A Smart Error Protection Scheme Based on Estimation of Perceived Speech Quality for Portable Digital Speech Streaming Systems

Jin Ah Kang and Hong Kook Kim

School of Information and Communications
Gwangju Institute of Science and Technology (GIST), Gwangju 500-712, Korea
{jinari,hongkook}@gist.ac.kr

Abstract. In this paper, a smart error protection (SEP) scheme is proposed to improve speech quality of a portable digital speech streaming (PDSS) system via a lossy transmission channel. To this end, the proposed SEP scheme estimates the perceived speech quality (PSQ) for received speech data, and then transmits redundant speech data (RSD) in order to assist speech decoder to reconstruct lost speech signals for high packet loss rates. According to the estimated PSQ, the proposed SEP scheme controls the RSD transmission, and then optimizes a bitrate of speech coding to encode the current speech data (CSD) against the amount of RSD without increasing transmission bandwidth. The effectiveness of the proposed SEP scheme is finally demonstrated using adaptive multirate-narrowband (AMR-NB) and ITU-T Recommendation P.563 as a scalable speech codec and a PSQ estimator, respectively. It is shown from experiments that a PDSS system employing the proposed SEP scheme significantly improves speech quality under packet loss conditions.

Keywords: Portable digital speech streaming systems, packet loss, error protection, perceived speech quality, redundant speech transmission.

1 Introduction

Due to the rapid development of Internet protocol (IP) networks over the past few decades, audio and video streaming services are increasingly available via the Internet. Moreover, as these services are extended to wireless networks, the quality of service (QoS) of audio and video streaming is becoming even more critical. Specifically, portable digital speech streaming (PDSS) systems require a minimum level of speech communication quality, where the speech quality is largely related to the network conditions such as packet losses or end-to-end packet delays [1]. When the speech streaming is performed via user datagram protocol/IP (UDP/IP) networks, however, packets may be lost or arrive too late for playback due to inevitable delays. In this case, a typical PDSS system can only tolerate a few packet losses for real-time services, where these packet losses frequently occur in wireless networks due to bandwidth fluctuations [2].

T.-h. Kim et al. (Eds.): UCMA 2011, Part II, CCIS 151, pp. 1–10, 2011.

Several packet loss recovery methods, implemented via the Internet and wireless networks, have been proposed for the speech streaming systems. For instance, the techniques proposed in [3] and [4] were sender-based packet loss recovery methods using forward error correction (FEC). In regards to wireless networks, the techniques proposed in [5] and [6] were based on unequal error protection (UEP) methods. In addition, the modified discrete cosine transform (MDCT) coefficients of audio signals were used as the redundant data in order to assist an MP3 audio decoder to reconstruct lost audio signals [7]. However, these methods did not take into account time-varying network conditions, i.e., packet loss rate (PLR). That is, in order to recover the lost packets based on the conventional FEC methods, the redundant data should be designed to be transmitted constantly even if the network conditions are declared as no packet losses.

Therefore, a smart error protection (SEP) scheme is needed that recovers packet losses efficiently according to the time-varying characteristic of PLR. Towards this goal, this paper proposes an SEP scheme that transmits redundant speech data (RSD) adaptively according to the estimation of perceived speech quality (PSQ). To this end, the PSQ estimation is performed in real-time for received speech data by using a single-ended speech quality assessment. Simultaneously, the PLR is estimated by a moving average method. In addition, a real-time transport protocol (RTP) payload format is newly suggested as a means of supporting the proposed SEP scheme. In other words, a speech packet combines the bitstreams of the current speech data (CSD) and the RSD when the PLR is assumed to be high by the estimated PSQ and PLR. Thus, even if a speech packet is lost, the speech decoder can reconstruct the lost speech signal by using the RSD bitstreams from the previous packet. On the other hand, when the PLR is assumed to be low by the estimated PSQ and PLR, a speech packet is organized using the CSD bitstreams alone that are encoded by a higher bitrate. The effectiveness of the proposed SEP scheme is finally demonstrated by using the adaptive multirate-narrowband (AMR-NB) speech codec [8] and ITU-T Recommendation P.563 [9] as a scalable speech codec and a single-ended speech quality assessment, respectively.

The remainder of this paper is organized as follows. Following this introduction, Section 2 presents the structure of a PDSS system based on the proposed SEP scheme and the RTP payload format for the proposed SEP scheme. Next, Section 3 describes the proposed SEP scheme in detail, and the performance of the proposed SEP scheme is discussed in Section 4. Finally, Section 5 concludes this paper.

2 A Portable Digital Speech Streaming System

2.1 Overview

A PDSS system extends traditional speech communication services over a public switched telephone network (PSTN) to wireless networks in order to provide various mobile communication services. To this end, the PDSS system samples a continuous speech signal to discontinuous speech frames, and it encodes the speech frames to bitstreams at a lower bitrate by using a compression algorithm. Then, it transmits the bitstreams using a real-time streaming protocol after packetizing. Meanwhile, at the

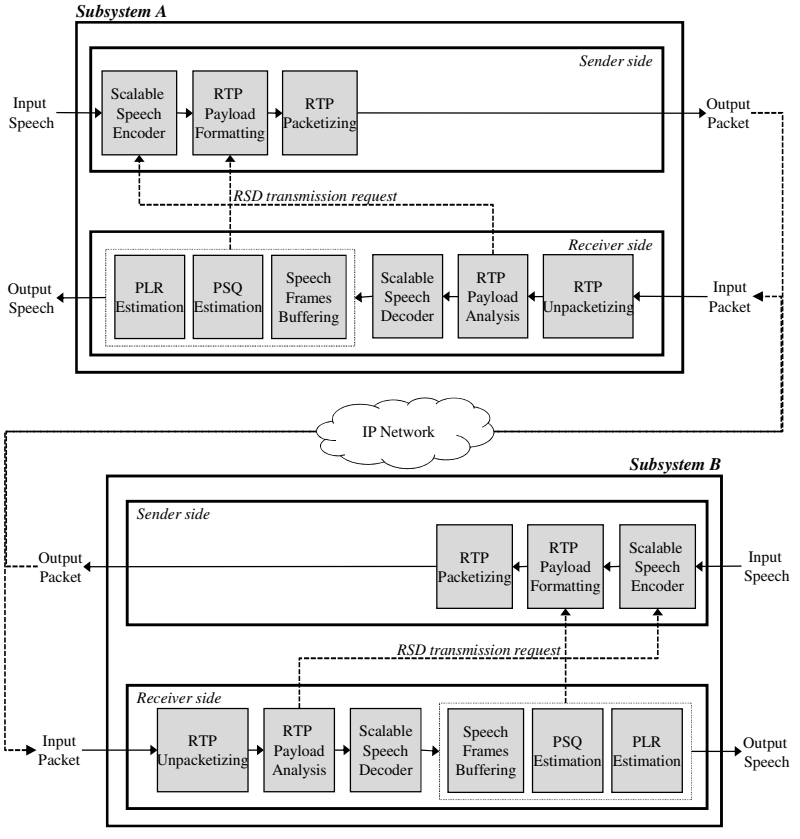

Fig. 1. Packet flow for a PDSS system employing the proposed SEP scheme, where *Subsystems A* and *B* represent the two communication parties

opposite PDSS system, the arriving packets are unpacketized to bitstreams, and the bitstreams are decoded to the speech frames. Finally, these speech frames are sent to an output device.

Fig. 1 shows a packet flow for the PDSS system implemented in this paper, where *Subsystems A* and *B* represent both parties of the speech stream communication employing the proposed SEP scheme. First, the sender side of *Subsystem A* performs scalable speech encoding for the input speech frame. Next, the sender side generates a packet according to an RTP payload format, where the packet includes the CSD bitstreams with the decision result whether or not the RSD transmission is needed. Note here that the RSD bitstreams should be incorporated in this payload when the RSD transmission is requested by *Subsystem B*. After that, the formatted RTP packet is transmitted.

Meanwhile, as the RTP packet arrives at the receiver side of *Subsystem B*, the receiver side analyzes the received packet according to the RTP payload format, and then extracts the CSD bitstreams and the decision result. In the case that the RTP payload format includes the RSD bitstreams, the RSD bitstreams are used to recover a

lost packet in the future. Next, the extracted CSD bitstreams are decoded using a scalable speech decoder and the decoded speech frames are stored in a speech buffer to be used for the PSQ estimation. Finally, the decision result regarding the RSD transmission is inserted into a RTP packet before a speech frame is sent to *Subsystem A*.

2.2 RTP Payload Format

As mentioned in Section 2.1, a PDSS system employing the proposed SEP scheme can have an indicator for a scalable bitrate of speech coding. Moreover, in order to deliver the feedback information from *Subsystem A* to *Subsystem B*, and vice versa, there should be any fields reserved in the format to accommodate the transmission of RSD bitstreams and feedback information. Thus, we first select the RTP payload format defined in IETF RFC 3267 for the AMR-NB speech codec [10], as shown in Fig. 2.

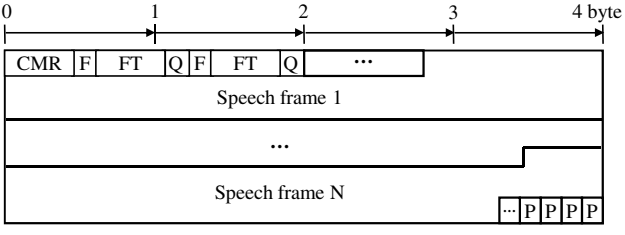

Fig. 2. Example of the RTP payload format for AMR-NB speech codec defined in RFC 3267

In the payload format, an 'F|FT|Q' sequence of control fields is used to describe each speech frame. Note here that a codec mode request (CMR) field is applied to the entire speech frame. In other words, a one-bit F field indicates whether this frame is to be followed by another speech frame (F=1) or if it is the final speech frame (F=0). In addition, an FT field, comprised of 4 bits, then indicates if this frame is actually coded by a speech encoder or if it is just comfort noise. That is, a number in this field is assigned from 0 to 7, corresponding to encoding bitrates of 4.75, 5.15, 5.90, 6.70, 7.40, 7.95, 10.2, and 12.2 kbit/s, respectively. However, if comfort noise is encoded, the assigned number ranges from 8 to 11. Note that the number 15 indicates the condition that there is no data to be transmitted, and that the numbers 12 to 14 are reserved for future use. Next, a Q field, indicating the speech quality with one-bit, is set at 0 when the speech frame data is severely damaged; otherwise, it is set at 1. Finally, the CMR field, comprised of 4 bits, is used to deliver a mode change signal to the speech encoder. For example, it is set to one out of eight encoding modes, corresponding to different bitrates of AMR-NB speech codec. At the end of the payload, P fields are used to ensure octet alignment. In order to realize the proposed SEP scheme in this payload format, two new frame indices for the RSD bitstreams and the feedback information are incorporated into the FT field, which are denoted using the numbers 12 and 13, respectively.

The use of the RTP payload format described above has several advantages. First, the control ability for a speech encoder, such as the CMR field, is retained by using the

RTP payload format for the speech codec employed in the implemented PDSS system. Next, the overhead of the control fields for each RSD bitstreams is required to be as small as 6 bits in 'F|FT|Q'. Finally, no additional transport protocol for the RSD transmission request is needed since this feedback is conducted using RTP packets that are used to deliver the speech bitstreams. Therefore, the transmission overhead for the RSD transmission request is significantly reduced, compared to existing transport protocols designed for feedback such as the RTP control protocol (RTCP) [11].

3 Proposed Smart Error Protection Scheme

3.1 Packet Loss Recovery and PSQ Estimation at the Receiver Side

Fig. 3 presents the procedure of packet loss recovery with the PSQ estimation at the receiver side of a PDSS system employing the proposed SEP scheme. First, a packet loss occurrence is verified through RTP packet analysis. Then, the received CSD bitstreams are decoded if it is decided that there is no packet loss. On the other hand, if it is decided that there is a packet loss, the lost speech signals are recovered by using the RSD bitstreams or by using the packet loss concealment (PLC) algorithm in the speech decoder, depending on the availability of the RSD bitstreams. Finally, the speech decoder reconstructs the speech frame data from the CSD bitstreams, and estimate PSQ and PLR with speech data once the amount of speech frames is enough to estimate a PSQ score.

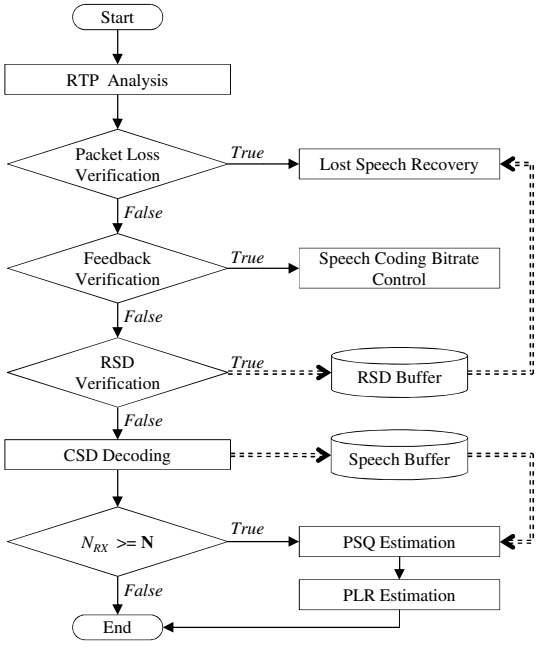

Fig. 3. Procedure of the packet loss recovery with the PSQ estimation at the receiver side

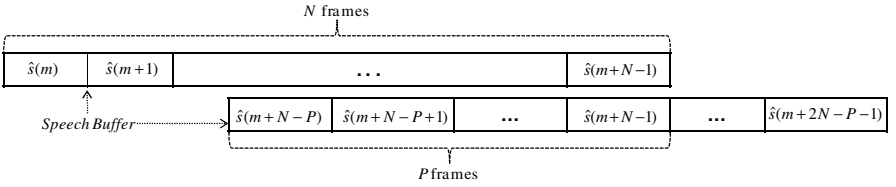

Fig. 4. Overlap of speech frames for the PSQ estimation at the receiver side

For the PSQ estimation, the speech data in a speech buffer are used by overlapping, as shown in Fig. 4. In the figure, $\hat{s}(m)$ is the m-th speech frame input to the speech buffer, N is the total number of frames to be used for the PSQ estimation, and P is the number of frames to be overlapped for the next PSQ estimation. In other words, the PSQ estimation is conducted when every (N-P) frames are newly received from the opposite PDSS system. In addition, the estimated PLR, $\hat{L}(k)$, is obtained by moving average for the previous PLR, $L(k-1)$, with the average PLR, $\overline{L}(0:k-1)$, as

$$\hat{L}(k) = (1-\alpha)\,\overline{L}(0:k-1) + \alpha\, L(k-1) \tag{1}$$

Finally, it is decided whether or not requesting the RSD transmission by comparing the estimated PSQ and PLR with each threshold. That is, the request for the RSD transmission, $RSD(k)$, is set to true or false according to the equation of

$$RSD(k) = \begin{cases} true, & if\ \hat{Q}(k) \le Thres_1\ \ and\ \ \hat{L}(k) \ge Thres_2 \\ flase, & otherwise \end{cases} \tag{2}$$

where $\hat{Q}(k)$ is the estimated PSQ score, and $Thres_1$ and $Thres_2$ are threshold for $\hat{Q}(k)$ and $\hat{L}(k)$, respectively.

3.2 Scalable Speech Coding and RSD Transmission at the Sender Side

Fig. 5 shows the procedure how to transmit scalable speech coding bitstreams and the RSD bitstreams at the sender side for the proposed SEP scheme. First, for given feedback information transmitted from the opposite PDSS streaming system, the sender side verifies the request for the RSD transmission and changes the bitrate of scalable speech coding according to the request. In other words, when the RSD transmission is not requested, the bitrate is set at the highest bitrate and then the CSD bitstreams are encoded alone with no additional RSD bitstreams. On the other hand, when the RSD transmission is requested, the bitrate is set at smaller bitrate than the current bitrate in order to assign the remaining bitrate for the RSD transmission. Thus, both of the CSD and RSD bitstreams are encoded. Finally, after the RTP payload format described in Section 2.2 is configured according to such adaptive RSD transmission, the RTP packets are transmitted to the opposite PDSS system.

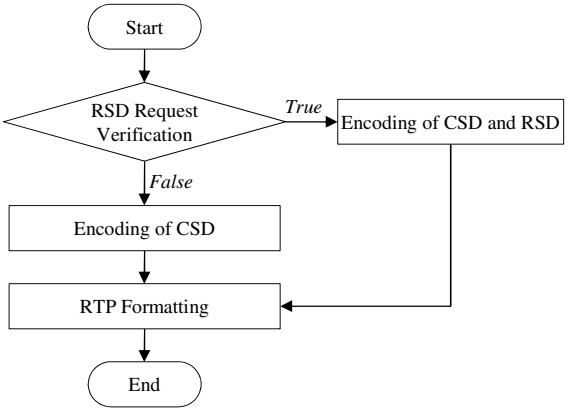

Fig. 5. Procedure of the scalable speech coding and the adaptive RSD transmission at the sender side

As described above, we can several advantages of the proposed SEP scheme as follows. First, the adaptive operation of the packet loss recovery according to the network conditions is effective since burst packet losses generally occur when the network is congested due to a sudden increase in the amount of data coming to the network [4]. Second, compared to the conventional redundant data transmission (RDT) methods that require additional network overhead, the proposed SEP scheme generates redundant data without increasing the transmission bandwidth by controlling the bitrate of a scalable speech codec. Third, in order to estimate the network conditions, the proposed SEP scheme conducts the estimation of PSQ. This is motivated by the fact that the PSQ measured as a mean opinion score (MOS) can be considered to be a clearer indicator of the speech quality than other parameters in the PDSS system.

4 Performance Evaluation

In order to demonstrate the effectiveness of the proposed SEP scheme, the PDSS system was first implemented by using the AMR-NB speech codec and the ITU-T Recommendation P.563 as a scalable speech codec and a PSQ estimator, respectively. Here, the speech signals were sampled at 8 kHz, and then encoded using the AMR-NB speech codec operated at 10.2 kbit/s. Thus, when the RSD transmission was needed, the bitrate of the CSD and RSD was set at 4.75 kbit/s, which was almost half the bitrate of 10.2 kbit/s. By considering the requirements of the ITU-T Recommendation P.563, N in Fig.4 was set to 200 frames for the PSQ estimation, which corresponded to 4 seconds. Moreover, P was set to 150 frames, thus the PSQ estimation was conducted whenever every new 50 frames were received. For the PLR estimation, we carried out performance evaluation of the proposed SEP scheme with the different value of α in Eq. (1), and then we set α to 0.4. Similarly, the thresholds for the estimated PSQ and PLR in Eq. (2), $Thres_1$ and $Thres_2$, were set to 4.0 MOS and 5%, respectively.

In the test, 48 speech files from NTT-AT speech database [12] were used, where each speech file was about 4 seconds long and sampled at a rate of 16 kHz. These speech signals were first filtered using a modified intermediate reference system (IRS) filter followed by an automatic level adjustment [13]. Then, the speech signals were down-sampled from 16 to 8 kHz. In order to show the effectiveness of the proposed SEP scheme under different PLRs including burst loss characteristics, we generated five different PLR patterns of 3, 5, 7, 9 and 11% by using the Gilbert-Elliot channel model defined in the ITU-T Recommendation G.191 [13]. Here, the burstiness of the packet losses was set at 0.5, and the mean and maximum consecutive packet losses were measured at 1.5 and 4.0 frames, respectively.

In order to demonstrate the effectiveness of the proposed SEP scheme, the speech quality of the PDSS system with the proposed SEP scheme was compared to that of the PDSS system with the regular RDT. The regular RDT was designed to transmit the RSD regularly for each speech frame by encoding the CSD and RSD bitstreams at a bitrate of 4.75 kbit/s in order to evaluate the performance without increasing transmission bandwidth.

In addition, the speech quality of the PDSS system by the PLC algorithm without the proposed SEP scheme or the regular RDT was also evaluated, where the PLC algorithm was operated at the highest bitrate of 10.2 kbit/s. Note here that the PLC algorithm embedded in the AMR-NB speech decoder was always applied without regarding to the RSD transmission. As the evaluation method for the recovered speech quality, the perceptual evaluation of speech quality (PESQ) defined in the ITU-T Recommendation P.862 [14] was used.

Table 1. Speech quality measured in MOS using PESQ for the different packet loss recovery methods, under PLRs ranging from 3 to 11%

MOS Score / Method	PLR (%)						Average
	0	3	5	7	9	11	
Without the regular RDT	3.70	3.16	2.97	2.83	2.61	2.52	2.96
With the regular RDT	3.14	3.07	2.97	2.91	2.80	2.78	2.94
With the proposed scheme	3.70	3.16	2.93	2.87	2.71	2.72	3.01

Table 1 compares speech quality measured in MOS using PESQ for the different packet loss recovery methods, under PLRs ranging from 3 to 11%. As shown in the table, the proposed SEP scheme first improved the speech quality for low PLRs as the PLC algorithm without using regular RDT did. In addition, the proposed SEP scheme provided better performance than without the regular RDT as the regular RDT did for high PLRs. Consequently, the proposed SEP scheme yielded the average speech quality of 3.01 MOS, which was 0.07 MOS higher than the regular RDT.

5 Conclusion

In this paper, we proposed a new smart error protection (SEP) scheme that guaranteed the speech quality without increasing transmission bandwidth for a portable digital speech streaming system (PDSS). To this end, the proposed SEP scheme was

designed to transmit redundant speech data (RSD) according to the estimation results for the perceived speech quality (PSQ) and packet loss rate (PLR), where a single-ended speech quality assessment and a moving average method were used to estimate PSQ and PLR, respectively. The proposed SEP scheme was applied to the receiver and sender sides of a PDSS system. In other words, the receiver side of the PDSS system first decided the RSD transmission based on the estimation of PSQ and PLR, and then sent feedback information on the decision result to the opposite PDSS system via real-time transport protocol (RTP) packets for speech bitstreams. On the other hand, the sender side of the PDSS system controlled the RSD transmission according to the received feedback, and subsequently optimized the speech coding bitrate in order to maintain the equivalent transmission bandwidth despite of the RSD bitstreams. Finally, we evaluated the speech quality recovered by the proposed SEP scheme under PLRs and compared it with that of the conventional redundant data transmission (RDT) method. From the results, the proposed SEP scheme improved the speech quality from 2.94 to 3.01 MOS compared than the conventional method for the PLRs ranged from 3% to 11%. Consequently, the proposed SEP scheme could be applied to the PDSS streaming systems in order to improve the speech quality degraded due to packet losses efficiently.

Acknowledgments. This work was supported in part by the "Fusion-Tech Developments for THz Information & Communications" Program of the Gwangju Institute of Science and Technology (GIST) in 2011, by the Mid-career Researcher Program through the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2010-0000135), and by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2010-C1090-1021-0007).

References

1. Wu, C.-F., Lee, C.-L., Chang, W.-W.: Perceptual-based playout mechanisms for multi-stream voice over IP networks. In: Proceedings of Interspeech, Antwerp, Belgium, pp. 1673–1676 (September 2007)
2. 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), 897–909 (2004)
3. Bolot, J.-C., Fosse-Parisis, S., Towsley, D.: Adaptive FEC-based error control for Internet telephony. In: Proceedings of IEEE International Conference on Computer Communications (INFOCOM), New York, NY, pp. 1453–1460 (March 1999)
4. Jiang, W., Schulzrinne, H.: Comparison and optimization of packet loss repair methods on VoIP perceived quality under bursty loss. In: Proceedings of 12th International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV), Miami, FL, pp. 73–81 (May 2002)
5. Yung, C., Fu, H., Tsui, C., Cheng, R.S., George, D.: Unequal error protection for wireless transmission of MPEG audio. In: Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS), Orlando, FL, pp. 342–345 (May 1999)
6. Hagenauer, J., Stockhammer, T.: Channel coding and transmission aspects for wireless multimedia. Proceedings of the IEEE 87, 1764–1777 (1999)

7. Ito, A., Konno, K., Makino, S.: Packet loss concealment for MDCT-based audio codec using correlation-based side information. International Journal of Innovative Computing, Information and Control 6, 3(B), 1347–1361 (2010)
8. ETSI 3GPP TS 26.101: Adaptive Multi-Rate (AMR) Speech Codec Frame Structure (January 2010)
9. ITU-T Recommendation P.563: Single-Ended Method for Objective Audio Quality Assessment in Narrow-Band Telephony Applications (May 2004)
10. IETF 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)
11. IETF RFC 1889: RTP: A Transport Protocol for Real-Time Applications (January 1996)
12. NTT-AT: Multi-Lingual Speech Database for Telephonometry (1994)
13. ITU-T Recommendation G.191: Software Tools for Speech and Audio Coding Standardization (November 1996)
14. 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 (February 2001)