Enhanced Muting Method in Packet Loss Concealment of ITU-T G.722 Employing Optimized Sigmoid Function

Bong-Ki Lee, Chungsoo Lim, Jihwan Park, Joon-Hyuk Chang

Department of Electronics Computer Engineering
Hanyang University, Seoul, Korea

bklee86@hanyang.ac.kr, chungsoo.lim@gmail.com, pjh0410v@hanyang.ac.kr, jchang@hanyang.ac.kr

10.21437/Interspeech.2013-757

Abstract

In this paper, we propose an improved adaptive muting method using a sigmoid function for the packet loss concealment algorithm of ITU-T G.722 Recommendation. The packet loss concealment algorithm performs an adaptive muting to prevent the generation of unnecessary noise during packet loss recovery. While muting is linearly and discontinuously performed according to packet errors, our muting approach is performed by the non-linear and continuous sigmoid function. The principal parameters of the sigmoid function are obtained based on training at which minimization between the desired signal and the reconstructed signal is performed. Experimental results show that this proposed muting technique can enhance the performance of the packet loss concealment algorithm of G.722 under various packet loss environments.

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

1. Introduction

In recent years, there has been a growing interest in voice over internet protocol (VoIP) services with increased demand for voice communication through the internet network [1]. VoIP applications perform a packet-based voice communication over IP network, operating with standard codecs such as ITU-T G.722, G.729, and G.723.1. Among them, we focus on the ITU-T G.722 coder, which is known to be high quality, low delay, and low complexity speech data encoding scheme [2]. Specifically, the ITU-T G.722 speech codec, which is a technique for compressing speech data below 64 kbps and high quality audio signals with 50-7000 Hz wideband, has been adopted as a standard by ITU-T.

However, VoIP applications still have some limitations regarding voice quality compared to some traditional technologies such as public switched telephone networks (PSTN). One of the major problems can be packet loss due to the delay and jitter in the process of transmitting speech data, whereby the quality of service (QoS) cannot be guaranteed [3]. Therefore, a packet loss concealment (PLC) algorithm, also known as frame erasure concealment algorithm, which extrapolates missing frames, is needed for VoIP applications in a packet loss environment [4]. Most of the standard speech codecs used in VoIP applications employ their own PLC algorithms to solve the problem of degraded speech quality due to the packet loss [5]. PLC algorithms are classified into sender-receiver based schemes and receiver based schemes. Sender-receiver based schemes include sending duplicate packets, or sending error correction bits in voice packets using forward error correction (FEC). In receiver-based schemes, lost packets are recreated by padding silence, by repeating the last received packet, or by performing waveform substitution based on previously received packets on each sub-band of linear-prediction (LP) residues.

PLC algorithms of ITU-T G.722 were standardized in 2006 as Appendix III and IV of ITU-T G.722 [6], [7]. The PLC algorithm described in Appendix IV is a receiver based scheme and meets the same quality requirements as the PLC algorithm in Appendix III, but with a lower complexity. In Appendix IV, the lost packets are extrapolated by using an information of previously received packets such as the LP coefficient (LPC), signal classification, and pitch period. Since the reconstruction of the missing frames causes an unnecessary noise or click sound especially in the case of consecutive packet losses (i.e., burst error), an adaptive muting method is used at the end of packet loss concealment. An adaptive muting factor with a value between 1 and 0 is multiplied by a pre-reconstructed speech signal, and as more consecutive packet losses occur, the muting factor is adaptively adjusted to a smaller value. This is applied differently depending on the class of the signal and applied linearly by using a pre-determined fixed curve.

In this paper, we present an improved adaptive muting method using a sigmoid function [8] to determine the adaptive muting curve. Two major parameters, which determine the shape of the sigmoid function, are chosen to be optimized values based on the error minimization between the desired signal and the reconstructed signal by using the sigmoid function. For training, the grid search technique [9] is used to find optimal values of the parameters within the search space especially for voiced, weakly voiced and unvoiced sound regions. The sigmoid function is then applied in the muting algorithm on missing frames for quality enhancement. Experimental results show that the proposed method outperforms the original adaptive muting method in terms of various speech quality measures. The rest of the paper is organized as follows. Section 2 briefly reviews the adaptive muting method, 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 G.722 Appendix IV

The PLC algorithm described in Appendix IV of ITU-T G.722 is a receiver based scheme, which uses an information of previously received packet in the receiver. Therefore, there is no change to the encoder, but the packet loss concealment mechanism is added to the decoder. Note that the terms “frame” and “packet” are considered to be equivalent in this paper. As the block diagram of the G.722 decoder with the PLC algorithm is shown in Fig. 1, the decoder includes additional blocks for the packet loss concealment, shown as the grey-shaded blocks

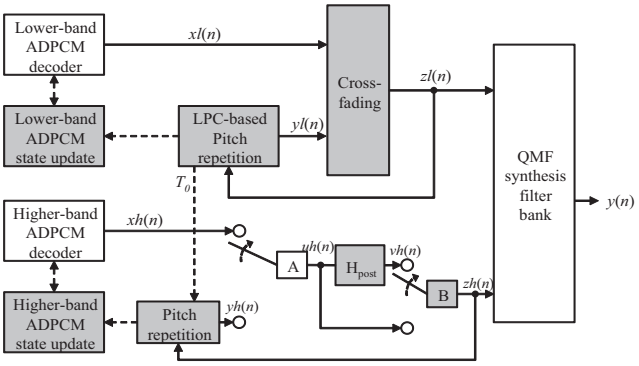

Figure 1: Block diagram of G.722 decoder with the PLC algorithm

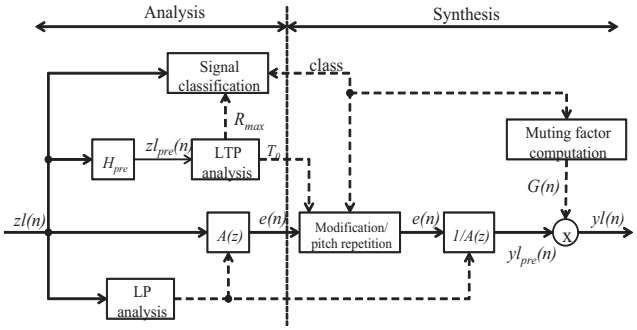

Figure 2: Lower-band LPC-based pitch repetition of G.722 decoder with the PLC algorithm

[7]. ITU-T G.722 codec uses the sub-band adaptive differential pulse code modulation (SB-ADPCM). In the SB-ADPCM, the frequency band is split into two sub-bands which are lower-band and higher-band. In the PLC algorithm, the operation of the higher-band is included in that of the lower-band. Therefore, we only describe the operation of the PLC algorithm based on the lower-band in this paper. First, reconstructed lower-band signal $yl(n)$ is extrapolated when packet loss occurs through the LPC-based pitch repetition block using the past valid lower-band signal $zl(n)$. After extrapolating the $yl(n)$, $xl(n)$ and $yl(n)$ are cross-faded.

Actually, Fig. 2 shows the lower-band LPC-based pitch repetition block diagram of G.722 decoder with the PLC algorithm [7]. In this figure, pre-reconstructed lower-band signal $yl_{pre}(n)$, which is prior to the adaptive muting, is synthesized by using $zl(n)$. However, an uncomfortable noise or click sound is generated especially for consecutive packet losses if the $yl_{pre}(n)$ is used directly. Therefore, the adaptive muting method is employed at the end of the PLC algorithm to reduce the effect of the uncomfortable noise or click sound. If we consider the adaptive muting mechanism, the reconstructed lower-band signal $yl(n)$ is represented as

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

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

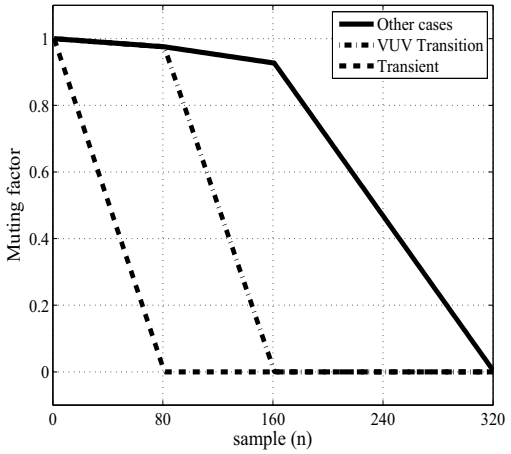

Figure 3: Muting factor according to three classes of the signal

Table 1: Adaptive muting parameters

Parameter	Speech class type		
	Transient	VUV Transition	Other cases
fac1	409	10	10
fac2p	409	10	20
fac3p	409	399	190

Lastly, the adaptive muting factor is applied differently depending on the class of the signal offered by G.722 decoder as shown in Fig. 3. While *transient* and *VUV transition* classes correspond to a transient period with large energy variation and a transition between voiced and unvoiced signals, respectively, *other cases* class includes unvoiced, weakly voiced, and voiced signal, which are the best candidates for extrapolation because the quality of reconstructed speech is dominantly affected by this type of signal [7]. Furthermore, the adaptive muting factor becomes zero after 320 samples (four packets in lower-band), which means silence, so that it prevents the generation of an uncomfortable noise or click sound when more than four packets are lost. As the parameters of the original adaptive muting method are shown in Table 1, those parameters depend on the class of signal. The adaptive muting factor is derived such that

$$G(n+1) = \begin{cases} G(n) - fac1 & , \ 0 \le n < 80 \\ G(n) - fac2p & , \ 80 \le n < 160 \\ G(n) - fac3p & , \ 160 \le n < 320 \\ 0 & , \ n \ge 320 \end{cases} \tag{2}$$

where $G(0) = 32768$. It is noted that since real systems use 16-bit signed integers, the value of one is represented as $2^{15} = 32768$ [7]. In (2), the adaptive muting factor is adapted sample-by-sample with *fac1* for the first lost packet, *fac2p* for the second lost packet, and *fac3p* for third and fourth lost packets. Moreover, the adaptive muting factor becomes zero after fourth lost packet. The original adaptive muting is applied linearly to each packet as stated above. This adaptive muting method is applied to the higher-band in the same way as to the lower-band. Finally, the reconstructed signals of the lower-band and higher-band are combined into wideband decoded signal $y(n)$ through the quadrature mirror filter (QMF) synthesis filterbank as shown in Fig. 1.

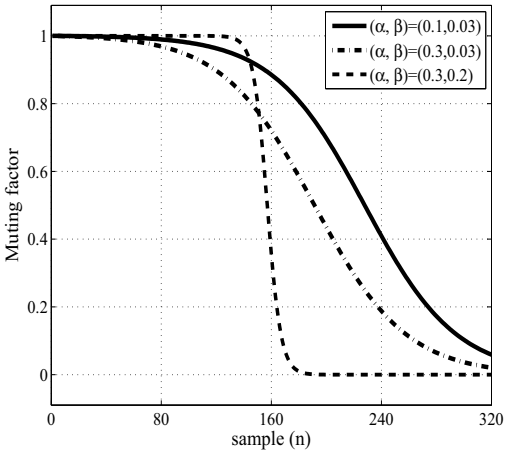

Figure 4: Three examples of the sigmoid function

3. Proposed Adaptive Muting Method

As explained in Sec. 2, the original adaptive muting method in Appendix IV of ITU-T G.722 is applied linearly and discontinuously between frames using the pre-determined curve to each packet. In this paper, we present an improved adaptive muting method which is applied non-linearly and continuously using sigmoid function. Indeed, optimal values of the parameters of the sigmoid function are selected according to the grid search [9], which is a simple exhaustive search method through a manually specified subset of the parameter space of a learning algorithm, guided by an error criterion. Specifically, we adopt the following two-parameter sigmoid function [8] such that

$$G(n) = \frac{1 + \alpha e^{-n_0 \beta}}{1 + \alpha e^{\beta(n - n_0)}} \quad , \quad 0 \leq n < 320 \qquad (3)$$

where α and β denote sloping parameters of the sigmoid function, and n_0 means an offset. Also, $G(n)$ becomes zero after 320 samples in common with the original method. This function, which has non-linear and continuous characteristics, can give more flexibility to the muting curve. Actually, n_0 is set to be 150, and the values of α and β are reasonably limited to be $0.1 \leq \alpha \leq 1.0$ and $0.01 \leq \beta \leq 0.20$, considering the reasonable shape of the sigmoid function, the slope of which is not rapidly decreasing or not zero. These search space makes it possible to prevent to the computation overflow when finding the optimal values of parameters of the sigmoid function. It is seen that the shape of the sigmoid function depends on α and β, as illustrated in Fig. 4 showing three example functions characterized by (3). The outputs of the three examples sigmoid functions presented in Fig. 4 are decreased with different speeds, which can be controlled by α and β. Since the *other cases* class has the greatest impact on the quality of reconstructed speech among the classes of the signal, we control the adaptive muting factor in *other cases* class only. Therefore, (3) is applied to the *other cases* class, while (2) is applied to the *Transient* and *VUV transition* classes. Using (1) and (3), the error between the desired signal and the reconstructed signal can be expressed as

$$\begin{aligned} \varepsilon(n) &= dl(n) - yl(n) \\ &= dl(n) - G(n) \cdot yl_{pre}(n) \\ &= dl(n) - \frac{1 + \alpha e^{-n_0 \beta}}{1 + \alpha e^{\beta(n - n_0)}} \cdot yl_{pre}(n) \end{aligned} \qquad (4)$$

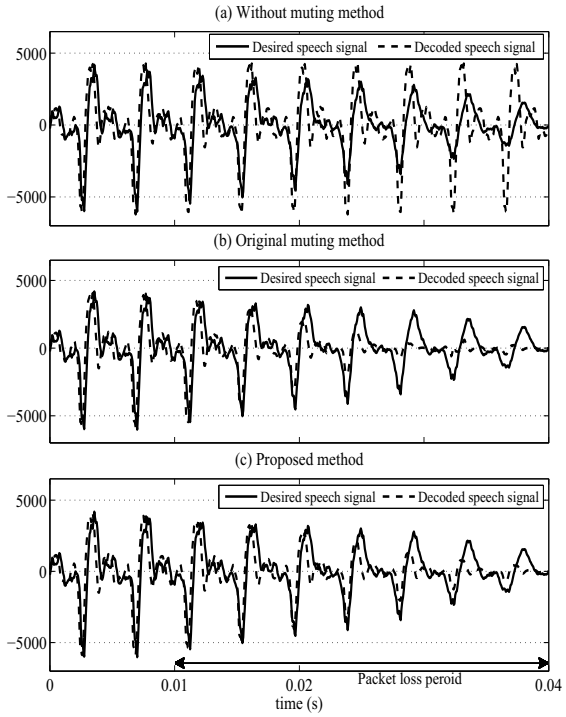

Figure 5: Waveform comparison between the without muting method, original method, and proposed method in terms of the desired speech signal and decoded speech signal during the packet loss period (from 0.01 sec to 0.04 sec)

Table 2: Optimal Points for Various Packet Loss Rates

Packet Loss Rate	Optimal Points	
	α	β
1%	0.90	0.02
5%	0.90	0.02
10%	0.90	0.02

where $dl(n)$ denotes the lower-band desired signal, which is decoded without any packet losses. Note that the cost function in (4) contains two unknowns, i.e., α and β, and then can be expressed as a function of α and β. From (4), the average of the mean square errors (MSEs) of all training files is expressed as a function of α and β.

$$\xi(\alpha, \beta) = \frac{1}{m} \sum_{k=1}^{m} E[\varepsilon_k^2(n)] \qquad (5)$$

where k and m denote training file and the total number of the training files for the grid search according to the processed speech by the proposed PLC algorithm, respectively.

$$(\hat{\alpha}, \hat{\beta}) = \arg \min_{0.1 \leq \alpha \leq 1.0, 0.01 \leq \beta \leq 0.20} \xi(\alpha, \beta) \qquad (6)$$

Based on (6), the optimal parameters, $\hat{\alpha}$ and $\hat{\beta}$, are obtained which minimize $\xi(\alpha, \beta)$. In a training step, to find the optimal parameters, we compute the average of MSEs over all training data in the speech materials with varying α and β. For this sigmoid parameter training, we used a number of the speech materials from the NTT database will be described in Sec. 4. Finally,

Table 3: Comparison of experimental results

Packet Loss Rate	Without Muting				G.722 App. IV				Proposed			
	SNR	Seg. SNR	PESQ	MOS	SNR	Seg. SNR	PESQ	MOS	SNR	Seg. SNR	PESQ	MOS
1%	21.147	29.654	3.968	3.731	21.173	29.896	3.974	3.853	**21.529**	**29.999**	**3.986**	**3.876**
5%	9.241	17.478	3.009	2.877	9.483	17.697	3.106	3.184	**9.858**	**17.926**	**3.126**	**3.241**
10%	5.656	10.788	2.313	2.273	6.132	11.354	2.639	2.651	**6.665**	**11.808**	**2.677**	**2.774**

we obtained the optimal points $(\hat{\alpha}, \hat{\beta})$ for various packet loss rates as shown in Table 2, and these are adopted into (3). It is noted that the same optimal points are obtained for three packet loss rates, meaning the optimal points are not sensitive to the packet loss rate. Also, this proposed adaptive muting method is applied to the higher-band in the same way as to the lower-band. As an example, Fig. 5 shows that the decoded speech signal using the proposed adaptive muting method is the most similar to the desired speech signal in comparison to the original adaptive muting method and without muting method.

4. Experiments and Results

To demonstrate the effectiveness of our proposed method, we compared the proposed method with original adaptive muting method in G.722 Appendix IV. In addition, we checked the results obtained without any muting mechanism in order to illustrate the importance of muting mechanisms in VoIP applications. For the experiments, we selected one hundred speech files spoken by four male and four female speakers from the NTT Korean speech database [10], and these data files were then partitioned into 30 percent test data files and 70 percent training data files without any overlap. Each file included two different meaningful sentences and the whole length of each file was 8 sec. In the algorithm we implemented, the speech data was sampled at 16 kHz and random packet losses were inserted at various rates with 10 ms by using error insertion device (EID) in ITU-T G.191 software tool [11].

The above methods were evaluated with objective speech quality measures including signal to noise ratio (SNR), segmental SNR [12], and wideband perceptual evaluation of speech quality (WPESQ) [13]. In addition, we performed a subjective quality test, which is called the 5-scale absolute category rating (ACR) and mean opinion score (MOS) [14] in order to validate the objective evaluation. For this test, subjective opinions were given by a group of ten Korean listeners (mean = 33.2 years) with normal hearing, where each listener determines one of the following scores for each test sentence: 5 (Excellent), 4 (Good), 3 (Fair), 2 (Poor), and 1 (Bad). The experimental results are shown in Table 3. As can be seen, without muting method yields the worst performance. This ensures that the muting technique is needed for the PLC algorithm. Summarizing all results as shown in Table 3, we see that the proposed method outperforms the other approaches in terms of SNR, segmental SNR, PESQ, and MOS. Also, the results of subjective quality test are coincident with that of objective quality test. It is evident that the proposed method ensures better speech quality than the original method. In particular, it is observed that performance gain becomes larger as the packet loss rate gets higher. This observation confirmed the superiority of the proposed algorithm at diverse network condition.

5. Conclusions

In this paper, we proposed an improved adaptive muting method using a sigmoid function for ITU-T G.722 Appendix IV using the sigmoid function. The principal contribution of this paper is the reduction of the error between the desired signal and the reconstructed signal by using the well-defined sigmoid function when packet losses occur. We selected optimal values of the two-parameter of the sigmoid function using on the grid search-based training step and applied the sigmoid function to the muting algorithm. The performance of the proposed approach has been found to be superior to that of the original method through the extensive objective and subjective quality tests.

6. Acknowledgements

This work was supported by National Research Foundation of Korea (NRF) grant funded by the Korean Government (MEST) (2012R1A2A2A01004895).

7. References

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

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

[3] J. H. James, C. Bing and L. Garrison, "Implementing VoIP: a voice transmission performance progress report," *IEEE Communications Magazine*, vol. 42, no. 7, pp. 36-41, Jul. 2004.

[4] J. A. Kang and H. K. Kim, "An adaptive packet loss recovery method based on real-time speech quality assessment and redundant speech transmission," *International Journal of Innovative Computing, Information and Control*, vol. 7, no. 12, pp. 6773-6783, Dec. 2011.

[5] N. I. Park, H. K. Kim, M. A. Jung, S. R. Lee, and S. H. Choi, "Burst packet loss concealment using multiple codebooks and comfort noise for CELP-type speech coders in wireless sensor networks," *Sensors*, vol. 11, no. 5, pp. 5323-5336, May 2011.

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

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

[8] S. Haykin, Adaptive Filter Theory, 3rd ed. Englewood Cliffs, NJ: Prentice-Hall, 1996.

[9] Aho A.V., Ullman J.V., Hopcroft J.E., Data Structures and Algorithms, New York, Addisson-Wesley, 1983.

[10] 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, iss. 1-2, pp. 17-31, Jun. 2004.

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

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

[13] ITU-T Rec. P.862.2, "Wideband extension to recommendation P.862 for the assessment of wideband telephone," Nov. 2007.

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