An Improved MELP Algorithm for Transition Frames

Qiu Tian
Communication University of China
Digital Engineering Center
Beijing, China，18611183394
qiutian06@cuc.edu.cn

Wang Feifei
Communication University of China
Digital Engineering Center
Beijing, China

Du Weitao
Communication University of China
Digital Engineering Center
Beijing, China

***Abstract*—The MELP vocoder has a good performance at 2.4 kb/s and is selected as the Federal Standard Vocoder of the United States. However, due to the simple U/V decision of MELP, desired synthesis performance is not achieved in transition frames. This paper presents an improved MELP algorithm in which transition frame is detected and processed on the basis of peakiness detection. Comparison results are given both in time domain waveform and PESQ tests. Better synthesized speech quality is provided by using the improved MELP algorithm.**

Keywords-Vocoder; MELP; U/V decision; transition frame;

I. Introduction

MELP (Mixed Excitation Linear Prediction) is based on the traditional LPC (Linear Prediction Coder) model with additional features including mixed excitation, aperiodic pulses, adaptive spectral enhancement, pulse dispersion, and Fourier magnitude modelling. It produces higher synthesized speech quality at a low bit rate and has replaced LPC-10 as the 2.4 kb/s Federal Standard Vocoder of the United States. However, due to the simple U/V decision in MELP, unnaturalness still exists in synthesized speech, especially in transition segments. This paper puts forward a new method of transition frame detection which makes frame type decisions more precisely. A new method to generate mixed excitation by dividing the transition frame into two sub-frames is also introduced in this paper, in order to improve the quality of synthesized speech.

II. The MELP Vocoder

A. Encoder

The MELP parameters which are quantized and transmitted are the pitch, the bandpass voicing strengths (Vbp_i, $i=1,2,\ldots,5$), the two gain values ($G1$ and $G2$), the linear prediction coefficients (a_i, $i=1,2,\ldots,10$), the Fourier magnitudes, and the aperiodic flag. The encoding process is shown in Fig.1.

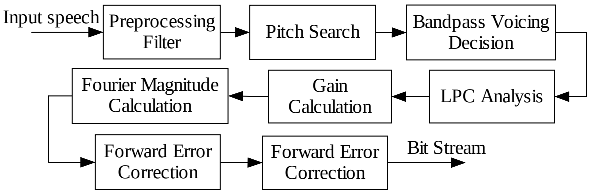


Figure 1. Block diagram of MELP encoder.

The first step in encoding process is to remove DC interference through a highpass filter with a cutoff frequency of 60Hz. Then use the normalized autocorrelation method to calculate the integer pitch value. The bandpass voicing analysis determines the five bandpass voicing strengths; meanwhile, the aperiodic flag is decided by voicing strength of the lowest band.

After bandpass analysis, a 10th order linear prediction analysis is performed on the input signal, and the autocorrelation analysis is implemented by the Levinson-Durbin recursion. Besides, the linear prediction residual signal should be calculated, whose peakiness contributes to the decision of the lowest band voicing strength.

Gains and Fourier magnitudes are calculated after getting the final pitch. The gain is the RMS value and is measured twice per frame using a pitch-adaptive window length. To measure the Fourier magnitudes of the first 10 pitch harmonics of the prediction residual, a 512-point FFT is performed in MELP algorithm.

B. Decoder

The structure of MELP decoder is shown in Fig.2.

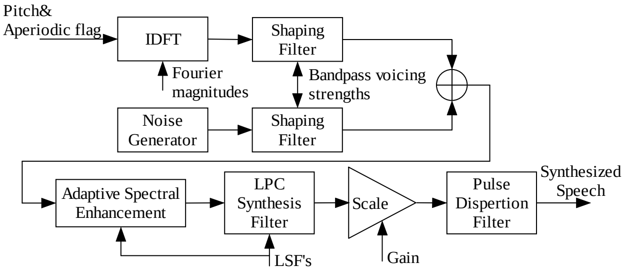


Figure 2. Block diagram of MELP decoder.

Parameter decoding is different for voiced and unvoiced modes. So the pitch is decoded first to get the mode information and to decide which decoding mode to choose. For quiet input signals, a small amount of gain attenuation is applied to both decoded gain parameters using a power subtraction rule. As speech is synthesized within a pitch period, all MELP parameters are interpolated pitch-synchronously for each synthesized pitch period, mostly are linear interpolations.

Generation of the mixed excitation is one of the most important parts in speech synthesis. The mixed excitation is the sum of the filtered pulse and noise excitations. Then an adaptive spectral enhancement filter is applied to the mixed excitation to solve the problem of formant matching.

The synthesis is to filter the mixed excitation using the LPC filter directly. After gain adjustment and pulse dispersion, a pitch period of speech is successfully synthesized.

III. Transition frame

A. Detection method

A transition frame contains both voiced and unvoiced segments. In this paper, if unvoiced segment occupies 1/4 to 3/4 of the entire frame and the rest of the frame is voiced segment, we judge this frame as a transition frame. According to MELP algorithm, transition frame will be encoded as unvoiced or voiced even if it's not. Thus the decoder will generate excitation in unvoiced or voiced mode, which in turn has negative effects on the quality of synthesized speech.

Based on original MELP algorithm, calculation of the energy difference between the two halves of a frame is added to the modified MELP method, which makes the detection more accurate. The specific detecting method is as follows:

- Peakiness detection. In MELP method, peakiness of LP residual signals *r(n)* is calculated over a 160 sample window centered on the last sample in the current frame. The peakiness value is:

$$\text{peakiness} = \frac{\sqrt{\frac{1}{160}\sum_{n=0}^{159} r^2(n)}}{\frac{1}{160}\sum_{n=0}^{159}\left|r(n)\right|} \tag{1}$$

  If the peakiness exceeds 1.34, the lowest band voicing strength is forced to 1. If it exceeds 1.6, then the lowest three band voicing strengths are all forced to 1.

  To detect the transition frame, another peakiness threshold is set to be 1.48 based on a large amount of test data. If the peakiness of residual exceeds 1.48, typical transition frames can be detected. However, the peakiness measure is inadequate to deal with all the cases, especially the case that the energy of beginning or end of sub-frame is quite low. Therefore, further steps should be made to ensure the detection results.

- Energy comparison between two sub-frames. There are two types of transition, unvoiced to voiced transition and voiced to unvoiced transition, represented by U-V and V-U in this paper for convenience.

  Equally divide the candidate transition frame from peakiness detection into two sub-frames, s_1 and s_2. Calculate the energy value M_1 and M_2 of sub-frames and their energy difference value △*M*.

$$M_{\mathrm{i}} = 10\log\left(\sum_{n=1}^{90} s_i(n)^2\right), \quad i=1, 2 \tag{2}$$

$$\Delta M = M_1 - M_2 \tag{3}$$

  Energy level of voiced speech is usually higher than unvoiced speech. When △M > 12dB, the encoder judges it as a V-U frame, and when △M < -12dB, it is a U-V frame.

  As a whole, one frame is set to be a transition frame if it meets the following conditions:

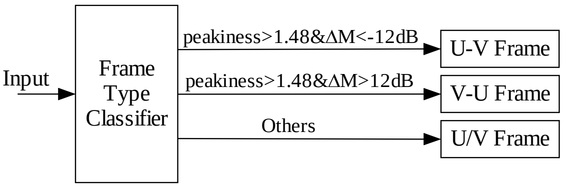


Figure 3. Transition frame detection.

By testing the detection method using a 172 frame length speech, 10 transition frames are detected with 1 error detection. Transition frames account for approximately 6% of the entire speech, proving that separately processing of transition frames has its practical significance.

B. Bit allocation

In MELP, parameters are quantized in 54 bits and transmitted at 2.4kb/s. After adding the transition frame type, the transition detection results need to be encoded, too. The transition flag uses 2 bits to indicate transition frames, 01 refers to U-V frame, 10 refers to V-U frame and 00 means current frame is not a transition frame. Flag 11 is used as error protection which appears when error detection happens.

C. Modification of the decoder

In MELP, the decoder decodes the mode information and synthesizes in unvoiced or voiced mode with relevant parameters. After modification, the decoder decodes the transition flag first. For voiced and unvoiced frames, the decoding process is just the same as MELP method. For transition frames, the main modification is in the generation of mixed excitation. If current frame is a U-V frame, the first sub-frame will be synthesized in unvoiced mode and the second sub-frame in voiced mode, synthesis of V-U frame shares the same method.

In unvoiced mode, default parameter values are used for the pitch, jitter, bandpass voicing, and Fourier magnitudes. The pitch is set to 50 samples, the jitter is set to 25%, all of the bandpass voicing strengths are set to 0, and the Fourier magnitudes are set to 1.

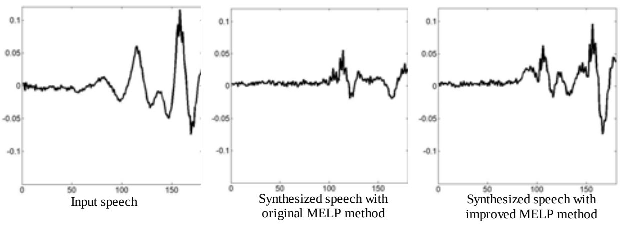
In voiced mode, when the lowest band voicing strength is 1, jitter is set to 25% if the aperiodic flag is 1; otherwise jitter is set to 0%. Other parameters are decoded in turn.

IV. Test results and evaluation

A. Test results

The improved MELP algorithm is simulated in MATLAB. A 172 frame length Chinese female speech is used as voice sample. The speech sample, the synthesized speech using original MELP method and the synthesized speech using improved MELP method are compared and analyzed. In the synthesized speech with improved method, unnaturalness caused by mutation between syllables is reduced, and subjective hearing experience is improved.

A U-V transition frame and a V-U transition frame are chosen from the speech sample, whose waveform comparisons are shown in Fig.4 and Fig.5.

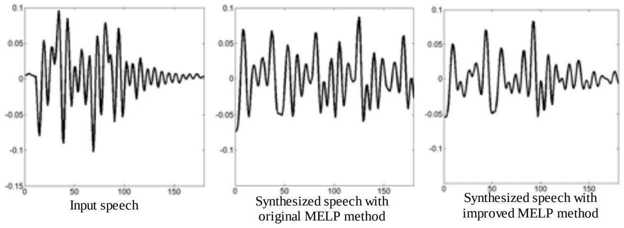


Input speech

Synthesized speech with original MELP method

Synthesized speech with improved MELP method

Figure 4. Simulation results of a U-V transition frame.



Input speech

Synthesized speech with original MELP method

Synthesized speech with improved MELP method

Figure 5. Simulation results of a V-U transition frame.

Viewing from the time domain waveforms, transition frames are synthesized as unvoiced or voiced frames in original MELP algorithm, leading to obvious waveform distortions. The improved synthesized waveform matches better with the speech sample, so the quality of synthesized speech gets better, too.

The test result shows that, the improved algorithm performs better in transition frames. Besides, with the addition of new algorithm, only 180 times of multiplications and additions are added to the computation of every frame. The processing time is almost the same.

B. PESQ test

PESQ (perceptual evaluation of speech quality) is an objective measurement for estimating subjective quality obtained in listening-only tests. We evaluate the performance of the improved algorithm through PESQ tests, using various voice samples including Chinese male, Chinese female, English male, and English female. The PESQ-MOS scores are shown in Table 1.

TABLE I. PESQ TEST RESULTS

Speech samples	*Original MELP*	*Improved MELP*
Male, Chinese	2.53	2.64
Female, Chinese	2.47	2.61
Male, English	2.93	2.99
Female, English	2.88	2.95

The test results show an increase in PESQ-MOS scores, indicating that the speech quality of the improved algorithm is better than the original one.

V. CONCLUSIONS

This paper proposes an improved frame type decision algorithm in MELP which classifies speech frames into three types: the unvoiced, the voiced, and the transition frame. Based on the original MELP algorithm, the improved method determines transition frame and its transition type by peakiness detection of LP residual and energy calculation of sub-frames. Test results show that the improved algorithm reduces distortions caused by simple U/V decision and inaccurate parameters of transition frames, and improves the quality of synthesized speech with a small additional of computation. It has a certain practicality in researches and engineering applications.

REFERENCES

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

[2] McCree A, Kwan T, George E B, and Viswanathan V, "A 2.4 kbit/s MELP coder candidate for the new U.S. Federal Standard," Acoustics, Speech, and Signal Processing, IEEE International Conference, vol.1, pp.200-203, 1996.

[3] Supplee L M, Cohn R P, Collura J S and McCree A, "MELP: the new Federal Standard at 2400 bps," Acoustics, Speech, and Signal Processing, IEEE International Conference, vol.2, pp.1591-1594, 1997.

[4] McCree A, Kwan T, George E B, Barnwell T P and Viswanathan V, "An Enhanced 2.4 Kbit/s Melp Coder", Speech Coding for Telecommunications, IEEE Workshop, pp.101-102, 1995.

[5] McCree A V and Barnwell T P, "A mixed excitation LPC vocoder model for low bit rate speech coding", Speech and Audio Processing, IEEE Transactions, vol.3, pp. 242-250, 1995.