HYBRID MELP/CELP CODING AT BIT RATES FROM 6.4 TO 2.4 KB/S

Jacek Stachurski [1], *Alan McCree* [1], *Vishu Viswanathan* [1], *Ari Heikkinen* [2]
Anssi Rämö [2], *Sakari Himanen* [2], *and Peter Blöcher* [3]

[1]DSP Solutions R&D Center, Texas Instruments, Dallas, Texas, USA
[2]Nokia Research Center, Tampere, Finland
[3]Ericsson Research, Nürnberg, Germany

ABSTRACT

This paper describes extensions of the 4 kb/s hybrid MELP/CELP coder, up to 6.4 kb/s and down to 2.4 kb/s. The baseline 4 kb/s coder uses three coding modes: MELP in strongly voiced speech frames, CELP with pitch prediction in weakly voiced frames, and CELP with stochastic excitation in unvoiced frames. To minimize switching artifacts between parametric MELP and waveform CELP coding, an alignment phase is encoded in MELP and zero-phase equalization is applied to the CELP target signal. The 6.4 kb/s extension uses the same three modes as the 4 kb/s coder, with improved MELP and CELP coders. The 2.4 kb/s extension uses only two modes: MELP for voiced frames and CELP synthesis with random excitation for unvoiced frames. The alignment phase is encoded in MELP frames for all bit rates so that time synchrony with input speech is always maintained. Alignment phase and zero-phase equalization enable smooth switching between coders at different bit rates. The hybrid MELP/CELP coding structure leads to coders that perform better at a given bit rate than MELP or CELP separately, and better than or equivalent to higher bit-rate ITU standards. Formal subjective tests show that for all-but-one tested conditions, the 6.4 kb/s hybrid coder is better than 8 kb/s G.729 and the 2.4 kb/s coder is equivalent to, or better than, 6.4 kb/s G.729 Annex D.

1. INTRODUCTION

Combining parametric and waveform coders to take advantage of their relative strengths in representing different speech regions has been shown to provide high-quality speech at low bit rates [1, 2, 3, 4]. In particular, the hybrid Mixed Excitation Linear Prediction (MELP) [5] and Code Excited Linear Prediction (CELP) [6] coding technique achieves near toll quality speech at 4 kb/s [7]. The goal of this work was to design a family of high quality switchable coders, from 6.4 kb/s to 2.4 kb/s, that would share most of the encoding framework of this hybrid MELP/CELP 4 kb/s coder. Since waveform coders tend to outperform parametric coders at higher bit rates and parametric coders tend to perform better at lower bit-rates, it may be intuitive to expect that a hybrid MELP/CELP coder would reduce to CELP and to MELP for higher and lower bit rates, respectively. Contrary to this expectation, informal and formal listening tests show that our best performing coder still maintains the hybrid coder structure for bit rates well above 4 kb/s: a hybrid coder at 6.4 kb/s comfortably outperforms CELP coders at comparable, and even higher, bit rates. Although at 2.4 kb/s the CELP analysis-by-synthesis does not seem to provide an advantage, we found it beneficial to continue using CELP synthesis and encode the unvoiced LP excitation with random entries of a stochastic codebook.

This paper presents 6.4 and 2.4 kb/s extensions that share most of the analysis, quantization, and synthesis framework with the baseline 4 kb/s coder. These extensions maintain the key elements of the hybrid MELP/CELP coder, the alignment phase encoded in MELP frames and the zero-phase equalization applied to the CELP target signal, to enable smooth switching between different bit rates. Section 2 provides an overview of a hybrid MELP/CELP coder. Section 3 summarizes the baseline 4 kb/s coder. Section 4 describes the coder extensions at 6.4 and 2.4 kb/s. The formal subjective test results of the coders are discussed in Section 5.

2. HYBRID MELP/CELP CODER

In the MELP/CELP hybrid coder [4], MELP is used to encode strongly voiced (SV) speech frames, while CELP encodes weakly voiced (WV) and unvoiced (UV) frames. For each speech frame, the mode decision is based on the estimated pitch confidence: frames with the most reliable pitch, marginal pitch, and those with least reliable or no pitch are classified as SV, WV, and UV, respectively. The parametric model of MELP provides a good representation of strongly periodic speech in SV frames. The waveform coding of CELP more effectively represents transitions and helps to prevent annoying pitch errors in regions with erratic pitch in WV frames, and better approximates unvoiced speech in UV frames.

To limit switching artifacts between MELP and CELP modes, an alignment phase is estimated and encoded in MELP frames. With the alignment phase, the MELP coder preserves time synchrony between the synthesized and the original speech, as is done in waveform-matching CELP. Conceptually, the alignment phase represents the linear phase (or time shift) that must be applied in the MELP decoder to the synthetic pitch-pulse waveform to achieve a match with the original speech at a given point in a frame. The fundamental frequency and the alignment phase are interpolated at the decoder so that the pitch and the phase equal the transmitted values at specific points within a frame.

Zero-phase equalization further reduces switching artifacts between MELP and CELP modes. This equalization is needed to minimize the waveform mismatch between the zero-phase (magnitude-only) MELP excitation and the CELP excitation that more closely reproduces the original waveform shape. The goal of the equalization is to remove from the CELP target signal the waveform phase (or shape) component not encoded in MELP. The equalization is performed by time-domain FIR filtering, with the filter coefficients generated from pitch pulses extracted from the LP residual. The pitch-pulse waveforms are time-reversed so that, when the filter is applied to the input speech, the LP residual corresponding to the phase-equalized output will be approximately zero-phase. The spectral magnitudes of the equalization filter are

set to one to ensure an all-pass filter characteristic. To preserve signal energy, the filter coefficients are normalized to unit gain and interpolated over time to maintain smooth evolution of applied filters. The phase equalization is performed at the encoder only and does not require bits to be transmitted. It has no effect in the MELP mode; however, it is run in MELP to update the signal memories. Alignment phase and zero-phase equalization greatly reduce switching artifacts between MELP and CELP modes.

3. BASELINE 4 KB/S HYBRID CODER

The 4 kb/s hybrid MELP/CELP coder submitted as a candidate for ITU standardization [7] was used as a baseline for the extensions presented in Section 4. This 4 kb/s coder uses high quality MELP [8] for SV frames, ACELP [9] with pitch prediction for WV frames, and CELP with stochastic codebook for UV frames. After noise suppression is performed using Smoothed Spectral Subtraction [10], standard LP analysis is done with a 25 ms Hamming window every 20 ms. Pitch and voicing are computed using a subframe correlation-based algorithm applied to the lowpass-filtered signal, and to five speech frequency bands, every 10 ms. The pitch and the pitch confidence used in the mode decision are generated from the normalized pitch correlations and a peakiness measure of the LP residual. The mode decision thresholds result in about 60%, 15%, and 25% of the active-speech frames classified as SV, WV, and UV, respectively.

The LP parameters are encoded in the LSF domain with switched predictive multi-stage vector quantization (VQ) with a Bark scale perceptual weighting function. The LSF codebooks include five stages with one bit indicating the choice of the weakly or strongly predictive codebook. The same codebooks are used in all modes, with all five stages used in MELP but only three stages used in CELP. Alignment phase is estimated, encoded, and transmitted in MELP frames, and zero-phase equalization is applied to the CELP target signal. In the MELP coder, the gain corresponding to a frame end and the mid-frame gain, and one pitch lag per frame are encoded. The Fourier magnitudes are coded with switched predictive multi-stage VQ with Bark scale weighting. These codebooks include five stages with one bit indicating the selection of the weakly or strongly predictive codebook. Two bandpass voicing vectors are encoded per frame. In weakly-voiced CELP, the overall frame gain and pitch are quantized. A fixed algebraic codebook is used with two pulses for each of the four sub-frames. Four pitch-prediction gains and four normalized fixed-codebook gains, making up an eight-dimensional vector, are vector-quantized. The unvoiced mode is indicated by reserved codewords of the encoded WV pitch. The LP excitation is coded with a stochastic codebook in five 4 ms sub-frames and the five stochastic-codebook gains, normalized by the overall gain, are vector-quantized. One bit encodes the MELP/CELP selection and one parity bit is used for protection against bit errors. The bit allocation of the 4 kb/s coder is summarized in Table 1.

4. HYBRID CODER EXTENSIONS

The 6.4 and 2.4 kb/s coder extensions share with the baseline 4 kb/s coder most of the encoder analysis, quantization tables, and decoder synthesis. All coders use the same noise suppression, LP analysis, pitch and bandpass voicing analysis, alignment phase estimation, zero-phase equalization, and estimation of the Fourier magnitudes. They share quantization tables for LSFs, Fourier magnitudes, bandpass voicing, stochastic excitation, and unvoiced-frame gains. A higher parameter-analysis rate is used for the LP coefficients, alignment phase, and Fourier magnitudes in the 6.4 kb/s extension, but the functional blocks that are used to estimate these parameters are common with the other coders. The differences between the coders are explained in the following sections.

4.1. 6.4 kb/s Coder

In the 6.4 kb/s extension, the SV, WV, and UV modes correspond to those of the 4 kb/s coder: MELP is used to encode SV frames and CELP is used in WV and UV frames. In improving the MELP mode, we evaluated the effects of quantization for each coded parameter. We also evaluated the coder performance with each parameter at a higher update rate. We concluded that increasing the rate of the LP parameters and Fourier magnitudes and quantizing each set with fewer bits provide a larger quality improvement than using more bits per set with once-per-frame encoding. The same was true for the alignment phase. To use the cubic interpolation of the alignment phase as is done in the 4 kb/s coder, the pitch is needed at the same rate as the phase; so the pitch update rate was also increased. The number of bits for the gain was increased, but the encoding of the bandpass voicing was unchanged.

The main aspects that limit the speech quality in the weakly voiced 4 kb/s CELP are only one pitch per frame and only two algebraic-codebook pulses per sub-frame. In the 6.4 kb/s extension, we addressed both of these problems. Four pitch values are encoded in this coder, one for each sub-frame. Bit saving is achieved by encoding pitch lag differences in all but the first sub-frame. The number of algebraic-codebook pulses was increased to four per sub-frame, and a larger codebook is used to quantize the pitch-prediction and fixed-codebook gains. In the unvoiced

	SV	WV	UV
LSFs	29	19	19
Frame gain(s)	8	5	5
Pitch	8	6	4
Bandpass voicing	6	–	–
Alignment phase	6	–	–
Fourier magn.	21	–	–
Fixed codebook	–	40	45
Codebook gains	–	8	5
MELP/CELP flag	1	1	1
Parity bit	1	1	1
Total bits per frame	80	80	80

Table 1. Bit allocation for 4 kb/s coder

	SV	WV	UV
LSFs	48	24	24
Frame gain(s)	10	5	5
Pitch	16	20	7
Bandpass voicing	6	–	–
Alignment phase	12	–	–
Fourier magn.	34	–	–
Fixed codebook	–	68	85
Codebook gains	–	9	5
MELP/CELP flag	1	1	1
Parity bit	1	1	1
Total bits per frame	128	128	128

Table 2. Bit allocation for 6.4 kb/s coder

CELP, the size of the stochastic codebook that could be used in the 6.4 kb/s coder is too large for a reasonable complexity. Therefore, we decided to use a two-stage stochastic codebook providing quality improvement with only a small increase in computational complexity and storage.

We experimented with replacing the MELP mode with CELP, but informal listening showed better performance for the hybrid approach. In fact, since we improved all modes by a similar margin, the same mode classification provided very good performance. Consequently, the mode decision thresholds were left unchanged and the mode usage statistics are the same as in the 4 kb/s coder.

To summarize, the main differences between the 6.4 kb/s extension and the baseline 4 kb/s coder are: in MELP, faster update rate for the LP parameters, alignment phase, pitch, and Fourier magnitudes; in weakly voiced CELP, separate pitch value for every sub-frame, more pulses in the algebraic codebook, and larger codebook for pitch-prediction and fixed-codebook gains; and a two-stage stochastic codebook in the unvoiced CELP. The bit allocation for the 6.4 kb/s coder is given in Table 2.

4.2. 2.4 kb/s Coder

At the 2.4 kb/s bit rate, the CELP analysis-by-synthesis does not appear to provide an advantage. In unvoiced frames, choosing a random entry from a larger stochastic codebook worked better in our experiments, particularly for noisy conditions, than performing analysis-by-synthesis with a smaller codebook that could be used at 2.4 kb/s. However, the CELP synthesis of randomly selected stochastic codebook entries, with the corresponding five sub-frame gains and an overall frame gain, worked better than the MELP sinusoidal synthesis with random excitation and the gain encoded twice per frame. Therefore, the CELP analysis is not used in unvoiced frames in the 2.4 kb/s extension, but the CELP synthesis with randomly selected codebook entries is used. We decided not to use the weakly voiced CELP at this bit rate.

Since CELP analysis-by-synthesis is not applied, the alignment phase and zero-phase equalization are not required in the 2.4 kb/s extension to reduce switching artifacts between modes, but they are needed for switching between coders at different bit rates. As zero-phase equalization does not require bits to be transmitted, it is performed to update the signal memories needed for switching between rates. However, at 2.4 kb/s, we could not afford to always encode the alignment phase with six bits as in the 4 kb/s coder. Therefore, we designed two bit assignments within the MELP mode: SV_1 used in the first MELP frame and in the frames with large pitch variations, and SV_2 used in all other MELP frames. In SV_1, the phase is encoded non-differentially with six bits, bandpass voicing is encoded with one bit (low/high voicing level), and only the weakly predictive Fourier-magnitudes codebook is used. In SV_2, the phase is encoded differentially with four bits, two sets of bandpass voicing vectors are jointly quantized, and the gain is encoded with fewer bits than in SV_1. The decoder recognizes the bit assignment from the mode sequence and the variation of decoded pitch. To make sure that the speech quality loss is minimized at this lower bit rate, we evaluated quality degradation from using a smaller number of bits for each MELP parameter. As a result, the number of bits is reduced for each parameter, most significantly for Fourier magnitudes, except for the non-differential quantization of the alignment phase for which the number of bits is unchanged. The motivation for including the alignment phase in the 2.4 kb/s coder was the ability to switch between different bit rates without annoying artifacts. However, informal listening tests indicated that the alignment phase also improves the performance of the 2.4 kb/s coder, particularly for noisy conditions.

	SV_1	SV_2	UV
LSFs	19	19	29
Frame gain(s)	7	6	5
Pitch	7	7	7
Bandpass voicing	1	3	–
Alignment phase	6	4	–
Fourier magn.	8	9	–
Codebook gains	–	–	5
Unused	–	–	2
Total bits per frame	48	48	48

Table 3. Bit allocation for 2.4 kb/s coder

As the weakly voiced mode is not used in the 2.4 kb/s coder, input speech is classified only into voiced and unvoiced frames. The mode decision thresholds are set so that about 70% of the active-speech frames are classified as voiced and 30% as unvoiced; the voiced frames are encoded with MELP, while in unvoiced frames CELP synthesis with random excitation is used. Within MELP frames, SV_1 and SV_2 are used about 10% and 90% of the time, respectively.

To summarize, the main differences between the 2.4 kb/s extension and the baseline 4 kb/s coder are: in MELP, two different bit assignments, SV_1 and SV_2; weakly voiced CELP mode not used; CELP synthesis with randomly selected stochastic codebook entries in unvoiced frames; and modified mode decision thresholds to produce only two coding modes. The bit allocation for the 2.4 kb/s coder is provided in Table 3.

4.3. Switching Between Bit Rates

The above hybrid coders were designed to facilitate switching between bit rates without perceptually annoying artifacts. Switching between two frames encoded at different bit rates is straightforward in MELP and in CELP. With our design, the measures used to limit switching artifacts between MELP and CELP in the hybrid coder are also in effect when the transition occurs between modes encoded at different bit rates. In all coder variations, the alignment phase is estimated and encoded in MELP frames so that the time synchrony is preserved, as it is done in CELP. The zero-phase equalization that further facilitates switching between the modes also reduces the artifacts in the MELP/CELP transitions when the modes have different rates.

5. SUBJECTIVE TEST RESULTS

Formal subjective tests of the 6.4 kb/s and 2.4 kb/s extensions were carried out in Swedish, for clean speech conditions, and in Finnish, for conditions with background noise. The experiments were designed based on the test plan used in the ITU 4 kb/s standardization. Two experiments were conducted with thirty-two naive listeners each. Thirty-two sentences spoken by four males and four females were used in each experiment. The ITU objectives for the performance of the future extensions to the 4 kb/s standard are: the 6.4 kb/s coder to be better than 8 kb/s G.729, and the 2.4 kb/s coder to be equivalent to 6.4 kb/s G.729 Annex D. The specific conditions used in our tests were inferred from the ITU 4 kb/s tests. For example, since the 4 kb/s coder in tandem was compared to three encodings of G.729, we also used in our tandem comparisons three encodings of G.729 and G.729 Annex D.

	G.729	6.4 kb/s	
high level	3.62	3.82	better
nominal	3.30	3.66	better
low level	2.95	3.35	better
tandem	2.36	3.20	better
car noise	-0.33	0.58	better
babble noise	-0.20	0.41	better
interf. talker	-0.14	-0.12	equivalent

Table 4. Summary of test results for 6.4 kb/s coder

	G.729 D	2.4 kb/s	
high level	3.21	3.23	equivalent
nominal	3.02	3.03	equivalent
low level	2.73	2.87	equivalent
tandem	1.79	2.21	better
car noise	-0.50	-0.15	better
babble noise	-0.47	-0.43	equivalent
interf. talker	-0.37	-0.93	worse

Table 5. Summary of test results for 2.4 kb/s coder

	2.4 kb/s	4 kb/s	6.4 kb/s
high level	3.23	3.61	3.82
nominal	3.03	3.43	3.66
low level	2.87	3.27	3.35
tandem	2.21	2.89	3.20
car noise	-0.15	0.14	0.58
babble noise	-0.43	0.17	0.41
interf. talker	-0.93	-0.43	-0.12

Table 6. Comparison of 2.4, 4, and 6.4 kb/s coders

An Absolute Category Rating (ACR) was used to test the performance of the coder for clean-speech conditions that included high level (-16 dBov), nominal level (-26 dBov), low level (-36 dBov), and tandem. A Comparison Category Rating (CCR) was used to evaluate the coder performance for car noise (at 15 dB), babble noise (at 30 dB), and interfering talker (at 20 dB). The ACR and CCR tests provide the Mean Opinion Score (MOS) and the Comparison Mean Opinion Score (CMOS), respectively. For each condition, the statistical equivalence between compared coders was verified with a 2-tailed t-test at the 95% confidence level.

The MOS and CMOS results for the 6.4 kb/s coder are summarized in Table 4. For clean speech, the 6.4 kb/s coder is statistically better than 8 kb/s G.729 for all tested conditions. It is also statistically better than G.729 for car and babble noise, and equivalent for interfering talker. The test results for the 2.4 kb/s coder are summarized in Table 5. The 2.4 kb/s coder is statistically equivalent to 6.4 kb/s G.729 Annex D for high, nominal, and low levels, and statistically better for the tandem condition. It is statistically better than G.729 Annex D for car noise, equivalent for babble noise, and worse for interfering talker. Table 6 summarizes the relative performance of the 2.4, 4, and 6.4 kb/s coders. The 6.4 kb/s coder is statistically better than the 4 kb/s coder for all tested conditions except for low levels where it scores higher but is not statistically better. The 4 kb/s coder is statistically better than the 2.4 kb/s coder for all tested conditions.

6. CONCLUSIONS

We have described high quality coders at 6.4 and 2.4 kb/s that share the coding framework with the 4 kb/s hybrid MELP/CELP coder. These coders form a close family: they share most of the encoder analysis, quantization tables, decoder synthesis, and they are switchable without perceptually annoying artifacts. In formal subjective tests, for all but one of the tested conditions, the 6.4 kb/s coder was statistically better than 8 kb/s G.729 and the 2.4 kb/s coder was equivalent to, or better than, 6.4 kb/s G.729 Annex D. Furthermore, the coders provide improved quality of speech with increasing bit rate. The performance of these coders shows that the hybrid MELP/CELP coding technique can lead to significantly better coders than MELP or CELP separately. The key components of the hybrid MELP/CELP coder, the encoding of alignment phase and the application of zero-phase equalization, not only reduce switching artifacts between modes, but also facilitate switching between coders at different bit rates.

7. REFERENCES

[1] D. L. Thomson and D. P. Prezas, "Selective Modeling of the LPC Residual During Unvoiced Frames; White Noise or Pulse Excitation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tokyo, Apr. 1986, pp. 3087–3090.

[2] W. B. Kleijn, "Encoding Speech Using Prototype Waveforms," *IEEE Trans. on Speech and Audio Processing*, vol. 1, no. 4, pp. 386–399, Oct. 1993.

[3] E. Shlomot, V. Cuperman, and A. Gersho, "Combined Harmonic and Waveform Coding of Speech at Low Bit Rates," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Seattle, May 1998, pp. 585–588.

[4] J. Stachurski and A. McCree, "A 4 kb/s Hybrid MELP/CELP Coder with Alignment Phase Encoding and Zero-Phase Equalization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Istanbul, June 2000, pp. 1379–1382.

[5] A. V. McCree and T. P. Barnwell III, "Mixed Excitation LPC Vocoder Model for Low Bit Rate Speech Coding," *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 4, pp. 242–250, July 1995.

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

[7] A. McCree, J. Stachurski, T. Unno, E. Ertan, E. Paksoy, V. Viswanathan, A. Heikkinen, A. Rämö, S. Himanen, P. Blöcher, and O. Dressler, "A 4 kb/s Hybrid MELP/CELP Speech Coding Candidate for ITU Standardization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Orlando, May 2002, pp. 629–632.

[8] J. Stachurski, A. McCree, and V. Viswanathan, "High Quality MELP Coding at Bit-Rates Around 4 kb/s," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Phoenix, May 1999, pp. 485–488.

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

[10] L. Arslan, A. McCree, and V. Viswanathan, "New Methods for Adaptive Noise Suppression," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Detroit, May 1995, pp. 812–815.