Lee and Chang *EURASIP Journal on Audio, Speech, and Music Processing* (2016) 2016:11
DOI 10.1186/s13636-016-0089-6

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Novel adaptive muting technique for packet loss concealment of ITU-T G.722 using optimized parametric shaping functions

Bong-Ki Lee[1] and Joon-Hyuk Chang[2*]

Abstract

Adaptive muting method using an optimized parametric shaping function as a part of the ITU-T G.722 Appendix IV packet loss concealment algorithm is proposed. The packet loss concealment algorithm incorporating an adaptive muting scheme is known to prevent the generation of unpleasant sounds during packet loss concealment. However, original muting uses a piece-wise linear muting curve according to packet errors so that our muting approach uses non-linear parametric shaping functions including sigmoid and raised-cosine. Training is substantially performed to determine the parameters of the two shaping functions in terms of objective speech quality measures, and optimal parameters are finally selected by subjective speech quality. Through extensive experiments, this proposed muting technique turns out to improve the performance of the reference muting mechanisms in the packet loss concealment algorithm of the G.722 Appendix IV under various experimental conditions.

Keywords: ITU-T G.722, Packet loss concealment, Adaptive muting, Parametric shaping function, Sigmoid, Raised-cosine function

1 Introduction

Recently, a variety of voice communication services through the internet have been building a growing interest in voice over internet protocol (VoIP) systems. VoIP applications are basically developed through a packet-based system over IP networks, which operate with the help of standard codecs such as ITU-T G.722, G.729, G.723.1, and adaptive multi-rate (AMR) [1]. Among them, ITU-T G.722 speech codec is to handle speech and audio signals of bandwidth up to 7 kHz, compared to 3.4 kHz in the case of narrow-band speech codecs [2] and designed to provide good speech quality at rate of 64 kbps. However, VoIP applications have a critical problem of packet loss due to delay and jitter during the transmission of the speech data so that the quality of service (QoS) cannot be acceptable in poor network conditions [3]. Hence, a packet loss concealment (PLC) algorithm, which extrapolates missing frames, is required for VoIP applications in the packet loss environment [4]. This type of problem has been considered many times over a long period of time [5].

In general, PLC algorithms can be classified into two categories: sender-based and receiver-based reconstruction schemes. There are several methods for the sender-based reconstruction schemes including packet retransmission [6], interleaving [7], and sending error correction bits in voice packets using forward error correction (FEC) technique [8]. However, these methods require considerable increase in bandwidth, longer end-to-end delay, or may require modifications on the sender side [9]. For example, the PLC algorithm in enhanced voice services (EVS) codec [10], recently standardized by 3rd generation partnership project (3GPP), transmits side information such as pitch lag to the decoder side so it requires a special transmission format from the encoder at specific encoding modes [11].

On the other hand, various works based on the receiver-based PLC scheme have been proposed. Early PLC algorithms recover from packet at the receiver stage by

*Correspondence: jchang@hanyang.ac.kr
[2]School of Electronic Engineering, Hanyang University, Seoul 04763, Republic of Korea
Full list of author information is available at the end of the article

inserting a silence, or noise, or previous packet [12]. Furthermore, a pitch waveform replication [13] and overlap-add interpolation [14] techniques have been developed as receiver-based PLC schemes. These techniques are easy to implement, but are known poor on speech quality at high packet losses. A linear prediction (LP)-based PLC algorithm which estimates the excitation signal of the missing segment by repeating the excitation signal of the previous received speech was described by Gunduzhan et al. [15]. More advanced approaches such as waveform similarity overlap-and-add (WSOLA) [16] and time-warping and re-phasing [17] methods have been proposed to produce a more natural sound. Recently, parametric regression approaches to PLC such as the Gaussian mixture model (GMM) [18] and hidden Markov model (HMM) [19] have been proposed where relevant codec parameters such as the pitch period, line spectral frequencies (LSFs), and codebook gain are estimated by using their specific models. However, these methods often introduce annoying artifacts for consecutive packet losses so that they cannot provide a high-quality output.

Note that, in 2006, ITU-T G.722 was revised to recommend standard PLC algorithms by adding Appendix III [20] and IV [21] to ITU-T G.722. The PLC algorithm described in Appendix III has higher quality but increases decoding computational complexity, while PLC algorithm described in Appendix IV brings almost no additional complexity compared with G.722 normal decoding. This paper is basically concerned with the G.722 Appendix IV. In the Appendix IV algorithm which is an LP-based PLC scheme, lost packets are extrapolated based on previously received packets with relevant information such as their LP coefficients (LPCs), signal classification, and the pitch period. Since reconstructing the missing frames often generates unpleasant sounds, especially in the case of successive packet losses (i.e., error burst), an adaptive muting method is included at the end of the PLC algorithm. The pre-reconstructed signal is multiplied by the pre-defined adaptive muting factor and this muting factor is gradually decreased during successive packet losses. Indeed, the muting is applied differently according to the class of the signal using a pre-determined fixed curve. There have been few studies on the muting mechanism though its importance in terms of perceptual listening quality. Therefore, it seems to be a worthwhile subject to study the muting mechanism, which improves the speech quality without increasing the delay.

In addition, Kovesi et al. proposed an improved version of muting curve in G.722 Appendix IV [22] at which the muting is performed more slowly than the original muting scheme using the piecewise linear muting curve. This minor change of the muting curve can lead to improve the speech quality without increasing worst-case complexity however, the optimization method used in [22] is not described and the performance might still not be sub-optimal. it is not optimized for listening quality, whereas our proposed muting technique is optimized for both objective and subjective speech quality measures.

In this paper, we present an improved adaptive muting method that determines the adaptive muting curve based on the optimized parametric functions. The parametric functions such as an exponential shaping function and raised-cosine function are chosen as the muting curve because they have superior freedom in the shape and are inherently characterized by core parameters, which are optimized to minimize the difference between the desired signal and the reconstructed signal. For this, in a training stage, the grid-search technique [23] is employed to determine optimal values of the parameters within the search space according to given error criteria. The exponential shaping function is firstly applied to the muting algorithm and used to enhance the quality of the reconstructed speech signal. This idea was originally presented in [24] where the sigmoid type function is solely applied to the muting curve and its two core parameters are optimized by using the grid-search method. Unlike [24], we first apply the raised-cosine function to offer better solution with an in-depth analysis. And then we try to find the optimal points of each function based on various error criteria including the mean square error (MSE), wideband-perceptual evaluation of speech quality (WB-PESQ) [25], segmental SNR (segSNR), and frequency weighted segmental SNR (fwSNRseg) [26], which are known to be objective ways to evaluate perceptual speech quality. Finally, we choose an optimal point among the optimal points as found earlier where the speech quality is successfully supported by the mean opinion score (MOS) test [27]. An extensive comparative study of the performance of each parametric shaping function as well as investigation between the objective speech quality measures and subjective speech quality in the muting mechanism are performed to highlight the contribution of our research compared to the previous research in [24].

According to the experimental results, it turns out that the proposed method outperforms the reference muting methods in terms of various speech quality measures. The rest of this paper is organized as follows. Section 2 briefly reviews the reference muting mechanisms, which are a baseline of the proposed algorithm, Section 3 introduces two parametric functions and describes the proposed adaptive muting method based on the optimization method, Section 4 presents simulation results, and Section 5 presents our conclusions.

2 Review of the reference muting mechanisms

The PLC algorithm specified in Appendix IV of ITU-T G.722 [21] corresponds to a receiver-based scheme as introduced in Section 1, where the used information is

originally based on the packet previously received. Therefore, the encoder does not have to be modified, but the decoder is slightly changed by adding a PLC mechanism. The readers note that the terms "frame" and "packet" are used interchangeably in this paper. Specifically, the G.722 decoder includes additional blocks for the PLC algorithm as shown in the grey-shaded blocks of Fig. 1 [21]. The ITU-T G.722 codec belongs to the type of sub-band adaptive differential pulse code modulation (SB-ADPCM), thereby splitting the frequency band into two sub-bands (a lower band and a higher band). It is noted that since the operation in the higher-band is included in that of the lower-band, we focus in this paper on describing how to operate the PLC algorithm at the lower-band only.

At first, if there are no packet errors, the cross-fade block does not change the reconstructed signal, i.e., $zl(n) = xl(n)$. Then, when packet loss actually occurs, the reconstructed lower-band signal $yl(n)$ is simply extrapolated through the LPC-based pitch repetition block using the past valid lower-band signal $zl(n)$. After extrapolation, $yl(n)$ and previous decoded signal $xl(n)$ are cross-faded to reduce the discontinuity between the frames as shown in Fig. 1.

For clear comprehension, Fig. 2 further shows the lower-band LPC-based pitch repetition block diagram of the G.722 decoder incorporating the PLC algorithm [21]. In this figure, the pre-reconstructed lower-band signal $yl_{\text{pre}}(n)$, which is prior to the adaptive muting, is synthesized by using $zl(n)$ and the LP analysis, long-term prediction (LTP) analysis, signal classification, and pitch repetition blocks. As mentioned above, unpleasant sounds are generated especially in successive packet losses if $yl_{\text{pre}}(n)$ is used inherently since the same reconstructed signal is repeated. Accordingly, the adaptive muting method is devised in the final step of the PLC algorithm to mitigate the effect of the unpleasant sounds. Considering the adaptive muting mechanism, the reconstructed lower-band signal $yl(n)$ is represented as

$$yl(n) = G(n) \cdot yl_{\text{pre}}(n), \tag{1}$$

where $G(n)$ denotes the adaptive muting factor, which has a value between 1 and 0. As given in (1), the pre-reconstructed lower-band signal $yl_{\text{pre}}(n)$ is multiplied by the adaptive muting factor on a sample-by-sample basis during the lost frames.

On the other hand, the adaptive muting factor is differently applied according to the class of the signal determined in the G.722 decoder as shown in Fig. 3a. While the *Transient* and *UV Transition* classes correspond to a transient period with large energy variation and a transition between voiced and unvoiced signals, respectively, the *Other Cases* class includes unvoiced, weakly voiced, and voiced signals, each of which is the superior candidate for extrapolation because the perceived quality of reconstructed speech largely depends on this type of the signal [21]. In addition, the adaptive muting factor decreases to zero after 320 samples (corresponding to four packets in the lower band), producing silence and thereby preventing the generation of unpleasant sounds when more than four packets are lost.

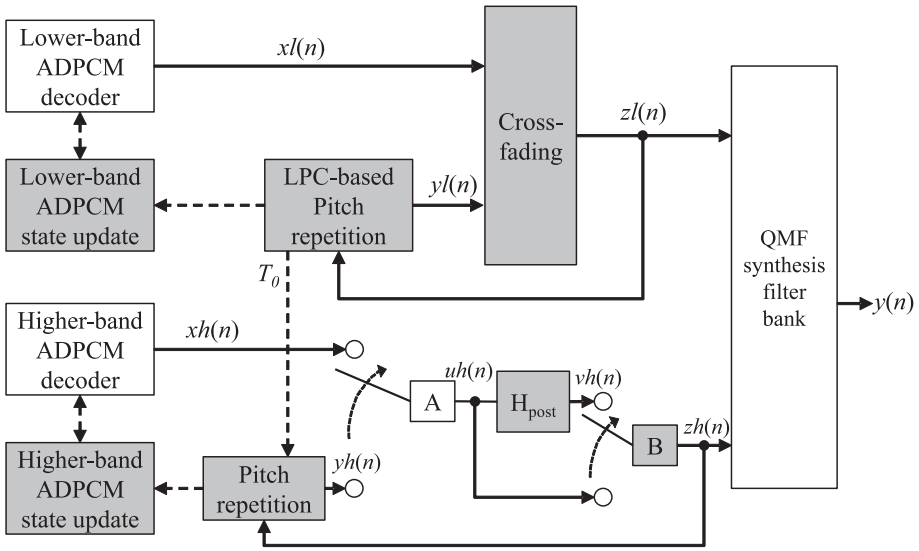

Fig. 1 Block diagram of G.722 decoder with the PLC algorithm

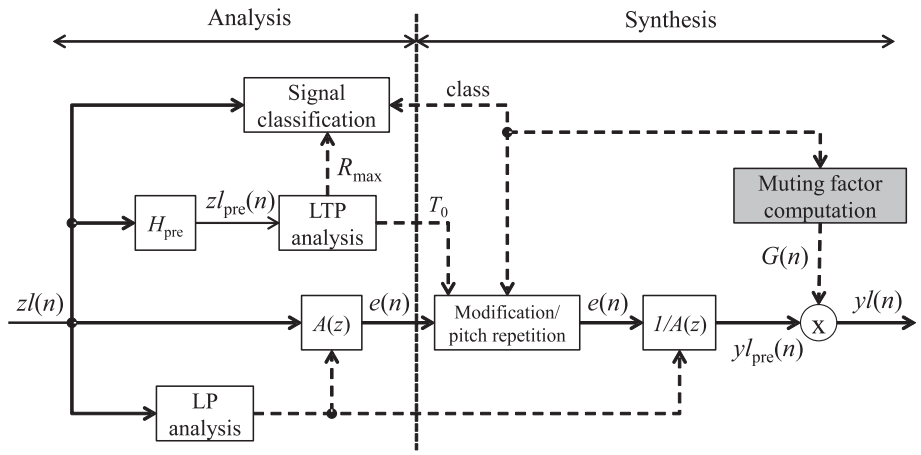

Fig. 2 Block diagram of lower-band extrapolation of missing frame

The actual parameters of the original adaptive muting method are shown in Table 1. Notice that these parameters depend on classes of the signal and the adaptive muting factor is consequently derived such that

$$G(n+1) = \begin{cases} G(n) - fac_1 & , 0 \le n < 80 \\ G(n) - fac_{2\mathrm{p}} & , 80 \le n < 160 \\ G(n) - fac_{3\mathrm{p}} & , 160 \le n < 320 \\ 0 & , n \ge 320 \end{cases} \quad (2)$$

where the adaptive muting factor is performed on a sample-by-sample basis with fac_1 for the first lost packet, $fac_{2\mathrm{p}}$ for the second lost packet, and $fac_{3\mathrm{p}}$ for the third and fourth lost packets. Indeed, the adaptive muting factor becomes zero after the fourth lost packet. Note that the original adaptive muting is linearly applied to each packet as stated previously. This adaptive muting method is inherently applied to the higher band in the same manner as the lower band. Last;y, the reconstructed signals of the lower band and higher band are combined into the wideband decoded signal $y(n)$ through the quadrature mirror filter (QMF) synthesis filterbank.

An improved version of the muting curve proposed by Kovesi et al. is shown in Fig. 3b [22]. In their work, the last linear part of the muting curve is changed in a way that the complete muting is achieved after 60 ms for the *Other cases* class and after 30 ms for the *UV transition* class in which the muting is slowly made than that of the original muting.

3 Proposed adaptive muting method

As explained in Section 2, the original adaptive muting method in Appendix IV of ITU-T G.722 is linearly applied between successive frames according to the predetermined curve. In the proposed algorithm, we present an improved adaptive muting algorithm which is applied non-linearly using the two parametric shaping functions such as the exponential function and raised-cosine function, commonly used for the logistic function. We firstly compare the performance of the two parametric shaping functions as for the muting curve according to the various error criteria. Then, optimal values of the parameters of the two parametric shaping functions are selected according to the grid-search [23], which is an exhaustive search method to find the optimal point in a manually specified subset of the parameter space of the learning algorithm, established by the given error criteria: the MSE and WB-PESQ, segSNR, and fwSNRseg. First, the sigmoid function is employed by incorporating three parameters as given by

$$G_s(n) = \frac{1 + 0.1 \cdot \alpha_s e^{-3\beta_s \gamma_s}}{1 + 0.1 \cdot \alpha_s e^{\beta_s (n - 3\gamma_s)}} \quad (3)$$

where α_s and β_s denote sloping parameters to control the shape of the sigmoid function, and γ_s denotes a shift parameter, respectively. Second, we consider the raised-cosine type function, which is commonly used for pulse shaping:

$$F(x) = \begin{cases} -1 & , x < -\frac{1+\beta_r}{2\alpha_r} \\ \alpha_r x - \frac{1-\beta_r}{2} & \\ \quad -\frac{\beta_r}{\pi} \cos\left[\frac{2\alpha_r x\pi + \pi}{2\beta_r}\right] & , -\frac{1+\beta_r}{2\alpha_r} \le x \le -\frac{1-\beta_r}{2\alpha_r} \\ 2\alpha_r x & , -\frac{1-\beta_r}{2\alpha_r} \le x \le \frac{1-\beta_r}{2\alpha_r} \\ \alpha_r x + \frac{1-\beta_r}{2} & \\ \quad +\frac{\beta_r}{\pi} \cos\left[\frac{2\alpha_r x\pi - \pi}{2\beta_r}\right] & , \frac{1-\beta_r}{2\alpha_r} \le x \le \frac{1+\beta_r}{2\alpha_r} \\ 1 & , x > \frac{1+\beta_r}{2\alpha_r} \end{cases} \quad (4)$$

where α_r and β_r determine the shape and the dynamic range of the function, and we modify this equation to the three-parameter raised-cosine type function which is a

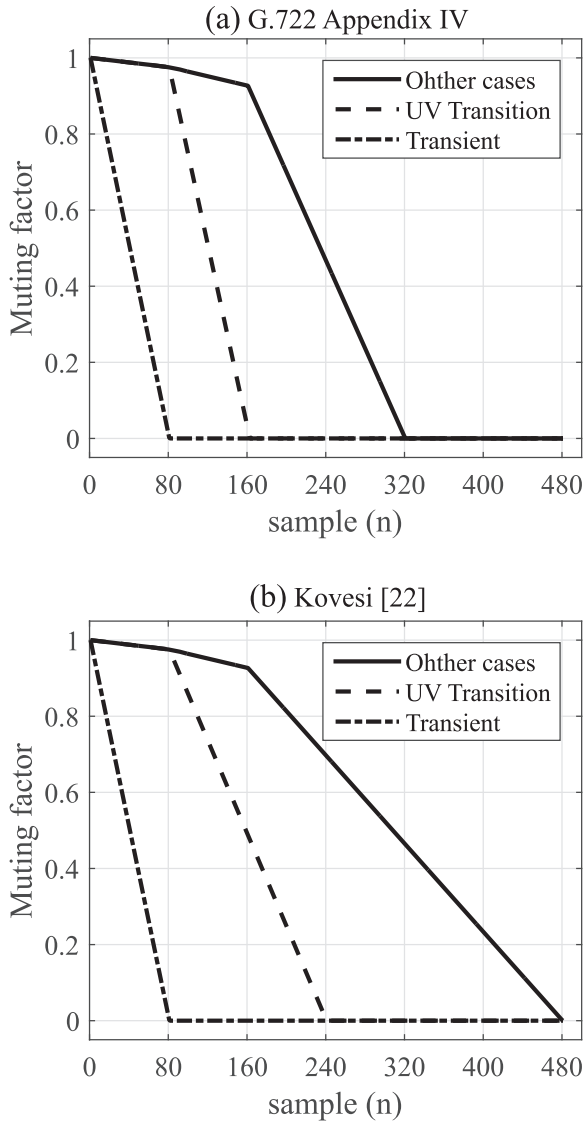

Fig. 3 Muting factor according to three classes of the signal in **a** G.722 Appendix IV and **b** Kovesi [22]

time-scaled and shifted version of (4) in accordance with the muting curve:

$$G_r(n) = \frac{1}{2}\left[F\left(\frac{-n+\gamma_r}{2\gamma_r}\right)+1\right] \tag{5}$$

where γ_r denotes a shift parameter. It is worthwhile noting that the muting curve $G(n)$ can be controlled by the core parameters (i.e., α, β, and γ) of the function. Also, in contrast with the original method, the muting factor $G(n)$ does not become zero after 320 samples for both sigmoid and raised-cosine functions to offer a great amount of flexibility for the muting curve.

As a consequence, these parametric shaping functions, which decrease monotonically, can offer more flexibility

Table 1 Adaptive muting parameters in G.722 Appendix IV

Parameter	Speech class type		
	Transient	*UV_transition*	*Other cases*
fac_1	409	10	10
fac_{2p}	409	10	20
fac_{3p}	409	399	190

to the shape of the muting curve than that of the reference muting curves [21, 22]. Since it is well-known that the *Other cases* class which includes unvoiced, weakly voiced, and voiced signals plays a dominant role in the perceived speech quality of reconstructed speech, we use the parametric shaping functions as the muting curve to the *Other cases* class only, while (2) is applied to the *Transient* and *UV transition* classes.

To estimate the difference between the desired speech signal and reconstructed speech signal, we adopt various error criteria: MSE, WB-PESQ, segSNR, and fwSNRseg. For considering the MSE criterion first, we use (1) so that the error between the desired signal and the reconstructed signal can be interpreted as

$$\begin{aligned}\varepsilon(n) &= dl(n) - yl(n)\\ &= dl(n) - G(n)\cdot yl_{\mathrm{pre}}(n),\end{aligned} \tag{6}$$

where $dl(n)$ denotes the desired lower-band signal, which is equal to a lower-band decoded signal without any packet losses, and $G(n)$ can be defined by (3) or (5). Thus, the MSE can be expressed as

$$J(\alpha,\beta,\gamma) = \sum_{n=1}^{N} E\left[\varepsilon(n)\right]^2, \tag{7}$$

where N denotes the total number of samples for a training data file. Note that the cost function in (7) contains three unknown parameters; α, β, and γ for both sigmoid (α_s, β_s, γ_s) and raised-cosine (α_r, β_r, γ_r) type functions so that they can be expressed as a function of α, β, and γ. From (7), the average of MSEs for training data files is expressed as

$$\xi(\alpha,\beta,\gamma) = \frac{1}{L}\sum_{l=1}^{L} J_l(\alpha,\beta,\gamma) \tag{8}$$

where l and L, respectively, denote the index of the training file and the total number of training files for the grid-search according to the processed speech by the proposed PLC algorithm. In (8), to find the optimal parameters, we compute the average of MSEs over all training data in the speech materials by varying α, β, and γ:

$$(\alpha^*,\beta^*,\gamma^*) = \arg\min_{\alpha,\beta,\gamma} \xi(\alpha,\beta,\gamma) \tag{9}$$

and taking the optimal parameters α^*, β^*, and γ^* to be those that minimize $\xi(\alpha,\beta,\gamma)$.

Since the packet losses actually affect the signal quality during speech periods, we also adopt the well-known objective speech quality measures such as WB-PESQ, segSNR, and fwSNRseg for the error criterion to measure the speech quality.

First, the segSNR, instead of working on the whole signal, is calculated by the average of the SNR values on short frames as given by

$$segSNR = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\frac{\sum_{n=Tm}^{Tm+T-1}x^2(n)}{\sum_{n=Tm}^{Tm+T-1}\{x(n)-x(n)\}^2}, \quad (10)$$

where T and M indicate the frame length (10 ms) and number of frames in the signal, respectively. And, the values for the upper and lower ratio limit are 35 and −10 dB, respectively.

Next, fwSNRseg is a weighted segSNR within a frequency band proportional to the critical band which can be defined as follows:

$$fwSNRseg = \frac{10}{M}\sum_{m=0}^{M-1}\frac{\sum_{j=0}^{K-1}W(j,m)\log_{10}\frac{X(j,m)^2}{\{X(j,m)-\hat{X}j,m\}^2}}{\sum_{j=0}^{K-1}W(j,m)}, \quad (11)$$

where $W(j,m)$ is the weight on the jth subband in the mth frame which is taken from the ANSI SII standard, K is the number of subbands, $X(j,m)$ is the spectrum magnitude of the jth subband in the mth frame, and $\hat{X}(j,m)$ is distorted spectrum magnitude. Previous studies have shown that segSNR and fwSNRseg show significantly higher correlation with subjective quality than the classical SNR.

In a similar manner to the MSE estimator, the score of speech quality measures F abovementioned is calculated based on training data files and an average value of it is computed by

$$\xi(\alpha,\beta,\gamma) = \frac{1}{L}\sum_{l=1}^{L}F_l(k). \quad (12)$$

Each file lasting 8 s consists of two different sentences and then processed by the proposed PLC algorithm in which $G_s(n)$ and $G_r(n)$ are applied to (1) as on adaptive muting mechanism. Since F depends on the reconstructed signal, ξ can be determined as a function of α, β, and γ. To find the optimal parameters, we compute the average of F in each objective measure over all training data in the speech materials by varying α, β, and γ and taking the optimal parameters α^*, β^*, and γ^* to be those that satisfy the below:

$$(\alpha^*,\beta^*,\gamma^*) = \arg\max_{\alpha,\beta,\gamma}\xi(\alpha,\beta,\gamma). \quad (13)$$

For parameter training of parametric shaping functions, we used a number of speech materials from the TIMIT database, as will be described further in Section 4. Then, the parameters according to the objective measures will be obtained and those will be applied to (3) and (5), respectively, for the test phase. Finally, the optimal parameters

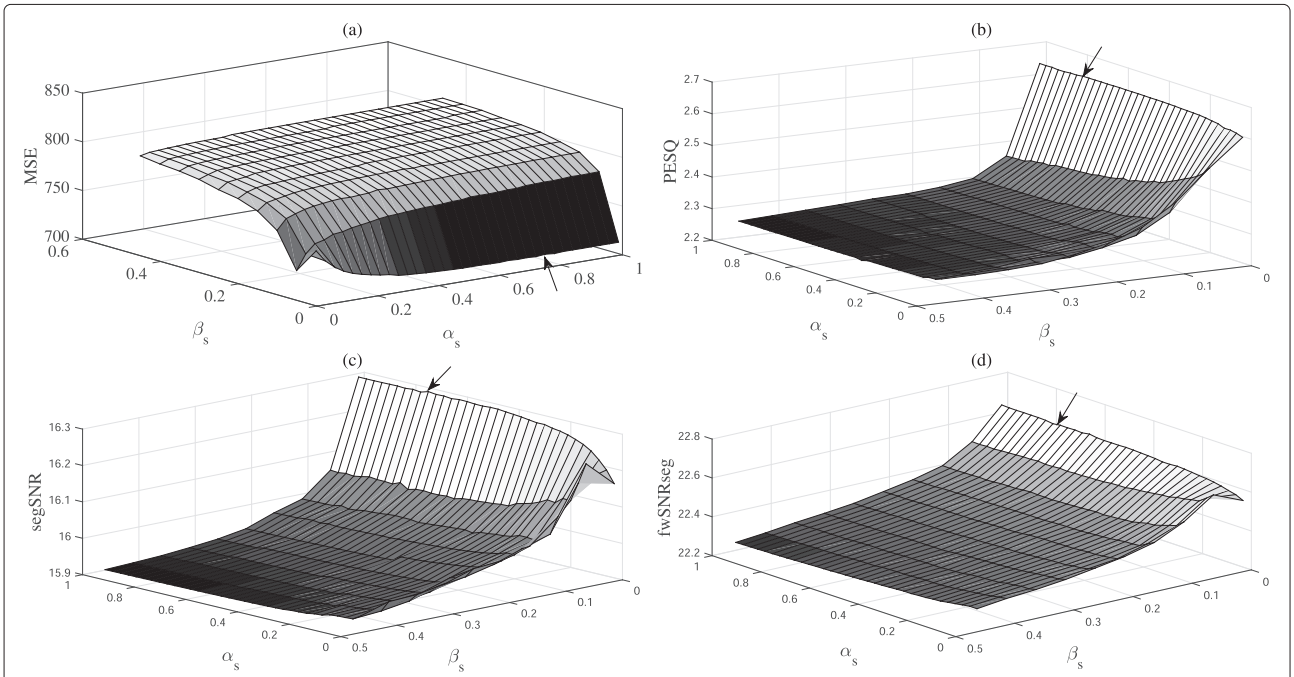

Fig. 4 3D *mesh curves* for the estimated optimal point using the sigmoid function (10 % of packet loss condition): **a** MSE, $\gamma_s = 0$, **b** WB-PESQ, $\gamma_s = 50$, **c** segSNR, $\gamma_s = 0$, and **d** fwSNRseg, $\gamma_s = 0$

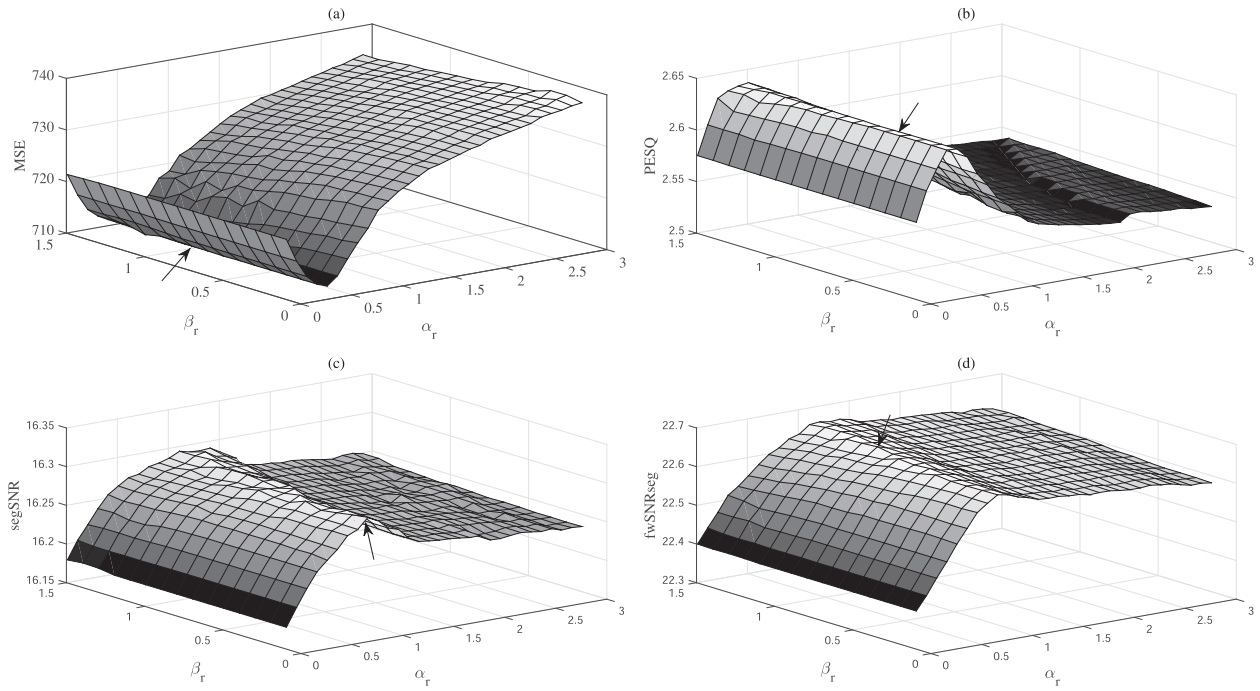

Fig. 5 3D mesh curves for the estimated optimal point using the raised-cosine function (10 % of packet loss condition): **a** MSE, $\gamma_r = 300$, **b** WB-PESQ, $\gamma_r = 350$, **c** segSNR, $\gamma_r = 200$, and **d** fwSNRseg, $\gamma_r = 200$

based on the MOS test will be chosen from the parameters obtained by the objective measures. It is noted that the proposed muting method does not cause any additional algorithmic delay and storage since finding the optimal parameters is based on the off-line training process. Also, like [21] and [22], worst-case complexity of bad frame processing is still lower than that of good frame processing, the overall worst-case complexity is unchanged. Finally, this proposed adaptive muting method is applied to the higher-band in the same way as to the lower-band.

4 Experiments and results

In this section, we find the optimal muting curve based on the objective and subjective speech quality. First, the parameters will be obtained according to the objective measures including MSE, WB-PESQ, segSNR, and fwSNRseg. Then, the optimal parameters can be chosen by the MOS test among the parameters obtained from the objective measures. For experiments, the standard TIMIT corpus [28] were used for the parameter training phase and NTT Korean speech database [29] were used for the test phase. In order to further verify the effectiveness of our proposed methods, we compared the proposed muting algorithm with the reference muting algorithms in G.722 Appendix IV [21] and Kovesi [22] under conditions of various packet loss rates. Furthermore, simple adjustment version of the original piecewise linear function in G.722 Appendix IV is also compared with the proposed muting method to check the performance difference between the piecewise linear curve and proposed

Table 2 Optimal parameters of the sigmoid function

Packet loss rate	MSE			WB-PESQ			segSNR			fwSNRseg		
	α_s	β_s	γ_s	α_s	β_s	γ_s	α_s	β_s	γ_s	α_s	β_s	γ_s
1 %	0.70	0.01	0	0.70	0.01	0	0.70	0.01	0	0.64	0.01	0
3 %	0.85	0.01	0	0.82	0.01	50	0.64	0.06	100	0.94	0.01	50
6 %	1.00	0.01	0	0.88	0.01	0	0.76	0.01	0	0.91	0.01	0
10 %	0.79	0.01	0	0.61	0.01	50	0.73	0.01	0	0.64	0.01	0
20 %	0.70	0.01	0	0.91	0.01	150	0.70	0.01	0	0.64	0.01	0
Average	0.81	0.01	0	0.78	0.01	50	0.71	0.02	20	0.75	0.01	10

Table 3 Optimal parameters of the raised-cosine function

Packet loss rate	MSE			WB-PESQ			segSNR			fwSNRseg		
	α_r	β_r	γ_r	α_r	β_r	γ_r	α_r	β_r	γ_r	α_r	β_r	γ_r
1 %	0.41	1.2	250	0.31	0.3	400	1.11	1.4	200	0.71	1.2	250
3 %	0.61	1.2	200	0.51	0.9	350	1.01	1.3	250	0.91	0.7	200
6 %	0.91	1.5	200	0.31	0.2	450	1.21	1.5	200	1.01	1.0	200
10 %	0.61	1.1	300	0.31	0.2	350	0.81	0.1	200	1.01	1.0	200
20 %	0.41	1.1	300	0.31	1.0	450	0.71	1.2	200	1.11	1.3	250
Average	0.59	1.3	240	0.35	0.52	400	0.97	1.1	210	0.95	1.04	220

non-linear curve. In the simple adjustment version of the muting curve which will be called 'G.722 Appendix IV+' in this paper, the last linear part of the muting curve is changed in a way that the complete muting is achieved after 80 ms for the *Other cases* class and the muting curves for *UV transition* and *Transient* classes are the same with the reference method in Kovesi [22]. In the algorithm we developed, the speech data was sampled at 16 kHz and input speech level was set to −26 dBov. The size of the packet (frame) was set to 10 ms (160 samples for the wideband signal) and random packet (frame) losses were inserted by using the error insertion device (EID) in ITU-T G.191 software tool [30], which uses Gilbert and Bellcore models, with zero bit error rate and 0.5 of burst factor in which 0 and 1 indicate the totally random and burst error, respectively. In the training and test processes, from 1 to 30 % of the packet loss rates are used [31] and values of the random seed were changed to ensure that performance evaluation is not biased. Various tests were performed to find an optimal muting curve and performance comparison in the next subsection.

4.1 Finding an optimal muting curve

For each parametric shaping function, we obtained the 3D mesh curves according to the objective measures as a function of the various values of α and β when γ is fixed as illustrated in Figs. 4 and 5 at which the packet loss rate was assumed to be 10 %. As a result, we obtained each of the optimal value $(\alpha^*, \beta^*, \gamma^*)$ for the various objective measures under various packet loss rates as summarized in Tables 2 and 3. Since it is hard to estimate the packet loss rate in the wireless network in advance and there are small deviations of the parameters obtained from various packet loss rates, we use average values of parameters for parametric shaping functions. Next, we choose the optimal parameters from the parameters described in Tables 2 and 3 by using absolute category rating (ACR) tests which use the five-point MOS score. For speech materials, we randomly selected total 30 files spoken by four male and female native Korean speakers from the NTT Korean speech database where each phrase consists of two different meaningful sentences and the duration of each sentence was 8 s. For this subjective test, 14 Korean listeners with normal hearing score their respective subjective opinions on the quality of each file, using one of the following points: 5 (Excellent),

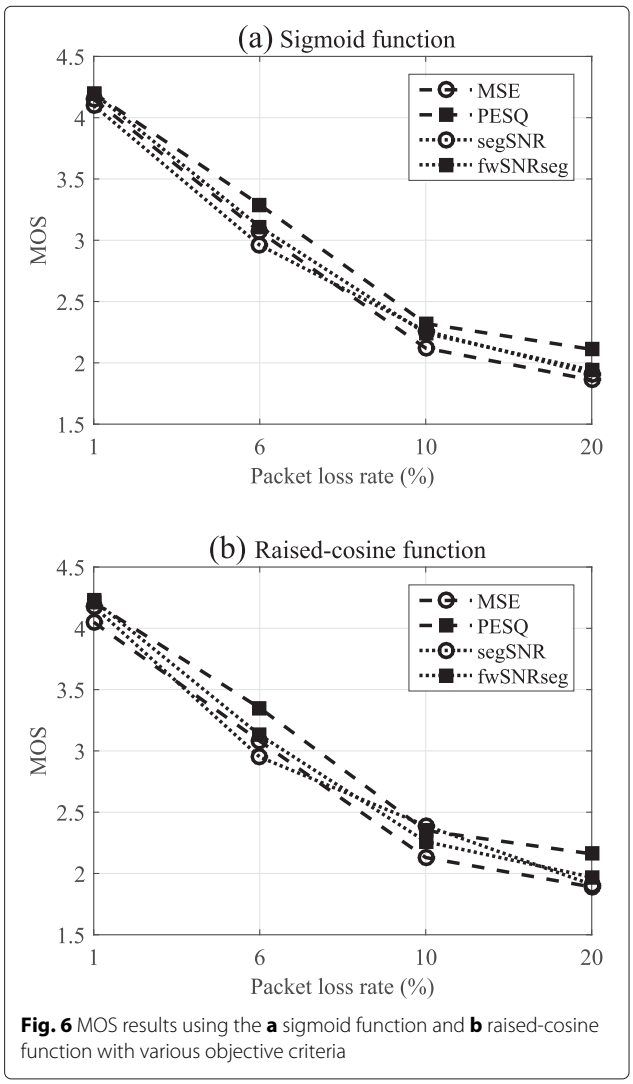

Fig. 6 MOS results using the **a** sigmoid function and **b** raised-cosine function with various objective criteria

4 (Good), 3 (Fair), 2 (Poor), and 1 (Bad). The test was conducted in a quiet room and the speech samples were diotically presented through a monitoring headphone. Although this number may not be sufficient for a formal assessment of speech processing systems, it would give an efficient indicator for checking their subjective performance. Figure 6 shows the MOS results according to the objective measures including the MSE, WB-PESQ, segSNR, and fwSNRseg under various packet loss rates. From this result, results obtained from the WB-PESQ criterion is the closest to the listening speech quality and then the parameters according to the WB-PESQ criteria are chosen for the optimal parameters. And, since the case of using the raised-cosine function is slightly better than the case of using the sigmoid function according to the MOS score, the raised-cosine function with the WB-PESQ criterion is finally chosen for the proposed muting curve. Figure 7 represents the muting curve in comparison with the Kovesi [22], G.722 Appendix IV+, and proposed methods including the sigmoid and raised-cosine functions in the case of *Other cases* class. The proposed muting of the raised-cosine is performed much slowly than the proposed muting of the sigmoid and reference muting methods. It is noted that the proposed muting curve is the same with the reference muting curves in [21, 22] in the case of *UV transition* and *Transient* classes.

4.2 Performance comparison

Prior to the performance comparison, we analyzed the results obtained without the muting scheme to illustrate the importance of muting mechanisms in VoIP applications. First, Fig. 8 illustrates the output waveforms of the desired speech signal which is decoded without packet loss and decoded speech signals using the reference muting methods in G.722 Appendix IV and Kovesi [22], G.722 Appendix IV+, and proposed muting method optimized by using the raised-cosine with WB-PESQ criterion. Figure 8 shows the example of the waveform comparison where dotted boxes indicate the packet loss periods in voiced speech segments (0.05 to 0.15 s). In Fig. 8a, b, the waveform is over-muted, which causes the degradation of speech quality. On the other hand, the waveforms of Fig. 8c, d are more similar to the desired speech signal in comparison to the reference methods. Furthermore, it is seen that the waveform of the proposed method in Fig. 8d is much closer to the desired speech signal than the G.722 Appendix IV+ in Fig. 8c while avoiding the annoying artifacts which will be verified by the speech quality measures.

Also, the above methods were evaluated with objective speech quality measures including log-likelihood ratio (LLR) [32] and WB-PESQ. As Table 4 summarizes the overall results, it is readily seen that the proposed method using the raised-cosine function opti-

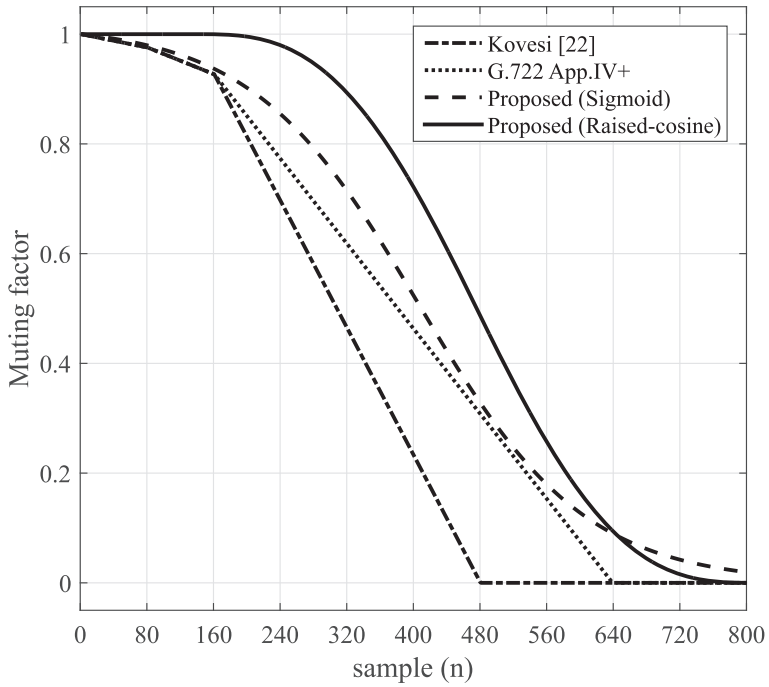

Fig. 7 Muting factor comparison in the case of *Other Cases* class by the G.722 App.IV, Kovesi [22], G.722 App.IV+, and proposed method using the raised-cosine with the WB-PESQ criterion

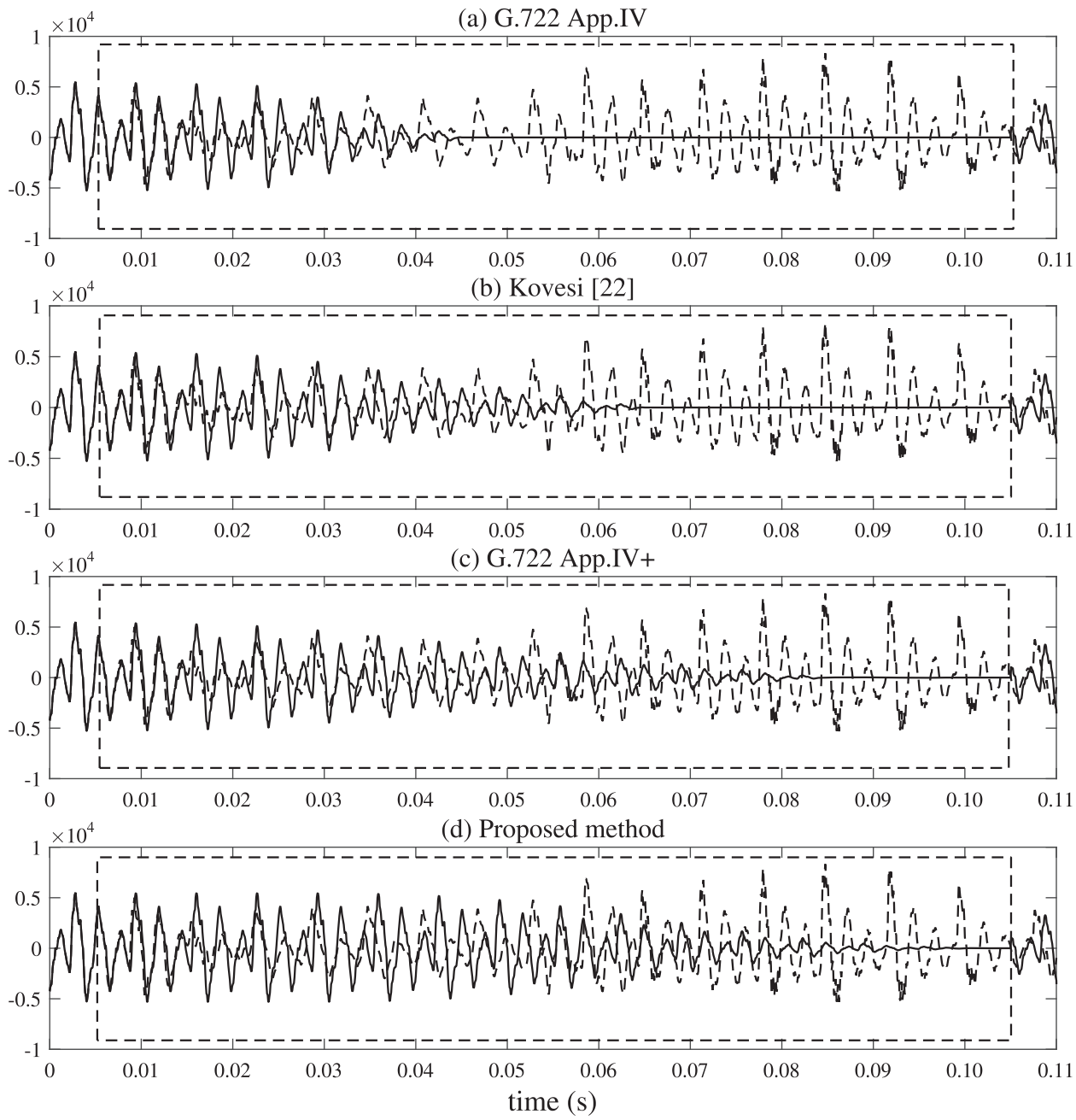

Fig. 8 Waveform comparison (*dash*: desired speech signal, *solid*: decoded speech signal) by the **a** G.722 App.IV, **b** Kovesi [22], **c** G.722 App.IV+, and proposed method using the raised-cosine with the WB-PESQ criterion during packet loss periods (0.05 to 0.15 s)

mized by WB-PESQ criterion outperformed the existing approaches in terms of the LLR and WB-PESQ over various packet loss rates. Especially, the performance of the G.722 Appendix IV+ is better than the G.722 Appendix IV and Kovesi [22], however, worse than the proposed method.

In addition, to validate the objective assessment, we performed a subjective quality test, namely MOS including ACR and comparison category rating (CCR) methods [27]. For the ACR test, total 90 files from the NTT Korean speech database are randomly selected and listening environments are the same with above MOS test. The result of the subjective quality test using ACR was similar to those of the objective quality test as Table 5 shows that the MOS results are statistically significant. Based on the ACR test, the proposed muting algorithm was compared

Table 4 Comparison of objective quality test (95 % confidence interval)

Packet loss rate	G.722 App. IV		Kovesi [22]		G.722 App.IV+		Proposed method	
	LLR	WB-PESQ	LLR	WB-PESQ	LLR	WB-PESQ	LLR	WB-PESQ
1 %	0.036	3.93 ± 0.01	0.036	3.93 ± 0.01	0.037	3.94 ± 0.03	0.036	3.94 ± 0.01
3 %	0.059	3.42 ± 0.03	0.058	3.44 ± 0.02	0.054	3.47 ± 0.03	0.052	3.47 ± 0.03
6 %	0.095	2.95 ± 0.02	0.091	2.98 ± 0.02	0.087	3.12 ± 0.05	0.083	3.15 ± 0.03
10 %	0.147	2.62 ± 0.04	0.139	2.64 ± 0.03	0.129	2.72 ± 0.02	0.121	2.77 ± 0.02
20 %	0.277	1.97 ± 0.03	0.263	2.10 ± 0.05	0.257	2.16 ± 0.04	0.246	2.23 ± 0.04

to the reference muting algorithms including Kovesi [22] and G.722 App. IV+ using CCR test to access speech processing that either degrades or improves the quality of the speech. In each test case, the listener was presented with two differently processed instances of the same sentence. For pairwise comparisons, each of the ten sentences from Korean speech corpus, which are not overlapped with materials of ACR test, were randomly presented to each listener and have durations between 3.5 and 7.5 s. In each of the ten sentences, 3, 6, 10, and 20 % of the packet losses are used for test and the other listening environments are the same with above ACR test. Figure 9 shows the distributions of listener ratings for each pair of processing types where the bars illustrate the relative frequencies of the scores given in the comparisons between the proposed and reference muting methods. Comparisons of the proposed and reference muting methods to the without muting method (w/o muting) show that muting is preferred in most cases. On average, the proposed muting method obtained higher preference scores than Kovesi and G.722 App. IV when compared to the w/o muting method. Also, the proposed muting method was considered substantially better than either Kovesi and G.722 App. IV+ muting methods. The mean score for each pair of processing types is shown on the horizontal axis together with the 95 % confidence interval. In all three comparisons, the mean score was significantly different from zero (t test, $p < 0.05$) indicating a statistical preference.

A comparison of overall simulation results suggests that the proposed method improved the speech quality compared to the reference methods including G.722 Appendix IV, Kovesi, and Appendix IV+. This result shows some improvement in perceptual speech quality and thus confirms the superiority of the proposed algorithm in adverse network environments. Furthermore, optimizing the piecewise linear curve by using the grid search technique yields much higher computational burden while the proposed non-linear curve which consists of two or three core parameters such as sigmoid or raised-cosine function is better to be used for muting curve optimization.

5 Conclusions

In this paper, we presented an improved adaptive muting method using parametric shaping functions for ITU-T G.722 Appendix IV. This method was based on the minimization of the error between the desired signal and reconstructed signal by using well-defined sigmoid and raised-cosine type functions to scale the muting factor when packet losses substantially occur. The parameters of each function were selected using the objective speech quality measures based on the grid-search and then applied to the sigmoid and raised-cosine type functions in the muting algorithm of the G.722 Appendix IV. Then, the optimal parameters are finally chosen by using the subjective speech quality measure. The performance of the proposed approach using the raised-cosine function optimized by WB-PESQ criterion was found best compared with those of the reference methods through objective and subjective quality tests.

Table 5 Overall MOS test results (95 % confidence interval)

Packet loss rate	Mean opinion score			
	G.722 App. IV	Kovesi [22]	G.722 App. IV+	Proposed method
1 %	3.64 ± 0.03	3.73 ± 0.04	3.75 ± 0.05	3.72 ± 0.04
5 %	3.12 ± 0.06	3.21 ± 0.05	3.25 ± 0.07	3.26 ± 0.06
10 %	2.57 ± 0.05	2.63 ± 0.07	2.64 ± 0.05	2.68 ± 0.07
20 %	1.88 ± 0.09	1.93 ± 0.08	1.98 ± 0.11	2.04 ± 0.13
30 %	1.13 ± 0.11	1.31 ± 0.08	1.36 ± 0.14	1.47 ± 0.12

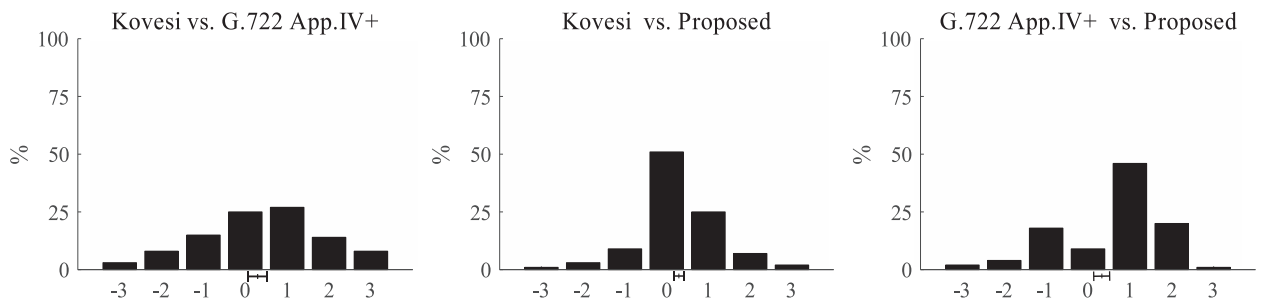

Fig. 9 Distributions of listener ratings in pair-wise comparisons between the proposed and reference muting methods in the CCR test

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2014R1A2A1A10049735) and the ICT R&D program of MSIP/IITP (R0126-15-1119, Development of a solution for situation-awareness based on the analysis of speech and environmental sounds).

Author details
[1]Department of Electronic and Computer Engineering, Hanyang University, Seoul 04763, Republic of Korea. [2]School of Electronic Engineering, Hanyang University, Seoul 04763, Republic of Korea.

Received: 25 May 2015 Accepted: 5 April 2016
Published online: 21 April 2016

References

1. B Goode, Voice over internet protocol (VoIP). Proc. IEEE. **90**(9), 1495–1517 (2002)
2. P Mermelstein, G.722: a new CCITT coding standard for digital transmission of wideband audio signals. IEEE Commun. Mag. **26**(1), 8–15 (1988)
3. U Tadeus, Quality of service in VoIP communication. AEU. Int. J. Electron. Commun. **58**(3), 178–182 (2004)
4. JH James, C Bing, L Garrison, Implementing VoIP: a voice transmission performance progress report. IEEE Commun. Mag. **42**(7), 36–41 (2004)
5. BW Wah, X Su, D Lin, in *Proc. Int. Symp. Multimedia Software Engineering*. A survey on error-concealment schemes for real-time audio and video transmissions over the internet (IEEE, Taipei, Taiwan, 2000), pp. 17–24
6. S Floyd, V Jacobson, S McCanne, L Ching-Gung, Z Lixia, A reliable multicast framework for light-weight sessions and application level framing. IEEE/ACM Trans. Networking. **25**(4), 342–356 (1995)
7. J Ramsey, Realization of optimum interleavers. IEEE Trans. Info. Theory. **16**(3), 338–345 (1970)
8. C Padhye, K Christensen, W Moreno, in *Proc. IEEE Int. Performance, Computing and Commun. Conf.* A new adaptive FEC loss control algorithm for voice over IP applications (IEEE, Phoenix, AZ, USA, 2000), pp. 307–313
9. T Chua, D Pheanis, QoS evaluation of sender-based loss-recovery techniques for VoIP. IEEE Netw. **20**(6), 14–22 (2006)
10. S Bruhn, H Pobloth, M Schnell, B Grill, J Gibbs, L Miao, K Jarvinen, L Laaksonen, N Harada, N Naka, S Ragot, S Proust, T Sanda, I Varga, C Greer, M Jelinek, M Xie, P Usai, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process (ICASSP)*. Standardization of the new 3GPP EVS codec (IEEE, South Brisbane, QLD, 2015), pp. 5703–5707
11. L Jeremie, V Tommy, B Stefan, S Hosang, P Ke, K Kei, W Bin, S Shaminda, F Julien, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process (ICASSP)*. Packet-loss concealment technology advances in EVS (IEEE, South Brisbane, QLD, 2015), pp. 5708–5712
12. J Suzuki, M Taka, Missing packet recovery techniques for low-bit-rate coded speech. IEEE J. Sel. Areas Commun. **7**(5), 707–717 (1989)
13. N Aoki, Modification of two-side pitch waveform replication technique for VoIP packet loss concealment. IEICE Trans. Commun. **E87-B**(4), 1041–1044 (2004)
14. J Lindblom, P Hedelin, in *Proc. IEEE Workshop Speech Coding*. Packet loss concealment based on sinusoidal modeling (IEEE, Ibaraki, Japan, 2002), pp. 65–67
15. E Gunduzhan, K Momtahan, A linear prediction based packet loss concealment algorithm for PCM coded speech. IEEE Trans. Speech Audio Process. **98**, 778–785 (2001)
16. MK Lee, SK Jung, HG Kang, YC Park, DH Youn, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process (ICASSP)*. A packet loss concealment algorithm based on time-scale modification for CELP-type speech coders (IEEE, Hong Kong, 2003), pp. 116–119
17. J Thyssen, R Zopf, J-H Chen, N Shetty, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process (ICASSP)*. A candidate for the ITU-T G.722 packet loss concealment standard (IEEE, Honolulu, HI, USA, 2007), pp. IV549–IV552
18. S Subasingha, M Murthi, S Andersen, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process (ICASSP)*. On GMM Kalman predictive coding of LSFS for packet loss (IEEE, Taipei, Taiwan, 2009), pp. 4105–4108
19. CA Rodbro, MN Murthi, SV Andersen, SH Jensen, Hidden Markov model-based packet loss concealment for voice over IP. IEEE Trans. Audio Speech Lang. Process. **14**(5), 1609–1623 (2006)
20. ITU-T Rec. G.722 Appendix III, *A high quality packet loss concealment algorithm for G.722*. (ITU-T, Geneva, 2006)
21. ITU-T Rec. G.722 Appendix IV, *A low-complexity algorithm for packet loss concealment with G.722*. (ITU-T, Geneva, 2006)
22. B Kovesi, S Ragot, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process (ICASSP)*. A low complexity packet loss concealment algorithm for ITU-T G.722 (IEEE, Las Vegas, NV, 2008), pp. 4769–4772
23. AV Aho, JV Ullman, JE Hopcroft, *Data Structures and Algorithms*. (Addisson-Wesley, NY, 1983)
24. B-K Lee, C Lim, J Park, J-H Chang, in *Proc. Interspeech*. Enhanced muting method in packet loss concealment of ITU-T G.722 employing optimized sigmoid function (ISCA, Lyon, France, 2013), pp. 3458–3462
25. ITU-T Rec. P.862.2, *Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs*. (ITU-T, Geneva, 2005)
26. Y Hu, P Loizou, Evaluation of objective quality measures for speech enhancement. IEEE Trans. Audio Speech Lang. Process. **16**(1), 229–238 (2008)
27. ITU-T Rec. P.800, *Methods for subjective determination of transmission quality*. (ITU-T, Geneva, 1996)
28. JS Garofolo, *Getting started with the DARPA TIMIT CD-ROM: an acoustic phonetic continuous speech database*. (NIST, Gaithersburg, 1993)
29. Multi-lingual speech database for telephonometry (1994). [Online]. Available: http://www.ntt-at.com/product/speech/. NTT Adv. Technol. Corp. Accessed 18 April 2016
30. ITU-T Rec. G.191, *Software tools for speech and audio coding standardization*. (ITU-T, Geneva, 2010)
31. S Lingfen, EC Ifeachor, in *Proc. IEEE Int. Conf. Communications (ICC)*. Perceived speech quality prediction for voice over IP-based networks (IEEE, NY, USA, 2002), pp. 2573–2577
32. S Quackenbush, T Barnwell, M Clements, *Objective Measures of Speech Quality*. (Prentice-Hall, Englewood Cliffs, NJ, 1988)