A 2.4 KBIT/S MELP CODER CANDIDATE FOR THE NEW U. S. FEDERAL STANDARD

Alan McCree Kwan Truong E. Bryan George Thomas P. Barnwell Vishu Viswanathan

1Personal Systems Laboratory, Texas Instruments, Dallas, TX.
2Atlanta Signal Processors, Inc., Atlanta, GA.

ABSTRACT
This paper describes our enhanced Mixed Excitation Linear Prediction (MELP) coder which is a candidate for the new U. S. Federal Standard at 2.4 kbits/s. The new coder is based on the MELP model, and it uses a number of enhancements as well as efficient quantization algorithms to improve performance while maintaining a low bit rate. In addition, the coder has been optimized for performance in acoustic background noise and in channel errors, as well as for efficient real-time implementation. Listening tests confirm that the enhanced 2.4 kbit/s MELP coder performs as well as the higher bit rate 4.8 kbit/s FS1016 CELP standard.

1. INTRODUCTION
The MELP coder reported in [1] has been one of the top performers in previous evaluations of 2.4 kbit/s speech coders performed by the U. S. Department of Defense. We have made significant improvements to this baseline MELP coder, and have demonstrated an unquantized enhanced MELP algorithm with quality comparable to the 8 kbit/s VSELP North American digital cellular standard. This coder has been quantized at 2.4 kbits/s, and its performance has been optimized for robustness and real-time implementation for submission to the U. S. Federal Standard competition. The improvements made to the baseline coder are in three areas: enhancing the MELP model, more efficient coding and quantization of parameters, and coder optimization.

2. MELP MODEL ENHANCEMENTS
The MELP model is based on the traditional LPC vocoder with either a periodic impulse train or white noise exciting an all-pole filter, but the model contains four additional features. As shown in Figure 1, the synthesizer has the following added capabilities: mixed pulse and noise excitation, periodic or aperiodic pulses, adaptive spectral enhancement, and pulse dispersion filter. These features allow the mixed excitation LPC vocoder to mimic more of the characteristics of natural human speech. We have investigated three primary approaches to improving this MELP model: Fourier series coding, mel-scale LPC, and faster frame rate.

2.1. Fourier Series Coding
Our previous work has shown that enhancing the model with additional information in the form of Fourier series magnitudes improves the speech quality of the MELP coder [1]. This enhanced model uses all of the features of the conventional MELP coder shown in Figure 1, as well as the Fourier series coefficients of the excitation signal. Fourier analysis is performed on the LPC residual signal computed using the quantized LPC inverse filter either by taking the FFT of an entire frame or by selecting a single pitch period for Fourier series analysis. Synthesis of each pitch period of the pulse train is done with an inverse DFT of exactly one period in length, using interpolated versions of the transmitted Fourier coefficients for consecutive frames.

We have experimented with both complex and real Fourier coefficients in the MELP coder. The complex Fourier series completely specifies a single period of the excitation signal in a fashion similar to Waveform Interpolation [2], and we have verified that this method can produce high quality speech in an unquantized system. However, the wide range of possible excitation signals makes quantization of the complex coefficients inappropriate for our low bit rate application. The real coefficients specify only the Fourier magnitudes of the excitation signal, assuming zero phase. The magnitudes are more straightforward to quantize than the complex coefficients, since the power spectrum of the LPC residual signal is already approximately flat. In addition, a subset of the Fourier magnitudes can be selected for quantization and transmission, and the remaining values can be set to a constant value without significant degradation.

2.2. Mel-Scale LPC
The improvements due to Fourier magnitude coding suggest that an improved LPC spectral fit would increase speech quality. In the ideal case, a very accurate LPC spectrum should result in a spectrally-flat residual signal so that Fourier magnitude coding would not be needed at all. We have conducted informal listening tests that confirm that using a 30th order LPC model does increase the quality of the speech output from the MELP coder for male speakers. However, it would require a high bit rate to accurately
quantize such a high-order model. McAulay has recently suggested that a mel-scale LPC model (MLPC) can efficiently capture this improved spectral fit [9]. Since the mel scale places greater importance on the lower frequencies, this matches well with our observation that Fourier magnitude information is more important for the lower harmonics. We have implemented MLPC within the MELP model, and our evaluations show that a 14th order MLPC can indeed produce an LPC residual signal with flatter Fourier magnitudes at the lower frequencies than a 14th order traditional LPC. Informal listening tests confirm that the use of MLPC does improve the output speech quality.

2.3. Faster Frame Rate
One significant source of distortion from the baseline MELP coder is the frame rate. We have demonstrated that decreasing the frame size from 22.5 to 10 ms results in a significant quality improvement. In fact, listening tests show that an unquantized 10 ms MELP system produces speech quality comparable to the 8 kbit/s VSELP coder. However, this system uses 14th order LPC and the first 18 harmonics of the Fourier magnitudes every 10 ms, so tradeoffs must be made in order to quantize this information at 2.4 kbit/s.

3. EFFICIENT CODING AND QUANTIZATION
We have developed more efficient means for quantizing the MELP parameters, and evaluated different schemes for quantizing the MELP enhancements. In addition, we have evaluated two different approaches for variable frame rate transmission of the MELP parameters [4].

3.1. LPC Filter
We have developed a 25-bit multi-stage vector quantizer (MSVQ) for the line spectral frequencies (LSF’s) that actually improves the speech quality of the MELP coder slightly in comparison to our existing 34-bit scalar quantizer, at a significantly lower bit rate. This MSVQ uses joint optimization for both codebook design and search, using an M-best algorithm [5]. This approach results in improved performance over sequential MSVQ, while still maintaining low complexity, both in terms of computation and memory requirements.

3.1.1. Initial Experiments
Since robustness across different input microphone responses is an important design goal for our quantization algorithm, we conducted initial experiments using completely disjoint training and test sets. Training was performed using 60,000 frames from the close-talking microphone TIMIT database, while testing was performed on studio microphone Diagnostic Acceptability Measure (DAM) material. A 24-bit MSVQ, using four stages of 6 bits each, M = 8, and the power spectral weighted LSF distance [6] was trained using an adaptive gradient version of the iterative sequential codebook design algorithm. Table 1 shows the average Spectral Distortion (SD) attained by this quantizer, labeled as MSVQ1, as well as the percentage of outliers with more than 2 dB of spectral error. While the widely accepted transparent quality level of 1 dB is not attained due to the mismatched training set, the MSVQ still attains SD performance that is clearly superior to the scalar quantizer. Unfortunately, listening tests in the MELP coder show that these two quantizers are essentially equivalent, contradicting the SD results.

<table>
<thead>
<tr>
<th>Quantizer</th>
<th>$SD$ (dB)</th>
<th>$&gt;2dB$ (percent)</th>
<th>$SD_{lw}$ (dB)</th>
<th>$&gt;2dB$ (percent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>34-bit scalar</td>
<td>1.44</td>
<td>10.4</td>
<td>1.21</td>
<td>3.3</td>
</tr>
<tr>
<td>24-bit MSVQ1</td>
<td>1.22</td>
<td>9.5</td>
<td>1.22</td>
<td>6.4</td>
</tr>
<tr>
<td>24-bit MSVQ2</td>
<td>1.23</td>
<td>6.4</td>
<td>1.44</td>
<td>16.1</td>
</tr>
<tr>
<td>24-bit MSVQ3</td>
<td>1.31</td>
<td>7.7</td>
<td>1.56</td>
<td>20.1</td>
</tr>
</tbody>
</table>

Table 1. Quantizer performance based on different SD measures.

3.1.2. Frequency-Weighted Spectral Distortion
Since traditional SD does not predict subjective performance of a quantizer in the MELP coder, we have developed a frequency-weighted spectral distortion ($SD_{lw}$) for quantizer evaluation [7]. This distortion is evaluated on the bark scale, which is based on equal critical bandwidths. To reduce complexity and avoid explicit resampling of the frequency axis, $SD_{lw}$ can equivalently be calculated using a linear frequency scale with an amplitude weighting given by the derivative of the bark scale. Finally, informal listening experiments indicate that to attain sufficient perceptual weighting either the average error should be used across frequencies (instead of the average squared error), or the amplitude weighting should also be squared. In practice, we use this second approach with a 256-point discrete sum approximation to the integral:

$$SD_{lw}(A_{4}(z), A(z)) = \sqrt{\frac{1}{W_{0}} \int_{f_{0}}^{4000} |W_{B}(f)|^{2} 10\log_{10} \frac{|A_{4}(f)|^{2}}{|A(f)|^{2}} df}$$

where $A_{4}(z)$ 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

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

Table 1 compares conventional SD with the new $SD_{lw}$ measure for the 34-bit scalar quantizer and the 24-bit MSVQ. The new measure supports our informal listening conclusion that the scalar quantizer and MSVQ provide similar performance. We have also used informal listening tests and $SD_{lw}$ to compare different LSF weighting functions for searching the MSVQ codebook. In our experiments with the MELP coder, the power-weighted LSF distance performs significantly better than the group-delay distance [8] or inverse harmonic mean [9]. These variations of MSVQ are shown in Table 1 as MSVQ1, MSVQ2, and MSVQ3, respectively.

3.1.3. Final Codebook Training
For final codebook training, we used a larger training set consisting of about 80,000 LSF vectors from studio microphone, close-talking microphone, and high-pass filtered speech. The simultaneous joint codebook design procedure was used. The 25-bit codebook consists of four stages of 7, 6, 6, and 6 bits, respectively. This quantizer attains approximately 1 dB of $SD_{lw}$. It performs better than the 34-bit scalar quantizer used in the baseline MELP coder over a wide range of acoustic inputs. This performance improvement has been verified by numerical measurements (LSF distance, SD, and $SD_{lw}$), informal listening, and formal intelligibility testing.
2. MSVQ performance for 14th order LPC.

3.1.4. Higher Order LPC

We have also generated MSVQ codebooks for 14th order LPC filters, using a modified form of the power weighted LSF distance for codebook design and search. Table 2 shows the SD and SD\textsubscript{eq} for MSVQ codebooks designed with 4 and 5 stages, requiring 24 and 30 bits, respectively. Informal listening tests support the conclusion that 24-bit quantization of the 14th order filter results in worse performance than for the 10th order case, and even the 30-bit quantizer produces audible quantization noise.

3.2. Fourier Magnitudes

We have found that only the first ten harmonic magnitudes are transmitted, then low bit rate quantization is very effective. In informal listening, direct vector quantization using 10 bits does not introduce significant degradation compared to the unquantized Fourier magnitudes.

We have also performed a number of experiments in order to find appropriate perceptual weighting functions for the Fourier magnitudes. We have compared two reasonable perceptual weighting functions for searching this VQ codebook: bark weighting and the perceptual weighting used in CELP coders. The bark weighting weights low frequencies more than high ones, while the CELP weighting is the LPC-based perceptual weighting filter commonly used in CELP coders. In informal listening, the bark weighting is superior to the CELP weighting, which is in turn superior to the unweighted Euclidean distance. In addition, using a fixed "typical" pitch period for the bark weighting does not affect performance, and greatly reduces computation since the weights are fixed so that the perceptually-weighted codebook can be precomputed and stored.

3.3. Mel-Scale LPC

We have trained and evaluated an MSVQ for the LSF's of the 14th order MLPC filter. This quantizer uses 30 bits, in 5 stages of 6 bits each, and was designed using a portion of the TIMIT database. Traditional SD is used to evaluate quantizer performance since the mel scale is already implicit in the filter representation. For MLPC, we find that using perceptual weighting based on the inverse harmonic mean results in lower SD than either unweighted Euclidean distance or power spectral weighting.

Using joint optimization, we have designed an MSVQ codebook for MLPC which attains an SD of 1.2 dB over the TIMIT training set. Unfortunately, this codebook is not nearly as robust across different microphones as are our traditional LPC MSVQ codebooks. In particular, the MLPC quantizer results in significantly higher SD (1.6 dB) over studio microphone material from the DAM speech files, and it also produces audible quantizer distortion. Apparently, MLPC is more sensitive to the low-frequency response of the microphone than is conventional LPC. As a result, we have decided not to use MLPC in the enhanced MELP coder.

3.4. Variable Frame Rate Transmission

In the unquantized enhanced MELP coder, increasing the frame rate by reducing the speech frame size from 22.5 to 10 ms results in a clear improvement in speech quality. Unfortunately, direct quantization at such a high frame rate is not feasible at low bit rates. Instead, we have experimented with two different versions of variable frame rate (VFR) systems. The first method, known as block VFR [4], organizes frames into N-frame blocks and then transmits the M frames which minimize the total error for the block. The parameters for untransmitted frames are filled in by interpolation, and the last frame of each block is always transmitted to minimize delay. Our second method [10] removes this constraint and allows any M frames to be selected, and uses a delayed decision approach based on dynamic programming to model the effect of future frames on the current block. The extra degree of freedom in this method allows shorter block sizes to be used with equivalent performance to block VFR. Decreased block size results in reduced coder delay, although some delay overhead is incurred by the delayed decision process.

With large block sizes, these approaches can significantly reduce the average frame rate while maintaining high quality, but they introduce significant delay. At more modest delay values (around 200 ms), the VFR systems still provide good performance. Unfortunately, the Federal Standard competition requires a total coder delay (including serial channel transmission delay) less than 180 ms, and also limits the complexity of the speech coder. Given these constraints, we find a fixed frame rate system to perform as well as the VFR approaches in the MELP coder.

4. CODER OPTIMIZATION

We have optimized the 2.4 kbit/s MELP coder for the Federal Standard competition. In particular, the robustness to channel errors has been improved and the coder has been implemented in real time. Finally, formal subjective testing has been performed for a number of test conditions.

4.1. Robustness to Channel Errors

Four steps have been taken to improve the performance of the MELP coder in the presence of channel errors. First, index reassignment using simulated annealing [11] has been applied to both the LSF and Fourier magnitude quantizers. In both cases, this results in a factor of two improvement in weighted mean-squared error for 1% bit errors and a significant improvement in perceived quality. Second, a frame repeat mode has been implemented for erasures. Third, forward error correction (FEC) has been added for unvoiced frames. Finally, an improved quantization procedure is used to code the speech gain.

The FEC uses the same approach as in the LPC-10 standard: for unvoiced frames, unneeded parameters are replaced by Hamming codes [12]. The Fourier magnitudes, bandpass voicing, and jitter bit are not used for unvoiced frames, so these 13 bits are used for three Hamming (7,4) codes and one Hamming (8,4) code. The (7,4) codes correct single bit errors, while the (8,4) code also detects double bit errors. The first stage of the LSF MSVQ (7 bits) and all gain information (8 bits) are protected. The unvoiced state is signaled by an all-zero pitch code, and pitch codes near in Hamming distance to the all-zero code are reserved to protect against bit errors. Unfortunately, informal listening tests show no significant improvement under bit errors when FEC is used. Presumably, this is because most of the distortion occurs in voiced frames.

During voiced frames, no error protection is applied and the performance of the coder is clearly degraded in bit errors. To reduce the dominant problem of audible gain ex-
Table 3. MELP coder bit allocation

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Voiced</th>
<th>Unvoiced</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSP's</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>Fourier magnitudes</td>
<td>8</td>
<td>-</td>
</tr>
<tr>
<td>Gain (2 per frame)</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Pitch and overall voicing</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Bandpass voicing</td>
<td>4</td>
<td>-</td>
</tr>
<tr>
<td>Aperiodic flag</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>Error protection</td>
<td>-</td>
<td>13</td>
</tr>
<tr>
<td>Sync bit</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total bits / 22.5 ms</strong></td>
<td><strong>54</strong></td>
<td><strong>54</strong></td>
</tr>
</tbody>
</table>

Recursions during steady-state vowels, we have developed a new method for coding the speech gain. In the baseline MELP coder, the gain is sent twice per frame. The gain for the second subframe is quantized with a uniform quantizer, with 5 bits covering the entire speech dynamic range. The first subframe is coded with 3 bits covering a smaller dynamic range based on neighboring subframe values. We have introduced a special code for the first subframe to denote the steady-state gain condition, when the gain does not change by more than a threshold. This introduces redundancy into steady-state frames, so that the receiver can detect invalid excursions in the second gain term when the interpolation code was sent. In informal listening tests, this new gain coding scheme provides a clear improvement during bit error conditions, while having no effect on clean channel performance.

4.2. Listening Tests

Formal subjective testing was performed on a 25 ms frame size version of the enhanced 2.4 kbit/s MELP coder, including A/B pairwise comparisons, DAM testing, and Diagnostic Rhyme Test (DRT). A/B testing was performed against both the baseline MELP coder and the 4.8 kbit/s FS1016 CELP. The A/B test was forced choice, with three listeners unfamiliar with these coders evaluating almost two hundred sentences over a wide range of acoustic conditions. The new MELP coder was preferred over the baseline coder 64% to 36%, and over FS1016 by 55% to 45%.

We also performed DAM and DRT testing over a smaller set of test conditions. These results were ambiguous, with no statistically significant difference between the enhanced and baseline MELP coders. One surprising result, however, was that the enhanced coder had lower intelligibility scores for female speakers. This was due to the 25 ms frame size, so the final MELP coder has been redesigned using 22.5 ms frames. A second DRT test performed on the 22.5 ms frame size MELP coder confirmed that the intelligibility for female speakers was preserved.

4.3. Implementation

The 2.4 kbit/s MELP candidate for the Federal Standard uses a 22.5 ms frame size with the bit allocation shown in Table 3. This coder has been implemented in real-time on a floating point DSP (60 MHz TMS320C31) using a mixture of C and assembly code. The coder requires a total of 20 MIPS, with 14 MIPS for the encoder and 6 MIPS for the decoder. Memory requirements are 9 kwords for the program, 6 kwords for tables, and 6 kwords of RAM.

5. CONCLUSION

We have developed an enhanced 2.4 kbit/s MELP coder as a candidate for the new U.S. Federal Standard. The new coder employs Fourier magnitude coding to improve the speech quality, as well as vector quantization techniques to efficiently encode the LPC and Fourier information. Listening tests show that this coder performs as well as the higher bit rate 4.8 kbit/s FS1016 CELP standard, and we expect this enhanced MELP coder to meet all of the requirements for the new Federal Standard.

6. ACKNOWLEDGEMENTS

The authors would like to thank Stephen McGrath for implementing the real-time system, Wai-Ming Lai for coordinating listening tests, and Wilf LeBlanc for assistance in training vector quantization codebooks.

REFERENCES


