A Pitch-Synchronous Pitch-Cycle Modification Method for Designing a Hybrid I-MELP/Waveform-Matching Speech Coder

Ali Erdem Ertan

DSP Solutions R&D Center,
Texas Instruments, Dallas, TX, U.S.A.
ertane@ti.com

Thomas P. Barnwell III

School of Electrical and Computer Engineering,
Georgia Institute of Technology, Atlanta, Georgia, U.S.A.
tom@ece.gatech.edu

Abstract

In this paper, we introduce a new approach to solve encoding method switching problems in a parametric/waveform-matching hybrid coder. This method only requires modification of the waveform-matching coder's target signal. For this purpose, we introduce a new pitch-synchronous cycle length modification method using the constant pitch transformation (CPT) and a frequency domain zero-phase equalization filter. An experimental I-MELP/PCM coder is used to explore the largest achievable quality improvement over a MELP coder. Test results show that encoding transitions with waveform-encoding techniques improves the quality over a fully parametric coder.

1. Introduction

In recent years, there has been a growing interest in designing a toll-quality speech coder at 4 kb/s. The main challenge for achieving this goal is that neither parametric speech coding techniques such as MELP nor waveform-matching coders such as CELP are suitable for this purpose. Although parametric coders have very high segmental speech quality in stationary segments, they often fail to represent transition regions and short events accurately. For this reason, the quality of these coders usually saturates at about 4 kb/s and does not approach toll-quality. On the other hand, although the waveform-matching coders are capable of encoding all kinds of signals including transition effects, they fail to achieve high-quality encoding of stationary segments at 4 kb/s, as they also try to capture the perceptually unimportant phase information. To solve this problem, several researchers [1, 2, 3] have proposed techniques that use the best of these two methods to encode different parts of speech signal: parametric speech coding methods for stationary segments and waveform-matching methods for transition segments.

The main problem with this approach is switching between these two encoding methods. As parametric coders do not encode phase information, the synthesized signal is not time-synchronous with the original signal and the waveform shapes of original and synthesized signals are quite different. On the other hand, waveform-matching coders preserve phase information, and as a result, the synthesized signal is time-synchronous with the original signal and the waveform shapes are similar. For this reason, switching between these two encoding methods usually introduces audible artifacts in the synthesized speech. Previously, several techniques were proposed as a solution for this problem. These techniques include (1) the modification of original signal to eliminate phase information so that waveform-matching coder's target signal and the synthetic signal obtained by the parametric coder has similar signal shapes [1, 2], (2) the modification of original signal so that modified signal will be time-synchronized with the synthetic signal obtained by the parametric coder [1], and (3) the addition of new parameters to a parametric coder such as alignment phase [2] or relative location of the pulse closest to the frame boundary [3] so that time-synchrony between the synthetic signal obtained by the parametric coder and input speech signal is preserved.

In this paper, we will present another approach to this problem. As we would like to use an existing pitch-synchronous parametric speech coder, particularly the 2.4 kb/s I-MELP coder [4], without making any modifications to either encoder or decoder, we designed an algorithm that replaces the input speech with nearly perceptually equivalent signal as the target signal for the waveform-matching coder. The algorithm ensures that this modified signal has both a similar shape with the synthetic speech obtained by the I-MELP decoder and that time-synchrony between modified signal and the synthetic speech obtained by the I-MELP decoder is preserved. For this purpose, we will introduce a new pitch-synchronous cycle length modification algorithm and a pitch-synchronous zero-phase equalization filter in this paper. The details of this algorithm will be given in the next section.

In Section 3, we introduce an experimental I-MELP/PCM coder to explore the most achievable amount of quality improvement over an I-MELP coder by using a hybrid coding technique. Finally, Section 4 will present the subjective listening test results; in this section we will first compare the test results of I-MELP/PCM coder to those of modified speech to evaluate the voiced segment encoding effectiveness of the I-MELP coder. Finally, to investigate the largest possible quality increase over the I-MELP coder, the test results of I-MELP/PCM coder will be compared to those of 2.4 kb/s I-MELP coder.

2. Pitch-Cycle Modification of Speech Signal for Designing a Hybrid MELP Coder

The basic idea of a pitch-cycle modification algorithm is to modify the input speech signal such that it becomes time synchronous with the synthetic speech obtained by the parametric coder and the waveform shapes of the signals become similar. For this purpose, we designed an algorithm that basically changes the location and the length of the pitch cycles in the original signal. In addition, it also applies a zero-phase equalization filter to the pitch cycles so that the resulting signal wave-

Ali Erdem Ertan had done this work when he was with School of Electrical and Computed Engineering, Georgia Institute of Technology, Atlanta, Georgia, U.S.A.

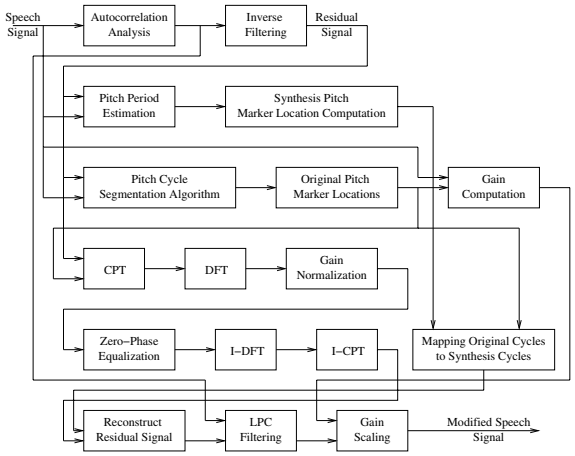

Figure 1: Flowchart of the pitch-cycle modification algorithm.

form is similar to the synthetic coder signal at the receiver. This algorithm is specifically designed to be used in conjunction with the synthesis method in the MELP model. However, it can also be used with any other fully-parametric coder whose decoder synthesizes the speech pitch-synchronously. All the changes are performed on the residual signal so as not to modify the spectrum shaping effects of the vocal-tract filter and to minimize possible audible artifacts at speech onsets.

The flowchart of this algorithm is shown in Figure 1. The algorithm first computes the location of *synthesis pitch cycles* in which the I-MELP uses to synthesize the signal pitch-synchronously in the decoder. Then, the input signal is segmented into individual pitch cycles using the pitch-cycle segmentation algorithm based on normalized correlation maximization described in [4, 5]. In this paper, these cycles are referred as *original pitch cycles*. The next -and arguably- the most important step is the mapping stage in which a suitable original pitch cycle is selected and mapped to the location of the synthesis pitch cycle. First, the pitch cycle in the original signal that includes the starting location of the synthesis pitch cycle is found. If this original pitch cycle overlaps completely with the synthesis pitch cycle in time or if significant portions of both this original pitch cycle and the synthesis pitch cycle overlap in time, this original pitch cycle is mapped to the synthesis cycle. Otherwise, the cycle just after this original pitch cycle is mapped to the synthesis cycle. This initial mapping is preserved as long as the lengths of the original pitch cycle and synthesis pitch cycle are close to each other. Otherwise, neighboring pitch cycles are also searched to find a more suitable pitch cycle for mapping. In the case that a suitable cycle cannot be found, cycles are combined to form longer cycles so that length of original pitch cycles and synthesis pitch cycles becomes closer [5].

To match the length of original pitch cycle to the synthesis one, one can simply apply a CPT [5] to the original signal to normalize its length, and then, apply an inverse CPT (I-CPT) with the synthesis pitch cycle length. However, to match the waveform shape of the modified signal with the synthetic coded signal in the decoder, we calculate the DFT of the normalized length signal first and apply zero-phase equalization filter in the frequency domain before I-CPT is applied. It should also be noted that to preserve the naturalness in the modified speech, both original pitch cycles and synthesis pitch cycles should have a resolution of 0.1 samples at 8 kHz sampling rate. As this resolution requires processing the signals at 10 times oversampled signal, the CPT not only allows cycle length modification but also reduces computational complexity significantly for the frequency domain zero-phase equalization filter. After the modified pitch cycles are concatenated to each other with the same length and location of synthesis pitch cycles, the signal is decimated back to 8 kHz and filtered with the LPC filter. This modified signal can now be used as the target signal for the waveform-matching coder. As the waveform shapes and cycle locations are the same in both this target signal and MELP synthesized speech, switching between these two encoding methods does not introduce any audible artifacts at frame boundaries.

One of the most important step in this algorithm is the frequency domain zero-phase equalization filtering. The purpose of this filter is to remove the phase component of the residual signal as it is also set to zero in synthetic signal obtained by the MELP decoder. However, simply setting the phase components of the frequency samples to zero introduces buzziness in the modified speech, as this procedure removes the noise in some of the bands in the residual signal. To preserve the correct amount of noise in the residual signal, the following method is used: The frequency samples in the bands with harmonic structure have constant phase in consecutive pitch cycles. On the other hand, the frequency samples in the bands with noisy structure have random phase. In speech signals with both kinds of excitations, both phase terms are present. Using this fact, the frequency domain representation of the K^{th} constant-length pitch cycle with a mixture of both types of excitation can be written as

$$X_K[k] = |X_K[k]|e^{j(\phi^c[k]+\tilde{\phi}^r_K[k])} \qquad k = 0, \cdots, \tau_C, \quad (1)$$

where $X_K[k]$ is the k^{th} frequency sample of the K^{th} pitch cycle, $\phi^c[k]$ is the phase term of the k^{th} frequency sample that is constant in consecutive pitch cycles, $\tilde{\phi}^r_K[k]$ is the random phase term of the k^{th} frequency sample, and τ_C is the constant cycle length. When the frequency samples of M consecutive pitch cycles are normalized to unity magnitude, and then averaged, the following equation is obtained:

$$\begin{aligned} Y[k] &= \sum_{m=1}^{M} e^{j\phi^c[k]} e^{j\tilde{\phi}^r_m[k]} \\ &= e^{j\phi^c[k]} \sum_{m=1}^{M} e^{j\tilde{\phi}^r_m[k]} \\ &= |Y[k]|e^{j(\phi^c[k]+\tilde{\phi}^r_M[k])} \end{aligned} \quad (2)$$

where $|Y[k]|$ is the magnitude of the averaged frequency sample k, and $\tilde{\phi}^r_M[k]$ is another random phase term obtained after averaging. Note that the zero-phase equalization filter can easily be computed by normalizing $Y[k]$ and inverting the phase components. The resulting filter can be written as

$$H_{0\phi}[k] = e^{-j(\phi^c[k]+\tilde{\phi}^r_M[k])}. \quad (3)$$

The zero-phase equalized frequency samples can be obtained as

$$\begin{aligned} \tilde{X}_K[k] &= X_K[k]H_{0\phi}[k] \\ &= |X_K[k]|e^{j(\tilde{\phi}^r_K[k]-\tilde{\phi}^r_M[k])} \\ &= |X_K[k]|e^{j\tilde{\phi}^r_{\tilde{K}}[k]}, \end{aligned} \quad (4)$$

where $\tilde{\phi}^r_{\tilde{K}}[k]$ is also a random phase component. As a result, using this technique, it is possible to eliminate the constant

phase term while replacing the original random phase with another random phase. In this pitch-cycle modification algorithm, the zero-phase equalization filter is computed for each original voiced pitch cycle from the frequency samples of the cycle itself and the two cycles adjacent to this cycle. As the zero-phase equalization filter is only applied to pitch cycles in voiced speech segments and voiced pitch cycles in transition segments, the unvoiced speech segments and the short events such as stop consonants are not modified.

An input speech signal segment, the same segment processed with this pitch-cycle modification algorithm and the same segment encoded by the 2.4 kb/s I-MELP coder are shown in Figure 2. The slight difference in the waveforms of the segment processed by the pitch-cycle modification algorithm and the segment encoded by the 2.4 kb/s I-MELP coder results from the adaptive spectral enhancement filter (ASEF) used in the 2.4 kb/s I-MELP coder. Fortunately, when the encoder separates the residual signal and linear prediction filter and transmitted them separately, and ASEF filter is applied to the signal continuously, no audible distortion is introduced as ASEF filter will also be applied at transition segments.

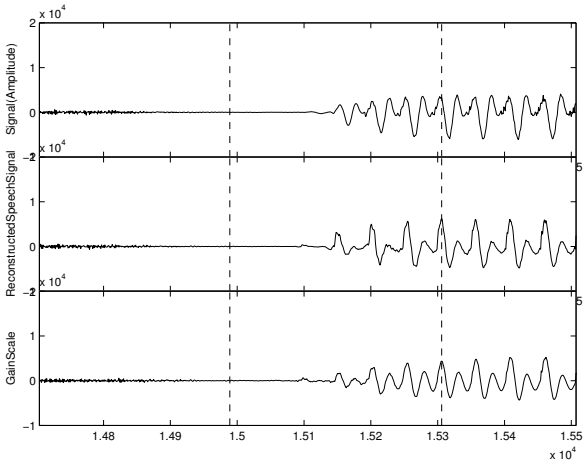

Figure 2: The input speech segment (top), the same segment processed by the pitch-cycle modification algorithm (middle) and the same segment encoded by the 2.4 kb/s I-MELP coder (bottom).

In the early informal listening tests performed on ten sentences from the TIMIT database, we observed that the modified input speech sounds very close to the input speech. Not surprisingly, it was also observed that any mistake in the pitch estimation algorithm introduces artifacts in the modified speech.

3. An Experimental Hybrid I-MELP/PCM Coder

The pitch-cycle modification algorithm described above modifies the input speech such that the modified signal can be used as the target for the waveform encoding method. As the cycle locations and signal shape is the same in both modified signal and the synthetic I-MELP coded signal, encoding method switching between the I-MELP coder and the waveform-matching coder does not introduce any audible distortion at frame boundaries. Since the quality of the synthesized voiced speech obtained by the MELP coder is very good even at 2.4 kb/s, it is logical to use the MELP model for the voiced speech segments, and encode the waveform of the signal for unvoiced and transition segments. In this initial implementation, the 2.4 kb/s I-MELP coder discussed in [4] is used to encode the voiced segments and the waveform of the residual signal in unvoiced and transition frames is transmitted to the decoder in the form of normalized-length cycles encoded with 16-bit PCM. The frame is always declared as unvoiced when the voicing strength of the frame is less than 0.8, and it is declared as transition when an unvoiced frame is followed by a voiced frame.

The unvoiced segments of the signal is transmitted to the decoder without any modification. As the signal is synthesized pitch-synchronously in the decoder (fixed cycle length is used in unvoiced frames), the signal between the starting location of the first synthesis pitch cycle and the ending location of the last synthesis pitch cycle is transmitted. For transition frames, each pitch-cycle in the modified signal is transmitted to the decoder in the form of constant-length pitch cycles. The inverse CPT is applied to the constant-length pitch cycles and the resulting pitch cycles, with a sampling rate ten times the original sampling rate, are copied to the corresponding segment locations of the excitation signal buffer. The segments in unvoiced frames and the constant-length pitch-cycles in transition frames are both quantized at 16 bits and encoded with PCM. In the voiced frames, the excitation signal is generated as in the 2.4 kb/s I-MELP coder. The rest of the MELP decoder, which includes the adaptive spectral enhancement filter, linear prediction filter and gain scaling, is used to synthesize the speech signal.

In the informal listening test, the quality of the synthesized speech using this method is slightly better than that of the 2.4 kb/s I-MELP coder. The synthesized speech sounds clearer, especially at onsets and transitions. However, to get full advantage of this parametric/hybrid encoding method, the voiced segments must be quantized more accurately, because the overall quality of the speech coder is still determined by the I-MELP coder. Stachurski et al. [2] proved that it is possible to achieve near-toll quality when a variation the MELP model operating at 4 kb/s is used to encode voiced parts of the speech signal. For this reason, the quality of this initial implementation can be improved with improved quantizers operating at higher bit rates.

4. Results of Formal Listening Tests

We evaluated the performance of the modified input speech, the experimental I-MELP/PCM coder and the 2.4 kb/s I-MELP coder using a degradation category rating (DCR) [6]. 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 modified input speech, the experimental I-MELP/PCM coder and the 2.4 kb/s I-MELP coder that uses the source dependent LPC and Fourier series estimation method [4] were evaluated. The reference signal in this test was 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.

Table 1 presents the results of the comparison between the original speech and the speech obtained with the modification algorithm presented in Section 2. As seen in this table, the quality of the modified speech signal is different from the un-

processed speech signal. Although the average DMOS score is around 4.0, which means that the degradation is not annoying, the modified speech signal should be very close to the unprocessed speech signal. However, as the main purpose of the DCR test is to determine the degradation in the processed signal, it is not surprising that even a slight difference between the unprocessed and the modified speech results in lower scores. When listening test results of individual subjects are inspected closely, we observed that there are a couple of sentences that are always scored very low. In these sentences, the pitch-cycle modification algorithm introduces severe audible artifacts because of a pitch-estimation mistake or a segmentation mistake. In addition, as the zero-phase equalization filter is only applied to voiced cycles, the algorithm introduces audible artifacts at onsets when the pitch-cycle segmentation algorithm makes mistakes in identifying the voicing state of a cycle. As these mistakes are significantly less in female speech than in male speech, a score difference of 0.4 DMOS was obtained between the two genders. Furthermore, as there are more pitch-cycles in a frame in female speech at transition regions than in male speech, the initial selection of the segmentation locations are more reliable for female speech than for male speech. Allowing additional delay will likely reduce this type of distortion especially in male speech as it is possible to have more reliable pitch estimates and to perform segmentation more accurately.

Table 1: Comparison of the original speech (reference) and the modified speech (test) in the DCR test.

Gender	X_{ref}	X_{test}	X_{diff}	Result (95% Conf.)
All	4.96	4.00	-0.96	Original speech better
Male	4.96	3.80	-1.16	Original speech better
Female	4.96	4.20	-0.76	Original speech better

The comparison of the modified speech and the I-MELP/PCM coder shown in Table 2 reveals interesting results. As the artifacts resulting from the switching between the excitation signal generated in the I-MELP coder and the transmitted waveform of the modified-residual signal is mostly eliminated, the quality difference between these two signal can only result from the encoding of the voiced part with the I-MELP coder. It was observed that the DMOS score difference between these two test cases is very small and statistically insignificant for the male speech. This result suggests that the I-MELP coder already encodes the voiced frames of the male speech very efficiently even at 2.4 kb/s. However, the large DMOS score difference for the female speech in the order of 0.44 DMOS also suggests that there is still a room for improvement in the speech quality of the I-MELP coder for female speakers. This result also suggests that the performance of the pitch-cycle modification is very crucial especially for male speakers. The quality of male speech will likely improve in I-MELP/PCM coder when the problems in the pitch-cycle modification algorithm is eliminated.

Table 2: Comparison of the modified speech (reference) and the experimental I-MELP/PCM speech coder (test) in the DCR test.

Gender	X_{ref}	X_{test}	X_{diff}	Result (95% Conf.)
All	4.00	3.75	-0.25	Modified speech better
Male	3.80	3.73	-0.07	Equal
Female	4.20	3.76	-0.44	Modified speech better

Finally, Table 3 presents the results of the comparison between the 2.4 kb/s I-MELP coder and the experimental I-MELP/PCM coder. As seen in this table, the quality of the I-MELP/PCM coder is statistically better than the 2.4 kb/s I-MELP coder with an average 0.22 DMOS score difference. The transmission of the modified-residual signal in the transition segments and unvoiced frames reduces the degradation in the synthesized speech and improves the quality. In addition, the degradation resulting from the pitch-estimation mistakes are also reduced in the I-MELP/PCM coder. For this reason, the DMOS score difference is larger in the male speech than the female speech.

Table 3: Comparison of the 2.4 kb/s I-MELP coder (reference) and the I-MELP/PCM speech coder (test) in the DCR test.

Gender	X_{ref}	X_{test}	X_{diff}	Result (95% Conf.)
All	3.53	3.75	0.22	I-MELP/PCM better
Male	3.47	3.73	0.27	I-MELP/PCM better
Female	3.59	3.76	0.18	Equal

These results shows that the quality of the hybrid coder is still largely determined by the parametric speech coder. Better quantization methods or bit-rate increase will most likely improve the quality especially for female speakers. In addition, Table 3 proves that encoding transition segments and short events by waveform encoding methods does improve the quality as well. We also believe that the quality of both purely parametric coders and hybrid coders can be improved by just relaxing the delay constraint. This will result in less pitch estimation errors and better segmentation accuracy at speech onsets, and therefore, eliminate the distortions significantly.

5. Acknowledgements

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

6. References

[1] E. Shlomot, V. Cuperman, and A. Gersho, "Hybrid coding: combined harmonic and waveform coding of speech at 4 kb/s," *IEEE Trans. Speech and Audio Processing*, vol. 9, pp. 632–646, 2001.

[2] J. Stachurski and A. McCree, "A 4 kb/s hybrid MELP/CELP coder with alignment phase encoding and zero-phase equalization," *Proc. IEEE Int. Conf. Acoust. Speech Signal Processing*, pp. 1379–1382, 2000.

[3] N. Katugampala and A. Kondoz, "A hybrid coder based on a new phase model for synchronization between harmonic and waveform coded segments," *Proc. IEEE Int. Conf. Acoust. Speech Signal Processing*, pp. 685–688, 2001.

[4] A.E. Ertan and T.P. Barnwell III, "Improving the quality of military standard MS-MELP speech coder using pitch synchronous analysis and synthesis techniques," *Proc. IEEE Int. Conf. Acoust. Speech Signal Processing*, pp. I761–I764, 2005.

[5] 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.

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