Research Article

Adaptive Speech Streaming Based on Speech Quality Estimation and Artificial Bandwidth Extension for Voice over Wireless Multimedia Sensor Networks

Jin Ah Kang, 1 Nam In Park, 2 Hong Kook Kim, 3 and Seong Ro Lee 4

1 Smart-Work Research Section, Wired & Wireless Convergence Research Department, Electronics and Telecommunications Research Institute (ETRI), 218 Gajeong-ro, Yuseong-gu, Daejeon 305-700, Republic of Korea
2 Digital Technology and Biometry Division, National Forensic Service (NFS), 10 Iphchun-ro, Wonju 220-170, Republic of Korea
3 School of Information and Communications, Gwangju Institute of Science and Technology (GIST), 123 Cheomdangwagi-ro, Buk-gu, Gwangju 500-712, Republic of Korea
4 Department of Information & Electronics Engineering, Mokpo National University, 1666 Youngsan-ro, Cheonggye-myeon, Muan-gun, Jeonnam 534-729, Republic of Korea

Correspondence should be addressed to Hong Kook Kim; hongkook@gist.ac.kr

Received 10 December 2014; Accepted 17 April 2015

Academic Editor: Antonio Lazaro

Copyright © 2015 Jin Ah Kang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In this paper, an adaptive speech streaming method is proposed to improve the perceived speech quality (PSQ) of voice over wireless multimedia sensor network (WMSNs). First of all, the proposed method estimates the PSQ of the received speech data under different network conditions that are represented by the packet loss rates (PLRs). Simultaneously, the proposed method classifies the speech signal as either an onset or a nononset frame. Based on the estimated PSQ and the speech class, it determines an appropriate bit rate for the redundant speech data (RSD) that are transmitted with the primary speech data (PSD) to help reconstruct the speech signals of any lost frames. In particular, when the estimated PLR is high, the bit rate of the RSD should be increased by decreasing that of the PSD. Thus, the bandwidth of the PSD is changed from wideband to narrowband, and an artificial bandwidth extension technique is applied to the decoded narrowband speech. It is shown from the simulation that the proposed method significantly improves the decoded speech quality under packet loss conditions in a WMSN, compared to a decoder-based packet loss concealment method and a conventional redundant speech transmission method.

1. Introduction

Because of the rapid development of low power and highly integrated digital electronic technologies, wireless multimedia sensor networks (WMSNs) are capable of retrieving audio and/or video streams as they interconnect sensor nodes equipped with multimedia devices such as cameras and microphones. Accordingly, they provide a wide range of potential applications needed to access audio or video data in real time such as environmental monitoring, human tracking, and security systems [1, 2]. However, it is difficult to guarantee seamless audio or video quality because those multimedia data are usually generated at a much higher bit rate than other sensor data. Furthermore, the reliability of transmission over WMSNs is apt to be degraded due to various resource constraints in WMSNs, compared to other networks [3, 4]. Thus, the quality of service (QoS) of multimedia streaming over WMSNs has become even more important. Specifically, applications of voice over WMSN (VoWMSN) require a minimum level of perceived speech quality (PSQ) [5–7].

To improve the PSQ of voice applications, various speech streaming methods have been proposed for use on IP networks. These methods are typically classified into either sender-based schemes or receiver-based schemes. Sender-based schemes consist of a collection of packet loss protection methods that provide error-robust transmission methods such as interleaving, forward error correction (FEC), and
redundant speech transmission (RST) [8–12]. On the other hand, receiver-based schemes consist of a collection of packet loss concealment (PLC) methods that compensate for lost speech signals using substitutable signals, for example, silence, previous good speech, or regenerated speech according to the analysis-by-synthesis criterion [13–16]. However, these two schemes can complement each other. That is, sender-based schemes are robust to higher packet loss rates (PLRs) because they often use redundant information to recover lost signals, which results in increased transmission bandwidth. Receiver-based schemes do not need to increase the transmission bandwidth because they conceal lost signals without using redundant information from the sender side; however, it is hard to prevent rapid PSQ degradation under high PLRs. It was also reported in [5] that in VoWMSN applications the receiver-based scheme could accommodate higher bit rates of speech coding under low PLR conditions, whereas the sender-based scheme was suitable for dealing with lower bit rates under high PLR conditions.

To take advantage of both schemes, a new method has been proposed in [17], which transmitted redundant speech data (RSD) adaptively according to the estimated PSQ under the current PLR condition and determined a suitable RST mode. On the basis of the mode, it generated bitstreams of primary speech data (PSD) and RSD using a scalable speech coder so as to maintain the equivalent transmission bandwidth. A lost speech signal was then recovered using the RSD for a high PLR. As a result, this method provided speech data (RSD) adaptively according to the estimated PSQ under the current PLR condition and determined a suitable RST mode. Next, the bitstreams of the PSD and RSD, PSD(n) and RSD(n), are generated using a scalable speech encoder according to the determined RST mode. After that, PSD(n) and RSD(n) are combined with the real-time transport protocol (RTP) payload format to obtain an RTP packet, PKT(n), which is transmitted to the receiver side over WMSN.

Meanwhile, at the receiver side, PSD(n) is depacketized from the RTP payload. In certain RST modes, if any RSD(n) exists in the payload, it is stored for use with any potential upcoming packet losses. The extracted PSD(n) is then decoded into PSD(n) using a scalable speech decoder. Finally, PSD(n) is sent to the output device. Note that if PSD(n) is only composed of a narrowband speech bitstream, the decoded narrowband speech signal, RSD, is artificially extended to a wideband speech signal to maintain a seamless PSQ. Since the PSQ estimation method (which will be described in Section 3.1) accepts only narrowband speech, RSD, is also used for the PSQ estimation.

### Table 1: Bit rate assignment according to different RST modes, where \( R_{p1} = R_{p0} - R_{r0} \) and \( R_{p2} = R_{p0} - R_{r1} - R_{r2} \)

| RST mode | PSD(n) (kbit/s) | RSD(n) (kbit/s) |
|----------|----------------|-----------------|
| \( M(n) \) | \( R_{p0} \) | \( R_{p1} \) | \( R_{p2} \) | \( R_{r0} \) | \( R_{r1} \) | \( R_{r2} \) |
| 0        | 32             | —              | —              | —              | —              | —              |
| 1        | —              | 16             | —              | 16             | —              | —              |
| 2        | —              | —              | 16             | —              | 8              | 8              |

**2. Voice over WMSN Based on Adaptive Speech Streaming**

### 2.1. Overall Structure

Figure 1 shows a block diagram and packet flow of the VoWMSN, which employs the proposed adaptive speech streaming method. Then, Section 3 proposes an adaptive speech streaming method for VoWMSNs. Section 4 evaluates the performance of the proposed method and compares it with those of a decoder-based PLC method and a conventional RST method. Finally, Section 5 concludes this paper.

**2.2. RTP Payload Format**

As mentioned above, the proposed adaptive speech streaming method can use an indicator for scalable bit rate speech coding. To deliver the estimated PSQ from the receiver side to the sender side, there should be a reserved field to accommodate the transmission of both the estimated PSQ and the RST bitstream. To this end, the RTP payload format defined in IETF RFC 4749 [19, 20], which is shown in Figure 2(a), is modified as shown in Figure 2(b). As shown in Figure 2(a), the “MBS | FT” sequence contains the payload header. In other words, the four-bit maximum bit rate supported (MBS) field is used to inform the sender side of the maximum bit rate that can be received. In the ITU-T Recommendation G.729.1 speech coder [21] that will be used to implement a VoWMSN system described in Section 4, the MBS is assigned a value from 0 to 11, corresponding to an encoding bit rate of 8, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, or 32 kbit/s, respectively. In addition, the frame-type (FT) field, which consists of four bits, indicates the actual encoding bit rate of the contained bitstreams. Thus, this field is also assigned a value from 0 to 11, corresponding to one of the encoding bit rates between 8 and 32 kbit/s. It should be noted that value 15 indicates a condition in which there is no data to be transmitted while values 12 to 14 are reserved for future use.

---

**International Journal of Distributed Sensor Networks**
3. Proposed Adaptive Speech Streaming Method

3.1. Speech Quality Estimation. The proposed adaptive speech streaming method begins by estimating the PSQ because the PSQ is a good indicator of both the current PLR and the bit rate of speech coding. To this end, the ITU-T Recommendation P.563 [22] is employed in this paper as an objective PSQ assessment method in order to monitor the PSQ of VoWMSN, and it estimates the PSQ as a mean opinion score (MOS) without using a reference speech signal. The proposed method requires that the PSQ estimate should be done in real time, thus the ITU-T Recommendation P.563 needs to be modified to have low-delay requirements, which is referred to as a nonintrusive perceived speech quality assessment (LD-QA) method in this paper.

Figure 3 shows an overall structure of the PSQ estimation method using three processing stages such as the preprocessing stage, the distortion estimation stage, and the perceptual mapping stage. To take into account various distortion factors during speech streaming, the model also combines three processing modules in the distortion estimation stage, which is based on the ITU-T Recommendation P.563. The first module models the vocal tract as a series of tubes with abnormal variations for degradation modeling and estimates the linear prediction coefficients (LPCs) within a certain range expected for a natural speech signal. The second module reconstructs a clean reference speech signal from the degraded speech signal and then evaluates the difference between the reconstructed clean speech and the degraded speech signal. The third module identifies and estimates specific distortions expected to be encountered in transmission channels.

In the perceptual mapping stage, the distortion effects estimated in the distortion estimation stage are linearly combined to estimate an MOS, which is denoted by $\tilde{Q}(n)$ in Figure 3. The PSQ estimation method described so far...
provides a single MOS for an input speech file whose length should be longer than approximately 4 s [22].

On the other hand, the proposed LD-QA method modifies the second stage so that the distortion effects are modeled using a minimal amount of speech data. In particular, each pitch mark in the first module of the second stage is extracted once every frame, where the frame size is 64 ms long, and the second module is also applied once every frame. In addition, the distortion-specific parameters for noise detection, temporal time clipping, and robotization are updated once every frame using speech signals of 500 ms, 64 ms, and 1 s long, respectively. Consequently, $\tilde{Q}(n)$ is calculated once every frame, which results in a processing delay of 20 ms for the LD-QA method.

3.2. Artificial Bandwidth Extension. When the estimated PLR is high, the bit rate of the RSD should be increased by decreasing the bit rate of the PSD, which is realized by changing the bandwidth of the PSD from wideband to narrowband. In this case, an ABE technique is used to overcome the performance degradation of the seamless PSQ by extending the bandwidth of speech signals from narrowband to wideband to improve the speech quality of the narrowband speech.

Figure 4 shows a block diagram of the ABE technique operated in the modified discrete cosine transform (MDCT) domain [18]. In this figure, if the RST mode is 2, narrowband speech, $\text{PSD}_{\text{NB}}(n)$, is segmented into a sequence of frames with a frame size of $N$. Next, each analysis frame is transformed into the frequency domain using a $2N$-point MDCT,
3.3. RST Mode Decision and Bit Rate Assignment. In the proposed adaptive speech streaming method, the PSQ is estimated at the receiver side using the LD-QA method and then is sent back to the sender side to make the RST mode decision. Figure 5 shows a block diagram of the RST mode decision based on the speech class and the estimated PSQ in the proposed method.

First, each frame is classified into one of six different classes, namely, silence/background noise, stationary unvoiced, nonstationary unvoiced, speech onset, nonstationary voiced, or stationary voiced [23]. Next, a preliminary experiment is carried out to investigate the relationship between each class and the RST mode under different PLR conditions. Consequently, it is found that the RST mode is most sensitive to the speech onset class. Thus, the proposed adaptive speech streaming method decides only whether or
not the $n$th frame is primarily made up of speech onset, such that

$$c(n) = \begin{cases} 
1, & \text{if PSD}(n) \text{ is onset}, \\
0, & \text{otherwise}. 
\end{cases} \quad (1)$$

Next, the RST mode, $M(n)$, is determined using both $\bar{Q}(n)$ described in Section 3.1 and $c(n)$ in (1) as

$$M(n) = \begin{cases} 
0, & \text{if } \bar{Q}(n) \geq \theta_{Q_1}, \\
1, & \text{if } \theta_{Q_1} < \bar{Q}(n) < \theta_{Q_2}, \\
2, & \text{otherwise}, 
\end{cases} \quad (2)$$

where $\theta_{Q_1}$ and $\theta_{Q_2}$ are the two predefined thresholds for estimating the speech quality degradation for the current PLR.

Figure 6 shows several bit rate assignments for the PSD and RSD bitstreams according to different RST modes, $M(n)$. As shown in the figure, if $M(n) = 0$, PSD($n$) is composed of the $n$th speech bitstream encoded at the highest bit rate, $R_{P0}$, with no RSD bitstream, which is denoted by PSD($n$) = [PSD($n$)@$R_{P0}$]. Otherwise, PSD($n$) is encoded at a lower bit rate, $R_{P1}$ or $R_{P2}$, and the remaining bit rate, $R_{R0}$, $R_{R1}$, or $R_{R2}$, is assigned to the RSD bitstream. However, if $M(n) = 1$, RSD($n$) is composed of the $n$th speech bitstream encoded at $R_{R0}$; that is, RSD($n$) = [PSD($n$ + 1)@$R_{R0}$]. In addition, if $M(n) = 2$, RSD($n$) is composed of the $(n + 1)$th and $(n + 2)$th speech bitstreams, which are encoded at $R_{R1}$ and $R_{R2}$, respectively; that is, RSD($n$) = [PSD($n$ + 1)@$R_{R1}$, PSD($n$ + 2)@$R_{R2}$]. Consequently, the total bit rate for the speech data is maintained.

4. Performance Evaluation

To demonstrate the effectiveness of the proposed adaptive speech streaming method, a VoWMSN system was created using the ITU-T Recommendation G.729.1 as a scalable speech coder as shown in Figure 7. In fact, the proposed method and other speech streaming methods were implemented in the application layer. For the evaluation, input speech signals were sampled at 16 kHz and encoded using the ITU-T Recommendation G.729.1 speech encoder at a bit rate of 32 kbit/s. The bit rate assignment for the PSD and RSD bitstreams according to different RST modes was performed as shown in Table 1. In addition, $\theta_{Q_1}$ and $\theta_{Q_2}$ in (2) were set to 3.75 and 4.15 MOS, respectively.

To compare the performance of the PSQ estimation within the equivalent transmission bandwidth, two conventional speech streaming methods were implemented: a decoder-based PLC method [21] and an RST method [10]. The decoder-based PLC method encoded speech signals using the ITU-T Recommendation G.729.1 encoder at 32 kbit/s with no RSD bitstream, and the conventional RST method also encoded speech signals using the ITU-T Recommendation G.729.1 encoder at a fixed bit rate of 16 kbit/s with an RSD bitstream of 16 kbit/s. In this experiment, 48 speech utterances were taken from the NTT-AT speech database [24], where each speech utterance was approximately 4 s long and was sampled at a rate of 16 kHz. Each utterance was filtered using a modified intermediate reference system (IRS) filter, followed by automatic level adjustment [25]. To evaluate the quality of the decoded speech for each method, scores of wideband perceptual evaluation of speech quality (WPESQ) were measured as defined by the ITU-T Recommendation P.862.2 [26]. For the simulation of WMSN conditions, the Gilbert-Elliot (GE) channel model [25] was used to simulate the packet loss conditions because it is able to characterize the fading of a wireless network [6, 27, 28]. In this paper, burst PLRs were generated from 0 to 25% at a step of 5% using the GE model. Note that the mean and maximum burst packet losses were measured at approximately 1.5 and 4 packets, respectively.

Figure 8 compares the WPESQ scores (in MOS) of decoded speech processed using different speech streaming methods under different PLR conditions. As shown in this figure, the WPESQ score of the proposed method was better than those of the conventional methods. In other words, the proposed method improved average WPESQ score by as much as 0.55 and 0.2 MOS, compared to the decoder-based PLC method and the RST method, respectively.

5. Conclusion

In this paper, an adaptive speech streaming method was proposed to improve speech quality of a voice over wireless multimedia sensor network (WMSN). To this end, the proposed method first classified each frame of input speech signals as either an onset frame or a nononset frame. Next, it estimated the perceived speech quality (PSQ) of the received speech data under packet loss conditions. On the basis of the estimated PSQ and the speech class, the proposed method
determined an appropriate bit rate for the redundant speech data (RSD) that was transmitted with the primary speech data (PSD) to assist the speech decoder in reconstructing the speech signals for a lost frame. In particular, an artificial bandwidth extension technique was applied to the narrow-band speech decoded when the estimated packet loss rate (PLR) was high. The effectiveness of the proposed method was demonstrated by implementing a voice over WMSN system employing the proposed method. A performance evaluation indicated that the proposed method significantly improved the decoded speech quality relative to the conventional methods under different PLR conditions ranging from 0% to 25%.

**Conflict of Interests**

The authors declare that there is no conflict of interests regarding the publication of this paper.

**Acknowledgments**

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the government of Korea (MSIP) (no. 2015R1A2A1A05001687), by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2009-0093828), and by the MSIP, Korea, under the ITRC (Information Technology Research Center) support program (IITP-2015-H8501-15-1016) supervised by the IITP (Institute for Information & Communications Technology Promotion).

**References**

[1] I. T. Almalkawi, M. G. Zapata, J. N. Al-Karaki, and J. Morillo-Pozo, "Wireless multimedia sensor networks: current trends and future directions," Sensors, vol. 10, no. 7, pp. 6662–6717, 2010.

[2] I. F. Akyildiz, T. Melodia, and K. R. Chowdhury, "A survey on wireless multimedia sensor networks," Computer Networks, vol. 51, no. 4, pp. 921–960, 2007.

[3] T. Melodia and I. F. Akyildiz, "Cross-layer QoS-aware communication for ultra wide band wireless multimedia sensor networks," IEEE Journal on Selected Areas in Communications, vol. 28, no. 5, pp. 653–663, 2010.

[4] J. A. Kang and H. K. Kim, "Adaptive redundant speech transmission over wireless multimedia sensor networks based on estimation of perceived speech quality," Sensors, vol. 11, no. 9, pp. 8469–8484, 2011.

[5] R. Mangharam, A. Rowe, R. Rajkumar, and R. Suzuki, "Voice over sensor networks," in Proceedings of the 27th IEEE International Real-Time Systems Symposium (RTSS ’06), pp. 291–300, Rio de Janeiro, Brazil, December 2006.

[6] N. I. Park, J. A. Kang, S. R. Lee, and H. K. Kim, "A packet loss concealment technique improving quality of service for wideband speech coding in wireless sensor networks," International Journal of Distributed Sensor Networks, vol. 2014, Article ID 852798, 8 pages, 2014.

[7] D. Brunelli, M. Maggiorotti, L. Benini, and F. L. Bellifemine, "Analysis of audio streaming capability of Zigbee networks," in Wireless Sensor Networks: 5th European Conference, EWSN 2008, Bologna, Italy, January 30–February 1, 2008. Proceedings, vol. 4913 of Lecture Notes in Computer Science, pp. 189–204, Springer, Berlin, Germany, 2008.

[8] F. Merazka, "Improved packet loss recovery using interleaving for CELP-type speech coders in packet networks," IAEING International Journal of Computer Science, vol. 36, pp. 1–5, 2009.

[9] W. Lizhong, W. Muqing, W. Lulu, and L. Moja, "An adaptive forward error control method for voice communication," in Proceedings of the 2nd International Conference on Networking and Digital Society (ICNDS '10), pp. 186–189, Wenzhou, China, May 2010.

[10] I. Kouvelas, O. Hodson, V. Hardman, and J. Crowcroft, "Redundancy control in real-time internet audio conferencing," in Proceedings of the International Workshop on Audio-Visual Services over Packet Networks (AVSPN ’97), pp. 195–201, Aberdeen, Scotland, September 1997.

[11] K. Park, Y. Oh, K. Lim, and K.-R. Cho, "A dynamic packet recovery mechanism for realtime service in mobile computing environments," ETRI Journal, vol. 25, no. 5, pp. 356–368, 2003.

[12] T.-Y. Wu, S. Guizani, W.-T. Lee, and P.-C. Huang, "An enhanced structure of layered forward error correction and interleaving for scalable video coding in wireless video delivery," IEEE Wireless Communications, vol. 20, no. 4, pp. 146–152, 2013.

[13] 3GPP TS 06.11, Substitution and Muting of Lost Frames for Full Rate Speech Channels, 2000.

[14] 3GPP TS 26.091, Mandatory Speech Codec Speech Processing Functions; AMR Speech Codec; Error Concealment of Lost Frames, 2010.

[15] 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, 2011.

[16] J. Huang, X. Zhang, and Y. Zhang, "Recovery of lost speech segments using incremental subspace learning," ETRI Journal, vol. 34, no. 4, pp. 645–648, 2012.

[17] J. A. Kang, H. K. Kim, S. H. Choi, and S. R. Kim, "Adaptive redundant speech streaming with scalable speech coding based on speech quality estimation," Information, vol. 17, no. 5, pp. 1921–1932, 2014.

[18] N. I. Park, Y. H. Lee, and H. K. Kim, "Artificial bandwidth extension of narrowband speech signals for the improvement of perceptual speech communication quality," Communications in Computer and Information Science, vol. 266, no. 2, pp. 143–153, 2011.

[19] A. Sollaud, "RTP payload format for the G.729.1 audio codec," IETF RFC 4749, 2006.

[20] IETF RFC 5459, G.729.1 RTP Payload Format Update: Discontinuous Transmission (DTX) Support, 2009.

[21] S. Ragot, B. Kovessi, R. Trilling et al., "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 the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP ’07), pp. IV529–IV532, Honolulu, Hawaii, USA, April 2007.

[22] L. Malfait, J. Berger, and M. Kastner, "P563-The ITU-T standard for single-ended speech quality assessment," IEEE Transactions on Audio, Speech and Language Processing, vol. 14, no. 6, pp. 1924–1934, 2006.

[23] Y. Gao, E. Shlomot, A. Benyassine, J. Thyssen, H. Su, and C. Murgia, "The SMV algorithm selected by TIA and 3GPP2 for
CDMA applications,” in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’01), pp. 709–712, Salt Lake City, Utah, USA, May 2001.

[24] NTT-AT, Multi-Lingual Speech Database for Telephonometry, 1994.

[25] ITU-T Recommendation G.191, Software Tools for Speech and Audio Coding Standardization, 2010.

[26] International Telecommunication Union (ITU), ITU-T Recommendation P862.2, Wideband Extension to Recommendation P862 for the Assessment of Wideband Telephone Networks and Speech Codecs, International Telecommunication Union (ITU), 2008.

[27] S. Khan, Y. Peng, E. Steinbach, M. Sgroi, and W. Kellerer, “Application-driven cross-layer optimization for video streaming over wireless networks,” IEEE Communications Magazine, vol. 44, no. 1, pp. 122–130, 2006.

[28] T. Nagano and A. Ito, “Packet loss concealment of voice-over IP packet using redundant parameter transmission under severe loss conditions,” Journal of Information Hiding and Multimedia Signal Processing, vol. 5, no. 2, pp. 286–295, 2014.