Enhanced Muting Method in Packet Loss Concealment of ITU-T G.722 Using Sigmoid Function with On-line Optimized Parameters

Bong-Ki Lee, Inyoung Hwang, Jihwan Park, and Joon-Hyuk Chang

Department of Electronics and Computer Engineering, Hanyang University, Seoul, Korea

bklee86@hanyang.ac.kr jchang@hanyang.ac.kr

10.21437/Interspeech.2014-590

Abstract

In this paper, we propose an enhanced adaptive muting method using a sigmoid function, which is based on a parameter tracking technique for the packet loss concealment algorithm of ITU-T G.722 speech codec. The packet loss concealment algorithm performs an adaptive muting to prevent the generation of unnecessary noises or clicks during packet loss recovery. While a conventional muting method applies the sigmoid function to the muting curve and the principal parameters of the sigmoid function are obtained by using a grid search-based training method, in the proposed muting algorithm, the parameters are substantially obtained from the previous good frames using the steepest descent algorithm, which minimizes the error between the desired signal and the reconstructed signal. From experimental results, the proposed adaptive muting method turns out to improve the performance of the conventional muting method under various experimental conditions.

Index Terms: VoIP, ITU-T G.722, packet loss concealment, adaptive muting, sigmoid function, steepest descent.

1. Introduction

Packet loss concealment (PLC), also known as frame erasure concealment, technology is essential to voice over packet switched networks such as an internet protocol (IP) network. In voice over IP (VoIP) network, the voice signal is packetized at the sender side with regular frame size (e.g., 10 ms) using an encoding algorithm such as ITU-T G.722, G.729, and adaptive multi-rate (AMR) speech codec [1]. The voice packet is then sent over the IP network to the receiver side where it is decoded. During the transmission, IP packets may be lost due to the delay and jitter so that the quality of service (QoS) cannot be guaranteed [2]. Since the voice quality is substantially impaired when a packet loss rate (PLR) is greater than 5%, a relevant PLC algorithm, which reconstructs missing frames, is essential for VoIP applications [3].

One of the standard coding algorithm, ITU-T G.722 speech codec [4] was revised to recommend standard PLC algorithms by adding Appendix III [5] and IV [6]. 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. In the G.722 Appendix IV algorithm, which is a linear prediction (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 involves clicks or unpleasant noises, especially in the case of consecutive packet losses (i.e., burst error), the PLC algorithm includes a method for adaptive muting at the end of the reconstructed frames. The pre-reconstructed speech signal is multiplied by the pre-defined adaptive muting factor and this muting factor is gradually decreased as more consecutive packet losses occur. Also, the muting is applied differently according to the class of the signal using a pre-determined fixed curve.

Recently, an adaptive muting method using a sigmoid function to determine the adaptive muting curve was developed [7] in which the shape of the sigmoid function is determined by core parameters chosen to minimize the error between the desired signal and the reconstructed signal during the training phase. In training, the grid search technique was employed to determine optimal values of the parameters within the search space in the off-line process. The optimized sigmoid function is then applied to the muting curve to enhance the quality of the reconstructed speech signal. Although this method improves the speech quality without introducing additional algorithmic delay and redundancies, the values of parameters obtained from training dominantly depend on the training database and are not changed after they are determined once so that they cannot characterize short-time variation in speech.

In this paper, we propose an enhanced adaptive muting method using the sigmoid function, which is determined based on the steepest descent criterion. The parameters of the sigmoid function obtained from previous good frames using the steepest descent algorithm [8], which minimizes the error between the desired signal and the reconstructed signal, are used for the current missing frame for adaptive muting. Main premise behind our technique is that the reconstruction is performed to find optimal parameters of the sigmoid function even in each good frame even though the packet loss does not occur during good frames. Notice that the core parameters of the sigmoid function are changing in time adaptively according to error minimization criterion and the training process is unnecessary, which indicates they do not depend on the established database. According to the simulation results, it is found that the proposed method outperforms the original muting method in G.722 Appendix IV and conventional muting method [7] in terms of various speech quality measures.

The rest of the paper is organized as follows: Section 2 briefly reviews the previous adaptive muting methods including the original method in G.722 App. IV and conventional method in [7], and Section 3 describes the proposed adaptive muting method. After simulation results are presented in Section 4, the paper is concluded in Section 5.

2. Review of Previous Methods

Since the proposed algorithm is mainly based on ITU-T G.722, we briefly review the previous PLC works. The PLC algorithm stated in Appendix IV of ITU-T G.722 [6] corresponds to a

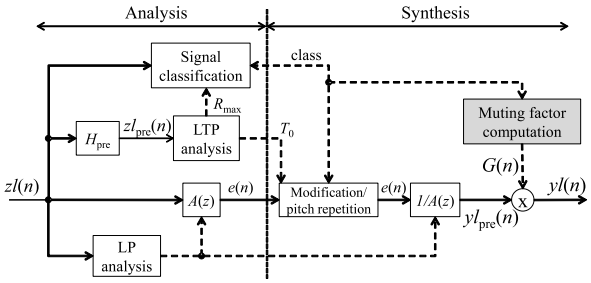

Figure 1: Lower-band LPC-based pitch repetition of G.722 decoder with the PLC algorithm

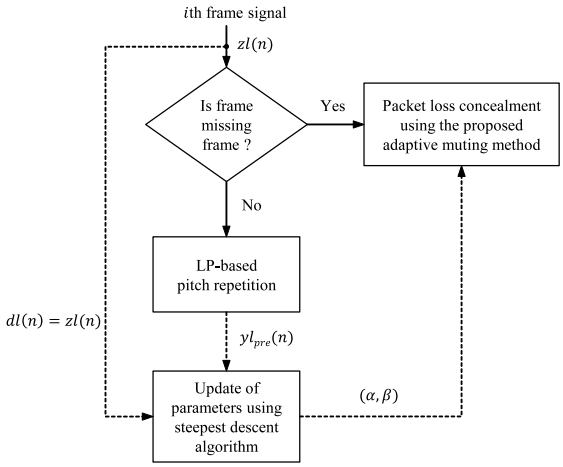

Figure 2: Overall flow chart of the proposed adaptive muting method

receiver-based scheme as introduced in Section 1, where the information is originated from the packet previously received. The encoder is thus not necessary to be modified, but the decoder is slightly changed by adding a PLC mechanism. Note that the terms "frame" and "packet" are used interchangeably in this paper. 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). Since the operation in the higher band is included in that of the lower band, we focus in this paper on describing the operation of the PLC algorithm at the lower band only. For easy comprehension, Fig. 1 is prepared to show the lower-band LPC-based pitch repetition block diagram of the G.722 decoder incorporating the PLC algorithm [6]. At first, when packet loss occurs, the reconstructed lower-band signal $yl(n)$ is extrapolated through the LPC-based pitch repetition block using the past valid lower-band signal $zl(n)$. Specifically, the pre-reconstructed lower-band signal $yl_{\text{pre}}(n)$, which is prior to the adaptive muting, is synthesized by using $zl(n)$. Therefore, unpleasant noises or clicks are generated especially in consecutive packet losses if $yl_{\text{pre}}(n)$ is used directly. Accordingly, the adaptive muting method is devised in the final step of the PLC algorithm to reduce the effect of the unpleasant noises or clicks. 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. In the original muting method [6], the adaptive muting factor is differently applied according to the class of the signal determined in the G.722 decoder. 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, which are the superior candidates for extrapolation because the perceived quality of reconstructed speech largely depends on this type of the signal [6]. Furthermore, 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 an unpleasant noises or clicks when more than four packets are missed.

Our recent algorithm [7] applies the two-parameter sigmoid function to the muting curve to offer more flexibility to the shape of the muting curve than that of the original muting curve, which is the linear and discontinuous. Optimal values of the parameters of the sigmoid function are then selected according to the grid search, which is a simple exhaustive search method to find an optimal point in a manually specified parameter space according to a given error criterion. The muting curve of the conventional method is applied linearly and continuously between successive frames with a higher freedom given in [7]. It is noted that the *other cases* class has the dominant impact on the quality of reconstructed speech among the classes of the signal, the sigmoid function as the muting curve is only used for *other cases* class to reduce the training time. Although this method effectively implements the adaptive muting so that speech quality substantially is improved without introducing additional algorithmic delay and redundancies, the principal parameters of the sigmoid function are dominantly affected by the speech database since they are obtained from the grid search-based training process. Furthermore, the parameters will not be changed after they are determined once and thus these may not be adequate for the short-time evolution of speech. Note that, these adaptive muting methods are inherently applied to the higher band in the same manner of the lower band. As a result, 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.

3. Proposed Adaptive Muting Method

We present an enhanced muting algorithm, which uses the sigmoid function incorporating the parameter tracking technique via the steepest descent criterion [8]. The parameters of the sigmoid function are updated by using the steepest descent method at the previous good frames by assuming intentionally the good frame as the missing frame. Subsequently, the parameters of the sigmoid function obtained from the previous good frames are used for the adaptive muting mechanism when packet loss occurs at the current frame. As shown in Fig. 2, the proposed adaptive muting method can be explained through the flow chart. If the signals at ith frame $zl(n)$ ($0 \leq n < 80$) is firstly good frame, it passes through the LP-based pitch repetition block where the lower-band pre-reconstructed signal $yl_{\text{pre}}(n)$ can be obtained as shown in Fig. 1. Since the signals at ith frame $zl(n)$ is the good frame, it is able to regard the signal $zl(n)$ as the desired signal $dl(n)$. Therefore, the parameters of

the sigmoid function can be updated by using the steepest descent criterion which minimizes the error between the desired signal $dl(n)$ and reconstructed signal $yl(n)$. On the other hand, if the signals at ith frame $zl(n)$ belongs to the missing frame the lower-band reconstructed signal $yl(n)$ is extrapolated through the PLC algorithm with the proposed adaptive muting method using the parameters of the sigmoid function updated at the previous good frame. For the muting curve, we adopt the following two-parameter sigmoid function such that

$$G(n) = \frac{1+\alpha e^{-\beta n_0}}{1+\alpha e^{\beta(n-n_0)}} \quad , \quad 0 \le n < 320 \tag{2}$$

where α and β denote sloping parameters of the sigmoid function, and n_0 means an offset, respectively. It is noted that $G(0) = 1$. Also, $G(n)$ becomes zero after 320 samples in common with the previous methods [6], [7]. This function, which has non-linear and continuous characteristics, can offer more flexibility to the shape of muting curve than the that of the original muting curve.

We then perform the update mechanism for the parameters α and β in order for the proposed adaptive muting curve to be a best model for the desired signal. If the every good frame before packet loss occurs is assumed intentionally as a missing frame, the desired signal is equal to the received signal at a good frame and the pre-reconstructed signal yl_{pre} is obtained from the LP-based pitch repetition. Since the desired signal is known, the estimation error $e(n)$ of the good frame can be expressed as

$$\begin{aligned} e(n) &= dl(n) - yl(n) \\ &= dl(n) - G(n)\cdot yl_{\text{pre}}(n) \\ &= dl(n) - \frac{1+\alpha e^{-\beta n_0}}{1+\alpha e^{\beta(n-n_0)}}\cdot yl_{\text{pre}}(n) \end{aligned} \tag{3}$$

where $dl(n)$ and $yl(n)$ denote the lower-band desired signal and the lower-band reconstructed signal, respectively. The cost function in terms of the squared error between desired signal and reconstructed signal incorporating the muting curve can be expressed as

$$\begin{aligned} J(\alpha,\beta) &= \big[e(n)\big]^2 \\ &= \left[dl(n) - \frac{1+\alpha e^{-\beta n_0}}{1+\alpha e^{\beta(n-n_0)}}\cdot yl_{\text{pre}}(n)\right]^2. \end{aligned} \tag{4}$$

Since (4) contains two unknowns, i.e., α, β, we apply the steepest descent criterion to minimize (4) for the two parameters. Thus, the following update formulas for α and β with the factors for controlling step size; μ_α and μ_β is obtained:

$$\begin{aligned} \alpha_{n+1} &= \alpha_n - \frac{\mu_\alpha}{2}\cdot\frac{\partial J(\alpha,\beta)}{\partial\alpha}\bigg|_{\alpha=\alpha_n} \\ &= \alpha_n + \mu_\alpha\cdot e(n)\cdot yl_{pre}(n)\cdot\frac{\partial G(\alpha,\beta)}{\partial\alpha}\bigg|_{\alpha=\alpha_n} \end{aligned} \tag{5}$$

where $\partial G(\alpha,\beta)/\partial\alpha$ is shown in (6)

$$\frac{\partial G(\alpha,\beta)}{\partial\alpha} = \frac{1}{e^{\beta n_0}\big(\alpha e^{\beta(n-n_0)}+1\big)} - \frac{e^{\beta(n-n_0)}\Big(\frac{\alpha}{e^{\beta n_0}}+1\Big)}{\big(\alpha e^{(n-n_0)}+1\big)^2} \tag{6}$$

and

$$\begin{aligned} \beta_{n+1} &= \beta_n - \frac{\mu_\beta}{2}\cdot\frac{\partial J(\alpha,\beta)}{\partial\beta}\bigg|_{\beta=\beta_n} \\ &= \beta_n + \mu_\beta\cdot e(n)\cdot yl_{pre}(n)\cdot\frac{\partial G(\alpha,\beta)}{\partial\beta}\bigg|_{\beta=\beta_n} \end{aligned} \tag{7}$$

In time, $\partial G(\alpha,\beta)/\partial\beta$ is given by

$$\begin{aligned} \frac{\partial G(\alpha,\beta)}{\partial\beta} &= \frac{-\alpha n_0}{e^{\beta n_0}\big(\alpha e^{\beta(n-n_0)}+1\big)} \\ &\quad - \frac{\alpha e^{\beta(n-n_0)}\Big(\frac{\alpha}{e^{\beta n_0}}+1\Big)(n-n_0)}{\big(\alpha e^{\beta(n-n_0)}+1\big)^2}. \end{aligned} \tag{8}$$

Note that the above iterations for α and β are performed in an on-line fashion, i.e., sample-by-sample basis, which implies for each input sample, α and β are updated once for each input sample. When α is updated, the value of β from the last iteration is used as required by (5) and (7), respectively. Since all the updates are given in an explicit form, the computation of α and β can be easily implemented, and these updated parameters are used for the adaptive muting mechanism when packet loss occurs. To simulate the proposed muting method using the steepest descent criterion without causing computation overflow, n_0 is set to be 150 which was determined based on the grid search method [7], and the values of α and β are limited to be $0.1 \le \alpha \le 1.0$ and $0.01 \le \beta \le 1.00$ by considering the reasonable shape of the sigmoid function, the slope of which is not rapidly decreasing or not zero. It is noted that the values of parameters of the sigmoid function are changed adaptively with the error minimization criterion so that the training process turns out to be unnecessary. Also, the proposed method can be applied to all signal class types, whereas the conventional method [7] is applied to the *other cases* class only.

4. Experiments and Results

To verify the performance of our proposed algorithm, we compared the proposed method with the original muting method in G.722 Appendix IV and the conventional muting method in [7] in condition of various PLRs. In addition, to illustrate the importance of muting mechanisms in VoIP applications, we analyzed the waveform results obtained from without the muting scheme. For these experiments, we chose 100 phrases spoken in Korean by four male and four female speakers from the NTT speech database [9]. In the algorithm we implemented, the speech phrases was sampled at 16 kHz and random packet (frame) losses were inserted at various rates by using the error insertion device in ITU-T G.191 software tool with zero bit error rate and 0.5 of burst factor [10]. Also, these phrases were separately partitioned into 30 percent used as test and 70 percent used as training for implementing the conventional method.

Fig. 3 shows that the waveforms of the desired speech signal, which is decoded without any packet losses and the decoded speech signal using the methods without muting scheme, the original muting method, the conventional muting method, and the proposed muting method during the packet loss period from 0.01 sec to 0.05 sec. In Fig. 3(a), the same waveform was repeated in the consecutive packet losses, possibly leading to perceptually annoying artifacts. In Fig. 3(b), the waveform was over-muted, which causes the degradation of speech quality. On the other hand, the waveform of Fig. 3(d) was much more similar to the desired speech signal in comparison to the other. Specifically, the waveform of the proposed muting method in Fig. 3(d) was slightly better in fitting to the desired speech signal than it of the conventional muting method depicted in Fig. 3(c), which implies the parameters of sigmoid function obtained from the previous good frame are better modeling for the muting curve than that obtained from training process incorporating the grid search.

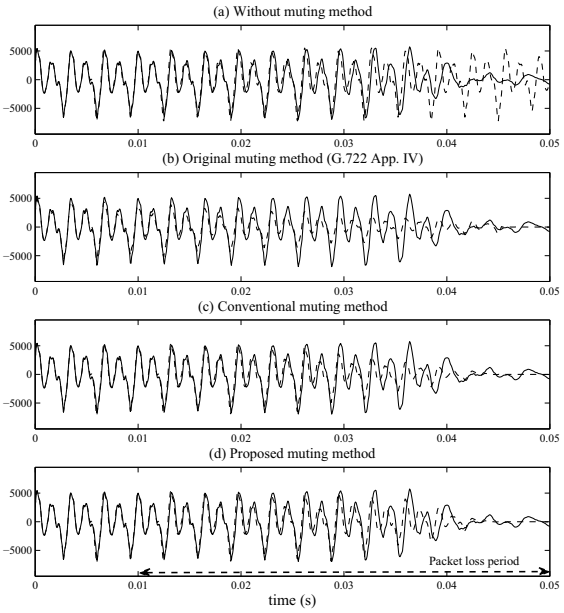

Figure 3: Waveform comparison (solid: desired speech signal, dash: decoded speech signal) of the (a) without muting method, (b) original muting method (G.722 App. IV), (c) conventional muting method [7], and (d) proposed muting method during the packet loss period (from 0.01 sec to 0.05 sec)

The above methods were evaluated with objective speech quality measures including segmental signal-to-noise ratio (SNR), perceptual evaluation of speech quality (PESQ), and composite measure (C_{ovl}) [11]. As Table 1 summarizes the overall results, it is seen that the proposed method outperformed the existing approaches in terms of the segmental SNR, PESQ, and composite measure (C_{ovl}) over various PLRs. This implies that the proposed algorithm using the on-line optimized parameters is better in modeling the sigmoid function-based muting curve. In addition, to validate the objective quality tests, we performed a subjective quality test, namely the mean opinion score (MOS) [12]. For this subjective test, ten Korean listeners with normal hearing score their respective subjective opinions on the quality of each sentence, using one of the following points: 5 (Excellent), 4 (Good), 3 (Fair), 2 (Poor), and 1 (Bad). The results of the subjective quality test were similar to those of the objective quality test as shown in Table 1. A comparison of overall simulation results shows that the proposed muting method yielded the improved quality compared to the original and conventional muting methods. In particular, the performance gain increased as the PLR increased. This observation confirmed the robust performance of the proposed algorithm in various network conditions.

5. Conclusions

In this paper, we proposed an enhanced adaptive muting algorithm, a part of ITU-T G.722 Appendix IV, using the sigmoid function. The principal contribution of this paper is minimizing the error between the desired speech signal and the reconstructed signal using the sigmoid function in the process of determining on the adaptive muting factor. To obtain the parameters of the sigmoid function, the steepest descent criterion is applied to the previous good frame and these parameters are used for adaptive muting when packet loss actually occurs. The performance of the proposed approach has been found superior to that of the previous methods through the extensive experiment results.

Table 1: Comparison of speech quality measure test in various packet loss environments (95% confidence interval)

PLR	Quality measure	Method		
		Original [6]	Conventional [7]	Proposed
3%	Seg. SNR	24.37±0.01	24.39±0.01	**24.42±0.01**
	PESQ	3.35±0.04	3.40±0.05	**3.42±0.04**
	C_{ovl}	3.69±0.05	3.75±0.04	**3.80±0.05**
	MOS	3.39±0.05	3.43±0.06	**3.45±0.06**
6%	Seg. SNR	21.08±0.01	21.11±0.02	**21.14±0.01**
	PESQ	3.01±0.04	3.09±0.04	**3.13±0.03**
	C_{ovl}	3.34±0.05	3.51±0.06	**3.57±0.06**
	MOS	3.13±0.06	3.17±0.08	**3.19±0.07**
10%	Seg. SNR	17.73±0.01	17.78±0.02	**17.83±0.02**
	PESQ	2.76±0.04	2.87±0.05	**2.92±0.04**
	C_{ovl}	2.86±0.08	3.12±0.06	**3.21±0.06**
	MOS	2.64±0.04	2.72±0.08	**2.80±0.06**
15%	Seg. SNR	13.21±0.02	13.29±0.02	**13.36±0.02**
	PESQ	2.32±0.03	2.47±0.04	**2.55±0.04**
	C_{ovl}	2.41±0.04	2.63±0.05	**2.76±0.03**
	MOS	2.16±0.09	2.27±0.08	**2.36±0.06**

6. Acknowledgements

This research was supported by National Research Foundation of Korea (NRF) grant funded by the Korean Government (MEST) (2012R1A2A2A01004895). This research was also supported by the MSIP (Ministry of Science, ICT & Future Planning), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2013-H0301-13-4005). This work was supported by Ministry of Science, ICT (Information and Communication Technologies) and Future Planning by the Korean Government (NRF-2011K2A2A6A00002)

7. References

[1] S. Karapantazis and F. Pavlidou, "VoIP: A comprehensive survey on a promising technology," *Computer Networks*, vol. 53, no. 12, pp. 2050-2090, Aug. 2009.

[2] A. Takahashi, H. Yoshino, and N. Kitawaki, "Perceptual QoS assessment technologies for VoIP," *IEEE Communications Magazine*, vol. 42, no. 7, pp. 28-34, Jul. 2004.

[3] N. Jayant and S. Christensen, "Effect of packet losses in waveform coded speech and improvement due to an add-even sample-interpolation procedure," *IEEE Trans. Communications*, vol. 29, pp. 101-109, Feb. 1981.

[4] ITU-T G.722, "7 kHz audio coding within 64 kbit/s," Nov. 1998.

[5] ITU-T Rec. G.722 Appendix III, "A high quality packet loss concealment algorithm for G.722," Nov. 2006.

[6] ITU-T Rec. G.722 Appendix IV, "A low-complexity algorithm for packet loss concealment with G.722," Nov. 2006.

[7] B.-K. Lee, C. Lim, J. Park, and J.-H. Chang, "Enhanced muting method in packet loss concealment of ITU-T G.722 employing optimized sigmoid function," in *Proc. Interspeech*, pp. 3458-3462, Aug. 2013.

[8] B. Widrow and S. Stearns, Adaptive Signal Processing, pp. 46-65, Prentice-Hall, 1985.

[9] S.-W. Yoon, H.-G. Kang, Y.-C. Park, and D.-H. Yoon, "An efficient transcoding algorithm for G.723.1 and G.729A speech coders: interoperability between mobile and IP network," *Speech Communication*, vol. 43, pp. 17-31, Jun. 2004.

[10] ITU-T Rec. G.191, "Software tools for speech and audio coding standardization," Mar. 2010.

[11] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio Speech and Language Processing*, vol. 16, no. 1, pp. 229-238, Jan. 2008.

[12] ITU-T Rec. P.800, "Methods for subjective determination of transmission quality," Jun. 1998.