Impact of network performance parameters on the end-to-end perceived speech quality

L.A.R. Yamamoto, J.G. Beerends

KPN Research, The Netherlands

Abstract

The evolution of computers and networks leaves no doubt about the important role of interactive multimedia services. The quality of these new services will be a key issue for their wide deployment, and this quality is determined by the opinion of the users. The best Quality of Service (QoS) is not the highest, but the most suitable to the different users' needs. In order to provide a suitable level of QoS, an application needs to know which relevant network parameters have impact on the quality as it is perceived by the users.

This paper presents a top-down approach to study the influence of network performance parameters on the user perceived quality, taking conversational speech as an example of a simple, essential and demanding application. Three types of network are considered: IP, ATM and IP over ATM. For each of these cases, the influence of the corresponding network performance parameters is studied.

1 Introduction

From all the multimedia components, audio is one of the most demanding in terms of QoS requirements, since distortions in the audio signal can be very annoying. In particular, speech conversations are notably critical due to the strict requirements on delay bounds.

The QoS concept defined in [5] includes many different aspects like end-to-end quality, billing, accessibility, security, etc. In this paper, only the influence of network performance parameters on the end-to-end quality of conversational speech will be treated. Many other parameters that influence the quality, for instance the design and performance of terminal equipment and application software, the surrounding environment, etc., have not been taken into account.

The main network performance parameters that affect the perceived quality are delay, delay variation (jitter) and data loss. In this paper, a top-down approach is used to study the influence of these parameters in the case of IP, ATM, and IP over ATM. This approach is depicted in Figure 1.

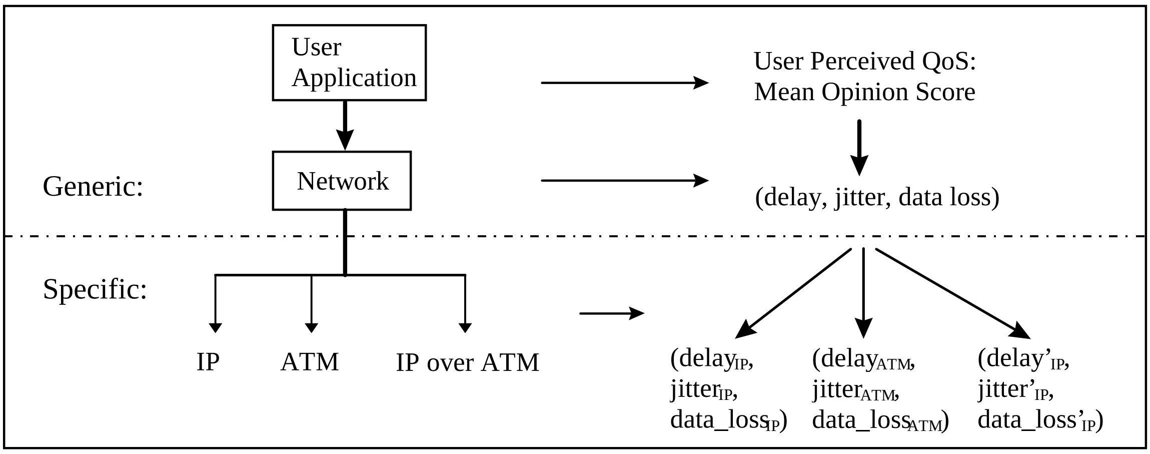

User Application
Network
Generic:
Specific:
IP
ATM
IP over ATM
User Perceived QoS: Mean Opinion Score
(delay, jitter, data loss)
(delay$_{IP}$, jitter$_{IP}$, data_loss$_{IP}$)
(delay$_{ATM}$, jitter$_{ATM}$, data_loss$_{ATM}$)
(delay'$_{IP}$, jitter'$_{IP}$, data_loss'$_{IP}$)

Figure 1: The approach used in this paper: First, the quality of speech is measured as a function of each generic parameter delay, jitter and data loss. Then, for each specific network (IP, ATM and IP over ATM), the translation from generic parameters to technology specific parameters is studied.

Although only a specific service is studied, the approach used here can in principle be applied to any other applications. The central ideas of this approach are:

- Given a certain level of QoS for a given application, measured at the user level, and relevant characteristics of the application (e.g. codec, etc.) determine which specific parameters the application should be able to adjust in order to achieve the specified QoS level.
- Given a specific perceived level of degradation in the QoS, what can be inferred about problems in the network.
- Given a certain amount of problems in the network (by measuring available performance parameters at the network level) what can be inferred about the QoS that can be obtained.

This paper aims at describing the QoS at the user level as a function of network parameters, from the point of view of an application making use of the network. It does not attempt to configure nor to dimension the network to meet specific application requirements.

In Section 2, the perceived quality of conversational speech as a function of each generic parameter is established. In Section 3, for each specific network (IP, ATM and IP over ATM), the translation from generic parameters to technology specific parameters is studied, which enables us to establish a direct relation between the QoS measurements at the user level and the specific parameters in each case.

2 User perceived quality

This section focuses on the perceived QoS of conversational speech as a function of each generic network performance parameters considered. Section 2.1 introduces a perceived QoS measurement method. Section 2.2 discusses the impact of each generic parameter - delay, jitter and data loss - on the measured QoS. Since the influence of data loss was not completely established, Section 2.3 shows an experiment to measure the QoS of speech as a function of data loss for various lengths of missing voice segments.

2.1 Perceived QoS measurement method

The end-to-end quality of conversational speech services, as perceived by the user, consists of three different quality aspects, the listening quality (how does the other party sound like), talker quality (how do I perceive my own voice), and conversational quality (how can we interact). All three can be assessed in subjective experiments where subjects judge each of the three aspects. A common scale used in these experiments is the Absolute Category Rating (ACR) five point opinion scale [8] given as: Excellent (5), Good (4), Fair (3), Poor (2), Bad (1). Averaging over all subjects in the experiment results in a score indicated as the Mean Opinion Score (MOS). It is important to stress that MOS values must be always interpreted with care, since they tend to vary between test occasions, countries, experimental task, etc.

2.2 Performance parameters

To be able to offer new services with an overall good quality, it is interesting that at least the same quality as the conventional services can be maintained in the new services. In case of the conversational speech service considered, the minimum level of quality desirable is equivalent to standard telephony quality (PSTN). For PSTN quality telephony the three components (listening quality, talker quality and conversational quality) usually have a MOS of around 4.0.

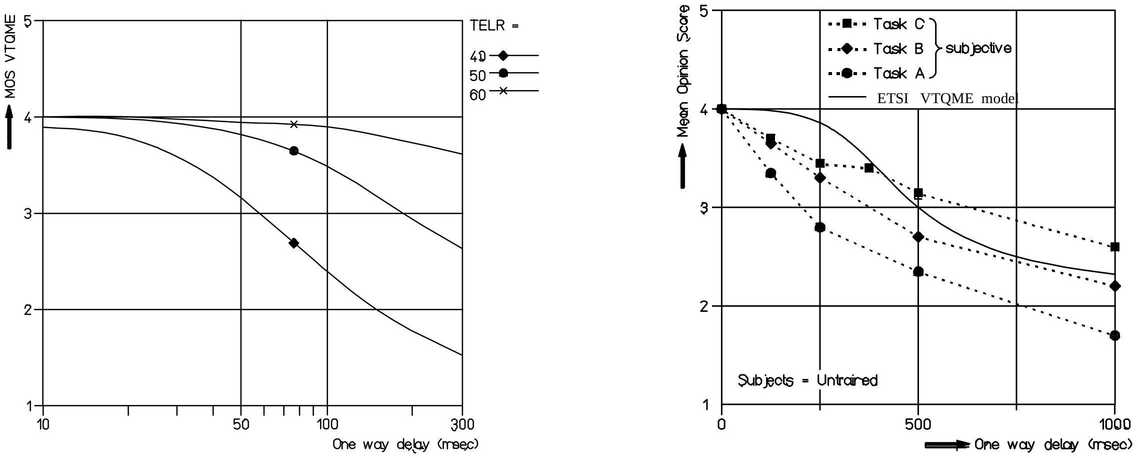
The QoS of telephone speech communication as a function of various distortions parameters is presented in the ETSI VTQME model [4]. For packet networks some of the distortions mentioned there are still relevant, e.g. echo, and the model can be used to predict the subjective quality. Other types of distortion are not included in that model, e.g. time clipping due to data loss (see Section 2.2.3) and time warping which compensates for delay variation (see Section 2.2.2).

The approach used in this paper tries to imitate the ETSI model approach in the sense that once the QoS at the user level is obtained as a function of relevant network performance parameters, it can be calculated by simply measuring these parameters, without further subjective experiments.

From all possible parameters that influence the three components of the quality, only a small set have a direct relation to network parameters. The main telephony quality parameters affected by network parameters are: speech delay, delay variation, and loss of speech signal (time clipping). The main related network parameters are, respectively: network delay, jitter, and data loss.

2.2.1 Delay

The influence of speech delay is two-fold. First it has an impact on the conversational quality in the sense of a decrease in interactivity level when the delay is increased. A one way end-to-end delay of up to 150 ms leads in general only to minor impairments in the conversational quality while delays up to 400 ms are still acceptable (G.114 [6]). Concerning the influence on the perceived disturbance of (talker) echoes the delay requirements are much stricter, end-to-end delays above 30 ms require echo canceling in order to maintain the quality. This is depicted in Figure 2.



(a) *Predicted values from the ETSI model for the talker MOS as a function of one way delay and Talker Echo Loudness Rating (TELR) [4]. A TELR of 40 dB is found in a four-wire digital connection using a normal telephone handset. Higher values can only be obtained using echo canceling*

(b) *Conversational MOS as a function of one way delay (without echo) for tasks containing different levels of interactivity (A=high, B=medium, C=low, [11]). The predicted values from the ETSI model [4] are also given.*

Figure 2: The influence of the delay in the MOS for a telephone connection.

Figure 2a shows how delay affects the talker MOS due to echo. Figure 2b shows how delay affects the conversational quality of a speech link depending on the amount of interaction between the participants. Task A is highly interactive, while task B presents medium interactivity and task C low interactivity. The curve that is estimated with the ETSI model is also shown.

2.2.2 Jitter

The influence of jitter is not perceived by the users when end systems use constant size buffers to compensate for the total delay. When the jitter is high some applications use adaptive (variable-size) buffers as jitter compensation mechanism, leading to a time warping effect, in which some parts of the original signal present different duration in the distorted signal, i.e., the time scale at the receiver is warped with respect to the original signal. Most of these compensations are carried out during silent periods only, thus the effect of time warping is generally inaudible. If the jitter is too high it also contributes to the data loss ratio, since packets that arrive later than the playback time will be discarded.

2.2.3 Data loss

Packet loss can lead to extra delay if additional mechanisms to recover from losses are implemented. Without any recovery mechanisms, data loss leads to a time clipping effect, in which fragments of the original signal are missing in the distorted signal. Since the redundancy present in the speech signal is very high, speech is very tolerant to time clipping. Hence it can tolerate a relatively small level of data loss in exchange for faster delivery. Indeed, a number of studies on this topic can be found in literature, mainly for ATM networks [12]. For IP networks a subjective experiment will be described in next section, where the influence of packet loss on the perceived quality of speech is studied.

2.3 User experiment: impact of data loss on the perceived QoS

The effect of speech delay and echo on the perceived quality is well known in literature, as well as the effect of ATM cell loss. The same can not be said about the effect of IP packet loss on the perceived one way speech quality in terms of time clipping, for different and potentially large packet sizes. Therefore, a subjective experiment was carried out. Although it was mainly focused on IP, the results can be applied to the three cases: IP, ATM, and IP over ATM.

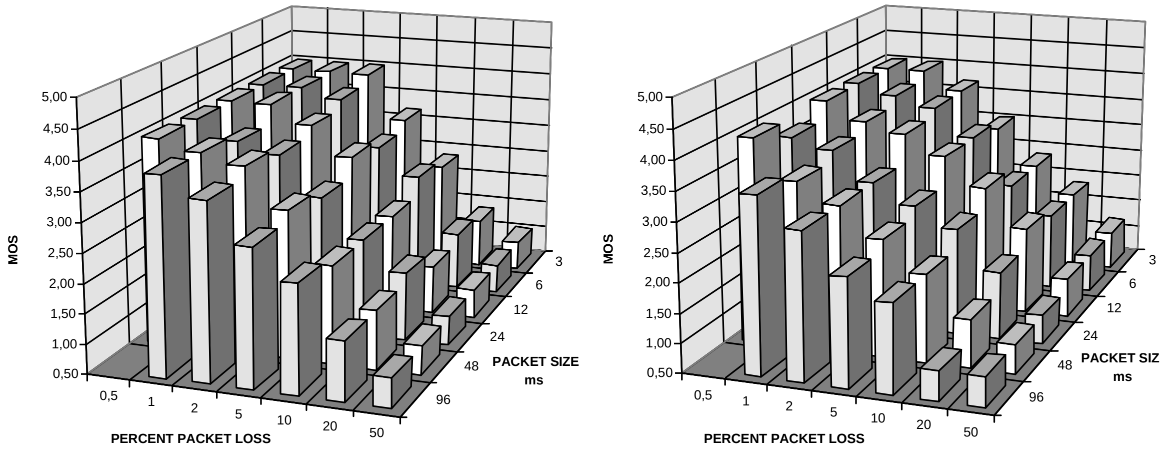
2.3.1 Experimental set up

The subjective listening experiment was performed in a quiet room (105 m^3, noise level below 25 dBA) with a reverberation time less than 500 ms for frequencies above 150 Hz. A standard telephone handset [7] was used at a normal loudness level (78 dBSPL equivalent). The speech material that was evaluated was recorded using the standard handset in the same quiet room. Speech was sampled at 8 kHz (narrow band) 16 bits linear PCM format using two voices, one male and one female, reading a sample text of about 30 seconds long. The duration was chosen as a compromise between using long files for statistical reliability (at least one packet loss must occur) and using short files to keep the experimental burden as low as possible. The speech material did not contain any long pauses (more than one second) in order to prevent, as much as possible, packet losses during silent intervals (resulting in inaudible losses).

The impairments were introduced by a packet loss simulation program, and consisted of regular losses from 0.5% to 50% of packets corresponding to speech segments between 3 and 96 ms. The lost voice segments were replaced by silent periods. The procedure for the subjective experiment was based on ITU-T recommendation P.800 [8]. In order to keep the effort minimal with maximum reliability a training procedure was established using reference degraded files. These files were obtained by degrading the two speech files with five different levels of noise using the ITU-T standard modulated noise reference unit [9] noise (Signal to Noise Ratios of 0, 5, 10, 15 and 25 dB). Subjects heard the five reference files and were informed of their MOS in the ACR scale. A total of 12 subjects were used in the assessment and each subject was presented with the degraded files in different random order.

2.3.2 Experimental results

The results for the male voice, given in Figure 3a, showed no significant effect of packet size on the perceived quality for a given packet loss ratio, while for the female voice (Figure 3b) a small effect of packet size was found. The difference was most probably caused by coincidental place of packet losses.



(a) *Results for male speech data. Packet size has no significant influence on perceived quality. For a given packet loss the maximum difference over all packet sizes is less than 0.5 MOS.*

(b) *Results for female speech data. Packet size only has a small, but relatively significant, influence on the perceived quality. For a given packet loss the maximum difference over all packet sizes is about 1.0 MOS.*

Figure 3: The subjective one way speech quality (in Mean Opinion Score) as a function of packet size and packet loss. The results show that packet loss dominates the subjective quality in both cases. The 95% confidence intervals are about ±0.4 MOS for all 35 data points in each case.

To model the influence of both packet size and packet loss a two-dimensional optimal linear regression was carried out on the male and female data separately. The fit was carried out in the logarithmic domain and resulted in two different regression lines that again showed a small but significant difference between male and female results. The final mapping used in this paper, derived by averaging the results, is given by:

$$\text{Predicted MOS} = 4.0 - 0.7\, ln(loss) - 0.1\, ln(size),$$

where *loss* is the data loss ratio in percentage (from 0.5% to 50%), and *size* is the packet size (actually the duration of the missing segment) in milliseconds (from 3 to 96 ms). Values larger than 4.0 and smaller than 1.0 are clipped to 4.0 (PSTN quality) and 1.0 respectively. The mapping of the *size* parameter onto the actual packet size in bytes depends on the bit rate used and will be treated in Section 3.

2.3.3 Conclusions from the subjective experiment

The subjective results show that the packet loss parameter dominates the subjective quality and that packet size has only little influence. The most important conclusion is that for packet sizes of less than 10 ms losses up to 1% still give near PSTN speech quality. This is in line with values found in literature for ATM cell loss experiments [12]. For larger packet sizes the degradation is dependent on the coincidental place of packet losses with respect to the original signal, but losses up to 1% give acceptable speech quality for all packet sizes up to 100 ms. It is interesting to notice that even in the presence of time clipping due to losses as high as 50%, the speech is in most of the cases still intelligible, in spite of the bad quality.

The results are valid for narrow band IRS filtered [7] speech (300-3400 Hz) at any bit rate as long as reasonable speech coding quality is used. Informal listening showed that when no IRS filtering is applied the disturbances become slightly larger. Wide band speech (50-7000 Hz) is expected to be significantly more critical.

Furthermore all results in the experiments are based on silence padding. If lost packets are not replaced by silent periods loud clicks may result, leading to severely more disturbing impairments. On the other hand if on the application level a smart speech interpolation scheme is used to replace the missing packets, degradations will become less disturbing. For the best possible interpolation, using pitch or noise replicated waveform substitution [3] [15], losses up to 10% may give acceptable speech quality.

3 Network Performance Mapping

In this section, the knowledge on the influence of each generic performance parameter as discussed in Section 2 is applied to each specific network (IP, ATM and IP over ATM). First of all, an overview of the QoS support offered in each case, together with the relevant specific performance parameters, is presented. Afterwards, the translation from generic parameters to technology specific parameters is studied, and direct relations between the QoS measurements at the user level and the specific parameters in each specific case are shown for the example speech service.

In order to map voice segments to IP packets or ATM cells, G.727 ADPCM at a data rate of 32 kbit/s will be taken as a default audio codec. This codec gives near PSTN quality.

3.1 IP

The new IP Integrated Services (IIS) model currently includes two QoS control services, in addition to the current default best effort service: Guaranteed Quality of Service (GS) and Controlled-Load Network Element Service (CLS).

When using GS [13] an application can reserve an amount of bandwidth and calculate a firm upper bound on the end-to-end delay, given its own traffic characteristics and some network parameters. Applications are not allowed to specify a desired bound on jitter, and the target loss ratio is no loss at all due to buffer overflow.

CLS [16] emulates the QoS achieved under best effort when the network elements on the path are not overloaded. An application requesting CLS specifies an estimation of its traffic, but does not quantitatively reserve resources. No quantitative delay or bandwidth guarantees are given.

Since high delays degrade the MOS of the speech application considered, delay guarantees are necessary if a MOS close to 4.0 is set as a target. Therefore, in the remaining of this section GS will be assumed for the mapping. Additionally, it is assumed that the voice data is transported using RTP/UDP/IPv6.

An application using GS needs to specify two items: TSpec (Traffic Specification) and Rspec. TSpec describes the traffic that the application expects to generate (or that the receiver can afford). RSpec (Service Request Specification) specifies the resources that the application requests from the network.

TSpec is specified in terms of a Token bucket TSpec parameter which contains the following fields: token rate (*r*), bucket depth (*b*), peak rate (*p*), minimum policed unit (*m*) and maximum datagram size (*M*). Tokens, produced at rate *r* (*r*>0) are stored in a bucket of capacity *b* (*b*>0), give the application the right to transmit a number of bytes at maximum rate $p \geq r$. The values of *b*, *m* and *M* are given in bytes while *r* and *p* are given in bytes per second. The application traffic is policed according to its TSpec. During policing actions all datagrams smaller than *m* will be counted as being of size *m*. Non-conforming traffic should be treated as best-effort according to the GS specification, rather than simply being discarded.

RSpec comprises two fields: a requested rate *R* (in bytes per second) and a slack term *S* (in microseconds), where $R \geq r$ and $S \geq 0$ (no condition is imposed between *R* and *p*). *S* expresses the difference between the maximum delay that the application can tolerate and the upper bound on the delay calculated.

3.1.1 Delay

According to the GS draft available at the moment of this writing [13] the end-to-end delay bound is given by:

$$d_{\max GS} = \frac{(b-M)}{R} \cdot \frac{(p-R)}{(p-r)} + \frac{M + C_{tot}}{R} + D_{tot} \qquad \text{, if } r \leq R < p \qquad (1)$$

$$d_{\max GS} = \frac{M + C_{tot}}{R} + D_{tot} \qquad \text{, if } r \leq p \leq R \qquad (2)$$

where:

> *b, r, p, M, R* have been defined above, and can be specified by the application;
>
> *Ctot* (in bytes) and *Dtot* (in microseconds) are accumulated error terms that contribute to increase the delays from all the nodes along the path. They are advertised by the network and are highly dependent on the implementation of each node.

Since *Ctot* and *Dtot* are calculated as a sum of the worst cases of every hop, the resulting dmaxGS can be largely overestimated (see [14][13]), and it might even happen that it is too high to be acceptable. Applications will probably still need to perform considerable adaptive buffering at the receivers in order to adjust the playback time to the worst case.

A more accurate method to calculate the end-to-end delay based on statistical approaches is presented in [14]. It is shown there that the delay can be kept below 100 ms in the case of 32 kbit/s ADPCM using GS, if the packet size is kept in the range of approximately 130 to 240 bytes, assuming a long distance connection over 30 hops and 10.000 km, 34 or 155 Mbits/s links loaded from 30% to 90% with traffic of the same type. This statistical method can not be applied to calculate the end-to-end delay in a real-time, hop-by-hop basis, but work in this area is expected to progress as the new Internet model is refined.

Assuming that the delay upper bound is acceptable, the application can control the parameters in TSpec and RSpec in order to achieve the desired service. For example, it could reduce *b* or *M*, or choose *R*>*p* to try to reduce the delay; choosing *R*<*p* could provide for a probably cheaper service (e.g. if charging is based on R) with somewhat higher delay but with the possibility of occasionally inserting some bursts at a rate *p* greater than the reserved rate. For the speech application considered, assuming a constant rate service (no silence suppression), a reasonable example of GS reservation is:

- *M* = *m* ≅ 100 to 400 bytes
- *R* = *r* = *M*drate* / (*M* - *hsize*) ≅ 8 down to 5 kbytes/second, as r is specified in bytes
- *b* = 2**M* ≅ 200 to 800 bytes
- $p \geq R$
- $S = 0$

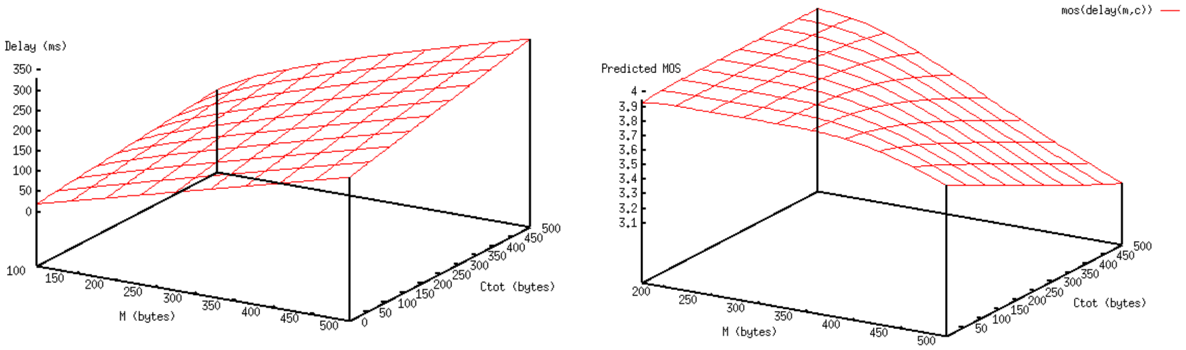
where:

M, m, R, r, b, P and *S* have been described above.

hsize is the size of the IP headers including RTP/UDP/IPv6 (no extra application specific headers are included in the calculations). In our case *hsize* = 60 bytes. The payload size is given by *M - hsize*.

drate is the data rate in which speech information is produced. Assuming 32 kbit/s ADPCM, drate = 32 kbit/s = 4 kbytes/s.

Since *m*, *R*, *r* and *b* are taken as a function of *M* in the example reservation above, and *p* has no influence on d_{maxGS} when *R*=*r*, it is possible to write d_{maxGS} as a function of *M* and the error terms. Assuming that echo canceling is available when necessary, the curve given in Figure 2b (Section 2.2.1) for the conversational MOS as a function of the delay according to the ETSI model can be applied to obtain the MOS as a function of *M* and the error terms, and thus study the impact of the packet size on the quality (the exact mathematical formulation on how to obtain the curve using the ETSI model is outside the scope of this paper). The result is plotted in Figure 4.



(a) *The end-to-end delay (minus Dtot) plotted as a function of M and Ctot (M can be controlled by the application while Ctot is given by the network).*

(b) *The predicted MOS plotted as a function of M and Ctot, following the reasoning:*
MOS = f(delay) using the ETSI model
delay = f'(M, Ctot) ⇒MOS = f''(M, Ctot)

Figure 4: Example of influence of GS parameters in the MOS.

Figure 4 shows the impact of GS parameters on the MOS for the given speech application. *M* can be chosen by the application while C_{tot} and D_{tot} can not. Since D_{tot} only shifts the curve vertically it is not plotted. The reason to include Ctot in the curve is that, as discussed earlier, if it is overestimated the resulting delay will also be overestimated and this will have a negative impact on the quality, out of the control of the applications.

From the figures above it can be seen that the quality degrades faster with the packet size if the network adds relatively large extra delay. In general, packets up to 500 bytes still give acceptable quality (MOS around 3.0), while packets around 200 bytes or smaller give near PSTN quality (MOS around 4.0).

3.1.2 Jitter

GS does not attempt to control the delay variation. The applications using GS can estimate the upper bound on delay variation by calculating d_{maxGS} as discussed in the previous section and then subtracting the latency of the path (the latter is normally advertised separately from GS, but can also be added to D_{tot} and advertised as zero). Since d_{maxGS} is likely to be overestimated, as pointed out in the previous section, the resulting delay variation is also overestimated, thus compensation mechanisms such as adaptive buffering at the receivers will still be necessary.

3.1.3 Data loss

At the moment the IIS model does not contemplate the possibility for the applications to specify a data loss ratio, but new service types including loss ratio might appear as the IIS model evolves, as pointed out in [1]. In any case, in order to predict the MOS of the speech application considered given the IP packet size (*M*), packet loss ratio (*loss*), and the coding data rate (*drate*), according to the results of the experiment described in Section 2.3 the following equation can be applied:

$$MOS = 4.0 - 0.7\ln(loss) - 0.1\ln((M\text{-}hsize) / drate)$$

Applying the equation above to the packet size ranges considered in Section 3.1.1 leads to an acceptable loss ratio in the range of 0.5% to 0.7% for a MOS close to 4.0. For normal operation of the network the loss ratio will be much less than this maximum acceptable, since GS targets at no loss due to buffer overflow, and loss due to link errors is expected to be low enough (of the order of 10^{-6} in the worst case), which leads to negligible impact on the MOS. In practice, however, temporary problems can affect normal operation introducing extra degradation, and the above equation can be used to understand the effect of those problems on the quality.

3.2 ATM

Since delay guarantees are necessary for the speech application considered, the DBR (Deterministic Bit Rate) capability combined with stringent QoS class is the most natural choice.

The traffic contract parameters for DBR are the PCR (Peak Cell Rate) and CDVT (Cell Delay Variation Tolerance). The applications can specify the PCR desired, while CDVT is determined by the network and is in general small. The PCR requested needs to be at least 86 cells/s for the data rate of the speech application considered (32 kbit/s), and cells must enter the network equally spaced, in order to comply with the traffic contract.

3.2.1 Delay and jitter

For ATM the mapping is straightforward: the generic delay parameter corresponds to CTD (Cell Transfer Delay) and the generic jitter corresponds to CDV (Cell Delay Variation). Both CTD and CDV depend on the connection path, but performance objectives for these parameters are available for ITU-T stringent QoS class. Mechanisms should be provided within ATM such that these parameters may be retrieved by the application. Indeed, CTD and CDV may be individually signalled for ATM Forum UNI 4.0.

In [14] an end-to-end delay of less than 70 ms for a 32 kbit/s ADPCM stream is estimated, in the case of a long haul ATM path of 30 switches and 10.000 km, 34 or 155 Mbits/s links loaded from 30% to 90% with traffic of the same type. In this case, the conversational MOS as estimated by the ETSI model (see Figure 2b Section 2.2.1) is close to standard PSTN quality.

3.2.2 Data loss

For ATM the mapping is again straightforward: the corresponding parameter is CLR (Cell Loss Ratio).

According to the results of Section 2.3, the MOS of the considered application given the ATM/AAL1 fixed payload size (*pload*=47 bytes), CLR, and the coding data rate (*drate =4 kbyte/s*) can be predicted as:

$$MOS = 4.0 - 0.7\ln(CLR) - 0.1\ln(pload / drate)$$

For a MOS target close to 4.0, the equation above leads to an acceptable loss ratio of approximately 0.7%. Since with ATM it is not possible to specify different loss tolerance values for each individual cell stream, streams are subjected to the same CLR. For stringent QoS class, the end-to-end objective for CLR is 3.10^{-7} as an upper bound, which leads to negligible impact on the MOS.

3.3 IP over ATM

In order to benefit from the inherent QoS support that ATM networks can offer, a mapping from the IP QoS model (IIS) to the ATM QoS model is necessary. This mapping includes, among other things, the translation of the IIS parameters for each QoS control service onto ATM signalling parameters.

A draft proposal for the translation of parameters from IIS to ATM is presented in [2]. In the case of Guaranteed Quality of Service (GS), the draft suggests mappings for DBR and SBR. Since delay guarantees are required for the example application considered, and also for simplicity purposes, this section covers only the DBR case.

At the IP layer the same values as those given in Section 3.1 are taken for the GS reservation parameters. In this case, PCR can be directly derived from the maximum datagram size *M* (since the rate *R*=*r* is chosen in function of *M*), by taking into account the additional protocol overhead and then converting the result from bytes per second to cells per second. Following the encapsulation schemes defined in RFC1483 [10] (VC multiplexing or

LLC encapsulation for IP over AAL5) up to 8 extra bytes of header overhead have to be taken in to account, in addition to the 8 bytes of the AAL5 trailer. PCR can be then calculated as follows:

$$PCR \geq \left\lceil \frac{M+16}{48} \right\rceil . \frac{drate}{M-hsize}$$

where *drate* is the codec data rate equal to 4.10^3 bytes/s for the speech application considered, *hsize* is the IP level total header size equal to 60 bytes in case of RTP/UDP/IPv6. The resulting PCR is in the range of 100 cells/s (for M=400 bytes) to 300 cells/s (for M=100 bytes).

The choice of *b* and *p* has to be taken into account by the local ATM shaper at the sender side, which is responsible for providing a compliant traffic to the ATM network.

3.3.1 Delay and jitter

From the IP point of view, ATM is regarded as a link layer technology, even if in fact ATM is much more than that. The total delay of the ATM portion of the path is considered as a link delay (thus fixed) and therefore added to the D_{tot} error term of the equations (1) or (2) from Section 3.1.1, because knowledge about the internals of the ATM cloud is difficult to obtain at the IP over ATM nodes.

The end-to-end delay depends of course on the number of IP and ATM elements in the network, and the proportion between the IP part and the ATM part. A simple case is where only the end hosts implement IP over ATM and all the intermediate nodes are part of the ATM cloud. In this case, the delay calculation is reduced to a solved problem where the ATM case is considered together with some additional delay due to segmentation/reassembly, shaping, etc., at the end hosts. In this simple case, the guarantees on delay and jitter are deterministic, and thus improved with respect to the pure IP solution. The worst case would be when all the IP elements are connected via disjoint ATM networks (no shortcuts possible). In this case little benefit can be taken from IP over ATM compared to pure IP or ATM.

3.3.2 Data loss

According to RFC1483 [10] an IP packet must be discarded in case of AAL5 CRC mismatch, which occurs due to lost ATM cells, or cells containing errors. Considering this, a worst case approximation on the IP packet loss ratio (*IP_plr*) as a function of ATM CLR (Cell Loss Ratio) and CER (Cell Error Ratio), given the maximum IP packet size (M, in bytes), for DBR (considering compliant traffic), is given by:

$$IP\_plr \leq \left\lceil \frac{M+16}{48} \right\rceil .(CLR+CER)$$

According to the current IP over ATM QoS mapping proposal [2], CLR has to be as low as possible, so that link errors rather than congestion can be source of cell loss. The stringent class objectives for the upper bounds on CLR and CER are 3.10^{-7} and 4.10^{-6}, respectively. For M in the range of 100 to 400 bytes, applying the equation above leads to a worst case IP packet loss ratio of the order of 10^{-5}, considering normal network operation. Recalling Section 3.1.3, this results again in no impact on the MOS of the given application. For a target MOS close to 4.0 the combined values of CLR and CER must be lower than 0.2%. The packet size has little impact on the MOS given a loss ratio. Consequently, the cascade effect of cell loss to packet loss has little influence on the quality of the considered speech application, during normal network operation. Intelligibility is affected but still possible in spite of the bad quality for IP losses of about 50%, which roughly correspond to 5% to 20% of losses in the ATM network, which are of course extremely high for any realistic network operation.

4 Conclusions and future work

A top-down approach to estimate the perceived quality of conversational speech as a function of specific network parameters was presented and applied to IP, ATM and IP over ATM. Although a traditional telephony style application was taken as a focus, the same approach is valid for any other application. The approach allows the calculation of the QoS at the user level by measuring specific network parameters, minimizing the need for further subjective experiments. Additionally, it helps applications to make optimal use of the network, by knowing what can be expected from the network, and understanding the extend to which lower network performance can degrade the quality.

The example application running over IP with Guaranteed Quality of Service (GS) can basically choose only the packet size. The other parameters are bounded by the application characteristics (constant rate traffic) such that only a reduced set of values make sense. The choice of the packet size is driven by the trade-off between delay and protocol overhead. The delay that the application can calculate based on GS can be kept low enough provided that the packets are kept small and the error terms of GS are not excessively overestimated. But small packets add extra protocol overhead thus extra bandwidth usage. With IP, the header overhead can be minimized by the use of header compression techniques, but this also introduces extra processing overhead and delay. The loss ratio can not be specified by the application, but the network is expected to provide much lower values. Given a loss ratio, the influence of packet size is limited.

One of the main problems of IP GS at the moment is the overestimation of the delay, which can affect the quality. This problem can be minimized while running the application over ATM (AAL1/DBR), where the QoS guarantees are fixed and trivial for the example application.

ATM was designed to provide QoS guarantees, but IP is still starting in this field and far from providing the same guarantees as ATM (and not aiming at). Using IP over ATM, IP applications can benefit from the QoS guarantees of ATM, and ATM can benefit from the interoperability and large number of applications available for IP. The delay can be improved with respect to the pure IP case, when the number of intermediate IP nodes is kept small (ideally via short cuts), provided that the translation from IP QoS to ATM QoS is accurately implemented. The cascade effect of cell loss to packet loss has little influence on the quality of the considered speech application, during normal network operation. On the other hand, IP over ATM is a complex technology, in which translating QoS parameters from IP to ATM constitutes only a small part of this complexity. The protocol overhead for small IP packets over ATM is considerably high, and interest in header compression techniques seems to be little up to the moment.

For the simple telephony-like service chosen, it is trivial that all the three transmission mechanisms considered can satisfy the requirements under normal conditions. In practice, however, network problems do occur, and understanding the impact of these problems on the perceived quality is crucial to the design of new services. The user experiment presented in Section 2.3 showed that relatively high loss rates degrade the quality but losses up to 1% give acceptable speech quality for all packet sizes considered. Furthermore intelligibility is still possible even with losses as high as 50%, in spite of the bad quality. Therefore, when the speech quality is affected to the point that intelligibility becomes difficult, it is an alert sign of extreme performance degradation, which is unlikely to come only from the network but might also come from the end systems, or from a mismatch between the application and the network. The design and performance of terminal equipment and application software has a decisive impact on the quality though out of the scope of this paper.

The approach to estimate the perceived quality as a function of specific network parameters presented here is based on computing the MOS separately for each parameter. The next step is to combine those parameters into a single function that would give the combined effect of the most relevant distortions on the perceived quality, as it is provided in the ETSI model. This model is based on the concept that “psychological factors on the psychological scale are additive”. The impact for each kind of impairment is computed and the result is added up together to form a “Transmission Rating Factor” which is then used to calculate the MOS (using an almost linear relation), among other measurements. Applying a similar combination technique for specific IP and/or ATM related impairments is left for further study.

Acknowledgments

The authors would like to thank Nicky van Foreest, Pascal Drabik, Kees van der Wal, and Joost Hermans for their comments, suggestions and/or continuous encouragement.

References

[1] M. Borden et al., *Integration of Real-time Services in an IP-ATM Network Architecture*, RFC 1821, August 1995.

[2] M Borden, M. Garrett, *Interoperation of Controlled-Load and Guaranteed-Service with ATM*, Internet Draft (work in progress), <draft-ietf-issll-atm-mapping-02>, March 1997.

[3] N. Erdöl, C. Castelluccia and A. Zilouchian, *Recovery of missing speech packets using the short-time energy and zero-crossing measurements*, IEEE Trans. On Speech and Audio Processing, July 1993, p.295-303.

[4] ETSI, *Transmission and Multiplexing (TM); Speech communication quality from mouth to ear of 3,1 kHz handset telephony across networks*, DTR/TM-05006, ETR250, France, July 1996.

[5] ITU-T Recommendation E.800, *Quality of service and dependability vocabulary*, August 1994.

[6] ITU-T, Recommendation G.114, *One-way transmission time*, February 1996.

[7] ITU-T, Recommendation P.48, *Specification for an intermediate reference system*, April 1989.

[8] ITU-T, Recommendation P.800, *Methods for subjective determination of transmission quality*, August 1996.

[9] ITU-T, Recommendation P.810, *Modulated Noise Reference Unit (MNRU)*, February 1996.

[10] J. Heinanen, *Multiprotocol Encapsulation over ATM Adaptation Layer 5*, RFC 1483, July 1993.

[11] N. Kitawaki, *Pure delay effects on speech quality in telecommunication*, IEEE J. Sel. A. Com. Vol.9 May 1991, pp. 586-593

[12] H. Nagabuchi, A. Takahashi and N. Kitawaki, *Speech quality degraded by cell loss in ATM networks,* NTT Review, July 1992, p. 45-51.

[13] S. Shenker/C. Partridge/R. Guerin, *Specification of Guaranteed Quality of Service*, Integrated Services WG Internet Draft (work in progress*)*, <draft-ietf-intserv-guaranteed-svc-08>, February 1997.

[14] J.C. van der Wal, M.R.H. Mandjes, H.J.M. Bastiaansen, *Delay Performance of the New Internet Service with Guaranteed QoS compared to ATM*, IEEE ATM workshop, Lisbon, May 1997.

[15] O.J. Wasem, D.J. Goodman, C.A. Dvorak and H.G. Page, *The effect of waveform substitution on the quality of PCM packet communications*, IEEE Trans. On Acoustics, Speech and Signal Processing, March 1988, p.342-347.

[16] J. Wroclawski, *Specification of the Controlled-Load Network Element Service*, Integrated Services WG Internet Draft (work in progress), <draft-ietf-intserv-ctrl-load-svc-05>, May 1997.

Authors' coordinates

Lidia Yamamoto
E-mail: L.A.R.Yamamoto@research.kpn.com
Tel.: +31 70 332 5092
John Beerends
E-mail: J.G.Beerends@research.kpn.com
Tel.: +31 70 332 2644
KPN Research
Postbus 421
2260 AK Leidschendam
The Netherlands
Fax: +31 70 332 6477