I.J. Image, Graphics and Signal Processing, **2010, 2, 46-52**
Published Online December 2010 in MECS (http://www.mecs-press.org/)

A NEW DISTORTION MEASURE FOR PARAMETER QUANTIZATION BASED ON MELP

Ye Li [1]
Shandong Computer Science Center
Shandong Provincial Key Laboratory of computer Network
Jinan, China

liye@keylab.net
Jingde Xu[2], Qinghua Li, Huijuan Cui, Kun Tang
EE Department, Tsinghua University
Beijing, China
xjd07@mails.tsinghua.edu.cn

Abstract—**Parameter quantization is very important for the synthetic speech quality of the vocoder. A new distortion measure for pitch as well as lsf quantization in ultra low bit rate Vocoder, whose parameters for several consecutive frames are grouped into a vector and jointly quantized to obtain high coding efficiency, is proposed based on mixed excitation linear prediction(MELP) vocoder. The product of sum of band pass voicing coefficients and gain parameter is used to denote the weighting factor of pitch as well as lsf parameters of current speech frame in the consecutive frames using weighted squared Euclidean distance measure to search the vector codebook. Comparing with the traditional method for a constant weighting factor by distinguishing Voiced/Unvoiced(UV) pattern of each speech frame, objective test results show that the quantization distortion of pitch is reduced by 3.3% and the mean opinion score (MOS) is increased by almost 0.1(3.5%).**

Index Terms—**speech coding; superframe; pitch quantization;lsf quantization**

I. INTRODUCTION

Low bit rate speech coding has always been one of the most important research areas in speech coding. Especially, 2.4kbp s、1.2kbps、0.6kbps vocoders based on MELP algorithm are widely used in shortwave communication, satellite communication and so on because of its high compression efficiency[1-3].

In 1988, Griffin D W et al. proposed 8kbps Multiband Excitation Vocoder (MBE)[4] which divides input speech into some bands based on the pitch value. Then it analysis the voicing coefficient of each band which can efficiently improve the quality of excitation and output speech.

Pitch is one of the most important parameters in low bit rate speech coding algorithms[5], such as MELP, poor quantization of which will seriously damage the quality of the reconstructed speech. Quantization of LPC coefficients is also an important research area6][7]. As LPC coefficients are not suitable for quantization transmission, generally, they are transformed to LSF coefficients firstly. In low bit rate speech coding, LSFs coefficients are usually quantized with multi frame joint vector quantization method based on interframe prediction. In very low speech coding algorithms, vector quantization (VQ) is adopted to quantize N consecutive parameters together efficiently .

Wei Xuan et al. proposed voiced/unvoiced(UV) classification recovery algorithm[8] in the speech decoder based on GMM, in which he fully considered the correlation between vocoder parameters, used LSFs and energy as a vector and recovered band pass voicing coefficients on the decoder with GMM. Test results show that recovery of UV parameters above in the decoder is accurate and it can efficiently save the UV quantization bit of UV which is of great significance in ultra low bit speech coding. The results also show that pitch、gain、UV and LSFs are highly correlated. Encoding efficiency and quality of synthesized speech can be improved using the correlation.

T.Wang et al. proposed a 1200bps vocoder based on MELP [9]. This paper divides pitch into different modes according to different UV values. In different mode, it adopts different bit allocation and codebook for the quantization of pitch and LSFs, which improves the quantization accuracy of parameters as a result. To the quantization of pitch, this paper also proposed a weighted squared Euclidean distance measure to search the codebook for the most approximate codeword and the weighting factors depend on the binary UV classification only. Product of weighting factor and power spectrum corresponding to LPC coefficients is used as weight to LSFs parameters.

Grant sponsor: Natural Sciences Foundation of China (Grant No. 60572081)

Based on [9], MELP-based joint optimization algorithm for multi-parameter codebook size is proposed in [10]. This algorithm divides multiple modes according to different quantization indexes of UV parameters and allocates different codebook sizes for pitch, gain and LSFs in different modes, especially in enhanced fully-voiced mode, This could efficiently improve the quantization performance and synthetic speech quality.

However, the method above does not consider the difference of variant voiced frames. Besides, it does not consider the gain parameter either, which is very important for subjective perception. In this paper, we propose a variable weighting factor based distortion measure for pitch VQ. Firstly, we denote each sub-band's voicing strength by band pass voicing coefficient (BPVC) and then compute the whole speech frame's voicing strength by the sum of BPVCs. Secondly, we use the whole speech frame's voicing strength above to describe the difference between variant voiced frames. At last, we use the multiplication of the gain parameter and the sum of five BPVCs as each frame's weighting factor for distortion measure. Simulation results show that the pitch distortion of reconstructed speech is reduced by almost 3.3% and the mean opinion score (MOS) is increased by almost 0.09(3.5%).

II. MIXED EXCITATION LINEAR PREDICTION (MELP).

Based on Linear Prediction Coder (LPC10) model, the mixed excitation linear prediction(MELP) vocoder is improved and it contains five differences: mixed excitation, pulse dispersion, aperiodic flag, adaptive spectral enhancement, and Fourier magnitude modeling [11-13][.

The mixed excitation means the vocoder divides one speech frame into five sub-bands at the bands of [0, 500]Hz, [500, 1000]Hz, [1000, 2000]Hz, [2000, 3000]Hz, [3000, 4000]Hz by five adaptive band pass filters. It could reduce the buzz associated with LPC10 vocoder effectively .

The adaptive spectral enhancement filter which gives a high quality to the synthetic speech is based on the poles of the linear prediction synthesis filter[13].

The decoder will use either periodic or aperiodic pulses to be the excitation when the input speech is voiced. During transition frames, the coder is often using aperiodic pulses which replace the periodic pulses and could reduce the tonal sounds efficiently. Periodic pulses are used in smooth frames while noise pulses are used when the input speech is unvoiced.

In the MELP based vocoders, the length of one frame is usually 22.5ms, which includes 180 samples. The parameters to be quantized in the encoder contains: the linear spectral frequency (LSF), the band pass voicing coefficients, the pitch, the Fourier magnitudes as well as the aperiodic flag. In the decoder, all the parameters are linear interpolated and the excitation passes the linear prediction synthesis filter to synthesize the final speech.

The principal of both the encoder and the decoder based on MELP could be depicted as Table 1 and Table 2.

TABLE 1. THE ENCODER OF MELP

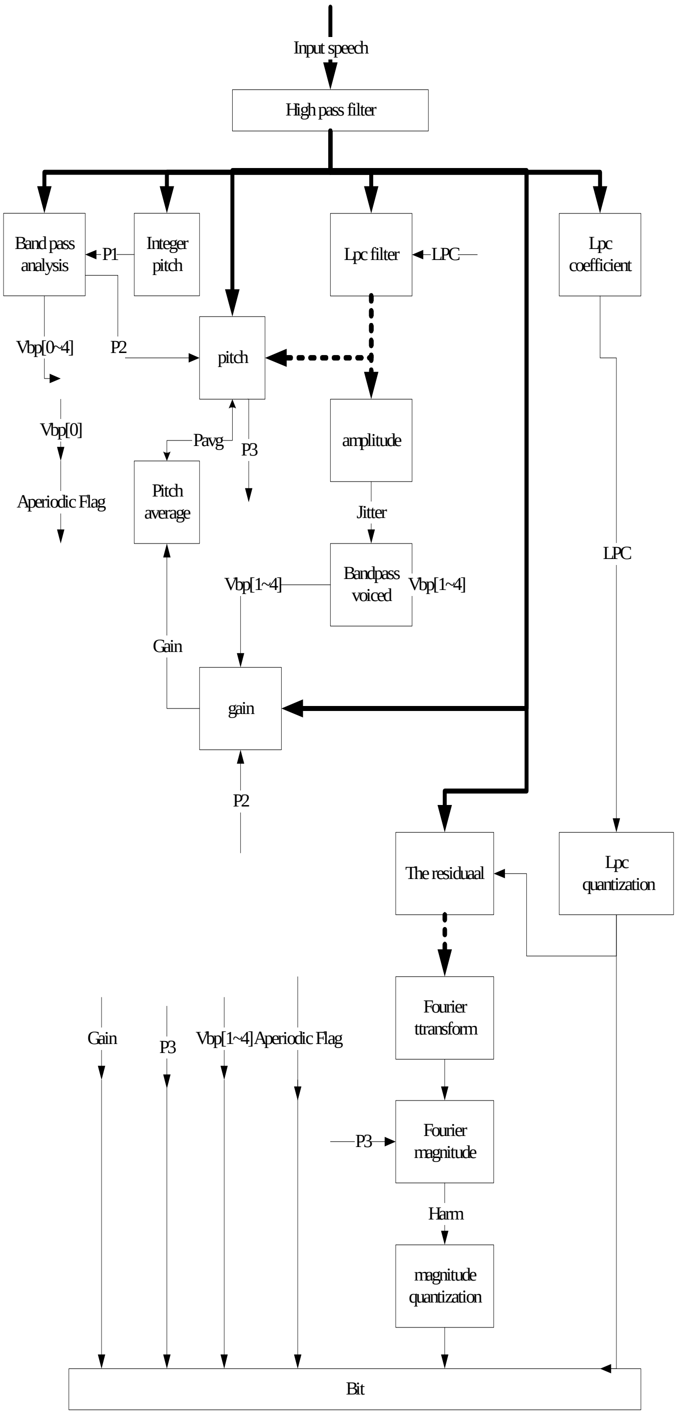
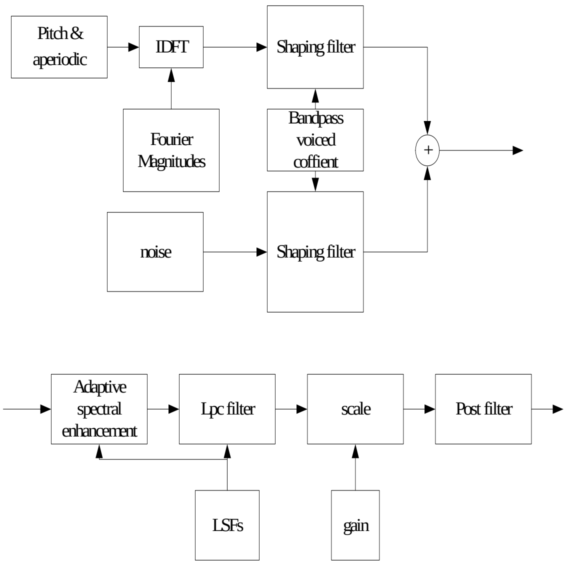


TABLE 2. THE DECODER OF MELP



III. RELATED WORK

Alan V. McCree, Thomas P. and Barnwell III [11][12] proposed a method to divide one speech frame into five sub-bands at the bands of [0, 500]Hz, [500, 1000]Hz, [1000, 2000]Hz, [2000, 3000]Hz, [3000, 4000]Hz by five bandpass filters, and then determine the UV state with computed BPVC for each sub-band-pass signals. In MELP, the first sub-band's UV state is the same as the speech frame's UV pattern. The weighting factor for pitch VQ is 1 for voiced frame and 0 for unvoiced frame [9]:

$$w = \begin{cases} 1.0, & \textit{for unvoiced frame} \\ 0.1, & \textit{for voiced frame} \end{cases} \tag{1}$$

For 10-dimension LSFs parameters, a commonly used distortion measure is as follows:

$$w(i,j) = \begin{cases} P_i(f_j)^{0.3}, & 1 \le j \le 8 \\ 0.64P_i(f_j)^{0.3}, & j = 9 \\ 0.16P_i(f_j)^{0.3}, & j = 10 \end{cases} \tag{2}$$

Where j denotes 1-10 parameter of LSFs, $P_i(f_j)$ denotes short time spectrum amplitude in the j coefficient frequency of the i frame in the superframe to be quantized.

However, the weighting factor for pitch vector codebook search takes only the UV pattern which is also the UV state of the first sub-band into account and the weighting factor for LSFs does not consider the difference between the consecutive speech frames. The mixtures of a periodic impulse train and white noise are used to excite an all-pole filter [14-17]. Generally speaking, voiced speech has a periodic impulse feature while unvoiced speech has a white noise-like feature. Unvoiced speech has a flat power spectrum without periodicity and its pitch parameter is not quantized. Therefore, unvoiced and voiced frames are treated with different weighting factors during the search in the codebook. However, this method does not consider the difference of variant voiced frames.

IV. NEW DISTORTION MEASURE

a)

New distortion measure of pitch: We assume that a superframe consists of consecutive N speech frames. Then we can quantize the parameters jointly to obtain high coding efficiency. In a superframe, the ith frame's pitch parameter is denoted by P_i and then transformed into logarithmic value which is denoted by p_i prior to the quantization, so the pitch vector of consecutive N frames is denoted by $\boldsymbol{p}$, which can be expressed as follow:

$$\boldsymbol{p} = [p_1, ..., p_N]_{1\times N} . \tag{3}$$

To search the codebook for the most approximate codeword, we use weighted squared Euclidean distance, expressed as follow:

$$D_w(\boldsymbol{p}, \bar{\boldsymbol{p}}) = \sum_{i=1}^{N} W_i (p_i - \bar{p}_i)^2 , \tag{4}$$

where p_i and $\bar{p}_i$ are the unquantized and quantized logarithmic pitch values of the ith speech frame respectively, W_i is the weighting coefficient of the ith speech frame.

In [11][12], where a low bit rate speech coding algorithm based on MELP was proposed, the excitation signals of the decoding section are the sum of five sub-band pass signals:

$$e(n) = e_p(n) \cdot \left(\sum_{i=1}^{5} \bar{b}_i \cdot h_i(n) \right) + noise(n) \cdot \left(\sum_{i=1}^{5} (1 - \bar{b}_i) \cdot h_i(n) \right) \tag{5}$$

where

$$e_p(n) = \sum_{k=1}^{K} p_k \cdot \cos[wkn + \varphi(n,k)] . \tag{6}$$

In (5), n is the time of samples, $e_p(n)$ are the harmonic signals synthesized by pitch parameter, $\bar{b}_1$、$\bar{b}_2$、$\bar{b}_3$、$\bar{b}_4$、$\bar{b}_5$ are the five interpolated BPVCs which are the voicing strength in each frequency band at the decoder part, $noise(n)$ are the normalized white noise signals，and $h_i(n)$ is the impulse response of five bandpass filters.

In (6), K is the total number of spectral components within the passband, w is the angular frequency responding to the interpolated pitch parameter, p_k is the interpolated Fourier Magnitudes of the kth harmonic, $\varphi(n,k)$ is the phase spectral components which can ensure the continuity at the margin of the speech frame.

From (5) and (6), all the five BPVCs can affect the excitation of reconstructed speech. The higher of the BPVC of the sub-band, the higher proportion of the harmonic signals in the ith sub-band excitation signals. So, the weighting factor should consider the sum of BPVCs, commonly be proportional to the sum of BPVCs:

$$W_i \propto \sum_{j=1}^{5} b_j \ , \tag{7}$$

where b_j is the BPVC of the jth sub-band of current speech frame, W_i is the weighting factor of current frame in the superframe for pitch VQ when using the weighted squared Euclidean distance measure to find out the codeword in the codebook.

Besides, the gain of a speech frame has an important influence on the subjective perception of the reconstructed speech. For a given speech frame, higher gain brings greater effect on subjective perception. Therefore, the pitch parameter should be quantized more accurately for the frames with higher gain. That is to say, the frames with higher gain must be treated with higher weighting factor, commonly be proportional to gain:

$$W_i \propto g_i \, , \tag{8}$$

where g_i is the gain parameter of the current frame.

From all above, the weighting factor for pitch VQ should consider both the sum of BPVCs and the gain parameter. When the sum of BPVCs is the same, the weighting factor for the speech frame with a higher gain should be higher in the superframe. Similarly, when the gain is the same, the weighting factor of the speech frame should be higher in the superframe if the sum of BPVCs is higher. Therefore, choosing the product of gain parameter and the sum of BPVCs to be the weighting factor of current frame in the superframe is a simple and effective method. The new distortion measure for pitch VQ is as follow：

$$D_w(\boldsymbol{p}, \bar{\boldsymbol{p}}) = \sum_{i=1}^{N} W_i (p_i - \bar{p}_i)^2 \, , \tag{9}$$

where

$$W_i = (\sum_{j=1}^{5} b_j) \cdot g_i \, . \tag{10}$$

b)

New distortion measure of LSFs: In the superframe comprised of N frames, LSFs parameters consist a 1× 10N vector:

$$\boldsymbol{f} = [\, f_{1,1}, f_{1,2}, ..., f_{N,10}]_{1\times 10N} \tag{11}$$

We adopt the following weighted Euclidean distance to search for an optimal codebook word:

$$D_w(\boldsymbol{f}, \bar{\boldsymbol{f}}) = \sum_{i=1}^{N} \sum_{j=1}^{10} W_{i,j} (f_{i,j} - \bar{f}_{i,j})^2 \tag{12}$$

As mentioned in the above chapter, UV parameters of 5 bands determine the composition of excitation signal, which is very important for synthesized speech signal. Therefore, the higher degree of voiced value, the higher weight should be given during LSFs parameter quantization.

$$W_{i,j} \propto \sum_{j=1}^{5} b_j \tag{13}$$

Meanwhile, as human ear is very sensitive to energy, the higher the energy of shortframe is ,the higher weight should be given during LSFs parameters quantization.

$$W_{i,j} \propto g_i \tag{14}$$

In conclusion, in searching for codebook of LSFs parameters, this paper chooses product of energy and voice degree as quantization weight:

$$D_w(\boldsymbol{f}, \bar{\boldsymbol{f}}) = \sum_{i=1}^{N} \sum_{j=1}^{10} W_{i,j} (f_{i,j} - \bar{f}_{i,j})^2 \tag{15}$$

Where, $b_{i,k}$ denotes voiced degree of k th subband of i th frame, j denotes1-10 parameter of LSF, $P_i(f_j)$ denotes short time spectrum amplitude in the j th coefficient frequency of the i th frame in the super frame to be quantized.

Here,

$$w(i,j) = \begin{cases} (\sum_{k=1}^{5} b_{i,k}) \cdot g_i \cdot P_i(f_j)^{0.3}, & 1 \le j \le 8 \\ (\sum_{k=1}^{5} b_{i,k}) \cdot g_i \cdot 0.64 P_i(f_j)^{0.3}, & j = 9 \\ (\sum_{k=1}^{5} b_{i,k}) \cdot g_i \cdot 0.16 P_i(f_j)^{0.3}, & j = 10 \end{cases} \tag{16}$$

V. SIMULATION RESULTS

a)

To compare the traditional weighting factor computation method in paper [9] with ours, both of them use their own weighting factor to train the pitch parameter VQ codebook according to simulated annealing algorithm [19].

The simulation platform utilizes a 300bps speech coding algorithm based on MELP. We form a superframe with consecutive 6 frames, and allocate 8 bits for pitch parameter VQ. Five standard speech files which contain the voice of several men and women are used for testing. MOS is tested according to the ITU P. 862 recommendation. The result is shown in TABLE 3.

TABLE 3. THE MOS COMPARISON OF RECONSTRUCTED SPEECH

Testing Speech	MOS with the traditional method	MOS with the proposed method
File 1	2.544	2.625
File 2	2.468	2.556
File 3	2.470	2.570
File 4	2.458	2.563
File 5	2.560	2.627
average	2.500	2.588

On account of the different distortion measure to search in the pitch vector codebook, it is difficult to find out a direct criterion to measure the quantization performance of these two methods. To measure the quantization performance of the two methods, we adopt the two different distortion measures to quantize pitch parameter and reconstruct speech. The average Euclidean distance (AED) is used to describe the degree of similarity between pitch parameter extracted from the reconstructed speech and the original speech:

$$D(\boldsymbol{P},\boldsymbol{P}^{'}) = \sum_{j=1}^{M}\sum_{i=1}^{N}(p_i - p_i^{'})^2 \Big/ M \quad ,$$

(17) where $\boldsymbol{P}$ is the pitch parameter vector extracted from the original speech, $\boldsymbol{P}$' is the pitch parameter vector extracted from the reconstructed speech , M is the number of superframes in the test speech. The result is shown as TABLE 4:

TABLE 4. THE AED COMPARISON OF RECONSTRUCTED SPEECH

Testing Speech	AED with the traditional method	AED with the proposed method
File 1	637.50	607.64
File 2	638.92	683.64
File 3	848.22	752.32
File 4	747.34	728.92
File 5	574.55	563.80
average	689.31	667.26

From TABLE II, compared with the traditional method, the average AED with the novel method decreases by almost 3.3%. It means that the degree of similarity between the pitch parameter extracted from the reconstructed speech and that from the original speech is also improved with the new method. It also means that using the new distortion measure to quantize the pitch parameter is more precise than the traditional method.

For the ultra low bit rate speech coding algorithm, it is difficult to improve the MOS since the number of quantization bits is limited. Using this method, the MOS of reconstructed speech raises by almost 0.09 and the listening quality of reconstructed speech is greatly improved.

b)

Based on 300bps MELP vocoder, this paper does simulation test to 5 standard Chinese sentences with MOS of ITU P.862. Results show as follows:

TABLE5. THE MOS COMPARISON OF RECONSTRUCTED SPEECH

Testing Speech	MOS with the traditional method	MOS with the proposed method
File 1	2.625	2.650
File 2	2.556	2.582
File 3	2.570	2.593
File 4	2.563	2.579
File 5	2.627	2.663
average	2.588	2.6143

To LSFs, the normal distortion measure is Spectral Distortion(SD). The SD of i the frame is below:

$$d_s(\mathbf{a}_i, \hat{\mathbf{a}}_i) = \sqrt{\frac{1}{F_u}\int_0^{F_u} 10\lg(\frac{S_i(f)}{\hat{S}_i(f)})^2 df}$$

Where F_u is 3k , $\mathbf{a}_i$ is the unquantized LSFs parameters and $S_i(f)$ denotes short time spectrum amplitude of the i th frame in the super frame to be quantized while $\hat{\mathbf{a}}_i$ is the quantized LSFs parameters and $\hat{S}_i(f)$ denotes quantized short time spectrum amplitude of the i th frame. Because Average Spectral Distortion(ASD) is usually representing the quality of the quantization of LSFs, we use ASD in this paper.

TABLE 6. THE AED COMPARISON OF RECONSTRUCTED SPEECH

Testing Speech	ASD with the proposed method	ASD with the traditional method
File 1	2.501	2.658167
File 2	2.514	2.664359
File 3	2.536	2.701203
File 4	2.550	2.712839
File 5	2.393	2.532182
average	2.499	2.653

c)

In order to illustrate the advantage of the proposed method compared the traditional method, we give the Fourier magnitude figures of synthesized speech waves based on 300bps MELP vocoder. It can be seen that pitch is smoother in figure 2 compared to that in figure 3. On the other hand, the Fourier magnitude in figure 4 is near to the original's and LSFs are quantized efficiently.

Fig. 1. The original speech

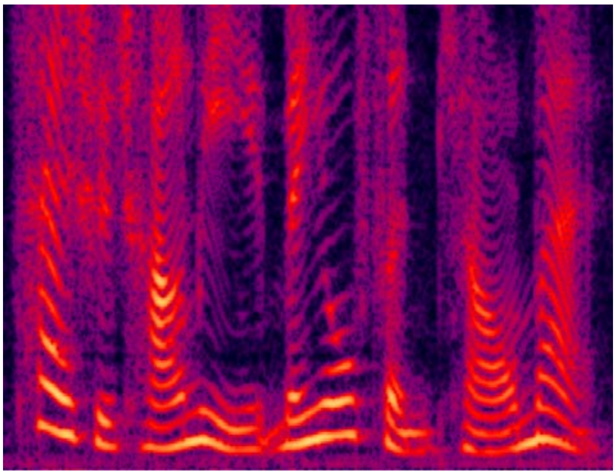


Fig. 2. The output speech used traditional method

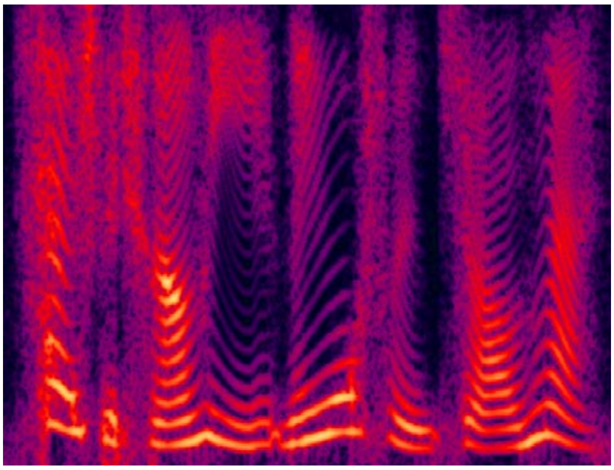


Fig. 3. The output speech used new distortion measure to pitch

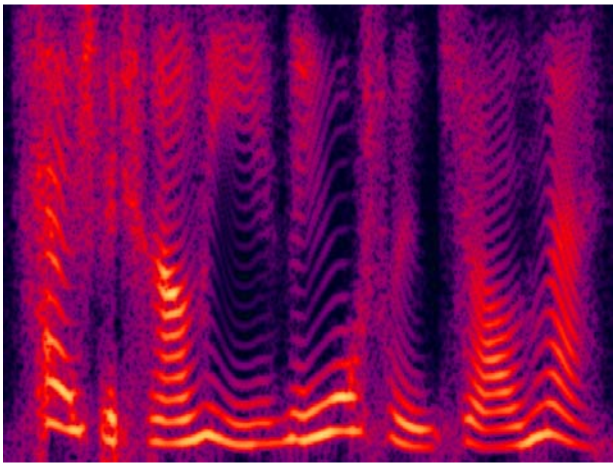
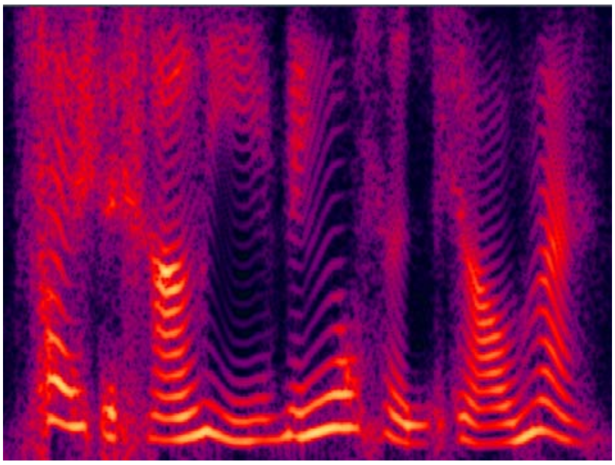


Fig. 4. The output speech used new distortion measure to LSFs



VI. Conclusion

In this paper, we propose a variable weighting factor based distortion measure for codebook search of pitch and LSFs VQ based on MELP. To search the codeword in the codebook using weighted squared Euclidean distance measure, we compute the weighting factor with the product of BPVCs' sum and gain parameter. The proposed method considers the importance of gain parameter to subjective perception and the difference of variant voiced frame by voicing strength, and improves the quantization precision of those frames of speech with higher gain and greater sum of BPVCs. Simulation results show that our method reduces the pitch distortion of reconstructed speech by 3.3% and increases the total MOS of the reconstructed speech by 0.1. This method has been successfully applied to joint vector quantized 300bps-1200bps speech coding algorithm based on MELP.

Acknowledgment

This work is supported by Natural Sciences Foundation of China (Grant No. 60572081).

References

[1] D.P. Kemp, J.S. Collura, T.E. Tremain, “Multi-frame coding of LPC parameters at 600-800 bps,” Proc. IEEE Inter. Conf. Acoustics, Speech and Signal Processing, vol. 1, pp. 609-612, 1991.

[2] Ching W S, Wong W C, Bay H S. A very low bit-rate matrix quantized speech coder with Gray coding. International Symposium on Speech, Image Processing and Neural Networks Proceedings ISSIPNN-1994. New York, NY, USA: IEEE Press, 1994. 468~471.

[3] Athaudage C N, Bradley A B; Lech M. Optimization of a temporal decomposition model of speech. Proceedings of the Fifth International Symposium on Signal Processing and its Applications ISSPA-1999. Brisbane, Qld., Australia: Queensland Univ. Technol, 1999. 471~474.

[4] Griffin D W, Lim J S. Multiband Excitation Vocoder. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1988, 36(8): 1223~1235.

[5] Ney H. A dynamic programming technique for nonlinear smoothing. ICASSP, Atlanta, USA: IEEE Press, 1981:62-65.

[6] Sung-Joo K, Yung-Hwan O. Efficient quantisation method for LSF parameters based on restricted temporal decomposition. Electronics Letters, 1999, 35(12): 962~964.

[7] Phu C N, Akagi M. Improvement of the restricted temporal decomposition method for line spectral frequency parameters. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP2002. Orlando, FL, USA: IEEE Press, 2002. 265~268.

[8] Wei X, Dang X Y, Cui H J, et al. Voiced/Unvoiced classification recovery in the speech decoder based on GMM. International Conference on Signal Processing, Beijing: IEEE Press, 2008:546-548.

[9] T. Wang, K. Koishida, V. Cuperman, A. Gersho, J. S. Collura, “A 1200 bps speech coder based on MELP,” Proc. IEEE Inter. Conf. on Acoustics, Speech, and Signal Processing, vol. 3, 5-9, June 2000, pp. 1375 – 1378

[10] XU Ming , LI Ye , Cui Hj , TANG K. Joint optimization algorithm for multi-parameter codebook size. The 9th International Conference on Signal Processing (ICSP), Peking, China, Oct 2008, Vol (1) : 514-517.

[11] A.V. McCree, T.P. Barnwell III. “A mixed excitation LPC vocoder model for low bit rate speech coding,” IEEE trans. Speech Audio Process., 1995, 3(4), pp. 242-250.

[12] McAulay R J, Quatieri T F. Multirate sinusoidal transform coding at rates from 2.4 kbps to 8 kbps. ICASSP, Dallas, USA: IEEE Press, 1987:1645-1648.

[13] Kohler M A. Comparison of the new 2400 Bps MELP federal standard with other standard coders. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-1997. Munich, Germany: IEEE Press, 1997. 1587~1590.

[14] McCree A V, Barnwell T P. A mixed excitation LPC vocoder model for low bit rate speech coding. IEEE Transactions on Speech and Audio Processing, 1995.

[15] T. Wang, K. Koishida, V. Cuperman, A. Gersho, J. S. Collura,, “A 1200/2400 bps coding suite based on MELP”, Speech Coding IEEE Workshop Proceedings, 6-9 Oct 2002, pp. 90 – 92.

[16] J. Makhoul, R. Viswanathan, R. Schwartz and A. W. F. Huggins, “A mixed-source model for speech compression and synthesis,” J. Acoust. Soc. Amer., vol. 64, pp. 1577-1581, Dec. 1978.

[17] S. Y. Kwon and A. J. Goldberg, “An enhanced LPC vocoder with no voiced unvoiced switch,” IEEE Trans. Acoust., Speech, Signal Processing,vol. ASSP-32, pp. 851-858, Aug. 1984.

[18] ZHAO Ming, Research on ultra low bit rate speech coding techniques and algorithms[D]. Tsinghua University, Beijing, 2004. (in Chinese)

[19] K. Zeger and A. Gersho, “Pseudo-Gray Coding,” IEEE Trans. On Communications, vol. 38, pp. 2147-2158, Dec. 1990.