SINE-WAVE AMPLITUDE CODING USING A MIXED LSF/PARCOR REPRESENTATION

Robert B. Dunn1, Robert J. McAulay1, Terry G. Champion2, Elliot Singer1 and Thomas F. Quatieri2
1MIT Lincoln Laboratory, Lexington, MA
2RL/ERT, Rome Laboratory, Hanscom AFB, MA

ABSTRACT
An all-pole model of the speech spectral envelope is used to code the sine-wave amplitudes in the Sinusoidal Transform Coder [1, 2]. While line spectral frequencies (LSFs) are currently used to represent this all-pole model, it is shown that a mixture of line spectral frequencies and partial correlation (PARCOR) coefficients [3] can be used to reduce complexity without a loss in quantization efficiency. Objective and subjective measures demonstrate that speech quality is maintained. In addition, the use of split vector quantization is shown to substantially reduce the number of bits needed to code the all-pole model.

1. INTRODUCTION
In the Sinusoidal Transform Coder (STC), the sine-wave amplitudes are not coded directly, but are transmitted via the parameters of an all-pole model of the spectral envelope. In this paper it is demonstrated that the computational complexity of STC can be reduced by using a mixture of line spectral frequencies (LSFs) and partial correlation (PARCOR) coefficients to represent the all-pole model instead of using only LSFs to represent the model. The mixed LSF/PARCOR representation has been used to increase the quantization efficiency of 10th order linear predictive coding [3], but in STC, improvements in quality have been obtained by using model orders much higher than the basic 10th order system. It is shown that by using the mixed LSF/PARCOR representation instead of a purely LSF representation, complexity is significantly reduced while quantization efficiency is maintained.

2. PARAMETER COMPUTATION
The LSFs for an all-pole model of order 8 and less can be computed in closed form [4] but if the model order is greater than 8, numerical root solving techniques must be used and these are computationally intensive. For example, computing LSFs for an 8th order model with the closed form solution requires 0.21 ms on a 40 MHz TMS320C30, but computing the LSFs using root solving requires 0.65 ms. When using the mixed LSF/PARCOR representation, the number of LSFs can be limited to 8 so the more efficient closed form solution can be used. In [3] the PARCOR coefficients of the mixed representation were computed from the quantized LSFs using the autocorrelation-lattice method [5], thus possibly compensating for some of the quantization error. But for a 14th order model (using 8 LSFs and 6 PARCOR coefficients) there is no loss in quantization efficiency if instead of using the autocorrelation-lattice method to compute the PARCOR coefficients, they are computed directly from the unquantized predictor coefficients using Durbin’s recursion. Using Durbin’s recursion instead of the autocorrelation-lattice method reduces the time needed to compute the parameters of the mixed representation by a factor of two.

Single precision floating point arithmetic is not sufficiently accurate for computing predictor coefficients from LSFs when the all-pole model order is 14 or greater, but real-time vocoders are implemented on digital signal processing (DSP) chips such as the TMS320C30 which have only single precision arithmetic. To demonstrate the problem, predictor coefficients for 4500 frames of speech were computed from LSFs using the TMS320C30 and the error in the spectral envelope caused by the limited accuracy of the TMS320C30 was observed. Table 1 shows the largest error in the spectral envelope observed when computing the spectral envelope with two different algorithms. For 8th and 10th order models the error is negligible but for 18th and 22nd order models the error is over 40 dB. The largest error observed over the 4500 frames for the 14th order system is about 1 dB, but listening tests performed over a larger database revealed that using single precision arithmetic for a 14th order LSF system occasionally results in speech frames with unacceptably large spectral error.

The precision of floating point numbers on the TMS320C30 can be increased from about 6 decimal digits to more than 13 decimal digits by using double length arithmetic [6]. The spectral error is reduced to a maximum of 0.002 dB (for model orders up to 22) when double length arithmetic is used, but use of double length arithmetic in-

<table>
<thead>
<tr>
<th>Model</th>
<th>Algorithm</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Order</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.025</td>
<td>0.030</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.092</td>
<td>0.060</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>1.108</td>
<td>0.898</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>43.63</td>
<td>55.87</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>72.58</td>
<td>81.12</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Largest spectral error (in dB) when computing predictor coefficients from LSFs using single precision arithmetic.
increases the computation time for the transformation by a factor of from 5 to 10. By using the mixed LSF/PARCOR representation, the order of the LSFs may be limited to 8 so that use of double length arithmetic can be avoided, while the overall model order may still be 14 or greater. The mixed LSF/PARCOR representation then requires only one fourth of the computation time needed by a purely LSF representation. Table 2 shows computation times for both representations as measured over 4500 frames of speech using a TMS320C30 with a 40 MHz clock.

### Table 2. Total time needed to compute all-pole model parameters and to transform coded parameters to predictor coefficients.

<table>
<thead>
<tr>
<th>Representation</th>
<th>Model Order</th>
<th>Time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSF</td>
<td>14</td>
<td>2.31</td>
</tr>
<tr>
<td>LSF/PARCOR</td>
<td>14</td>
<td>0.56</td>
</tr>
<tr>
<td>LSF</td>
<td>22</td>
<td>4.25</td>
</tr>
<tr>
<td>LSF/PARCOR</td>
<td>22</td>
<td>0.73</td>
</tr>
</tbody>
</table>

3. PARAMETER CODING

The quantization efficiency of the mixed LSF/PARCOR representation was compared to that of the 14th order LSF representation currently used in STC for speech coding at 2400 bits per second. The quality of the speech synthesized from both systems was evaluated using a perceptually weighted spectral distortion measure and using blind AB listening tests. The perceptually weighted spectral distortion measure is

\[ SD_w(i) = \frac{1}{\pi} \int_0^\pi W(\omega)[10 \log_{10} P_i(\omega) - 10 \log_{10} \tilde{P}_i(\omega)]^2 d\omega \]  

where \( P_i(\omega) \) is the power spectrum of the all-pole model on frame \( i \), \( \tilde{P}_i(\omega) \) is the power spectrum of the quantized all-pole model on frame \( i \) and \( W(\omega) \) is a perceptually based weighting function that improves correlation of the objective measure with subjective test results [7].

Scalar quantization of the mixed LSF/PARCOR representation was compared to the current version of STC which uses scalar quantization of a purely LSF representation for the all-pole model of the sine-wave amplitudes. Vector quantization of the purely LSF representation and the mixed representation were also evaluated. In low-rate STC frame-fill techniques are used to encode the all-pole model on alternate subframes, however, in the comparisons below every subframe is quantized because frame-fill techniques for the mixed LSF/PARCOR representation have not yet been studied. The scalar quantizers were designed as in [7, 8] using 50 bits and split vector quantizers were designed as in [9] using 45 bits.

Quantization efficiency was measured using a database of 6 speakers (3 male and 3 female) with 10 sentences per speaker. The difference in mean distortion between the conditions tested, shown in Table 3, is not more than 0.10 dB and in all cases less than 2.0% of the speech frames have distortions greater than 2 dB. This demonstrates that both forms of the mixed representation perform about as well as the 14th order LSF system. Blind AB listening tests were done with 5 subjects using a subset of the test speech material. The results of these tests showed that there is no loss in quality when the mixed LSF/PARCOR representation is used instead of a purely LSF representation, and that 45 bit vector quantization of the mixed representation did not reduce quality compared to 50 bit scalar quantization of a purely LSF representation.

### Table 3. Weighted spectral distortion (in dB) of quantized 14th order all-pole models.

<table>
<thead>
<tr>
<th>Parameter Coding</th>
<th>Bits</th>
<th>Mean ( SD_w ) in dB</th>
<th>% ( SD_w ) &gt; 2 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSF-SQ</td>
<td>50</td>
<td>0.69</td>
<td>0.79</td>
</tr>
<tr>
<td>LSF/PARCOR-SQ</td>
<td>50</td>
<td>0.73</td>
<td>0.82</td>
</tr>
<tr>
<td>LSF-VQ</td>
<td>45</td>
<td>0.76</td>
<td>0.62</td>
</tr>
<tr>
<td>LSF/PARCOR-VQ</td>
<td>45</td>
<td>0.79</td>
<td>1.33</td>
</tr>
</tbody>
</table>

4. SUMMARY

The computational complexity of sine-wave amplitude coding for STC has been reduced by representing the all-pole model in terms of a mixture of LSFs and PARCOR coefficients. Using objective and subjective measures of performance, it has been shown that this was accomplished without a reduction in speech quality. In addition, it has been demonstrated that the use of split vector quantization can reduce the number of bits needed to code the spectrum from 50 to 45. The extra bits can then be used to code other parameters with more fidelity, thereby increasing the overall quality of the system. In future work, the frame-fill properties of the mixed LSF/PARCOR representation will be investigated.

REFERENCES


