In Proceedings of International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS'04), San Jose, California, USA, July 2004,

A Perceptual Quality Model for Adaptive VoIP Applications*

Christian Hoene, Holger Karl, Adam Wolisz
Technical University of Berlin, TKN, Sekr. FT 5-2
Einsteinufer 25, 10587 Berlin, Germany
hoene@ieee.org

Abstract

Quality models predict the perceptual quality of services as they calculate subjective ratings from measured parameters. In this paper we present a new quality model that evaluates VoIP telephone calls in order to control their transmission at run-time. In addition to packet loss rate, coding mode and delay it takes into account the impairments due to changes in the transmission configuration (e.g. switching the coding mode or re-scheduling the playout time). It is also computation efficient and open source.

To demonstrate its potential, we apply our model to select the ideal coding and packet rate in bandwidth-limited environments. Furthermore we decide, based on model predictions, whether to delay the playout of speech frames after *delay spikes*. Delay spikes often occur after congestion and cause packets to arrive too late. We show a considerable improvement in perceptual speech quality if our model is applied.

Keywords: internet telephony, adaptive VoIP, perceptual quality model, E-model, PESQ, packetization, delay spikes.

1 Introduction

Recent studies show that a significant number of Internet backbone links do not provide *toll quality* (Markopoulou et al., 2002) — the lowest quality of classic PSTN based telephone calls — when used for Voice over IP (VoIP) applications. To overcome this shortcoming, application level control is a promising approach. It can be used to complement or substitute QoS mechanisms like over-provisioning (Fraleigh et al., 2003) or DiffServ. Voice over IP applications can adapt the *VoIP configuration* to the current state of the network. In recent years, several algorithms have been proposed which dynamically tune the configuration to current packet delays and losses. These algorithms change the size of the playout buffer, the coding rate and the amount of forward error correction in order to maximize the VoIP quality. But how is the quality of a (VoIP or standard PSTN) telephone call measured?

It is obvious that the quality of telephone calls should be measured by the users: Humans should evaluate the *perceived* Quality of Service (QoS). Of course such subjective measurement campaigns are time consuming and costly if statistically meaningful results are to be obtained (ITU P.800, 1996). On the other hand VoIP applications cannot measure subjective impressions but only directly observable, network- or transport-layer metrics like packet loss rates, round trip times, and packet delay distributions. These metrics for *networking QoS*, however, do not reflect the perceived quality precisely.

An efficient way to correlate perceived QoS and networking QoS are *quality models* that simulate human rating behavior. Quality models calculate a perceptual quality rating using given networking metrics. In the last years considerable efforts have been made to predict human rating behavior using precisely measurable parameters. We will describe in short the most common quality models for telephony:

The Perceptual Assessment of Speech Quality (*PESQ*) algorithm predicts the speech quality of narrowband speech transmission. The PESQ algorithm is standardized in ITU P.862 (2001b). It compares the original and the degraded version of a speech sample to assess the speech quality with a mean opinion score value (MOS), which scales from 1 (bad) to 5 (excellent).

The quality of a telephone call is not entirely judged by the speech quality. Further factors have to be con-

*This work has been partly supported by the Deutsche Forschungsgemeinschaft (DFG) via the priority program "Adaptability in heterogeneous communication networks with wireless access".

sidered. The *E-Model* (ITU G.107, 2000) takes into account various other impairments like delay and echoes to calculate the so-called *R factor*. A higher R factor corresponds to a better telephone quality, zero being the worst value, 70 toll quality, and 100 excellent quality. One novel feature of the E-Model is the assumption that sources of impairment which are not correlated to each other can be added on a psychological scale. This allows to trade off different sources of impairment (e.g. loss versus delay) against each other.

The main drawbacks of ITU's quality models are the following: The PESQ algorithm is not able to predict the speech quality at run-time nor does it take into account end-to-end delays. As well as being computationally complex it is also patented. On the other hand the E-Model considers operational parameters which are not known or not relevant to the application. It does not consider the impairment due to dynamic adaptations. Furthermore it assumes tandem coding (transcoding) conditions (ITU G.108, 1999) and as a results leads to an imprecise correlation between loss rate and speech quality. Thus, neither quality model is suitable for adaptive VoIP applications because they work under different operational conditions and lack particular features which are demanded by adaptive VoIP applications.

In this paper we present a perceptual quality model that is primarily intended for adaptive VoIP applications. It is based on the same subjective measurements as the PESQ and E-Model. Our main contributions are the following: 1. We measured the coding distortion of the commonly used codecs with PESQ for different loss rates and loss patterns without considering tandem coding. 2. We measured the impairment of speech quality when the packet playout schedule is adjusted and determined the detrimental effect caused by switching between different coding rates. Contrary to the generally accepted view, switching coding modes does noticeably harm the speech quality. 3. We developed a formula which converts MOS values to R factors and included it in our quality model. The ITU approved this formula as a standard extension. Our quality model is open source and available on the internet (Hoene, 2004).

This model can be used in several circumstances. In particular its on-line nature enables its use within applications to judge the actual or potential benefits of modifying protocol parameters. We shall describe two such examples in a later section in more detail.

This paper is structured as follows: In Section 2 we discuss the two common quality models, the VoIP architecture and describe the requirements for an application layer quality model. Next we describe our quality model. In Section 4 we present measurement results on the coding performance of common codecs and parameterize our quality model. We also included two examples of the application of the quality model in Section 5. In the conclusion we discuss further research issues.

2 Background

2.1 VoIP System

Internet Telephony allows to offer telephony services across networks using the Internet protocols and is an alternative to the classic telephone system (PSTN). IP Telephony consists of signaling and transmission protocols. The signaling protocols (H.323 or SIP) establish, control and terminate a telephone call. In the following we will discuss the principle components of the of VoIP system, which cover the end-to-end transmission of voice (Fig. 1).

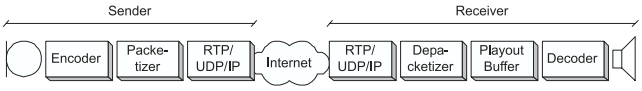


Figure 1: VoIP System

Digitized human speech is encoded. Encoding algorithms compress the audio signal. Most speech encoding schemes compress segments of speech and generate frames. The common, standardized encoding algorithms (G.711, G.723.1, G.726, G.729, GSM, AMR, AMR-WB) differ in their coding rate (bits/s), frame rate (frames/s), algorithmic latency (ms), complexity and speech quality (MOS). An important optimization opportunity for speech codecs is the fact that human speech consists of periods of voice activity and silence (Chuah and Katz, 2002). Some coding schemes lower the packet rate during silence to send only background noise descriptions (SID). This operating mode is called discontinuous transmission (DTX).

One or multiple speech frames are concatenated in one packet. RTP, UDP, and IP packet headers are added to the speech segments before the packets are sent the to receiver. Optionally, forward error correction (FEC) can be included in the packet (Perkins et al., 1997). FEC adds redundancy to the transmission so that lost packets can be recovered, as long as the following packets are received successfully. Redundancy can be either media-independent or media-dependent. Media-dependent FEC (Hardman et al., 1995; Bolot and Vega-Garcia, 1998) uses multiple coding modes to

compress the content at different rates (e.g. both G.711 and G.723.1).

The network transmits packets from the sender to the receiver. In the Internet packets can get lost because of congestion or (wireless) transmission errors. The transmission delay of packets, the time needed to transmit a packet from the sender to the receiver, is variable and depends on the current network condition and the routing path (Bolot, 1993; Bolot et al., 1995; Markopoulou et al., 2002; Kaj and Marsh, 2003). VoIP packets may be transmitted in parallel over multiple paths (Liang et al., 2001).

At the receiver, protocols process the packets and deliver them to the de-jittering buffer which temporily stores packets so that they can be played out in a timely manner. If packets are too late to be played out on time, they are usually regarded as lost. Consequently, losses as seen by the application are in fact a superposition of real losses and excessive delays, where excessive is used in terms of play-out buffer dimensioning (non-standardized algorithms). After the playout buffer the speech frames are decoded. If a frame is lost, the decoder conceals the lost frame and extrapolates the last successfully received frame (Perkins et al., 1998) into the gap. Finally the digital signal is transformed into an acoustic signal.

Application-layer adaptation can enhance the quality of VoIP because it changes VoIP configuration so that it matches the current state of the network best (Bolot and Vega-Garcia, 1996). For example, in cases of congestion, it has been proposed to change the coding rate (Yin et al., 1990; Barberis et al., 2001; Servetti and Martin, 2003). Thus, the bandwidth of a VoIP flow is lowered and the probability of further packet losses due to congestion is decreased.

Whereas congestion control in general avoids extensive packet losses, it does not avoid packet losses at all. Therefore an error control scheme should be used (Podolsky et al., 1998; Bolot and Vega-Garcia, 1998; Bolot et al., 1999). FEC is a good candidate for end-to-end error control of interactive speech transmission. The IETF has standardized a FEC scheme that adds a redundant copy of speech frames to the following packets (Perkins et al., 1997). If a packet is lost, the receiver reconstructs the lost speech frame after receiving the following frames. Of course FEC increases both bandwidth and delay. Thus, it is beneficial to jointly optimize adaptive FEC and playout scheduling (Rosenberg et al., 2000; Boutremans and Boudec, 2003).

The number of frames in one packet can be changed to adapt the packet rate and link utilization. For example, Veeraraghavan et al. (2001) have used an adaptive packetization for Voice over WLAN. Alternatively Kim et al. (2002) uses frame grouping to combine multiple voice flows in a single IP packet.

A couple of publications (Ramjee et al., 1994; Moon et al., 1998; Pinto and Christensen, 1999; Sreenan et al., 2000; Laoutaris and Stavrakakis, 2002) study how to choose the ideal time of playing out the received frame. The size of the dejittering buffer should be adjusted so that most packets are not received too late and packet losses are minimized. Furthermore the dejitter buffer for VoIP should adapt immediately to short increases in the transmission delay. The scheduling of playout can be adjusted most easily during silence because then it is not notable. Adjustments during voice activity require more sophisticated concealment algorithms (Liu et al., 2001; Liang et al., 2003).

Requirements If a perceptual quality model should be applied for the application layer of VoIP parameters, certain requirements have to be met. In general application layer control can be divided in two parts, an acoustic and a transmission control part. Although the acoustic processing is highly important we shall not discuss it in the present paper[1]. A quality model has to cope only with the static and variable impairments due to coding distortion, packet loss and delay. In the following we will discuss whether PESQ and the E-Model fulfill these requirements and whether they can be used in an adaptive VoIP application.

2.2 Perceptual Assessment

PESQ is a model for perceptual evaluation of speech quality. PESQ compares an original speech sample with its transmitted and hence degraded version. It implements a cognitive model which emulates the psychoacoustics of human hearing (Beerends et al., 2002). One novel feature of PESQ is the identification of transmission delays (Rix et al., 2002). First PESQ adjusts the degraded version to be time aligned. Then a psychoacoustic model assesses the distortion between original

[1] The acoustic processing is highly important for the perceptual quality and often neglected in the implementation of a VoIP phone. It is required to regulate the gain of the input and output signal in order to guarantee a constant and pleasant loudness of the audio signal (adaptive gain control). Another aspect is the presence of background noise, which deteriorates the performance of many encoding algorithms. Therefore an appropriate background noise suppression has to be implemented so that the human voice of the talker is filtered from the acoustic signal. Last not least, often the acoustic output is fed back to the microphone, so that a talker echo is notable. A local echo cancellation is hence required if no headset is used.

and degraded sample.

PESQ can identify both constant delay offset and variable delay jitter. Constant delays are not considered in the calculation of the MOS value, but delay variations change the rating of the speech quality.

One should note that PESQ can only be applied for distortions which have been known before its development. These are coding distortions due to waveform codecs and CELP/hybrid codecs, transmission/packet losses, multiple transcoding, environmental noise and variable delay. Benchmark tests of PESQ have yielded an average correlation of 0.935 with the corresponding MOS values under these conditions. PESQ may have to be changed before it can be applied for low-rate vocoders (below 4kbit/s), digital silence, dropped words or sentences, listener echo and wideband speech.

Even though the PESQ model can be downloaded free of charge from the ITU web page, using PESQ requires an expensive license agreement. Furthermore the computational complexity of PESQ is high. Thus, PESQ cannot be used in real-time nor it can be integrated into open-source software.

The E-Model (ITU G.107, 2000) is a computational model that can be used as a transmission-planning tool for telecommunication systems. A detailed description can be found in (Möller, 2000). One novel feature of the E-Model is the assumption that the psychological effect of uncorrelated sources of impairment is additive. The assumption is based on empirical results in the field of psychophysical research, which relate physical stimulus magnitudes to perceptual magnitudes (Allnatt, 1983).

The transmission rating factor R range from 0 to 100 and is composed of five terms which subsume different types of impairments. The terms I refer to impairment factors.

$$R = R_o - I_s - I_d - I_e + A \qquad (1)$$

R_o represents the transmission rating of the basic signal-to-noise ratio. Circuit noise, room noise at sender and receiver, sidetone, which is the sound of the speaker's own voice as heard in the speaker's telephone receiver, and noise floor, which is generated by the device itself, are factors that are taken into account. The default value of Ro equals 93.2 (Sun and Ifeachor, 2003).

The factor I_s is the sum of all impairments which occur simultaneously with the voice transmission: A too loud voice signal, quantizing distortion (A/D and D/A conversion, logarithmic PCM coding, ADPCM coding) and a non-optimum talker sidetone.

Transmission delay also impairs the quality of a telephone system. The factor I_d represents this delay impairment, which is strongly effected by talker and listener echoes. If echoes are present, the delay can be noticed more easily.

Whereas the previous I factors cover mainly classic PSTN related quality impairments, I_e takes into account all impairments caused by more complicated, new equipment. It is mainly used for predicting the coding distortion of low-rate speech codecs. Because the influence of frame losses depends largely on the type of coding and loss concealment, the frame loss rate influences I_e, too. The value of I_e can be gathered from subjective auditory tests.

The last factor A is based on the knowledge that the quality of telephone call is judged differently if the user has an advantage of access. Wireless, cellular, and satellite connections might be valued higher. For example, cellular phone users do not expect the same quality level as in PSTN telephone calls. If the Internet access is cheap or even free, VoIP might have an advantage of access, too. Typical values of A range from 0 to 20.

3 New quality model

Because both PESQ and E-Model do not fulfill all requirements (overview in Table 1) we introduce a new quality model. It takes into account coding distortion, packet loss and delay to predict the perceptual quality but it assumes an optimal acoustic processing. We split the quality model into *source* and *sink* side. The source controls the transmission of voice, based on a periodic but delayed feedback of mean packet delays and loss rates. On the other side the receiver has to react to received packets immediately. For example, the playout time may have to be adjusted to a late packet. Our quality model has to take into account both these time scales.

Features/Impairment	PESQ	E-Model	our model
coding distortion	yes	yes	yes
mean packet loss rate	yes	yes	yes
absolute delay	no	yes	yes
delay variations	yes	no	yes
single packet loss	yes	no	yes
switching the coding mode	yes	no	yes
computational complexity	high	low	low
works at real time	no	yes	yes
license free	no	yes	yes
acoustic impairments	many	many	

Table 1: Properties and features of quality models

3.1 Source Side

In the following only parameters being available at the source are considered. Equation 2 is based on the E-Model. However, if the acoustic processing is optimal, we can simplify the E-Model to fewer parameters with$c$ describing the codec, dtx the DTX mode, cr the coding rate, lr the mean packet loss rate, pack the packetization time, and t the end-to-end delay. The computation of R is then given by:

$$R = \mathrm{MOS_2R}\left(\mathrm{MOS}\left(\mathrm{c}, \mathrm{dtx}, \mathrm{cr}, \mathrm{lr}, \mathrm{pack}\right)\right) - I_d\left(t\right) \quad (2)$$

If neither talker nor listener echoes are present, the delay impairment I_d can be reduced to the term of I_{dd}: For an end-to-end delay $0 < T_a \leq 100\,\mathrm{ms}$, I_{dd} is 0. For any $100\,\mathrm{ms} \leq T_a < 500\,\mathrm{ms}$ is

$$I_{dd}\left(T_a\right) = 25\left(\left(1+X^6\right) - 3\left(1+\left(\frac{X}{3}\right)^6\right)^{\frac{1}{6}} + 2\right) \quad (3)$$

with $X = \frac{-2+\lg T_a}{\lg 2}$. The mean opinion score can be obtained from the R Factor with a conversion formula. For $6.5 < R < 100$, this conversion formula can be inverted:

$$\mathrm{MOS_2R}\left(x\right) = \frac{20}{3}\left(8 - \sqrt{226}\cos\left(h + \frac{\pi}{3}\right)\right) \quad (4)$$

$$h = \frac{1}{3}\arctan 2\left(18566 - 6750x,\right.$$
$$\left. 15\sqrt{-202500x^2 + 1113960x - 903522}\right)$$

In section 4 we derive $\mathrm{MOS}\left(\mathrm{c}, \mathrm{dtx}, \mathrm{cr}, \mathrm{lr}, \mathrm{pack}\right)$ values from PESQ measurements. In a real implementation, the values would typically be stored in a table for efficiency reasons. If the table does not contain a parameter but only a next higher and next lower value, the MOS value is calculated by linear interpolation of available values.

3.2 Sink Side

At the receiver we like to introduce a novel view on quality: The quality is degraded by a continues flow of *impairment events* that relate directly to a single psychophysical stimulus. An impairment event decreases the quality of the transmission temporally. It starts at some point in time t_{start} and lasts until t_{end}, when it is not notable anymore. In a VoIP system, three different events cause an impairment. First if one or multiple consecutive frames get lost, the quality decreases as the receiver-side concealment algorithm can not extrapolate the acoustic signal. Second, if the playout scheduler changes the playout time, the speech may be impaired (Fig. 5). Last, switching the coding mode or coding rate can cause "clicking" sounds (Fig. 4).

Impairment events can be measured by the duration and the strength of distortion. Let us define a measure, which has been applied in a similar context (Hoene et al., 2003). 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 (as 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 $\mathrm{MOS}(s, c)$. 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 perceptual quality. If impairment events occur, the resulting quality is described by $\mathrm{MOS}(s, c, e_1, e_2, \ldots)$. The values of e_x describe the impairment events at position x.

The impairment of an event is defined as the difference between the quality due to coding loss and the quality due to coding loss and the change of VoIP configuration, times the length of the sample:

$$\mathrm{Imp}\left(s, c, ev\right) = \left(\mathrm{MOS}\left(s, c\right) - \mathrm{MOS}\left(s, c, e\right)\right) \cdot t\left(s\right) \quad (5)$$

In Section 6 we will show how our new quality model, the measure of impairment, can be used to trade off packet loss bursts against playout adjustments.

4 Tuning the Quality Model

In the previous section we introduced the abstract notion of our quality model. Still, the absolute parameters and variables are to be defined. For example, we introduce the function MOS(...) , which stores MOS values for various operating conditions. Also, we introduce the notion of an impairment events. The objective of the following speech quality measurements is to determine the curve and values these functions so that the quality model developed here can use these values. To limit the length of this paper we will constrain the number of codecs to one, the Adaptive-Multi-Rate coding (AMR), which is the default codec for third generation WCDMA systems (3GPP, TS 26.071, 2002).

4.1 Measurement Setup

We followed the recommendation ITU P.833 (2001a), which describes how to derive the equipment impairment factor I_e from listening-only tests, but we used fewer test cases and instrumental assessment tools. Each single measurement consists of five steps and is repeated several times with different configurations (see Fig. 2).

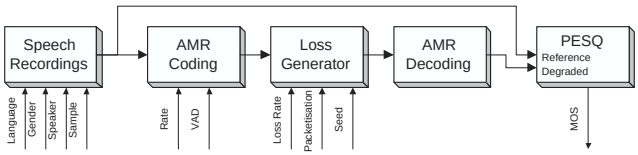


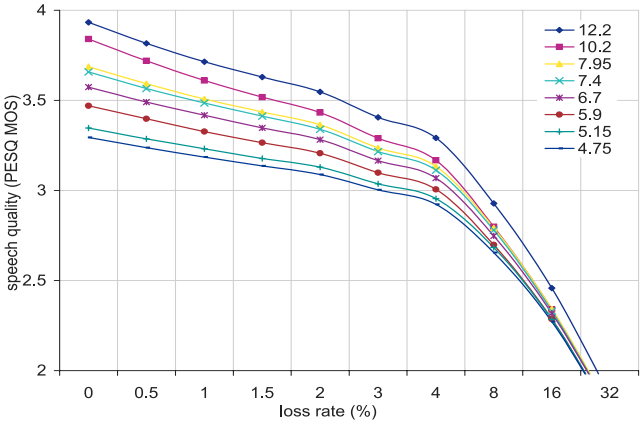
Figure 2: Measurement Set-Up

First, a speech recording is selected from a data base. We used the ITU P.suppl 23 data base (1998) that contains 832 samples from different languages, speakers and sentences. Each sample has a duration of 8s. Additional background noise is not present. Second, the ITU reference implementation of AMR compresses the sample. AMR generates speech frames. Each *frame* contains 20ms of speech and can be encoded with an coding rate of 4.75, 5.15, 5.75, 6.7, 7.2, 7.95, 10.2 or 12.2 kbit/s. Third, a loss generator simulates the packet losses depending on the loss probability, packetization and random seed. Next, the AMR decoder generates a degraded version of the speech sample and conceals lost frames. Finally the ITU reference implementation of the PESQ algorithm compares the degraded speech sample with the reference sample to calculate the MOS value.

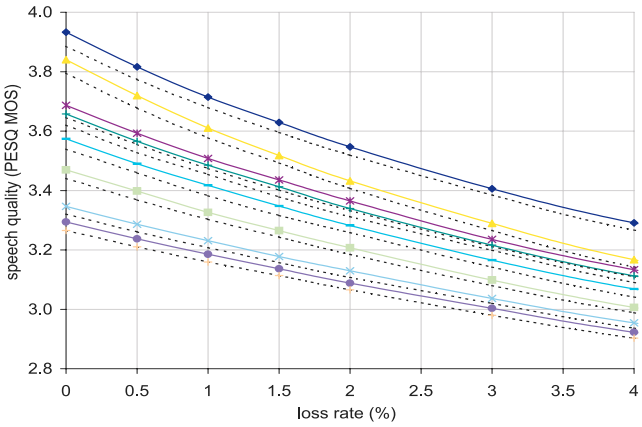
4.2 Results

We study the impact of single random losses on the instrumental speech quality. Figure 3a shows the impact of loss and coding rate on the objective speech quality with a packetization of one frame per packet. A lower coding rate and a high loss rate decrease the speech quality. Figure 3b displays the distortion due to silence compression, which is present but low. Figure 3c shows that a higher packetization does not change the speech quality to a large extent.

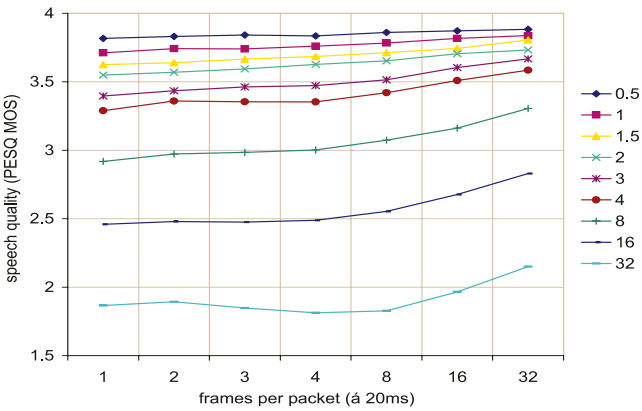
In the following we show the distortion caused by frequent switching of the coding rate (Fig. 4) versus the mean coding rate. During the encoding of a sample, which has a length of 8 s, we switch the coding rate six



(a) without silence suppression (DTX)



(b) with DTX (dotted line) and without (straight)



(c) Impact of packetitzation vs. packet loss rate

Figure 3: Impact of coding rate and loss rate

times. Also we calculate the average coding rate. Figure 4 also contains cases without any mode switching, which have an impairment of zero.

Because playout schedulers adjust the playout time of speech frames, we measured those adjustments as well. We consider one adjustment within a 8 s sample and distinguish between adjustments during voice activity (Fig. 5a) and silence (Fig. 5b). A *positive* adjustment extends the degraded sample. The resulting gap is concealed by the decoder's concealment algorithm. A *negative* adjustment shortens the degraded sample. As a comparision we also measured the impairment caused by a *loss burst* which has the same length as the positive adjustment's gap. During silence PESQ does not consider adjustments to up to one second as harmful. Adjustments during voice activity decrease the speech quality and increase the impairment.

5 First Example: Limited Bandwidth

In the this example we apply our quality model to the problem of adapting a VoIP flow to limited available bandwidth. To our best knowledge the problem of how to adapt both coding rate and packet rate to limited bandwidth has never been studied in published literature. Our parameterized quality model allow us to analyse this question. We assume that the capacity of a connection remains constant and is known. The transmission delay of a packet is given and remains constant for each packet. The question to answer is how to choose the optimal coding rate and packetization under these circumstances. We discuss this issue on a circuit switched link and a packet-switched, Ethernet-like link.

Circuit Switched Link: Let us assume a channel that has a limited bandwidth and carries one stream of AMR coded frames. If the coding rate exceeds the bandwidth of the channel, frames are dropped. The loss rate L depends on the bandwidth of channel B_c and the bandwidth of the flow B_f, which is equal to the coding rate B_s (see equation 6).

$$L = \begin{cases} B_c > B_f : & 0 \\ B_c \leq B_f : & 1 - \frac{B_c}{B_f} \end{cases} \quad (6)$$

Clearly there is a tradeoff between coding rate and loss because both decreasing coding rate and increasing loss rate will lower the speech quality. In Fig. 6 the tradeoff between loss rate and coding mode is displayed taking into account Equation 6 and the measurement data of Fig. 3. If the loss rate exceeds a value of about

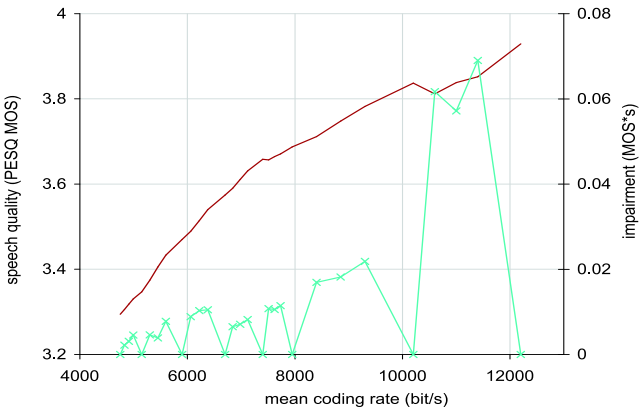
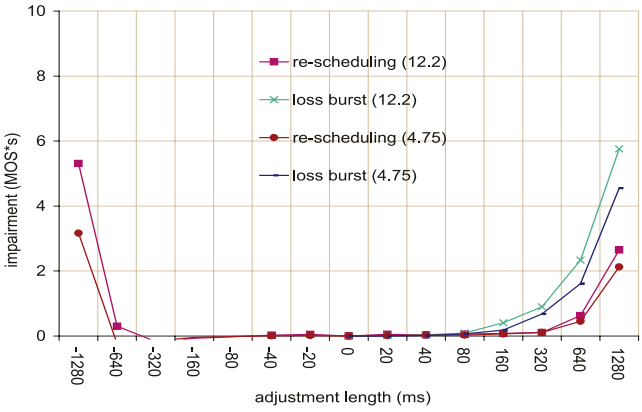
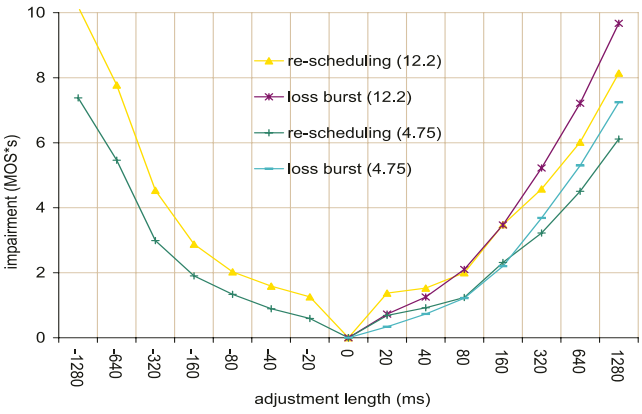


Figure 4: Impairment of switching the coding mode



(a) During silence



(b) During voice activity

Figure 5: Impact of playout re-scheduling

0.5%, a better speech quality is achieved by a lower coding rate — the drop in MOS is very sharp, if the coding rate exceeds the available bandwidth. As expected, voice flows are highly sensitive to losses and packet losses should be avoided by switching to a lower coding rate.

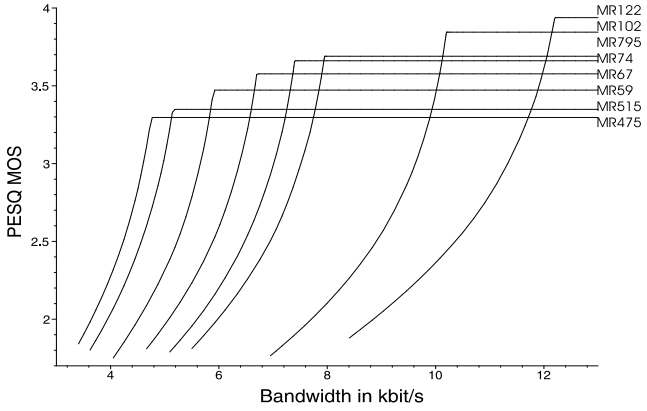


Figure 6: MOS vs. bandwidth and coding rate (MR475=4.75kbps, MR122=12.2kbps).

Full-Duplex Ethernet Link: Next we assume a full-duplex, switched Ethernet link, which bypasses the CSMA/CD medium access protocol and has a capacity of B_c. Speech frames are generated at rate of r. A packet consists of f speech frames. In addition VoIP packets contain protocol headers: The Ethernet header is 26 bytes long (8 bytes preamble, 14 bytes header and 4 bytes CRC), IP (8 bytes), UDP (20 bytes) and RTP (12 bytes). A short header is added (6 bits) in front of each speech frame (Sjoberg et al., 2002). The size of a packet p is rounded to the next byte, if its size is a fraction of a byte:

$$p = 8 \left\lceil \frac{628 + (B_s/r + 6)\, f}{8} \right\rceil \tag{7}$$

We can calculate the flow bandwidth B_f using the packet size p, the number of frames per packet f and the frame rate r.

$$B_f = \frac{p \cdot r}{f} \tag{8}$$

The loss rate depends on the bandwidth of the flow B_f and of the channel B_c as described in Equation 6. In addition to the impairment due to loss, multiple frames in a packet introduce an additional packetization delay which we have to consider. Thus, we apply equation 2 to take into account both loss and delay and obtain the following equation. The system delay t_{sys} is the end-to-end transmission delay without the packetization delay.

$$\begin{aligned} R = \quad & \mathrm{MOS_2R}\left(\mathrm{MOS}\left(\mathrm{c}, \mathrm{dtx}, \mathrm{cr}, \mathrm{lr}, \mathrm{pack}\right)\right) \\ & - I_{dd}\left(f/r + t_{\mathrm{sys}}\right) \end{aligned} \tag{9}$$

In Fig. 7 we show the optimum VoIP configuration (as rated by the *R factor*) if both packet and coding rate are ideally chosen under limited bandwidth. We assume the AMR codec (50 packets per second) and 150 ms system delay. The figure shows that the packetization increases if the available bandwidth drops. Only at a very low bandwidth the coding rate decreases also. In the figure we do not plot the packet loss rate because it is zero nearly all the time.

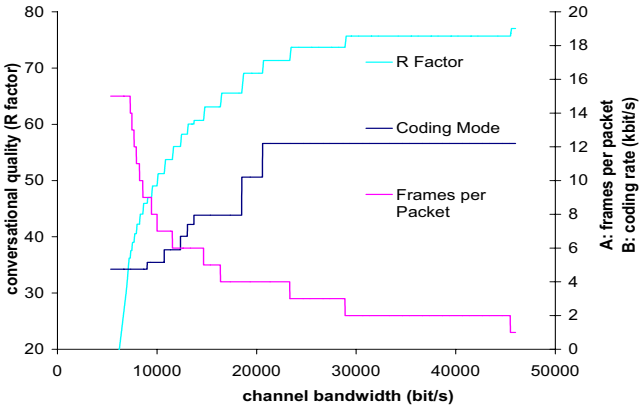


Figure 7: Choosing optimal coding rate and packetization on packet-switched link.

6 Second Example: Delay Spikes

As an example on how to use our quality model on the sink side, we apply it to an adaptive playout algorithm. The size of a playout buffer should be chosen in a way that both the number of too late frames and the additional delay are low. Common playout buffer algorithms adapt the size of the playout buffer to the transmission history to find an optimal trade-off between losses and delay. However, analysis of Internet traces show that sometimes packet delays show a sharp, spike-like increase (Ramjee et al., 1994) which cannot be predicted in advance. After a spike, packets are received at a high frequency. Soon afterwards the jitter process returns to normal. We like to consider the question whether to adjust the playout of speech frame to delay spikes, using

the quality model introduced in this paper. We concentrate on the non-trivial case of delay spikes during voice activity.

Frame F_n arrives too late to be played out on time. No consecutive frames F_i with $i > n$ have been received so far. The scheduled playout time of frame F_n is $t^n_{playout}$, but the frame has arrived at $t^n_{arrive} > t^n_{playout}$. At the arrival time the decoder has already concealed all frames F_i with $t^i_{playout} < t^n_{arrive}$, because they have been considered as lost. Should the playout times be increased by $t_{gap} = t^n_{arrive} - t^n_{playout}$ temporarily so that the too late frames are still played out?

Because adjustments have a different impact according to the current speech property, it is important to know whether the F_i frames $(i > n)$ contain silence or voice. The voice activity of frame F_n is known, because it has already arrived. Thus, we know the speech quality impairment of the adjustment, which delays the playout.

But when to re-adjust the playout to its previous value again? Clearly as soon as the voice falls silent the playout should be changed because during silence the adjustment is not hearable. But how long will the talker speak? The speech properties of the consecutive frames are not known, because they have not been received so far.

But there is hope in statistics: The ITU has standardized an artificial voice model (P.59, 1993), which uses exponentiation distributions to describe the length of talk-spurts and silence periods. The mean lengths are $t_{\text{on}} = 1.004\,\mathbf{s}$ and $t_{\text{off}} = 1.587\,\text{s}$. Because of the exponentiation distribution of the talk-spurt length, a negative adjustment can be made in the mean after 1.004s.

To calculate the quality rating of a delay spike without an adjustment of the playout buffer time, we apply the ITU E-Model. The R factor is calculated from the speech quality measurements (Fig. 5b) and the mouth-to-ear delay:

$$R_{loss} = \text{MOS}_2\text{R}\left(\text{MOS}_{loss}\left(t_{gap}\right)\right) - I_d\left(t_{m2e}\right) \qquad (10)$$

To calculate the quality rating of the adjustment, we use the results from Fig.5b which refer to samples with $t_{sample} = 8s$. To calculate delay impairment, we sum and weighten the quality of adjusted period and the normal period. The quality impairment of the adjustment during silence is not considered because it is virtually zero:

$$R_{adjust} = \text{MOS}_2\text{R}\left(\text{MOS}_{pos}\left(t_{gap}\right) - \frac{I_d(t_{m2e})(t_{sample}-t_{on})+I_d(t_{m2e}+t_{gap})t_{on}}{t_{sample}}\right) \qquad (11)$$

In Fig. 8 we calculated $R_{adjust} - R_{loss}$ for different gap lengths and mouth-to-ear delays. It can be seen that loss bursts up to 80 ms can be tolerated in any case. Loosing frames up to 160 ms might be better for calls which have a high mouth-to-ear delay. Otherwise, the playout buffer should be adjusted to playout the too late frames.

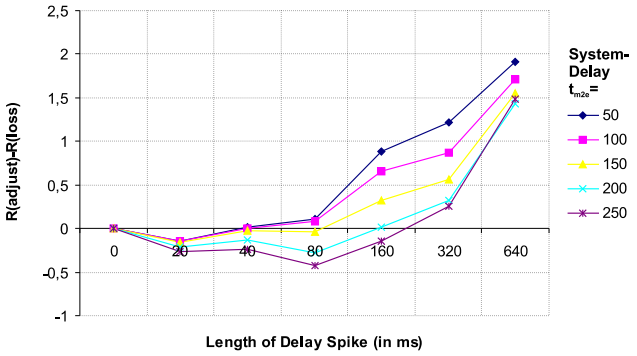


Figure 8: Whether to adjust the playout to late packets (positive values) or to drop the late packets (negative values)

7 Conclusion

We presented a new quality model for voice. Its main purpose is to parameterize adaptive VoIP applications and algorithms so that they can achieve high perceptual quality ratings.

1.) One the contributions of this paper is that the coding mode must not be switched too often because it harms the speech quality. Consequently, media-dependent FEC is not feasible. Media-dependent FEC tries to improve speech quality by switching to an other coding mode. However, switchomg to the coding mode reduces speech quality because it introduces clicking sounds.

2.) We demonstrated that as soon as bandwidth gets limited it is more efficient to change the packet rate instead of the coding rate. Pervious approaches to rate-adaptive voice only considered the coding rate.

3.) Our results indicate that a playout buffer should adjust its playout to delay spikes if they cause frames to arrive at least 80 ms after their scheduled playout time or even later.

One should consider that the measurement results of our work are based on an objective perceptual model which only approximates the real rating behavior of human beings. Thus, we have conducted subjective tests

to prove and to enhance the accuracy of our objective quality model (Hoene et al., 2004). We continue our work on quality models to include the effects of single packet losses.

References

Allnatt, J. (1983). *Transmitted-picture Assessment*. John Wiley & Sons, New York, USA.

Barberis, A., Casetti, C., Martin, J. C. D., and Meo, M. (2001). A simulation study of adaptive voice communications on IP networks. *Computer Communications*, 24(9):757–767.

Beerends, J. G., Hekstra, A. P., Rix, A. W., and Hollier, M. P. (2002). Perceptual evaluation of speech quality (PESQ), the new ITU standard for end-to-end speech quality assessment, part II - psychoacoustic model.

Bolot, J.-C. (1993). End-to-End Packet Delay and Loss Behavior in the Internet. In *ACM SIGCOMM '93*, volume 23, pages 289–298, Stanford, CA, USA.

Bolot, J.-C., Crepin, H., and Garcia, A. V. (1995). Analysis of audio packet loss in the internet. In *Network and Operating System Support for Digital Audio and Video (NOSSDAV)*, pages 154–165.

Bolot, J.-C., Fosse-Parisis, S., and Towsley, D. F. (1999). Adaptive FEC-based error control for internet telephony. In *Proceedings of IEEE Infocom*, pages 1453–1460.

Bolot, J.-C. and Vega-Garcia, A. (1996). Control mechanisms for packet audio in the internet. In *Proceedings of IEEE Infocom*, pages 232–239, San Franisco, CA, USA.

Bolot, J.-C. and Vega-Garcia, A. (1998). The case for FEC-based error control for packet audio in the internet. *ACM Multimedia Systems*.

Boutremans, C. and Boudec, J. Y. L. (2003). Adaptive joint playout buffer and FEC adjustment for internet telephony. In *Proceedings of IEEE Infocom*, San-Francisco, CA, USA.

Chuah, C.-N. and Katz, R. H. (2002). Characterizing packet audio streams from internet multimedia applications. In *Proceedings of IEEE International Conference on Communications (ICC 2002)*, volume 2, pages 1199–1203.

ETSI (2002). Universal Mobile Telecommunications System (UMTS), AMR Speech Codec, General Description. 3GPP TS 26.071 Version 5.0.0 Release 5.

Fraleigh, C., Tobagi, F., and Diot, C. (2003). Provisioning IP backbone networks to support latency sensitive traffic. In *Proceedings of IEEE Infocom*, San Francisco, CA, USA.

Hardman, V., Sasse, M. A., Handley, M., and Watson, A. (1995). Reliable audio for use over the Internet. *Proceedings of gs Internet Society's International Networking Conference (INET)*, pages 171–178.

Hoene, C. (2004). A perceptual quality model for adaptive VoIP applications: Software distribution. URL: http://www.tkn.tu-berlin.de/research/simquamol/.

Hoene, C., Rathke, B., and Wolisz, A. (2003). On the importance of a VoIP packet. In *Proceedings of ISCA Tutorial and Research Workshop on the Auditory Quality of Systems*, Mont-Cenis, Germany.

Hoene, C., Wiethoelter, S., and Wolisz, A. (2004). Predicting the perceptual service quality using a trace of voip packets. In *Proceedings of Fifth International Workshop on Quality of future Internet Services (QofIS'04)*, Barcelona, Spain. To appear.

ITU (1993). Artificial conversational speech. ITU-T Recommendation P.59.

ITU (1996). Methods for subjective determination of transmission quality. ITU-T Recommendation P.800.

ITU (1998). ITU-T coded-speech database. ITU-T Recommendation P.Supplement 23.

ITU (1999). Application of the E-model: A planning guide. ITU-T Recommendation G.108.

ITU (2000). The E-Model, a computational model for use in transmission planning. ITU-T Recommendation G.107.

ITU (2001a). Methodology for derivation of equipment impairment factors from subjective listening-only tests. ITU-T Recommendation P.833.

ITU (2001b). Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. ITU-T Recommendation P.862.

Kaj, I. and Marsh, I. (2003). Modelling the arrival process for packet audio. In *Quality of Service in Multiservice IP Networks*, pages 35–49, Milan, Italy.

Kim, H., Chae, M.-J., and Kang, I. (2002). The methods and the feasibility of frame grouping in internet telephony. *IEICE Transactions on Communications*, E85-B(1):173–182.

Laoutaris, N. and Stavrakakis, I. (2002). Intrastream synchronization for continuous media streams: A survey of playout schedulers. *IEEE Network Magazine*, 16(3).

Liang, Y. J., Färber, N., and Girod, B. (2003). Adaptive playout scheduling and loss concealment for voice communication over ip networks. *IEEE Transactions on Multimedia*, 5(4):532–543.

Liang, Y. J., Steinbach, E. G., and Girod, B. (2001). Real-time voice communication over the internet using packet path diversity. In *ACM Multimedia*, pages 431–440.

Liu, F., Kim, J., and Kuo, C.-C. J. (2001). Adaptive delay concealment for internet voice applications with packet-based time-scale modification. In *Proceedings IEEE ICASSP*.

Markopoulou, A. P., Tobagi, F. A., and Karam, M. J. (2002). Assessment of VoIP quality over internet backbones. In *Proceedings of IEEE Infocom*, New York, NY, USA.

Moon, S. B., Kurose, J., and Towsley, D. (1998). Packet audio playout delay adjustments: performance bounds and algorithms. *ACM/Springer Multimedia Systems*, 27(3):17–28.

Möller, S. (2000). *Assessment and Prediction of Speech Quality in Telecommunications* Kluwer Academic Publishers.

Perkins, C., Hodson, O., and Hardman, V. (1998). A survey of packet loss recovery techniques for streaming audio. *IEEE Network*, 12:40–48.

Perkins, C., Kouvelas, I., Hodson, O., Hardman, V., Handley, M., Bolot, J., Vega-Garcia, A., and Fosse-Parisis, S. (1997). RTP payload for redundant audio data. IETF RFC 2198.

Pinto, J. and Christensen, K. J. (1999). An algorithm for playout of packet voice based on adaptive adjustment of talkspurt silence periods. In *in Proceedings of the IEEE 24th Conference on Local Computer Networks (LCN)*, pages 224–231.

Podolsky, M., Romer, C., and McCanne, S. (1998). Simulation of FEC-based error control for packet audio on the internet. In *Proceedings of IEEE Infocom*, pages 505–515, San Francisco, CA, USA.

Ramjee, R., Kurose, J. F., Towsley, D. F., and Schulzrinne, H. (1994). Adaptive playout mechanisms for packetized audio applications in wide-area networks. In *Proceedings of IEEE Infocom*, pages 680–688, Toronto, Canada.

Rix, A. W., Hollier, M. P., Hekstra, A. P., and Beerends, J. G. (2002). Perceptual evaluation of speech quality (PESQ), the new ITU standard for end-to-end speech quality assessment, part I - time alignment. volume 50.

Rosenberg, J., Qiu, L., and Schulzrinne, H. (2000). Integrating packet FEC into adaptive voice playout buffer algorithms on the internet. In *Proceedings of IEEE Infocom*, pages 1705–1714, Tel Aviv, Israel.

Servetti, A. and Martin, J. C. D. (2003). Adaptive interactive speech transmission over 802.11 wireless LANs. In *Proceddings IEEE Int. Workshop on DSP in Mobile and Vehicular Systems*, Nagoya, Japan.

Sjoberg, J., Westerlund, M., Lakaniemi, A., and Xie, Q. (2002). Real-time transport protocol (RTP) payload format and file storage format for the adaptive multi-rate (AMR) and adaptive multi-rate wideband (AMR-WB) audio codecs. IETF RFC 3267.

Sreenan, C., Chen, J.-C., Agrawal, P., and Narendran, B. (2000). Delay reduction techniques for playout buffering. *IEEE Transactions on Multimedia*, 2(2):88–100.

Sun, L. and Ifeachor, E. C. (2003). Prediction of perceived conversational speech quality and effects of playout buffer algorithms. In *Proceedings of IEEE International Conference on Communications (ICC 2003)*, pages 1–6, Anchorage, USA.

Veeraraghavan, M., Cocker, N., and Moors, T. (2001). Support of voice services in IEEE 802.11 wireless LANs. In *Proceedings of IEEE Infocom*, pages 488–497, Los Alamitos, CA, USA.

Yin, N., Li, S.-Q., and Stern, T. E. (1990). Congestion control for packet voice by selective packet discarding. *IEEE Transactions on Communications*, 38(5):674–683.