Optimization of Voiced Excitation Model by MVF Algorithm

Bing Xue and Xiaoqun Zhao(✉)

College of Electronic and Information Engineering,
Tongji University, Shanghai, China
{xuebing8529, zhao_xiaoqun}@tongji.edu.cn

Abstract. In mixed excitation linear prediction (MELP) vocoder incentive model, the decision error of traditional maximum voicing frequency (MVF) algorithm is relatively large. In this paper, it proposes a method to optimize MVF algorithm. Firstly, in many MVF algorithms, considering about algorithm accuracy, applicability and real-time, It selects the most appropriate algorithm (cumulative harmonic scoring algorithm) for very low rate speech coder. And then, the method optimize the definition and noise immunity of the algorithm, which use adaptive MVF value to divided spectrum into two sub-bands. Finally, using pitch as performance parameters of incentive model, it simulates voiced incentive model, and compares the pitch error rate of two band excitation (TBE) model and MELP model. A conclusion is generated that TBE model is closer to the original speech features, and its performance is better than that of the original MELP incentive model.

Keywords: MVF · Residual signal · Pitch · Voicing excitation model · TBE

1 Introduction

In the (very) low rate speech coding, as an important part of speech in the model, incentive model is one of the important breakthrough. The speech model of audio segment consists of periodic and aperiodic component. The spectrum ratio of those two parts has two different calculation ways. One is multiband; the other is dual-band. In the dual-band calculation, low frequency band is regard as periodic components, with the method of impulse excitation; high frequency band is regard as aperiodic components, with the method of noise excitation. The demarcation point of the dual-band is called maximum voicing frequency (MVF). This way of division is closer to actual phonetic pronunciation characteristics. Two band excitation (TBE) model can effectively improve the quality of synthesized speech [1], at the same time, its performance depends on MVF extraction accuracy.

MVF often used in harmonic and noise model (HNM) [2, 3], deterministic and stochastic model (DSM) [4], excitation signal model and so on. In recent years, it has sprung up a large number of MVF calculation methods, including peak valley value (P2V) [5], the amplitude spectrum-phase spectrum [2], multi parameter correlation method, iterative judgment method, cumulative harmonic scoring (CHS) [6, 7] and so on. Considering the accuracy, applicability and real time performance of MVF, it is the

X.-L. Huang (Ed.): MLICOM 2016, LNICST 183, pp. 53–61, 2017.
DOI: 10.1007/978-3-319-52730-7_6

most appropriate algorithm that CHS is applied to mixed excitation linear prediction (MELP) vocoder incentive model.

Based on the peak of the energy spectrum, CHS method respectively calculate the scoring of periodic part and aperiodic part and the weighted value from each harmonic to this two parts. The final total harmonic score is used to filter the MVF values.

2 The Optimization of MVF Algorithm

2.1 Optimize the Definition of MVF Algorithm

Amplitude spectrum-phase spectrum method use hanning window, but CHS method uses rectangular window. Although the rectangular window may bring side lobe leakage, the algorithm that we use mainly focused on characteristics of main lobe. At the same time, compared to hanning window, rectangular window's main lobe energy is more concentrated, so it uses rectangular window in optimized MVF algorithm.

For silent voice segment, MVF is set to 0 Hz. For the speech segment, the window length is twice the pitch period length. Assuming each frame of the pitch period is f_0, the estimate of each frame' MVF is mf_0 (m is the largest number of harmonic). The original speech signal is $s(k), k = 1, 2, \ldots 2N + 10$ (N is pitch period length).

Ten samples before the current sample is used to predict the current sample, which is express as below:

$$s_{pre}(k) = -\sum_{i=1}^{10} a_i s(k-i),\ k = 1,\ 2, \ldots,\ 2N \tag{1}$$

In order to facilitate representation, the expression of predictive value selects negative sign. After linear prediction (LPC) treatment, residual signal ($\mathrm{res}(k)$, $k = 1$, $2, \ldots 2N$) is defined as below:

$$\mathrm{res}(k) = \mathrm{s}(k) - \mathrm{s}_{pre}(k) = \sum_{i=0}^{10} a_i \mathrm{s}(k-i) \tag{2}$$

As for periodic signal, its spectrum is discrete, therefore, after the discrete Fourier transform processing, the direct component of residual signal ($\mathrm{res}(i)$) is removed.

$$R(k) = \sum_{i=1}^{2N} res(i) e^{-\frac{\pi j(k-1)(i-1)}{N}},\ k = 1,\ 2, \ldots,\ 2N \tag{3}$$

Energy spectrum density is defined as: $P(k) = |\mathrm{R}(k)|^2, k = 1, 2 \ldots 2N$. It increases the range of spectrum amplitude and makes peak features more dramatically. It is assumed that A is an even number, then, $P(k), k = 1, 3, \ldots 2N - 1$ is the pitch harmonic power components of current frame residual signal ($\mathrm{res}(i)$); $P(k), k = 2$, $4 \ldots, 2N$ is noise power components. Because $P(k)$ is a point-symmetrical with respect to N, it only need to deal with $P(k), k = 1, 2 \ldots, N$, at the same time, the largest number

of harmonic m should be determined in the harmonic power components in $k = 1, 3 \ldots, N-1$. The normalized power function is defined as follow:

$$P_n(k) = \begin{cases} 1, k = 1 \\ \dfrac{[P(k) + P(k+2)]/2}{[P(k) + P(k+2)]/2 + P(k+1)}, k = 2, 4 \ldots, N \\ \dfrac{P(k)}{[P(k-1) + P(k+1)]/2 + P(k)}, k = 3, 5 \ldots, N-1 \end{cases} \tag{4}$$

The comprehensive accumulated energy E consists of the accumulation of harmonic energy E_h and noise energy E_a.

$$E_h(x) = \sum_{k=3,5 \ldots 2x+1} \max[0, P(k)(P_n(k) - P_n(k-1))], x = 1, 2 \ldots N/2 \tag{5}$$

$$E_a(x) = 2 \sum_{k=2x, 2x+2 \ldots, N} P(k) P_n(k), x = 1, 2 \ldots, N/2 \tag{6}$$

$$E(x) = E_h(x) + bE_a(x), x = 1, 2 \ldots N/2 \tag{7}$$

Among them, A is adjustment parameter. The scope of it generally takes to: $0.2 \sim 0.6$ [8]. MVF is the xth harmonic frequencies when accumulated comprehensive energy E obtain the maximum value. $m = x$.

2.2 Optimize the Noise Immunity of MVF Algorithm

Original speech, $s(k), k = 1, 2, \ldots N$, in which N is the length of pitch period, consists of effective voice information $\mathrm{u}(k)$ and noise information $\mathrm{n}(k)$. In the short-time signal processing, $\mathrm{u}(k)$ is periodic signal, $\mathrm{n}(k)$ is aperiodic signal, therefore, $\mathrm{s}(i)$ is processed with DFT, as shown below.

$$\mathrm{S}(k) = \sum_{i=1}^{N} \mathrm{u}(i) e^{-j\frac{2\pi ki}{N}} + \sum_{i=1}^{N} \mathrm{n}(i) e^{-j\frac{2\pi ki}{N}}, k = 1, 2, \ldots N \tag{8}$$

Harmonic components are mutually orthogonal, so the energy spectrum is as follow:

$$\mathrm{P}(k) = |\mathrm{S}(k)|^2 = \sum_{i_1=N_x}^{N'} \mathrm{u}(i_1)^2 + \sum_{i_2=N_y}^{N''} \mathrm{n}(i_2)^2 + \sum_{i_3=N_z}^{N'''} [\mathrm{u}(i_3) + \mathrm{n}(i_3)]^2 \tag{9}$$

As is shown in the above type, N times harmonic is divided into three types of interval: i_1, i_2 and i_3. i_1 represents a harmonic interval which only distributes effective information. i_2 represents a harmonic interval which only distributes noise information. i_3 represents an interval which distributes both effective and noise information. In the

harmonic interval of i_2 and i_3, the amplitude of $S(k)$ can accumulate noise information, and the amplitude of energy spectrum will change. It will also affect the outcome when the noises accumulate to a certain extent. As shown in Table 1, we assume that it happens in the noise environment. N_0 is the harmonic which correspond to the standard value of MVF. E_0 is comprehensive accumulated energy. In interval i_1, N_1 is some harmonic and E_1 is comprehensive accumulated energy. Similarly, N_2 and E_2 is in i_2. N_3 and E_3 is in i_3.

Table 1. Compare the differences of adding noise verdict

	Comprehensive accumulated energy	The largest harmonic number	Result of judgment	Comment
Noise environment 1	$E_1 = E_0$	N_1	Accuracy	The signal is not affected by noise
Noise environment 2	$E_2 > E_0$	N_2	Serious miscalculation	Adding some other harmonic components, noise can impact excitation signal.
Noise environment 3	$E_3 > E_0$	$N_3 > N_0$ $(N_3 < N_0)$	High (Low)	

The original speech signal is processed with LPC (linear predictive coding) and inverse filtering, and then, predicting residual signal can be got. In the processing of add noise speech, LPC can analyze and extract format, sound loudness, pitch period and some other key messages. Inverse filter can filter high frequency noise; therefore, residual signal is used for frequency domain processing instead of the original speech signal. Residual signal effectively avoid some of the noise. As is shown in Fig. 1, adding automobile noise, SNR = −4 dB, residual signal fully retain the periodicity and peak character of the original speech without noise. However, with the loss of SNR, time-domain signals of the original speech with noise will be seriously deformative and noise interference will also be serious.

The anti-noise performance of modified MVF algorithm is superior to primitive CHS algorithm. Compare the MVF values of voiced frame before and after adding noise. For example, in automobile noise, as shown in Fig. 2, it is spectrum difference before and after adding noise. Even under low SNR circumstance, peak characteristics of harmonic component are still obvious. MVF of the original speech frame is 1440 Hz. MVF of the adding noise speech frame which is calculated by original CHS algorithm is 240 Hz. However, it is 1440 Hz which is calculated by modified MVF algorithm. It can be concluded that using residual signal, modified MVF upgrade the anti-noise performance of the algorithm.

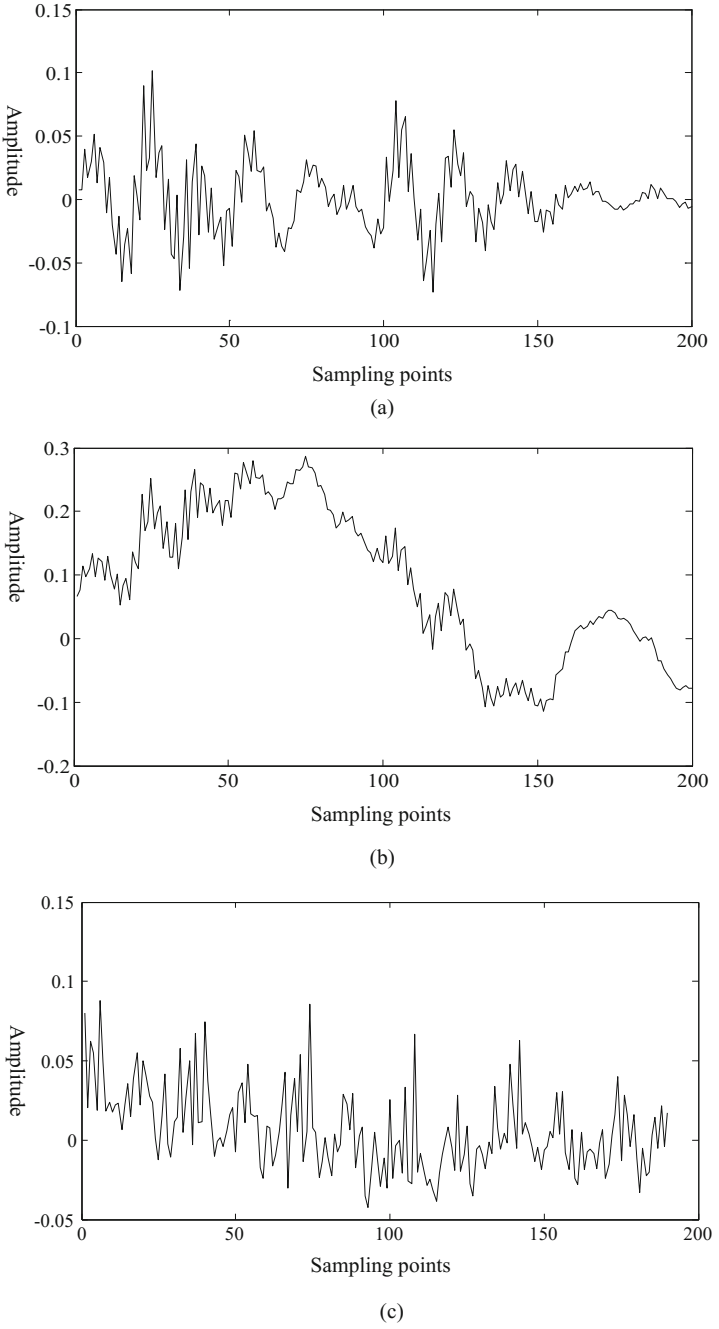

Fig. 1. Time domain waveform of voiced frame ((a) original speech signal with no noise (b) original speech signal with noise (c) residual signal with no noise)

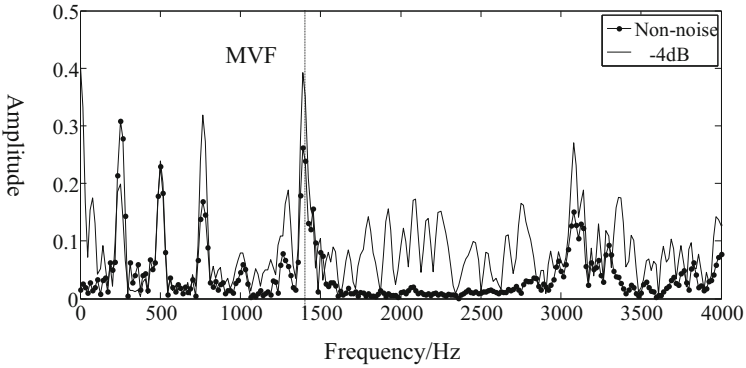

Fig. 2. Single frame spectrum with/without noise

To improve the anti-noise performance of modified MVF algorithm, the frequency spectrum distribution of noise should be considered. In principle, different frequency spectrum distribution of noise has some differences between low frequency and high frequency. In the processing of frequency domain, when different kinds of noise are superposed in target voice, it can be achieved by the improved MVF algorithm to suppress noise.

The improved MVF algorithm has good noise resistance to the general background noise, but human background noise has potential interference. When human background noise exists, UVS Judgment can be used to filter out some interference.

3 The Performance Simulation of Residual Excitation Model

The effect of motivation model, produced in vocoder, can provide the original voice frequency and sound loudness features. It can also be termed pitch. From the point of synthetic speech's subjective judgment, pitch is also one of the intuitive factors to decide the voice quality. It can directly affect the hearing feeling. For example, bass and soprano have obvious difference on the pitch. Except for accurate pitch period, it can effectively improve the naturalness of synthesized speech that coding pitch information of the closer original voice and compound incentive model in the decoder. In addition, the change of pitch directly affects the meaning expression, which is Chinese tone. Different tones correspond to different pitch changes. In the incentive model, the accuracy of extracting pitch corresponds to the naturalness and intelligibility of the synthetic speech, so pitch error rate can be used to evaluate the performance of incentive model. The accuracy of pitch is mainly manifested on the tone and volume. Figures 3 and 4 respectively are spectrograms which are different in tone and volume.

Exciting signal of TBE and MELP can be respectively calculated from original speech signal. The pitch of exciting signal and original speech signal can be extracted and compared in spectrograms. As is shown in Fig. 5, TBE exciting signal and MELP

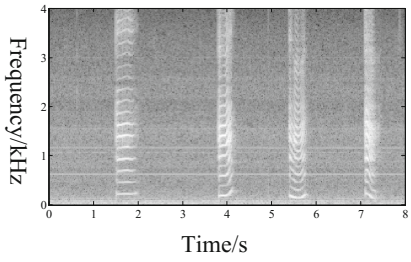

Fig. 3. Spectrogram of different tones

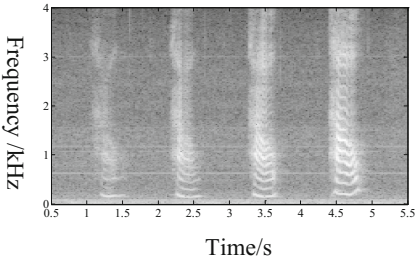

Fig. 4. Spectrogram of different volume

exciting signal both retain the information of pitch change and four tones do not have some confusing phenomena. However, for the voiced frame of speech segment, MELP exciting signal just retains the periodic information under 2 kHz. The exciting signals above 2 kHz are replaced by random signals. As is shown in the spectrogram, there are no clear white horizontal stripes. In the spectrogram of original speech signal, speech segments in 0.7 $\sim$ 1 s, 3.2 $\sim$ 3.5 s and 2.2 $\sim$ 2.6 s both have periodic information in 0 $\sim$ 4 kHz. In other words, the pitches of those speech segments are at 4 kHz. TBE model can not only accurately extract the pitch of voiced frame, but also the pitch of devoiced frame. As an example, the devoiced frame in the vicinity of 0.5 s should contain the third-harmonic component. TEB spectrogram has three clear horizontal white stripes, but in the MELP spectrogram, the harmonic in sub band of 500 Hz is intercepted and stimulated by pulse signal. It only has two horizontal white stripes, including fundamental frequency and second harmonic. Therefore, there are some errors in MELP exciting signals.

Comparing with the spectrograms of some different speech segments and the frequency distribution of their exciting signals, the pitch error rate of TBE exciting signal is far less than that of the MELP exciting signal in both of devoiced and voiced frames. The fixed subband division method cannot accurately extract the pitch information of speech signal, but the method of adaptive MVF can overcome this problem. TBE incentive model is better than original MELP incentive model.

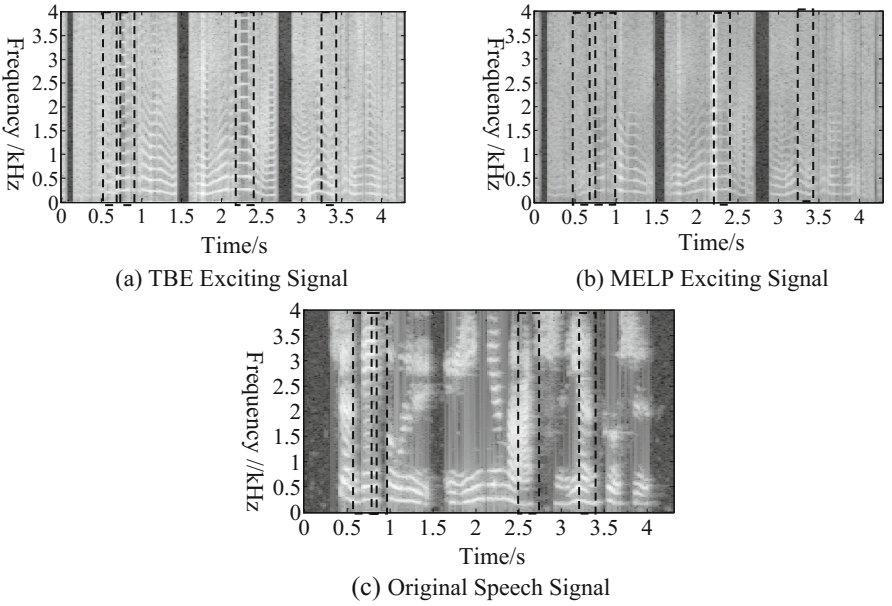

Fig. 5. Spectrogram of exciting signal and original speech signal

4 Conclusion

Contraposing the shortcomings of MELP vocoder incentive model, in this paper, it adopt the way of dynamic segmentation double band and implement TBE incentive model. Firstly, CHS is chosen in many MVF algorithms, which is one of the most suitable algorithms for very low bit-rate speech coder. Then, using residual signal to analyze the performance of algorithms, it improve the definition and antinoise performance of MVF algorithm, and put forward using pitch as a parameter of ruling the performance of incentive model. Finally, through the simulation on the performance of the incentive model, a conclusion can be drawn that the pitch error rate of TBE exciting signal is far less than that of the MELP exciting signal. Therefore, TBE model is closer to original speech features and it can effectively improve the performance of MELP vocoder incentive model.

References

1. Sang-Jin, K., Jong-Jin, K., Minsoo, H.: HMM-based Korean speech synthesis system for hand-held devices. IEEE Trans. Consum. Electron. **52**(4), 1384–1390 (2006)
2. Erro, D., Sainz, I., Navas, E., et al.: Harmonics plus noise model based vocoder for statistical parametric speech synthesis. IEEE J. Sel. Top. Signal Process. **8**(2), 184–194 (2014)
3. Drugman, T., Dutoit, T.: The deterministic plus stochastic model of the residual signal and its applications. IEEE Trans. Audio Speech Lang. Process. **20**(3), 968–981 (2012)

4. Stylianou, Y.: Applying the harmonic plus noise model in concatenative speech synthesis. IEEE Trans. Speech Audio Process. **9**, 21–29 (2001)
5. Drugman, T., Stylianou, Y.: Maximum voiced frequency estimation: exploiting amplitude and phase spectra. IEEE Signal Process. Lett. **21**(10), 1230–1234 (2014)
6. Hermus, K., Van Hamme, H., Irhimeh, S.: Estimation of the voicing cut-off frequency contour based on a cumulative harmonicity score. IEEE Signal Process. Lett. **14**(11), 820–823 (2007)
7. Hermus, K., Girin, L., Van Hamme, H. et al.: Estimation of the voicing cut-off frequency contour of natural speech based on harmonic and aperiodic energies. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4473–4476 (2008)
8. Edge, R.L.: Measuring speech naturalness of children who do and do not stutter: the effect of training and speaker group on speech naturalness ratings and agreement scores when measured by inexperienced listeners [DB/OL]. https://getd.libs.uga.edu/pdfs/edge_robin_l_201208_phd.pdf