Packet Loss Robust Speech Streaming Based on Speech Quality Estimation and Bandwidth Extension of Narrowband Speech

Jin Ah Kang', Nam In Park', Seong Ro Lee[2], and Hong Kook Kim'

'School of Information and Communications
Gwangju Institute of Science and Technology (GIST), Gwangju 500-712, Korea
{jinari, nipark, hongkook} @gist.ac.kr
[2] School of Information Engineering
Mokpo National University, Jeollanam-do 534-729, Korea
srlee@mokpo.ac.kr

Abstract. In this paper, a packet loss robust speech streaming technique is proposed to improve the perceived speech quality (PSQ) of a speech streaming system. To this end, the proposed technique estimates PSQ for the received speech data, and then determines a proper redundant speech transmission mode in terms of the amount and bit-rate of the redundant speech. According to this decision, the proposed technique conducts rate control using a scalable speech coder to transmit primary and redundant speech data within the equivalent transmission bandwidth. To guarantee seamless PSQ, despite the speech bandwidth changing from narrowband to wideband due to the rate control, a bandwidth extension technique is incorporated. The effectiveness of the proposed technique is then demonstrated using ITU-T Recommendation G.729.1 as a scalable speech coder. It is shown from the experimental results that the proposed technique significantly improves PSQ under various packet loss conditions.

Keywords: Redundant speech transmission, Speech quality estimation, Bandwidth extension, ITU-T Recommendation G.729.1

1 Introduction

As audio and video streaming services are increasingly being extended to wireless networks, the quality of service becomes even more critical. In particular, speech streaming services such as e-learning require a minimum level of speech quality. However, when speech streaming is performed over wireless networks, packets may be lost due to variable network constraints [1]. Thus, this paper proposes a packet loss robust speech streaming (LR-SS) technique that conducts adaptive redundant speech transmission (ARST) based on the estimation of the perceived speech quality (PSQ). In addition, bandwidth extension of speech from narrowband to wideband is incorporated in the proposed LR-SS technique to overcome the degradation of seamless PSQ when the speech bandwidth varies due to rate control conducted for ARST.

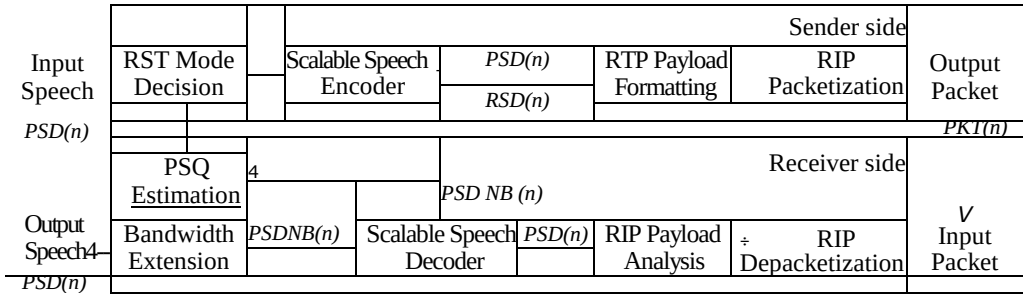

Fig. 1. Packet flow of a speech streaming system employing the proposed packet loss robust speech streaming technique.

2 Proposed Packet Loss Robust Speech Streaming

Fig. 1 shows a packet flow of a speech streaming system that employs the proposed LR-SS technique. The sender side generates bitstreams of primary and redundant speech data, *PSD(n)* and *RSD(n),* using a scalable speech encoder according to the redundant speech transmission (RST) mode. The RST mode is determined by the estimated PSQ delivered from the receiver side as feedback information. After combining *PSD(n)* and *RSD(n)* in the payload, the RTP packet, *PKT(n),* is sent to the receiver side. At the receiver side, $\overline{PSD(n)}$, is extracted from the payload of the received RTP packet. If *RSD(n)* exists in the payload, then this bitstream is stored for use during future packet loss recovery. *PSD(n)* is then decoded using a scalable speech decoder, and this decoded speech is used to estimate PSQ. The PSQ of the n-th speech frame, *Q(n),* is then estimated as a five-point mean opinion score (MOS).

A low-delay non-intrusive method of speech quality assessment, which is a simplified and low-delay version of the ITU-T Recommendation P.563 [2], is proposed in this paper. The RST mode, *M(n),* is determined in terms of the number of RSD frames according to

$$M(n)=\begin{cases} 0, & \text{if } 0(k) \; {}^{\circ}Q_1 \; 2, \text{ if} \\ O(k) < {}^{9}Q2 \; 1, & \\ \text{otherwise} & \end{cases} \qquad (1)$$

where $^{9}Q1$ and $^{\circ}Q2$ are the predefined thresholds for *Q(n).*

In addition, bandwidth extension from narrowband to wideband is performed using two different approaches that are applied to the 4-4.6 kHz and 4.6-7 kHz bands, respectively. Specifically, the 4-4.6 kHz band is extended through harmonic spectral band replication and correlation-based replication techniques [3], while the 4.6-7 kHz band is extended using a spectral folding technique [4].

Table 1. Speech quality measured in MOS using WPESQ for the different packet loss recovery techniques under PLRs ranging from 0% to 24%.

PLR (%) / Technique	0	3	6	9	12	15	18	21	24	Avg.
Decoder-based PLC technique	4.12	3.61	3.36	3.15	2.97	2.82	2.65	2.63	2.50	3.09
RST technique	3.88	3.84	3.86	3.78	3.64	3.51	3.45	3.37	3.21	3.62
Proposed LR-SS technique	4.12	3.81	3.81	3.75	3.67	3.61	3.64	3.63	3.45	3.72

3 Performance Evaluation

In order to demonstrate the effectiveness of the proposed LR-SS technique, a speech streaming system was implemented using the ITU-T Recommendation G.729.1 [5]. The bit-rate of the PSD bitstream was set at 32 kbit/s if *M(n)* was 1. If *M(n)* was 2, bit-rates for both the PSD and RSD bitstreams were all set at 16 kbit/s. If *M(n)* was 3, the bit-rate of the PSD bitstream was set at 16 kbit/s and that for the two RSD frames were set at 8 kbit/s. In addition, 9 and *9Q2* were set at 4.26 and 4.10 MOS, respectively, according to the results of a preliminary experiment.

To compare the speech quality, two conventional packet loss recovery techniques were implemented: an *RST technique* [6] and a *decoder-based PLC technique* [7]. The RST technique encoded speech signals using the ITU-T Recommendation G.729.1 encoder at a fixed rate of 16 kbit/s with the RSD bitstream of 16 kbit/s, but the decoder-based PLC technique encoded speech signals using the ITU-T Recommendation G.729.1 encoder at 32 kbit/s with no RSD bitstream. In this experiment, 40 speech utterances were selected from NTT-AT database [8]; each utterance was about 4 sec long and was sampled at a rate of 16 kHz. To evaluate speech quality, the wideband perceptual evaluation of speech quality (WPESQ) defined in the ITU-T Recommendation P.862.2 [9] was used. For the packet loss conditions, packet loss rates (PLRs) from 0% to 24% at a step of 3% were generated using the Gilbert-Elliott channel model defined in the ITU-T Recommendation G.191 [10].

Table 1 compares the quality of the speech recovered using the different packet loss recovery techniques under different PLRs. As shown in the table, the proposed LR-SS technique yielded an average speech quality of 3.72 MOS, which was higher than that yielded by the PLC and the RST techniques by as much by as 0.63 and 0.1 MOS, respectively.

4 Conclusion

In this paper, a packet loss robust speech streaming technique was proposed to transmit redundant speech according to the estimated PSQ under the current network condition. The PSQ was estimated for the received speech data in terms of MOS using a

low-delay non-intrusive method of speech quality assessment. In addition, bandwidth extension was conducted as a post-processing technique to guarantee seamless PSQ despite variation in the speech bandwidth from wideband to narrowband generated by the rate control. It was shown from the experiments that the proposed technique increased average WPESQ scores under different PLRs in comparison to the conventional techniques.

Acknowledgments. This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2012-010636), and by Priority Research Centers Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0022980).

References

1. Zhang, Q., Wang, G., Xiong, Z., Zhou, J. Zhu, W.: Error robust scalable audio streaming over wireless IP networks. IEEE Transactions on Multimedia, 6, 6, (2004) pp. 897-909.
2. Malfait, L., Bergerand, J., Kastner M.: P.563—the ITU-T standard for single-ended speech quality assessment. IEEE Transactions on Audio, Speech, and Language Processing, 14, 6, (2006) pp. 1924-1934.
3. Park, N. I., Kang, J. A., Kim, H K • A packet loss concealment algorithm based on artificial bandwidth extension. In: Proceedings of International Conference on Advanced Signal Processing (ASP), Seoul, Korea, (2012) pp. 96-101.
4. Pulakka, H., Laaksonen, L., Vainio, M., Pohjalainen **J.,** Alku, P.: Evaluation of an artificial speech bandwidth extension method in three languages. IEEE Transactions on Audio, Speech, and Language Processing, 16, 6, (2008) pp. 1124-1137.
5. ITU-T Recommendation G.729.1: An 8-32 kbit/s Scalable Wideband Coder Bit-stream Interoperable with G.729, (2006).
6. Kouvelas, I., Hodson, **0.,** Hardman, V., Crowcroft, J.: Redundancy control in real-time internet audio conferencing. In: Proceedings of International Workshop on Audio-Visual Services over Packet Networks (AVSPN), Aberdeen, Scotland, (1997) pp. 195-201.
7. Ragot, S., Kovesi, **B.,** Trilling, R., Virette, **D.,** Duc, N., Massaloux, D., Proust, E., Geiser, B., Garter, M., Schandl, S., Taddei, H., Gao, Y., Shlomot, E., Ehara, H., Yoshida, K., Vailancourt, T., Salami, R., Lee, M. S., Kim, D. Y.: ITU-T G.729.1: an 8-32 kbit/s scalable coder interoperable with G.729 for wideband telephony and voice over IP. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Honolulu, HI, (2007) pp. 529-532.
8. NTT-AT: Multi-Lingual Speech Database for Telephonometry, (1994).
9. ITU-T Recommendation P.862.2: Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs, (2007).
10. ITU-T Recommendation G.191: Software Tools for Speech and Audio Coding Standardization, (1996).