In Proceedings of Fifth International Workshop on Quality of future Internet Services (QofIS'04), Barcelona, Spain, September 2004, (corrected Oct.15,2004).

Predicting the Perceptual Service Quality Using a Trace of VoIP Packets *

Christian Hoene, Sven Wiethölter, Adam Wolisz

Telecommunication Networks Group, Technical University Berlin
Einsteinufer 25, 10587 Berlin, Germany
Email: *hoene@ieee.org*

Abstract. We present an instrumental approach on how to assess the perceptual quality of voice transmissions in IP-based communication networks. Our approach is end-to-end and uses combinations of common codecs, loss concealment algorithms, playout schedulers, and ITU's quality assessment algorithms E-Model and PESQ. It is the first method that takes the impact of playout rescheduling and *non-random packet loss distributions* into account. Non-random packet losses occur if a rate-distortion optimized multimedia streaming algorithm forwards packets dependent on the packets' importance.
Our approach is implemented in open-source software. We have conducted formal listening-only tests to verify the accuracy of our quality model. In the majority of cases, the human test results show a high correlation with the calculated predictions.

Keywords: VoIP, quality assessment, playout scheduling, rate-distortion optimized streaming

1 Introduction

Instrumental perceptual assessment methods predict the behavior of humans rating the quality of multimedia streams. The ITU has standardized a psycho acoustic quality model called PESQ, which predicts the human rating of speech quality and calculates a mean opinion score (MOS) value [1]. Another model, the E-Model [2], evaluates the configurations of telephone systems. Among other factors it takes coding mode, packet loss rate, and absolute transmission delay into account to give an overall rating of the quality of telephone calls. Both models consider most sources of impairment which could occur in a telephone system. For example, they can predict the impact of the mean packet loss rate on speech quality. However, they do not consider packet losses if the loss depends on the packets' content or importance. Furthermore, they cannot be directly applied to traces of VoIP packets, which are produced by experimental measurements or network simulations.

* This work has been supported by Deutsche Forschungsgemeinschaft (DFG) via the AKOM program.

To overcome these deficiencies we have developed a systematic approach that combines ITU's E-model, ITU's PESQ algorithm, and various implementations of codecs and playout schedulers. Our software encodes a speech sample, analyzes a given trace of VoIP packets, simulates multiple playout schedulers, and finally assesses the quality of telephone services (coding distortion, packet loss, transmission delay and playout rescheduling). Thus, it can determine the final packet loss rate, speech quality, mean transmission delay and conversational call quality. In this regard, we have achieved the following contributions, which this paper subsumes and describes as they are presented in relation to each other.

- We developed a formula in [3] on how to include PESQ into the E-Model. The ITU approved this formula as a standard extension.
- We conducted formal listening-only tests to verify the prediction performance of PESQ for impairments due to non-random packet losses [4] and playout rescheduling, which are caused by rate-distortion optimized streaming and adaptive playout scheduling respectively. The overall correlations are R=0.94 and R=0.87 respectively.
- Finally we implemented the most common playout schedulers and provide them to the research community as open-source software [5]. Because perceptual speech quality assessment is computational complex, we provide a tool which runs the calculations in parallel.

Our approach outperforms previous algorithms because it does not only consider the impact of playout rescheduling but also takes transmission delay, speech quality and non-random packet loss distribution into account. Altogether, we are able to predict the quality of VoIP transmissions at a high precision that has not been reached before.

The paper is structured as follows: In section 2 we present the technical background and discuss related work. How to combine PESQ, E-Model and playout schedulers is explained in section 3. The next section contains the results of listening-only tests. Finally, in section 5, we draw conclusions.

2 Background

2.1 Internet Telephony

The principle components of a VoIP system, which covers the end-to-end transmission of voice, are displayed in Fig. 1. First, at the source the analogue processing, digitalization, encoding, packetization, and protocol processing (RTP, UDP, and IP) are conducted. Then, the resulting packets are transmitted through the network, consisting of an Internet backbone and access networks. At the receiver, protocols process the packets and deliver them to the playout scheduler/buffer. In the next step, the multimedia frames are decoded and played out. Because telephony consists of bidirectional transmission a similar transmission is presented in the reverse direction. In the following paragraphs we will discuss some components in more detail to show how they cause the service quality to degrade.

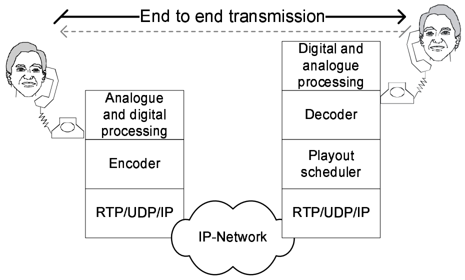

Fig. 1. VoIP transmission of a telephone call

Network: On the Internet and in the access networks packets can get lost because of congestion or (wireless) transmission errors. The packet loss process can be controlled to optimize the perceived service quality:

Chou et al. [6] suggest to forward multimedia packets according to their estimated distortion and error propagation. He proposed a rate-distortion optimized multimedia streaming framework for packetized and lossy networks.

De Martin [7] proposed an approach called Source-Driven Packet Marking, which controls the priority marking of speech packets in a DiffServ network. If packets are assumed to be perceptually critical, they are transmitted at a premium traffic class.

Sanneck used a modified Random-Early-Dropping (RED) at packet forwarding nodes [8]. If a node is congested, the probability of packet dropping should depend on the packet markings. Additionally, Sanneck proposes to mark G.729 coded voice packets according to their estimated importance.

All three algorithms handle packets in a content-sensitive manner. Therefore dropping of packets might depend on their marking and content. Thus, it is inadequately to measure only the mean packet loss rate for predicting the speech quality.

Playout scheduler: At the receiver, a playout buffer stores packets so that they can be played out in a time-regular manner, concealing variations in network delay (jitter). As the playout buffer contributes to the end-to-end delay it should not store packets longer than necessary. Instead, the playout buffer should drop packets that arrive too late to be played out at the scheduled time.

The playout scheduling can be static: If packets exceed a given transmission time they will be discarded (we will refer to this scheme as *fixed playout buffer*). Alternatively *adaptive playout buffers* re-define the playout time in accordance to the delay process of the network [9,10]. We refer to this kind of adaptation as *rescheduling*. The playout schedule can be adjusted easily during silence because then it is not notable. Adjustments during voice activity require more sophisticated concealment algorithms [11].

2.2 Quality Assessment

The perceived quality of a service can be measured with subjective tests. Humans evaluate the quality of service according to a standardized quality assessment process [12]. Often the quality is described by a *mean opinion score (MOS)* value, which scales from 1 (bad) to 5 (excellent). *Listening-only tests* are time consuming. Especially if many tests have to be made, the effort of subjective evaluation is prohibitive. Fortunately, in the last years, considerable effort has been made to develop instrumental measurement tools, which predict the human rating behavior. We will explain shortly the approaches used in this paper.

The *perceptual assessment of speech quality (PESQ)* algorithm predicts human rating behavior for narrow band speech transmission [1]. It compares an original speech fragment with its transmitted and thus degraded version to determine an estimated MOS value. For multiple known sources of impairment (typical for analogue, digital and packetized voice transmission systems) it shows a high correlation (about 0.94) with human ratings.

The quality of a telephone call cannot be judged by the speech quality alone. The ITU *E-Model* [2] additionally considers end-to-end delay, echoes, side-tones, loudness and other factors to calculate the so called *R-factor*. A higher R-factor corresponds to a better telephone quality, being 0 the worst value, 70 the minimal quality of telephone calls (“toll quality”), and 100 the best value.

2.3 Related Work

Markopoulou et al. [13] measured the performance of a couple of Internet backbone links and analyzed them with ITU’s E-Model. Their findings include not only that the quality of VoIP depends largely on the provider’s link quality but also on the playout buffer scheme.

Hammer et al. [14] suggest to use PESQ to assess the speech quality of a VoIP packet trace. He proposes to split the trace into overlapping subparts. The benefit of this approach is that different coding schemes and also packet marking algorithms can be judged. Also, FEC or different playout schedulers can be supported in principle.

An approach that also considers interactivity is presented by Sun and Ifeachor [15]. The authors suggest to combine the E-Model and PESQ and describe a set of equations, which they derived by linear approximations to the rating behavior of PESQ, E-Model and the correlation between packet loss rate and speech quality.

3 Combining E-Model, PESQ and Playout Schedulers

Considering the characteristics of VoIP packet transmissions on the one side and the capability of perceptual models on the other side, we identify the following aspects as incomplete:

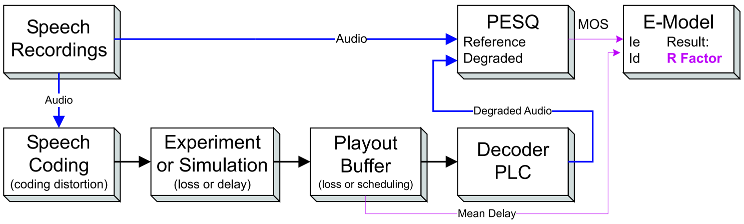

Fig. 2. Speech and delay assessment

– Perceptual quality assessment has to take into account the entire processing chain from source to sink, including encoding, routing across the Internet, de-jittering, decoding and playing at the receiving side, because only this reflects human-to-human conversation. Thus, when studying a transmission of VoIP packets the entire transmission system has to be considered.
– The end-to-end quality depends largely on the playout buffer scheme [13]. However, until now an "ideal" playout scheduler has not been identified and any implementer of a VoIP phone is free to choose any scheme. Thus, to predict the impact of playout scheduling one has to consider all, and if not possible, the most common playout schedulers.
– Rescheduling of adaptive playout schedulers harms the speech quality because of temporal discontinuities. The E-Model does not take the dynamics of a transmission into account because it relies on static transmission parameters. PESQ instead considers playout adaptation but does not include the absolute delay into its rating algorithm. PESQ has been designed to judge the impact of playout scheduling but has not been validated yet for this purpose [16].
– The E-Model does not consider non-random packet losses and PESQ has not been verified for this kind of distortion and the prediction accuracy is unknown.

To overcome these shortcomings we combine the E-Model, PESQ and playout schedulers as shown in Fig. 2: First, a set of the most common playout scheduler schemes (including fixed-deadline and adaptive algorithms [9,10] of Van Jacobsen, Mills, Schulzrinne, Ramjee, and Moon) calculates the packets' playout times and the mean transmission delay. One should note that only speech frames during voice activity are considered because during silence a human cannot identify the transmission delay.[1]

Next, PESQ calculates the speech quality that depends on coding distortion, non-random packet loss and playout rescheduling. Because PESQ has not been verified for non-random packet loss and playout rescheduling, we conduct formal

[1] Indeed, some playout schedulers change the playout time at the start of a talk spurt. Others change it at the beginning of silence periods. Both have to be considered as equal with respect to the transmission delay.

listening tests to verify its accuracy (see section 4). Last, both the speech quality and the mean transmission delay are fed into the E-Model. We assume the acoustic processing as optimal [3]. Therefore we can simplify the E-Model to a model with only few parameters. The computation of R_{factor} is then given by:

$$R_{factor} = \mathrm{MOS_2R}\left(\mathrm{MOS_{PESQ}}\right) - I_{dd}\left(t\right) \tag{1}$$

Reference [3] describes the function $\mathrm{MOS_2R}$ and the conditions under which (1) can be applied. For a definition of the function I_{dd} we like to refer to [2].

Software Package: We have implemented the approach discussed above and provide it as open-source to the research community. The software covers the digital processing chain of VoIP. To be fully operational, the PESQ algorithm and a G.729 codec have to be bought from its rights owners. Alternatively they can be downloaded at no costs from ITU's web page for trials only. Further information can be found at our web page and in the manual [5].

We try to verify the correctness of our software by several means. Publishing of this software together with its source code ensures that more users are going to use it and to study its code. Thus, the pace of finding potential errors will be increased. Last not least, we have tested our tool-set on various projects which include the assessment of voice over WLAN, the impact of handover and wireless link scheduling. Overall, we are confident of the correctness of our implementation.

4 Listening-Only Tests

PESQ has not been verified for all causes of impairment. In [4] we have conducted formal listening-only test to determine PESQ's prediction performance in cases of single or non-random packet losses. In this section we verify whether PESQ can measure the impairment of playout rescheduling. This verification is important because PESQ was not designed for this kind of impairments and operates outside the scope of its operational specification [16].

To verify PESQ, we construct artificially degraded samples and conduct both listening-only tests and instrumental predictions. If both PESQ and human tests yield similar results for the samples, PESQ is verified. Usually, the results of speech quality tests are compared via correlation. Thus, the amount of correlation (R) between subjective and objective speech quality results is our measure of similarity. R=1 means that the results are perfectly related. If no correlation is present, R is equal to zero. To compare absolute subjective and instrumental MOS values, we apply linear regression to one set of values which is a usual practice. The correlation R does not change after linear regression.

Sample Design: Analysis of Internet traces has shown that sometimes packet delays show a sharp, spike-like increase [9,13] which cannot be predicted in advance. Delay spikes are a short increase of the packet transmission times which usually occur after congestion or on a wireless link after fading. Soon after the spike

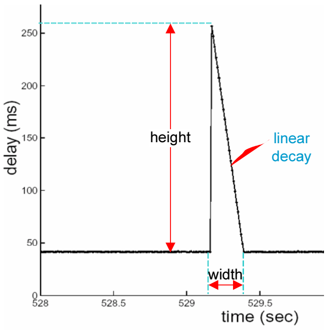

Fig. 3. Delay spike

the following packets arrive shortly one after the other until the transmission delay has returned to its normal value (Fig. 3). We like to consider the question whether to adjust the playout of speech frames to delay spikes by concentrating on the non-trivial case of delay spikes during voice activity.

For constructing the samples we have used the software package described in this paper. It generates artificial packet traces that contain delay spikes. One can control the frequency, the height and weight of delay spike. Further, three different playout strategies are analyzed: First, we drop every packet that is affected by the spike and thus arrives too late. Second, in case of a delay spike, the playout is re-scheduled so that no packet will be dropped. As a consequence, the playout delay will be increased. The last strategy is similar to the second, but after any spikes, the playout delay is adjusted during silence periods until the playout delay returns to normal.

We construct 220 samples (length approx. 5-10s), containing samples encoded with G.711, G.729 and containing one delay spike with a height of 50 to 300ms and a width of 55 to 330ms.

Formal Listening Tests: The listening-only tests followed closely the ITU recommendations [12], Appendix B, that describes methods for subjective assessment of quality. The tests took place in a professional sound studio ($46\,m^2$, low environmental noise, etc.). Nine persons judged the quality of 164 samples. The samples' language is German, which all listeners understand.

We do not follow the ITU's recommendations when scientific results suggest changes that improve the rating performance. For example, we have used high quality studio headphones instead of an Intermediate Reference System, because headphones have a better sound quality. Further, multiple persons were in the room at the same time to reduce the duration of the experiment.

Last but not least we do not apply the "Absolute Category Rating" because a discrete MOS makes it difficult to compare two only slightly different samples. The impact of a single frame loss is indeed very small. We allow intermediate val-

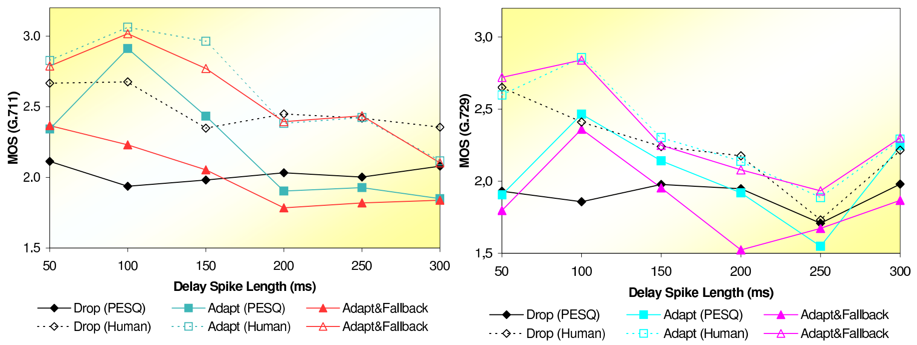

Fig. 4. Playout strategy: delay spike height vs. speech quality.

ues and use a linear MOS scale. PESQ calculates a MOS value with a resolution of up to 10^{-6} at the MOS scale.

Results: Ten persons gave a total of 2210 judgments. We could use only 2033 judgments because some test persons failed to get track with the sample number. The rating performance during the second half of the test was significantly worse than during the first half. We also compared a group of native speakers and a group of foreign students. Both have shown a similar rating performance. This leads us to the conclusion that being concentrated is more important than being a native speaker.

Figure 4 displays the speech quality versus the spike height. We show the rating results of humans and PESQ for different adaptation policies. The black lines (drop) refer to the dropping of any late packet during the delay spike. The blue lines (adapt) display the results when delaying the playout after a delay spike. Last, the red lines (adapt&fallback) include the effect of falling back to the original playout time as soon as possible. The later rescheduling occurs only during periods of silenced speech.

Analysis: In the experiments the sample content is varied for each different delay spike height. Because the sample content has a large influence on the speech quality ratings, one cannot compare absolute MOS values on the horizontal axis (that displays the delay spike height) in Figure 4. However, the playout strategies can be ranked against each other if the delay spike height remains the same.

If the delay-spike's height is 200 ms or larger, dropping packets is more beneficial than delaying the playout. Further, the "fallback" adjustments during silence degrade the speech quality. The "adapt" algorithm performs always better.

Table 1 displays the prediction performance of PESQ. The overall correlation is $R = 0.866$. If one considers only samples that contain modulated noise (MNRU), the correlation is nearly perfect ($R = 0.978$). Also, we identify a coherency between the MOS variance and the correlation. For a given sample set,

Table 1. Listening-only test: delay spikes

Selection		MNRU	Coding		Spike Height			Playout strategy		
criteria:	all	yes	G.711	G.729	100ms	200ms	300ms	drop	adapt	&fallback
Samples	113	13	33	63	15	18	18	32	32	32
MOS	2.518	3.013	2.565	2.284	2.844	2.228	2.280	2.220	2.453	2.469
PESQ MOS	2.280	2.823	2.277	2.028	2.680	1.873	1.998	2.039	2.243	2.058
PESQ var.	0.723	1.015	0.564	0.373	0.882	0.141	0.246	0.088	0.717	0.541
Correlation	0.866	0.978	0.856	0.668	0.906	0.737	0.768	0.476	0.838	0.799

the more the samples differ the higher is the correlation. Considering the "drop" strategy for example, both the PESQ variance and the correlation are low. We assume that humans cannot distinguish degraded samples which are only slightly different.

Comparing the absolute MOS values in Table 1, one can see that there exists a constant offset between instrumental and subjective MOS values. As we can not understand the reasons for this offset, we assume that it can be explained due to the social behavior and emotions of our listening personal. Their ratings are severer or more indulgent as compared to the ratings used during the development of PESQ.

5 Conclusions

In this paper we have presented an approach on how to assess the quality of VoIP transmissions. We identify important sources of quality degradation that can occur in a VoIP system; especially the impact of playout rescheduling and non-random packet losses has not been considered in previous approaches.

We combine PESQ, the E-model, different coding schemes and playout schedulers to analyze VoIP packet traces. The ITU approved the mathematical combination of PESQ and E-Model as a standard extension.

PESQ is verified with formal listening-only tests to identify its prediction accuracy[2]. The listening-only tests lead to manifold results. They show that PESQ indeed predicts in general the speech quality well. However, we identified that in the same cases, PESQ has to be improved. These improvements are beyond the scope of this paper. To enable other researchers the verification as well as the tuning of their algorithms, the complete experimental data including all samples and ratings are available on request.

Beyond the scope of this paper are also various performance evaluations that our approach enables. For example, the assessment of playout schedulers can be used to identify the ideal one. Also, Internet backbone traces can be assessed and novel VoIP over WLAN systems can be developed. Especially, if the importance of speech frames is utilized [4] and non-random packet losses are enforced, we will show that the performance of VoIP over wireless can be enhanced significantly.

[2] We like to thank our students L. Abdelkarim and T. Dulamsuren-Lalla

References

1. 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 (2001)
2. ITU-T Recommendation G.107: The E-model, a Computational Model for Use in Transmission Planning (2000)
3. Hoene, C., Karl, H., Wolisz, A.: A perceptual quality model for adaptive VoIP applications. In: SPECTS, San Jose, California, USA (2004)
4. Hoene, C., Dulamsuren-Lalla, E.: Predicting performance of PESQ in case of single frame losses. In: Measurement of Speech and Audio Quality in Networks Workshop (MESAQIN), Prague, CZ (2004)
5. Hoene, C.: Simulating playout schedulers for VoIP - software package. URL:http://www.tkn.tu-berlin.de/research/qofis/ (2004)
6. Chou, P., Mohr, A., Wang, A., Mehrotra, S.: Error control for receiver-driven layered multicast of audio and video. IEEE Transactions on Multimedia **3** (2001) 108–122
7. Martin, J.D.: Source-driven packet marking for speech transmission over differentiated-services networks. In: IEEE ICASSP. Volume 2. (2001) 753–756
8. Sanneck, H., Tuong, N., Le, L., Wolisz, A., Carle, G.: Intra-flow loss recovery and control for VoIP. In: ACM Multimedia. (2001) 441–454
9. Ramjee, R., Kurose, J.F., Towsley, D.F., Schulzrinne, H.: Adaptive playout mechanisms for packetized audio applications in wide-area networks. In: IEEE Infocom, Toronto, Canada (1994) 680–688
10. Moon, S.B., Kurose, J., Towsley, D.: Packet audio playout delay adjustments: performance bounds and algorithms. ACM/Springer Multimedia Systems **27** (1998) 17–28
11. Liang, Y.J., Färber, N., Girod, B.: Adaptive playout scheduling and loss concealment for voice communication over IP networks. IEEE Transactions on Multimedia **5** (2003) 532–543
12. ITU-T Recommendation P.800: Methods for subjective determination of transmission quality (1996)
13. Markopoulou, A., Tobagi, F., Karam, M.: Assessing the quality of voice communications over internet backbones. IEEE/ACM Transactions on Networking **11** (2003) 747– 760
14. Hammer, F., Reichl, P., Ziegler, T.: Where packet traces meet speech samples: An instrumental approach to perceptual QoS evaluation of VoIP. In: IEEE IWQoS, Montreal, Canada (2004)
15. Sun, L., Ifeachor, E.: New models for perceived voice quality prediction and their applications in playout buffer optimization for VoIP networks. In: IEEE ICC, Paris, France (2004)
16. Rix, A.W., Hollier, M.P., Hekstra, A.P., Beerends, J.G.: Perceptual evaluation of speech quality (PESQ), the new ITU standard for end-to-end speech quality assessment, part I - time alignment. Journal of the Audio Engineering Society **50** (2002) 755