HIGH QUALITY MELP CODING AT BIT-RATES AROUND 4 KB/S

Jacek Stachurski *Alan McCree* *Vishu Viswanathan*

DSP Solutions R&D Center, Texas Instruments, Dallas, TX
E-mail: [jacek | mccree | vishu]@csc.ti.com

ABSTRACT

Recently, a number of coding techniques have been reported to achieve near toll quality synthesized speech at bit-rates around 4 kb/s. These include variants of Code Excited Linear Prediction (CELP), Sinusoidal Transform Coding (STC) and Multi-Band Excitation (MBE). While CELP has been an effective technique for bit-rates above 6 kb/s, STC, MBE, Waveform Interpolation (WI) and Mixed Excitation Linear Prediction (MELP) [1, 2] models seem to be attractive at bit-rates below 3 kb/s. In this paper, we present a system to encode speech with high quality using MELP, a technique previously demonstrated to be effective at bit-rates of 1.6–2.4 kb/s. We have enhanced the MELP model producing significantly higher speech quality at bit-rates above 2.4 kb/s. We describe the development and testing of a high quality 4 kb/s MELP coder.

1. INTRODUCTION

Two approaches have been pursued to obtain coded speech of high quality at bit-rates around 4 kb/s. In the first approach, an attempt is made to lower the bit-rate without lowering the speech quality in coders that are capable of achieving the required quality at higher bit-rates. A number of CELP-based algorithms have been developed with reported results close to the specified goals.

In the second approach, the efforts are directed at improving the quality of encoded speech for coders whose performance tends to saturate above 3 kb/s. Modifications to the Prototype Waveform Interpolation [3] technique resulted in a Waveform Interpolation (WI) model which has been reported to achieve toll quality reconstructed speech provided that sufficiently high parameter rate is used [4]. Coders based on STC and MBE are also believed to be close to achieving, and under certain conditions actually achieve, toll quality around 4 kb/s. Other frequency domain techniques such as the Mixed Sinusoidally Excited Linear Prediction (MSELP) [5] have shown promise. Competitive with CELP at 4 kb/s, these techniques seem more likely to eventually achieve toll-quality speech at even lower bit-rates.

In this paper, we investigate the MELP technique which has been shown effective at bit-rates below 3 kb/s. We present the results of our experiments aimed at improving the performance of the MELP model. In previous MELP implementations, the Fourier coefficients, which represent the periodic part of the LP excitation, are specified by their amplitudes only. We examine a MELP model in which complex coefficients are used in the synthesis. We analyze the importance of individual MELP parameters based on the quality loss introduced when the parameter is encoded less often or with fewer bits. A 4 kb/s implementation of a MELP coder is described and the results of subjective listening tests are presented.

2. MELP CODING

2.1. The MELP Model

In MELP synthesis, the Linear Prediction (LP) all-pole filter is excited by a signal which is constructed from periodic and noise contributions [1, 2]. At the encoder (Fig. 1), the LP parameters are determined and the LP residual is obtained. The pitch is estimated from low-pass filtered speech and the voicing strengths are evaluated based on the correlation maxima of the band-pass filtered signal. The voicing strengths determine how much the periodic and the noisy parts contribute to the LP excitation in specific frequency bands. They describe, in effect, the periodicity present in the signal as a function of frequency. The Fourier coefficients define the spectral characteristics of the periodic part of the LP excitation. In previous MELP coders, the Fourier coefficients were implemented as amplitudes only, although both amplitudes and phases can be estimated. The Fourier coefficients are usually calculated from the FFT of the windowed LP residual, evaluated at the harmonic frequencies specified by the identified pitch. Gain analysis can be performed either on the LP residual or directly on the speech signal, pitch-synchronously or with a fixed length window. We obtain good results by calculating the gain from the windowed LP residual.

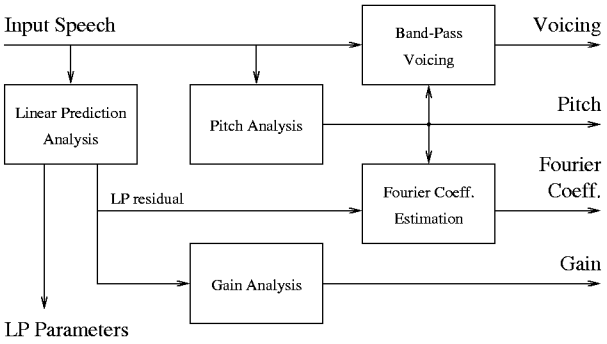

Figure 1: MELP analysis

At the decoder (Fig. 2), the periodic part of the excitation is generated from the interpolated Fourier coefficients. Fourier synthesis is applied to spectra in which the Fourier coefficients are placed at the harmonic frequencies derived from the interpolated pitch. This synthesis is described by the formula

$$x[n] = \sum_k X_n[k]\, e^{jk\phi_n}$$

with

$$\phi_n = \phi_{n_0} + \sum_{n_0+1}^{n} \omega_i$$

or, equivalently,

$$\phi_n = \phi_{n-1} + \omega_n$$

where $X_n[k]$ and ω_n are the Fourier coefficients and the normalized fundamental frequency, respectively, both interpolated for time n. In earlier versions of MELP, the Fourier coefficients and the fundamental frequency were assumed constant within one pitch period, thereby transforming the above equations to a pitch-synchronous inverse DFT. In our current implementation, the Fourier coefficients and the pitch are interpolated sample-by-sample[†].

The noisy part of the excitation is generated from white noise. The frequency bands of the periodic and the noise signals are shaped through time-domain filtering according to the transmitted voicing information. The two components of the excitation are added and the signal is scaled by the encoded gain. Finally, linear prediction synthesis is performed.

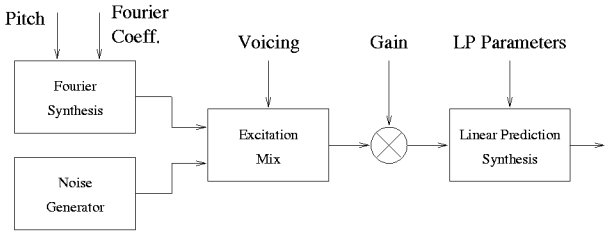

Figure 2: MELP synthesis

2.2. WI versus MELP

Conceptually, the Waveform Interpolation (WI) [4] and MELP technique have many similarities. Both techniques use a combination of periodic and noisy signals to excite a time-varying all-pole filter. At the decoder, the spectra corresponding to the periodic part of the excitation are interpolated. From the series of interpolated spectra, the time-domain LP excitation is created. In WI, the noisy and the periodic parts of the excitation are usually added in the frequency domain prior to the Fourier synthesis. In MELP, the Fourier synthesis is applied to the spectra which represent the periodic contribution with the periodic and the noisy parts mixed in the time domain.

The major difference between WI and MELP is in the way the periodic and noisy parts of the LP excitation are estimated at the encoder. In WI, the periodic part of the excitation is represented by Slowly Evolving Waveforms (SEWs), and the noisy part by Randomly Evolving Waveforms (REWs). The SEWs are estimated by filtering individual harmonics of DFT spectra with respect to their evolution in time. The REWs are set to the difference between the unfiltered spectra and the SEWs. In effect, the periodic part of the LP excitation is obtained from the frequency harmonics averaged over time.

In MELP, the periodic part of the excitation is derived from harmonics estimated from Fourier coefficients calculated via FFT. The length of the FFT window is relatively long with respect to one pitch period. The estimated harmonics can therefore be interpreted as averages over time, as in the WI technique. The MELP Fourier coefficients contain harmonics which were averaged over time but, unlike the SEWs, their overall energy does not fluctuate. The energy of the coefficients is normalized to one, and only the voicing strengths define their relative contribution to the LP excitation. These voicing strengths determine the spectral shaping of the periodic and noisy parts of the excitation. They correspond to the energy ratio of the SEWs and the REWs as a function of frequency.

In WI, filtering a series of spectra provides, at a higher computational expense and increased coder delay, not only representation of the periodic and noisy parts of the LP excitation but also, via the SEW phase, the "fine structure" of pulses present in the voiced excitation. The fine structure of the pulses is missing in the versions of MELP which implement the Fourier coefficients as amplitude only. At low bit-rates, the number of bits available is not sufficient to encode magnitude and phase of the Fourier harmonics. At higher bit-rates, however, at least partial encoding of both components may lead to improved performance.

Based on these observations, we believe that, like the WI model, the MELP model is capable of providing progressively higher speech quality as we increase the bit-rate.

2.3. Complex Fourier Coefficients in MELP

Tests with complex Fourier coefficients representing the periodic part of the LP excitation in MELP have been reported in [2]. Complex coefficients were not used then due to a low bit-rate of the system (2.4 kb/s). When more bits are available, some encoding of the phase information might be beneficial. Therefore, we tested the advantages of using the complex Fourier coefficients as opposed to Fourier magnitudes only. We estimated the amplitudes, as before, from the FFT coefficients. The corresponding phases were derived from pitch-synchronous DFT coefficients. A linear phase was introduced to the estimated phases to produce maximum correlation between consecutive complex spectra.

Such a system, however, was found to be not as good as the magnitude-only version. With the addition of the phase, an excessive coarseness was introduced into the synthesized signal.

[†]In a continuous-signal description the quantity $\sum_{n_0+1}^{n} \omega_i$ translates into an integral $\int_{t_0}^{t} \omega(\tau)d\tau$. In our discrete formulation ω is constant between samples.

Smoothing of the consecutive phases was then implemented. In one approach the phase was extracted from an average of a few spectra. The performance of the modified system was better than the amplitude-only version for some speaker-sentence pairs only. For many sentences, the amplitude-only synthesis was still preferred.

In another approach, the maximum variation of the phase was limited to a predetermined constant. When the constant was set to zero, the system was phase-locked. Increasing this constant had the effect of improving the perceptual quality for some synthesized sentences, but the improvements were not consistent.

Further research is warranted to take advantage of the combined amplitude and phase synthesis in the MELP model. The variations in phase introduce a non-periodic component into the periodic part of the excitation. The system might benefit, therefore, from voicing-strength estimation which takes into account the phase variations.

3. THE 4 KB/S MELP CODER

With no clear advantage of using complex Fourier coefficients within the tested MELP model, we decided to represent the coefficients using magnitudes only. The other coded parameters are: LP coefficients, pitch, voicing, and gain.

3.1. Parameter Rate

All parameters were estimated every 10 ms (100 Hz). To quantize the parameters at this rate, about 6 kb/s are needed. To reduce the rate we investigated the relative importance of every parameter by slowly decreasing its rate and linearly interpolating for the intermediate values. We measured the SNR of the modified system with respect to the reference (all parameters at 100 Hz rate). The average segmental SNR for the LP parameters at 50 Hz was 10 dB and the SNR for the Fourier magnitudes at 50 Hz was as high as 30 dB. However, the lower rate of either parameter resulted in little, if any, audible distortion. The worst degradation was observed when pitch and voicing analysis (combined in the system) was slowed down to 50 Hz. Although it might be possible to reduce the rate of the pitch, a high rate for the voicing strengths seems necessary.

3.2. Quantization

The LP parameters are quantized in the LSF domain with switched predictive multi-stage vector quantizer (MSVQ), similar to that described in [6]. Two jointly optimized codebooks are used, one with a strong predictor and one with a weak predictor (both first order). One bit is transmitted to indicate the codebook selected. The codebook with weak predictor supplies vectors whenever a relatively large change in the spectral envelope occurs (in our case roughly one third of the time), while the codebook with strong predictor is selected when the evolution of the spectrum is smooth.

Predictive quantization is often avoided if the coder is to be used in an environment in which frame erasures occur. We improved the performance of our coder by devising a simple protection for such a case. In the presence of frame erasures, an error will propagate throughout the series of frames for which the strongly predictive codebook is consecutively selected. If the error occurs in the middle of the series, the exact evolution of the spectral envelope is compromised but only limited perceptual distortion is introduced. When a frame erasure happens within a region where the weakly predictive codebook is consistently selected, the effect of the error will be localized by the the weak predictor. The largest degradation in the reconstructed speech quality is observed for erasures in a weakly-predictive frames followed by a series of strongly-predictive frames. In this case the evolution of the spectral envelope is "built up" on the spectrum which is very different from the one which is supposed to start the evolution. To prevent that, the first frame which, based on error minimization, would be encoded with a strongly predictive codebook is forced to use the weakly predictive one. In case an error occurs in that frame, the repeated spectral parameters from the previous frame (based on which the further evolution is "built") supply similar characteristics.

The gain and the pitch are quantized with a uniform scalar quantizer in the logarithmic domain. The parameters are quantized within full range at the frame boundaries. The intermediate values are quantized within an adaptive range determined from the adjacent quantized parameters.

The voicing strengths are estimated in five frequency bands based on maximum correlation values of band-pass filtered speech [6]. Each frequency band is classified as highly voiced, medium voiced, weakly voiced, or unvoiced. In the encoding of all the combinations, a variant of a cut-off frequency rule is applied in which higher frequencies cannot be assigned a higher voicing level than lower frequencies. We verified experimentally that with three bits, 97% of the time the voicing-strength estimates are encoded exactly.

The Fourier magnitudes are quantized, similarly to the LSFs, with switched predictive MSVQ. A weighted error criterion which favors more accurate quantization of the lower harmonics is used [2]. We do not interpolate the harmonic-magnitudes vector to a fixed dimension. Given that the interpolated values carry less information about the estimated harmonics, we are not convinced of the benefits of this process. Instead, we classify the Fourier magnitude vectors into shorter and longer groups, less than 55 and more than 45 harmonics respectively (the ranges overlap so that some spectra are included in both groups). Two codebooks (one strongly predictive and one weakly predictive) were trained for the two groups, providing four codebooks in total. In the coding procedure, a longer vector is truncated to the size of the shorter one. If the truncated vector is the one to be coded, it is then extended to its original size with constant entrants so that the average energy of the vector elements is equal to one. A set of two codebooks (one strongly and one weakly predictive) is chosen based on the quantized pitch value.

3.3. Bit Allocation

The bit allocation of the 4 kb/s coder is presented in Table 1. The LP parameters and the Fourier magnitudes are quantized every

20 ms. We allocated four bits to represent the interpolation paths of those parameters. We tested three combinations for the bit assignment: all four bits used for the interpolation of the LP parameters, all four used for the interpolation of the Fourier magnitudes, two bits for the former and two bits for the latter. The best results were obtained when all four bits were assigned for the interpolation of the LP parameters.

Parameter	Bits/frame	Frame size (ms)	Bit Rate (kb/s)
LSF coeff.	24 + 4	20	1.40
Gain	5 + 3	20	0.40
Pitch	8 + 5	20	0.65
Voicing	3	10	0.30
Fourier magn.	22	20	1.10
Parity bits	2	20	0.10
Total			4.00

Table 1: Bit allocation in the 4 kb/s coder

The gain and the pitch are estimated every 10 ms. A bit reduction in representing those parameters is obtained by assigning a smaller number of bits to quantize every other parameter value. The smaller number of bits span a reduced quantization range. The range is determined from the adjacent values of the quantized parameters.

The performance of the coder was found to be relatively better in the presence of random frame erasures than random bit errors. Two parity bits are therefore used to protect the most sensitive bits. If a parity error is identified, we treat it as a frame erasure and handle it accordingly.

Two significant differences are worth noting with respect to the bit allocation reported in [5]. We found that representation of the voicing strengths at a higher rate is very important. It was found beneficial to represent the Fourier magnitudes less often with all the available bits rather than more often but less accurately.

4. SUBJECTIVE EVALUATIONS

In initial informal A/B (pairwise) listening tests, five listeners compared the MELP model with the MSELP model, which was found, in similar tests, to be at least as good as the 32 kb/s ADPCM [5]. Sixteen sentences spoken by two male and two female speakers were used. In these tests, the performance of the MELP model was determined to be as good as, or better than, the MSELP model. In another A/B test, the unquantized MELP was found to be as good as the ITU toll-quality 8 kb/s standard G.729.

A more extensive set of listening tests was then conducted using 15 listeners and 24 sentence pairs spoken by two males and two females. In these tests, for clean input speech, the 4 kb/s MELP coder performed better than the GSM Full Rate 13 kb/s standard but not as well as the ITU's G.729. However, the 4 kb/s coder was as good as G.729 in the presence of background car noise and under the condition of 3% random frame erasures.

5. CONCLUSIONS

We have developed a MELP coder design with the goal of producing significantly higher quality as we increase the bit-rate above 2.4 kb/s. From an extensive speech quality optimization study, we have found the following results. A high transmission rate for the voicing strengths and an accurate encoding of the LP parameters are perceptually important. The Fourier coefficients, on the other hand, can be encoded less often and using fewer bits without significantly degrading the speech quality.

Predictive coding of the LP parameters and the Fourier magnitudes, when combined with switched prediction and appropriate error protection, produces satisfactory coder performance in the presence of frame erasures. Classifying the Fourier coefficients into groups of similar lengths and coding the groups separately without interpolation to a fixed length vector is a viable solution for dealing with the problem of variable dimension vectors.

Finally, we have shown that the MELP coder is capable of delivering high quality synthesized speech at 4 kb/s. With further research, we hope to be able to achieve toll-quality speech at this bit-rate.

6. ACKNOWLEDGMENTS

We would like to thank Wai-Ming Lai for help in conducting the subjective listening tests.

7. REFERENCES

[1] 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, pp. 242–250, July 1995.

[2] A. McCree, K. Truong, E. B. George, T. P. Barnwell III, and V. Viswanathan, "A 2.4 kbit/s MELP Coder Candidate for the New U.S. Federal Standard," in *Proc. IEEE Int. Conf. ASSP*, (Atlanta), pp. 200–203, May 1996.

[3] W. B. Kleijn, "Encoding Speech Using Prototype Waveforms," *IEEE Trans. on Speech and Audio Processing*, vol. 1, pp. 386–399, Oct. 1993.

[4] W. B. Kleijn and J. Haagen, "Waveform Interpolation for Coding and Synthesis," in *Speech Coding and Synthesis*, pp. 175–208, Elsevier, 1995.

[5] S. Yeldener, J. C. De Martin, and V. Viswanathan, "A Mixed Sinusoidally Excited Linear Prediction Coder at 4 kb/s and Below," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, (Seattle), May 1998.

[6] A. McCree and J. C. De Martin, "A 1.7 kb/s MELP Coder with Improved Analysis and Quantization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, (Seattle), May 1998.