On the Importance of a VoIP Packet

Christian Hoene, Berthold Rathke, Adam Wolisz

Technical University of Berlin
hoene@ee.tu-berlin.de

Abstract

If highly compressed multimedia streams are transported over packet networks, losses of individual packets can impair the perceptual quality of the received stream in different degrees, depending on the content and context of the lost packet. In this paper, we investigate the impact of individual packet loss on the perceptual speech quality in Voice-over-IP systems using three popular coding types and receiver-side loss concealment algorithms. We set up a testing environment to measure the impairment of individual packet losses and define an appropriate quality metric. We evaluate published algorithms on packet loss quality prediction (DTX, Source-Driven Packet Marking and SPB-DiffMark) and identify their strengths and weaknesses. The quality of a VoIP telephone call can be enhanced significant, if a precise packet-loss quality model decides for each VoIP packet, how it should be forwarded throughout the network.

1. Introduction

The Internet develops to the ubiquitous communication network, which supports all kind of multimedia communications including telephony. Telephony still accounts for major usage times and large parts of revenue, being the most important form of telecommunication between humans. Most telephone calls are conducted on PSTN systems. Internet Telephony [1] – on the other side – still struggles to fulfill quality requirements and expectations. Therefore, enhancing the quality of telephone calls over packet networks, especially the Internet, is a worthwhile goal. Quality of Service (QoS) architectures have been introduced to guarantee quality levels. They usually assign a high priority to telephone calls and a lower priority to data transmission. Emerging QoS architectures like DiffServ [2] can treat, forward and drop packets according to their pre-defined priority. Novell approaches request different priorities for individual packets within a single flow [3],[4],[5]. This leads to the problem to classify the importance of each packet correctly, so that the overall service quality can be optimized.

The human perceived quality of a telephone call should the main optimization criterion, because most calls are between humans. The quality of telephone calls compromises many aspects: One important factor is the quality of speech transmission. The perceived speech quality is often measured in the metric mean opinion source (MOS). Metrics like MOS, which are based on human based quality judging, are difficult to apply to plan and control a communication networks because they require that humans perform the quality evaluation. Networking based quality metrics like packet loss rate, throughput and delay are easier to measure, to control and to guarantee. But they do not reflect the experienced user quality precisely.

Let us give an example: Packet losses decrease the quality significantly and are one of the major sources of impairment in a Voice over IP system. Packet losses occur if networking nodes are congested, if (wireless) links have transmission errors, or if packets have to be dropped at the receiver because they arrived too late to be played out on time [5]. The relation between mean packet loss rate and the MOS value is well studied (see section 2.2). It depends on the particular speech coding and decoding algorithm, the concealment and the rate and burstiness of frame losses. The relation between packet loss rate and quality is only valid, if – as often assumed – the Internet drops each packet with the same likelihood. This assumption does not hold for the emerging QoS solutions because the packet loss probability might depend on the packet's priority.

A single packet loss influences the quality in a wide range (see figure 4) because the characteristics of speech vary over time. The importance of a packet depends on its content, surrounding speech context, the performance of the decoder's concealment algorithm, and whether previous packets have been lost. Especially if highly compressing codecs are used, a packet loss might desynchronize the internal state of the decoder. In that case, the impairment is not only limited to the lost segment but will last for a notable period until the decoders state is synchronized again [3].

Therefore, just relating mean packet loss rates to quality is not sufficient. Instead, a more precise *quality model* has to be developed, which takes into account the content of packets to calculate the influence of packet losses.

In principle, a packet loss quality model can be developed for two different usage scenarios. In the first scenario, a telephone call or the transmission of a sample is evaluated after the entire transmission is completed. The complete transmitted speech samples and all packet losses are known. Such quality model just has to predict the human quality rating. The second usage scenario is more demanding: The quality model predicts the importance of a packet during the transmission (e.g. during the encoding of the packet). Consequently this quality model does not know, how the speech will progress nor whether previous packets have been lost. Thus, is does not only predict the human rating behavior, but also the progression of human speech and the loss process of the network.

In this paper, we focus on the second class of quality models. The contributions of this paper are the following: We develop a testing methodology that allows to measure and to study the impact of individual packet losses. We define a quality metric, which can be applied to measure the importance of a packet but also to control the packet loss process actively. Using our testbed and our metric, we evaluated three published packet-loss quality models. Our results show, that none of those

algorithms show a suitable prediction performance. Therefore, we intend to develop a quantitative quality model for the importance of VoIP packets, which will be applied to enhance the transmission performance of VoIP.

2. Related Work

2.1. Discontinuous Transmission (DTX)

One of classic application of the temporal characteristics of speech is the suppression of packet transmissions during silence. Periods of active speech alter with periods of background noise (virtual silence). Periods of silence are less important for the perceptual quality of speech transmission. Therefore, DTX interrupts the constant flow of frames until new audio content has to be transmitted again. Normally, DTX is part of the speech encoder.

The main problem of DTX is the clipping of speech. The DTX's silence detection algorithm might misunderstand voice as background noise and vice versa. The clipping of speech can reduce the speech quality significantly. Furthermore, human prefer some background noise against total, digital silence. Therefore, decoders (like ITU's G.729 Appendix B [7]) implement a comfort noise generator, which generates background noise. The speech quality can be even enhanced, if the decoder's background noise is similar to that on the encoding side. Therefore, modern codecs transmit background noise descriptions at regular interval during silence.

Packet based networks like the Internet benefit from DTX. If they transmit multiple voice streams in parallel, the different frame rates are distributed equally among the flows. Thus, the total bandwidth of all streams is less than the sum of all maximal coding rates, causing a statistical multiplexing gain, which saves bandwidth. In cellular mobile network, DTX is used to save transmission energy. If the transmission is interrupted the energy consumption during this period is reduced. Thus, the running time of a battery powered mobile phone is prolonged. Furthermore, the average interference on the air is reduced, thus allowing higher cell capacity, which – in some sense – is a statistical multiplexing gain, too.

2.2. Mean Packet Loss Rate

The relation between mean packet loss rate and speech quality has been extensively studied in [8][9]. It depends on the particular speech coding and decoding algorithm, the concealment and the burstiness of frame losses. As stated before, the main problem is the assumption that the Internet drops each packet with the same likelihood. This assumption is correct for the classical Internet, but does not apply to emerging Quality of Service architectures, which can treat, forward and drop packets according their pre-defined priority.

2.3. Source-Driven Packet Marking

De Martin [4] has proposed an approach called Source-Driven Packet Marking, which controls the priority marking of speech packets in a DiffServ [2] network. If packets are assumed to be perceptually critical, they are transmitted at a premium traffic class. All other packets are sent using the normal best-effort traffic class.

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. (It assumes, that no previous speech frames got lost.) First, only speech frames with a minimal level of energy are considered to be mark as premium. Next, the marking algorithm takes the coding parameter (e.g. gains, linear prediction filter, codebook indices) and computes the parameter, which the concealment algorithm would produce if the packet will be lost. It compared both parameter sets to compute the perceptual quality degradation in case of loss.

If any of the following perceptual distance parameters exceed a given threshold, the packet is mark as premium. Depending the voice/unvoiced state of the previous frame (as measured at the decoder), these thresholds are used for voiced frames:

- Adaptive-codebook index difference > 20%
- Adaptive-codebook gain difference > 5 dB
- Spectral distortion > 4dB

Instead, if the decoder expects an unvoiced frame, only the fixed-codebook gains and the spectral distortion are used:

- Fixed-codebook gain difference > 5dB
- Spectral distortion > 4 dB.

De Martin has conducted formal listening tests, which showed that the source-driven packet marking enhances speech quality from MOS 3.4 to MOS 3.7 during at a loss rate of 5%. During conditions of no loss, G.729 has a speech quality of MOS 4.0. It is sufficient, if 20% of all packets are marked as premium.

2.4. SPB-DIFFMARK

Sanneck [3] 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. Multiple consecutive losses have a higher impact on the quality degradation than to single, isolated losses.

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 internal state of the decoder might be de-synchronized for up to next 20 following frames [10]. Furthermore, voiced packets are more important than unvoiced packets. As a consequence, Sanneck proposed to mark packets with +1 (foreground), 0 (best-effort), and –1 (background traffic) depending of their speech properties. The so called Speech Property-Based Differential Packet Marking (SPB-DIFFMARK) algorithm marks after an unvoiced/voiced transition at most the next N packets with +1 and stops the marking with +1, if the packet is classified as unvoiced. All packets, which are not marked with +1, are marked with 0 or –1, depending whether the number of +1 and of –1 are equal again. The number of fore- and background packets should level over the long term due fairness requirements.

Sanneck proposes to use a modified Random-Early-Dropping (RED) at packet forwarding nodes. If a node is congested, the probability of packet dropping should depend on its marking. Packets with a +1 will be dropped at a low probability, packets marked with –1 with the highest probability. This algorithm has been evaluated under a couple of different loss patterns using objective speech quality evaluation algorithms (MNB and EMBSD). The authors showed, that the SPB-DIFFMARK algorithm increases perceptual quality of VoIP, compared to alternative algorithms, like the alternating or random marking of packets.

3. Quality Metric

The MOS is an intuitive measure to compare the quality of audio transmissions. It is widely applied as listening quality scale. To determine the MOS value, humans evaluate a degraded sample. A MOS value between 1 (bad) and 5 (excellent) is assigned to rate the quality of the degraded sample in respect to the expected quality of the original sample.

If a sample is encoded, transmitted and decoded, the maximal achievable quality of transmission is limited to the coding performance, which depends on the codec algorithm, its implementation and the sample content. Some samples are more suitable to be compressed than others. For a sample *s*, which is coded with the encoding and decoder implementation *c*, the quality of transmission is MOS(*s*,*c*). The sample s has a length of t(*s*) seconds. One should note, that the length of a sample excludes the leading and subsequent periods of silence, which are usually not relevant to evaluate perceptual quality.

In a VoIP system, the quality is not only degraded by encoding but by packet losses, too. If frame losses occur, the resulting quality is described by MOS(*s*,*c*,{l_1,l_2,…}). The values of l_x describe the loss of a speech frame at position *x*. Let us define *the importance of frame losses as the difference between the quality due to coding loss and the quality due to coding loss and frame losses, times the length of the sample:*

$$\begin{aligned}&\mathrm{Imp}\left(s,c,\{l_1,l_2,\ldots\}\right)=\\&\left(\mathrm{MOS}\left(s,c\right)-\mathrm{MOS}\left(s,c,\{l_1,l_2,\ldots\}\right)\right)\cdot \mathrm{t}\left(s\right)\end{aligned} \tag{1}$$

3.1. Loss Event

In general, the impairment due to a frame loss i is temporally limited and starts at t^i_{start} (the starting time of the frame i) and any point in time t^i_{end}, until the perceptual relevant error propagation has vanished. The impairment period is the period [t^i_{start}; t^i_{end}] and has a duration of t^i_{start} - t^i_{end} seconds. The losses of two frames are correlated, if the impairment periods overlap. If two or more losses are correlated, we talk about a single *loss event*.

3.2. Additively Property of Importance

If a quality model should be applied to mark VoIP packets, it should be possible to give a statement like "Packet A and packet B are as important as packet C" and "Packet A is three times more important than packet B". In a mathematic sense, the requirement is called additively property. In the following, we develop two ways, how importance values can be added. One solution is based on the MOS scale; the other applies the psychological scale of the ITU E-Model.

a) Assuming that two loss events le_1 and le_2 are not correlated than the following equation is valid:

$$\mathrm{Imp}\left(s,c,le_1\cup le_2\right)\approx \mathrm{Imp}\left(s,c,le_1\right)+\mathrm{Imp}\left(s,c,le_2\right) \tag{2}$$

b) Studies on the quality evaluation of video signals showed, that if two sources of impairments are not correlated, they are additive on a psychological scale. ITU E-Model [11] introduced a psychological scale to compute the expected quality of telephony systems. The primary output of E-Model is the "Rating Factor" *R*. It ranges from 0 (worst) to 100 (best). The E-Model adds the impairments due to echo, delay, coding, speech losses, and other factors to calculate the R-Factor.

We apply the additive property of the R-Factor scale for losses: If two losses are not correlated temporal, the speech quality impairments can be added by converting the MOS values to a psychological scale, calculate the importance with a formula similar to (1) and adding the importances. Assuming that R is a function to convert MOS values to a psychological scale, we define the importance on the psychological scale as

$$\begin{aligned}&\mathrm{Imp}_R\left(s,c,le\right)=\\&\left(\mathrm{R}\left(\mathrm{MOS}\left(s,c\right)\right)-\mathrm{R}\left(\mathrm{MOS}\left(s,c,le\right)\right)\right)\cdot \mathrm{t}\left(s\right)\end{aligned} \tag{3}$$

E-Model described only a converting formula to obtain the Mean Option Score (MOS) from the R Factor in ITU-T Rec. G.107 (2000) (Equation B-4) [11]:

$$MOS\left(R\right)=\begin{cases}R\le 0: & 1\\ 0\le R\le 100: & 1-\frac{7}{1000}\mathrm{R}+\frac{7}{6250}\mathrm{R}^2-\frac{7}{1000000}\mathrm{R}^3\\ R\le 100: & 4.5\end{cases}$$

This formula cannot be inverted [12] because it calculates for each *R*-value between 0 and $80-30\sqrt{6}\approx 6{,}5$ a MOS value below 1. In order to calculate one R factor from a given MOS value we limit the value range to $[6.5;100]$ and inverted it with the Candono's Formula to the following equation, which – in the mean time – as been adopted by the ITU:

$$R\left(MOS\right)=\frac{20}{3}\left(8-\sqrt{226}\cos\left(h\left(MOS\right)+\frac{\pi}{3}\right)\right) \tag{4}$$

with

$$\begin{aligned}&h\left(MOS\right)=\tfrac{1}{3}\arctan\left(18566-6750MOS,\right.\\&\left.15\sqrt{-903522+1113960MOS-202500MOS^2}\right)\end{aligned}$$

4. Measurement Setup

In order to study the impact of packet loss events on the speech quality, we have conducted extensive simulations using the scenario illustrated in figure 1:

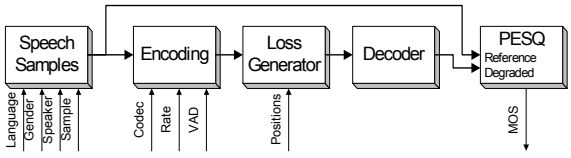

Figure 1: Evaluation set-up

4.1. Speech Recordings

We have used speech recordings, which are taken from an ITU coded speech database [13] that consists of 832 files, each 8s long, 16 speakers, half female and half male, spoken in four different languages, without background noise. We have chosen this database to limit the influence of specific languages, speakers, or samples.

To suppress surrounding influences that are not the scope of our study, background noise is not added to the speech samples.

4.2. Speech Coding and Decoding

We have chosen three common speech-coding algorithms: ITU's G.711 and G.729, and ETSI's Adaptive-Multirate (AMR).

4.2.1. ITU G.711

ITU G.711 [14] is applied for encoding telephone audio signal at a rate of 64 kbit/s with a sample rate of 8 kHz and 8 bits per sample. G.711 can operate in two modes a-law (European) and µ-law (US). LAW. The µ-law mode is applied in this paper. We added the packet loss concealment (PLC) from ITU G.711 Appendix I [15], which limits the impact of transmission losses. The PLC algorithm works on a frame size corresponding to 10ms.

4.2.2. ITU G.729

ITU G.729 [7] is an algorithm for the coding of speech signals at 8 kbit/s using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP). The coder operates on speech frames of 10 ms using a sampling rate of 8000 samples per second. For every 10 ms frame, the speech signal is analysed to extract the parameters of the CELP model (linear-prediction filter coefficients, adaptive and fixed-codebook indices and gains). These parameters are encoded. At the decoder, these parameters are used to retrieve the excitation and synthesis.

An error concealment procedure is included in the decoder to reduce the degradation in the reconstructed speech in case of frame losses. The concealment strategy extrapolates the current frame, based on previously received information. The method replaces the missing excitation signal with one of similar characteristics, while gradually decaying its energy. Depending whether last good frame is classified as periods or not, the concealed frame is being forced to be periodic or non-periodic as well.

4.2.3. AMR

The Adaptive Multi-Rate (AMR) speech codec by the third Generation Partnership Project (3GPP) [16] is the mandatory codec for UMTS. The encoding scheme 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. The decoder implements a loss concealment algorithm [17]. The concealment decreases the gains with each lost frame, and continues to apply other coding parameters unchanged.

4.3. Loss Generator

We simulated packet losses at different positions in the voice stream. The loss positions are in the middle of the sample to ensure that in any case surrounding audio segments are present. Packet losses cover a period of 1, 2, 3, and 4 speech frames.

We set up a large database of the quality impairments due to losses, which contains following the value of $\mathrm{MOS}(s,c,le)$: Samples $s = 1\ldots832$, Codecs $c \in \{G711, G729, AMR4.75, AMR12.2\}$. Loss events $le \in \{\varnothing, le^1, le^2, le^3, le^4\}$ at the positions $le^1 \in \{\{l_{3500ms}\},\{l_{3520ms}\},\ldots,\{l_{4480ms}\}\}$ and $le^2 \in \{\{l_{3480ms}, l_{3500ms}\},\{l_{3500ms}, l_{3520ms}\},\ldots,\{l_{4460ms}, l_{4480ms}\}\}$ and le^3, le^4 respective. All together, the database covers 690000 different speech quality measurements in case of single, individual loss events. Furthermore, we simulated mean packet loss rates at randomly distributed packet losses at different loss rates and packet lengths.

4.4. Speech Quality Evaluation

To evaluate the speech quality, we applied the ITU's PESQ algorithm [17], which predicts fairly well the rating behavior of human beings. It compares the original speech sample with the degraded speech sample to calculate a MOS value.

The psychoacoustics models like PESQ are an abstraction of the human rating behavior. The parameters of those mathematical models are selected so that the model shows a high correlation to subjective test results. A model performs well, if it predicts the subjective results even for unknown samples. PESQ has been verified for impairment due to coding loss and normally distributed packet losses. However, it has not been designed to measure the impact of specific packet losses, thus it might fail for this kind of specific impairment. Currently, no subjective listening tests are known, which compare the different kinds of packet loss impairments (e.g. clipping, silence, etc.) Therefore, a proper fine-tuning or verification of PESQ is not possible, yet.

In that sense, PESQ can be used only as a tool, which provides a first impression about possible packet loss importances. The actually perceptual impairment of specific kind of losses has to be proven in future subjective tests. We started to verify selected test condition by informal listening tests (samples can be heard at [19]).

4.4.1. Accuracy of PESQ

Even for test cases, which PESQ has been designed for, it might not perform perfectly. In [20], the author measured the prediction performance of PESQ. He compared the speech quality prediction of PESQ with human conducted subjective tests, covering test conditions with impairments due to coding distortion and packet losses.

The difference between the PESQ MOS value and the MOS from subjective tests is called the residential error: $\left|MOS_{human} - MOS_{PESQ}\right|$. The author showed, that the residential error is below 0.25 for 70% and below 0.50 for 90% of all test conditions. The correlation between subjective and objective tests is about 0.93. If two different test conditions are compared to conduct a competitive analysis of speech quality, PESQ was able to identify the difference (A is better than B or A is worse than B or no difference) in about 70% of all outcomes. In 30% of all comparisons PESQ detected the difference not at all, wrongly, or detected a difference without any reason.

The author concluded, that the predicting performance of PESQ is only high for the correlation between test results. Absolute MOS value might be very imprecise. Thus, PESQ might not be able to predict correctly, whether a minimal service quality (e.g. MOS=4.0) is fulfilled.

Consequently, we use PESQ based MOS results only for comparison with each other. Thus, we are able to benefit from the relative predictor accuracy, without to suffer from the absolute measurement error.

4.4.2. Resource requirements

To limit the impact of measurement error of a single test condition, we conducted a large test campaign including different samples and loss positions to yield better statistical findings. Our testing campaigns consists out of 690000 different test conditions, each has a length of 8s. It would take about 2 years to conduct all those tests by humans, which is not practical at all. Even using PESQ, we had to develop a parallel processing network [21], which conducts the evaluation of test conditions on multiple workstation in parallel. Thus, the calculation time has been limited to a couple of days.

4.5. Packet Selection Algorithms

In addition to the impairment of different losses, the speech properties of each frame are added to the database. The encoders and decoder are extended to provide speech properties of each frame. The parameters include

- The voice activity as measured by the voice activity detecting algorithms, which are included in the G.729 and AMR encoders.
- The voicing decision of the G.729 and AMR decoder.
- The marking criteria of the Source-Driven-Packet-Marking algorithm (4.5.1)
- The marking criteria of the SPB-DiffMark algorithm (4.5.2)

To obtain the relation between given set of speech properties and the importance, the frames with those properties are selected to form a subgroup of test conditions. The importance of this subgroup is analyzed in respect to the mean, the variance, the distribution, and the extremes. Thus, we can correlate speech properties with the packet importance.

4.5.1. Source-Driven-Packet-Marking implementation

De Martin has described an algorithm, how to mark speech packets as important. We re-implemented his algorithm on basis of his publication, because the original implementation is not available anymore. In the following, we will describe the source-driven packet-marking algorithm, as we interpreted and understood it.

a) As a minimal requirement, only frames during voice activity are considered as importance packets.

b) For those active frames, the decoder decodes the transmitted frame, and secondly conceals the same frame position, as if the frame would have been lost. (Before decoding and concealing the correct internal decoder state is ensured). Both decoded and concealed frames are compared.

c) The linear prediction filter describes the spectral envelop of the decoded signal. If the spectral distance of the spectral envelope is larger than 4 dB, packets are marked as important. To calculate the spectral distortion, we use the LPC vector Az from the first of two subframes. The power at a given frequency f is calculate with

$$P\left(\overline{Az}, f\right) = \left| \sum_{i=0}^{M} Az[i] e^{\frac{2\pi \cdot f \cdot i}{8000}} \right|^{-1}$$

We assume a frequency band ranging from 125 Hz to 3125 Hz, which is the telephone frequency band. We determine the spectral distortion with the following equation being Az_1, Az_2 the LPC vectors of the transmitted and the concealed frame.

$$SD\left(Az_1, Az_2\right) = \frac{\sqrt{\sum_{f=125, incr.=30}^{3125} \left(10 \lg\left(\frac{P(Az_1, f)}{P(Az_2, f)}\right)\right)^2}}{100}$$

d) The next marking criterion is the difference between the adaptive codebook indices. If the previous frame is considered as voiced and the value of $\Delta P_i = \left|1 - \frac{P_i^{loss}}{P_i}\right|$ is greater than 20%, the corresponding frame is marked. As value of the adaptive codebook index we use the pitch delay parameter P1 from first subframe.

e) If a frame is voiced, the adaptive-codebook gains of the second subframe are used to calculate the gain difference: $\Delta AG_i = 10 \lg\left(\frac{AG_i}{AG_i^{loss}}\right)$. If it exceeds a value of 5dB the frame is marked. Similar, the fixed-codebook gain difference is calculated for unvoiced frames: $\Delta FG_i = 10 \lg\left(\frac{FG_i}{FG_i^{loss}}\right)$. The same threshold is applied for the unvoiced frames, too.

One should note, that our algorithm might differ from the De Martin's implementation. However, the author has confirmed, that our implementation complies to his algorithms but does not seem to mark all the low-energy packets as best-effort.

4.5.2. The SPB-DiffMark implementation

The implementation of SPB-DiffMark is publicly available [22]. It uses a modified decoder to obtain the information, whether a frame is voiced or unvoiced. As in the original implementation, we do not treat packets different if they are silenced. Thus, some packets are identified as voiced, even those the voicing decision is based on a frame that has a very low energy.

5. Results

5.1. Mean Packet Loss Rate

As a reference we measured the impact of randomly distributed speech frame losses on the speech quality. Our results show a high correlation with subjective listening test results (e.g. from in [2]). If the coding rate is high, the MOS value is high, too. As expected, G.711 provides the best speech compression, followed by the AMR 12.2, G.729 and AMR 4.75 codecs. However, the codec with a lower coding have a better bandwidth efficiency.

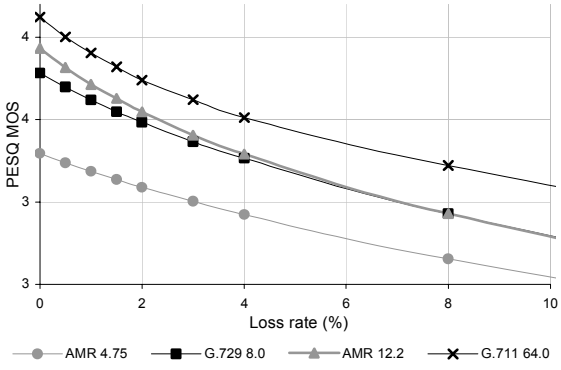

Figure 2: Impact of Loss Rate on Speech Quality

5.2. Adding importance of frame losses

If we apply the additivity property of importance (formula (2)), mean importance of a frame loss can be calculated by sum of N speech frame importances divide by N.

$$\mathrm{Imp}(s,c,le_{mean}) = \frac{\sum_{i=1}^{N} \mathrm{Imp}(s,c,\mathrm{le}_i)}{N} \qquad (5)$$

In section 5.1 we measured the mean speech quality due to randomly distributed losses. If one assumes that the most losses are not correlated, which is true at least at low loss rates, and that each loss is a single loss event, we can write the following equation to obtain the mean importance of a speech frame.

$$\mathrm{Imp}(s,c,le_{mean}) = \left(MOS(s,c) - MOS(s,c,\{l_1,\ldots,l_N\})\right) \cdot \frac{t(s)}{N} \qquad (6)$$

In the following table, we plot the mean importances on a speech frame depending of the low loss rate. One should note, that an AMR frame has a length of 20ms and G.729 and G.711 a length of 10ms. Therefore AMR frames are more important.

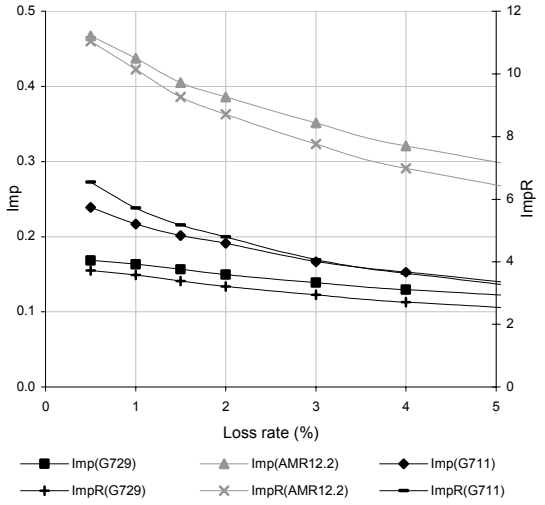

Figure 3: Loss rate (%) vs. importance

If the number of losses increases, the importance decreases. If the additive property of the importance is given and the losses are not correlated, the importance remains the same. The results show, however, that calculating the importance on the MOS scale instead of the R factor scale approximates the additive property better. Furthermore, the importance depends on the loss rate.

5.3. DTX

If frames are lost during silence, the impairment is scarcely audible. In table 1, the importance of mean active and silence frames is listed. In general silence frame are hundred times less important than active. Thus, the DTX algorithms perform well and are a good indicator of unimportant frames.

Table 1: Voice activity vs. importance

Voice	AMR 4.75	G.729	AMR 12.2	G.711
All[1]	0.113	0.173	0.269	0.393
Active	0.389	0.655	0.923	1.338
silence	0.003	0.004	0.008	0.016
Number of active	28%	31%	28%	28%

Taken all speech frames, that contain active voice, we plot a histogram of the frequency of the importance (figure 4). Even through the mean importance of an active frame is quite low, there is still a large amount of frames that are highly important.

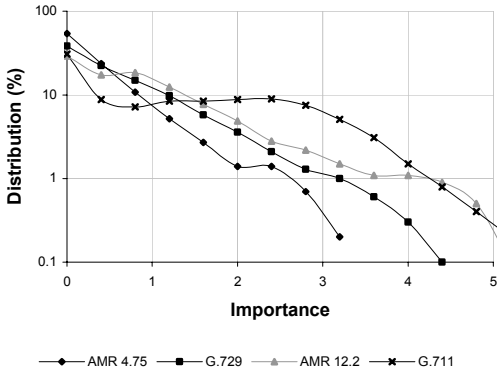

Figure 4: Histogram of frequency of importances (frames during voice activity)

5.4. Loss of a VoIP packet

In a VoIP system, one or multiple speech frames are put together in one packet. If the packet got lost, one or multiple speech frames are lost, too. According to our definition in section 3.1, multiple frame losses are seen as one loss event. In table 2, the importance of a loss event is listed.

Table 2: Loss length vs. importance of VoIP packet

Length of loss	AMR 4.75	G.729	AMR 12.2	G.711
10 ms		0.173		
20 ms	0.113	0.410	0.269	0.393
30 ms		0.591		
40 ms	0.229	0.700	0.462	0.992
60 ms	0.329		0.630	1.209
80 ms	0.411		0.764	1.331

Clearly, the importance of a loss event increases with the number of dropped frames. The importance per single frame depends on the packet's length. For example, losing two 20ms G.711 packets is better than losing on 40ms packet. On the other side, losing two 40ms packets is worse than losing on 80ms packet. Similar rules are valid for the other coding schemes, too.

1 The importance of mean packet differ to figure 3 because of the different amount of silence periods.

5.5. Source-Driven Packet Marking

The Source-Driven Packet Marking applies four criteria to mark as premium. In the following table, the importance and frequency of speech frames are listed.

Table 3: De Martin: Importance of marked frame, listed for each selection criterion

G.729	Voiced		Unvoiced	
	Imp.	Count	Imp.	Count
All	0.611	100%	0.353	100%
$SD > 4dB$	1.113	0.73%	1.773	1.36%
$\Delta P > 20dB$	0.435	24.8%	-	-
$\Delta AG > 5dB$	0.364	6.66%	-	-
$\Delta FG > 5dB$	-	-	1.358	8.02%

The prediction performance of De Martins algorithm is high for unvoiced packets. For voiced packet, using the difference of the codebook indices and the difference in the gains are less promising, but the spectral distance is a good predicting parameter for voiced frames, too. If the selection criteria for important packets are combined, the Source-Driven Packet Marking algorithm behaves as shown in the following table. As described in the paper, about 20% of all frames are marked.

One drawback of De Martins algorithm is, that just the quality impairment of the current frame is analyzed. Any error propagation is not taken into account.

Table 4: De Martin: Importance of unmarked and marked speech frames

DeMartin	Importance	Count
All frames	0.504	100%
Normal (unmarked)	0.484	78.5%
Premium (marked)	0.577	21.5%
Voiced (as reference)	0.611	58.5%
Unvoiced (as reference)	0.354	41.5%

5.6. SPB-DIFFMARK

We measured the importances of loss after and before an unvoiced/voiced transition (figure 5). None of our coding schemes shows that the concealment schemes perform badly during an unvoiced/voiced transition.

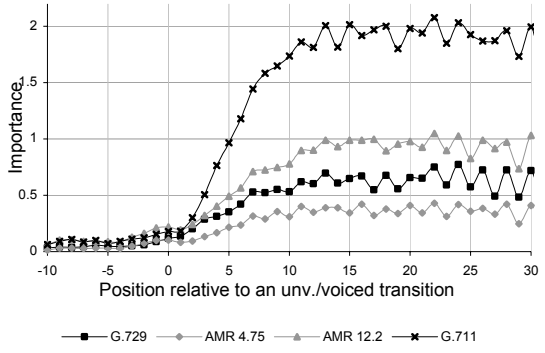

Figure 5: Impact of an unvoiced/voiced transition

SPB-DiffMark marks packets as –1, 0 or +1 depending on the history of speech properties. Using PESQ and our speech sample, the algorithm performs as following:

Table 5: Sanneck: Importance of with –1, 0 and +1 marked speech frames

Sanneck	Importance	Count
All frames	0.217	100%
	0.246	56.3%
0	0.105	15.0%
-1	0.066	28.8%
Active (as reference)	0.655	31%
Silence (as reference)	0.004	69%

As expected, +1 packets are most important followed by packets that are marked with 0 or –1. The number of +1 is not equal to the number of –1. This fact is due to the length of our samples, we just analyzed short periods of speech. Most unvoiced/voiced transitions occur during the end of the analyzed period. Therefore the SPB-DiffMark cannot level the number of –1 and +1 packets.

The SPB-DiffMark algorithm benefits by using the voicing criterion as a quality prediction. Its predicting performance can be easily enhanced, if active/silence detection is added (as originally intended). We cannot verify a high impairment after an unvoiced/voiced transition, if we use PESQ as objective speech evaluation tool. However, we can confirm that, if the EMBSD tool is used, the unvoiced/voiced transition has a higher impact. These circumstances lead us to the conclusion, that strength of single packet loss impairments depends largely on the perceptual model, which is applied for evaluation. Further subjective tests need to be conducted to verify and to enhance the predicting performance of objective speech quality evaluation tools.

6. Conclusions

The importance of a packet or the impairment, what its loss will cause, can be predicted by two classes of quality models. The first class knows the entire transmission of a sample including all losses. The second predicts the importance of packets during run-time.

To evaluate the entire transmission we applied PESQ even though its prediction performance is verified only for mean and not for specific packet losses. Using PESQ as reference, we evaluated three algorithms that predict the importance of a packet at run-time. Each algorithm has its strengths and weaknesses. It is straight forward to combine the best from all three algorithms to get an better model: using DTX to identify unimportant frames, using De Martin's spectral distance and gain difference, and using Sanneck's voicing distinction.

Our future research will focus on enhancing the prediction performance of packet loss quality models. We plan to develop an algorithm, which is based on subjective listening tests. It has to predict the concealment at the decoder like De Martin's algorithm. Furthermore, up-to-date perceptual models (like PESQ) have to been applied to measure the distortion between transmitted frame and the concealed data. The amount of error propagation has to be predicted by heuristic means (similar to Sanneck's approach). A neural network or a hidden-markov-model might be applied to approximate the length and strength of the error propagation. But even the existing algorithms are good enough to enhance VoIP in a QoS capable network by a distinguished packet treatment, as early research results have shown. We expect that our quality model will enable further performance gains.

7. References

[1] Schulzrinne, H. and Rosenberg, J., "Internet telephony: architecture and protocols - an IETF perspective", *Computer Networks and ISDN Systems,* Vol. 31, pp. 237-255, Feb. 1999.

[2] Blake, S. et al., *An architecture for differentiated services*. RFC 2475, Network Working Group, Dec 1998.

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

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

[5] Hoene, C., Carreras I., and Wolisz A., "Voice Over IP: Improving the Quality Over Wireless LAN by Adopting a Booster Mechanism - An Experimental Approach", In P. Mouchtaris, editor, *Proc. of SPIE 2001 - Voice Over IP (VoIP) Technology*, pp. 157-168, Denver, Colorado, USA, Aug. 2001.

[6] Tobagi, F., Markopoulou, A., and Karam, M., "Is the Internet ready for VoIP?" *Proc. of IWDC,* 2002.

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

[8] ITU-T Recommendation G.113*: Transmission impairments due to speech processing*, Feb. 2001.

[9] Markopoulou, A., Tobagi F., and Karam, M., "Assessment of VoIP quality over internet backbones." *Proc. Infocom'02*, New York, USA, 2002.

[10] Roseberg, J: "G.729 error recovery for internet telephony," *Project report, Columbia University*, 1997.

[11] ITU-T Recommendation G.107 : *The E-Model, a computational model for use in transmission planning*, Jul. 2003.

[12] ITU-T Recommendation P.833: *Methodology for derivation of equipment impairment factors from subjective listening-only tests*, Feb. 2001.

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

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

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

[16] 3GPP TS 26.090: *Mandatory Speech Codec speech processing functions AMR speech codec; Transcoding functions*. Jun. 1999.

[17] 3GPP TS 26.091: *Mandatory Speech Codec speech processing functions AMR speech codec; Error concealment of lost frames*, Apr. 1999.

[18] 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.

[19] Speech Degradation due to packet loss. URL http:// www-tkn.ee.tu-berlin.de/~hoene/singlelossevent/

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

[21] Hoene, C. , BatchDistribution Software, http://www-tkn.ee.tu-berlin.de/equipment/bd/, Apr. 2002

[22] Sanneck, H., "Software for the simulation of SPB protection for VoIP streams", URL http://www.sanneck.net/research/voice/spb-fec/ content.html, 2003.