A Packet-layer Quality Assessment System for VoIP

Liangliang Jiang[1] and Fuzheng Yang[2]

Abstract. A packet-layer quality assessment system is proposed to monitor the quality of VoIP service in this paper. The proposed system is composed of three modules, where the quality assessment module evaluates the voice quality in terms of the information extracted by the packet header analysis module and silence detection module. It is noteworthy that the voiced segments and silence segments in the voice streaming are distinguished and treated differently in the evaluation of voice quality. Taking adaptive multi-rate (AMR) encoded stream over RTP/UDP/IP (Real-time Transport Protocol/ User Datagram Protocol/Internet Protocol) as an example, the process of the voice quality assessment employing the proposed system is described in detail. Experimental results reveal that the estimation model in the proposed system achieves superior performance over the compared models.

Keywords: voice quality assessment· quality of experience· packet loss· VoIP

1 Introduction

In recent years, voice over IP (VoIP) has gained wide popularity owing to its features of high bandwidth efficiency and low costs. However, VoIP provides best-effort service, and the quality of experience (QoE) is greatly difficult to guarantee. Consequently, an effective objective method for voice quality assessment is indispensable for QoE planning and quality monitoring.

According to the used information, objective voice quality assessment methods can be classified into five categories: parametric planning model, packet-layer

[1]L. Jiang (✉)
State Key Laboratory of Integrated Service Networks, Xidian University，Xi'an, China
e-mail: lljiang@stu.xidian.edu.cn

[2]F. Yang
State Key Laboratory of Integrated Service Networks, Xidian University，Xi'an, China
e-mail: fzyang@mail.xidian.edu.cn

model, bitstream-layer model, media-layer model and hybrid model [1]. Specifically, the parametric planning model is initially designed for QoE planners to estimate the voice quality crudely [2], it requires prior information on the measured system. As the name implies, the packet-layer model predicts the voice quality only employing the information from packet headers without resorting to any media-related payload information [3][4]. The bitstream-layer model formulates an evaluation of voice quality with the information carried by the encoded bitstream apart from packet-layer, whose performance significantly depends on the level of access to the bitstream. Generally speaking, the more information can be employed, the better performance will be gained. The media-layer model is also known as signal-based model, which exploits the reconstructed signal to assess the voice quality. The hybrid model is a combination of the aforementioned models, i.e. it can freely utilize any information exploited in the aforementioned models.

Regarding networked voice, one of the most important operational issues for providing end users with stable and satisfactory QoE is to enable real-time quality monitoring, where the packet-layer model serves as a good solution [5]. Since only analyzing the packet headers, the packet-layer model is very efficient in computational complexity, and is quite useful at network inter-nodes. Another advantage is that the packet-layer model does not need decryption and voice decoding, making it favorable when the packet payloads are encrypted. The packet-layer model has been extensively discussed in ITU-T SG12 and is to be standardized as a new recommendation named as P.NAMS [5]. As a typical example, ITU-T recommendation G.107, i.e., the E-model, is usually used as a packet-layer model by extracting model parameters from packet header [2]. Specific packet-layer models have been proposed to assess the voice quality [3][4]. However, the significance of the voice content in different parts may be different, which has not been taken account of when evaluating the voice quality. In this paper, we propose a packet-layer quality assessment system to predict the quality of networked voice, where the significant parts and insignificant parts in the voice streaming are distinguished and treated differently in the estimation model of voice quality. The proposed system provides a comprehensive evaluation of transmission quality for VoIP by only utilizing the information extracted from packet headers. Taking adaptive multi-rate (AMR) encoded stream over RTP/UDP/IP (Real-time Transport Protocol/ User Datagram Protocol/Internet Protocol) as an example, the detailed process of the quality assessment in the proposed system is described in this paper.

2 Framework of The Proposed Voice Quality Assessment System

The reconstructed voice signal on the client side can be regarded as being mainly impaired by two factors: lossy compression and packet loss. Consequently, the quality assessment for networked voice should take account of the distortions caused by these two factors, and the information about the coding and packet loss

parameters is highly significant. Accordingly, the first step of the proposed system is to extract the information from packet headers by the packet header analysis module.

Specially, the voice sequence consists of two types of segments with different significance, i.e., voiced segments and silence segments, where voiced segments are nearly the whole part to influence the QoE [6]. Therefore the proposed system introduces a silence detection module to distinguish voiced segments and silence segments with resorting to the extracted information from packet headers.

Given the information extracted by the packet analysis module and silence detection module, the quality assessment module predicts the voice quality only in terms of the coding and packet loss parameters of voiced segments. The framework of the proposed system is illustrated in Fig.1.

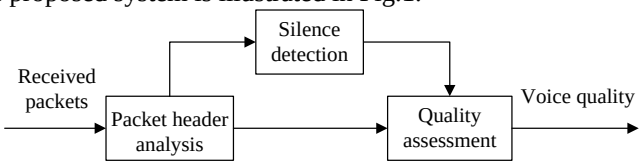


Fig.1 Framework of the packet-layer voice quality assessment system

3 Information Extraction

3.1 Packet Header Analysis

It is worth pointing that the VoIP service selects RTP/UDP/IP instead of HTTP/TCP/IP (Hypertext Transfer Protocol/Transmission Control Protocol/Internet Protocol) as transport protocols, since the backoff and retransmission mechanism in TCP may lead to uncomfortable delays, which can't meet the real-time requirement in voice communication. The packet structure under RTP/UDP/IP transport protocols is shown in Fig.2. Since the lengths of both UDP header and RTP header are known, the length of the payload can be obtained according to the length field of UDP header, which indicates the total length of the UDP header, RTP header and payload. The timestamp field of RTP header is an important mark, which reflects the sampling instant of the first octet in the RTP packet. Thus the duration of the measured voice (excluding the last packet) can be calculated in terms of the timestamp of the first packet and the last packet, and the duration of the last packet can be estimated with the duration of its previous packet. Moreover, the timestamp can be also used to combine the packets associated with the same frame, because when a voice frame is partitioned and packeted into several packets to adapt the network environment, the timestamp are same for

these packets. By means of the analysis above, the bit-rate and frame size can be deduced.

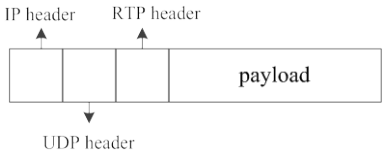


Fig.2 Packet structure under RTP/UDP/IP transport protocols

According to the rule of RTP, the sequence number field of RTP header increases by one for each RTP data packet sent. Therefore, packet loss can be detected by monitoring the sequence number field of RTP header, meanwhile the number of lost packets is directly determined.

3.2 Silence Detection

The assessment parameters in the proposed system are computed only depending on the information of voiced packets, it is therefore crucial to detect the packet type. Given the packet header information, the frame type can be predicted according to the apparent difference between the size of a voiced and a silence frame. Specifically, the AMR codec utilizes the voice activity detection and comfort noise generation techniques to reduce bandwidth usage during speech pauses. The size of silence frame is 1 or 6 bytes, while the minimum size of voiced frame is 13 bytes in the AMR codec, significantly larger than the size of silence frame [7]. The frame size can be obtained by packet header analysis module.

The adjacent packets usually have the same type because of intensive short-time relativity in voice signal. For a lost packet, the packet type is estimated based on the type of its two adjacent non-lost packets [6]. If the two packets are both voiced, the lost packet is judged as voiced. On the other hand, if the two packets are both silence, the lost packet is judged as silence. It is noted that sometimes the two adjacent packets have different types, and then the lost packet may be voiced or silence. In this case, it is always judged as voiced in the proposed system since voiced packets account for larger percentage than silence packets in a voice sequence.

4 Voice Quality Assessment

In the proposed system, after the analysis on packet headers and the detection of silence segments, the quality assessment module executes to estimate the quality of the voice streaming, as illustrated in Fig.3. The perceived quality regarding coding distortion is first predicted, based on which the quality degradation due to packet loss is further estimated by making use of the information about lost voiced packets. Finally, the overall quality of the voice streaming is obtained through integrating the coding distortion and the distortion due to packet loss.

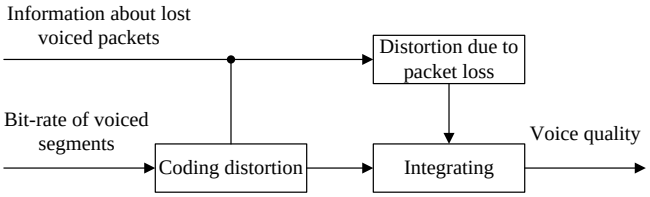


Fig.3 Framework of the quality assessment module

Subject to the limited available information in packet-layer level, the overall bit-rate is a major parameter to predict the coding distortion [6]. Considering the great significance of voiced segments, the coding distortion is estimated based on the average bit-rate of voiced segments in the proposed system. The AMR codec consists of eight coding modes with bit-rates of 4.75, 5.15, 5.90, 6.70, 7.40, 7.95, 10.2 and 12.2 kb/s, these bitrates are exactly the bitrates of voiced segments. For each bit-rate, the corresponding quality regarding coding distortion is shown in Fig.4. Here the quality regarding coding distortion is the average of the scores of the compressed training sequences under each bit-rate, which are measured by the perceptual evaluation of speech quality (PESQ) algorithm [8]. The selection of the training sequences will be discussed in detail in the next section. Since the AMR codec is capable of dynamically selecting the bit-rate to adapt the channel conditions, the bit-rate of voiced segments may be time-varying during the measured period. Consequently, it is necessary to capture the relationship between the quality regarding coding distortion and the average bit-rate of voiced segments, which is shown in Fig.4. The fitting curve is formulated as follows:

$$Q_c = a_1 \cdot \ln\left(Br\right) + a_2 \tag{1}$$

where Q_c is the voice quality regarding coding distortion, Br is the average bit-rate of voiced segments and a_1 and a_2 are model parameters obtained under the least square error criterion.

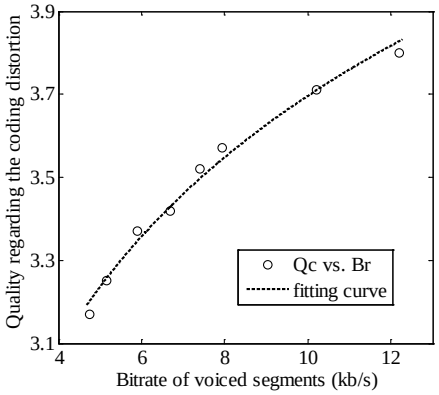


Fig.4 Relationship between the quality regarding coding distortion and the bit-rate of voiced segments.

Without resorting to any media-related payload information, packet-layer model usually evaluates the distortion due to packet loss in terms of packet loss rate [3].

Nevertheless, the packet loss frequency and the average burst length were used to obtain outstanding performance in [4], which is denoted as the PA model for convenience in this paper. The packet loss frequency and the average burst length are the average number of packet loss blocks per second and the average number of packets in each packet loss block in the voice streaming, respectively, where the packet loss block is defined as one group of consecutive lost packets.

The distortion due to packet loss is formulated following this spirit in the proposed system, as

$$l_v' = a_3 \cdot (l_v - 1) + 1 \tag{2}$$

$$DF = (1 - a_4) \cdot \exp\left(-\frac{l_v' \cdot f_v}{a_5}\right) + a_4 \cdot \exp\left(-\frac{l_v' \cdot f_v}{a_6}\right) \tag{3}$$

where DF is the quality degradation factor due to packet loss, f_v is the packet loss frequency of voiced packets, l_v is the average burst length of lost voiced packets, and l_v' is an intermediate variable. Besides, a_3, a_4, a_5 and a_6 are model parameters, which can be obtained by utilizing the training method provided in [4] with the chosen training sequences. Different from the PA model, the proposed system only considers the information of lost voiced packets when calculating the assessment parameters. Once the lost packets are determined as silence, their loss is neglected because the loss of silence packets leads to little quality impairment.

Through the joint integration of the perceived quality regarding coding distortion and the quality degradation caused by packet loss, the overall voice quality is calculated as follows:

$$Q_a = 1 + (Q_c - 1) \cdot DF \tag{4}$$

where Q_a is the overall voice quality, which is the output of the proposed system.

5 Experimental Results

In the experiments, the reference model was performed by the PESQ algorithm [9][10], which is well recognized for its good performance to estimate subjective scores. The pre-processed material from experiment three in ITU-T P-series supplement 23 database was used to provide voice sequences for training and test [11]. Under the criterion that the number of female voice sequences is equal to that of male voice sequences, we randomly select 80 sequences for training model parameters, and other 60 for model performance evaluation. The test sequences were encoded under eight coding bit-rates from 4.75 to 12.2kb/s by AMR codec and the encoded bit-streams were packetized. In the simulation, the 4-state markov model was introduced to model the packet loss distribution [12], and the used packet loss rates were set to 0%, 1%, 3%, 5%, and 10%, respectively. Finally, 2400 degraded sequences (60 sequences × 8 bit-rates × 5 packet loss rates) were generated to evaluate the performance of the proposed system.

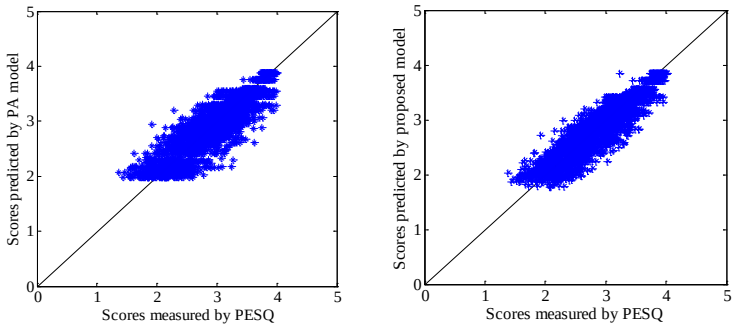
The parameters of the proposed model, i.e., a_1, a_2, a_3, a_4, a_5 and a_6, are 0.664, 2.168, 0.36, 0.43, 1.63 and 0.43, respectively. In the experiments, the estimation

model in the proposed system was compared to the PA model [4] and the E-model [2]. Two performance criteria were used to evaluate the performance of the compared models. That is, the Pearson correlation coefficient (PCC) for the linearity and the root mean squared error (RMSE) for the accuracy. Table 1 gives the specified performance comparison, from which it can be seen that the estimation model in the proposed system outperforms the PA model and the E-model. For example, the performance gain over the PA model is 0.0358 in PCC and the RMSE reduction reaches 0.0376.

Table 1 Performance comparison

Assessment Model	PCC	RMSE
PA model	0.8788	0.2649
E-model	0.8779	0.2651
Proposed model	0.9146	0.2273

In order to illustrate the performance comparison intuitively, two scatter plots are provided in Fig.5, where (a) and (b) show the scatter plots of the scores measured by the PESQ algorithm and that predicted by PA model and the proposed model, respectively. It can be directly observed that the points in (b) are closer scattered around the diagonal, i.e. the scores obtained by the proposed model are more consistent with that acquired by the PESQ algorithm. Obviously, the experimental results demonstrate the effectiveness of the proposed model.



(a)PESQ vs. PA model (b) PESQ vs. proposed model

Fig.5. Scatter plots of scores measured by PESQ vs scores predicted by PA model and the proposed model.

6 Conclusions

In this paper, a packet-layer quality assessment system has been proposed for monitoring the quality of networked voice. Through a simple analysis on received packet performed by the packet header analysis module, useful information is extracted directly from the packet headers. According to the prediction results acquired by the silence module, only the information about voiced segments is sent to the quality assessment module, where the voice quality regarding coding distor-

tion is first estimated utilizing the average bit-rate of voiced segments, and the overall voice quality is further evaluated by taking account of the impact of packet loss. The estimation model in the proposed system achieves superior performance over the PA model and E-model. Since only exploiting the packet headers, the proposed system is featured by its high efficiency and therefore well suited for real-time networked voice applications.

Acknowledgment This work was supported by the Fundamental Research Funds for the Central Universities (72115612), and the 111 Project (B08038).

References

1. Takahashi, Yoshino H. and Kitawaki N. (2004). Perceptual QoS assessment technologies for VoIP. *IEEE Communications Magazine*, *42*(7), 28-34.
2. ITU-T Recommendation G.107 (2002). The E-model, a computational model for use in transmission planning.
3. Clark A.D. (2001). Modeling the effects of burst packet loss and recency subjective voice quality. *Proceeding of 2nd IP Telephony Workshop*, New York, USA, 123-127.
4. Egi N., Hayashi T. and Takahashi A. (2010). Parametric packet-Layer model for evaluation audio quality in multimedia streaming services. *IEICE Trans. Communications*, *E93-B*(6), 1359-1366.
5. ITU-T SG12 Temporary Document TD 297 (2010). Updated draft terms of reference for P.NAMS. Geneva, Switzerland.
6. Yang F., Jiang L. and Li X. (2012). Real-time quality assessment for voice over IP. *Concurrency and Computation: Practice and Experience*, *24*(11), 1192-1199.
7. 3GPP TS 26.101 v6.0.0 (2004). Adaptive multi-rate (AMR) speech codec frame structure. Valbonne, France.
8. ITU-T Recommendation P.862 (2001). Perceptual evaluation of speech quality (PESQ), an objective method for end to end speech quality assessment of narrowband telephone networks and speech codecs.
9. Radhakrishnan K., Larijani H. and Buggy T. (2010). A non-intrusive method to assess voice quality over internet. *Proceeding of the 2010 International Symposium on Performance Evaluation of Computer and Telecommunication Systems*, Ottawa, Canada, 380-386.
10. Roychoudhuri L. and Al-Shaer E. S. (2005). Real-time audio quality evaluation for adaptive multimedia protocols. *Proceeding of the 8th IFIP/IEEE International Conference on Management of Multimedia Networks and Services*, Barcelona, Spain, 133-144.
11. ITU-T Recommendation P-series supplement 23 (1998). ITU-T coded-speech database.
12. ITU Rep. COM12-D97-E (2003). Packet loss distributions and packet loss models. Geneva, Switzerland.