An MELP Vocoder Based on UVS and MVF

Tangle Lu and Xiaoqun Zhao(✉)

Tongji University Shanghai, Shanghai, China
{lutangle, zhao_xiaoqun}@tongji.edu.cn

Abstract. Mixed excitation linear prediction (MELP) vocoder is generally used in low bit-rate vocoder, whose target now focuses on overall coding scheme, decrease of coding rate and improvement of robustness. Unvoiced/voiced/silence detective algorithm (UVS) possesses certain robustness and anti-noise property, while voiced excitation model based on maximum voicing frequency algorithm (MVF) is closer to the original speech characteristics. In this paper, the original excitation model of MELP vocoder is replaced and UVS is joined so that an improved 2.4 kbps coding rate vocoder is accomplished. Compared with MELP of federal standards, the improved vocoder owns better synthetic speech quality and robustness.

Keywords: Signal processing · MELP · UVS · MVF · Speech evaluation

1 Introduction

Vocoder makes it possible to transmit speech with limited bandwidth; and receiving end has high intelligibility and naturalness. Military communications need to reduce power consumption; multimedia communication systems need to reduce storage costs; satellite communications are quite short of channel resources in poor communication conditions; underwater communications possess serious signal attenuation. As a result, speech signal should be low bit-rate coded [1, 2].

Research on low bit-rate vocoder focuses primarily on the overall scheme, decrease of coding rate and improvement of robustness. Generally the research object is representative MELP vocoder, which nevertheless has poor synthetic speech naturalness, hum and low robustness in noise environment. Excitation model's performance has a significant impact on synthetic speech quality in encoder, of which multi-band mixed excitation model is superior with small time and space complexity. However sub-band sound intensity error-detection (multi-band causes) can have serious consequences, and literature [3] indicated multi-band model does not match the actual excitation. A better excitation model is required.

Based on the original MELP vocoder, in this paper, new UVS is joined, and excitation model based on MVF closer to original speech characteristic is used. Compared in performance with the one of Federal standard, improved MELP vocoder improves synthetic speech quality and robustness.

X.-L. Huang (Ed.): MLICOM 2016, LNICST 183, pp. 44–52, 2017.
DOI: 10.1007/978-3-319-52730-7_5

2 Principle of MELP Vocoder

2.1 Vocoder Model

By linear prediction [4] speech waveform is analyzed to create channel excitation and parameters of transfer function so that speech waveform coding turns to parameter coding and the amount of data of speech transmission greatly reduces. The improved MELP vocoder retains original part of parameter extraction and quantification and improves the performance of its excitation model, as shown in Fig. 1. Encoder extracts these parameters, including UVS, pitch period, line spectral pairs (LSF), gain, Fourier series values, MVF, which are vectorial or scalar quantized. Decoder interpolates parameter, generates excitation signal, enhances adaptive spectrum, linear predicts, and generates synthetic speech. With gain suppression, noise suppression, pulse discrete filtering and so on, decoder improves the quality of synthetic speech.

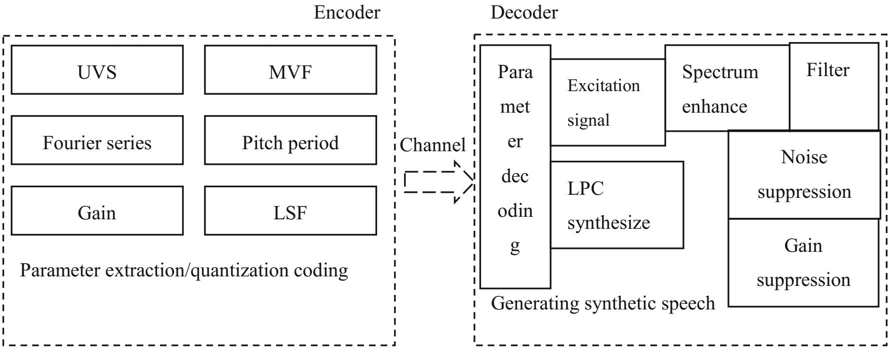


Fig. 1. Improved MELP vocoder

UVS and MVF are extracted according to method in literature [5, 6]. After extracting pitch period, integer pitch period is calculated by normalized correlation function, which fraction pitch period amends to search eventual integer pitch period by recursion. LSF parameter is produced by 15 Hz bandwidth spread on linear forecast coefficients. 10 LSF parameters' interval is at least 50 Hz. If interval is smaller, computation offset needs to be increased between former and latter component. Gain parameter used mean square value of windowing signal. Window's length is pitch period; centers respectively are located in two position of current frame. Calculate gain value two times to prevent frame gain located in the transition part; Fourier series retain the top 10 largest harmonics value, obtained by FFT calculation of residual signal.

Parameters extracted in encoder need further compression at a relatively small number of bits. Pitch period, gain, and Fourier series use scalar quantization; while LSF uses four-stage vector quantization. UVS judgment results and MVF value in improved algorithm use scalar quantization, and UVS uses only 2 bits to fully express three types of frame. According to the distribution of harmonic number, MVF value can be uniform quantized by interval. For high probability numerical interval [0, 10], quantitative interval is 1; interval (10, 30] 4. If harmonic number is greater than 30, MVF value can be considered as full band distribution, so 16 numerical distributions are quantized by 4 bits. Combining UVS judgment results, MVF quantitative bit numbers are further reduced: MVF value of silent frame is 0 and only needs 1 bit to quantize; unvoiced frame ranges from [1, 4] and only 2 bits; voiced frame [8, 24] and 3 bit. MVF with UVS judgment results quantitation can reduce numbers of bits.

Decoder synthesizes speeches. Firstly, parameter interpolation converts frame parameter to pitch period for synchronization interpolation with pitch period. Secondly, excitation signal filtered by adaptive spectrum enhancement is constructed, which makes synthetic speech better match formant waveform. LPC coefficients restored by LSF parameters construct synthetic filter. In addition, synthetic speech's coherence is improved by gain correction, pulse shaping filter.

2.2 Performance of Vocoder

Improved vocoder adds UVS and MVF to enhance anti-noise property of the vocoder. UVS judgement filters some stationary noise interference from time domain. At the same time, quality of synthetic speech in noise environment is improved according to unvoiced frame length that can help preliminarily decide background noise or unvoiced frame of objective speech. From frequency domain MVF technologies filter aliasing noise in speech frames. Harmonics as excitation signal are extracted to avoid impact on synthetic speech from noise component of other frequencies. In addition, for high frequency noise harmonic component, MVF technologies filter by counting harmonic component.

Coding rate of the vocoder remains 2.4 kbps; however, improved schedule owns more redundant bits that can effectively enhance data accuracy in decoder. Tables 1 and 2 respectively are bits allocation of quantized parameters of the original MELP and the improved. Total number of bits of the original MELP vocoder is 54 bits. The improved is only 53 bits in unvoiced frame and about 11 bits in silent frame. Redundant bits that are essential are for error-correcting coding of critical parameters to effectively enhance the robustness of the vocoder.

Table 1. Bits allocation of the original MELP vocoder

	Voiced	Unvoiced
UV	1	1
Pitch period	7	7
LSF	25	25
Gain	8	8
Fourier series	8	–
Sound intensity of sub-band	4	–
Non-cyclical sign	1	–
Forward error correcting coding	–	13

Table 2. Bits allocation of improving the MELP vocoder

	Voiced	Unvoiced	Silent
UVS	2	2	2
Pitch period	7	7	–
LSF	25	25	–
Gain	8	8	8
Fourier series	8	8	–
MVF	3	2	1
Non-cyclical sign	1	1	–
Forward error correcting coding		1	

Computational complexity of the improved vocoder is slightly lower than the original, mainly from the following aspects:

(1) Eliminates the process of five sub-bands, that is to say, sound intensity pf five sub-bands does not need to be calculated.
(2) Reduces the computation of pitch period of silent frames, LSF, Fourier series and other parameters.
(3) Only half frame after some low-complexity operation such as add, minus and square is used to extract UVS judgement eigenvalue, including mean and variance.
(4) For the calculation of peaks, the original MELP used dual hard decision of each sub-band; MVF uses soft decision, only calculating harmonic peaks.
(5) Combining with the UVS judgement, MVF consider judgement result as a prior probability, which greatly reduces computational complexity.

Delay of speech information is reduced because of low computational complexity of the vocoder. In addition, in the process of UVS judgement and MVF excitation model, intermediate variables need less storage space. For UVS, single frame can decide the frame type, where CAMDF mean, CAMDF variance and 4 thresholds are stored. For MVF, space complexity of sub-band voiced/unvoiced decision is close to the original. Intermediate variables are mainly accumulated energy value of candidate harmonics, whose number is about half length of the pitch period.

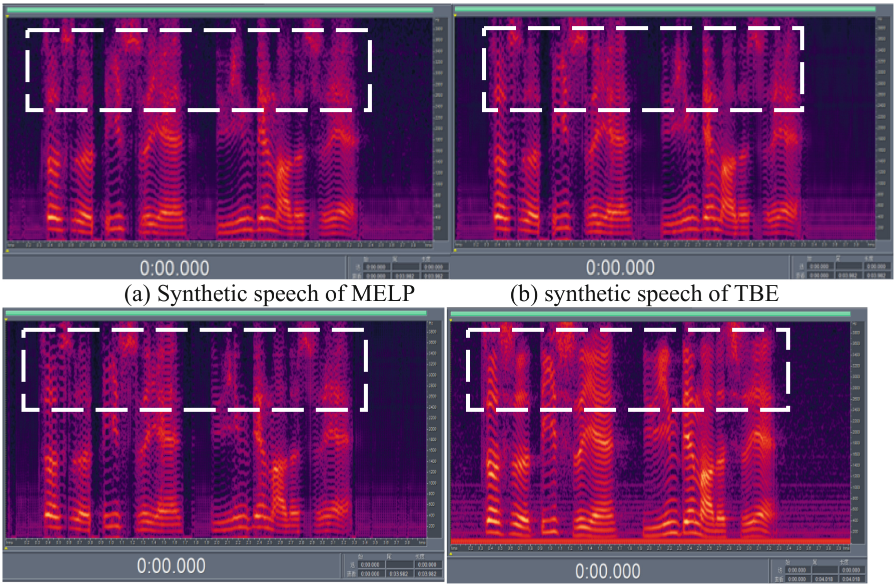
3 Experiment and Analysis

3.1 Simulation

Different coding schemes are implemented in Matlab, including coding schemes of federal standards MELP, TBE as well as improved MELP. By comparing the quality of synthetic speech, optimization of the improved vocoder is proved.

For all speech in corpus, synthetic speech is obtained by three coding schemes above. As shown in Fig. 2, speech of improved MELP can better restore the original speech with its higher similarity with original speech than TBE. At the same time, as shown in dashed line of Fig. 2, part of defect of 2.4 kbps encoder of the Federal standard is effectively improved, with improved the MELP accurately enhancing default high frequency components of voiced frames of the original MELP vocoder.

Speeches in corpus are pure. Through MELP, TBE, and improved MELP vocoder, subjective MOS [7] scores are respectively 3.94, 4.12 and 4.29. If noise library speech is added to pure speech with certain SNR, MOS scores turn to 2.95, 3.21 and 3.32. PESQ [8, 9] scores are also in line with scores rule above, respectively 2.27, 2.30 and 2.31. For noise speech, scores are 1.90, 1.96, and 1.97.



(a) Synthetic speech of MELP

(b) synthetic speech of TBE

(c) Synthetic speech of improved MELP

(d) original speech

Fig. 2. spectrogram of synthetic speech and original speech

3.2 Data Analysis

Robustness of encoder is analyzed from perspectives of gender of the speaker, speech length and noise characteristics.

(1) Gender robustness

Figures 3 and 4 represent PESQ score distribution of part of female and male speaker's speech respectively. Ordinate of each sample corresponds to PESQ scores of three coding schemes.

According to the scores of female speakers, improved MELP coding schemes generally owns higher scores than original MELP coding scheme, with sometimes difference about 0.4, while scores of TBE coding scheme are between MELP and improved MELP, slightly lower than improved MELP. Difference can be less than 0.001 for part of sample points. This is because the part of speech owns voiced features obviously, and few unvoiced frame, leading to similar advantage between UVS judgment and original voiced/unvoiced judgment.

Compared with female speakers, male speakers' synthetic speech of improved MELP is similar to the original MELP, mainly because base frequency of male speaker is lower than female speaker. Most high harmonic of male speaker concentrates about 2 kHz. Voiced frame for high frequency component cannot be accurately divided, so, adaptive MVF and spectrum division of fixed sub-band is close to performance of pure male speech.

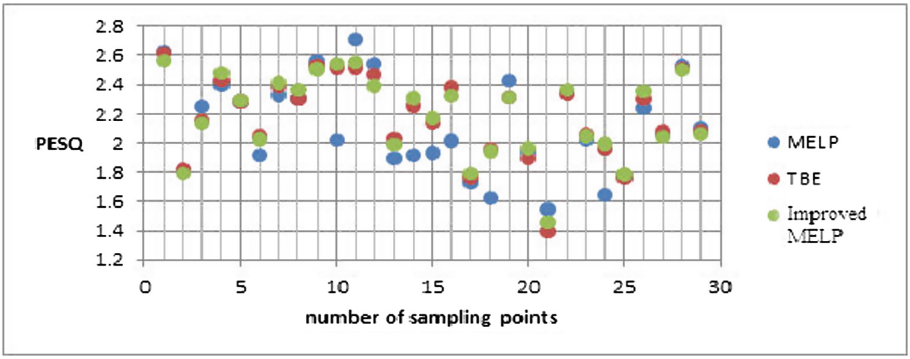


Fig. 3. Distribution of PESQ scores (female speaker)

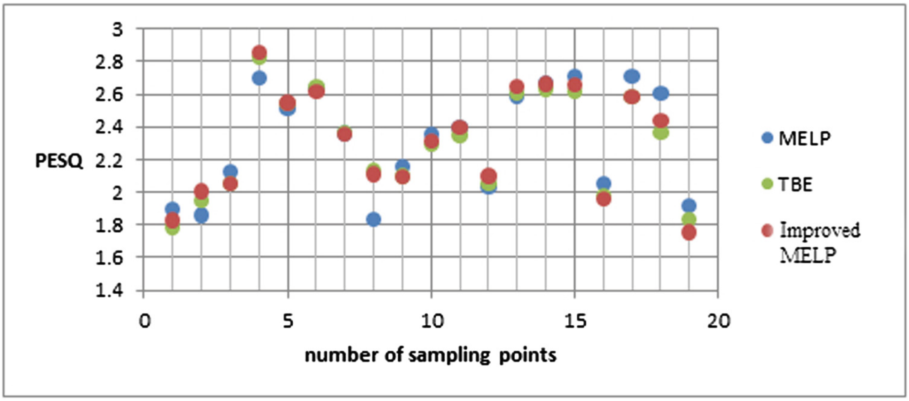


Fig. 4. Distribution of PESQ scores (male speaker)

(2) Time length robustness

Compared are PESQ scores of synthetic speech in corpus of different time length through MELP, TBE and improved MELP vocoder. For example, PESQ of long speech (120 s) of schemes above are respectively 2.87, and 2.88 and 2.89. PESQ of short speech (5 s), which are a little lower than long speech, are respectively 2.20, and 2.23, and 2.24.

For time differences, MELP, TBE and improved MELP vocoder embody the same rule: PESQ score difference is 0.6. For improved MELP vocoder, a long speech is useful for MVF threshold training, which makes the spectrum boundary of voiced/unvoiced frame closer to the original speech. And there is a lot of codebook training of vocal tract parameters in the original MELP vocoder. Long speech is conducive to optimal codebook search, of which improved MELP takes advantage.

In actual communication, length of speech is generally about 5 min [10]. Therefore, improved MELP coding scheme can be applied in the area with good performance.

(3) Noise robustness

Figure 5 and 6 shows PESQ scores in daily noise and military noise with SNR 15 dB to −10 dB. Because of military secrets military noise involved, only F16 cockpit noise data is listed, similar to other military noise in noise robustness.

Data is also divided into male and female, and then scores of the synthetic speech in different noise type are compared. It can be seen that female speech's score is higher than male. This is because the pitch frequency of female speech is higher. In the frequency domain, harmonic intervals of female speech are greater than those of male speech, so noise effect in low frequency for random scatter is relatively small.

MELP vocoder suffers the greatest impact by the noise type of babble and volvo, while improved MELP coding scheme enhance communication quality in these noise environments. For noise type of white and pink environment, improved MELP also enhance score of original MELP, which verifies improved MELP has better noise robustness. For noise type of factory1 F16 where males often appear, improved MELP has performance advantages.

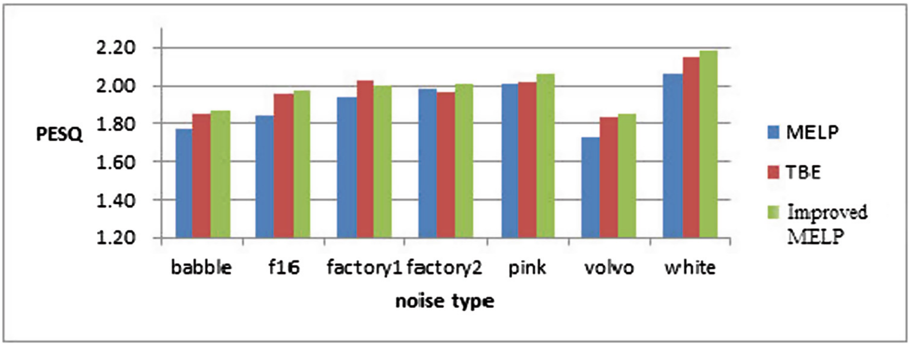


Fig. 5. PESQ in noise environment (female speaker)

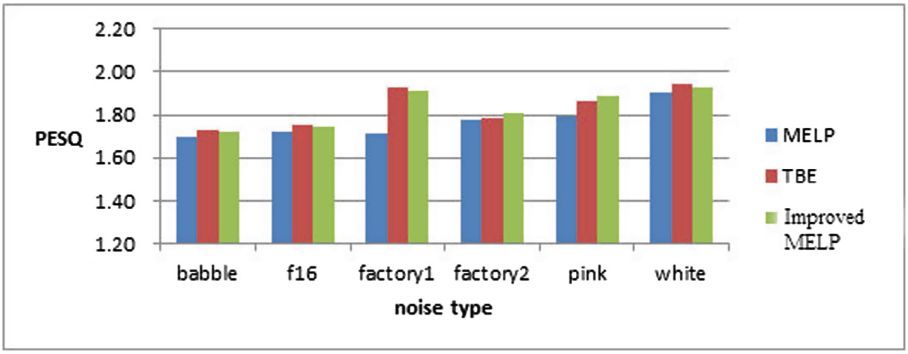


Fig. 6. PESQ in noise environment (male speaker)

4 Conclusion

This paper introduces an improved coding scheme from encoder and decoder, involving coding principle and application of the key technology, including the UVS judgement and MVF excitation model.

Differences in performance between improved MELP coding scheme and the original MELP vocoder are compared from many aspects: anti-noise property, coding rate, computational complexity, and delay and space complexity. In theory the performance advantage of the improved MELP vocoder is verified.

Schemes of MELP, TBE, and improved MELP are implemented in Matlab. Corpus uses PKU-SRSC, noiseX-92, and recorded speech, including diversities of gender, age, rate, intensity, duration, noise environmental and so on. Three coding schemes respectively encode and decode speeches, give synthetic speech PESQ scores, classify/evaluate scores, and verify the robustness of improved MELP coding scheme.

References

1. C114 communications network in China. Development and application of low bit-rate speech coding (2015). http://market.c114.net/154/a190256.html
2. Underwater acoustic communication. http://wiki.dzsc.com/info/7374.html
3. Degottex, G., Stylianou, Y.: Analysis and synthesis of speech using an adaptive full-band harmonic model. IEEE Trans. Audio Speech Lang. Process. **21**(10), 2085–2095 (2013)
4. Zongfu, L.: Multimedia Technology. Tsinghua University Press, Beijing (2009)
5. Jingyun, X., Xiaoqun, Z., Rongyun, L., Jiao, W.: Vocoder excitation model based on voicing cut-off frequency of speech. J. Beijing Univ. Posts Telecommun. **03**, 28–33 (2015)
6. Rongyun, L.I., Xiaoqun, Z., Jingyun, X.U.: Adaptive anti-noise unvoiced/voiced/silence detection algorithm. J. Yanshan Univ. **02**, 133–138 (2015)
7. Rothauser, E.H., et al.: IEEE recommended practice for speech quality measurements. IEEE Trans. Audio Electroacoust. **17**, 227–246 (1969)

8. Conway, A.E.: Output-based method of applying PESQ to measure the perceptual quality of framed speech signals. Wirel. Commun. Netw. Conf. **4**, 2521–2526 (2004)
9. Hines, A., Skoglund, J., Kokaram, A., et al.: Robustness of speech quality metrics to background noise and network degradations: comparing ViSQOL, PESQ and POLQA. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3697–3701 (2013)
10. Wei, D.: Establishment and application of call duration model. Telecommun. Technol. **10**, 58–60 (2001)