Special issue Performance Evaluation of Wireless Networks and Communications of the Computer Communications Journal, Elsevier, May 2005

A Perceptual Quality Model Intended for Adaptive VoIP Applications[⋆]

C. Hoene, H. Karl, A. Wolisz

Technical University of Berlin, TKN, Sekr. FT 5-2
Einsteinufer 25, 10587 Berlin, Germany
Phone: +49 30 31423819 Fax: +49 30 31423818

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 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). Moreover, this model can be used at run time to control the transmission of such calls. It is also computationally efficient and open-source.

To demonstrate the potential of our model, we apply it 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 to control VoIP transmissions.

Key words: internet telephony, adaptive transmission, perceptual quality model, E-model, PESQ, packetization, delay spikes.

1 Introduction

Recent studies (Markopoulou et al., 2002) show that a significant number of Internet backbone links do not provide *toll quality* — the lowest quality of classic PSTN based telephone calls — when used for Voice over IP (VoIP)

[⋆] This work has been supported by the Deutsche Forschungsgemeinschaft (DFG). It is an enhanced and extented version of (Hoene et al., 2004).

Email address: hoene|hkarl|awo@ieee.org (C. Hoene, H. Karl, A. Wolisz).

applications. To overcome this shortcoming, application level control of the transmission of voice calls 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 telephone call measured? More precisely, how can an adaptive VoIP application assess the quality of its transmission and the predict the impact of its adaptation actions?

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); they are also evidently not applicable for on-line adaptation. On the other hand, VoIP applications measure 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 service 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 briefly summarize here 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 (2001). It compares the original and the degraded version of a speech sample to assess the speech quality and assign a mean opinion score value (MOS), which ranges from 1 (bad) to 5 (excellent).

The quality of a telephone call is not entirely captured by the speech quality alone. Further factors have to be considered. 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 is 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. The E-Model, on the other hand, considers operational parameters which are not known or not relevant to the application (see section 2.1.2). It does not consider the impairment due to dynamic adaptations. Furthermore, it assumes tandem coding (transcoding) conditions (ITU G.108, 1999) and as a result 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.
(4) 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 and 6. 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 protocol (H.323 or SIP) establish, control and terminate a telephone call. In the following, we will discuss the principal components of the of VoIP system, which cover the end-to-end transmission of voice (Fig. 1).

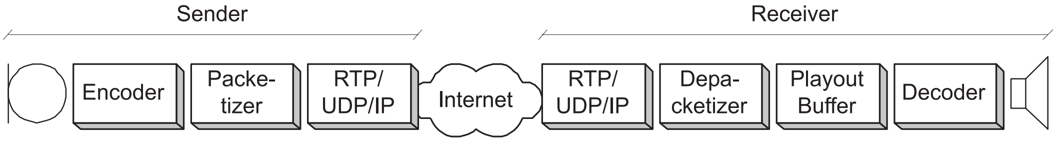


Fig. 1. VoIP system

Digitized human speech is encoded by encoding algorithms, which 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 (Hz), 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 to the 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 due to 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 dejittering buffer (also called playout scheduler) 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.

2.1.1 Application-layer adaptation

This form of adaptation can enhance the quality of VoIP because it changes the configuration of the VoIP system 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 excessive 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 an 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 the required bandwidth and the algorithmic 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. Kim et al. (2002) alternatively use 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).

2.1.2 Requirements for a quality model

This brief survey has shown a plethora of possible complications that a quality model has to deal with. If a perceptual quality model should be applied to the adaption of VoIP parameters in the application layer, 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]. Our 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, the E-Model and other approaches fulfill these requirements and whether they can be used in an adaptive VoIP application.

2.2 Perceptual Assessment

2.2.1 PESQ

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). PESQ can identify transmission delays (Rix et al., 2002). First, PESQ adjusts the degraded version to be aligned in time with the original version. Then, a psychoacoustic model assesses the distortion between original 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

[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 speaker 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.

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 can it be integrated into open-source software.

2.2.2 The E-Model

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 distinguished 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 magnitudes of physical stimuli to perceptual magnitudes (Allnatt, 1983).

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

$$R = R_o - I_s - I_d - I_e + A \tag{1}$$

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, quantization (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 affected 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 a telephone call is judged differently if the user has an advantage of access. For example, wireless, cellular, and satellite connections might be valued higher than standard tethered access. 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.

2.2.3 Related Work

Beside PESQ and the E-Model other option models have been proposed. Clark (2001) identified the effect that humans do remember the quality only for a curtain time until the former impressions are overwhelmed by recent. More precise, the impact of a distortion decays expontentially with time. He also introduced the notion of packet loss event, which we adopt in this paper. Mohammed et al. (2001) proposed a model to measure speech quality at real-time in order to control the transmission at run-time. The author uses a neural network which is trained to calculate speech quality. However, transmission delay is not considered. Takahashi et al. (2004) verified the E-Model and has to proposed an enhanced model which improves the rating performance from $R = 0.763$ to $R = 0.793$. The author identifies further potenial areas of improvement.

Jiang and Schulzrinne (2002) studied the impact of both media-dependent and -independent FEC on the speech quality for bursty losses. They conducted listening-only tests. Media-indepenent FEC shows better performance and is to be preferred. Also neither positive nor negative impact of loss burstiness is clearly given. The observations confirm the results given in this work.

Humans can judge the quality of a speech sample even without the knowledge of the original whether as PESQ requires both degraded and original. Recently, a new psychoaccoustic model called 3SQM (Schmidmer, 2004; ITU-T P.563, 2004) has been developed and standardilized which so not require the original sample. It rating performance achieved nearly the same level as PESQ.

Table 1
Properties and features of quality models

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	-

The combination of both E-Model and PESQ rating results is proposed both by Sun and Ifeachor (2004). As in this work, Sun uses stored values to correlate random packet loss rates. Instead of the E-Model an approximated linear model is applied.

3 A new quality model

Because PESQ, E-Model and other published quality models 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* sides. 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.

3.1 Source Side

In the following, only parameters being available at the source are considered. Equation 2 is based on the E-Model. If the acoustic processing is optimal, however, 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:

$$R = \text{MOStoR}\left(\text{MOS}\left(c, dtx, cr, lr, pack\right)\right) - I_d\left(t\right) \tag{2}$$

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 0s $< T_a \leq$ 0.1s, I_{dd} is 0. For any 0.1s $\leq T_a <$ 0.5s, I_{dd} is

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

with $X = \frac{-2+\lg\left(T_a 10^3\right)}{\lg 2}$. 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:

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

with

$$\begin{aligned} h = \tfrac{1}{3}\arctan 2\,(&\text{x} = 18566 - 6750m, \\ &\text{y} = 15\sqrt{-202500m^2 + 1113960m - 903522}\Big) \end{aligned}$$

In Section 4, we derive MOS (c, dtx, cr, lr, pack) 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 cannot extrapolate the acoustic signal. Second, if the playout scheduler changes the playout time, the speech may be impaired (Fig. 6). Last, switching the coding mode or coding rate can cause "clicking" sounds (Fig. 5).

Impairment events can be measured by the duration and the strength of distortion. Let us define a measure that 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 encoder 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:

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

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 $\mathrm{MOS}(\cdot)$, which maps MOS values for various operating conditions. We also introduce the notion of an impairment event. The objective of the following speech quality measurements is to determine the concrete curve and values of these functions so that the quality model developed here can use these values. To limit the length of this paper we will confine ourselves to a single codec, the Adaptive-Multi-Rate coding (AMR), which is the default codec for third generation WCDMA systems (ETSI, 2002).

4.1 Measurement Setup

We followed the recommendation ITU-T P.833 (2001), 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).

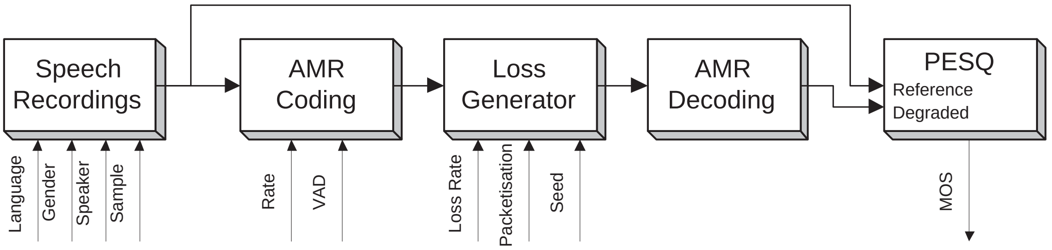


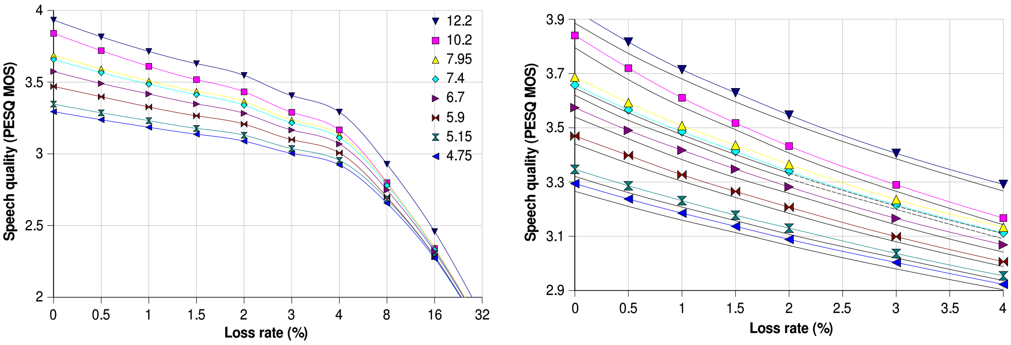
Fig. 2. Measurement Set-Up

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 20 ms of speech and can be encoded with a 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. Fig. 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. Fig. 3b displays the distortion due to silence compression, which is present but low. Fig. 4 shows that a higher packetization does not change the speech quality to any large extent.

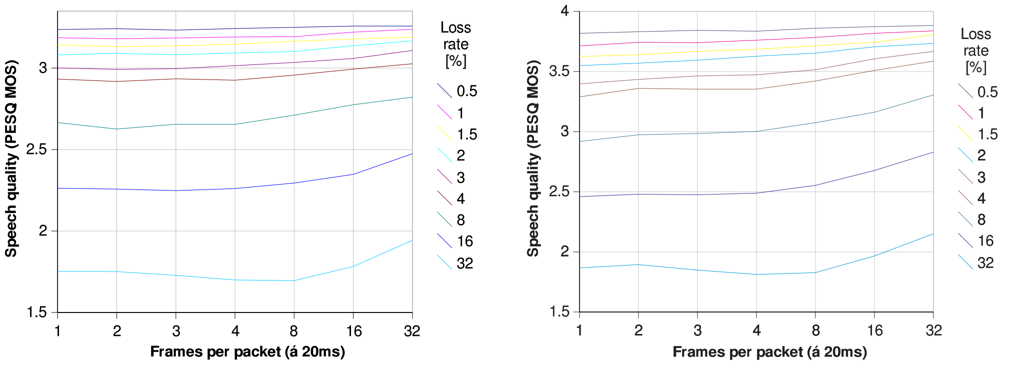
In the following, we show the distortion caused by frequent switching of the coding rate (Fig. 5) versus the mean coding rate. During the encoding of a sample, we switch the coding rate several times at different rates. For example,



(a) without silence suppression (DTX)

(b) without DTX (straight line) and – slightly lower – with DTX (dotted line)

Fig. 3. Impact of coding rate and loss rate.



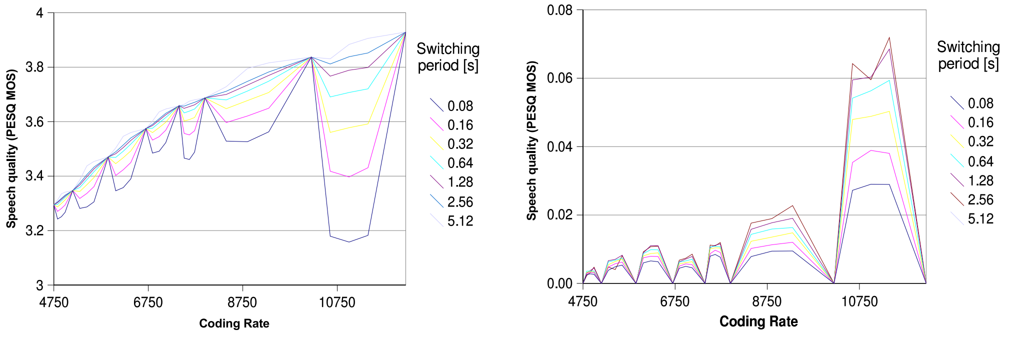
(a) AMR 4.75 kbit/s

(b) AMR 12.2 kbit/s

Fig. 4. Impact of packetization (frames per packet) vs. packet loss rate.

switching period is 80ms, the lower coding rate is selected for 20, 40, or 60ms and the higher coding rate is selected for 60, 40, or 20ms respectively. In Fig. 5a we display the resulting speech quality for a average coding rate and the impairments due to switching the coding mode. Fig. 5b 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 8s sample and distinguish between adjustments during voice activity (Fig. 6c) and silence (Fig. 6b). 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 comparison, we also measured the impairment caused by a *loss burst* that has the same length as the positive adjustment's gap (Fig. 6a). During silence, PESQ does not con-

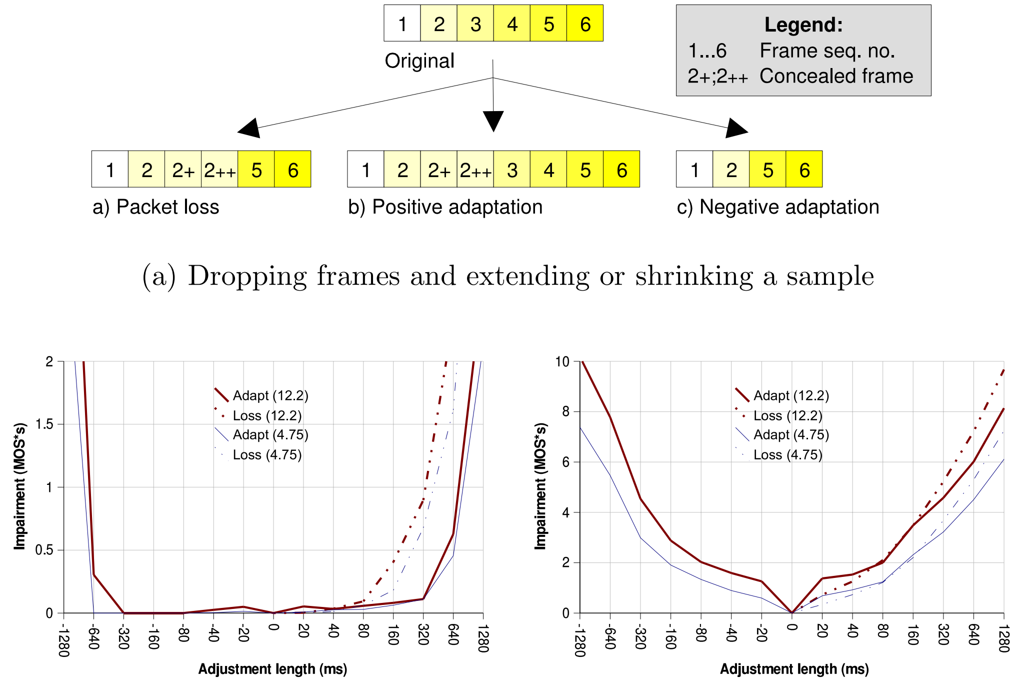


(a) PESQ MOS

(b) Impairment MOS*s

Fig. 5. Impact due to of switching the coding mode (at different frequencies).

sider adjustments to up to one second as harmful. Adjustments during voice activity decrease the speech quality and increase the impairment.



(a) Dropping frames and extending or shrinking a sample

(b) During silence

(c) During voice activity

Fig. 6. Impact of playout re-scheduling

5 First Example: Limited Bandwidth

In this example, we apply our quality model to the problem of adapting a VoIP flow to a channel with limited available bandwidth that is described by a maximal data rate. 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 allows 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.

5.1 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 the channel B_c and the bandwidth of the flow B_f, which is equal to the coding rate B_s (see Equation 6).

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

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. 7 the tradeoff in MOS between available channel bandwidth and, thus, loss rate on the one hand and coding mode on the other hand is displayed, taking into account Equation 6 and the measurement data of Fig. 3. If the loss rate exceeds a value of about 0.5 % (i.e., the available bandwidth becomes too small compared to the coding rate), 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.

5.2 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

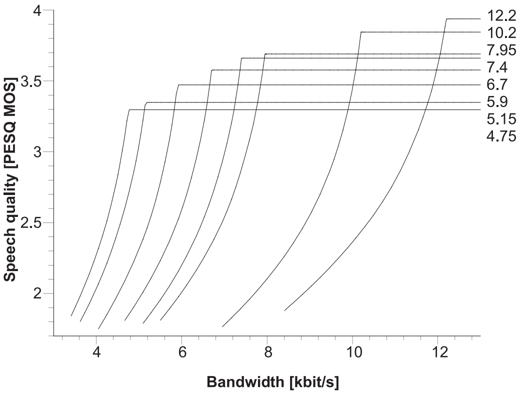


Fig. 7. MOS vs. channel bandwidth and coding rate (MR475=4.75kbps, MR122=12.2kbps).

are generated at a 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:

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

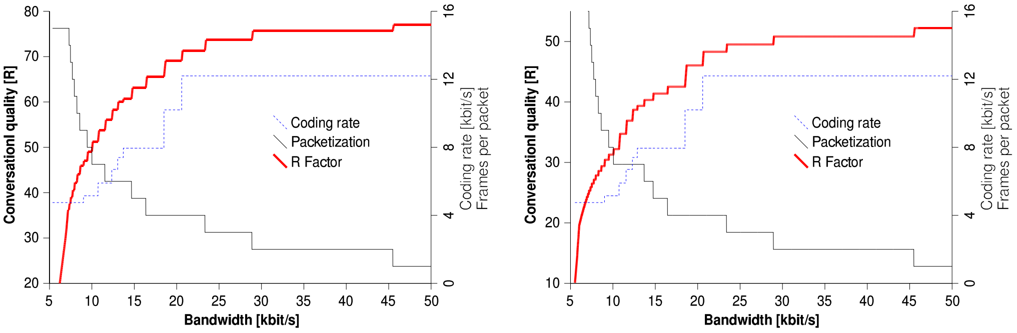
We can calculate the flow bandwidth B_f using the packet size p, the number of frames per packet f and the frame rate r.

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

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.

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

In Fig. 8, 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 and 400 ms system delay. The figures show that the packetization increases if the available bandwidth drops. Only at a very low bandwidth the coding rate decreases too. In the figure, we do not plot the packet loss rate because it is zero nearly all the time.



(a) System delay: 150ms

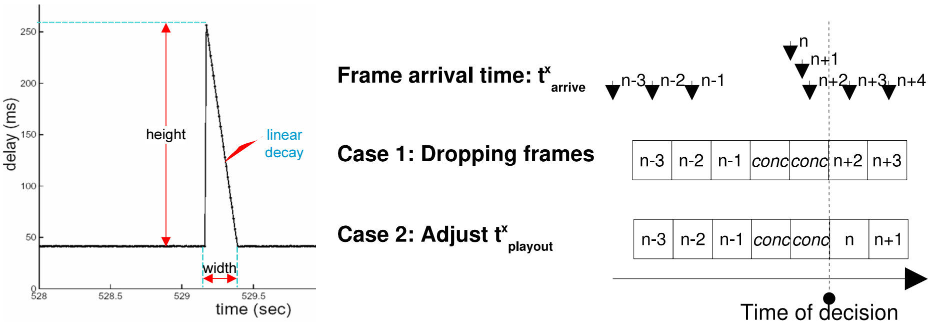
(b) System delay: 400ms

Fig. 8. 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 consider a open issue in the design of adaptive playout algorithms. The size of a playout buffer should be chosen in a way that both the number of too late frames and the additional delay imposed by the buffer 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 shows that sometimes packet delays show a sharp, spike-like increase (Ramjee et al., 1994; Markopoulou, 2002) which cannot be predicted in advance. After a spike, packets are received at a high frequency. Soon afterwards, the jitter process returns to normal (Fig. 9a). We like to consider the question whether it is advantageous to delay the playout of speech frame to included such spikes or to drop any too late packets (Fig. 6b). We concentrate on the non-trivial case of delay spikes during voice activity using the quality model introduced in this paper.

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



(a) Increased transmission delay in IP backbones

(b) Drop frames or delay the playout?

Fig. 9. Playout scheduling in cases of delay spikes

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: Brady discovered that both talkspurt and silence periods of digitised voice can be approximated by an exponential distribution (Brady, 1969). A commonly accepted and standardilzed ITU-T P.59 (1993) model for a artificial voices is a continuous-time, discrete state Markov chain with two states refering to talk spurt (ON) and silence (OFF) periods. The holding time in each state is exponentially distributed with mean $1/\lambda$ and $1/\mu$. Hence, the transitional rates from the ON to OFF state is λ and μ, respectively. We apply this model to predict when a negative adjustment can be made. (For simplicity reasons, we assume in the following that the next silence frames will be in exactly 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 (MOS$_{loss}$ from Fig. 6c) and the mouth-

State	Mean Duration
talk-spurt	$t_{\text{on}} = \mu^{-1} = 1.004\text{s}$
silence	$t_{\text{off}} = \lambda^{-1} = 1.587\text{s}$

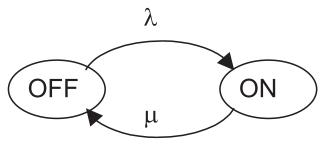


Fig. 10. Voice model: two-state markov model

to-ear delay (t_{m2e} is usually estimated or measured):

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

To calculate the quality rating of the adjustment, we use the results MOS_{adapt} from Fig. 6c which refer to samples with a length of $t_{sample} = 8\,s$. To calculate delay impairment, we sum and weight 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:

$$R_{adjust} = \text{MOStoR}\left(\text{MOS}_{adapt}\left(t_{gap}\right) - I_d^{mean}\right) \tag{11}$$

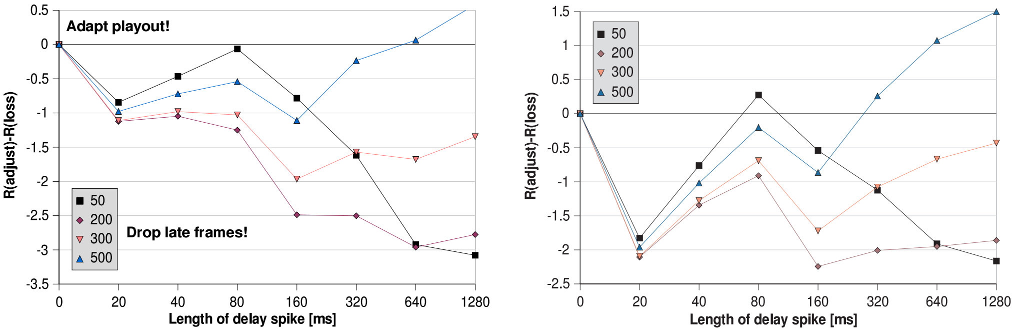
with

$$I_d^{mean} = I_d\left(t_{m2e}\right) \frac{t_{sample} - t_{on}}{t_{sample}} + I_d\left(t_{m2e} + t_{gap}\right) \frac{t_{on}}{t_{sample}}$$

In Fig. 6c we have shown that the rescheduling of speech frames harms the speech quality less than losing the frames as long as the gap is larger than 80ms. In Fig. 11 we also consider the impact of transmission delay and calculated $R_{adjust} - R_{loss}$ for different gap lengths and mouth-to-ear delays. It can be seen that delay spikes in general should be dropped and an adaptation is not required. However, one should consider that is fact only account for single delay spikes. Often a delay spike is only a first indication for a up coming period of further high transmission delays (Markopoulou, 2002). Further sudies are required to identify the impact of multiple delay spikes that occur shortly one after the other.

7 Conclusion

If a quality model is being developed, the area and context of its application is highly importance and has a large impact on design decisions: 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.



(a) AMR 4.75kbps

(b) AMR 12.2kpbs

Fig. 11. Whether to adjust the playout to late packets or to drop late packets

One of the conclusions 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 another coding mode. However, switching to the coding mode reduces speech quality because it introduces clicking sounds.

We demonstrated that as soon as bandwidth gets limited it is more efficient to change the packet rate instead of the coding rate. Previous approaches to rate-adaptive voice only considered the coding rate. Also our results also indicate that a playout buffer should not adjust its playout to delay spikes if they occur singular.

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, subjective tests to verify and to enhance the accuracy of these results are required. We are continuing our work on quality models to include the effects of single packet losses (Hoene et al., 2005).

References

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

Barberis, A., Casetti, C., Martin, J. C. D., Meo, M., May 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., Hollier, M. P., Jun. 2002. Perceptual evaluation of speech quality (PESQ), the new ITU standard for end-to-end speech quality assessment, part II - psychoacoustic model. Journal of the Audio Engineering Society 50.

Bolot, J.-C., 1993. End-to-end packet delay and loss behavior in the internet.

In: Conference proceedings on Communications architectures, protocols and applications. ACM Press, pp. 289–298.

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

Bolot, J.-C., Fosse-Parisis, S., Towsley, D. F., Apr. 1999. Adaptive FEC-based error control for internet telephony. In: Proceedings of IEEE Infocom. New York, NY, USA, pp. 1453–1460.

Bolot, J.-C., Vega-Garcia, A., Apr. 1996. Control mechanisms for packet audio in the internet. In: Proceedings of IEEE Infocom. San Franisco, CA, USA, pp. 232–239.

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

Boutremans, C., Boudec, J.-Y. L., Apr. 2003. Adaptive joint playout buffer and FEC adjustment for internet telephony. In: Proceedings of IEEE Infocom. Vol. 1. San-Francisco, CA, USA, pp. 652–662.

Brady, P. T., Sep. 1969. A model for generating on-off speech patterns in two-way conversation. The Bell System Technical Journal 48 (9), 2445–2472.

Chuah, C.-N., Katz, R. H., Apr. 2002. Characterizing packet audio streams from internet multimedia applications. In: Proceedings of IEEE International Conference on Communications (ICC 2002). Vol. 2. College Park, MD, USA, pp. 1199–1203.

Clark, A., Mar. 2001. Modeling the effects of burst packet loss and recency on subjective voice quality. In: Internet Telephony Workshop. New York, NY, USA, pp. 123–127.

ETSI, Jun. 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., Diot, C., Apr. 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., Watson, A., Jun. 1995. Reliable audio for use over the Internet. Proceedings of gs Internet Society's International Networking Conference (INET), 171–178.

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

Hoene, C., Karl, H., Wolisz, A., Jul. 2004. A perceptual quality model for adaptive VoIP applications. In: Proceedings of International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS'04). San Jose, California, USA.

Hoene, C., Rathke, B., Wolisz, A., Apr. 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., Schäfer, G., Wolisz, A., 2005. Predicting the importance of a speech frame, under submission.

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

ITU-T G.108, Sep. 1999. Application of the E-model: A planning guide.

ITU-T P.563, Mar. 2004. P.SEAM. Draft.

ITU-T P.59, Mai 1993. Artificial conversational speech.

ITU-T P.800, Aug. 1996. Methods for subjective determination of transmission quality.

ITU-T P.833, Feb. 2001. Methodology for derivation of equipment impairment factors from subjective listening-only tests.

ITU-T P.862, Feb. 2001. 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 P.Supplement 23, Feb. 1998. ITU-T coded-speech database.

Jiang, W., Schulzrinne, H., 2002. Comparison and optimization of packet loss repair methods on voip perceived quality under bursty loss. In: NOSSDAV. pp. 73–81.

Kaj, I., Marsh, I., Feb. 2003. Modelling the arrival process for packet audio. In: Quality of Service in Multiservice IP Networks. Milan, Italy, pp. 35–49.

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

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

Liang, Y. J., Färber, N., Girod, B., Dec. 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., Girod, B., 2001. Real-time voice communication over the internet using packet path diversity. In: ACM Multimedia. pp. 431–440.

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

Markopoulou, A. P., Oct 2002. Assessing the quality of multimedia communications over internet backbones. Ph.D. thesis, Stanford University, USA.

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

Mohammed, S., Cercantes-Perez, F., Afifi, H., Apr. 2001. Integrating networks measurements and speech quality sujective scroes for control purposes. In: Proceedings of Infocom 2001. Vol. 2. Anchorage, AK, USA, pp. 641–649.

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

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

Perkins, C., Hodson, O., Hardman, V., Sep. 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., Fosse-Parisis, S., Sep. 1997. RTP payload for redundant audio data. IETF RFC 2198.

Pinto, J., Christensen, K. J., Oct. 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). Lowell, MA, USA, pp. 224–231.

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

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

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

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

Schmidmer, C., Mar. 2004. 3SQM - a breakthrough in single ended voice quality testing. In: Proceedings of the 30th German Convention on Acoustics (DAGA). Strasbourg, France.

Servetti, A., Martin, J. C. D., Apr. 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., Xie, Q., Jun. 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., Narendran, B., Jun. 2000. Delay reduction techniques for playout buffering. IEEE Transactions on Multimedia 2 (2), 88–100.

Sun, L., Ifeachor, E., Jun. 2004. New models for perceived voice quality prediction and their applications in playout buffer optimization for VoIP networks. In: Proceedings of IEEE International Conference on Communications (ICC 2004). Paris, France, pp. 1478 – 1483.

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

Takahashi, A., Yoshino, H., Kitawaki, N., Jun. 2004. Perceptual QoS assessment technologies for VoIP. IEEE Communications Magazine, 28–34.

Veeraraghavan, M., Cocker, N., Moors, T., Apr. 2001. Support of voice services in IEEE 802.11 wireless LANs. In: Proceedings of IEEE Infocom. Los Alamitos, CA, USA, pp. 488–497.

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