# AN IMPROVED MIXED EXCITATION LINEAR PREDICTION (MELP) CODER

*Takahiro Unno*[†], *Thomas P. Barnwell III*[†], *Kwan Truong*[‡]

[†]Center for Signal and Image Processing, Georgia Institute of Technology, Atlanta, GA

[‡]Atlanta Signal Processors, Inc., Atlanta, GA

[takahiro|tom]@ee.gatech.edu, kwan@aspi.com

## ABSTRACT

This paper presents an improved Mixed Excitation Linear Prediction (MELP) coder. The MELP is the linear-prediction-based speech coder that was recently chosen as the new 2400 bps U.S. Federal Standard. Even though the MELP is quite good, there are still some perceivable distortions, particularly around non-stationary speech segments and for some low-pitch male speakers. The key features of our new coder include a robust pitch detection algorithm, a new plosive analysis/synthesis method, and a post processor for the Fourier magnitude model. Formal quality tests are used to show that the new MELP improves the quality of the U.S. Federal Standard MELP coder while requiring only a small increase in algorithmic delay and while also retaining compatibility with the Federal Standard MELP bit-stream specification.

## 1. INTRODUCTION

The MELP coder is a form of the traditional linear-prediction-based speech coder that was recently chosen as the new 2400 bps U.S. Federal Standard [1, 2, 3]. As part of this competition, the MELP's quality was shown to be better than that of the 4.8 kbps FS1016 CELP standard [2, 3]. The MELP has five features that differentiate it from traditional pitch-excited vocoders. These include mixed excitation, aperiodic pulses, adaptive spectral enhancement, pulse dispersion, and Fourier magnitude modeling. In the MELP, these features combine with efficient parameter quantization algorithms to make it a good quality speech coder at a low bit rate.

Careful listening tests over a variety of speech samples, however, showed that three types of distortions were still observed in the MELP-coded speech. The first is the artificial-sounding artifacts (best described as "roughness") particularly perceivable around non-stationary speech segments. The second is a lack of clarity, a slight "muddiness". The source of this distortion seems to be concentrated mostly in transitions, but it disrupts the clarity of the entire sentence. The third is a slight highpass-filtered quality for some low-pitch male speakers. This causes the coded speech for some male speakers whose fundamental frequencies are less than 100 Hz to sound too synthetic.

To solve these problems, this paper introduces three additional features to the MELP coder: (1) a robust pitch detection algorithm that significantly reduces the artificial noise in non-stationary speech segments, (2) a plosive analysis/synthesis method that enhances the clarity of the coded speech, (3) a post processor for the Fourier magnitude model that improves the overall speech quality for the low-pitch male speakers.

With these features, the new coder provides better quality than that of the Federal Standard MELP coder while maintaining the basic MELP properties such as low bit rate and short algorithmic delay, and compatibility with the Federal Standard MELP coder in terms of the bit-stream specification.

## 2. THE IMPROVED MODEL

### 2.1. Robust Pitch Detection

As noted above, the artificial "roughness" occurs mostly at the beginning of a vowel segment, the end of a vowel segment, or the junction of two vowel segments. Figure 1(b) shows an example of the speech in which the artificial noise is observable. As shown in Figure 1(a), an irregular pitch tract has occurred at the junction between vowel segments in the input speech signal. The resulting low pitch correlation for the frame containing an irregular pitch period forced the MELP algorithm to mark the frame as unvoiced. As a result, noise excitation was applied at the junction of vowel segments, causing the rough sound. Similarly, low pitch correlation may also cause pitch doubling and result in a different type of rough sound. In both of the cases, the irregular pitch affects not only the frame itself but also the following and the preceding frames because the pitch and the voicing decisions are linearly interpolated with those of the adjacent frames.

In the Federal Standard MELP coder, the initial pitch estimate is the pitch lag $T$ between 40 and 160 samples that provides the highest normalized autocorrelation for the lowpass-filtered input speech. The fractional pitch is searched around the initial pitch lag, and the bandpass voicing decision is made using the pitch correlation corresponding to the fractional pitch lag. In our new coder, a sliding pitch analysis window method is used. This method seeks the pitch analysis window position that provides the highest pitch correlation by sliding the window around the original position. This is equivalent to using a nearby, more stationary vowel segment instead of using the signal containing an irregular pitch for pitch analysis. Using this more stationary signal can avoid inappropriate voicing decisions and pitch estimates, and can reduce the artificial noise in non-stationary vowel segments (see Figure 1(c)). The pitch correlation provided by the sliding window method is given

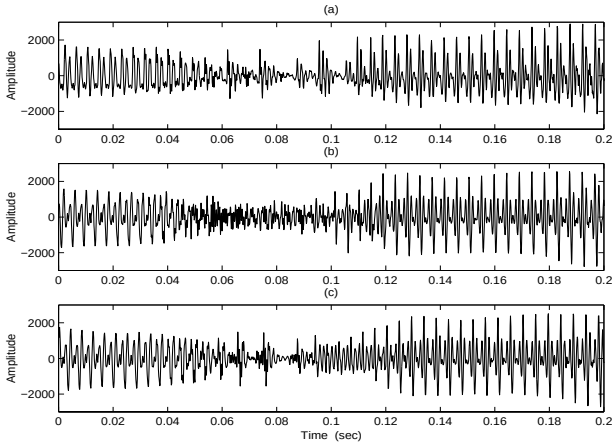


Figure 1: Elimination of the artificial noise at the junction of two vowel segments. (a) Input speech signal. (b) Synthesized speech signal by the Federal Standard MELP. (c) Synthesized speech signal by the improved MELP.

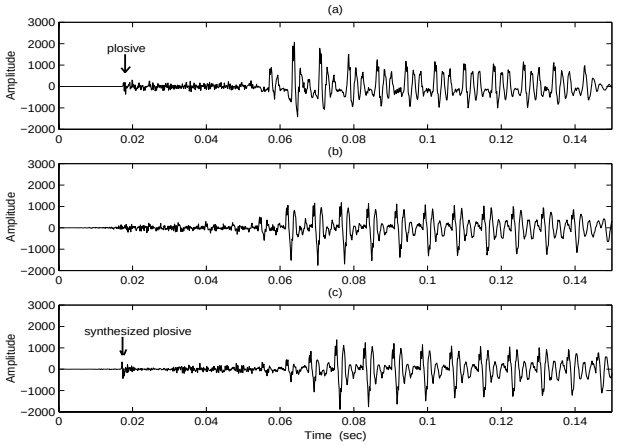


Figure 2: Reproduction of the plosive signal in the synthesized speech. (a) Input speech signal. (b) Synthesized speech signal by the Federal Standard MELP. (c) Synthesized speech signal by the improved MELP.

by:

$$R(T) = \max_{i=-T_s}^{T_s-1} [\max_{T} R_i(T)] \quad (1)$$

$$R_i(T) = \frac{C(i, T+i)}{\sqrt{C(i,i)C(T+i,T+i)}} \quad (2)$$

where $T_s$ is the maximum sliding range, and $C(k,l)$ is the autocorrelation given by:

$$C(k,l) = \sum_{n=0}^{N-1} s(n+k)s(n+l) \quad (3)$$

where $N$ is a frame size and $s(n)$ is the lowpass speech signal. In the new coder, the initial pitch estimate is the pitch lag corresponding to $R(T)$. This initial pitch lag and the signal in the window that provides $R(T)$ are used for the fractional pitch search, bandpass voicing decision, LPC analysis, and the gain estimate. A direct implementation of Equation (2) for all values of $i$'s results in a significant increase in the computational complexity. Our approach to reduce the computational complexity is to utilize the following recurrence equation to compute the autocorrelation:

$$C(k,l) = C(k-1,l-1) + s_{k+(N-1)}s_{l+(N-1)} - s_{k-1}s_{l-1} \quad (4)$$

### 2.2. Plosive Analysis/Synthesis Method

The mixed excitation allows the MELP to have considerable freedom for the voicing decision. However, the mixed noise and pulse excitation is not capable of reproducing isolated pulse-like signals such as those seen in plosive sounds. Figure 2 shows an example of speech containing a plosive sound. The input speech signal (a) contains a /p/ sound preceding the vowel segment. In the synthesized speech signal (b), a noise excitation is applied to the full band of the segment associated with the plosive sound /p/ and degrades the clarity of the speech quality. To provide a clearer speech quality for the sentence containing a plosive sound, we propose a new algorithm for the production of plosives. Our algorithm can be separated into three parts: the detection of the plosive location, the modeling of the plosive signal, and the synthesis of the plosive signal.

#### *2.2.1. Plosive Detection*

To identify the plosive signal, the peakiness value of the LPC residual signal is used. The peakiness value is the ratio of the L2 norm to the L1 norm of the LPC residual signal. Since the peakiness value is sensitive to the phase of the plosive signal, a sliding window is used to find the frame position which maximizes the peakiness value. The peakiness value with the sliding window is given by:

$$P = \max_{i=-T_s}^{Ts-1} P_i \quad (5)$$

$$P_i = \frac{\sqrt{\frac{1}{N}\sum_{n=0}^{N-1} r_{n+i}^2}}{\frac{1}{N}\sum_{n=0}^{N-1} |r_{n+i}|} \quad (6)$$

where $N$ is a frame size, and $T_s$ is the maximum sliding range that is also used in Equation (1). The peakiness value with a sliding window is illustrated in Figure 3 along with that of the fixed position window and the corresponding input speech waveforms. Clearly, the peakiness value with the sliding window can distinguish the frame containing the plosive signal. In addition to the peakiness value, the lowpass energy is computed and used to distinguish the rapid onset of vowel from the plosive signal. The detected plosive sounds include /p/, /k/, /t∫/, /t/, /d/, /b/, /th/, /g/, and /v/. It was found that most of the detected plosive signals were non-aspirated while most of non-detected plosive signals were aspirated. This result is preferable to detecting all plosive signals because only the non-aspirated plosive sounds need the plosive model.

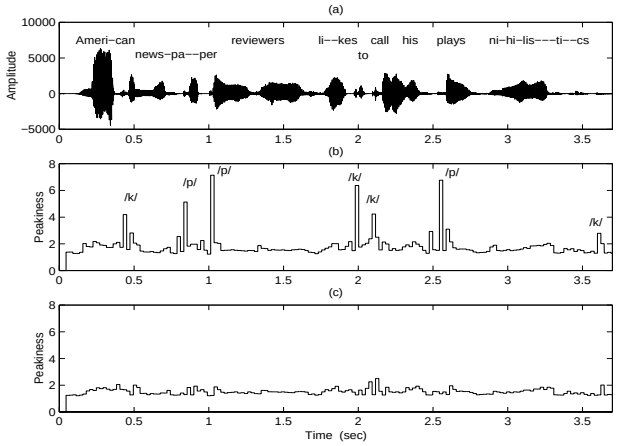


Figure 3: Plosive signal detection. (a) Input speech signal. (b) Peakiness value with sliding window. (c) Peakiness value with fixed position window.

#### 2.2.2. Plosive Modeling

Multipulse excitation has been shown to produce an isolated pulse-like signal while requiring a relatively high bit rate [4]. The proposed alternative model provides perceptually transparent quality for plosive sounds but maintains a low bit rate. To determine the critical characteristics of plosive sounds which need to be modeled, a preliminary experiment was conducted by replacing the plosive signals in a set of sentences for other types of plosive signals(e.g. /p/ for /k/). As a result of this experiment, it was determined that transparent replacement requires only a rough spectral fit between the plosive signal.

In our proposed model, all plosive signals $p(n)$ are produced by scaling and LPC-synthesis-filtering a single prestored template LPC residual signal $v(n)$:

$$p(n) = g_p v(n) + \sum_{i=1}^{p} a_i p(n-1) \tag{7}$$

where $g_p$ is the scaling factor based on the energy of the input plosive signal, and $a_1, \cdots, a_p$ are the LPC coefficients computed from the input plosive signal. The template plosive signal is chosen arbitrarily and filtered with a 14th order inverse LPC filter to produce the template LPC residual signal $v(n)$. Since the preliminary experiment showed that only a rough spectral fit between the input and the synthesized plosive signals is required to reproduce transparent speech quality, an accurate LPC analysis is not needed for the input plosive signal. In fact, listening to synthesized plosive sounds using different LPC orders, the 6th order analysis was found to be sufficient. In our implementation, this additional 6th-order LPC analysis for the plosive signal is not even used in the improved MELP coder. Instead, the 10th-order LPC parameters corresponding to the frame containing the plosive signal are used for the reproduction of the plosive signal in order to minimize the additional bits for the plosive model.

#### 2.2.3. Plosive Synthesis

In our final implementation, the plosive signal is reproduced independently and then added back to the synthesized speech signal in the MELP decoder. The length of the synthesized plosive signal is fixed to half of the frame length of 90 samples (11.25 ms). The position of the plosive signal is identified by seeking the maximum amplitude position in the detection window and quantized to one bit, i.e. either the first half or the second half of the current frame. Hence, the final synthesized signal, $s'(n)(0 \le n \le 179)$, for the frame containing the plosive signal is given by:

$$s'(n) = \begin{cases} s(n) + p(n - N_p) & \text{if } N_p \le n \le N_p + 89 \\ s(n) & \text{otherwise} \end{cases} \tag{8}$$

where $p(n)$ and $s(n)$ are the plosive signal and the synthesized speech signal respectively, and $N_p = 0$ if the maximum amplitude of the plosive signal is detected in the first half of the frame, and $N_p = 90$ if detected in the second half. In the frame containing the plosive signal, the gain parameter is modified to suppress the energy just before the surge of energy:

$$\begin{array}{rcll} g_i(0) & = & g_{i-1}(1) & \text{if } N_p = 0 \\ g_i(1) & = & g_i(0) & \text{if } N_p = 90 \end{array} \tag{9}$$

where $g_i(j)$ is the $j$th gain parameter $(j = 0, 1)$ in the $i$th frame. Figure 2(c) shows the synthesized speech including the plosive synthesis.

### 2.3. Post Processor for the Fourier Magnitude Parameters

The Fourier magnitude model allows the MELP to reconstruct the lower frequency spectrum more accurately and improve the speech quality, particularly for male speakers. However, coded speech for some low-pitch male speakers still has highpass-filtered quality. Figure 4 shows the Fourier magnitudes of original speech and coded speech in low frequencies for a low-pitch male speaker. In Figure 4, it is observed that the first, second, and third harmonic magnitudes of the coded speech are smaller than those of the original speech. It is caused by two different types of filtering: one is an adaptive spectral enhancement filter (ASEF); the other is a preprocessing highpass filter with a cutoff frequency of 60Hz. The ASEF is designed to emphasize the formant peaks and also to suppress magnitudes of spectral valleys. The preprocessing highpass filter is applied to the input speech signal to remove very low frequency noise. Since the first three harmonics in Figure 4 reside in the spectral valley, these magnitudes are suppressed by the ASEF. In addition, the first harmonic which resides around 90Hz is affected by the highpass filter.

To improve the coded speech quality for the low-pitch male speakers, the harmonic magnitudes in the low frequencies are adaptively modified by removing the effect of the two filters including the ASEF and the highpass filter. The modified harmonic magnitude $|\tilde{S}(e^{j\omega_i})|$ is given by:

$$|\tilde{S}(e^{j\omega_i})| = |S(e^{j\omega_i})| \frac{\sqrt{G}}{H(e^{j\omega_i})} \tag{10}$$

where $\omega_i$ is the $i$th harmonic frequency, $G$ is the average Fourier spectrum energy, and $S(e^{j\omega_i})$ is the non-modified

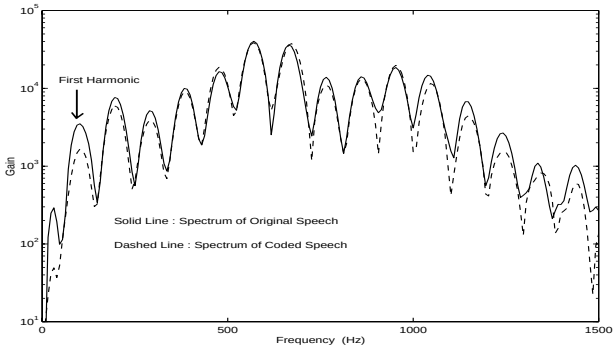


Figure 4: Fourier magnitudes of speech for a low-pitch male speaker. A solid Line and a dashed line are the spectrum of the original and the coded speech respectively.

Fourier magnitude of the $i$th harmonic. The new coder uses the MELP Fourier Magnitude parameter, which is the Fourier magnitude of the LPC residual signal, for the harmonic magnitude modification rather than using the harmonic magnitude of the coded speech $S(e^{j\omega_i})$. The magnitude response of the filter $|H(e^{j\omega})|$ is given by:

$$|H(e^{j\omega})| = |H_1(e^{j\omega})||H_2(e^{j\omega})| \tag{11}$$

where $|H_1(e^{j\omega})|$ and $|H_2(e^{j\omega})|$ are the magnitude responses of the ASEF and the preprocessing highpass filter respectively. To avoid losing the advantage of the ASEF, Equation (10) is applied to only the harmonics that are 200Hz less than the first formant frequency. The first formant frequency $F_1$ is roughly estimated using quantized Line Spectrum Frequencies(LSF's) as follows:

$$F_1 = \begin{cases} \frac{\hat{f}_1+\hat{f}_2}{2} & \text{if } \hat{f}_2 - \hat{f}_1 < \hat{f}_3 - \hat{f}_2 \\ \frac{\hat{f}_2+\hat{f}_3}{2} & \text{otherwise} \end{cases} \tag{12}$$

where $\hat{f}_i$ is the $i$th quantized LSF.

## 3. RESULTS

### 3.1. Subjective Test

To evaluate the quality of the improved MELP coder, we conducted an A/B comparison test with 32 sentence pairs uttered by 32 different speakers. The Federal Standard MELP coder was used as a reference. The test material included only clean speech and was presented to 14 listeners. An overall result shows that the improved MELP coder was preferred over the Federal Standard MELP coder by 65 % to 35 %. To analyze the performance of the new coder statistically, the paired-sample sign test was applied to the score of the A/B comparison test [5]. Since the listeners were forced to choose one of the coders as better for each sentence in the test, Z-factor can be based on the binomial distribution as follows:

$$Z = \frac{N_a - (N \cdot p_0)}{\sqrt{N \cdot p_0 \cdot (1 - p_0)}}, \; p_0 = 0.5 \tag{13}$$

where $N$ is the total number of samples, and $N_a$ is the number of the samples in which the new coder is preferred. The overall result shows $Z = 6.200$, and it indicates that the new coder provides better quality than that of the Federal Standard MELP coder with significance level of 1 %.

### 3.2. Bit Rate and Delay

The bit rate of the improved MELP coder is the same as that of the Federal Standard MELP coder. The frames containing plosive information require an additional four bits including one bit for the flag, one bit for the position, and two bits for the gain. However, these bits can be packed into the bit stream of the Federal Standard MELP coder without increasing the bit rate by forcing the plosive frames to be voiced and using the Fourier-magnitude bits for the plosive synthesis. The new pitch detection algorithm and the post processor for the Fourier magnitude parameter do not affect the bit allocation. The new pitch detection algorithm and the plosive detection algorithm do require an additional algorithmic delay of 15 ms.

## 4. CONCLUSION

We have added three new features to the Federal Standard MELP coder. The robust pitch detection algorithm reduces or removes the artificial noise in transitions. The plosive analysis/synthesis method reproduces plosive sounds and provides better clarity in the coded speech. The post processor for the Fourier magnitude model improves the speech quality for low-pitch male speakers. The improved MELP coder was shown to provide better quality than that of the Federal Standard MELP coder in subjective tests while requiring only a short additional algorithmic delay and while maintaining the bit-stream specification of the Federal Standard.

## 5. REFERENCES

[1] A. V. Macree and T. P. Barnwell III, "A Mixed Excitation LPC Vocoder Model for Low Bit Rate Speech Coding," *IEEE Trans. Speech and Audio Processing*, vol.3, pp. 242-250, July 1995.

[2] A. V. Macree, K. Truong, E. B. George, T. P. Barnwell III, and V. Viswanathan, "A 2.4 kbits/s MELP Coder Candidate for the New U.S. Federal Standard," in *Proc. ICASSP*, pp. 200-203, vol. 1, May 1996.

[3] L. M. Supplee, R. P. Cohn, J. S. Collura, and A. V. Macree, "MELP: The New Federal Standard at 2400 pbs," in *Proc. ICASSP*, pp. 1591-1594, vol. 2, April 1997.

[4] S. Singhal and B. S. Atal, "Amplitude Optimization and Pitch Prediction in Multipulse Coders," *IEEE Trans. on ASSP*, vol. 37, no. 3, pp. 317-327, March 1989.

[5] F. Chiou, "User-Interactive Speech Enhancement Using Fuzzy Logic," *PhD thesis*, Georgia Institute of Technology, 1998.