Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 55 (2013) 619–640

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Compressed domain speech enhancement method based on ITU-T G.722.2

Bingyin Xia, Changchun Bao *

Speech and Audio signal Processing Lab, School of Electronic Information and Control Engineering, Beijing University of Technology, Beijing 100124, China

Received 6 August 2012; received in revised form 24 January 2013; accepted 6 February 2013
Available online 13 February 2013

Abstract

Based on the bit-stream of ITU-T G.722.2 speech coding standard, through the modification of codebook gains in the codec, a compressed domain speech enhancement method that is compatible with the discontinuous transmission (DTX) mode and frame erasure condition is proposed in this paper. In non-DTX mode, the Voice Activity Detection (VAD) is carried out in the compressed domain, and the background noise is classified into full-band distributed noise and low-frequency distributed noise. Then, the noise intensity is estimated based on the algebraic codebook power, and the *a priori* SNR is estimated according to the noise type. Next, the codebook gains are jointly modified under the rule of energy compensation. Especially, the adaptive comb filter is adopted to remove the residual noise in the excitation signal in low-frequency distributed noise. Finally, the modified codebook gains are re-quantized in speech or excitation domain. For non-speech frames in DTX mode, the logarithmic frame energy is attenuated to remove the noise, while the spectral envelope is kept unchanged. When frame erasure occurs, the recovered algebraic codebook gain is exponentially attenuated, and based on the reconstructed algebraic codebook vector, all the codec parameters are re-quantized to form the error concealed bit-stream. The result of performance evaluation under ITU-T G.160 shows that, with much lower computational complexity, better noise reduction, SNR improvement, and objective speech quality performances are achieved by the proposed method comparing with the state-of-art compressed domain methods. The subjective speech quality test shows that, the speech quality of the proposed method is better than the method that only modifies the algebraic codebook gain, and similar to the one with the assistance of linear domain speech enhancement method.

Keywords: Speech enhancement; Compressed domain; CELP; G.722.2; Parameter modification

1. Introduction

For the mobile communication system which is often operated in the complex environments, the background noise is the main impairment to the speech quality. So it is necessary to adopt speech enhancement module in the mobile communication system to reduce the effect of noise and improve the quality of speech communication.

Generally, the state-of-the-art speech enhancement algorithms can be classified into two categories, Linear Domain (LD) speech enhancement and Compressed Domain (CD) speech enhancement.

Linear domain speech enhancement methods are often applied on the input speech signal in PCM format. This kind of method is generally a pre-processing module in front of speech codec, which is in the terminal devices of mobile communication network. However, due to the limitation of power consumption, storage space and cost, the performance of LD speech enhancement method used in terminal devices may not meet the requirement from users. On the other hand, when the LD method is used in the network equipment such as base station or media gateway, the noisy speech is first decoded, then processed by the LD enhancement method, and finally re-encoded to get the

* Corresponding author. Tel.: +86 10 67391635; fax: +86 10 67391625.
E-mail address: baochch@bjut.edu.cn (C. Bao).

0167-6393/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2013.02.001

output bit-stream. As there are full decoding and re-encoding processes involved, the additional delay, computational complexity, and speech quality degradation are usually not acceptable in practical applications.

Compressed domain speech enhancement method, on the other hand, is operated on the encoded bit-stream of noisy speech. In the CD method, only some codec parameters will be decoded, modified under a certain rule, and finally re-quantized and written back to the bit-stream. The CD method could achieve modest amount of noise reduction and speech quality improvement, while the computational complexity is relatively small, and no additional delay is introduced. As a result, it is suitable for the application in the network equipments at the base station or media gateway.

According to the above analysis, compressed domain speech enhancement based on the modification of codec parameters is an appropriate solution of speech enhancement in the network equipments of mobile communication.

In the resent years, researchers have paid more and more attention to the compressed domain speech enhancement. Code-Excited Linear Prediction (CELP) (Schroeder and Atal, 1985) is the most widely used model in low bit-rate speech coding, which is often adopted in the mobile networks. Most of the research works are focused on this model.

The block diagram of CELP speech codec with two-stage codebook structure is shown in Fig. 1.

The generation of speech signal is represented by the source-filter model in CELP codec. The transfer function of human vocal tract is modeled by an all-pole model with a certain order. The excitation signal of synthesis filter is formed by the weighted average of the adaptive codebook and the fixed codebook vectors. The adaptive codebook represents the periodic components of speech signal, and the fixed codebook, which represents the stochastic components, is composed of some excitation vectors. In

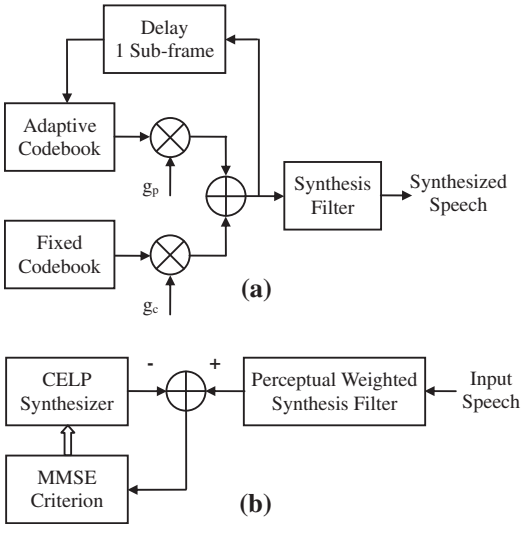

Fig. 1. Block diagram of CELP codec with two-stage codebook structure. (a) synthesizer; (b) analyzer.

the CELP analyzer, the excitation vector with the minimum subjective distortion is obtained by minimizing the perceptual weighted error between the input speech and the synthesized speech using closed loop optimization. There are four kinds of parameters extracted and transmitted by the CELP codec, including short-time spectral parameters (such as Linear Predictive Coding (LPC) coefficients, or Line Spectral Frequency (LSF), or Immittance Spectral Frequency (ISF)), pitch, fixed codebook index, and the codebook gains of adaptive and fixed codebooks.

The research on the compressed domain speech enhancement algorithm based on CELP model started from 2000. Until now, the research is mainly focused on the modification of codebook gains, and some reported work tries to reduce the effect of noise by the modification of LPC coefficients.

In 2000, Ravi Chandran proposed the first compressed domain method (Chandran and Marchok, 2000). The noise intensity is estimated by the assistance of VAD in speech codec. The algebraic codebook gain is modified under the rule in which the noise reduction and speech distortion are considered at the same time. And the adaptive codebook gain is slightly modified to avoid the loss of signal power. In high SNR conditions, this method could achieve modest amount of noise reduction, and the subjective quality is improved to some extent, whereas there are still some distortions introduced to the enhanced speech.

In (Duetsch et al., 2004; Taddei et al., 2004), Herve Taddei proposed two compressed domain speech enhancement methods that only the fixed codebook gain is modified. In (Taddei et al., 2004), the fixed codebook gain corresponding to the noise is estimated by the method of Minimum Statistics (MS) (Martin, 1994). Then the *a priori* SNR is estimated by the decision-directed approach. The modified fixed codebook gain is obtained by a scaling factor with the form of spectral subtraction or Wiener filter. And a post filter is adopted to minimize the loss of speech power. Since no compensation is performed on the adaptive codebook gain, and the post-filtering is not effective in the preservation of speech power, the voiced speech segment is likely to be over-attenuated, and the subjective listening quality is degraded.

In the method described in the patent which was published by Sukkar et al. (2006), the noisy speech is decoded and processed by a LD speech enhancement method. Then a scaling factor for the codebook gains is calculated based on the noisy and enhanced speech signals. The adaptive codebook gain is modified first, and the fixed codebook gain is obtained by maintaining the power of excitation signal. This method is not an actual compressed domain method, and the computational complexity is too high. On the other hand, the speech signal in a single frame is modified as a whole, so there is not sufficient noise reduction in the voiced speech segments.

In 2007, Emmanuel Thepie Fapi et al. proposed a CD speech enhancement method based on the modification of LPC parameters (Fapi et al., 2008), while the excitation

parameters are kept unchanged. The noisy speech is decoded, and then the speech and noise segments are discriminated by VAD algorithm in speech codec. In the noise period, the LPC and autocorrelation coefficients of noise are estimated, and the spectral envelope is whitened to reduce the noise effect. The LPC coefficients of clean speech are estimated using the relationship between the LPC of speech and noise signal. This method has a very high computational complexity. The spectral damping in noise period is likely to change the characteristics of background noise. The frequent modification on the poles of LPC filter and the poor smoothness of spectral envelope between adjacent frames will introduce some artifacts into the enhanced speech.

At present, most of the research on the compressed domain speech enhancement does not involve the discontinuous transmission (DTX) and frame erasure concealment (FEC) functions. When the DTX function is adopted, by the assistance of the VAD method, speech segments are encoded at a high bit-rate whereas non-speech segments are encoded at a low bit-rate. As a result, the transmission efficiency is improved evidently. On the other hand, the FEC function is used to recover the lost parameters when frame erasure occurs. Speech enhancement in the compressed domain should have the compatibility with these two functions as they are often adopted in practical applications.

In this paper, a compressed domain speech enhancement method based on ITU-T G.722.2 speech codec is proposed. The proposed method can be used in all the coding modes of ITU-T G.722.2 speech codec, and is compatible with DTX and FEC functions. The result of performance evaluation shows that, in comparison with the state-of-the-art CD enhancement methods, with a relatively lower computational complexity, the proposed method could provide larger amount of noise reduction and SNR improvement, and the objective and subjective speech quality is improved evidently at the same time.

The rest of this paper is organized as follows. In Section 2, we will have a brief review of ITU-T G.722.2 codec. Then the effect of noise on the speech codec parameters is discussed in Section 3. The proposed compressed domain speech enhancement method in non-DTX mode, DTX mode and frame erasure condition are described in Section 4, 5, 6, respectively. The performance evaluation and discussion are presented in Section 7. And finally we come to the conclusion.

2. Overview of ITU-T G.722.2 codec

ITU-T G.722.2 (ITU-T, 2003) is a wideband speech codec used in the mobile communication systems, which has 9 bit-rates from 23.85 kbps to 6.60 kbps. G.722.2 is based on Algebraic Code-Excited Linear Prediction (ACELP), which is a variation of the CELP model. The stochastic noise codebook in the CELP model is replaced by the stochastic codebook with some kind of algebraic structure. And the algebraic codebook does not need space to store. The block diagram of G.722.2 encoder is shown in Fig. 2.

The sampling rate of the input speech signal is 16 kHz. The frame length of the encoder is 20 ms, which is divided into four sub-frames of 5 ms each. The encoder performs the analysis of LPC, long-term prediction and fixed codebook parameters at the sampling rate of 12.8 kHz. At each frame, the speech signal is analyzed to extract the parameters of the CELP model, including linear prediction filter coefficients, adaptive and fixed codebooks' indices and gains. In addition to these parameters, high-band gain indices are computed in 23.85 kbps mode. These parameters are encoded and transmitted at the encoder. At the decoder, these parameters are decoded and the speech is synthesized by filtering the reconstructed excitation signal through linear prediction synthesis filter.

The parameters transmitted by the ITU-T G.722.2 codec include: Immitance Spectral Frequency (ISF), pitch, algebraic codebook indices, and the gains of adaptive and algebraic codebooks. ISF parameters are used to present the LPC synthesis filter, which describes the characteristics of spectral envelope. The pitch parameter, which consists of integer and fractional parts, describes the periodicity of speech signal. The algebraic codebook describes the stochastic characteristics of speech signal, and the codebook vector at different rates is constructed by placing a certain number of signed pulses in the tracks. Algebraic codebook indices present the pulse positions and signs in the codebook vector. Adaptive and algebraic codebook gains present the energy information of excitation signal, and they are quantized by vector quantization using 6 or 7 bits at different bit-rates.

3. The effect of noise on the codec parameters

Due to the introduction of noise, when the noisy speech is encoded by the speech codec, there will be significant difference between the parameters extracted from the noisy and clean speech samples. As a result, clarifying the noise's influence on the parameters of CELP model, and finding the proper way to modify the codec parameters are the key problems for compressed domain speech enhancement.

Among the codec parameters of CELP model, pitch lag is the most robust one to the noise. According to our experiments, except for the pitch doubling and halving phenomena occurred in a small number of frames, the pitch curve still remains smooth.

The algebraic codebook index is the most badly affected parameter by the noise. The pulse positions and signs of codebook vector searched from the noisy residual signal contain little information of clean speech.

As a result, the modification of pitch and algebraic codebook indices is not helpful for removing the noise and improving the quality of speech.

On the other hand, the effect of noise on spectral envelope of speech will be reflected on the ISF parameters,

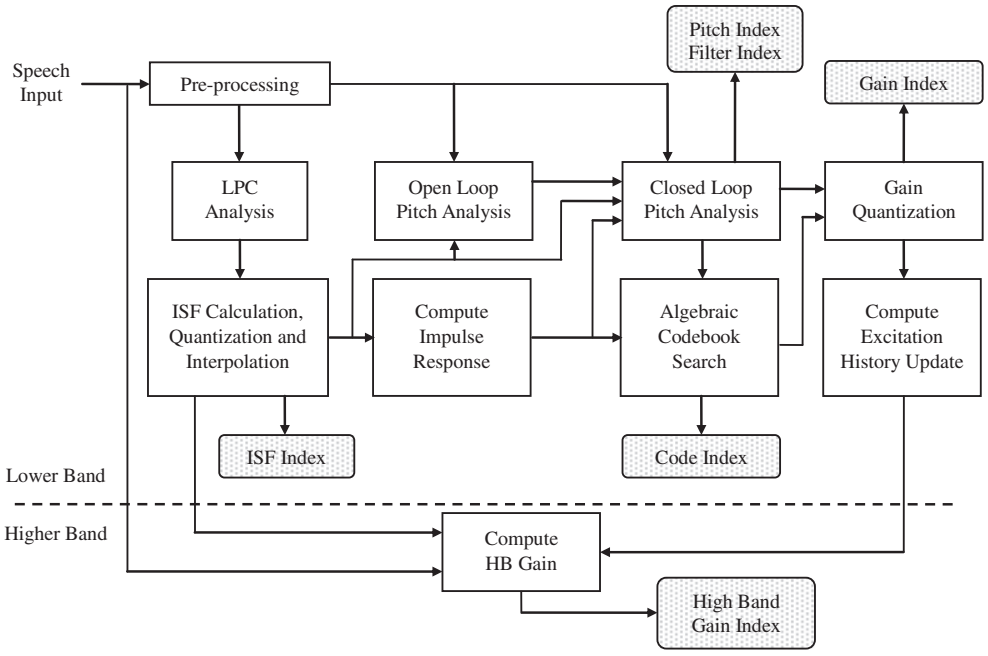

Fig. 2. Block diagram of ITU-T G.722.2 encoder.

and the effect on the signal level will be shown directly on the adaptive and algebraic codebook gains. So the proper modification of these two parameters could result in certain amount of noise reduction.

In this section, the effect of noise on the ISF and codebook gains will be analyzed to help the derivation of CD speech enhancement method. The noisy speech samples in white and car interior noise with an SNR of 6 dB are used in this analysis.

ISF is an equivalent representation of the LPC coefficient. In the codec of ITU-T G.722.2, there are 16 dimensions in ISF parameters. The first 15 dimensions are the normalized immitance frequencies, and the last one is the 16th LPC coefficient.

The statistical distributions of ISF parameters for clean and noisy speech under the car interior noise and white noise are shown in Fig. 3. As the last ISF parameter is the last LPC coefficient, only the distributions of the first 15 dimensions are illustrated here. The corresponding scatter diagrams of the first two dimensions of ISF parameters are illustrated in Fig. 4.

From the statistical histograms and the scatter diagrams illustrated in Figs. 3 and 4, it can be inferred that different kinds of noise have different effects on ISF parameters.

From the comparison of statistical histograms shown in Fig. 3(a) and (b), we can see that, under the car interior noise, the ISF distribution remains nearly unchanged in the high frequency regions, only the lowest two dimensions of ISF parameters are shifted towards lower frequency regions. As shown in Fig. 4(a) and (b), the main part of scatter diagram of ISF is not changed evidently. The difference is caused by the weak speech components which are easily contaminated by the noise signal. Generally, the car interior noise is a low-frequency distributed noise (the majority of noise power is concentrated in the low frequency region), and it has a small effect on the overall

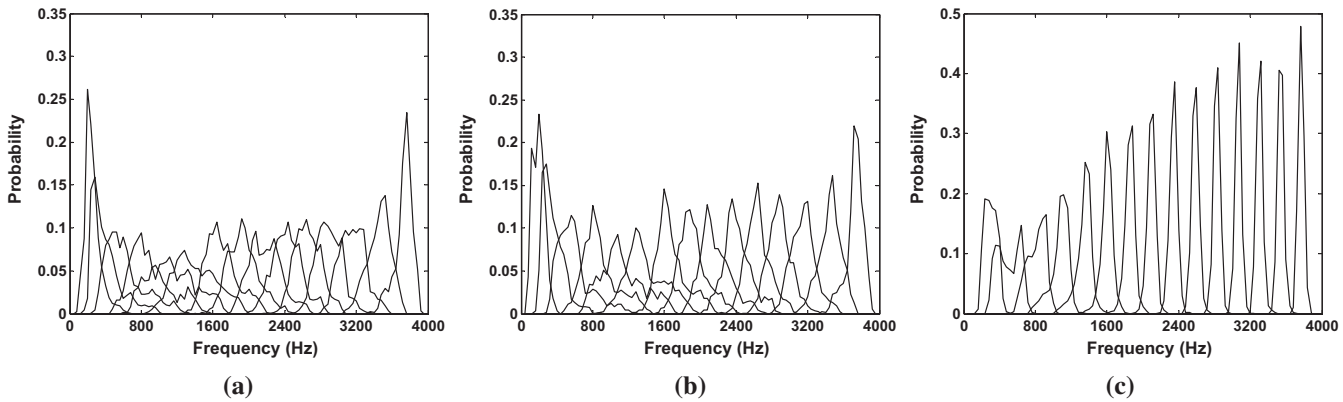

Fig. 3. The statistical histograms of ISF parameters. (a) clean speech; (b) noisy speech under the car interior noise; (c) noisy speech under the white noise.

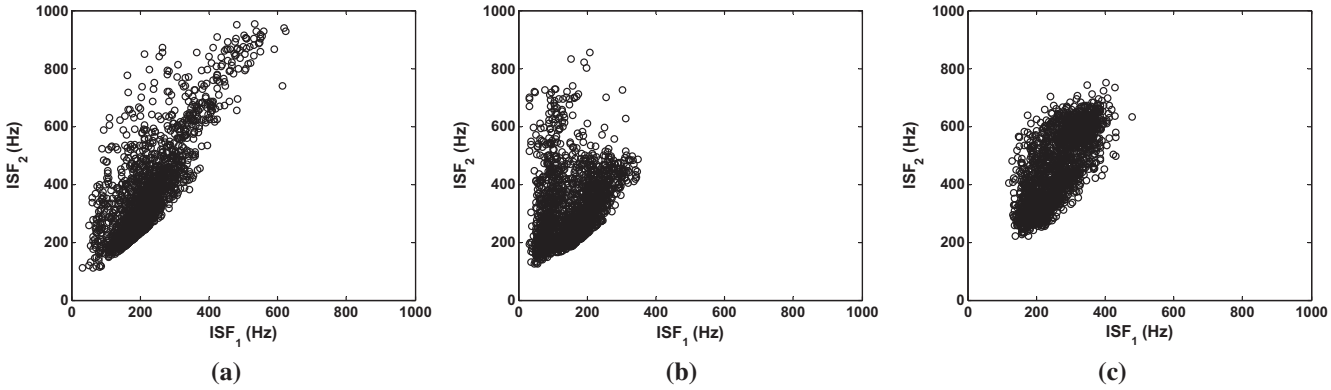

Fig. 4. The scatter diagrams of the first two dimensions of ISF parameters. (a) clean speech; (b) noisy speech under the car interior noise; (c) noisy speech under the white noise.

ISF distributions. Based on the above discussion, it is not necessary to modify the noisy ISF parameters in low-frequency distributed noise condition.

Similarly, comparing the statistical histograms illustrated in Fig. 3(a) and (c), it is obvious that, all 15 dimensions of ISF parameters are shifted towards high frequency regions. The variance of each dimension of ISF parameter gets smaller, which can also be observed from the scatter diagram shown in Fig. 4(c). In the high frequency regions with low SNR, the ISF distributions of noisy speech and noise signal are quite similar. Generally, since white noise is a full-band distributed noise, it has apparent effect on all the dimensions of ISF parameters. Based on these observations, we can infer that, the appropriate modification of ISF parameters may result in some noise reduction in full-band distributed noise.

From the above analysis, we can conclude that the full-band distributed noise signal has a significant effect on the ISF parameter of speech signal. However, there are only 16 coefficients in ISF parameters to represent the whole spectral envelope, so a slight modification of an individual ISF coefficient will have a magnificent effect on the overall spectral structure. Also, the fine tune of spectral envelope could not be achieved by the modification of ISF with only 16 dimensions. As a result, the modification of spectral envelope parameters is not considered in this paper.

The adaptive and algebraic codebook gains represent the amplitude information of excitation signal in CELP model. So after the speech signal is corrupted by noise, the change of signal amplitude will be directly reflected on these two parameters.

The statistical histograms of codebook gains for clean and noisy speech are shown in Fig. 5. The noise's effect on the codebook gains is analyzed as follows:

For the algebraic codebook gain, from Fig. 5(a)–(c), the statistical histograms are shifted towards larger values, and the offset is closely related to the noise type and intensity. For instance, in the same SNR condition, the low-frequency distributed noise like car interior noise has small effect on the codebook gain. While for the full-band distributed noise like white noise, whose energy is concentrated in the algebraic codebook excitation, there is a strong effect on the codebook gain.

From Fig. 5(d)–(f) we can see that, after the speech is contaminated by the noise, the value of adaptive codebook gain g_p is decreased in both the car interior noise and white noise. The white noise has a relatively larger effect on the adaptive codebook gain. But the pattern of its change is not clear.

From these observations, the algebraic codebook gain is more sensitive to the noise and the rule of its change is more straightforward. Based on this consideration, the algebraic codebook gain is modified first in this paper to remove the effect of noise.

4. The CD speech enhancement method in non-DTX mode

When the DTX mode is not adopted, the block diagram of the proposed CD speech enhancement method is illustrated in Fig. 6.

First the codec parameters of noisy speech are extracted from the input bit-stream by partial decoder I, including ISF, algebraic codebook gain g_c, adaptive codebook gain g_p, the corresponding excitation signals $c(n)$ and $d(n)$, and some assistant parameters like the voicing factor r_v.

The ISF parameters are used to calculate the low frequency power ratio (LFPR) of spectral envelope, which is used in the noise type classification. And the smoothed voicing factor is utilized in the compressed domain VAD method to classify the input sub-frame into two types, one for voiced speech, the other for unvoiced speech and background noise.

Based on the codebook gains and the excitation signals of noisy signal extracted from partial decoder I, and the excitation signals $c'(n)$ and $d'(n)$ derived from partial decoder II, considering the results of compressed domain VAD and noise type classification, the adaptive and algebraic codebook gains are joint modified. Then the adaptive comb filtering is used as a post-processing to reduce the residual noise between the harmonics in the voiced speech segments.

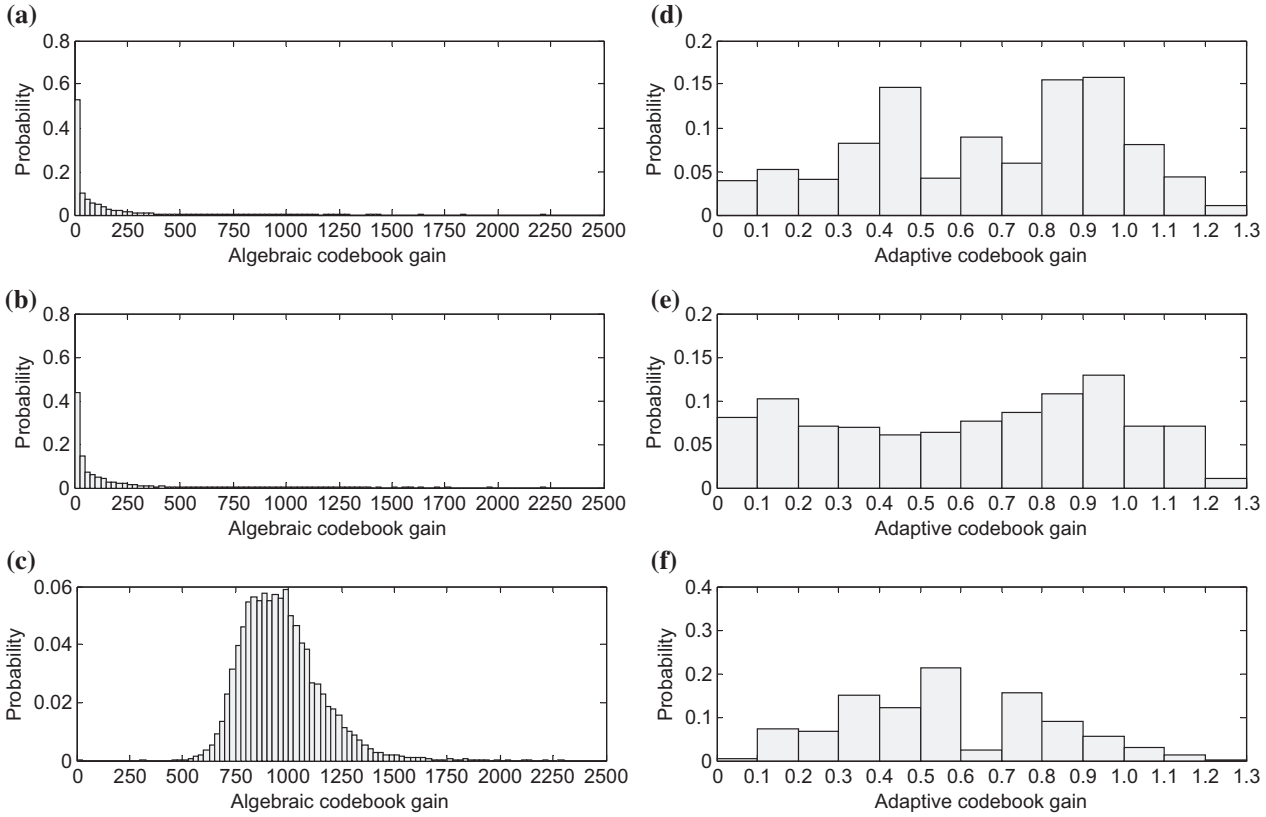

Fig. 5. Statistical histograms of the adaptive and algebraic codebook gains. Algebraic codebook gains for (a) clean speech, (b) car interior noise (6 dB) and (c) white noise (6 dB). Adaptive codebook gains for (d) clean speech, (e) car interior noise (6 dB) and (f) white noise (6 dB).

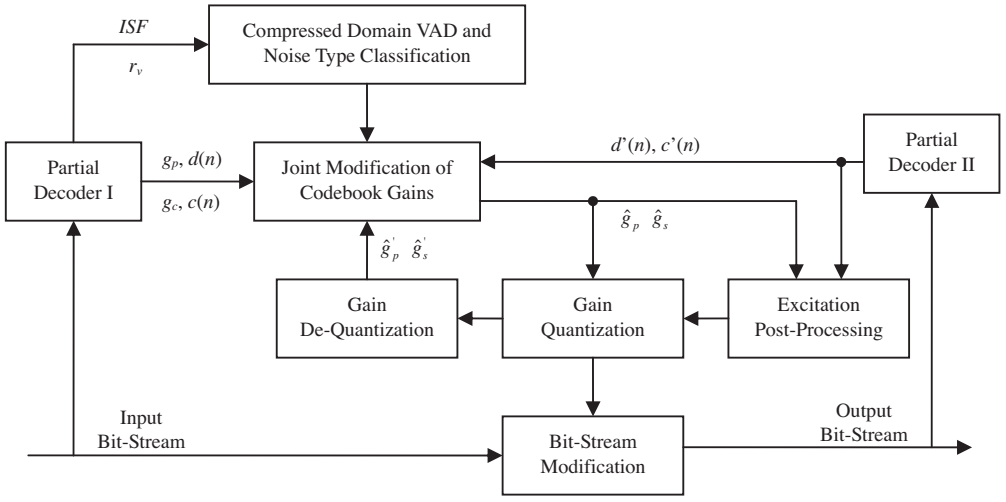

Fig. 6. Block diagram of the proposed compressed domain speech enhancement method.

Finally, the modified gain parameters are jointly quantized, and the quantized version of the codebook gains and the modified excitations are updated to partial decoder II.

The quantization index of the modified gain parameters is written back into the bit-stream to replace the corresponding parts of the input bit-stream, and we can get the output bit-stream at last.

In the next subsections, the compressed domain VAD and noise classification method are introduced first, then the joint modification method of codebook gains and the excitation post-processing method are described, and at last the gain re-quantization method is presented.

4.1. Compressed domain VAD method

The proposed compressed domain VAD is used to assist the joint modification procedure of codebook gains. The input noisy speech frame is classified into two categories,

one for voiced speech, and the other for the unvoiced speech and background noise. The VAD procedure is accomplished using the voicing factor parameter from partial decoder I. Since the linear domain VAD is not adopted, relatively accurate result can be achieved with much lower computational complexity.

The voicing factor $r_v(m)$ (m is the sub-frame index) is defined by the powers of adaptive and algebraic codebook vectors. The value of $r_v(m)$ is between -1 and 1. The value of $r_v(m)$ is related to the amount of voicing with a value of 1 for purely voiced segments and a value of -1 for purely unvoiced segments.

First the voicing factor is smoothed recursively along the time by the following relationship:

$$\bar{r}_v(m) = \alpha_v \bar{r}_v(m-1) + (1-\alpha_v) r_v(m) \quad (1)$$

where $\bar{r}_v(m)$ is the smoothed voicing factor, and $\alpha_v = 0.9$ is the smoothing factor.

Comparing $\bar{r}_v(m)$ with the predefined threshold T_v, if its value is greater than T_v, the current sub-frame will be classified as voiced speech, otherwise it will be classified as unvoiced speech or background noise. An example of compressed domain VAD is shown in Fig. 7. The noisy speech in white noise with an SNR of 12 dB is used. As illustrated in Fig. 7, using the threshold $T_v = -0.65$, we can get accurate VAD results.

The range of voicing factor varies with different kinds of noises, so the threshold should be adjusted according to the type of noise background. For low-frequency distributed noise, the threshold value should be larger, while smaller for the full-band distributed noises.

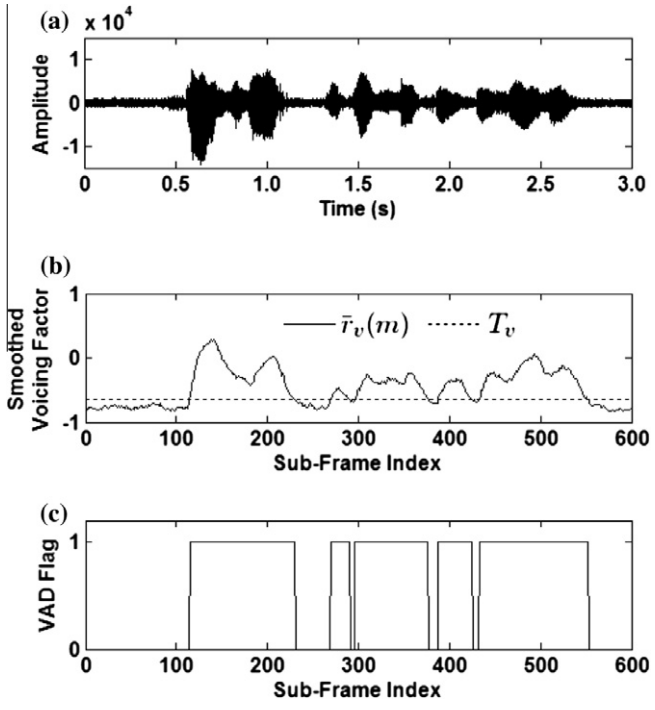

Fig. 7. An example of voicing factor based VAD method. (a) waveform of noisy speech; (b) smoothed voicing factor and the threshold; (c) VAD result.

4.2. Noise type classification

According to the analysis in Section 3, different kinds of noise have different effects on the parameters of CELP model. In order to get optimized performance, the appropriate modification methods should be utilized.

The noise type classification is used to classify the background noise into two categories, the first one is the full-band distributed noise like white noise, and the other one is the low-frequency distributed noise like car interior noise. The low frequency power ratio (LFPR) of the spectral envelope is used as a feature in the classification process.

In the speech segments with VAD = 0, the LPC spectral envelope is calculated by 256-point FFT and the power ratio of the lowest $N_{low} = 5$ frequency bins is obtained as:

$$R_{en\_low} = \frac{\sum_{i=0}^{N_{low}-1} E_{lpc}(i)}{\sum_{i=0}^{N_{FFT}/2} E_{lpc}(i)} \quad (2)$$

where N_{FFT} is the length of FFT, E_{lpc} is the LPC spectral envelope calculated from the noisy ISF parameters, and R_{en_low} is the LFPR of spectral envelope.

The power of the low-frequency distributed noise is concentrated in low frequency regions, which results in larger value of R_{en_low}. While the full-band distributed noise has a relatively flat power spectrum, which results in much smaller value of R_{en_low}. Calculating the long term average of R_{en_low} in several frames, the average value of R_{en_low} is compared with the threshold $T_R = 0.3$. If R_{en_low} is greater than the threshold, then the background noise is classified as low-frequency distributed noise, otherwise the speech is considered to be contaminated by the full-band distributed noise.

4.3. The joint modification method of codebook gains

The joint modification of codebook gains is the essential module of the proposed CD speech enhancement algorithm. The block diagram is shown in Fig. 8.

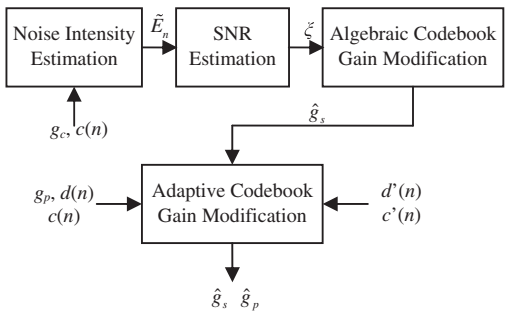

Fig. 8. Block diagram of joint modification of codebook gains.

Based on the noisy algebraic codebook gain g_c and the codebook vector $c(n)$, the power of excitation signal $\tilde{E}_n$ corresponding to the noise is estimated by MS method. Then the *a priori* SNR of the current sub-frame is adaptively estimated according to the result of noise type classification. Next the clean speech's algebraic codebook gain $\hat{g}_s$ is obtained by the rule of Wiener filtering. Finally, based on $\hat{g}_s$ and the excitation parameters from partial decoder I, together with the adaptive and algebraic codebook excitations ($d'(n)$ and $c'(n)$), the modified adaptive codebook gain is obtained under the rule of energy compensation.

4.3.1. Noise intensity estimation method

According to the analysis in Section 3, after the speech is contaminated by noise, the rule of algebraic codebook gain's change is more straightforward than the adaptive codebook gain. The statistical histogram is shifted towards larger values, and the offset is directly related to the noise intensity.

As a result, if we can get the local minimum value of noisy algebraic codebook gain g_c in the time domain by a certain method, then the relatively accurate estimation of noise excitation energy can be obtained by searching the minimum value of algebraic codebook power and applying appropriate bias compensation.

First, the MS algorithm described in Martin (1994) is used to estimate the algebraic codebook gain $\hat{g}_n$ corresponding to the noise (Taddei et al., 2004).

In order to deal with the fast fluctuation of g_c along the time, it is first smoothed by the first order recursive averaging as follows:

$$S(m) = \alpha_s S(m-1) + (1-\alpha_s) g_c^2(m) \tag{3}$$

where $S(m)$ is the smoothed value of g_c^2, $\alpha_s = 0.8$ is the smoothing factor, and m is the sub-frame index.

Then the minimum value of $S(m)$ is searched by MS method. And the algebraic codebook gain $\hat{g}_n(m)$ of noise is estimated by the following relationship:

$$\hat{g}_n(m) = B_{over} \cdot \sqrt{\min\{S(m)...S(m-D)\}} \tag{4}$$

where $D = 80$ is the window length of minimum search, and B_{over} is the over-estimation factor used for removing the bias introduced by minimum search.

The noise power in the total excitation signal is estimated as:

$$\hat{E}_n(m) = \hat{g}_n^2(m) \sum_n c_m^2(n) \tag{5}$$

In the codec of ITU-T G.722.2, the pulse positions in the algebraic codebook vector may overlap with each other. Consequently, the algebraic codebook energy $\sum_n c_m^2(n)$ without the gain information is not a constant. Then, the noise power estimation $\hat{E}_n(m)$ still has some fluctuations along the time.

To solve this problem, in the proposed method, another minimum search is performed on $\hat{E}_n(m)$ to get $\tilde{E}_n(m)$, which is the final estimation of noise excitation energy.

An example of noise power estimation is illustrated in Fig. 9, where E_{code} is the algebraic codebook energy of noisy speech. We can see that, $\hat{E}_n(m)$ has many over-estimation phenomena occurred in the speech segments. After the second term of minimum search, the final estimation $\tilde{E}_n(m)$ appears to be much more stationary.

4.3.2. SNR estimation method

The noise power in excitation signal is used in this section to calculate the *a posteriori* SNR and *a priori* SNR, which are further utilized in the computation of modification factor for the algebraic codebook gains.

First, two types of the *a posteriori* SNR estimation methods are proposed in this paper, which are defined as follows:

$$\gamma_{exc}(m) = \frac{E_{fcb\_before}(m)}{\tilde{E}_n(m)} \tag{6}$$

$$\gamma_{subframe}(m) = \frac{E_{subframe}(m)}{E_{subframe\_min}} \tag{7}$$

where $E_{fcb\_before}(m)$ is the algebraic codebook power before speech enhancement, $E_{subframe}(m)$ is the speech power in the mth sub-frame, $E_{subframe\_min}$ is the local minimum of sub-frame speech power. In order to get more stationary results, the length of minimum search for $E_{subframe}$ is set to 180 sub-frames, which is much longer than the one used for g_c.

The results of SNR estimation under the car interior noise and white noise are shown in Figs. 10 and 11, respectively. We can see that, the two kinds of the *a posteriori*

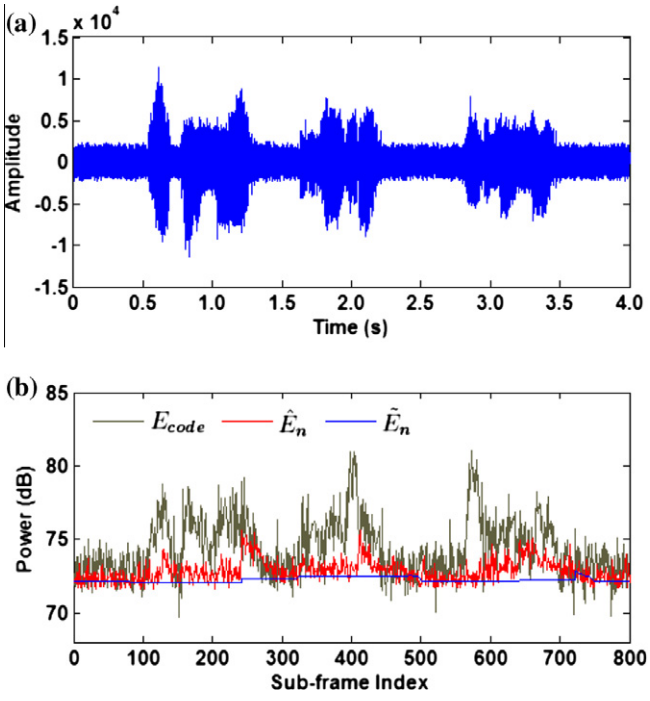

Fig. 9. An example of noise power estimation. (a) waveform of noisy speech; (b) algebraic codebook power and the noise power estimation.

SNR estimation have different characteristics. $\gamma_{exc}(m)$, which is based on the algebraic codebook power, is more accurate in the unvoiced segments, and performs well in the full-band distributed noise like white noise (as shown in Fig. 11(b)). While in the low-frequency distributed noise conditions, SNR under-estimation is likely to take place in the voiced segments (like the weak speech components around the 440th sub-frame in Fig. 10(b)). On the other hand, $\gamma_{subframe}(m)$, which is based on the speech sub-frame power, is suitable for various kinds of noise conditions, and could get reliable estimation results in both unvoiced and voiced segments (as shown in Fig. 10(c) and Fig. 11(c)). However, over-estimation often occurs for this form of estimation, so the direct use of $\gamma_{subframe}(m)$ will result in audible and annoying residual noise in the enhanced speech.

According to the above discussion, an adaptive SNR estimation method based on the result of noise type classification is proposed in this paper. The two kinds of the *a posteriori* SNR estimation are combined under different rules in full-band distributed noise and low-frequency distributed noise.

In the low-frequency distributed noise conditions, two forms of the *a posteriori* SNR are combined using the relationship in Eq. (8)

$$\xi(m)=\begin{cases}\max(\beta\xi(m-1)+(1-\beta)\max(0.5\gamma_{exc}(m)+0.5\gamma_{subframe}(m)-1,0),\xi_{\min}), & \bar{r}_v(m)>-0.5\\ \max(\beta\xi(m-1)+(1-\beta)\max(\gamma_{exc}(m)-1,0),\xi_{\min}), & \textit{Otherwise}\end{cases} \tag{8}$$

In Eq. (8), $\bar{r}_v(m)$ is the smoothed voicing factor in the mth sub-frame, β is the smoothing factor in SNR estimation, and $\xi_{\min}$ is the minimum value of the *a priori* SNR estimation.

From Fig. 10, we can see that, in the frame with larger voicing factor, the combination of $\gamma_{exc}(m)$ and $\gamma_{subframe}(m)$ by weighted average can prevent the *a priori* SNR from being under-estimated in weak speech segments. While in the frame with smaller voicing factor, the direct use of $\gamma_{exc}(m)$ could achieve large amount of noise reduction.

In full-band distributed noise condition, *a Posteriori* SNR Controlled Recursive Averaging (PCRA) method is proposed in this paper. The *a posteriori* SNR is used to estimate the speech presence probability in the current sub-frame, which is utilized to control the updating rate of the *a priori* SNR.

The *a posteriori* SNR $\gamma_{subframe}(m)$ based on speech sub-frame energy is smoothed first using the following relationship:

$$\bar{\gamma}_{subframe}(m)=\alpha_\gamma\bar{\gamma}_{subframe}(m-1)+(1-\alpha_\gamma)\gamma_{subframe}(m) \tag{9}$$

where $\alpha_\gamma=0.8$ is the smoothing factor of the *a posteriori* SNR.

Comparing $\bar{\gamma}_{subframe}(m)$ with a predefined threshold, if it is larger than the threshold, the speech presence flag $I(m)$ is set to one, otherwise, $I(m)=0$.

Then the speech presence probability is calculated as:

$$p(m)=\alpha_p p(m-1)+(1-\alpha_p)I(m) \tag{10}$$

where $\alpha_p=0.8$ is the smoothing factor.

Next, the smoothing factor β for the *a priori* SNR is determined using the speech presence probability:

$$\beta=\beta_{\min}+(\beta_{\max}-\beta_{\min})(1-p(m)) \tag{11}$$

where $\beta_{\max}=0.9$ and $\beta_{\min}=0.8$ are the maximum and minimum values of the smoothing factor, respectively.

Finally, the *a priori* SNR is estimated as:

$$\xi(m)=\max(\beta\xi(m-1)+(1-\beta)\max(\gamma_{exc}(m)-1,0),\xi_{\min}) \tag{12}$$

For lower bit-rate modes of codec, due to the reduction of pulse numbers in the algebraic codebook vector, it is less sufficient in describing the non-stationary components of speech signal. So the *a posteriori* SNR estimation $\gamma_{exc}(m)$ based on excitation power is lack of accuracy. To settle the problem of under-estimation in speech segments, a small amount of $\gamma_{subframe}(m)$ is compensated into the *a priori* SNR estimation.

From Fig. 11(b), it is obvious that there are many fluctuations for the excitation energy based *a posteriori* SNR γ_{exc}. But as shown in Fig. 11(d), by using sub-frame energy based *a posteriori* SNR $\gamma_{subframe}$ to control the update rate of the *a priori* SNR, the more accurate estimation results are obtained with less fluctuations in noise periods.

According to the decision-directed approach (Ephraim and Malah, 1984), after the process of speech enhancement, the *a priori* SNR should be updated for the next sub-frame, i.e.,

$$\xi(m)=\frac{E_{fcb\_after}(m)}{\tilde{E}_n(m)} \tag{13}$$

where $E_{fcb\_after}(m)$ is the algebraic codebook energy after the process of speech enhancement.

4.3.3. The modification of algebraic codebook gain

The algebraic codebook gain of the enhanced speech $\hat{g}_s$ is obtained by multiplying gain g_c of noisy speech with a modification factor $G_{gc}(m)$:

$$\hat{g}_s(m)=G_{gc}(m)g_c(m) \tag{14}$$

where $G_{gc}(m)$ is a modification factor with the form of Wiener filtering, which is defined by the *a priori* SNR as follows:

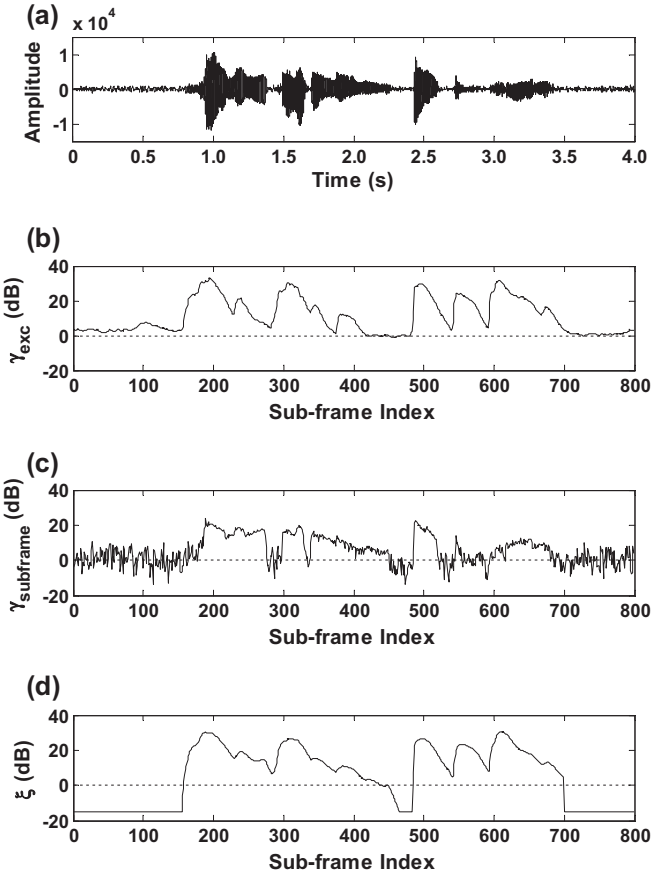

Fig. 10. SNR estimation under the car interior noise. (a) waveform of noisy speech; (b) γ_{exc}; (c) $\gamma_{subframe}$; (d) the *a priori* SNR estimation.

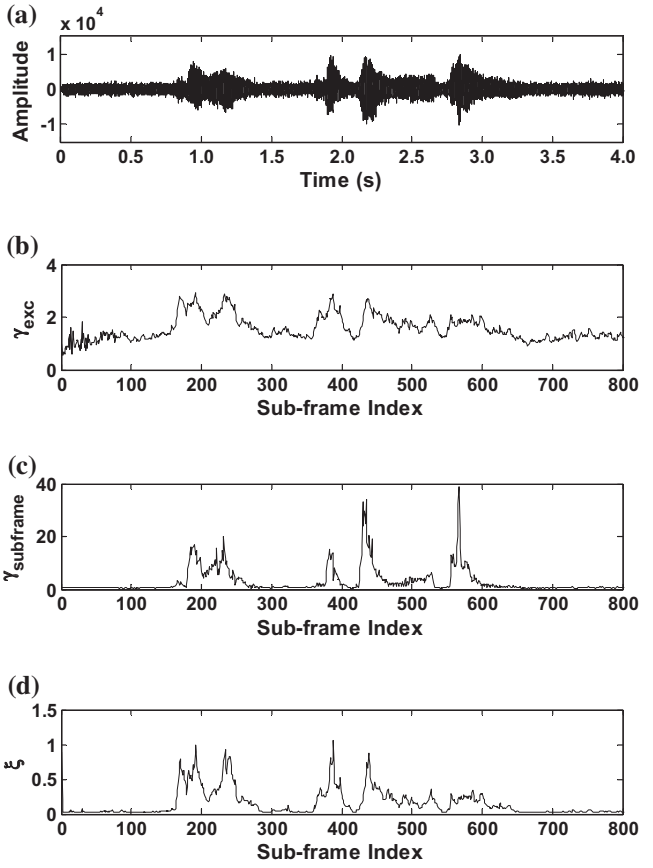

Fig. 11. SNR estimation under the white noise. (a) waveform of noisy speech; (b) γ_{exc}; (c) $\gamma_{subframe}$; (d) the *a priori* SNR estimation.

$$G_{gc}(m) = \frac{\xi(m)}{1+\xi(m)} \tag{15}$$

The value of modification factor $G_{gc}(m)$ is between 0 and 1. When the *a priori* SNR $\xi(m)$ is large, the attenuation applied on g_c becomes relatively small to prevent speech distortions. On the other hand, a heavier attenuation will be applied to reduce the noise effectively.

4.3.4. The modification of adaptive codebook gain

Most of the compressed domain speech enhancement methods, like the one described in Taddei et al. (2004), are focused on the modification of algebraic codebook gain. These methods could achieve large amount of noise reduction, but there is a severe loss in the level of speech components. The reason is that, in the CELP model shown in Fig. 1, the adaptive codebook excitation is closely related to the algebraic codebook vector. The reduction of algebraic codebook gain may remove some noise components, while the power of the total excitation signal is attenuated at the same time. This kind of effect will be reflected back to the adaptive codebook excitation through the long-term predictor. If no modification is performed on the adaptive codebook gain, it will finally result in the loss of speech power.

Since the power loss of speech components usually occurs in the voiced segments, we can use different methods to obtain the modified adaptive codebook gain according to the result of compressed domain VAD.

If the input sub-frame is classified as unvoiced speech or background noise, the adaptive codebook gain is kept unchanged to avoid the additional amplification of noise components. Otherwise, the adaptive codebook gain needs to be compensated. The compensation rule is based on keeping the power of modified excitation signal equal to the scaled version of the noisy one, which can be expressed as:

$$E_{after}(m) = \alpha_e(m) E_{before}(m) \tag{16}$$

where E_{before} and E_{after} are the powers of the total excitation signal before and after the modification of codebook gains, $\alpha_e(m)$ is the power scaling factor of the mth sub-frame. E_{before} and E_{after} are defined as:

$$E_{before}(m) = \sum_n \left(g_p(m) d_m(n) + g_c(m) c_m(n)\right)^2 \tag{17}$$

$$E_{after}(m) = \sum_n \left(\hat{g}_p(m) d'_m(n) + \hat{g}_s(m) c'_m(n)\right)^2 \tag{18}$$

In the proposed method, $\alpha_e(m)$ is equal to the scaling factor $G_{gc}(m)$ for the algebraic codebook gain. Then $\alpha_e(m)E_{before}(m)$ can be viewed as an approximate estimation of clean excitation power. In comparison with the method given in Sukkar et al. (2006), there is no need to get the scaling factor from LD speech enhancement method anymore.

By rewriting Eq. (16) with the adaptive codebook excitation, algebraic codebook excitation and the corresponding gain parameters, we can get the following relationship (for the sake of brevity, the sub-frame index m is omitted):

$$\sum_n (\hat{g}_p d'(n) + \hat{g}_s c'(n))^2 = \alpha_e \sum_n (g_p d(n) + g_c c(n))^2 \tag{19}$$

Then we can get:

$$E_a \hat{g}_p^2 + 2\hat{g}_s E_c \hat{g}_p + \hat{g}_s^2 E_u - \alpha_e E_{before} = 0 \tag{20}$$

where

$$E_a = \sum_n (d'(n))^2 \quad E_u = \sum_n (c'(n))^2$$
$$E_c = \sum_n d'(n)c'(n) \tag{21}$$

E_a and E_u are the powers of adaptive and algebraic codebook vectors from partial decoder II, respectively. E_c is the inner product of two codebook vectors.

Eq. (20) can be viewed as a quadratic function with respect to $\hat{g}_p$. By solving for the roots of Eq. (20), we can get the modified adaptive codebook gain. If there are two real and positive roots, the larger one will be used. In some rare cases, there is no real root for the function, and then no modification will be applied on g_p. In other words, we will set $\hat{g}_p = g_p$. Also, the modified adaptive codebook gain is constrained under the maximum value in the gain quantization codebook.

4.4. Post-processing of excitation signal

In the proposed method, the comb filtering is used as a post-processing method on the total excitation signal.

Comb filtering is crucial to improve speech quality in low-frequency distributed noise. It is helpful for reducing the noise in extreme low frequency, removing the residual noise between harmonics and reconstructing some of the harmonic structures. As a result, the comb filtering is adopted in the low-frequency distributed noise condition.

First, the total excitation signal is reconstructed using the modified codebook gains, that is,

$$u_m(n) = \hat{g}_p(m)d'_m(n) + \hat{g}_s(m)c'_m(n) \tag{22}$$

where $\hat{g}_p(m)$ and $\hat{g}_s(m)$ are the modified adaptive and algebraic codebook gains, respectively, $d'_m(n)$ and $c'_m(n)$ are the adaptive and algebraic codebook vectors derived from partial decoder II, respectively, and m is the sub-frame index.

The adaptive comb filtering used in this paper has the following form:

$$H_c(z) = G_c \frac{1 + az^{-T}}{1 - bz^{-T}} \tag{23}$$

where T is the integer pitch lag in the current sub-frame, which can be extracted from the input bit-stream, a and b are the filter coefficients, $G_c = (1 - b)/(1 + a)$ is a scaling factor to avoid undesired amplification or attenuation of the excitation signal.

The filter coefficients, a and b, control the shape of frequency response and the relative amount of attenuation between spectral peak and valley. When the value of a or b increases, a heavier attenuation will be applied.

In order to control the inter-frame effect of comb filter, the value of b is set to zero in our method. For the purpose of improving the performance of comb filter, the value of a is selected adaptively according to the SNR condition and the value of voicing factor as follows:

a. When the voicing factor and SNR are all small, this condition corresponds to the noise periods. The value of a is set to zero, and no filtering operation is used to avoid the additional harmonic components.
b. When the voicing factor is large and the SNR is small, this condition corresponds to the weak voiced segments which are badly corrupted by the noise. Then, a large value of a is used to apply intensive filtering.
c. When there are large values for both the voicing factor and SNR, this condition corresponds to strong voiced speech components. Then, a smaller value of a is adopted, and a smaller amount of filtering is applied to avoid additional attenuation to the speech.

The filtered total excitation signal is expressed as:

$$\hat{u}_m(n) = h_c(n)^* u_m(n) \tag{24}$$

where $h_c(n)$ is the impulse response of comb filter.

The short-time spectrum of the excitation signal for a voiced speech sub-frame, and the corresponding frequency response of the adaptive comb filter are shown in Fig. 12

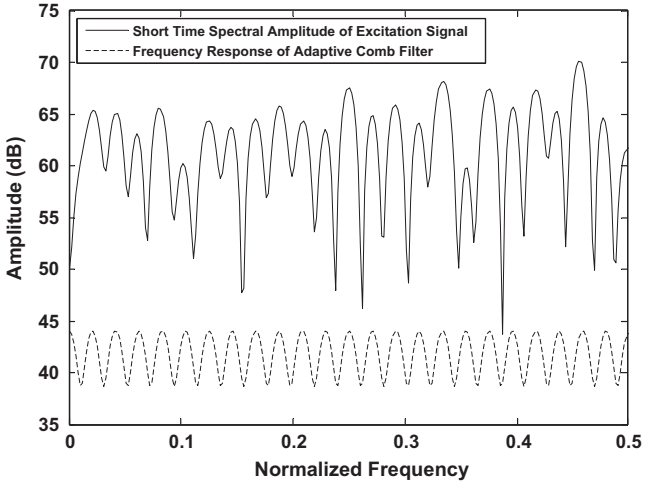

Fig. 12. Short-time spectrum of excitation signal and the corresponding frequency response of comb filter.

(For convenience, the curve of frequency response is shifted up). In Fig. 12, with $a = 0.3, b = 0$, the frequency response of the comb filter is well fitted to the harmonic structure of excitation signal, and it will be helpful for reducing residual noise between the harmonic components.

The fine structure of the excitation signal in partial decoder II could not be modified directly by the compressed domain method. To solve this problem, we need the assistance of gain quantization to reflect the effect of comb filtering to the destination decoder. Under the minimum mean square error criterion on the excitation signal or on the reconstructed speech signal, the total excitation signal in the destination decoder will get close to the output of comb filter to some extent.

4.5. Re-quantization of the codebook gains

In order to get the output bit-stream, it is necessary to develop an efficient re-quantization method for the codebook gains. In this paper, different gain quantization criterions are used in different kinds of noise conditions. The MMSE criterion on the excitation signal is adopted for the full-band distributed noise, which is referred to as quantization rule 1. For the low-frequency distributed noise, the MMSE criterion on the reconstructed speech is utilized, and referred to as quantization rule 2.

In quantization rule 1, the gain codebook is searched by minimizing the mean square error of the total excitation signal. The quantization error is expressed as follows (for the sake of brevity, the sub-frame index m is omitted):

$$E = \sum_n \left[x_e(n) - (\hat{g}^k_{p\_q} d'(n) + \hat{g}^k_{s\_q} c'(n)) \right]^2 \tag{25}$$

where $\hat{g}^k_{p\_q}$ and $\hat{g}^k_{s\_q}$ are the adaptive and algebraic codebook gains corresponding to the kth codeword in the gain codebook, respectively. $x_e(n)$ is the target vector in the excitation domain, and can be expressed as:

$$x_e(n) = \hat{g}_p d'(n) + \hat{g}_s c'(n) \tag{26}$$

For quantization rule 2, the gain codebook is searched by minimizing the mean square error of the synthesized speech. The quantization error is calculated as follows:

$$E = \sum_n [x(n) - (\hat{g}^k_{p\_q} y(n) + \hat{g}^k_{s\_q} z(n))]^2 \tag{27}$$

where $x(n)$ is the target vector in the speech domain, which is expressed as a convolution of the pulse response $h(n)$ of LPC synthesis filter with the total excitation signal $\hat{u}_m(n)$ after comb filtering:

$$x(n) = h(n)^* \hat{u}_m(n) \tag{28}$$

$y(n)$ and $z(n)$ are the output signals that the adaptive and algebraic codebook excitations pass through LPC synthesis filter, which can be expressed as follows, respectively,

$$\begin{aligned} y(n) &= h(n)^* d'(n) \\ z(n) &= h(n)^* c'(n) \end{aligned} \tag{29}$$

By comparing the aforementioned two quantization rules, we can find that the rule 2 is similar to the one used in the gain quantization of CELP model, and it has high quantization accuracy. However, since the adaptive, algebraic, and total excitation signals have to be filtered through the LPC synthesis filter, the process is relatively complex. On the other hand, the accuracy of quantization rule 1 is slightly lower than rule 2, but it is still acceptable for our application, and the complexity is much lower.

According to the above discussion, in the proposed method, rule 2 is adopted in the low-frequency distributed noise condition to achieve high speech quality, while in the full-band distributed noise, rule 1 is utilized to get a compromise between speech quality and computational complexity.

5. Compressed domain speech enhancement method in DTX mode

The hangover scheme of DTX mode (ITU-T, 2002a,b) proposed in ITU-T G.722.2 codec is shown in Fig. 13.

In the DTX mode, by the assistance of VAD, when the speech segment is ended (the VAD result turns from 1 to 0), a hangover period of 7 frames is activated. The frame type in the hangover period is still set to SPEECH (S). The first frame after the hangover period is referred to as the First Silence Insertion Descriptor (SID_FIRST, F). The noise segment starts from the SID_FIRST frame. The noise information is transmitted every few frames (in SID_UPDATE frames), while in the other frames, no data is transmitted (called NO_DATA frames). The first SID_UPDATE frame is the third frame after the SID_FIRST frame, after this, the SID_UPDATE frames are transmitted every 8th frame.

The Comfort Noise Generation (CNG) parameters transmitted in the SID_UPDATE frames include:

(1) The weighted averaged ISF parameter vector f_{mean}: the weighted average of the ISF parameters of the eight most recent frames, which represents the spectral envelope of noise background.
(2) The averaged logarithmic frame energy en^{mean}_{log}: the average of the logarithmic energy of the eight most recent frames, which represents the intensity of noise background.

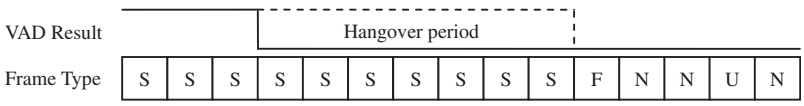

Fig. 13. Hangover scheme of DTX mode in ITU-T G.722.2.

From the characteristics of DTX mode, the CD speech enhancement method in DTX mode is only necessary in the SID_UPDATE frames, while the method described in Section 4 can be used in the SPEECH frames.

In order to keep the characteristics of the noise spectral envelope, the ISF parameters are not modified in the proposed algorithm, and only the logarithmic frame energy in the noise frame is attenuated. In order to get stationary residual noise background, the modification factor of log frame energy is set to be equal to the minimum attenuation factor of g_c in speech frames, which can be expressed as:

$$G_{DTX} = G_{\min} = \frac{\xi_{\min}}{1 + \xi_{\min}} \tag{30}$$

where $\xi_{\min} = -15dB$ is the minimum value of the *a priori* SNR estimation.

The modified log frame energy is re-quantized and written back to the bit-stream.

6. Compressed domain speech enhancement method when frame erasure occurs

When frame erasure occurs, the basic idea of CD speech enhancement method is to take advantage of the FEC module (ITU-T, 2002c) in ITU-T G.722.2 codec to recover all the codec parameters, modify part of them to remove the speech distortion introduced by the frame erasure, then re-quantize these parameters and write them back to the bit-stream. The block diagram is shown in Fig. 14.

The averaged algebraic codebook gain is modified first, while the adaptive codebook gain is kept unchanged. Then, using the reconstructed algebraic codebook vector and the adaptive codebook vector from partial decoder II, the gain parameters are re-quantized. Finally, the quantized gains, the total excitation signal and the algebraic codebook vector are updated into partial decoder II.

The basic principle of algebraic codebook gain modification is to increase the amount of attenuation with an exponential rule from the first erased frame. There are two parameters needed from the previous good frames, including the averaged algebraic codebook gain $\bar{g}_c$ of the recent four sub-frames, and the modification factor G_{gc_old} of the previous good sub-frame.

The number of successive erased sub-frames is denoted as L_{FEC}. Then the algebraic codebook gain is modified as:

$$\hat{g}_s = \bar{g}_c G_{FEC}^{L_{FEC}} G_{gc\_old} \tag{31}$$

where $G_{FEC} = 0.9$ is an exponential attenuation factor.

From Eq. (31), we can see that, the modification factor in the current sub-frame is related to the one in the previous good frame, and the amount of attenuation will increase when successive frame erasure occurs.

The recovered algebraic codebook vector from the FEC module in G.722.2 decoder is composed of 64 random numbers in the range of $[-1, 1]$. It is not in line with the basic structure of the algebraic codebook in the standard. As a result, the algebraic codebook vector should be reconstructed before quantized.

The reconstruction of algebraic codebook vector is carried out as follows:

a. Generate certain number of random positions and signs for the pulses according to the current coding mode;
b. Encode the pulse positions and signs and write back to the bit-stream;
c. Reconstruct the algebraic codebook vector for gain quantization and the memory update of synthesis filter.

The gain quantization method when frame erasure occurs is the same as described in sub-Section 4.5.

7. Performance evaluation

The performance evaluation in this paper includes five aspects: the test under ITU-T G.160 (ITU-T, 2008), the subjective speech quality test, the computational complexity test, and the performance tests for DTX mode and frame erasure condition.

In all the tests, the clean speech sequences are chosen from NTT database. The additive noise signals are selected from ITU noise database and NoiseX-92 database (Varga

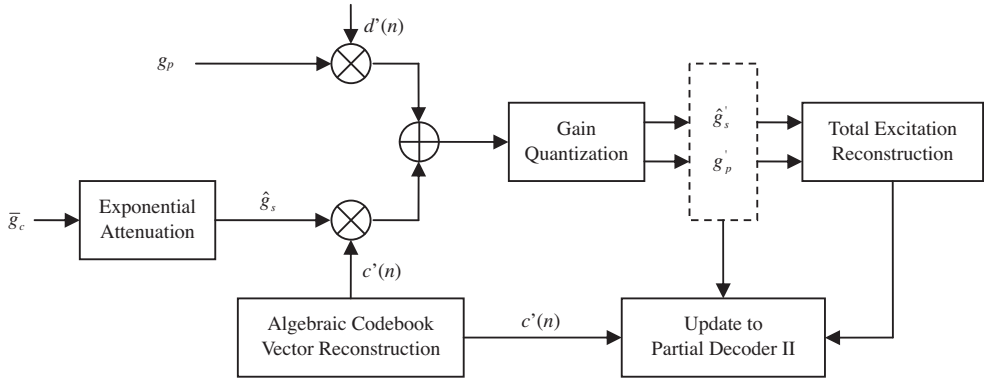

Fig. 14. Block diagram of the compressed domain speech enhancement method when frame erasure occurs.

and Steeneken, 1993). The sampling rate of noise signals is adjusted to 16 kHz before the test.

Since the proposed method is operated on the encoded bit-stream of noisy speech, the reference signals in the tests, including the clean and noisy speech, should go through the process of encoding and decoding to remove the effect of quality degradation and possible level change which are introduced by speech codec.

The production of test signals is shown in Fig. 15.

Since there are 9 codec modes in ITU-T G.722.2 codec, for the sake of brevity, the 9 modes are denoted as Mode 0 to Mode 8 with the increasing coding rates, i.e. the 6.6 kbps mode is referred to as Mode 0, and the 23.85 kbps mode is denoted as Mode 8.

7.1. G.160 test

ITU-T G.160 is a performance test standard used for Voice Enhancement Devices (VED) in digital network-based equipment. The purpose of this test is to evaluate the performance of speech enhancement in terms of the amount of noise reduction and SNR improvement, the convergence time and the objective speech quality.

In this paper, the reference algorithms in G.160 test are the one only modifies the algebraic codebook gain (Taddei et al., 2004) (referred to as Ref1) and the one with the assistance of LD speech enhancement (Sukkar et al., 2006) (referred to as Ref2).

The noise reduction test in white noise is used to ensure that the noise reduction method could provide specified level of noise reduction, and the level change of speech components remains in the acceptable range.

Q_m is the specified level of noise reduction which is determined by the noise reduction test in purely white noise. There are three parameters in this test, including Q_{n1}, Q_{n2} and Q_s. Q_{n1} and Q_{n2} are the noise reduction factors in the noise periods in the front and the end of the test sequence. Q_s is the level difference of speech components before and after speech enhancement. If the values of Q_{n1} and Q_{n2} are in the range of $Q_m \pm 3$ dB, and the value of Q_s is between -3 dB and 2 dB, then the method under test fulfills the request of G.160.

The test results are summarized in Table 1. From the test results, the performance of Ref1 and Ref2, including the parameters of Q_m, Q_{n1} and Q_{n2}, does not change a lot in different codec modes. While for the proposed CD enhancement method, the amount of noise reduction under the white noise gets smaller with the decrease of coding rates. The main reason is that, as described in Section 4.3.2, due to the reduction of pulse numbers in algebraic codebook in lower coding rates, it is less sufficient to describe the stochastic components of speech signal. As a result, the *a priori* SNR based on excitation power is over-estimated in noise segments, which results in lower amount of noise reduction.

The expected amount of noise reduction Q_m of the proposed method reaches 29 dB in Mode 8, while reduces to 19 dB in Mode 0. In all the codec modes, Q_m of the proposed method is much larger than that of Ref1. And the proposed method could provide larger amount of noise reduction in the higher 8 codec modes than Ref2, while in Mode 0 it is slightly smaller than Ref2. On the other hand, the noise reduction factors for the proposed method, Q_{n1} and Q_{n2}, can reach the requirement of G.160. And the

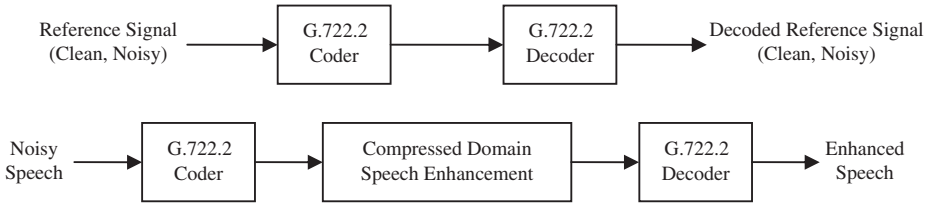

Fig. 15. The production of test signals.

Table 1
The results of noise reduction test under the white noise.

Codec mode	Ref1				Ref2				Proposed			
	Q_m (dB)	Q_{n1} (dB)	Q_{n2} (dB)	Q_s (dB)	Q_m (dB)	Q_{n1} (dB)	Q_{n2} (dB)	Q_s (dB)	Q_m (dB)	Q_{n1} (dB)	Q_{n2} (dB)	Q_s (dB)
0	8.65	8.82	8.85	5.76	20.32	20.47	20.30	0.65	19.11	21.41	21.52	0.50
1	8.73	8.88	8.85	7.28	20.32	20.63	20.35	0.84	21.43	22.30	21.44	0.27
2	8.59	8.66	8.63	6.63	20.51	20.59	20.30	0.81	23.14	22.74	21.69	−0.29
3	8.59	8.65	8.65	6.94	20.47	20.61	20.14	0.90	24.94	24.08	24.55	−0.42
4	8.56	8.70	8.67	6.94	20.53	20.82	20.48	0.87	26.04	27.04	27.12	−0.33
5	8.57	8.63	8.75	6.76	20.59	20.77	20.43	0.91	28.40	29.16	28.88	−0.74
6	8.57	8.63	8.70	6.63	20.57	20.79	20.67	0.90	28.46	29.25	29.74	−0.67
7	8.58	8.77	8.72	6.53	20.66	20.83	20.62	0.89	29.53	30.11	30.14	−1.02
8	8.56	8.69	8.69	6.47	20.67	20.81	20.78	0.90	29.11	29.79	29.88	−0.49

effect on the speech level, which is reflected by Q_s, is much smaller than two reference methods.

Since Ref1 only modifies the algebraic codebook gain, it is impossible to achieve heavy noise reduction, and Q_s is much larger due to the severe level loss of speech components. In Ref2, the modification factor of codebook gains is calculated by the assistance of LD enhancement method. As a result, its noise reduction performance depends on the adopted LD method. In our experiment, Weighted Euclidean Distortion Measure (WEDM) estimator (Loizou, 2005) is used. On the contrary, the proposed method solves the problem of speech level loss by the joint modification of codebook gains, while achieves large amount of noise reduction without the use of LD enhancement methods, and the effect on speech components is much lower.

The noise reduction test in colored noise is designed to measure the ability of noise reduction and SNR improvement, and the effect on speech level in colored noise. There are three test parameters, including Signal-to-Noise Ratio Improvement (SNRI), Total Noise Level Reduction (TNLR), and SNRI to NPLR Difference (DSN). Here, NPLR refers to Noise Power Level Reduction.

According to the requirement of ITU-T G.160, if the value of SNRI is larger than 4 dB, TNLR is less than −5 dB, and the value of DSN is between −4 dB and 3 dB, then the speech enhancement method under test meets the requirement of G.160. The larger SNRI, smaller TNLR, and DSN that is close to zero correspond to better speech enhancement performance.

This test is carried out under the street and factory noise, with the SNR of 6 dB, 12 dB and 18 dB, respectively. And the results are averaged over all test conditions. The test results are shown in Table 2.

From the test results listed in Table 2, the SNRI of Ref1 is below 4 dB, the TNLR is around −8 dB, and DSN is around −3 dB. This could not fulfill the requirement of G.160. For Ref2, the value of SNRI is around 10 dB, TNLR is between −16 dB and −18 dB, and the DSN is between −1.5 dB and −2 dB. On the other hand, for the proposed method, the performance drops when the coding rate decreases. The SNRI parameter could meet the requirement of G.160, the value of TNLR is between –10 dB and –18 dB, and DSN is below 0.4 dB.

In all the codec modes, the proposed method could provide greater SNR improvement than Ref1, the amount of noise reduction is much heavier, and the effect on the speech level is much lower.

The noise reduction ability of Ref2 is similar to the LD method used. The proposed method is a complete CD method, and it is an intrinsic drawback that the CD method is not capable for the colored noise. So it is reasonable that the performance is slightly lower than Ref2. From the result in Table 2, the SNRI for the proposed method is slightly smaller than that of Ref2, and the difference is between 2 dB and 4 dB. In the higher four modes, the TNLR of the proposed method is similar to the one of Ref2. In Mode 2, 3 and 4, the TNLR of the proposed method is slightly smaller, the difference is around 2 dB. In the lowest two modes, the difference of TNLR between the proposed method and Ref2 is about 5 dB. Meanwhile, the absolute value of DSN for the proposed method is much smaller, which means there are lower distortions on the speech level.

The convergence test is used to ensure that the speech enhancement method could provide expected amount of noise reduction in response to the sudden change of noise power after a maximum allowed convergence time.

The convergence time is defined as the time from the change of noise power to the instant when the amount of noise reduction is in the range of $Q_m \pm 3$ dB. This test is carried out under the white noise. There are three step changes of noise level during the test, and the corresponding convergence times are denoted as T_1, T_2 and T_3, respectively. If the convergence time is within 3 s, the method under test can meet the demand of G.160 standard.

The results of convergence test are summarized in Table 3.

From the results in Table 3, the convergence time of the proposed method under the white noise is within 2 s, which can meet the requirement of G.160. In comparison with Ref1, as more complex noise estimation method is adopted to get more stationary results, the convergence time of the proposed method is longer. The convergence time of Ref2 is determined by the LD method adopted, and it is slightly shorter than the proposed method.

Table 2
The results of noise reduction test under the colored noise.

Codec mode	Ref1			Ref2			Proposed		
	SNRI (dB)	TNLR (dB)	DSN (dB)	SNRI (dB)	TNLR (dB)	DSN (dB)	SNRI (dB)	TNLR (dB)	DSN (dB)
0	1.46	−7.49	−2.97	8.13	−16.32	−1.63	3.74	−10.75	0.05
1	2.13	−7.83	−2.92	9.02	−16.96	−1.79	4.13	−11.81	0.28
2	3.21	−8.08	−2.86	9.70	−17.70	−1.91	5.72	−15.26	0.40
3	3.23	−8.10	−2.89	9.79	−17.91	−1.93	5.95	−15.66	0.31
4	3.43	−8.12	−2.90	10.05	−18.05	−1.93	6.01	−16.08	0.36
5	3.66	−8.23	−2.98	10.25	−18.26	−1.99	6.74	−17.14	0.33
6	3.46	−8.24	−3.06	10.13	−18.30	−2.02	6.75	−17.45	0.28
7	3.86	−8.28	−2.98	10.51	−18.43	−2.03	7.48	−18.29	0.32
8	3.85	−8.27	−3.03	10.37	−18.42	−2.03	8.05	−18.05	0.26

Table 3
The results of convergence test.

Codec mode	Ref1			Ref2			Proposed		
	T_1 (s)	T_2 (s)	T_3 (s)	T_1 (s)	T_2 (s)	T_3 (s)	T_1 (s)	T_2 (s)	T_3 (s)
0	0.62	0.62	0	1.33	1.16	0	1.78	1.44	0
1	0.62	0.64	0	1.42	0.99	0	1.88	1.53	0
2	0.62	0.65	0	1.37	1.03	0	1.73	1.45	0
3	0.61	0.63	0	1.4	1.25	0	1.76	1.48	0
4	0.62	0.64	0	1.36	1.11	0	1.74	1.44	0
5	0.62	0.63	0	1.48	1.25	0	1.75	1.46	0
6	0.62	0.65	0	1.35	1.14	0	1.74	1.47	0
7	0.62	0.64	0	1.51	1.05	0	1.74	1.45	0
8	0.61	0.65	0	1.31	1.28	0	1.73	1.45	0

The objective speech quality test is used to measure the quality improvement produced by speech enhancement method. The test method is not specified in the standard of G.160. Perceptual Evaluation of Speech Quality (PESQ) (ITU-T, 2001) is used in this paper.

This test is carried out under the ITU noise database (four noise types, including babble, office, etc.), and the NoiseX-92 noise database (12 noise types, including F16, factory, etc.). Three SNR conditions (6 dB, 12 dB and 18 dB) are used in this test.

The PESQ scores of the noisy speech, the enhanced speech produced by the proposed and reference methods in 16 noise conditions and 9 codec modes are illustrated in Figs. 16–19.

From the test results, in the noise conditions of Buccaneer1, Buccaneer2, Destroyer engine, F16, Factory2,

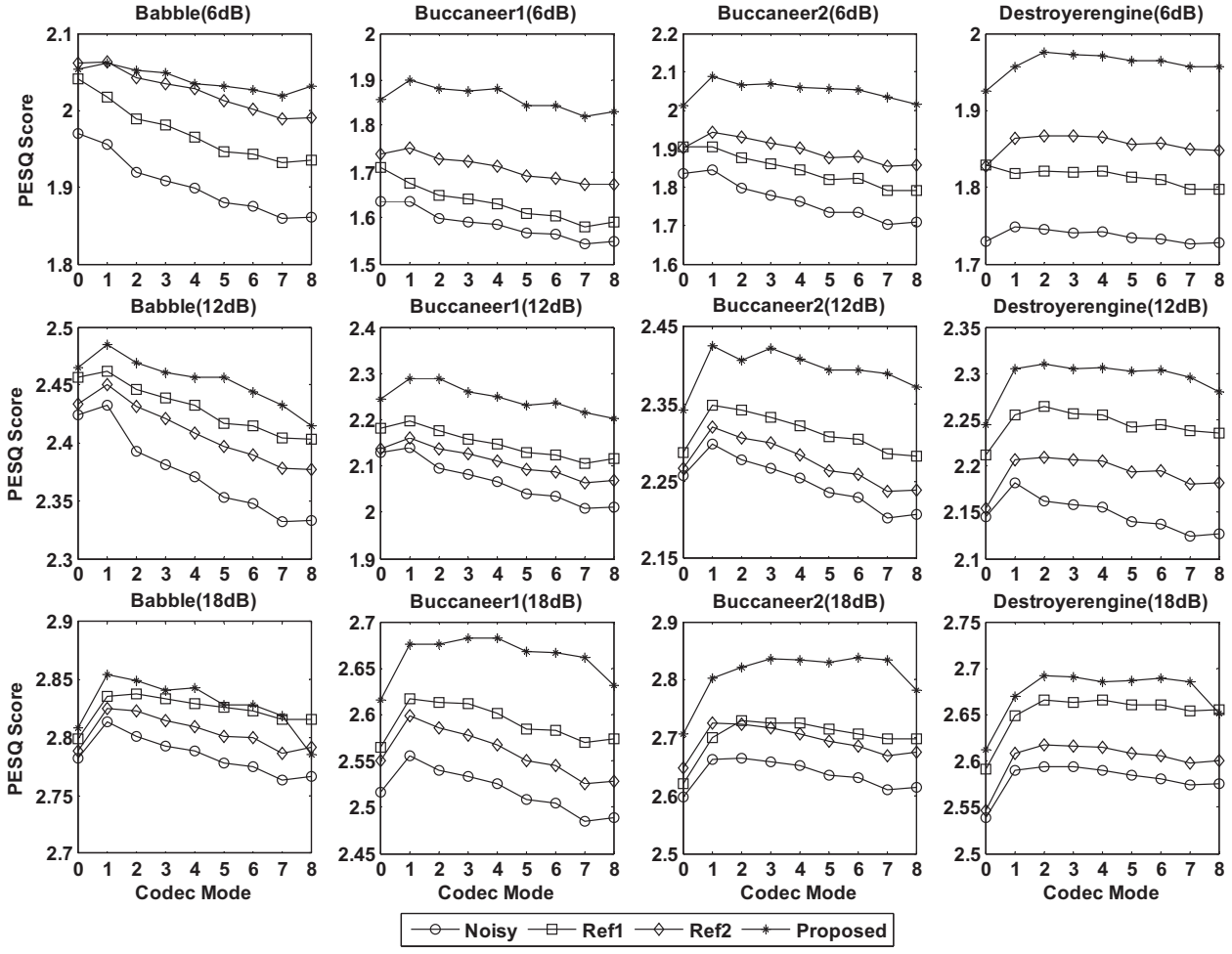

Fig. 16. The results of objective speech quality test in Babble, Buccaneer1, Buccaneer2 and Destroyerengine noise conditions.

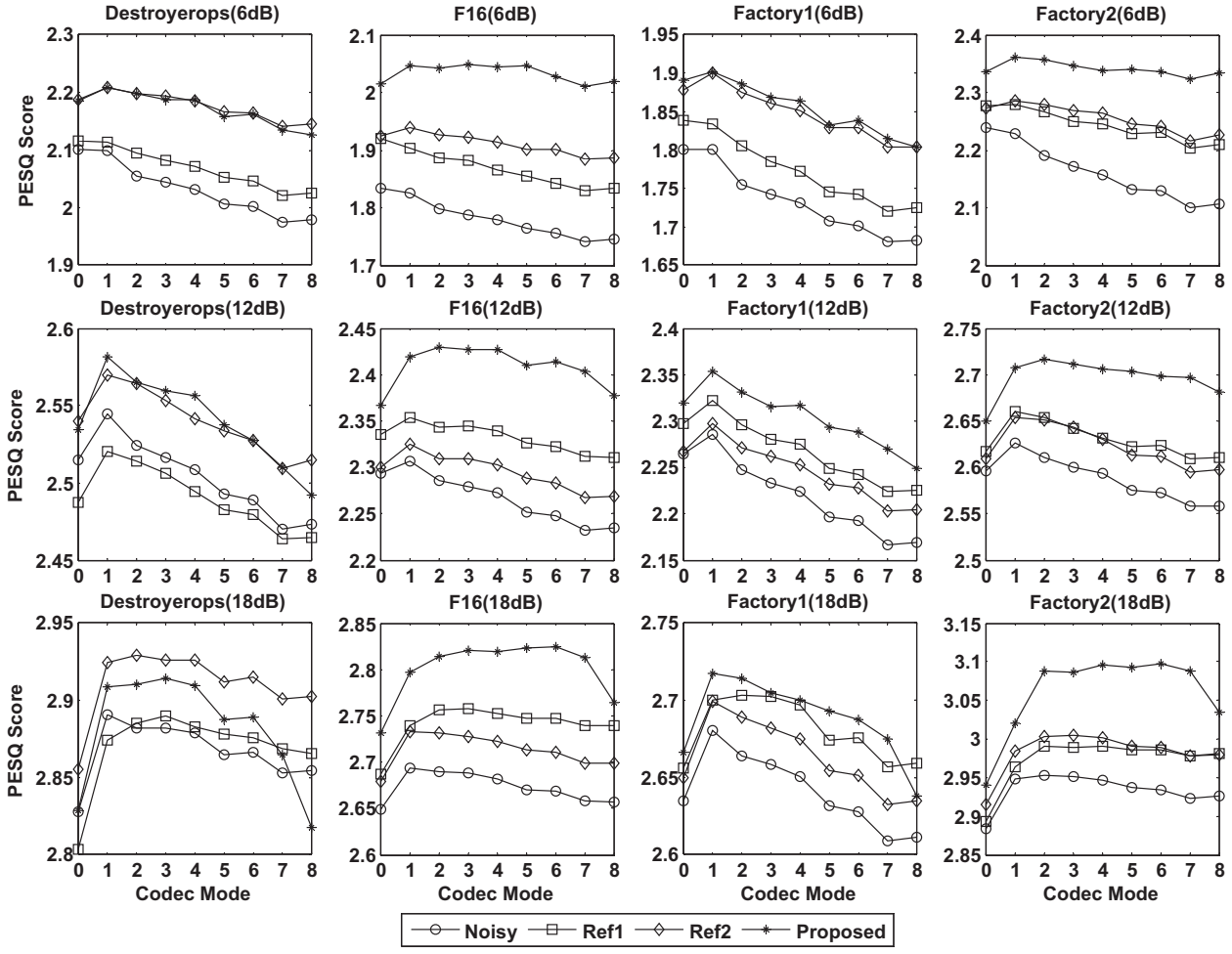

Fig. 17. The results of objective speech quality test in Destroyerops, F16, Factory1 and Factory2 noise conditions.

Hfchannel, Pink and White, in comparison with Ref1 and Ref2, the average PESQ scores of the proposed method are improved by about 0.05, and in some conditions the improvement is greater than 0.1. From the spectrum characteristics of noise background, these noise types belong to the full-band distributed noise, and the noise power spectrum is stationary along the time. In this kind of noise condition, the objective speech quality of the proposed method is much better than the reference methods.

In the noise conditions including Babble, Factory1 and Tank, the PESQ improvement of the proposed method comparing with Ref1 and Ref2 is within 0.05. These kinds of noise belong to the full-band distributed noise, and there are some non-stationary components existed in the noise spectrum. The proposed method performs slightly better than the reference methods.

In the noise conditions like Destroyerops and Office, the PESQ scores of the proposed method are slightly lower than the reference methods. These kinds of noises belong to the full-band distributed noise with strong non-stationary components like transient and speech-like components. The performance of the proposed method remains to be improved in this kind of noise conditions.

In the noise conditions like Leopard, Street and Volvo, except for the SNR conditions of 6 dB and 12 dB in Leopard noise, and the SNR condition of 12 dB in Street noise, the PESQ scores of the proposed method are similar to Ref1 and Ref2. These kind of noise types belong to the low-frequency distributed noise. The state-of-art CD speech enhancement methods are not very effective in this kind of noise condition. The PESQ improvement is not significant in this condition.

Generally, though Ref1 could improve the objective speech quality to some extent, the subjective quality of speech will not be improved evidently due to the severe loss of speech power. By the assistance of LD speech enhancement method, Ref2 could get stationary residual noise in the noise period, but the noise reduction in speech period is not sufficient and results in strong and annoying residual noise. Comparing with the reference methods, the proposed method can suppress the noise efficiently in noise period, while some of the noise is removed in the speech

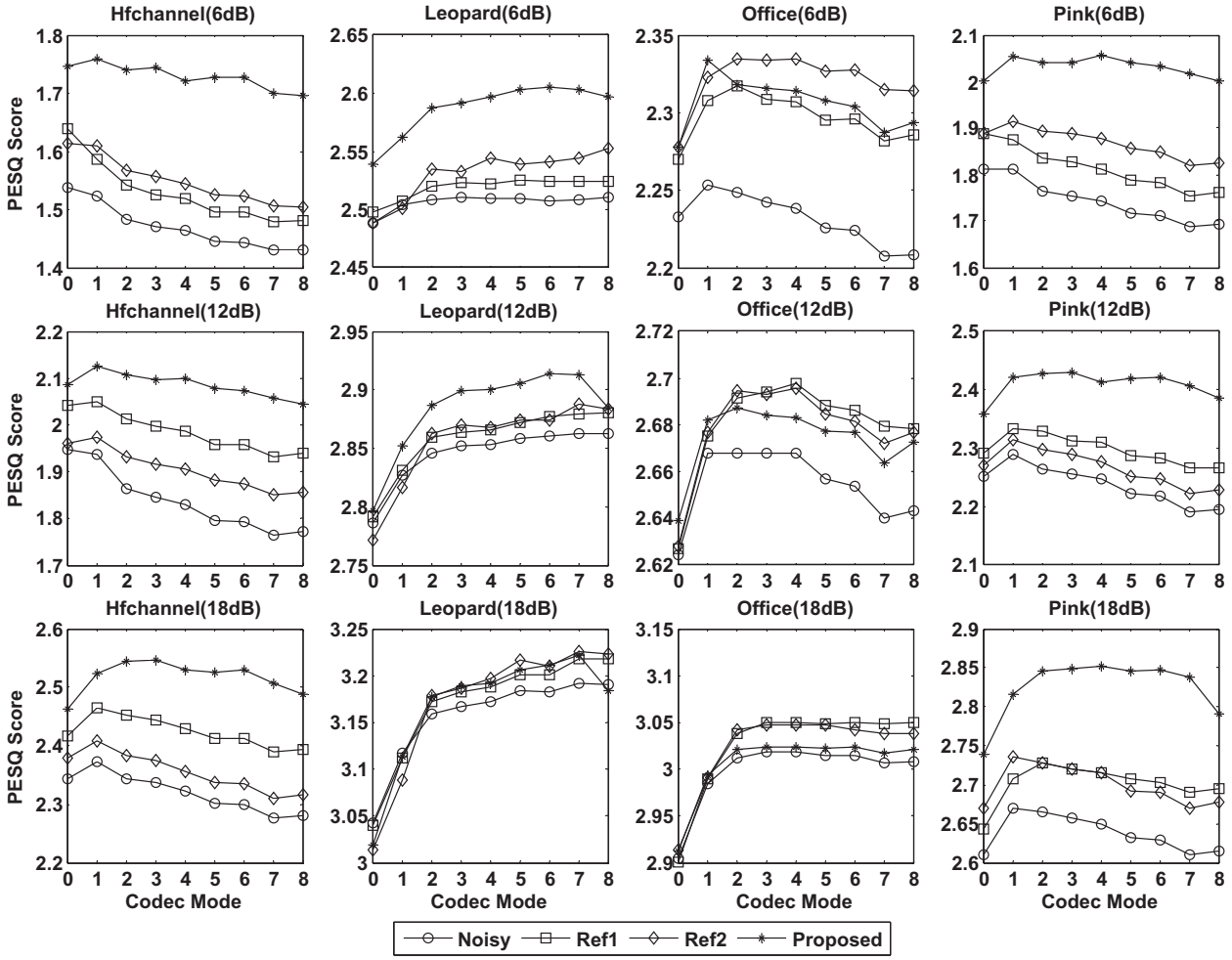

Fig. 18. The results of objective speech quality test in Hfchannel, Leopard, Office and Pink noise conditions.

period, which results in much better objective speech quality.

7.2. The subjective speech quality test

The purpose of this test is to assess the subjective quality of enhanced speech by the proposed method.

The reference methods are also the one that only modifies the algebraic codebook gains (Taddei et al., 2004) and the one jointly modifies the codebook gains by the assistance of LD speech enhancement method (Sukkar et al., 2006).

Six types of noise, including Babble, F16, Factory, Volvo, White, and street noise, are used in this test. The first five noise signals are chosen from NoiseX-92 database, and the street noise is selected from ITU noise database. The 48 clean speech samples are taken from NTT database.

In this test, the SNR conditions of 6 dB, 12 dB and 18 dB are used, respectively. There are 4 speech samples in each SNR condition, which results in 72 test sequences totally. Every test sequence consists of one noisy speech segment and two enhanced speech segments from the reference and test methods, respectively. Each speech segment is 3 s long. The enhanced speech segments by the proposed and reference method are placed randomly in the test sequence. During the test, the participants will listen to the noisy speech first, and then select the enhanced speech segment with better overall speech quality. Totally, 8 listeners (4 males and 4 females) participated in this test.

The results of the subjective speech quality test are summarized in Tables 4 and 5.

From the result listed in Table 4, comparing with Ref1, the subjective speech quality of the proposed method in Babble, F16 and White noise is much better. In Factory and Street noise, the quality of the proposed method is slightly better. While in Volvo noise, the reference and proposed methods have comparable speech quality. Generally speaking, the proposed method has a better speech quality than Ref1.

From the result given in Table 5, comparing with Ref2, the subjective speech quality of the proposed method is slightly worse in Babble, F16, Factory and Street noise conditions. But in the other two noise conditions, the percentage of preference is larger for the proposed method. In general, the preference difference of the proposed method and Ref2 is within 5%, and the percentage of

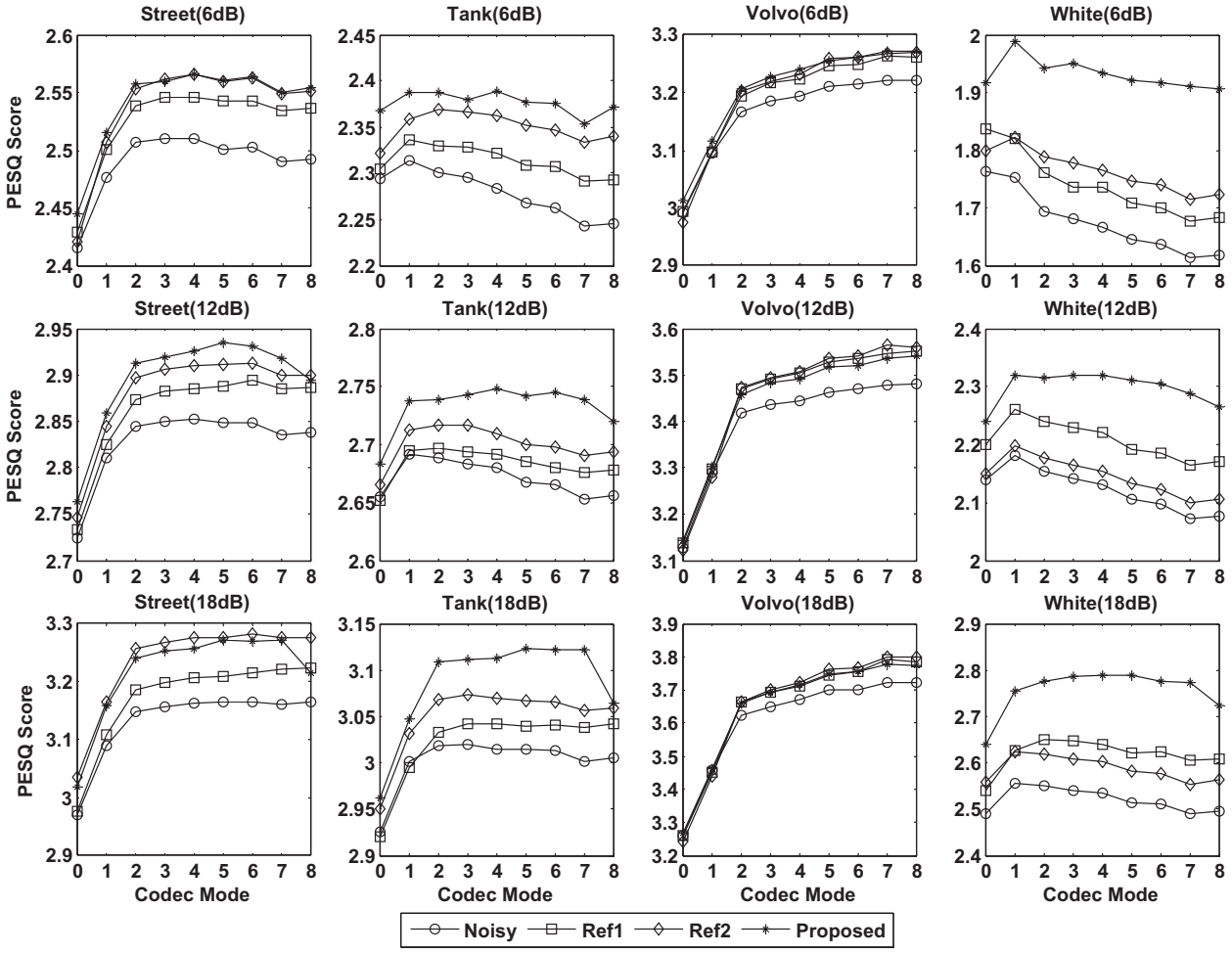

Fig. 19. The results of objective speech quality test in Street, Tank, Volvo and White noise conditions.

Table 4
The A/B test results comparing with Ref1.

	Babble (%)	F16 (%)	Factory (%)	Street (%)	Volvo (%)	White (%)
Prefer Ref1	25.00	34.38	29.17	21.88	20.83	35.42
Prefer the proposed method	44.79	44.79	31.25	26.04	19.79	45.83
No preference	30.21	20.83	39.58	52.08	59.38	18.75

Table 5
The A/B test results comparing with Ref2.

	Babble (%)	F16 (%)	Factory (%)	Street (%)	Volvo (%)	White (%)
Prefer Ref2	35.42	30.21	30.21	21.88	15.63	32.29
Prefer the proposed method	33.33	28.13	26.04	19.79	17.71	34.38
No preference	31.25	41.67	43.75	58.33	66.67	33.33

No-Preference is the highest in most of the test noise conditions. So there is no significant difference between the proposed method and Ref2 in term of subjective speech quality.

Ref1 only modifies the algebraic codebook gain, which results in severe loss of speech level in full-band distributed noise, and finally leads to the degradation of subjective speech quality. Ref2 removes little noise from the speech segments, but with the assistance of LD enhancement method, the residual noise in noise periods is more stationary. This is an advantage over the proposed method, which makes the subjective quality of these two methods unable to be distinguished.

In conclusion of the above discussion, the subjective speech quality of the proposed method is better than Ref1, and similar to Ref2.

Table 6
The test results of computational complexity.

Codec mode	Ref1		Ref2		Proposed	
	Avg. (WMOPS)	WorstCase (WMOPS)	Avg. (WMOPS)	WorstCase (WMOPS)	Avg. (WMOPS)	WorstCase (WMOPS)
0	7.577	7.639	16.329	16.634	6.527	7.532
1	7.466	7.591	15.738	16.233	6.474	7.48
2	7.532	7.678	15.377	15.475	6.527	7.566
3	7.553	7.707	15.401	15.473	6.552	7.593
4	7.573	7.728	15.425	15.524	6.576	7.616
5	7.612	7.761	15.462	15.561	6.614	7.655
6	7.64	7.785	15.483	15.582	6.638	7.674
7	7.686	7.839	15.538	15.636	6.693	7.727
8	7.677	7.845	16.13	16.243	6.683	7.731
Average	7.591	7.730	15.654	15.818	6.587	7.619

7.3. The computational complexity test

In our research, the proposed and reference methods are realized using the fixed-point C language, and the computational complexity is calculated by the tools in STL2005 under the standard of ITU-T G.191 (ITU-T, 2005).

The test material is composed of the noisy speech in Babble, Street, Volvo, Factory and White noise with the SNR of 6 dB and the total length of about 10 min. The computational complexity includes two aspects, the average complexity and the worst case complexity.

The computational complexity of the proposed method and the reference methods are summarized in Table 6.

From the test results, the averaged computational complexity of the proposed method is 1.004 WMOPS and 9.067 WMOPS smaller than Ref1 and Ref2, respectively. And it is only 42.1% of Ref2. The worst case complexity of the proposed method is similar to Ref1, and it is only 48.2% of Ref2.

Ref1 has a similar structure to the proposed method, and its complexity is concentrated in the re-quantization of gain parameters. It is necessary for Ref2 to perform full decoding on the input bit-stream, and it needs the assistance of LD enhancement method, so the computational complexity is much higher than the other two methods. The proposed method is more complex than Ref1 in algorithm, but the average complexity is much lower by using adaptive gain quantization rules.

7.4. Performance test in DTX mode

As described in Section 5, when DTX function is adopted in ITU-T G.722.2 speech codec, compressed domain speech enhancement method in non-DTX mode is used in speech frames, while the method for DTX mode is utilized in noise segments. As a result, the difference between the performances of non-DTX and DTX modes are focused on the noise segments.

The same performance tests, as described in the previous sub-sections, are performed on the proposed CD speech enhancement method in DTX mode. In comparison with the method in non-DTX mode, similar results are obtained in the colored noise reduction test, the convergence test, and the speech quality improvement test. Meanwhile, the noise reduction performance in white noise is improved in DTX mode, which results from the heavy noise attenuation in noise segments.

Since DTX mode is only activated in high SNR conditions, a noisy speech sample in white noise with an SNR of 30 dB is used as an example. The spectrograms of the clean speech, noisy speech and enhanced speech are shown in Fig. 20(a)–(c), respectively.

From the spectrograms shown in Fig. 20, we can see that the speech components are well preserved by the non-DTX mode speech enhancement method in speech segments. On

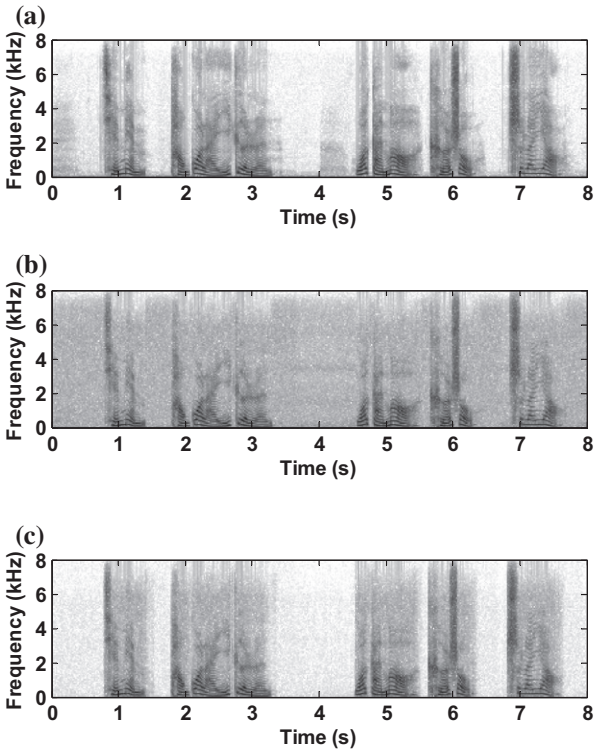

Fig. 20. Spectrogram comparison in DTX mode. (a) clean speech; (b) decoded noisy speech; (c) enhanced speech.

the other hand, by the assistance of DTX mode, the noise is removed efficiently in noise segments, and the residual noise remains stationary.

7.5. Performance test in frame erasure condition

When frame erasure occurs, the compressed domain speech enhancement focuses on the recovery of lost speech frames, and tries to remove the artifacts introduced by the FEC procedure.

From the results of performance test in the condition of 3% frame error, the amount of noise reduction under the white and colored noise, and the convergence time are equivalent to the performance when no frame erasure occurs. Meanwhile, the speech quality is relatively lower when frame error occurs. The reason is that, the effect of frame erasure is mainly reflected on the lost frame and the several following frames. If the frame error happens in the speech period, the speech quality is very likely to be degraded. On the other hand, the overall noise reduction and noise tracking ability will not be affected evidently.

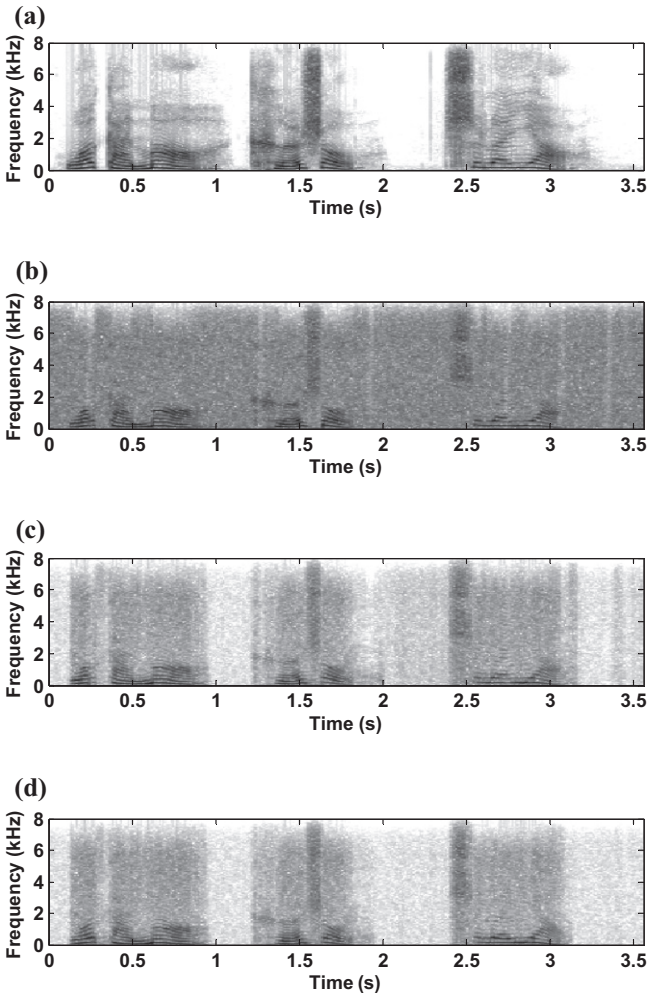

Fig. 21. Spectrum comparison when frame erasure occurs. (a) clean speech; (b) decoded noisy speech; (c) enhanced speech; (d) enhanced speech when no frame erasure occurs.

An example of noisy speech in white noise with 3% frame error is used in this paper. The spectrograms of clean speech, noisy speech, enhanced speech with and without frame erasure are shown in Fig. 21(a)–(d), respectively.

In Fig. 21(b), the light vertical lines indicate the positions of frame erasure. By comparing the spectrograms in Fig. 21(c) and (d), we can conclude that, frame erasure occurred in noise segments will bring relatively more residual noise in the enhanced speech, while most of the discontinuity and artifacts in speech segments are removed by the proposed method.

8. Conclusions

In order to realize efficient speech enhancement features in the network equipment of mobile communication system, a compressed domain speech enhancement method based on the modification of codec parameters is proposed based on ITU-T G.722.2 codec. This method can operate in all the codec modes of G.722.2, and is compatible with the DTX mode and the situation when frame erasure occurs. In non-DTX modes, the compressed domain VAD and noise type classification are performed first. Then, based on the algebraic codebook power, the noise intensity is estimated, and the *a priori* SNR is estimated by an adaptive method based on the noise type. Next, the adaptive and algebraic codebook gains are jointly modified. Especially, for the low-frequency distributed noise, the residual noise between the harmonic components is removed by the comb filter. Finally, the modified codebook gains are re-quantized and written back to the bit-stream. In DTX mode, the spectral envelope of the noise frame is kept unchanged, and the log frame energy is attenuated to remove the effect of noise. When frame erasure occurs, the codec parameters are recovered by the FEC module in the decoder, then the algebraic codebook vector is reconstructed, and the algebraic codebook gain is exponentially attenuated.

The performance evaluation is carried out under the standard of ITU-T G.160. In all the codec modes, the noise attenuation ability of the proposed method in white noise is better than the reference CD methods. In colored noise conditions, the noise reduction is much heavier than the reference method that only modifies algebraic codebook gain, but smaller than the one with the assistance of LD methods, while the loss of speech components is much lower. Comparing with the reference methods, the proposed method provides much better speech quality in most of the full-band distributed noise conditions. In the low-frequency distributed noise, the speech quality is slightly better or equivalent than the reference methods. The subjective speech quality test shows that, the speech quality of enhanced speech produced by the proposed method is better than the one that only modifies the algebraic codebook gains, and is similar to the one with the assistance of LD methods. In conclusion, with much lower computational complexity,

the proposed method can remove both the full-band distributed noise and low-frequency distributed noise effectively, and improve the objective and subjective speech quality evidently.

Acknowledgement

This work was supported by the Beijing Natural Science Foundation Program and Scientific Research Key Program of Beijing Municipal Commission of Education (No. KZ201110005005), the Funding Project for Academic Human Resources Development in Institutions of Higher Learning under the Jurisdiction of Beijing Municipality, the 10th Postgraduate Science Foundation of Beijing University of Technology (ykj-2012-7284), and Huawei Technologies Co., Ltd.

References

Schroeder, M.R., Atal, B.S., 1985. Code-excited linear prediction (CELP): High-quality speech at very low bit rates. In: Proc. IEEE Internat. Conf. Acoust., Speech, Signal Process. (ICASSP), vol. 3, pp. 937–940.

Chandran, R., Marchok, D.J., 2000. Compressed domain noise reduction and echo suppression for network speech enhancement. In: Proc. 43rd IEEE Midwest Symp. Circuits Systems, pp. 10–13.

Duetsch, N., Taddei, H., Beaugeant, C., Fingscheidt, T., 2004. Noise reduction on speech codec parameters. In: Proc. 5th, ITG Fachber., pp. 357–362.

Taddei, H., Beaugeant, C., De Meuleneire, M., 2004. Noise reduction on speech codec parameters. In: Proc. IEEE Internat. Conf. Acoust., Speech, Signal Process. (ICASSP), vol. 1, pp. I497–I500.

Martin, R., 1994. Spectral subtraction based on minimum statistics. In: Proc. EUSIPCO-94, vol. 2, pp. 1182–1185.

Sukkar, R.A., Younce, R.C., Zhang, P., 2006. Method and apparatus for noise reduction. United States Patent Application. Publication Number: US 2006/0217970 A1.

Fapi, E.T., Beaugeant, C., Taddei, H., Pastor, D., 2008. Noise reduction within network through modification of LPC parameters. In: Proc. 7th Internat. Conf. Source and Channel Coding.

ITU-T, 2003. ITU-T G.722.2, Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB). Internat. Telecomm. Union (ITU), Series G.

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. ASSP-32 (6), pp. 1109–1121.

ITU-T, 2002. ITU-T G.722.2 Annex A, Comfort noise aspects. Internat. Telecomm. Union (ITU), Series G.

ITU-T, 2002. ITU-T G.722.2 Annex B, Source controlled rate operation. Internat. Telecomm. Union (ITU), Series G.

ITU-T, 2002. ITU-T G.722.2 Appendix I, Error concealment of erroneous or lost frames. Internat. Telecomm. Union (ITU), Series G.

ITU-T, 2008. ITU-T G.160, Voice enhancement devices. Internat. Telecomm. Union (ITU), Series G.

Varga, A., Steeneken, H.J.M., 1993. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech Comm. 12 (3), pp. 247–251.

Loizou, P., 2005. Speech enhancement based on perceptually motivated Bayesian estimators of the speech magnitude spectrum. IEEE Trans. Speech Audio Process. 13 (5), 857–869.

ITU-T, 2001. ITU-T P.862, Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. Internat. Telecomm. Union (ITU), Series P.

ITU-T, 2005. ITU-T G.191, Software tools for speech and audio coding standardization. Internat. Telecomm. Union (ITU), Series G.