are detected reliably using the speech power estimate alone, this estimate is less useful for distinguishing steady-state speech because the background noise estimate becomes biased by a succession of high-energy subframes. In order to make the system more robust, an estimate of the short-term variability of the average subframe power over an estimation interval of a few frames is also used. Because speech is highly nonstationary relative to the usual noise backgrounds, an increase in the subframe to subframe variation in signal power is a powerful indicator of speech. However, because the length of the interval needed to estimate variability imposes delay, this estimate is not useful for speech onset detection.

The test statistics in each frequency band are thresholded and speech is declared present in a subframe if any statistic exceeds its threshold. A frame is encoded at full rate if speech was declared present in any of the subframes of the current frame or the lookahead interval. Because of the delay imposed by the variability detector, no additional smoothing or hangover is needed.

5. FRAME ERASURE CONCEALMENT

Poor channel conditions within a CDMA system manifest themselves as frame erasures in which the information for one or more 20 ms frames is totally lost. The decoder is informed that a received frame has been irretrievably corrupted, and incorporates a recovery strategy based on extrapolating spectral information and energy contours.

Spectral extrapolation is accomplished as follows: The formant frequencies for the last good frame are preserved, and bandwidth expansion is performed with each succeeding erased frame to allow the vocal tract filter to gradually decay to a flat spectrum. The long-term predictor values used in the first erased frame of a series of erased frames represents the average values of both delay (corrected for pitch doubling) and gain. Long term predictor gains are reduced with each additional erased frame. The treatment of the fixed codebook contribution depends on whether the last good frame was voiced or unvoiced, as determined by comparing the average long-term predictor gain to a threshold. If the frame was voiced, the zero vector is chosen for the fixed codebook component. If the frame was unvoiced, a random fixed codebook index is selected, and the average fixed codebook gain from the last good frame is used in the first missing frame. The gain is reduced with each succeeding erased frame. The energy contour is extrapolated through erased frames by scaling the decoded speech by an appropriate gain. In the first erased frame, this gain is just the ratio of the average amplitudes of the decoded speech in the last good frame and the current missing frame. This calculation is repeated in each successive missing frame, but the resulting gain is attenuated to allow the signal to decay. Since the decoder state memories have been corrupted during an error burst, it is necessary to continue scaling the decoded speech signal by the ratio of the desired amplitudes even once the error burst has ended.

6. SUBJECTIVE QUALITY

A preliminary version of the coder was tested in a MOS test using 40 listeners and 5 male and 5 female talkers. In this test the listeners rate the quality of each speech sample on a 5 point scale using the adjectives excellent (5), good (4), fair (3), poor (2) and bad (1). The arithmetic average of all responses is used to compute the MOS score. The playback was done over headphones using only one ear, and the frequency response of the playback system was flat between 200 and 3600 Hz. In each experiment several MNRU reference conditions were used in addition to the reference coders 32 kb/s ADPCM (G.726) , 16 kb/s G.728 LDCELP[1] and 13 kb/s GSM[4].. A total of three experiments were conducted. The first experiment tested clean and tandem conditions. The speech had either a flat or a IRS frequency shaping. The tandem was simulated using asynchronous digital tandem, in which a time shifted version of the first encoding is used as input for the second encoding. Only the flat speech material was tested for 2 encodings. The results

Table 4: MOS results for experiments 1 and 3.

Coder	src	G.726	G.728	GSM	CELP
Rate (kb/s)	128	32	16	13	13
Flat Speech	4.24	3.71	3.77	3.44	3.94
IRS Speech	3.96	3.61	3.65	3.38	3.79
Tandem	4.24	3.43	3.56	3.03	3.68
15 dB car	3.70	3.61	3.54	3.16	3.54
20 dB babble	4.06	3.91	3.92	3.62	3.85
20 dB classic	4.29	4.05	4.12	3.62	4.08
20 dB vocal	4.26	4.11	4.20	3.85	4.02

Table 5: MOS results for frame erasures.

FER (%)	0	1	2	3
MOS	3.83	3.62	3.42	3.21

are shown in table 4. The second experiment tested frame erasures. The frame erasures were generated using a 6-th order Markov model [12], which means that error bursts of up to 5 20 ms frames can occur. The results are summarized in table 5. The third experiment tested the performance of the coders for speech with background noise. Different types of background noise were used such as car noise at 15 dB, babble noise at 20 dB, and two types of music signals at 20 dB, classical and vocal music. The second part of Table 4 summarizes the results. From these results it can be seen that the proposed CELP coder receives similar scores as G.728 and G.726 and hence meets all the requirements of table 1. Further improvement of the performance for music signals can be obtained by increasing the order of the short-term predictor to 16 and using a split-VQ to quantize the LSFs using 48 bits. However, such a change increases the complexity of the coder, and it was decided that the performance with the 12-th order predictor was adequate.

Although no formal test results are available for the coder operating at 6.5 kb/s, it was found from informal listening tests that at that rate the performance is comparable to 8 kb/s IS-96 QCELP. Future tests will include variable rate versions of the coder, but informal tests have indicated that running the coder in variable rate mode results in an average bit rate of about 7 kb/s with very little degradation in quality compared to the coder operating at 13 kb/s.

7. CONCLUSION

This paper presented a high quality variable rate coder intended for enhanced cellular CDMA services. At its highest rate (13 kb/s) the performance is equivalent (within the statistical error of the test) to 16 kb/s G.728 and 32 kb/s G.726. At its half rate the performance is equivalent to 8 kb/s IS96 QCELP. The 1/8 rate coder, uses an amplitude modulated noise excitation to drive a 4-th order LPC filter. Since these parameters are updated frequently, it allows good tracking of nonstationary noises. A rate selection mechanism based on measuring the variance of band-limited energy contours and on estimating instantaneous SNR in selected frequency bands has proven to be simple and efficient and produces an average bit rate of about 7 kb/s. Currently, the coder is being implemented in fixed and floating point DSP's and based on preliminary results it was found that the complexity of the 13 and 6.5 kb/s coders is about the same as IS96 QCELP operating at 8 kb/s.

8. REFERENCES

[1] *Recommendation G.728, Coding of speech at 16 kb/s using low-delay code excited linear prediction.* ITU, 1992.

[2] W. Gardner, P. Jacobs, and C. Lee. QCELP: A variable rate speech coding for cellular networks. In B.S. Atal, V. Cuperman, and A. Gersho, editors, *Speech and Audio Coding for Wireless and Network Applications*, pages 85–92. Kluwer Academic Publishers, Boston, MA, 1993.

[3] I.A. Gerson and M.A. Jasiuk. Vector sum excited linear prediction (VSELP). In B.S. Atal, V. Cuperman, and A. Gersho, editors, *Advances in Speech Coding*. Kluwer Academic Publishers, Boston, MA, 1990.

[4] K. Hellwig, P. Vary, D. Massaloux, J.P. Petit, C. Galand, and M. Rosso. Speech codec for the european mobile radio system. In *Proc. GLOBCOM*, pages 1065–1069, 1989.

[5] A. Kataoka, T. Moriya, and S. Hayashi. Implementation and performance of an 8 kb/s conjugate structure CELP speech coder. In *Proc. ICASSP*, pages II–93–II–96, 1994.

[6] W.B. Kleijn. Personal communication.

[7] P. Kroon and B.S. Atal. Predictive coding of speech using analysis-by-synthesis techniques. In S. Furui and M.M. Sondhi, editors, *Advances in speech signal processing*, pages 141–164. Marcel Dekker Inc., New York, 1990.

[8] G. Kubin, B.S. Atal, and W.B. Kleijn. Performance of noise excitation for unvoiced speech. In *Proc. IEEE Speech coding workshop*, pages 35–36, 1993.

[9] R. Salami, C. Laflamme, and J-P. Adoul. 8 kb/s ACELP coding of speech with 10 ms speech frame: a candidate for ITU standardization. In *Proc. ICASSP*, pages II–97–II–100, 1994.

[10] F.K. Soong and B.H. Juang. Optimal quantization of line spectrum pair (LSP) parameters. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, page 30.S.7, 1988.

[11] CB. Southcott, D. Freeman, G. Cosier, and et al. Voice control of the pan-european digital mobile radio system. In *Proc. GLOBECOM*, pages 1070–1074, 1989.

[12] V.K. Varma. Testing speech coders for usage in wireless communication systems. In *Proc. IEEE Speech coding workshop*, pages 93–94, 1993.

[13] W. Verhelst and P. Nilens. A modified-superposition speech synthesizer and its applications. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 2007–2010, 1986.