A 1.7 KB/S MELP CODER WITH IMPROVED ANALYSIS AND QUANTIZATION

Alan McCree and Juan Carlos De Martin

DSPS R&D, Texas Instruments, Dallas, Texas
E-mail: [mccree|demartin]@csc.ti.com

ABSTRACT

This paper describes our new Mixed Excitation Linear Predictive (MELP) coder designed for very low bit rate applications. This new coder, through algorithmic improvements and enhanced quantization techniques, produces better speech quality at 1.7 kb/s than the new U.S. Federal Standard MELP coder at 2.4 kb/s. Key features of the coder are an improved pitch estimation algorithm and a Line Spectral Frequencies (LSF) quantization scheme that requires only 21 bits per frame. With channel coding, this new MELP coder is capable of maintaining good speech quality even in severely degraded channels, at a total bit rate of only 3 kb/s.

1. INTRODUCTION

The Mixed Excitation Linear Predictive (MELP) coder [1] was recently adopted as the new U.S. Federal Standard at 2.4 kb/s. Although 2.4 kb/s is generally considered to be a low bit rate, there are a number of applications where an even lower bit rate is necessary. One such application is wireless digital transmission of speech, where channels with poor signal-to-noise ratios require the insertion of a considerable amount of redundancy in order to preserve acceptable speech quality, thereby reducing the number of bits available to the source coder.

In this paper, we describe a MELP coder which requires only 1.7 kb/s and delivers speech quality superior to that of the Federal Standard at 2.4 kb/s for both clean and noisy speech. Properly protected with convolutional codes and with adequate handling of frame-erasures, our new MELP coder is capable of preserving the base quality even in 5% random errors.

2. CODER DESCRIPTION

The 1.7 kb/s MELP coder, like the new Federal Standard, uses the MELP model described in [2]. This model is based on the traditional LPC vocoder with either a periodic impulse train or white noise exciting an all-pole filter, but contains four additional features. As shown in Figure 1, the synthesizer has the following capabilities: mixed pulse and noise excitation, periodic or aperiodic pulses, adaptive spectral enhancement, and a pulse dispersion filter. There are three significant differences between the 1.7 kb/s MELP coder and the 2.4 kb/s Federal Standard: model improvements, more efficient quantization, and channel coding.

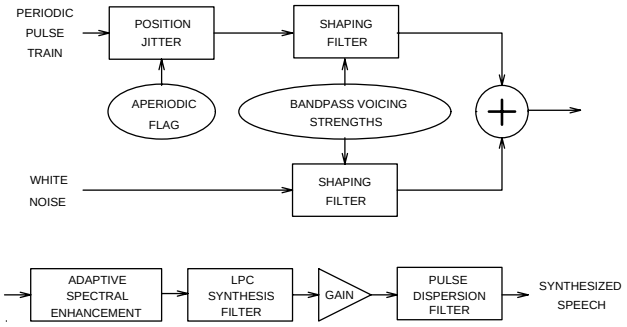

Figure 1: MELP synthesizer.

2.1. Model Improvements

Improvements to the MELP model have come in three areas. First, the pitch and voicing estimation has been improved. Second, a noise suppression front-end has been added to improve performance in acoustic background noise. Finally, the frame size has been decreased from 22.5 to 20 ms, resulting in an overall increase in speech quality.

2.1.1. Pitch Estimation

We have developed a subframe-based pitch estimation algorithm that significantly improves performance compared to the frame-based approach used in the Federal Standard. The objective is to find the pitch track through a speech frame that minimizes the pitch-prediction residual energy over the frame, assuming that the optimal pitch prediction coefficient will be used for each subframe lag T_s. Formally, this error can be written as a sum over N_s subframes:

$$E = \sum_{s=1}^{N_s} E_s = \sum_{s=1}^{N_s} \left[\sum_n x_n^2 - \frac{\left(\sum_n x_n x_{n-T_s}\right)^2}{\sum_n x_{n-T_s}^2} \right]$$

where x_n is the n^{th} sample of the input signal and the sum over n includes all the samples in subframe s. Minimizing this error is equivalent to maximizing the normalized corre-

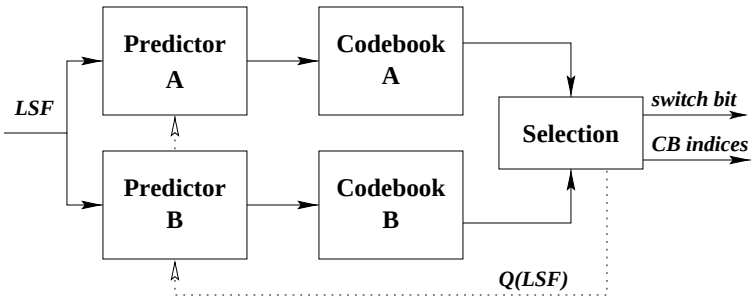

Figure 2: Switched-predictive LSF quantizer block diagram.

lation coefficient ρ given by

$$\rho^2 = \frac{\sum_{s=1}^{N_s} \frac{\left(\sum_n x_n x_{n-T_s}\right)^2}{\sum_n x_{n-T_s}^2}}{\sum_{s=1}^{N_s} \sum_n x_n^2} = \frac{\sum_{s=1}^{N_s} P_s \rho_s^2}{\sum_{s=1}^{N_s} P_s}$$

where $P_s = \sum_n x_n^2$ and ρ_s is the traditional normalized correlation coefficient within the subframe s. We now force a pitch track by imposing the constraint that each subframe pitch lag must be within a certain range of an overall pitch value T:

$$\rho_s(T) = \max_{T_s=T-\Delta}^{T+\Delta} \frac{\sum_n x_n x_{n-T_s}}{\sqrt{\sum_n x_n^2 \sum_n x_{n-T_s}^2}}$$

where Δ is the amount of pitch variation allowed across subframes within a frame. Note that without the pitch tracking constraint, the overall prediction error is minimized by finding the optimal lag for each subframe independently. Also, this method differs from the autocorrelation-based approach presented in [3] in that it incorporates the energy variations from one subframe to the next.

We use this subframe-based algorithm for both pitch and voicing estimation. For pitch estimation, we vary T over the entire pitch range and find the highest normalized correlation ρ of the lowpass filtered speech signal, with additional pitch doubling logic. For bandpass voicing analysis, we apply the algorithm to estimate the correlation strength at the pitch lag for each frequency band of the input speech. Experimentally, we find that this subframe-based pitch and voicing analysis performs better than the frame-based approach of the Federal Standard, particularly for speech transitions and regions of erratic pitch such as vocal fry.

2.1.2. Noise Suppression

Our Smoothed Spectral Subtraction (SSS) noise suppression method is based on traditional spectral subtraction, where an estimate of the noise power spectrum is subtracted from the spectrum of the noisy speech, but involves three separate improvements [4]. First, a clamp is applied to the noise suppression filter $H(w)$ so that it cannot go below a minimum value of -10 dB. This prevents the noise suppression filter from fluctuating around very small gain values, and also reduces potential speech signal distortion. Second, the noise power spectrum estimate is artificially increased by a small margin (5 dB) so that small errors in noisy signal spectral estimates do not lead to fluctuating attenuations. Third, instead of using the FFT-derived estimates of the noisy speech and noise spectra directly in the attenuation rule, we use smoothed versions of the power spectra. We use a moving average smoothing in frequency; a smoothing window size of 32 (for an FFT size of 256) was found to work well. This smoothing reduces the variance of the spectral estimates, which prevents musical noises from occurring. As a combined result of these three improvements, the SSS algorithm is able to attenuate the acoustic background noise by 10 dB without introducing any musical noise artifacts.

2.2. Quantization

The major bit rate reduction in the new MELP coder comes from the new LSF quantization scheme, which lowers the number of bits needed to represent the LPC filter from 25 to 21 bits, at no extra cost in terms of storage or complexity. More efficient quantization of pitch, voicing, and gain saves an additional three bits per frame. In order to reduce the overall data rate, the Fourier series magnitudes transmitted in the Federal Standard coder are eliminated, saving eight bits per frame.

2.2.1. LSF Quantization

We have designed a 21-bit switched predictive quantization scheme with better performance than the 25-bit quantizer used in the Federal Standard. Most of this efficiency improvement is due to the use of predictive quantization, but there is additional performance gain from using a theoretically optimal LSF weighting function.

We use a switched-predictive multi-stage vector quantization (MSVQ) of the LSF's, as shown in Figure 2. For each speech frame, both predictor/codebook pairs are tried, and the one that provides the best quantization performance is selected for transmission along with one bit to represent the switch information. We have a found a significant advantage to using two different codebooks rather than sharing

a single codebook [5], without any increase in complexity compared to the non-predictive case. The use of separate codebooks allows each to be separately optimized, as in safety-net VQ [6], while still utilizing prediction for both. Since both of the two 4-stage, 20-bit MSVQ codebooks are less than half the size of the 25-bit non-predictive version, both the storage and search complexity are actually reduced in the new scheme, and we can increase the search depth of our M-best MSVQ search from $M = 8$ to $M = 12$ for equivalent complexity.

For training, we use an extension of the iterative sequential MSVQ training procedure [7], in which we alternate between training the predictor coefficients given the codebook and training the codebook given the predictor coefficients. The closed-loop switching mechanism is also included in the training procedure. This implements a full closed-loop optimization for both the predictor coefficients and the codebook.

In addition to switched prediction, we also use a new LSF weighting function to approximate the frequency-weighted spectral distortion (SD_{fw}) defined by [1]

$$SD_{fw}(A_q(z), A(z)) = \sqrt{\frac{1}{W_0}\int_{f=0}^{4000} |W_B(f)|^2 10log_{10}\frac{|A_q(f)|^2}{|A(f)|^2}df}$$

where $A_q(z)$ and $A(z)$ represent the quantized and unquantized LPC filters, W_0 is a normalization constant, and the Bark weighting $W_B(f)$ is defined by

$$W_B(f) = \frac{1}{25 + 75(1 + 1.4(\frac{f}{1000})^2)^{0.69}}$$

We have previously found this perceptual weighting function based on the Bark scale to better predict listener preference in the MELP coder, and we now present an LSF weighting function which optimizes this form of SD.

At high rates, the optimal LSF weighting to minimize unweighted SD is the sensitivity matrix of the LSF's [8]:

$$\left.\frac{\partial^2 SD(a(\omega), a(\bar{\omega}))}{\partial\omega_k \partial\omega_l}\right|_{\omega=\bar{\omega}} = 4\beta j_{\omega_k}^T R_A j_{\omega_l}$$

where j_{ω_k} is the kth column of the Jacobian matrix for the LSF's, R_A is the autocorrelation matrix of the impulse response of the LPC synthesis filter, and β is a scale factor. Using the principles of linear filtering, it is straightforward to show that the optimal LSF weighting for a perceptually-weighted form of SD can be computed by replacing the matrix R_A with R'_A, the autocorrelation matrix of the perceptually-weighted impulse response of the LPC filter. In practice, we use an 8th order all-pole model approximation to the Bark weighting function $W_B(f)$. We find experimentally that this optimal weighting function results in a consistent but modest improvement in SD_{fw} over the empirical weighting described in [9], and an improvement of more than 0.05 dB compared to the power-weighted LSF distance [10] used in the Federal Standard.

Quantizer	SD_{fw} (dB)	$> 2dB$ (percent)
25-bit	1.06	2.4
21-bit switched	0.97	0.81

Table 1: LSF quantizer performance for flat input speech.

Parameters	2.4 kb/s	1.7 kb/s
LSF's	25	21
Fourier magnitudes	8	0
Gain	8	5
Pitch and overall voicing	7	6
Bandpass voicing	4	2
Aperiodic flag	1	0
Sync bit	1	0
Total bits / frame	54	34

Table 2: Bit allocations for 2.4 kb/s Federal Standard and 1.7 kb/s MELP coder. The frame sizes of the two coders are 22.5 ms and 20 ms, respectively.

The weighted spectral distortion for the Federal Standard quantization and the switched-predictive version is shown in Table 1. The test set is flat input speech that was not included in the training set. The 21-bit switched-predictive quantizer is clearly superior to the 25-bit non-predictive version, both in terms of average distortion and number of outliers. We have also observed that for severely filtered speech, which is not well represented in the training set, the switched-predictive scheme outperforms the non-predictive version. This suggests that the use of prediction reduces the sensitivity of the quantizer to mismatches between training and test sets due to filtering of the speech material.

2.2.2. Quantization of Remaining Parameters

Table 2 shows the bit allocation for the 1.7 kb/s MELP coder as compared to the 2.4 kb/s Federal Standard. In addition to the savings of four bits in LSF quantization and eight bits by not transmitting Fourier series magnitudes, there is an additional savings of eight bits from the remaining parameters. First, the gain is only transmitted once per frame rather than twice as in the Federal Standard, since the frame size is now shorter. Also, we have found that 6 bits are sufficient to quantize the pitch and overall voicing when Fourier magnitudes are not used. In addition, the number of bits required for bandpass voicing information is reduced to two by selecting from a catalog of four possible partial voicing patterns. The aperiodic flag is replaced by a functionally equivalent pitch contour perturbation technique, which does not require explicit transmission. In this approach, the encoder introduces pitch jitter by ensuring that the pitch contour changes rapidly for frames that are classified as aperiodic.

2.3. Channel Coding

Forward-error correction (FEC) codes are used to improve the performance in channel errors. Every 40 ms, two frames worth of data are grouped and encoded with a convolutional code of rate 3/5. To reduce the total bit rate, the perceptually less important fourth stage of the LSF's is left unprotected. Counting a 4-bit CRC protecting the most significant bits and a 6-bit tail, the overall bit rate on the channel is 3 kb/s. At the receiver side, a Viterbi decoder accepts soft inputs from the demodulator and performs Maximum-Likelihood decoding. If the CRC signals an error, a frame erasure algorithm extrapolates reasonable values for the parameters of the current frame from the past history.

3. SUBJECTIVE TEST RESULTS

We conducted formal subjective listening tests of an earlier version of this coder [11]. This evaluation consisted of forced choice A-B comparison tests with 102 sentence pairs, uttered by 10 different speakers, and with the 2.4 kb/s Federal Standard as reference coder. The test material included clean speech, both flat and IRS filtered, as well different kinds of noise (traffic, office, babble and truck). The pairs were randomized and presented to five different listeners. Overall the new low rate MELP coder was preferred over the Federal Standard, with a clear preference in five of the six test conditions. Only for clean flat speech was the Federal Standard preferred, probably due to the presence of the Fourier Series magnitudes.

We also informally assessed the performance of the system in channel errors. For 5% random errors on the channel, the FEC scheme described in Section 2.3 results in a post-decoding error rate of only $3 \cdot 10^{-4}$ and a frame-erasure rate of 0.25%. This results in virtually flawless performance. Even at error rates as high as 7%, the quality is quite good, with very few annoying artifacts in the output speech. Only at error rates approaching 9% is the performance seriously degraded, since the rate 3/5 coding begins to fail.

4. CONCLUSIONS

We have presented a new MELP coder which, through model and quantization improvements, outperforms the new Federal Standard at a significantly lower bit rate, making it an attractive candidate for wireless communications and other low data rate applications. A new subframe-based pitch and voicing algorithm provides good performance even for difficult speech signals, while switched-predictive MSVQ allows more accurate quantization of the LSF's at a lower data rate. Finally, a channel coding scheme based on convolutional codes preserves output quality even under very degraded channel conditions.

5. REFERENCES

[1] A. McCree, K. Truong, E. B. George, T. P. Barnwell, and V. Viswanathan, "A 2.4 kbit/s MELP Coder Candidate for the New U.S. Federal Standard," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 200–203, May 1996.

[2] A. V. McCree and T. P. Barnwell III, "A Mixed Excitation LPC Vocoder Model for Low Bit Rate Speech Coding," *IEEE Trans. Speech and Audio Processing*, vol. 3, pp. 242–250, July 1995.

[3] B. Kleijn, P. Kroon, L. Cellario, and D. Sereno, "A 5.85 kb/s CELP Algorithm for Cellular Applications," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, (Minneapolis), pp. II596–599, 1993.

[4] L. Arslan, A. McCree, and V. Viswanathan, "New Methods for Adaptive Noise Suppression," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 812–815, May 1995.

[5] S. Wang, E. Paksoy, and A. Gersho, "Product Code Vector Quantization of LPC Parameters," in *Speech and Audio Coding for Wireless and Network Applications*, pp. 251–258, Norwell MA: Kluwer Academic Publishers, 1993.

[6] T. Eriksson, J. Linden, and J. Skoglund, "Exploiting Interframe Correlation in Spectral Quantization - A Study of Different Memory VQ Schemes," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, (Atlanta), pp. 765–768, 1996.

[7] W. P. LeBlanc, B. Bhattacharya, S. A. Mahmoud, and V. Cuperman, "Efficient Search and Design Procedures for Robust Multi-stage VQ of LPC Parameters for 4 kb/s Speech Coding," *IEEE Transactions on Speech and Audio Processing*, vol. 1, pp. 373–385, October 1993.

[8] W. Gardner and B. Rao, "Theoretical Analysis of the High-Rate Vector Quantization of LPC Parameters," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 367–381, September 1995.

[9] R. P. Cohn and J. S. Collura, "Incorporating Perception into LSF Quantization – Some Experiments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, April 1997.

[10] K. K. Paliwal and B. S. Atal, "Efficient Vector Quantization of LPC Parameters at 24 Bits/Frame," *IEEE Trans. Speech and Audio Processing*, vol. 1, pp. 3–14, Jan. 1993.

[11] A. McCree and J. C. DeMartin, "A 1.6 kb/s MELP Coder for Wireless Communications," in *IEEE Workshop on Speech Coding for Telecommunications*, pp. 23–24, 1997.