Adaptive Speech Streaming Based on Packet Loss Prediction Using Support Vector Machine for Software-Based Multipoint Control Unit over IP Networks

Jin Ah Kang, Mikyong Han, Jong-Hyun Jang, and Hong Kook Kim

An adaptive speech streaming method to improve the perceived speech quality of a software-based multipoint control unit (SW-based MCU) over IP networks is proposed. First, the proposed method predicts whether the speech packet to be transmitted is lost. To this end, the proposed method learns the pattern of packet losses in the IP network, and then predicts the loss of the packet to be transmitted over that IP network. The proposed method classifies the speech signal into different classes of silence, unvoiced, speech onset, or voiced frame. Based on the results of packet loss prediction and speech classification, the proposed method determines the proper amount and bitrate of redundant speech data (RSD) that are sent with primary speech data (PSD) in order to assist the speech decoder to restore the speech signals of lost packets. Specifically, when a packet is predicted to be lost, the amount and bitrate of the RSD must be increased through a reduction in the bitrate of the PSD. The effectiveness of the proposed method for learning the packet loss pattern and assigning a different speech coding rate is then demonstrated using a support vector machine and adaptive multirate-narrowband, respectively. The results show that as compared with conventional methods that restore lost speech signals, the proposed method remarkably improves the perceived speech quality of an SW-based MCU under various packet loss conditions in an IP network.

Keywords: Software-based multipoint control unit, Adaptive speech streaming, Packet loss prediction, Redundant speech transmission, Support vector machine, AMR-NB.

Manuscript received Apr. 29, 2016; revised Oct. 28, 2016; accepted Nov. 3, 2016.

Jin Ah Kang (jakang@etri.re.kr), Mikyong Han (mkhan@etri.re.kr), and Jong-Hyun Jang (jangjh@etri.re.kr) are with the 5G Giga Communication Research Laboratory, ETRI, Daejeon, Rep. of Korea.

Hong Kook Kim (corresponding author, hongkook@gist.ac.kr) is with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Rep. of Korea.

I. Introduction

Studies presented during the last year showed that worldwide IP traffic has increased more than fivefold in the last five years and is predicted to increase nearly threefold over the next five years. Particularly noteworthy are findings showing that busy-hour IP traffic is growing more rapidly than average IP traffic. In addition, traffic from wireless and mobile devices will exceed traffic from wired devices in roughly three years [1]. This situation indicates that real-time multimedia communication services such as video conferencing could suffer more from increased packet losses owing to transmission errors, congestion control, or long delays [2], [3]. Accordingly, adaptive streaming methods providing the best audio/video quality under current network conditions are needed. In this regard, this paper proposes an adaptive speech streaming method for a software-based multipoint control unit (SW-based MCU) to improve the perceived speech quality (PSQ) of video conference services.

To improve the PSQ of multimedia streaming services over IP networks, a range of speech streaming methods have been proposed. These methods are generally classified into either sender- or receiver-based schemes. Sender-based schemes are composed of an aggregate of packet loss protection methods for providing an error robust transfer method, for example, interleaving, forward error correction, or redundancy speech transmission (RST) [4]–[8]. A receiver-based scheme is composed of an aggregate of packet loss concealment (PLC) methods that compensate for lost speech signals using substitutable signals such as silence, previous good speech, or

generated speech based on an analysis-by-synthesis criterion [9]–[12]. However, these two schemes complement each other. In other words, a sender-based scheme is robust to higher packet loss rates (PLRs) because it often uses redundant information to restore lost speech signals, which increases the transmission bitrate. A receiver-based scheme does not increase the transmission bitrate because it restores lost speech signals without using redundant information; however, it is difficult to prevent rapid PSQ degradation under a high PLR.

To utilize the advantages of both schemes, an adaptive speech streaming method (PSQE-based ASSM) was proposed to adaptively transmit redundant speech data (RSD) based on a real-time PSQ estimation under current network conditions [13]. The PSQE-based ASSM determines a suitable RST mode based on the PSQ estimation for the current PLR, and then generates the bitstreams of primary speech data (PSD) and RSD using a multirate speech coder to maintain the equivalent transmission bitrate. Accordingly, a lost speech signal is reconstructed using the RSD under a high PLR. As a result, this method provides an improved overall PSQ under various PLRs within equivalent transmission bitrates. Despite its advantages, this method may respond rather slowly to variations in the PLR because the PSQ is estimated using speech signals in which previous speech signals of as long as 4 s are included to gather a minimal amount of speech signals for an estimation.

Accordingly, to provide a more improved PSQ against variations in the PLR, in this paper, a new adaptive speech streaming method (PLP-based ASSM) based on a packet loss prediction (PLP), which predicts the loss of each packet before sending it, is proposed. In particular, the proposed method aims to improve the PSQ of a SW-based MCU. To this end, a machine learning approach is applied to learn the pattern of packet losses in an IP network, and then predict the loss of a packet to be transmitted over that IP network. The proposed PLP-based ASSM also transmits bitstreams adaptively combined with PSD and RSD according to the decision of the RST mode in a manner similar to that of the PSQE-based ASSM. That is, the RST mode is determined using both the speech classification and PLP results.

The remainder of this paper is organized as follows. Section II describes the PSQE-based ASSM, which is the basis of the proposed method. Next, Section III proposes the PLP-based ASSM. Section IV then provides a performance evaluation of the proposed PLP-based ASSM by measuring the prediction accuracy for packet losses and comparing the PSQ to that of a decoder-based PLC method and a conventional RST method. Finally, Section V provides some concluding remarks regarding this research.

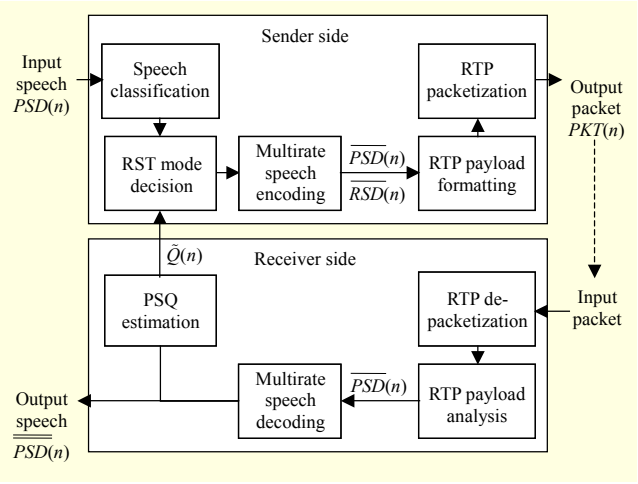

Fig. 1. Block diagram and packet flow of speech streaming system employing the PSQE-based ASSM.

II. Previous Work: Adaptive Speech Streaming Method Based on Perceived Speech Quality Estimation

1. Overall Structure

Figure 1 shows a block diagram and the packet flow of a speech streaming system employing PSQE-based ASSM [13]. As the speech signal $PSD(n)$ comes into an input device at the sender side of a speech streaming system, it is classified as either an onset frame or a non-onset frame. The classification result is then used to determine a suitable RST mode together with the estimated PSQ delivered from the receiver side. Next, the bitstreams of the PSD and RSD, $\overline{PSD}(n)$ and $\overline{RSD}(n)$, are generated using a multirate speech encoder based on the determined RST mode. Then, $\overline{PSD}(n)$ and $\overline{RSD}(n)$ are combined using a real-time transport protocol (RTP) [14] payload format to obtain an RTP packet $\overline{PKT}(n)$, which is transmitted to the receiver side over an IP network.

At the receiver side, $\overline{PSD}(n)$ is de-packetized from the RTP payload. In certain RST modes, if any $\overline{RSD}(n)$ exists in the payload, it is stored for use with any potential upcoming packet loss. The extracted $\overline{PSD}(n)$ is then decoded into $\overline{\overline{PSD}}(n)$ using a multirate speech decoder. Finally, $\overline{\overline{PSD}}(n)$ is sent to the output device; it is also used for the PSQ estimation.

2. RTP Payload Format

As mentioned above, the PSQE-based ASSM can use an indicator for multirate speech coding. To deliver the estimated PSQ from the receiver side to the sender side, there should be a

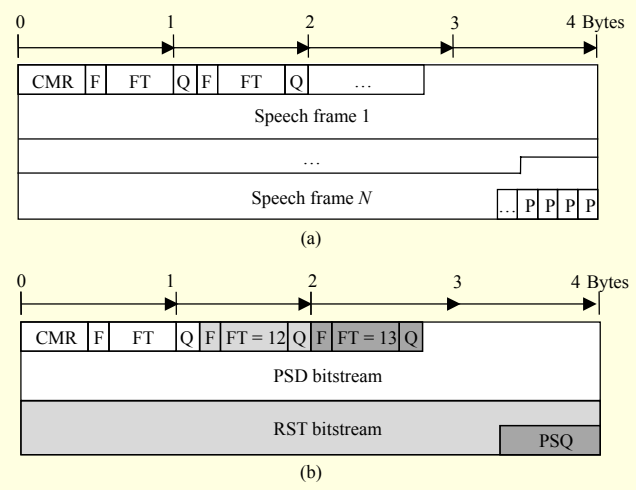

Fig. 2. Comparison of RTP payload formats: (a) format defined in IETF RFC 3267 and (b) modified format for PSQE-based ASSM.

reserved field to accommodate the transmission of both the estimated PSQ and the RSD bitstream. To this end, the RTP payload format defined in IETF RFC 3267 [15], which is shown in Fig. 2(a), is modified as shown in Fig. 2(b). Note that the adaptive multirate narrowband (AMR-NB) [16] is selected as a speech coder for the SW-based MCU in order to support legacy mobile phones. Thus, the PSQ of the narrowband speech signal decoded by the AMR-NB can be estimated using ITU-T Recommendation P.563 [17].

As shown in Fig. 2(a), the "CMR|F|FT|Q" sequence contains the payload header. In detail, the 4-bit codec mode request (CMR) field asks the sender side to change the encoding bitrate. In the AMR-NB, the CMR is assigned a value from zero to 7, corresponding to an encoding bitrate of 4.75 kb/s, 5.15 kb/s, 5.90 kb/s, 6.70 kb/s, 7.40 kb/s, 7.95 kb/s, 10.2 kb/s, and 12.2 kb/s, respectively. In addition, a 1-bit "F" field is set to 1 or zero in order to indicate whether this frame is to be followed by another piece of speech frame data. The "FT" field, which consists of four bits, then represents the actual encoding bitrate of the included bitstream. Therefore, this field is also assigned a value from zero to 7, corresponding to one of the bitrates between 4.75 kb/s and 12.2 kb/s. However, the assigned value changes from 8 to 11 when comfort noise is encoded. Note that a value of 15 indicates a condition in which there are no data to be transmitted. At the end of the payload, the "P" field is used for an octet alignment.

On the other hand, in a modified RTP payload format, two fields (that is, FT = 12 and FT = 13) are added to indicate the RSD bitstream and estimated PSQ, respectively, as shown in Fig. 2(b). Moreover, the main field for the speech frames, as shown in Fig. 2(a), is split into three fields representing the PSD bitstream, RSD bitstream, and estimated PSQ.

3. PSQ Estimation and RST Mode Decision

The PSQE-based ASSM begins by estimating the PSQ, which is a good indicator of the current PLR. Therefore, a low-delayed version of the nonintrusive PSQ assessment method defined in ITU-T Recommendation P.563 is used to estimate the PSQ as a mean opinion score (MOS) without using a reference speech signal [13].

The estimated PSQ, $\tilde{Q}(n)$ in Fig. 1 is then sent back to the sender side for the RST mode decision. The RST mode decision is conducted as follows. First, the input speech signal $PSD(n)$ in Fig. 1 is classified into one of four different classes: silence, onset, unvoiced, or voiced [18]. Note that the RST mode is found to be the most sensitive to the speech onset class under different PLR conditions according to our study [19], where each speech frame was declared as four different classes, such as silence, onset, unvoiced, and voiced, and the PSQ was the lowest when speech frames belonging to speech onset were lost. Thus, the PSQE-based ASSM decides only whether the n-th speech frame $PSD(n)$ is primarily made up of a speech onset, such that

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

Next, the RST mode $M(n)$ is determined using both $\tilde{Q}(n)$ and $C(n)$ in (1) as

$$M(n)=\begin{cases}0, & \text{if } \tilde{Q}(n)\ge\theta_{Q2},\\ 1, & \text{if } \theta_{Q1}\le\tilde{Q}(n)<\theta_{Q2} \text{ and } C(n)=0,\\ 2, & \text{otherwise,}\end{cases} \tag{2}$$

where θ_{Q1} and θ_{Q2} are the predefined thresholds for estimating the degradation of speech quality under the current PLR. θ_{Q1} and θ_{Q2} are set to 2.3 and 4.0 MOS, respectively, so that the performance of the PSQE-based ASSM is maximized for $2.0\le\theta_{Q1}\le3.0$ and $4.0\le\theta_{Q2}\le4.5$ for the performance evaluation in Section IV-2.

Figure 3 shows several bitrate assignments for the PSD and RSD bitstreams according to different $M(n)$. If $M(n)$ = 0 such as in Fig. 3(a), then $\overline{PSD}(n)$ is composed of the n-th speech bitstream encoded at the highest bitrate R_{P0} with no RSD bitstream, which is denoted by $\overline{PSD}(n)=[PSD(n)@R_{P0}]$. Otherwise, $\overline{PSD}(n)$ is encoded at a lower bitrate R_{P1} or R_{P2}, and the remaining bitrate (R_{R0}, R_{R1}, or R_{R2}) is assigned to the RSD bitstream. That is, if $M(n)$ = 1, such as in Fig. 3(b), $\overline{RSD}(n)$ is composed of the n-th speech bitstream encoded at R_{R0}, namely, $\overline{RSD}(n)=[PSD(n+1)@R_{R0}]$. If $M(n)$ = 2,

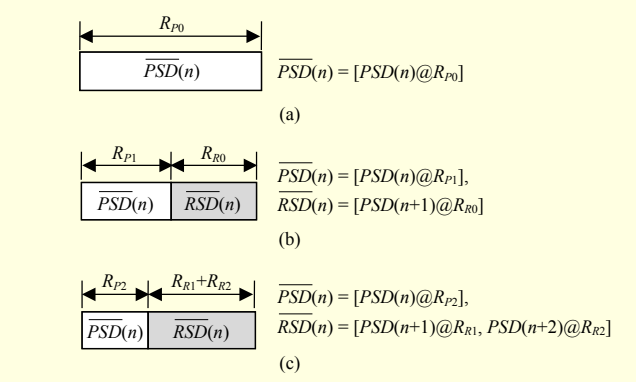

Fig. 3. Bitrate assignment according to different RST modes when RST mode $M(n)$ equals (a) zero, (b) 1, and (c) 2.

such as in Fig. 3(c), $\overline{RSD}(n)$ is composed of the $(n+1)$-th and $(n+2)$-th speech bitstreams, which are encoded at R_{R1} and R_{R2}, respectively, namely, $\overline{RSD}(n) = [PSD(n+1)@R_{R1}, PSD(n+2)@R_{R2}]$. Consequently, the total bitrate for the speech data is maintained.

III. New Adaptive Speech Streaming Method Based on Packet Loss Prediction for Software-Based Multipoint Control Unit

In the PSQE-based ASSM, the PSQ estimation performs well as an indicator of the current packet loss condition. Nevertheless, the PSQ estimation can be considered a long-term estimation, relative to the variations in packet loss conditions, because the PSQ is estimated based on the received speech signals (of as long as 4 s), whereas packet losses often occur in a burst fashion at short time intervals. Note here that the length of a speech segment for the PSQ estimation was set to 4 s according to ITU-T Recommendation P.563, in which the minimum length of active speech is 3 s and the speech activity ratio should be from 25% to 75% [20]. Moreover, each speech sample in the NTT-AT database [21] is composed of two short utterances of 4 s long, resulting in an actual length of 8 s for each speech sample, and their speech activity is approximately 75%.

In this section, a new PLP method is proposed to predict the loss of each packet before sending it by adopting a machine learning approach. In addition, using the proposed PLP, a new adaptive speech streaming method, called PLP-based ASSM, is proposed to ensure the robust performance of an SW-based MCU against variations of the packet loss condition.

1. Overall Structure

Figure 4 shows a block diagram and packet flow of the speech processor for an SW-based MCU that employs the proposed PLP-based ASSM. As shown in this figure, the speech processor of the SW-based MCU conducts real-time speech mixing, which receives speech signals $S_{in}(i)$ from all connected clients. It then mixes these signals $S'_{in}(i)$ after decoding. After encoding, it sends the mixed signals $S_{out}(i) = \sum_{t=1, t\neq i}^{T} S'_{in}(t)$ back to the clients. Thus, the SW-based MCU can offer speech communication functions in video conferencing applications.

At the receiver side of the speech processor in Fig. 4, as the

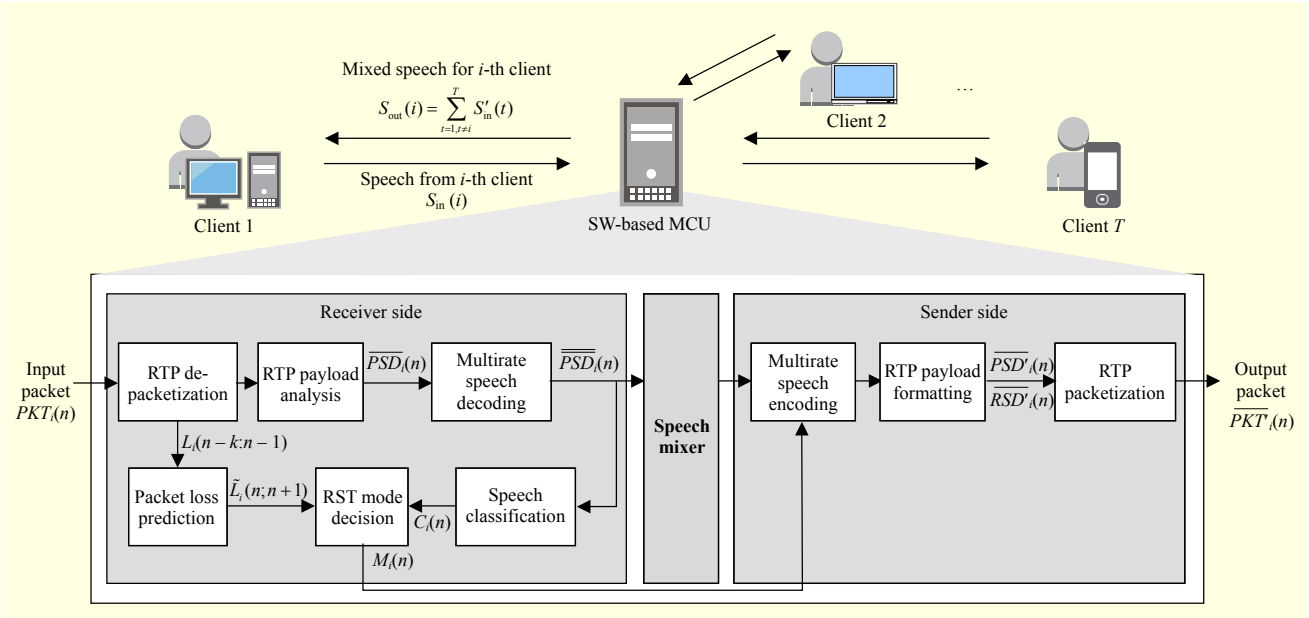

Fig. 4. Block diagram and packet flow of speech processor for SW-based MCU employing proposed PLP-based ASSM.

speech packet $PKT_i(n)$ is transmitted from the i-th client, $\overline{PSD_i}(n)$ is de-packetized from the RTP payload. If any $\overline{RSD_i}(n)$ exists in the payload, it is stored for use with any potential upcoming packet losses in the same fashion as the PSQE-based ASSM. The extracted $\overline{PSD_i}(n)$ is then decoded into $\overline{\overline{PSD_i}}(n)$ using a multirate speech decoder, and sent to the speech mixer to be mixed with speech signals transmitted from the other clients. In addition, $\overline{\overline{PSD_i}}(n)$ is also used for speech classification. Its result, $C_i(n)$, is used to determine an appropriate RST mode $M_i(n)$ when combined with the packet loss prediction result $\tilde{L}_i(n:n+1)$.

At the sender side of the speech processor, the mixed speech signal $S_{out}(i)$ for the n-th frame is encoded into the bitstreams of the PSD and RSD, $\overline{PSD'}_i(n)$ and $\overline{RSD'}_i(n)$, using a multirate speech encoder based on the determined RST mode. Then, $\overline{PSD'}_i(n)$ and $\overline{RSD'}_i(n)$ are combined with the RTP payload format to obtain an RTP packet $\overline{PKT'}_i(n)$, which is transmitted to the i-th client over an IP network. The RTP format used to transmit these bitstreams is defined in the same manner described in Section II, except that it also provides fields to deliver the result of a PSQ estimation.

2. Packet Loss Prediction

A. System Structure

The proposed PLP-based ASSM starts by predicting the loss of a packet to be transmitted in order to determine an appropriate RST mode. To this end, a support vector machine (SVM), which is a popular supervised learning model that finds an optimal hyperplane to analyze the data used for classification and regression [22]–[24], is adopted. Note that the SVM is trained using a radial basis function (RBF). The SVM has been shown to be a good real-time classifier [25], [26] and is also used to conceal lost speech packets [27], [28]. In other words, the SVM generates a model for packet losses in an IP network in which an SW-based MCU is located. Next, the SVM predicts the loss of each packet to be transmitted over that IP network based on the generated packet loss model. In this paper, these two steps of modeling and prediction of the packet loss are referred to as off-line training and on-line prediction, respectively.

Figure 5 shows a detailed diagram of the proposed PLP. For the off-line training for predicting whether the n-th packet is lost, a sequence of packet loss indicators of the previous packets is grouped together. The RTP de-packetization process provides an indicator by comparing the RTP sequence number of the received packet with those of the previously received packets. That is, $L(k) = 1$ if the k-th received packet is lost; otherwise, $L(k) = 0$. Thus, $L(n-k{:}n-1)$ in the figure is the k-dimensional binary sequence of indicators from the $(n-k)$-th packet to the $(n-1)$-th packet. Next, $L(n-k{:}n-1)$ is used to extract a feature vector $X_n(m)$ for the packet loss prediction of the n-th packet. A detailed component of $X_n(m)$ will be described in Section III2C. After a sufficient number of features are input into the pattern training, a packet loss model is generated.

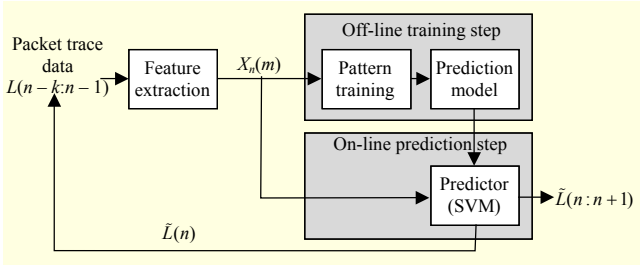

Fig. 5. Block diagram of proposed PLP based on machine learning provided by SVM.

Next, the on-line prediction step is started with the input of $L(n-k{:}n-1)$. Then, $X_n(m)$ is input into the predictor, in which the input features are analyzed to find the closer class within the packet loss model. Finally, the loss $\tilde{L}(n)$ of a packet to be transmitted over the trained network is predicted. To support a certain RST mode that includes two RSD bitstreams, the loss of the next $(n+1)$-th packet, $\tilde{L}(n+1)$, is also predicted using $\tilde{L}(n)$ together with $L(n-k+1{:}n-1)$.

B. Dataset

To generate a packet loss model through machine learning using an SVM, a sufficient number of packet traces is needed for off-line training. These traces are also needed to evaluate the performance of the proposed method in a real network environment. To this end, an open dataset [29] that contains end-to-end measurements of the one-way RTP traffic generated from a server to the measurement hosts located in a residential area is used. The server was located at the University of Glasgow. The residential measurement hosts were located in the United Kingdom and Finland. The dataset contains a maximum 3,800 RTP packet traces varying between 1 to 10 min in duration, when transmitting data at ranges of 1 to 8.5 Mbit/s according to the edge link capacity. Within the traces, a range of different characteristics on packet losses was shown. For example, the packet loss behavior changed according to the time of day, and there were no packet losses at all for many of the traces [29], [30].

C. Features

As the features representing the packet loss characteristics, a

Table 1. Feature set configured for proposed PLP.

Notation	Description (unit)
PLR_{avg}	Average packet loss rate (%)
BPL_{avg}	Average number of packet losses that occurred in the form of a burst (packets)
BPL_{min}	Minimum number of packet losses that occurred in the form of a burst (packets)
BPL_{max}	Maximum number of packet losses that occurred in the form of a burst (packets)
PL_d	Number of consecutive packets received without the loss from previous packet loss (packets)

feature set is configured as shown in Table 1. This set consists of five features. The first feature is defined as the average packet loss rate PLR_{avg}. The next three features are related to the burst characteristics of the packet loss, that is, the average, minimum, and maximum numbers of packet losses occurring in the form of a burst. These features are represented by BPL_{avg}, BPL_{min}, and BPL_{max}, respectively. The last feature is defined to take into account the statistical distance between a loss stream and lossless stream, that is, the number of consecutive packets received without a loss from previous packet loss, PL_d. Consequently, $X_n(m)$ in Fig. 5 consists of five components: PLR_{avg}, BPL_{avg}, BPL_{min}, BPL_{max}, and PL_d.

3. RST Mode Decision

To apply the proposed PLP instead of the PSQ estimation to the adaptive speech streaming, the RST mode, which was defined through (2) in Section II, is now modified by combining the PLP results of $\tilde{L}(n)$ and $\tilde{L}(n+1)$ with the speech classification results of $C(n)$ and $C(n+1)$, such that

$$M(n)=\begin{cases}0, & \text{if } \tilde{L}(n)=0 \text{ and } \tilde{L}(n+1)=0,\\ 1, & \text{if } \tilde{L}(n)=1 \text{ and } \tilde{L}(n+1)=0 \text{ or if } C(n)=1,\\ 2, & \text{if } \tilde{L}(n+1)=1 \text{ or if } C(n+1)=1.\end{cases} \tag{3}$$

In addition, to maintain the total bitrate for the speech data, the bitrate assignments for the PSD and RSD bitstreams according to different RST modes are conducted in the same manner as the PSQE-based ASSM described in Section II.

IV. Performance Evaluation

1. Accuracy of Packet Loss Prediction

To verify the effectiveness of the proposed PLP, the prediction accuracy for the packet loss was measured. Therefore, the RTP traffic dataset [29] described in the previous section was used. In particular, between the two datasets A and B provided from this RTP traffic dataset, the more recently measured dataset B was chosen for this experiment. In the dataset, the RTP traffic was measured during different dates and times for several networks. That is, a number of traces were provided for each network. Approximately three quarters were assigned to the off-line training, and the remaining were assigned to the prediction test.

Table 2. Number of lossless and loss packets used for off-line training and prediction test for performance evaluation of proposed PLP under different network conditions.

Network/sent RTP traffic (Mbps)	Packet loss state	Train (packets)	Test (packets)
ADSL4/1	Lossless	903,512	230,981
	Loss	100	8
ADSL4/2	Lossless	1,970,882	492,587
	Loss	141	202
ADSL4/5	Lossless	4,666,180	1,166,116
	Loss	460	553
ADSL5/1	Lossless	863,597	226,240
	Loss	553	85
ADSL5/2	Lossless	2,002,420	486,974
	Loss	3,503	307
ADSL5/5	Lossless	5,112,589	1,279,784
	Loss	12,287	1,386

Table 2 shows the detailed use of dataset B for off-line training and the conducting of a prediction test of the proposed PLP, where the number of packets for training and testing the proposed PLP was decomposed into that of lossless and loss packets, respectively. Note that the mean and maximum burst packet losses were measured to be approximately 1.4 and 22 packets, respectively. The SVM was trained using LIBSVM [23], where the gamma in a RBF was set to 0.2 by an exhaustive search. The number of packet loss information used to extract features (k in Fig. 5) was set to five packets.

Table 3 shows the performance of the proposed PLP for a given lossless or loss packet state under six different network conditions. The performance was measured by the prediction accuracy defined by the ratio of the number of packets predicted correctly to the total number of packets. In case of the lossless packet state, the number of packets that were predicted as lossless was counted, and it was divided by the total number of packets that were actually lossless. Similarly, the number of packets that were predicted as lost, out of actually lost packets, was counted for the lost packet state. The two leftmost columns indicate the input conditions, that is, the observed packet loss

Table 3. Prediction accuracies of proposed PLP under different network conditions.

Observation		Prediction accuracy (%)	
Network/Sent RTP traffic (Mbps)	Packet loss state	Lossless	Loss
ADSL4/1	Lossless	**100**	0
	Loss	0	**100**
ADSL4/2	Lossless	**99.999**	19.307
	Loss	0.001	**80.693**
ADSL4/5	Lossless	**100**	4.529
	Loss	0	**95.471**
ADSL5/1	Lossless	**100**	9.412
	Loss	0	**90.588**
ADSL5/2	Lossless	**99.999**	26.71
	Loss	0.001	**73.29**
ADSL5/5	Lossless	**99.994**	40.765
	Loss	0.006	**59.235**
Average	Lossless	**99.999**	16.787
	Loss	0.001	**83.213**

states for a given network. The two rightmost columns show the prediction results from the proposed PLP for each given network. As a result, the average prediction accuracy exceeded 99.99% and 83.21% for the lossless and loss packets, respectively.

As shown in Table 3, the prediction accuracy of predicting lossless packets was much higher than that of predicting loss packets for all network conditions. In particular, the prediction accuracy for the ADSL4 with a bitrate of 1 Mbps was perfect. This resulted from the small number of loss packets under this network condition, that is, there existed only eight packets, as shown in Table 2.

2. Improvement of Speech Quality

To demonstrate the effectiveness of the proposed PLP-based ASSM, an SW-based MCU was created using the AMR-NB as a multirate speech coder. The proposed method and other speech streaming methods were implemented. For the evaluation, input speech signals were sampled at 8 kHz and encoded using the AMR-NB speech encoder at a bitrate of 10.2 kb/s. The bitrate assignment for the PSD and RSD bitstreams according to the different RST modes was applied as shown in Table 4.

The performance of the proposed PLP-based ASSM within the equivalent transmission bandwidth was compared with those of two conventional speech streaming methods: a decoder-based PLC method [10] and a sender-based RST method [6], and the PSQE-based ASSM [13]. The decoder-based PLC method encoded speech signals using the AMR-NB encoder at 10.2 kb/s with no RSD bitstream. The sender-based RST method operated in two modes according to the number of RSD bitstreams; it first encoded speech signals using the AMR-NB encoder at a fixed bitrate of 7.95 kb/s with one RSD bitstream of 4.75 kb/s (such as $M(n) = 1$ in Table 4) and then encoded speech signals at a fixed bitrate of 4.75 kb/s with two RSD bitstreams of 4.75 kb/s (such as $M(n) = 2$ in Table 4). In this experiment, 24 speech samples were taken from the NTT-AT database. Each speech sample was composed of two utterances of approximately 4 s long, resulting in an actual length of 8 s, and was sampled at a rate of 16 kHz. Each speech utterance was filtered using a modified intermediate reference system (IRS) filter, followed by an automatic level adjustment [31]; they were subsequently downsampled from 16 kHz to 8 kHz.

Table 4. Bitrate assignment for different RST modes.

RST mode $M(n)$	$\overline{PSD}(n)$ (kb/s)			$\overline{RSD}(n)$ (kb/s)		
	R_{P0}	R_{P1}	R_{P2}	R_{R0}	R_{R1}	R_{R2}
0	10.2	N/A	N/A	N/A	N/A	N/A
1	N/A	7.95	N/A	4.75	N/A	N/A
2	N/A	N/A	4.75	N/A	4.75	4.75

To evaluate the quality of the decoded speech for each method, perceptual evaluation of speech quality (PESQ) scores were measured as defined by ITU-T Recommendation P.862 [32]. To transmit each speech sample under the various PLRs, each trace assigned for the test, as described in Table 2, was prepared to have packet loss information for as long as 8 s. In addition, to evaluate the various packet loss conditions, we randomly chose trace files whose PLR ranged from 1% to 11% in steps of 1%: from here, the maximum PLR was 11% because the dataset showed much lower overall PLRs. Note that the mean and maximum burst packet losses were measured to be approximately 1.2 and 22 packets, respectively.

Figure 6 compares the PESQ scores (in MOS) of the decoded speech processed using different speech streaming methods under different packet loss conditions. In the figure, each bar was drawn by averaging the MOS scores over all speech samples for each PLR, and the vertical line at the top of each bar denotes the standard deviation for a statistical significance test. As shown in the figure, the PESQ score of the proposed method was significantly better than those of the conventional methods for all PLRs. Moreover, the proposed PLP-based ASSM improved the average PESQ score over PLRs by as much as 0.65, 0.44, 0.61, and 0.53 MOS, as

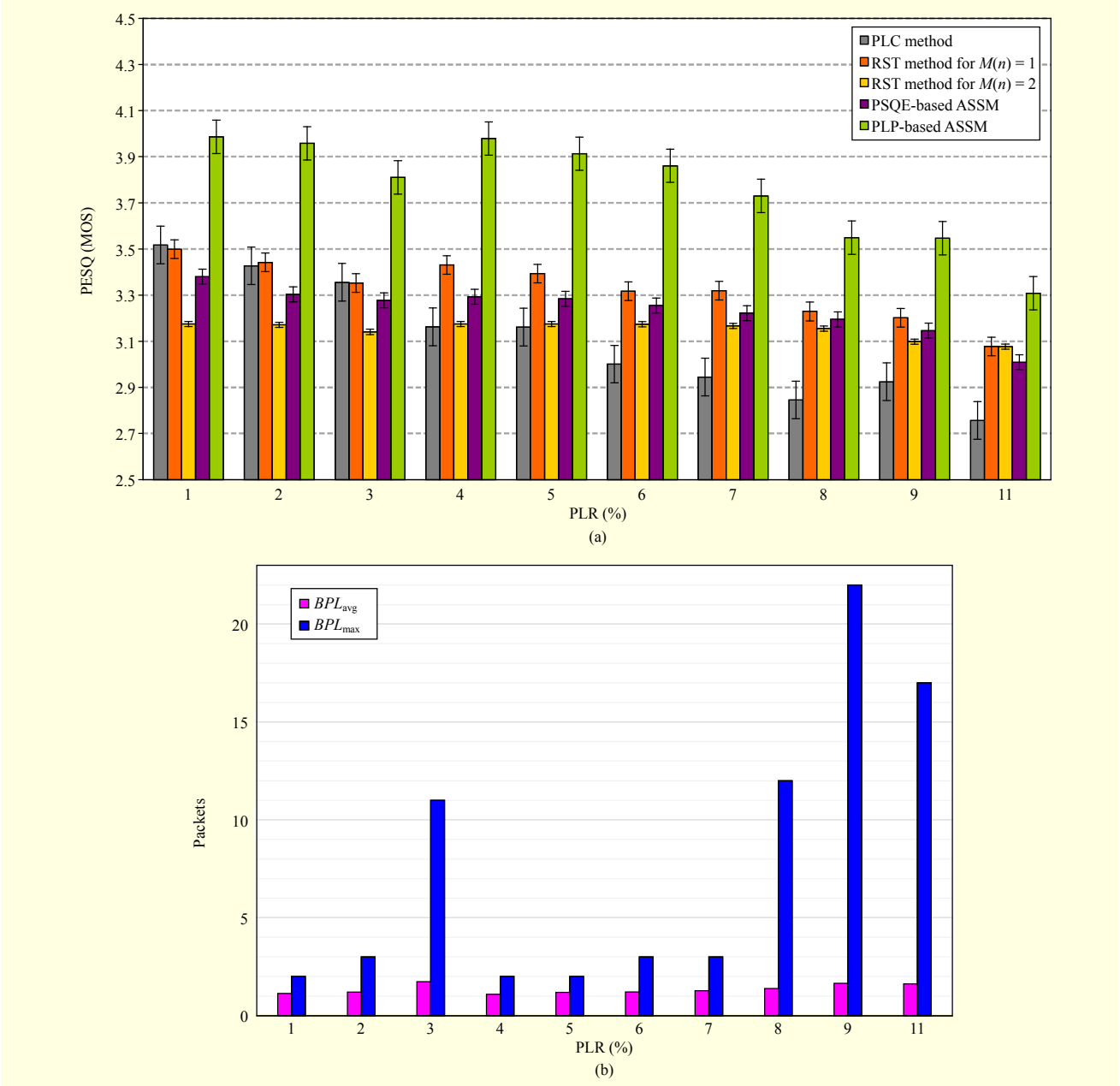

Fig. 6. Comparison of (a) PESQ scores measured in MOS for different speech streaming methods under various PLRs ranging from 1% to 11%, and (b) for burstiness of loss packets represented by BPL_{avg} and BPL_{max}.

compared to the PLC method, the RST method with two different modes, and the PSQE-based ASSM, respectively. Meanwhile, the PESQ scores did not decrease monotonically as PLR was increased. This was a result of the characteristics of the traces used in the experiment, where the packet losses occasionally occurred in a severe burst fashion even if the average PLR of a trace was low, as shown in Fig. 6(b).

V. Conclusion

In this paper, an adaptive speech streaming method was proposed to improve the perceived speech quality of an SW-based MCU over an IP network. To this end, the proposed method first predicts whether a speech packet to be transmitted is lost. Next, it classifies each frame of the input speech signals as either an onset frame or a non-onset frame. Using packet loss prediction and the speech class, the proposed method determines an appropriate bitrate for the RSD that was sent with the PSD to assist the speech decoder in restoring the speech signals of any lost packets.

The effectiveness of the proposed method was demonstrated by measuring the prediction accuracy for packet losses

observed in various packet traces, as well as by implementing an SW-based MCU employing the proposed method. A performance evaluation indicated that the proposed method provides a good level of performance for packet loss prediction with an average accuracy exceeding 99.99% and 83.21% for lossless and loss packets, respectively. In addition, the proposed method significantly improves the decoded speech quality relative to conventional methods under different PLR conditions ranging from 1% to 11%.

Acknowledgement

This work was supported by the "The Cross-Ministry Giga KOREA Project" grant from the Ministry of Science, ICT and Future Planning, Korea, Rep. of Korea (GK16P0100, Development of Tele-Experience Service SW Platform Based on Giga Media).

References

[1] Cisco, "Visual Networking Index: Forecast and Methodology, 2014–2019," Cisco Systems, Inc., San Jose, CA, USA, May 2015.

[2] Cisco, "TelePresence Packet Loss and Poor Audio/Visual Quality in One Direction," Cisco Systems, Inc., San Jose, CA, USA, Nov. 2014.

[3] 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, Aug. 2011, pp. 8469–8484.

[4] F. Merazka, "Improved Packet Loss Recovery using Interleaving for CELP-type Speech Coders in Packet Networks," *IAENG Int. J. Comput. Sci.*, vol. 36, Feb. 2009, pp. 1–5.

[5] W. Lizhong et al., "An Adaptive Forward Error Control Method for Voice Communication," *Int. Conf. Netw. Digital Soc.*, Wenzhou, China, May 30–31, 2010, pp. 186–189.

[6] I. Kouvelas et al., "Redundancy Control in Real-Time Internet Audio conferencing," *Int. Workshop Audio-Visual Services over Packet Netw.*, Aberdeen, UK, Sept. 14–15, 1997, pp. 195–201.

[7] K. Park et al., "A Dynamic Packet Recovery Mechanism for Realtime Service in Mobile Computing Environments," *ETRI J.*, vol. 25, no. 5, Oct. 2003, pp. 356–368.

[8] T. Wu et al., "An Enhanced Structure of Layered Forward Error Correction and Interleaving for Scalable Video Coding in Wireless Video Delivery," *IEEE Wireless Commun.*, vol. 20, no. 4, Aug. 2013, pp. 146–152.

[9] 3GPP TS 06.11, *Substitution and Muting of Lost Frames for Full Rate Speech Channels*, Nov. 2000.

[10] 3GPP TS 26.091, *Mandatory Speech Codec Speech Processing Functions; AMR Speech Codec; Error Concealment of Lost Frames*, Jan. 2010.

[11] N.I. Park, et al., "Burst Packet Loss Concealment using Multiple Codebooks and Comfort Noise for CELP-Type Speech Coders in Wireless Sensor Networks," *Sensors*, vol. 11, no. 5, May 2011, pp. 5323–5336.

[12] J. Huang, X. Zhang, and Y. Zhang, "Recovery of Lost Speech Segments using Incremental Subspace Learning," *ETRI J.*, vol. 34, no. 4, Aug. 2012, pp. 645–648.

[13] J.A. Kang et al., "Adaptive Speech Streaming Based on Speech Quality Estimation and Artificial Bandwidth Extension for Voice over Wireless Multimedia Sensor Networks," *Int. J. Distrib. Sensor Netw.*, vol. 2015, Apr. 2015, pp. 1–8.

[14] RFC 3550, *RTP: A Transport Protocol for Real-Time Applications*, July 2003.

[15] RFC 3267, *Real-Time Transport Protocol (RTP) Payload Format and File Storage Format for the Adaptive Multi-rate (AMR) and Adaptive Multi-rate Wideband (AMR-WB) Audio Codecs*, June 2002.

[16] 3GPP TS 26.101, *Mandatory Speech Codec Speech Processing Functions; Adaptive Multi-rate (AMR) Speech Codec Frame Structure*, Jan. 2010.

[17] L. Malfait, J. Bergerand, and M. Kastner, "P.563—the ITU-T Standard for Single-Ended Speech Quality Assessment," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 6, Nov. 2006, pp. 1924–1934.

[18] Y. Gao et al., "The SMV Algorithm Selected by TIA and 3GPP2 for CDMA Applications," *IEEE Int. Conf. Acoustics, Speech, Signal Process.*, Salt Lake City, Canada, May 7–11, 2001, pp. 709–712.

[19] J.A. Kang, *Packet Loss Robust Speech Streaming Techniques Based on Speech Quality Estimation*, Ph.D. Dissertation, School of Information and Mechatronics, Gwangju Institute of Science and Technology, Rep. of Korea, 2012.

[20] ITU-T Recommendation P.563, *Single-Ended Method for Objective Speech Quality Assessment in Narrow-Band Telephony Applications*, May 2004.

[21] NTT-AT, *Multi-lingual Speech Database for Telephonometry 1994*, NTT Advanced Technology Corp., Kanagawa, Japan, 1994.

[22] C. Cortes and V. Vapnik, "Support-Vector networks," *Mach. Learning*, vol. 20, no. 3, Sept. 1995, pp 273–297.

[23] C. Chang and C. Lin, "LIBSVM: A Library for Support Vector Machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, Apr. 2011, pp. 1–27.

[24] H. Yoon et al., "Improved Two-Phase Framework for Facial Emotion Recognition," *ETRI J.*, vol. 37, no. 6, Dec. 2015, pp. 1199–1210.

[25] A. Ameri et al., "Support Vector Regression for Improved Real-Time, Simultaneous Myoelectric Control," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 22, no. 6, Nov. 2014, pp. 1198–1209.

[26] J. Sun, J. Sun, and P. Chen, "Use of Support Vector Machine

Models for Real-Time Prediction of Crash Risk on Urban Expressways," *J. Trans. Res. Board*, vol. 2432, 2014, pp. 91–98.

[27] T. Nagano and A. Ito, "Packet Loss Concealment of Voice-over IP Packet using Redundant Parameter Transmission under Severe Loss Conditions," *J. Inform. Hiding Multimedia Signal Process.*, vol. 5, no. 2, Apr. 2014, pp. 286–295.

[28] H. Hsu et al., "Speech Attribute Classifier using Support Vector Machine for Speech Packet Loss Concealment," *Int. Committee Co-ordination Standardization Speech Databases Assessment Techn.*, Macau, China, Dec. 9–12, 2012, pp. 68–71.

[29] M. Ellis, C. Perkins, and D. Pezaros, "End-to-end and Network-Internal Measurements of Real-Time Traffic to Residential Users," *ACM Multimedia Syst. Conf.*, Santa Clara, CA, USA, Feb. 23–25, 2011, pp. 111–116.

[30] M. Ellis et al., "A Two-Level Markov Model for Packet Loss in UDP/IP-based Real-Time Video Applications Targeting Residential Users," *Comput. Netw.*, vol. 70, Sept. 2014, pp. 384–399.

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

[32] ITU-T Recommendation P.862, *Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs*, Feb. 2001.

Jin Ah Kang received her BS and MS degrees in Telecommunications Engineering from Jeju National University, Rep. of Korea, in 2001 and 2005, respectively, and her PhD degree in Information and Communications from Gwangju Institute of Science and Technology, Rep. of Korea, in 2012. Since 2014, she has been with ETRI, where she is now a senior researcher. Her main research interests include speech/audio signal processing and multimedia streaming with applications in immersive services.

Mikyong Han received her MS degree in Computing Engineering from the School of Electronics and Information, Kyung Hee University, Seoul, Rep. of Korea, in 1993. She joined ETRI in 1993 and is currently a principal research member as well as a director of the Real & Emotional Sense Convergence Service Platform Research Laboratory. From March 2012 to February 2013, she was a visiting professor at the School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, USA. Her major research interests include multimedia service platforms, emotional service platforms, and immersive media service platforms.

Jong-Hyun Jang received his PhD degree in Computer Science and Engineering from Hankuk University of Foreign Studies, Yongin, Rep. of Korea, in 2004. Since 1994, he has been with the ETRI, where he is currently a principal researcher as well as an executive director of the Giga Service Research Department. He has worked on several projects for the development of programming environments and PCS since 1994. His research interests are real-time middleware for telecommunication systems, home networking systems, and giga media services.

Hong Kook Kim received his BS degree in Control and Instrumentation Engineering from Seoul National University, Rep. of Korea, in 1988, and his MS and PhD degrees in Electrical Engineering from the Korea Advanced Institute of Science and Technology, Daejeon, Rep. of Korea, in 1990 and 1994, respectively. He was a senior researcher at the Samsung Advanced Institute of Technology, Seoul, Rep. of Korea, from 1990 to 1998. During 1998–2003, he was a senior technical staff member with the Voice Enabled Services Research Lab at AT&T Labs-Research, Florham Park, USA. Since August 2003, he has been with the School of Electrical Engineering and Computer Science at Gwangju Institute of Science and Technology, Rep. of Korea, as a professor. During 2014-2015, he was a visiting professor at the City University of New York, USA. He is currently an IEEE Senior Member and an affiliate member of the IEEE Speech and Language Technical Committee. Since 2012, he has served as an editorial committee member and area editor of Digital Signal Processing. His current research interests include large vocabulary speech recognition, audio coding and speech/audio source separation, and embedded algorithms and solutions for speech and audio processing for handheld devices.