In Proc. MESAQIN 2004, Prague, CZ, June 2004 MESAQIN 2004 - Measurement of Speech and Audio Quality in Networks

Predicting Performance of PESQ in Case of Single Frame Losses

Christian Hoene, Enhtuya Dulamsuren-Lalla
Technical University of Berlin, Germany
Fax: +49 30 31423819
Email: hoene@ieee.org

Abstract

ITU's objective evaluation algorithm PESQ predicts the quality of speech transmissions. In this work we verify whether PESQ can measure the impact of single frame losses – a source of impairment for which PESQ has not been designed. To construct samples for experimental tests, we develop a tool that controls the loss of specific frames, e.g. only important or voiced frames. We conduct subjective, formal listening-only tests to verify PESQ's prediction performance. The human ratings correlate with PESQ at a degree of R=0.94. Given the precision of speech quality measurements we show the equality of subjective and instrumental results.

Keywords

PESQ, single frame loss, formal listening-tests

1 Introduction

To assess the speech quality of telephone or communication systems the ITU has defined the quality model Perceptual Evaluation of Speech Quality (PESQ) [9]. It compares an original speech sample with the corresponding transmitted and degraded version to calculate a Mean Opinion Source (MOS). The MOS value scales from 1 (bad) to 5 (excellent) and describes the level of speech quality.

PESQ is only a psychoacoustic model of the human hearing. Thus, it only simulates the human rating behaviour and it is – as a matter of principle – less precise than humans. On the other side, when humans rate the speech quality in listening-only tests, the results are precise only if the tests are carefully conducted. The ITU has set up a detailed description [1] on how to conduct listening-only tests in such a manner that they achieve a highest degree of accuracy. These tests are referred as formal tests. This paper describes the results for formal listening tests which verify the prediction performance of PESQ in the presence of a special kind of distortion, namely single frame losses.

PESQ has been designed to take the impairment due to multiple frame losses into account. Frame (or packet) losses occur if networks are congested or (wireless) links have transmission errors. PESQ measures the impact of frame losses well. It shows a high correlation with the results of formal tests (R=0.93) [10]. But one should note that this statement is only true if randomly distributed frame losses occur. It does not hold if single, specific frame losses are to be measured.

In our previous work [11] we have shown that objective quality models (such as EMBSD [14] and PESQ) rate single frame losses largely differently. In this work we verify whether PESQ measures the importance of single frame losses similar as humans do. This verification is important because PESQ has not been designed for this kind of measurement and operates outside the scope of its operational specification. Knowing the importance of multimedia packets is required if rate-distortion optimized multimedia transmission algorithms shell enhance the efficiency of the communication systems [15].

The difficulty of the listening tests is the fact that humans often can not hear the impairment of one frame loss. Humans can judge only the impact of multiple frame losses. Thus, if we want to verify PESQ's rating of single frames, we have to construct samples containing multiple losses of the same frame. However, it is not possible to generate samples which contain multiple losses of the same frame, because at least the frame's context will be different. Thus, we drop multiple, similar frames. If both PESQ and human tests yield same results for multiple but similar frame losses, PESQ is verified single losses[1].

[1] As long as frame losses do not occur shortly one after the other, we can assume that PESQ results scale linear with the number of lost frames [11].

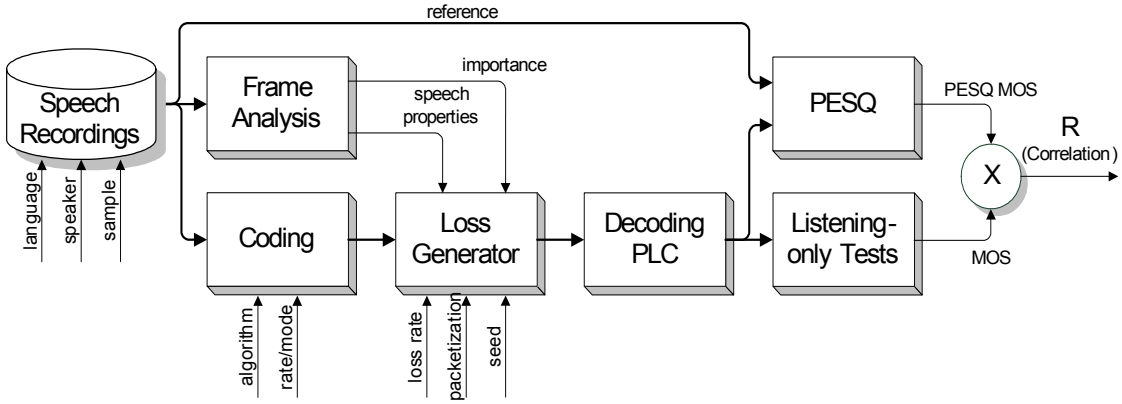

Figure 1: Test design

To identify similar frames, a packet classification is required. Thus, to verify PESQ's ability to classify frame losses, we need a proper classification of frames. This circular problem definition makes verifications difficult. Colloquial speaking it is a classic chicken-and-egg problem. Anyhow, we have decided to classify frames according to their importance, as measured with PESQ, and to their different speech properties, (silence, active, voiced and unvoiced sounds). We also vary the coding scheme.

To generate the samples for the human based test, we have implemented the tool Mongolia, which generates samples with specific frame losses. As a gadget we also have set up a public web service interface [13]. We have conducted formal listening tests judging 164 different samples by 9 persons. Our listing tests show a correlation of 0.94 with the predictions of PESQ. We can conclude that we can use PESQ to predict the impact of single packet losses.

This paper is structured as follows. First, we discuss related work. Then, we describe our tool Mongolia. Last, we present the results of the listening-only tests which are finally concluded.

2 Related Work

Speech frames differ greatly. A classic application of the temporal characteristics of speech is the suppression of the packets' transmission during silence. Discontinuous Transmission (DTX) interrupts the constant flow of frames until new audio content has to be transmitted again. DTX drops only frames which are not important for speech quality. DTX has been verified by listening-only tests.

De Martin [3] has proposed a packet classification scheme, which marks 20 percent of all speech frames as important. The others are marked as normal. The author describes a packet-marking algorithm for the ITU G.729 coding. For each frame it computes the expected perceptual distortion, as if the speech frame were lost. De Martin has conducted formal listening tests which have shown that the source-driven packet marking algorithm, if applied on a Diff-Serv network, enhances speech quality from MOS 3.4 to MOS 3.7, if 5% of all frames are lost.

Sanneck [2] analyzed the temporal sensitivity of VoIP flows if they are encoded with µ-law PCM and G.729: Losses in PCM flows have some but weak sensitivity to the current speech properties. The concealment performance of G.729, on the other hand, depends largely on the change of speech properties. If a frame is lost shortly after unvoiced/voiced transition, the loss is over-proportional notable. Furthermore voiced packets are more important than unvoiced packets. Sanneck used objective speech quality evaluation algorithms (MNB and EMBSD) to assess the packet classification.

In our previous work [11] we determined importance of single speech frames. We applied PESQ to measure the impact of losing single speech packets. We benchmarked the packet classification DTX, De Martin's, and Sanneck's algorithms.

3 Experimental Design

To verify PESQ, we construct artificially degraded samples and conduct both subjective and objective listening-only tests. Figure 1 displays the testing procedure.

3.1 Sample Design

The tool Mongolia (Figure 2) helps to generate degraded samples. The tool can be tested remotely on our web page [13]. It works as follows: First, a reference sample is selected from ITU's database P.suppl 23 [5]. Each sample has a length of 8s. Background noise is not present. If requested, samples (and their degraded) versions can be played loudly. Next, a coding algorithm compresses the

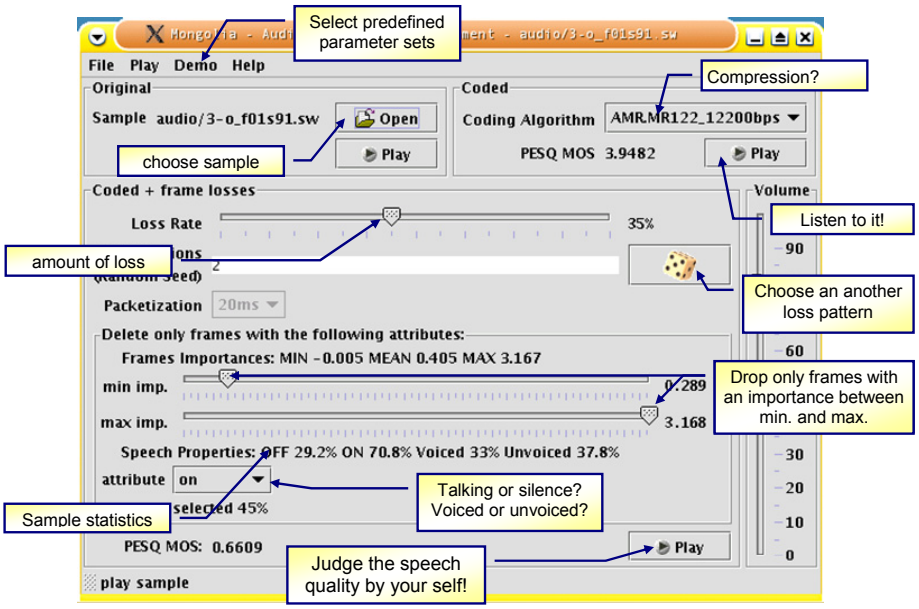

Figure 2: Design tool Mongolia: http://www.tkn.tu-berlin.de/research/mongolia

reference sample and the PESQ calculates the degraded sample's MOS value. The tool supports the three coding modes:

- G.711 [6] μ-law encoded narrow-band speech with a rate of 64kbit/s. We use the packet loss concealment (PLC) algorithm G.711 Appendix I [7], which works on frame sizes of 10ms.
- ITU G.729 [4] uses a Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP) algorithm to compress speech to frames of 10ms and at a rate of 8 kbit/s.
- The Adaptive Multi-Rate (AMR) [8] speech codec applies an Algebraic Code Excited Linear Prediction coding (ACELP) to support eight coding rates, ranging from 4.75 to 12.2 kbit/s, and generates a frame each 20ms. We support coding rates of 4.75 and 12.2 kbit/s.

Next, the overall frame loss rate controls, how many frames are dropped. The packet length controls the burstiness of frame losses. The later effect refers to packetised transmission of speech because a VoIP packet can contain multiple voice frames. A random seed value controls the positions of the losses. The user can select whether important or less important frame are dropped. The *importance* of a frame is the quality degradation that the frame's loss would cause. In [11] we described in detail how the importance of a packet is calculated. High values refer to more important frames. Next, frames are selected according to their speech property: a) frames containing during silence or b) active voice or active frame containing c) unvoiced and d) voiced sounds. Last, the packet loss statistics and the PESQ value are displayed.

For our listening-only tests we construct samples from four English language speakers (male, female). We drop 3% of all packets but only during voice activity. In this paper we do not analyse the trivial case of dropping silenced frames. We select all four coding modes and choose the shortest packet length (10ms or 20ms). We force the loss of either all, voiced or unvoiced segments. We also drop frames from either all, the most or the least important half of the packets. Altogether this test design consists of 4·4·3·3=144 samples. As a reference we also generate 20 samples containing modulated noise reference units (MNRU) as described in [16].

3.2 Formal Listening Only Tests

The listening-only tests followed closely the ITU recommendations [1], Appendix B that describes methods or subjective assessment of quality. The tests took place a professional sound studio (46 m^2, low environmental noise, etc.). Nine persons judged the quality of 164 samples. The samples' language is English, which all listeners understand.

We do not follow the ITU's recommendations if scientific results suggest changes that improve the rating performance. For example, we use high quality studio headphones instead of an Intermediate Reference System, because headphones have a better sound quality. Also, multiple persons are 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 values and use a linear MOS-LQS[2] scale. PESQ calculates a MOS-

[2] LQS refers to listening-only subjective tests; LQO are objective tests to determine the speech quality.

LQO value with a resolution of up to 10^{-6} at the MOS scale, too.

Finally, we analyse the results. We calculate the correlation of subjective and objective listening-only results to get a measure for similarity (R). R=1 means that the results are perfectly related. If no correlation is present, R equals zero. If we compare absolute subjective and objective MOS values, we apply a linear regression to one set of values. The correlation R does not change after linear regression.

4 Results

First, we present the MNRU listening-only results. In Figure 3 we present MOS values from PESQ, our listening-only tests and from tests described in [12]. We also included MOS-LQS values after linear regression, which fit closely the PESQ MOS-LQO values (Figure 3). Subjective and objective results have a correlation of 0.999.

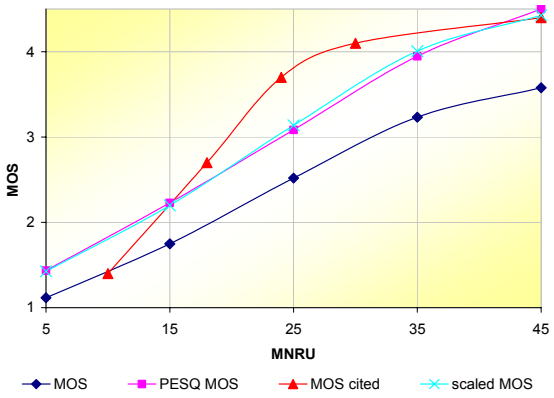

Figure 3: Reference tests: MNRU vs. MOS

Next, we show the MOS values excluding the MNRU results. We calculate the mean values of all listeners MOS values and all different reference samples (totally 4*9=36 trials). Table 3 contains the MOS values. In Figure 4 we display PESQ MOS-LQO vs. MOS-LQS to get an impression of the measurement performances.

Table 1 contains the correlation between MOS-LQS and MOS-LQO values. We analyse the prediction performance for difference kinds of impairment. In general the correlation depends on the variation of the sample (see Figure 5). If the samples are largely different (e.g. silenced noise and loud additional noise) both humans and PESQ rate the speech quality similar.

For example, PESQ predicts rather bad the impact of packet losses considering only samples, which are equally encoded (especially AMR 4.75, G.711, and G.729). On the other side, those samples differ only slightly and their variance is low. Thus, this effect might be explained by "measurement noise" being present in subjective tests.

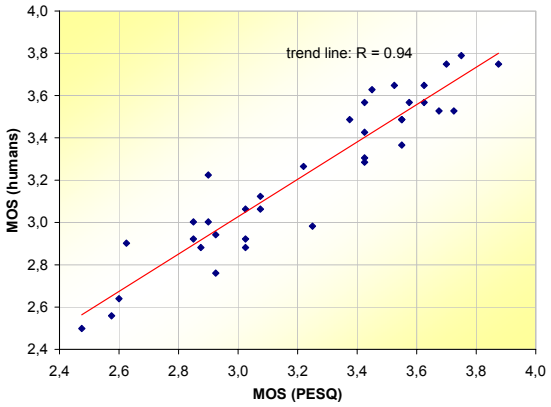

Figure 4: Comparison of MOS and PESQ MOS

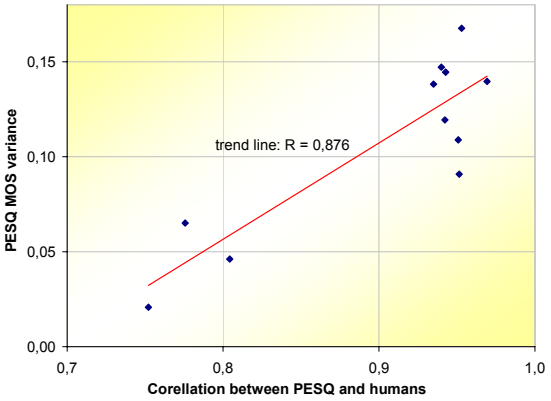

Figure 5: Sample variance vs. prediction performance

5 Summary

Speech frames differ great in their importance. If important frames are lost, the transmission quality of speech is significant degraded. On the other side, some frames – even during voice activity – are hardly worth transmitting.

In our previous publication we have developed a method which can measure the importance of frames or packets. This method is based on the objective quality assessment tool PESQ. The aim of this paper is to verify the accuracy our PESQ to measure the impact of single frame losses.

We have developed the tool Mongolia, which demonstrates how strong the importance of frames differs. It can be accessed and trailed via a public web interface. We used our tool to construct test samples, which helps to verify PESQ. We have conducted

formal listening-only tests, which show a correlation of 0.94 with results of PESQ. These tests prove that ITU's PESQ algorithm predicts the impact of single frames losses precisely.

If different sources of impairment (e.g. frame loss, coding distortion or noise) are to be compared, PESQ does not allow precise trade-off decisions to be made because absolute MOS values differ. In addition, informal listening-tests show that PESQ might not judge the effect of clipping – shortly before an ON-OFF transition – precisely. Further studies are required to identify problematic packet loss patterns.

6 Acknowledgement

We like to thank Prof. Noll and Prof. Wolisz for their valuable comments, our colleagues and friends for rating our samples and Prof. Hobohm and Folkmar Hein for providing the studio.

7 References

[1] ITU-T Recommendation P.800: *Methods for subjective determination of transmission quality*, Aug. 1996.

[2] H. Sanneck, L. Le, and A. Wolisz, "Intra-flow Loss Recovery and Control for VoIP", *Proc. Of ACM MULTIMEDIA*, pp. 441-451, Ottawa, Canada, Sep. 2001.

[3] J.C. De Martin, "Source-Driven Packet Marking for Speech Transmission Over Differentiated-Services Networks", *Proc. Of IEEE ICASSP 2001*, Salt Lake City, USA, May 2001.

[4] ITU-T, Recommendation G.729: *Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear-prediction* (CS-ACELP), Mar. 1996.

[5] ITU-T. Recommendation P.Suppl 23: *ITU-T coded-speech database*, Feb. 1998.

[6] ITU-T Recommendation G.711: *Pulse code modulation (PCM) of voice frequencies*, Nov. 1988.

[7] ITU-T Recommendation G.711 Appendix I: *A high quality low-complexity algorithm for packet loss concealment with G.711*, Sep. 1999.

[8] 3GPP TS 26.090: *Mandatory Speech Codec speech processing functions AMR speech codec; Transcoding functions*. Jun. 1999.

[9] ITU-T Recommendation P.862: *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, Feb. 2001.

[10] S. Pennock, "Accuracy of the Perceptual Evaluation of Speech Quality (PESQ) algorithm", Proc. Of MESAQIN, 2002.

[11] C. Hoene, B. Rathke, and A. Wolisz, "On the Importance of a VoIP Packet", In *Proc. Of ISCA Tutorial and Research Workshop on th Auditory Quality of Systems*, Herne, Germany, Apr. 2003.

[12] Y. J. Liang, N. Färber, and B. Girod, "Adaptive playout scheduling and loss concealment for voice communication over IP networks," *IEEE Transactions on Multimedia*, Dec. 2003.

[13] C. Hoene, "Software Tool Mongolia", URL http://www.tkn.tu-berlin.de/research/mongolia, April 2004.

[14] W. Yang, "Enhanced Modified Bark Spectral Distortion (EMBSD): An Objective Speech Quality Measure Based on Audible Distortion and Cognition Model", Dissertation, Temple University, Philadelphia, USA, May 1999.

[15] P. A. Chou, Z. Miao, "Rate-distortion optimized streaming of packetized media," Microsoft Research Technical Report MSR-TR-2001-35, February 2001.

[16] ITU-T Recommendation P.810: *Modulated noise reference unit (MNRU)*, Feb. 1996.

Table 1: Accuracy of PESQ

Condition	Correlation (R)	Number of trials	Mean MOS	Mean norm. MOS	Mean PESQ MOS	PESQ MOS variance
All but MNRU	0,940	1296	3,189	3,235	3,235	0,147
MNRU	0,999	180	2,439	2,738	3,039	na
AMR 12.2	0,951	324	3,218	3,254	3,292	0,109
AMR 4.75	0,804	324	2,545	2,808	2,778	0,046
G.711	0,752	324	3,828	3,657	3,617	0,021
G.729	0,776	324	3,167	3,220	3,252	0,065
Both voiced and unvoiced	0,969	432	3,210	3,248	3,243	0,140
Voiced	0,943	432	2,984	3,098	3,144	0,145
Unvoiced	0,953	432	3,375	3,357	3,317	0,168
Importance All	0,942	432	3,230	3,261	3,239	0,119
Importance Upper half	0,935	432	2,928	3,061	2,998	0,138
Importance Lower half	0,951	432	3,410	3,381	3,467	0,091

Table 2: MOS Results for modulated noise (MNRU)

MNRU	MOS	Norm. MOS	PESQ MOS	MNRU	MOS [12]
5	1,12	1,43	1,44	10	1,4
15	1,75	2,20	2,23	18	2,7
25	2,52	3,14	3,08	24	3,7
35	3,23	4,01	3,95	30	4,1
45	3,58	4,43	4,50	none	4,4

Table 3: Listening-only test results

Imp.	Speech Property	Codec	MOS	MOS scaled	PESQ MOS	PESQ MOS – MOS sca
Min 50%	Voiced	AMR 12.2	3,387	3,366	3,550	**0,2**
All			3,022	3,124	3,075	0,0
Max 50%			2,656	2,882	2,875	0,0
Min 50%		AMR 4.75	2,473	2,761	2,925	**0,2**
All			2,169	2,559	2,575	0,0
Max 50%			2,077	2,498	2,475	0,0
Min 50%		G.711	3,814	3,648	3,525	-0,1
All			3,784	3,628	3,450	**-0,2**
Max 50%			3,692	3,567	3,575	0,0
Min 50%		G.729	3,266	3,285	3,425	0,1
All			2,809	2,982	3,250	**0,3**
Max 50%			2,656	2,882	3,025	0,1
Min 50%	Unvoiced	AMR 12.2	3,631	3,527	3,725	0,2
All			3,570	3,487	3,375	-0,1
Max 50%			2,930	3,063	3,025	0,0
Min 50%		AMR 4.75	2,930	3,063	3,075	0,0
All			2,839	3,003	2,850	**-0,2**
Max 50%			2,687	2,902	2,625	**-0,3**
Min 50%		G.711	3,966	3,749	3,875	0,1
All			4,027	3,789	3,750	0,0
Max 50%			3,692	3,567	3,625	0,1
Min 50%		G.729	3,570	3,487	3,550	0,1
All			3,479	3,426	3,425	0,0
Max 50%			3,174	3,224	2,900	**-0,3**
Min 50%	voice active (both unvoiced and voiced)	AMR 12.2	3,631	3,527	3,675	0,1
All			3,296	3,305	3,425	0,1
Max 50%			2,839	3,003	2,900	-0,1
Min 50%		AMR 4.75	2,717	2,922	3,025	0,1
All			2,717	2,922	2,850	-0,1
Max 50%			2,291	2,639	2,600	0,0
Min 50%		G.711	3,966	3,749	3,700	0,0
All			3,814	3,648	3,625	0,0
Max 50%			3,692	3,567	3,425	-0,1
Min 50%		G.729	3,570	3,487	3,550	0,1
All			3,235	3,265	3,220	0,0
Max 50%			2,748	2,942	2,925	0,0