IMPROVING THE 2.4 KB/S MILITARY STANDARD MELP (MS-MELP) CODER USING PITCH-SYNCHRONOUS ANALYSIS AND SYNTHESIS TECHNIQUES

Ali Erdem Ertan and Thomas P. Barnwell III*

School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332–0250
Email: ertane,tom@ece.gatech.edu

ABSTRACT

In this paper, we presented new pitch-synchronous analysis and synthesis methods for improving the quality of the 2.4 kb/s U.S. military standard-MELP coder. These improvements include: a new pitch-estimation algorithm; a pitch-cycle segmentation algorithm; an adaptive linear-prediction analysis and Fourier series computation selection method; and speech synthesis with fractional cycle lengths. In listening tests, we found that these new techniques improve the quality for female speech significantly and they completely eliminate the quality difference for the MS-MELP coder between male and female talkers.

1. INTRODUCTION

Over the past decade, advances in DSP processor technology have allowed the implementation of state-of-the-art real-time speech coders operating at 2.4 kb/s. In the early 1990s, the competition for a new 2.4 kb/s U.S. military speech coder standard accelerated these efforts. Combining more complex speech models with complex analysis and quantization methods improved the quality of low-bit rate coders significantly over the traditional LPC-10 standard. Mixed excitation linear predictive (MELP) coding was one of these techniques and it was selected as the new 2.4 kb/s U.S. military standard speech coder in 1996 [1]. A variation of this coder, MELPe, that uses a high-quality noise suppression algorithm as a front-end was later also selected as the new 2.4 kb/s NATO standard [2].

Although the MELP coder has high quality, the subjective tests performed in the standard competition's selection phase also revealed that there is a significant quality difference between male and female speakers [3]. Our experiments on this coder showed that the pitch-period estimation algorithm makes more pitch-estimation errors for female speakers than for male speakers, especially at onsets. Furthermore, because the decoder constrains the length of pitch cycles to be an integer, the evolution of the pitch-period is not natural in the synthetic speech signal. However, as the length of pitch cycles in male speech is usually longer than that in female speech, this problem mostly impacts female speakers. Finally, although the pulse-dispersion filter improves the quality slightly for male speakers, it makes the synthesized speech distinctly noisy for female speakers.

*Ali Erdem Ertan is now with DSP Solutions R&D Center in Texas Instruments.

In this paper, we present the details of a speech coding algorithm that improves the quality of female speech by using pitch-synchronous analysis and synthesis techniques. The pitch estimation mistakes for female speakers are decreased substantially by a new sub-frame based pitch tracking algorithm. In addition, we introduce a new pitch-cycle segmentation algorithm that segments the speech signal into individual pitch cycles with fractional cycle lengths. These segments are used as inputs to a circular linear prediction (CLP) algorithm and also to a constant pitch transform (CPT). The problem of the non-natural evolution of pitch periods for female speakers is significantly reduced by synthesizing cycles with fractional lengths in the decoder (using the CPT). Finally, the elimination of the pulse dispersion filter also reduces the noisy characteristic of female speech while reducing the quality of male speech only slightly. Subjective listening tests revealed that these enhancements improve the quality of female speech significantly. A version of this coder was successfully used in a 2.4 kb/s coder designed for harsh acoustic noise environments [4]. The techniques used in this coder are detailed in the following sections.

2. THE NEW 2.4 KB/S IMPROVED-MELP (I-MELP) CODER

The new features in the MELP model eliminate the unnatural quality of the LPC vocoders, improve intelligibility and provide some degree of robustness to background noise. However, the MS-MELP coder that uses this model still has problems as discussed in the introduction section. The new 2.4 kb/s I-MELP coder is designed to reduce these problems and to improve the quality of encoded speech of female speakers. In this work, we approached these problems by using new analysis and synthesis methods. For this reason, the parameter quantization remains the same as in the 2.4 kb/s MS-MELP coder. Therefore, both encoder and decoder of the 2.4 kb/s I-MELP coder are cross-compatible with the 2.4 kb/s MS-MELP coder.

2.1. The Encoder

The new I-MELP encoder extracts the same parameters as the MS-MELP coder. To improve the performance of the coder, three new elements were introduced: a new pitch-estimation algorithm; a pitch-cycle segmentation algorithm; and an adaptive linear prediction analysis and Fourier series estimation selection method.

2.1.1. The New Pitch Estimation Algorithm

The voiced/unvoiced decision and pitch period are arguably the most important parameters affecting the synthesized speech quality in a parametric speech coder. Incorrect estimation of these parameters not only results in audible artifacts, but also affects the estimation accuracy of other parameters such as Fourier series magnitudes. Because of the use of pitch-synchronous analysis methods in this work, the correct estimation of pitch period is even more crucial.

The pitch-period estimation algorithm in MS-MELP coder uses normalized correlation as the correlation measure. The normalized correlation, ρ_τ, at lag τ is defined as

$$\rho_\tau = \frac{\sum_{n=0}^{L-1} x[n]x[n+\tau]}{\sqrt{\sum_{n=0}^{L-1} x[n]^2 \sum_{n=0}^{L-1} x[n+\tau]^2}}, \quad (1)$$

where $x[n]$ is the analyzed signal and L is the number of samples used in the calculation. This correlation measure is robust to energy variations and small amount of background noise, and it is always bounded by ± 1. Naturally, the pitch lag that maximizes this measure is selected as the pitch-period of the signal. However, to reduce pitch doubling and pitch halving mistakes, MS-MELP computes the normalized correlation function many times using various low-pass filtered speech and residual signals combined with pitch-doubling elimination logic. Although this algorithm is very effective, pitch-doubling and voicing-estimation mistakes still occur at onsets, especially for female speakers. Unno et al. [5] reported that the placement of the estimation window becomes crucial in these instances. McCree et al. [6] showed that a similar but subframe based approach improves estimation accuracy, especially at onsets. This approach is based on finding a pitch track within some number of subframes that minimizes the pitch-prediction residual energy over the frame. This method is exactly equivalent to finding the integer lags that maximize the normalized correlation in the subframes. We also used this approach in our work.

Our algorithm begins by computing a pitch track for each allowed integer pitch lag assumed as the average pitch period. In each pitch track, the pitch variation is bounded by $\pm 15\%$ of the assumed average pitch period to eliminate improper pitch variations in the frame. For each pitch track, the track's normalized correlation and the true average pitch period are computed as the energy weighted mean of the subframe normalized correlation coefficients and integer pitch lags, respectively. Most of the time, the pitch track with the largest track's normalized correlation results in the pitch track whose average pitch period is the true pitch period of the frame. However, to eliminate possible pitch-doubling mistakes, the allowed integer pitch lags that are assumed as the initial average pitch period are partitioned into octave regions and a seperate pitch track is obtained for each octave region. When the ratio between the normalized correlation of a pitch track in a lower octave region and that in a higher octave region exceeds a threshold, the track from the lower octave region is selected even if it's normalized correlation is not the largest.

To eliminate the pitch-halving mistakes, it is beneficial to find the pitch track using the residual signal. However, the residual signal usually becomes very noisy at the end of the words. In these cases, the pitch track obtained from the speech signal is more reliable. For this reason, two pitch tracks are computed, one from each signal. In addition, as the high-frequency part of the partially-voiced speech spectrum is often very noisy, both signals are filtered with a low-pass filter to eliminate the effect of noise in the computation of normalized correlation.

The results from these two signals are combined using decision logic. Usually, the frame's pitch period and voicing degree is obtained from the average pitch period of the pitch track with the largest normalized correlation. However, when the correlation levels of both signals are weak, the one whose average pitch period is closer to the speech's average pitch period is selected. The speech's average pitch period is computed only when the signal has strong correlation.

In addition to these features, we also employ a pitch-smoothing algorithm when the pitch is computed more than twice in a frame. This procedure is effective in eliminating isolated pitch-estimation mistakes.

Experimentally, we observed that this algorithm estimates pitch period and voicing degree correctly almost all the time. The remaining errors mostly occurred at onsets for low-pitched speakers, especially when less than half of the analysis frame is in the voiced region. Throughout our test set, no pitch-halving problem were observed.

2.1.2. The New Pitch-Cycle Segmentation Algorithm

The correct estimation of pitch period alone does not guarantee the correct working of the pitch-synchronous algorithms; the pitch-cycle boundaries must also be located accurately. In addition, since both the CLP method and the CPT of the pitch cycles requires near-exact periodicity [7,8], the pitch-cycle boundaries must be located in fractional steps within the critically sampled narrowband speech signal.

In our algorithm, we use the normalized correlation once more as the periodicity measure, but this time, the length of the segment is also synchronized to the pitch lag for which the normalized correlation is computed (i.e. L is set to τ in (1).) This measure is maximized when τ is equal to a pitch-cycle length. Note that this algorithm assumes that the starting location of pitch cycle, $x[0]$, is known. When the speech signal is segmented continuously, this assumption is always satisfied.

Our algorithm mainly uses the low-pass filtered residual signal, as the residual signal does not contain the spectral shaping effects of the vocal-tract filter and the high-frequency noise of the partially-voiced speech sounds. However, when the largest normalized correlation is low, the algorithm switches to the low-pass filtered speech signal to find the cycle boundaries. The algorithm first searches a range of pitch lags around the average pitch of the frame obtained by the procedure described above. When the lag with largest correlation is found, the same algorithm is repeated on the N times upsampled signal within one sample distance around the integer lag estimate to find the cycle length in $\frac{1}{N}$ sample resolution. As the use of CLP and the CPT of pitch cycle requires 10 times upsampled signal [7], N is set to 10 in the I-MELP coder.

Onsets are handled differently in our algorithm. When a possible onset is detected by the pitch estimation algorithm or the energy of the current frame is increased by 6 dB with respect to that of previous frame, the signal is blindly segmented many times, once for each integer pitch lag within the allowed integer pitch periods assumed as the average. As the segmentation algorithm finds the length of pitch cycle when segmenting the signal, the blind segmentation method can also be seen as finding a pitch track for each integer pitch lag assumed as the average pitch period just as in the pitch estimation algorithm. As a result, the track's nor-

malized correlation can also be computed and the frame can be declared as a real onset when the largest track's normalized correlation exceeds a threshold. When such an onset is detected, an initial segmentation location is chosen so that the segment boundaries always reside in low-energy parts of residual, i.e. between pitch pulses. Then, the rest of the frame is re-segmented by the algorithm discussed above using the average pitch of the pitch track with largest normalized correlation. We observed that occasional voicing-estimation mistakes made by the pitch-estimation algorithm because of very rapid pitch changes at onsets are corrected with this technique.

In experiments, we observed no segmentation mistakes with this algorithm. The segment locations obtained by this method are most suitable for making pitch-cycle modifications. However, they can also be used in the CLP analysis as discussed in [7].

2.1.3. Adaptive Linear-Prediction and Fourier Series Estimation Method Selection

In the I-MELP, the encoder chooses between the autocorrelation method and the multi-cycle CLP (M-CLP) method as the linear-prediction analysis method depending on the correlation level of the signal. In addition, the number of cycles used in the M-CLP method is also chosen adaptively. The autocorrelation method is always used when one of the following conditions occurs:

- The voicing strength of the frame is less than a voicing threshold, ρ_{vc}.
- The correlation between the pitch cycle that overlaps with the last sample of the frame, *main pitch cycle*, and the pitch cycle just after the main pitch cycle is less than a cycle-correlation threshold, ρ_{cc}.
- The correlation of only a single cycle in the frame exceeds ρ_{cc} and there is more than 6 dB difference between the energy of the first cycle in the frame and the energy of this single cycle.

The first two criteria avoid the use of the M-CLP method when the speech is either unvoiced or generated by erratic glottal pulses. In these cases, the prediction filter obtained by the CLP method is not reliable and may result in audible artifacts in the synthetic speech. In addition, when there is a large energy variation between two adjacent pitch cycles such as in a vowel-nasal transition, the cycle boundaries may not be properly selected because of low correlation between adjacent pitch cycles. The last criterion deals with this problem and avoids the use of the CLP method in such cases. When the autocorrelation method is not used, the number of cycles used in the M-CLP method is selected according to the correlation of the adjacent cycles. The cycle just before and just after the main pitch cycle is also used in the M-CLP analysis if their correlation with the next pitch cycle exceeds ρ_{cc}. These steps ensure the use of multiple cycles when the speech signal is stationary and the use of single cycle when the cycle is on a transient segment (the correlation between adjacent cycles are low in this case.) In early informal listening tests, it was observed that the choice of the thresholds, ρ_{vc} and ρ_{cc}, is very important. Setting these thresholds too low results in audible artifacts in the synthesized speech. In these cases, the boundaries of the pitch cycles cannot be obtained reliably, and the prediction-filter coefficients obtained by the CLP method result in audible artifacts in the synthetic speech. On the other hand, when these thresholds are set too high, the transition regions are always analyzed with the autocorrelation method and the estimated prediction filter is inferior to the one estimated by the CLP method using properly segmented pitch cycles. Experimentally, we determined that when both ρ_{vc} and ρ_{cc} are set to a moderate voicing level, 0.75, no prediction-filter estimation related distortion is observed in the synthetic speech while the transition regions are still modeled with the CLP method.

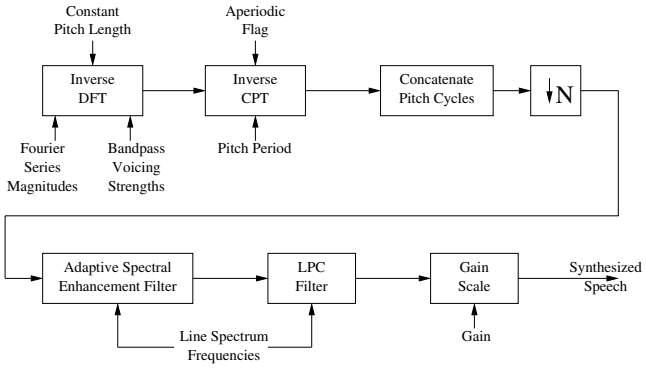

Fig. 1. The decoder of the new 2.4 kb/s improved MELP coder.

The Fourier series magnitudes are also obtained by two different methods depending on the method used in the linear-prediction filter estimation. When the autocorrelation method is used to obtain the prediction filter, the Fourier series magnitudes are obtained by the peak-picking of residual signal's FFT method used in the MS-MELP coder. When the M-CLP method is used, the main pitch cycle is circularly filtered with the inverse of the prediction filter, and then, the CPT is applied to normalize its length. The Fourier series magnitudes are obtained as the magnitudes of the first ten frequency samples of this constant-length circular residual signal's DFT. This method allows the encoder to capture the magnitudes of the harmonics reliably at onsets and transition regions from a single cycle. Note that the main pitch cycle in this case has also a sampling rate of ten times the original sampling rate.

2.2. The Decoder

The main improvement of the I-MELP decoder over the MS-MELP decoder is the generation of the excitation signal at ten times the original sampling rate, which allows us to synthesize fractional length cycles at 8 kHz. The excitation signal is then decimated prior to filtering with the adaptive spectral enhancement filter. Using this technique, we are able to synthesize speech signal with more natural variation of pitch period, especially for female speakers. The new decoder is illustrated in Fig. 1.

Similar to the MS-MELP decoder, this decoder also finds the starting locations of the new pitch cycles and synthesizes the speech pitch-synchronously. The only exception is that the interpolated pitch-cycle length is first multiplied by ten and rounded to the nearest integer to find the upsampled pitch-cycle length. The DFT of the pitch cycle is then generated at constant pitch length using the interpolated Fourier series magnitudes and the band-pass voicing strength coefficients. To reduce the computational complexity and eliminate the delay resulting from band-pass filtering, the mixed excitation is generated in the DFT domain as in [5]. The constant length mixed-excitation pitch cycle is obtained by the inverse DFT.

Then, the inverse CPT is applied to the constant length pitch cycle to modify its length to the upsampled pitch-cycle length. After all pitch cycles in the frame are synthesized, they are concatenated, low-pass filtered with the decimation filter, and then decimated to obtain the excitation signal at 8 kHz.

The rest of the synthesis process is the same as for the MS-MELP decoder except that the pulse dispersion filter is not applied to final synthetic speech. Experimentally, we observed that this filter slightly enhances the quality of the male speech, but makes the female speech distinctly noisy. Moreover, we also observed that the distortion related to pitch mistakes become less severe when this filter is applied. As the new pitch-estimation algorithm reduces the pitch mistakes significantly and the quality loss in male speech due to removal of this filter may be compensated by the new improvements, we removed this filter from the I-MELP decoder.

In the informal subjective listening tests, it was observed that the proposed 2.4 kb/s I-MELP coder's speech quality is similar to that of the MS-MELP coder for male speakers. However, the quality of the new coder for female speakers is distinctly better than that of the MS-MELP coder. The results of the formal listening test comparing the 2.4 kb/s MS-MELP coder and the new 2.4 kb/s I-MELP coder are presented in the next section.

3. RESULTS OF FORMAL LISTENING TESTS

We evaluate the performance of the new I-MELP coder by a degradation category rating (DCR) [9]. In the DCR test, the subjects rate the amount of degradation in the processed signal with respect to a reference signal using a 5 point scale (5=degradation inaudible, 1= degradation very annoying). The overall combined score is referred as degradation mean opinion score (DMOS). The test was conducted by 22 subjects. Subjects were presented 12 phrases each from a set of 64 phrases. Eight conditions, three of which are the 2.4 kb/s I-MELP coder, the 2.4 kb/s MS-MELP coder and the 2.4 kb/s I-MELP that only uses autocorrelation method for linear-prediction analysis, referred as I-MELP[AC], were evaluated. The reference signal in this test is the unprocessed critically sampled narrowband speech signal. The results were analyzed using a two-sided t-test to determine whether the average scores of the two test cases were the same or different with 95% confidence level.

The DCR test results that compare the I-MELP and the MS-MELP coders are shown in Table 1, in which the reference coder is the MS-MELP coder and the test coder is the I-MELP coder. These results clearly show that the female speech quality of the I-MELP coder is significantly better than that of the MS-MELP coder. These tests also verified that there is indeed a large quality difference between genders for the MS-MELP coder. However, this behavior was not seen for the I-MELP coder. We also found that overall quality of the I-MELP coder is significantly better that that of the MS-MELP coder. In these experiments, we also observed that the quality of both coders is similar for male speech when coders do not make pitch estimation mistakes. Unfortunately, when there is a pitch mistake in the I-MELP coder, the distortion becomes more severe because of the absence of the pulse dispersion filter and because of the additional distortion resulting from a faulty linear-prediction estimation with the M-CLP method. As shown in Table 2, when only autocorrelation method is used, the quality of the I-MELP coder is very close to that of MS-MELP coder for male speakers. These results also showed that the quality improvements of the CLP in transition regions are subtle, and therefore, hidden in other distortions resulting from the quantization at 2.4 kb/s.

Table 1. Comparison of the MS-MELP coder (reference coder) and the I-MELP coder (test coder) in the DCR test.

Gender	X_{ref}	X_{test}	X_{diff}	Result (95% Conf.)
All	3.34	3.53	-0.19	I-MELP Better
Male	3.63	3.47	0.16	Equal
Female	3.05	3.59	-0.54	I-MELP Better

Table 2. Comparison of the MS-MELP coder (reference coder) and the I-MELP[AC] coder (test coder) in the DCR test.

Gender	X_{ref}	X_{test}	X_{diff}	Result (95% Conf.)
All	3.34	3.60	-0.26	I-MELP[AC] Better
Male	3.63	3.64	-0.01	Equal
Female	3.05	3.56	-0.51	I-MELP[AC] Better

4. ACKNOWLEDGEMENT

The authors would like to thank Texas Instruments for supporting this work, and Robert Morris for his subjective listening test software.

5. REFERENCES

[1] U.S. Department of Defense, "Specifications for the analog to digital conversion of voice by 2,400 bit/second mixed excitation linear prediction," 1998.

[2] T. Wang, K. Koishida, V. Cuperman, A. Gersho, and J.S. Collura, "A 1200/2400 bps coding suite based on MELP," *Proceedings of the Speech Coding Workshop*, pp. 90–92, 2002.

[3] M.A. Kohler, "A comparison of the new 2400 bps MELP federal standard with other standard coders," *Proc. IEEE Int. Conf. Acoust. Speech Signal Processing*, pp. 1587–1590, 1997.

[4] T.P. Barnwell III, M.A. Clements, D.V. Anderson, E. Moore, M. Lee, A.E. Ertan, V. Krishnan, S. Kamath, W. Choi, J. Hu, C. Demiroglu, P.S. Whitehead, and A.S. Durey, "Low bit rate coding of speech in harsh conditions using non-acoustic auxiliary device," *Special Workshop in Maui*, pp. 290–294, 2004.

[5] T. Unno, T.P. Barnwell III, and K. Truong, "An improved mixed excitation linear prediction (MELP) coder," *Proc. IEEE Int. Conf. Acoust. Speech Signal Processing*, pp. 245–248, 1999.

[6] A.V. McCree and J.C. De Martin, "A 1.7 kb/s MELP coder with improved analysis and quantization," *Proc. IEEE Int. Conf. Acoust. Speech Signal Processing*, pp. 593–596, 1998.

[7] A.E. Ertan and T.P. Barnwell III, "Spectral estimate performance of circular linear prediction modeling for real-speech signals," *Proc. of* 38^{th} *Asilomar Conference on Signals, Systems and Computers*, p. To be published, 2004.

[8] Ali Erdem Ertan, *Pitch-Synchronous Processing of Speech Signal for Improving the Quality of Low Bit-Rate Speech Coders*, Ph.D. thesis, Georgia Institute of Technology, 2003.

[9] ITU, "Recommendations p.80 methods of subjective determination of transmission quality," 1993.