

COVER PAGE

Document downloaded by @DAEL

Tue Nov 10 13:09:25 2026
For personal use

When automatic English translation is provided, only the original document is authentic.
The EAA cannot be held responsible of any translation error

Bibliographical reference

Influence of Sender Parameters and Network Architecture on Perceived Audio Quality, Jan Erik Voldhaug, Erik Hellerud, Astrid Undheim, Erling Austreim, Peter U. Svensson and Peder J. Emstad, Acta Acustica **vol. 94** (Number 1), 2008, pp. 1-11

DOI
https://doi.org/10.3813/AAA.918002

ACTA ACUSTICA UNITED WITH ACUSTICA
Vol. 94 (2008) 1 – 11

Scientific Papers
DOI 10.3813/AAA.918002

Influence of Sender Parameters and Network Architecture on Perceived Audio Quality

Jan Erik Voldhaug, Erik Hellerud, Astrid Undheim, Erling Austreim, U. Peter Svensson, Peder J. Emstad
Centre for Quantifiable Quality of Service in Communication Systems*, Norwegian University of Science and Technology, Trondheim, Norway. {voldhaug, erih, astrid, austreim, svensson, peder}@q2s.ntnu.no

Summary
Network simulations and subjective testing have been used to evaluate the effects of frame lengths, packetization settings, and the advantages of using a Differentiated Services (DiffServ) network architecture for music streaming over Internet Protocol (IP) networks. The simulations show that using smaller packet sizes can result in lower packet loss ratios, even though the total bit rate is increased due to network overhead. The network architecture simulations show that DiffServ make the packet loss process less bursty when compared to the current Best Effort (BE) network architecture. These network simulations were followed up with a subjective test to investigate the effects more thoroughly. The subjective test combined three audio clips, three packet loss ratios (3%, 5%, 7%) and two Error Concealment (EC) algorithms with two packet loss processes. Results from the subjective test show statistically significant effects of audio clip, EC, and packet loss ratio on perceived audio quality as well as interaction effects between the network architectures and both audio clip and packet loss ratio. It was found that for a loss ratio of 3%, the loss pattern from a DiffServ network lead to a significantly higher perceived audio quality than the loss pattern from a BE network. However, for higher loss ratios no differences can be seen between the two network architectures.

PACS no. 43.10.Pr, 43.60.Dh

1. Introduction

Audio streamed over a packet switched network may be distorted by delayed and lost packets. Packets discarded by congested routers, packets lost due to link failures, and packets arriving too late for playback all contribute to the total packet loss. Potential real-time requirements inhibit retransmission of lost packets and limit acceptable packet delay. For interactive applications like Voice over IP (VoIP) and ensemble plays (musicians playing together over network), total end-to-end delays of 150 ms [1] and 20–60 ms [2, 3] respectively are considered acceptable. Coding and packetizing schemes must be chosen carefully for such applications. In addition, queue management in the network might have a significant influence on the packet loss pattern, which in turn might influence the performance of the Error Concealment (EC) and thereby the perceived quality.

Few studies have been carried out that consider the effects of network distortions on perceived audio quality for CD quality audio. The knowledge of such effects is crucial when designing systems for interactive audio applications in resource-constrained network environments. The aim of this work is to evaluate and compare the effects of packet loss on the perceived audio quality. For completeness, the behaviour of both current Best Effort (BE) and future Differentiated Services (DiffServ) network architectures are considered, along with potential differences between them. This work also investigates how the choice of packetization (e.g. packet size and time between packets) in the sender, may affect the packet delay and packet loss ratio for a given audio stream.

It is very hard to define a typical number for both packet loss and delay. The amount of packet loss in the core network will often be close to 0%, so the total amount is mostly dependant on the access network. In a congested wireless network, more than 20-30% packet loss can easily occur. Measurements can be done in real networks, but results are valid only at the time of measurement as they depend on the traffic situation. This work investigates packet loss ratios that are high enough to enable subjective testing of temporal distributions and at the same time low enough to limit the quality degradations. Simulations of bottleneck fixed links are used to establish realistic models of the loss processes for the different architectures. Only packet loss caused by discarded packets in routers are considered in the simulations. Wireless scenarios are not considered in

Received 1 March 2007,
accepted 12 November 2007.

* "Centre for Quantifiable Quality of Service in Communication Systems, Centre of Excellence" appointed by The Research Council of Norway, funded by The Research Council, NTNU and UNINETT. http://www.q2s.ntnu.no

© S. Hirzel Verlag · EAA

Document downloaded by @DAEL, 2026-11-10 13:09:27 - personal use only

this work, but are likely to introduce more isolated losses, see e.g. [4]. Also, for simplicity, no compression is employed in the subjective test, i.e. possible interaction effects between packet loss and codec distortions are not considered. The effect of codec distortions combined with packet loss is discussed in [5].

Section 2 presents some basics concerning audio over IP. Packetization simulations are described in section 3, while section 4 covers effects of network architectures as well as the results from the subjective test carried out in this study. Concluding remarks are given in Section 5.

2. Audio over IP

In this section we give a brief overview of techniques necessary for transmitting audio over IP networks. We also seek to identify factors that influence the perceived audio quality when transmitting audio over an IP network.

2.1. Techniques for audio over IP

In order to transmit audio over IP networks, the signal has to be coded and packetized before transmission. In addition, in situations with packet loss some processing on the receiver side is necessary. A short introduction to these techniques is given in this section.

2.1.1. Encoding

As mentioned in the introduction, interactive applications impose strict delay requirements on the transmission of speech and particularly high quality audio (music). The available audio coding techniques differ quite a bit with respect to how much delay they introduce.

Some audio coders are designed specifically for such applications, with low delay. Low Delay Advanced Audio Coding (AAC) introduces an algorithmic delay of at least 20 ms [6] whereas Fraunhofer's Ultra Low Delay codec achieves an algorithmic delay of under 10 ms [7]. APT's [8] APT-X offers less than 3 ms delay at 48 kHz sampling frequency, at the price of a higher bit rate.

The frame size plays a major part in the encoding delay. In MPEG-4 AAC [9] the framing and filterbank will add a delay of 2048 samples. The transform block size is 2048 for stationary signals, but the encoder can switch to 8 blocks of 256 samples if transients are detected within the frame. This feature adds another 576 samples to the total delay [6]. It should also be noted that the use of a bit reservoir will add significant amounts of delay. In Low Delay AAC the delay is reduced by using shorter frames, and removing the possibility for transform block switching.

Restricting the encoder to use short frames and correspondingly short transform lengths, will in general lead to a decreased coding gain for stationary signals. However, short frames enable the use of smaller and more frequent packets, which in turn may affect the loss and delay characteristics.

One interesting property is that all frames have to overlap with 50% to achieve perfect reconstruction with the Modified Discrete Cosine Transform (MDCT), which is used in AAC. This means that if one frame is lost, 2048 samples will be affected. However, since the frames are windowed before the transformation, this will create a natural fade-in and fade-out [10] if a frame is lost.

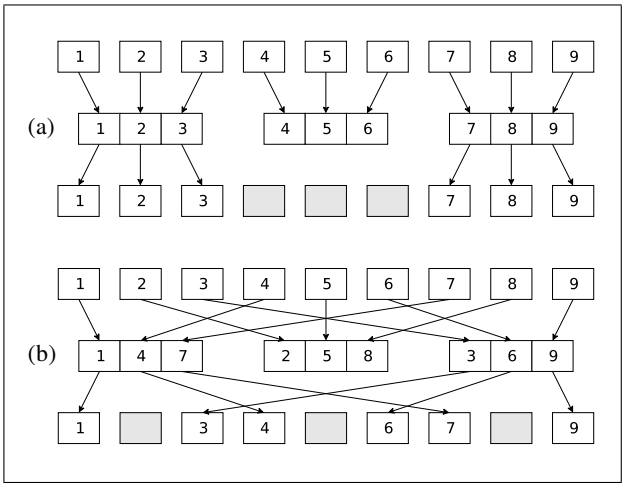


Figure 1. Illustration of the consequences of packet loss both with and without interleaving. (a) Without interleaving: Three consequtive AUs are lost. (b) Interleaving: Lost AUs are spread out.

2.1.2. Packetization

When AAC encoded content is prepared for streaming by the sender, it is packetized according to RFC 3640 [11]. The bit stream resulting from one frame is called an MPEG-4 Access Unit (AU). In normal mode, consecutive AUs are placed in the same IP-packet as long as the total packet size does not exceed the Maximum Transmission Unit (MTU) for the network being used. The MTU is normally 1500 bytes for Ethernet. In Figure 1a the second packet is lost, which leads to the loss of AU 4-6. Three consecutive frames will create a large gap, which can be difficult for the receiver to conceal satisfactorily. Another possibility is to use interleaving, as illustrated in Figure 1a. Using interleaving, two consecutive AUs will not be put in the same IP packet. A packet loss will lead to multiple small gaps instead of one large gap, which according to [12] increases the perceived quality significantly. However, interleaving will also increase the total delay. The sender has to buffer up AUs before they are packetized, and the receiver has to buffer AUs before they are played out.

The packet size and time between packets depends on the encoder, the chosen bit rate as well as packetization settings. The use of short frames and heavy compression will lead to smaller packets. Depending on the output from the encoder, data is placed in packets along with packet headers. For real-time or near real-time multimedia streaming it is common to use the Real-time Transport Protocol (RTP) [13] as application layer protocol and User Datagram Protocol (UDP) [14] as transport layer protocol. The overhead for each RTP/UDP/IP packet then becomes 40 (12/8/20) bytes, when disregarding the link layer header bytes. If less audio data is placed in each packet,

Document downloaded by @DAEL, 2026-11-10 13:09:27 - personal use only

more packets are required to transmit the same data, and the overhead rate increases.

2.1.3. Network Architecture

With the current Best Effort (BE) Internet, no guarantees regarding packet loss, packet delay or variations in delay are provided. Further, all types of traffic are treated the same without any service differentiation. In contrast, DiffServ [15] defines an architecture for implementing scalable service differentiation in the Internet. Scalability is achieved by aggregating the traffic into a small number of classes called behavior aggregates. A set of functional elements is then implemented in the network, including elements for packet classification and traffic conditioning at network boundary nodes. A set of Per-Hop-Behaviors (PHB) is defined for the aggregated traffic classes, specifying the forwarding behavior in the routers. This way, reasonable buffer space and bandwidth share are allocated to each class. Two PHBs have been defined: Expedited Forwarding (EF) PHB, which provides bounds on the end-to-end delay, and the Assured Forwarding (AF) PHB group with four different AF classes. For each AF class, additionally three different drop probabilities are defined, further enabling packet prioritization within an AF class.

Different queue management schemes are possible in both BE and DiffServ networks, with tail drop and Random Early Detection (RED) [16] as the most common. To avoid long-term congestion, an active queue management scheme is used in the AF classes in DiffServ, and RED was proposed in [17]. With tail drop, packets are dropped when arriving at a full queue, while with RED, packets are dropped with increasing probability as a function of the average queue size, reducing the possibility that losses occur in bursts.

Previously, Muppala et al. investigated VoIP performance on a DiffServ enabled network and found significant improvements with regard to delay and jitter, compared to BE [18]. However, their findings were not followed up with a subjective test. Aspects of video transmission over DiffServ have also been investigated, e.g. by Zhu *et al.* [19]. An evaluation of packet loss distortion on wide-band audio has also previously been carried out [20].

2.1.4. Processing in the receiver

When a packet is lost, it will in general lead to the loss of one or more audio frames, which have to be concealed by the receiver using some Error Concealment (EC) algorithm. EC techniques will replace missing audio samples with estimates based on knowledge of the transmitted signal and the codec, and thereby increase the perceived audio quality. Techniques range from basic schemes like silence insertion to computationally complex interpolation algorithms in the frequency domain [21]. In general, more complex EC techniques will result in higher perceived audio quality at the expense of increased computational complexity.

If redundancy is used in the packet stream, effects of packet losses can be reduced, but adding redundancy will increase the total bit rate. This redundancy can e.g be a separate stream with a lower bit rate [22], information to reconstruct parts identified by the sender as important [23], or simply sending parts of the bit stream multiple times [24].

2.2. Quality Considerations of audio over IP

Algorithms exist for predicting the perceived quality of speech and audio that is transmitted over a network. One category of algorithms compares the sent reference signal to a received distorted version. In this category, the PESQ [25] predicts quite successfully the perceived quality of speech samples, even with packet loss distortion. A corresponding algorithm for general audio signals, PEAQ [26], has not been developed for handling packet loss distortion, and consequently performs poorly for predicting such distortion [20]. Another category of algorithms uses a parametric approach, that is, the quality prediction is based on transmission channel parameters rather than a specific distorted sound clip. The E-model has been developed for speech [27, 28] and handles packet loss distortion as well [29, 30]. However, no similar objective method has been developed for general audio, or music signals. Hence, subjective testing were carried out in this work. Below, the various factors that influence the perceived quality of audio over IP are discussed.

2.2.1. Packet loss ratio

When routers become congested (full queues), arriving packets have to be dropped. For non real-time applications (web browsing, file downloads, email etc.) this is not necessarily a problem since protocols such as the Transmission Control Protocol (TCP) will make sure that lost packets are retransmitted. However, for real-time applications there is often no time to wait for retransmission, and the data in a lost packet is indeed lost.

The impact of a lost packet on perceived audio quality depends both on the content of the lost packet and the content of surrounding packets. If the lost data consists of an onset or a transient, the use of neighboring packets for estimation becomes difficult. For more stationary sounds, the content of neighboring packets can be used to effectively estimate the missing data. How the audio is packetized and encoded will also influence the quality, and the frame length can be an important factor as well. Also, if consecutive audio frames are placed in the same packet, this will lead to a large gap, as illustrated in Figure 1a. Evaluation of packet loss distortions for music has been performed in [5, 20] and for speech in [29]. These studies have shown that that the quality decreases quickly when even small amounts of packet loss are introduced. Here, we define the packet loss ratio as the ratio of packets that are lost, to the total number of transmitted packets.

2.2.2. Packet delay

Interactive applications can have end-to-end delay requirements as low as 20 ms, as discussed in section 1 [2, 3]. Overly delayed packets that do not arrive at the receiver

Document downloaded by @DAEL, 2026-11-10 13:09:27 - personal use only

in time for playout translate into lost packets. For non-interactive applications (moderate) delays need not to be a major concern. Varying delay (jitter) will, however, still be a cause of concern. It is the jitter, rather than the average delay, that determines the required size of the receiver's playout buffer.

2.2.3. Temporal loss distribution

The temporal loss distribution also affects the perceived audio quality. If the distribution is bursty, it can be problematic to reconstruct the lost data satisfactorily, but on the other hand, longer periods of error free transmission can be expected. As previously mentioned, a tail-drop strategy in the router will typically lead to a bursty loss distribution, while a RED strategy will reduce the burstiness.

In [30, 31] it was found that a bursty packet loss process leads to larger impairments for speech when compared to random packet loss, but that the effect is most dominant for loss ratios of 5% and higher.

3. Effects of Packetization

First, we investigated how encoding and packetization affects the network and also how the network distorts the studied flow. Frame lengths, packet sizes, and time between packets were varied in the simulations. The different packetization settings are shown in Figure 3. Changes in cross traffic audio flow characteristics were also simulated.

3.1. Packetization Simulations

Simulations were carried out using a trace driven simulator of a bottleneck node implemented in J-sim [32]. All simulations used the set-up depicted in Figure 2, where the studied audio stream, or flow, was combined with cross traffic.

3.1.1. Constant payload bit rate

A router with an output link capacity of 2 Mbps was simulated for queue sizes of 50 and 100 kB. The cross traffic in the simulations consisted of both TCP and multimedia (UDP) traffic. We based our encoding simulation on the ISO/MPEG Advanced Audio Coding (AAC) standard.

Three different audio cases were simulated, as illustrated in Figure 3. The first case, *large packets*, was a "standard" 128 kbps AAC flow, with 1024 samples in each frame (43 frames per second). In the second case, *medium packets*, rather than attempting to fill the packets, only one AU is placed in each packet. The third case, *small packets*, used the same packetization scheme as the previous, but the frame length from encoding was reduced so that each packet contained only 192 new samples. It should be noted that this is not possible within the AAC standard, but this case was added to evaluate how a codec with this possibility would function in the network. This means that there were 230 frames, and hence packets, per second rather than 43. As the total number of header bytes per packet

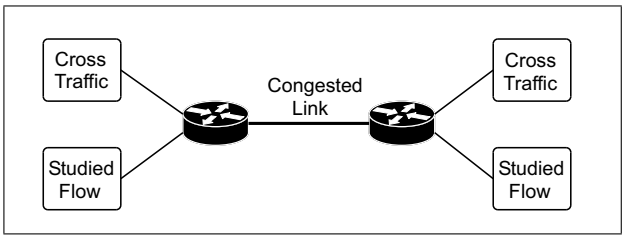


Figure 2. Simulation set-up.

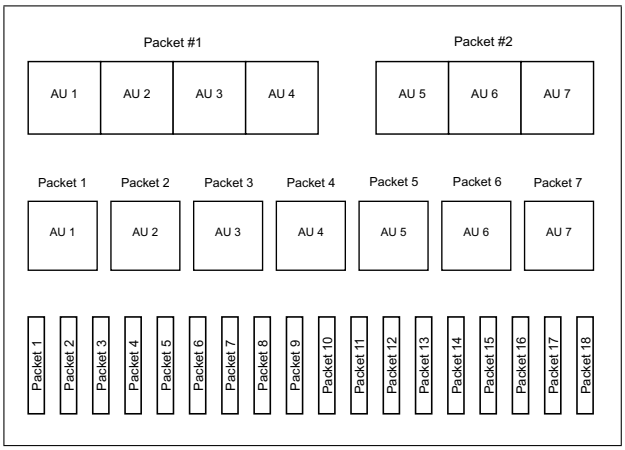


Figure 3. Packetization schemes. Case I (top): Large packets. With 128 kbps each IP packet will typically contain 3–5 AUs. Case II (middle): Medium packets. One AU in each IP packet, leading to approximately 43 packets per second for a sampling rate of 44.1 kHz. Case III (bottom): Small packets. Each IP packet contains 192 samples which means just under 230 packets per second for 44.1 kHz. Case I and case II are based on long audio frames, while case III is based on short audio frames.

Table I. Efficiency for the different packetization schemes and two payload rates. The efficiency for *Large* is given as the upper bound, that is if all packets have a payload of 1460 bytes.

Packetization Scheme	Payload Rate	Total Rate	Efficiency
Large	192 kbps	197.3 kbps	97.3%
Large	128 kbps	131.5 kbps	97.3%
Medium	192 kbps	205.8 kbps	93.3%
Medium	128 kbps	141.8 kbps	90.3%
Small	192 kbps	265.5 kbps	72.3%
Small	128 kbps	201.5 kbps	63.5%

is constant, increasing the total number of packets will result in an increase in the total network overhead. If the payload / encoding rate is kept constant, as done here, this also leads to an increase in the total bit rate (Table I).

3.1.2. Constant total bit rate

Simulations were also performed with a constant total bit rate, independent of frame length and packetization. Keeping the total bit rate constant while reducing the packet size implies a reduction of the payload rate. For these simulations, the output link of the router was set to 4 Mbps

Document downloaded by @DAEL, 2026-11-10 13:09:27 - personal use only

and the queue size was 50 kilobytes. Flows of 170 kbps and 230 kbps total bit with both *large* and *small* packets were simulated. All simulations were performed twice, with changes in the audio flows in the cross traffic as the only difference. Initially, the audio cross traffic consisted of audio flows with all three packetization settings (mixed packetization). This was then changed so that all audio cross traffic flows adopted the same packetization as the studied flow. It should be noted that the cross traffic consisted of both TCP and video traffic in addition to audio.

3.2. Packetization Simulation Results

The studied flow contributed to only a small part of the total router throughput and the various settings for the studied flow did not appear to have any great effect on the router performance. The router appeared to become congested at approximately the same time regardless of the characteristics of the studied flow, i.e. this was largely determined by the cross traffic.

From the simulations it was found that the average packet delay depends heavily on the router queue length, as is self-evident, but not on the packet size / inter-arrival times of the studied flow, see Figure 4a. The packet loss ratio, however, depends heavily both on the queue length and the packetization, as seen in Figure 4b. The packet loss ratio experienced by our flow is significantly lower when many small packets are used instead of fewer large packets. The packet loss ratio are reduced when the queue length is increased.

Results from the second round of packetization simulations are shown in Figure 5. Again we see, in Figure 5a, that the average packet delay caused by the router is fairly independent of the packet size and inter-arrival time of the studied flow. We also see that even if the difference in overhead-rate for the case with small packets is compensated for by keeping the total rate of the flows equal, the difference in experienced packet loss ratio is still large. The packet loss ratio is reduced by a factor of about 25 when we switch to small packets. Changing the characteristic of the audio flows in the cross traffic does not seem to have any effect on neither the packet loss ratio nor the packet delay.

The difference in packet loss ratio between the cases with small and large packets in the studied flow is found to be very large. In these simulations tail-drop queue management is used in the router, meaning that if the router is in congestion (the router queue is full) when a packet arrives, the packet is discarded. It is, however, easier for a small packet to be accepted into a router queue than for a larger packet which requires more free space. It is observed that the packet loss ratio is not affected if the audio flows in the cross traffic adopt the same strategy with many small packets, which might seem surprising. It is, however, likely that the packetization strategy of the audio cross traffic would be more significant had it constituted a larger part of the total cross traffic.

Keeping the total rate constant while decreasing the frame length and packet size results in a decreased payload

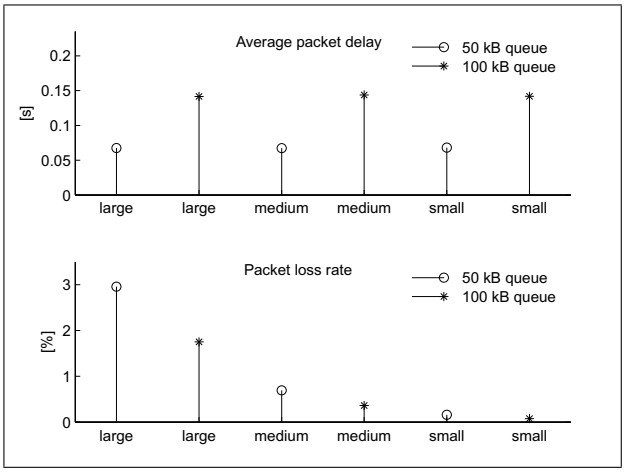


Figure 4. Simulations with a constant payload rate: (a) Average packet delay and (b) packet loss ratios for different packetizations and queue lengths. The three simulated cases from left to right: Large, medium and small packets.

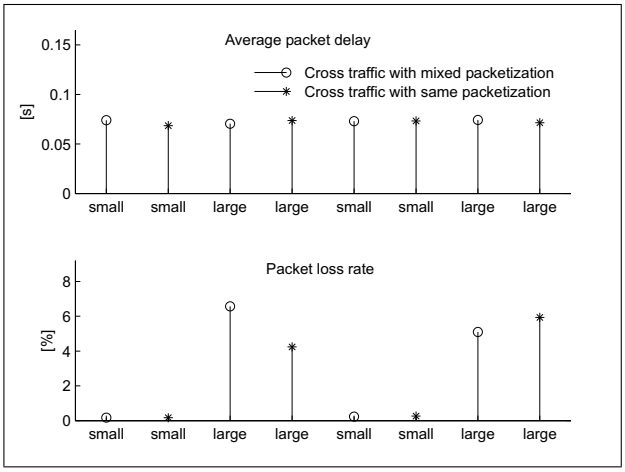


Figure 5. Simulations with a constant total rate: (a) Average packet delay and (b) packet loss ratios for different packetizations and audio cross traffic packetizations. The four cases from left to right: short frames (170 kbps), long frames (170 kbps), short frames (230 kbps), and long frames (230 kbps). Rates are given as the total bit rate for the flow.

rate as the network overhead increases. In other words, there is a trade-off between degradations from packet loss and compression. Transmitting the audio in small packets is found to decrease the packet loss ratio, but will inevitably result in either an increase in the total flow bit rate or a decrease in the payload rate. It is shown in [5] that a reduction of the encoding rate from 128 to 64 kbps may not affect the acceptability in situations with packet loss. It should also be noted that from a receiver's point of view it is generally easier to conceal the loss of smaller packets (short frames) than larger packets.

4. Effects of Network Architecture

Simulations were combined with subjective tests to investigate the effect of network architecture on perceived audio

Document downloaded by @DAEL, 2026-11-10 13:09:27 - personal use only

quality. Audio flows were studied in two configurations; one using Best Effort routers, the other using DiffServ routers. In these simulations, the loss ratio is kept the same for both architectures, although it should be noted that the use of DiffServ generally will result in fewer packet losses. The simulation results were used to generate stimuli for the subjective test.

4.1. Network Architecture Simulations

In a BE network there is no service differentiation, so the cross traffic here consisted of both Transmission Control Protocol (TCP) and User Datagram Protocol (UDP) multimedia traffic with tail drop as the queue management scheme in the routers. With DiffServ, high quality audio traffic will typically be put in a separate service class and hence will mostly be affected by the traffic in this class. Therefore, in the DiffServ simulations, the cross traffic consisted solely of audio traffic. The studied audio flow and the cross traffic were both classified as AF traffic and RED was used for queue management in the routers.

Simulations were performed to find loss patterns for loss ratios of 3, 5 and 7%. These loss ratios are rather high, but required in order to evaluate the differences between loss distributions. Lower packet loss ratios will be problematic simply due to the low number of lost packets. To achieve the given loss ratios, the capacity of the bottleneck link was varied. In addition, for the DiffServ simulations, the parameters of the RED curve [16] were varied such that losses occurred because of random drop instead of buffer overflow. It must be noted that the DiffServ architecture is designed to avoid long-term congestion for non BE traffic. A traffic stream classified as AF should then experience lower packet loss in a DiffServ network than in a BE network.

4.2. Simulation Results

In the simulations, time instants when losses occurred were recorded. Times between packet losses were then compared with known distributions using quantile-quantile (QQ) plots, in order to find a good approximation for the distribution of the time between losses. QQ plots verify the distribution similarity between two data sets. As an example, Figure 6 shows the QQ plot for the DiffServ case with 3% average loss ratio against the Gamma distribution with the same mean and variance. As the points follow the 45 degree reference line quite closely, the distribution of the time between losses (in number of packets) can be approximated by the Gamma distribution. The same distribution was found to fit the other DiffServ cases with loss ratios of 5 and 7%. The Gamma distribution is a function of two parameters, α and β, with mean $\mu = \alpha\beta$ and variance $\sigma^2 = \alpha\beta^2$. The parameters for the fitted models are given in Table II.

Losses occur more in bursts for the BE case, and the loss process was found to be best represented by a three-state model, as shown in Figure 7. The states represent the loss period, and short and long periods without any loss. The transition probabilities q and $1 - q$ were determined from the simulations. QQ plots showed that distributions of both the time in each state, and time between losses in the loss period, could be approximated by the Gamma distribution. Distribution parameters are shown in Table III.

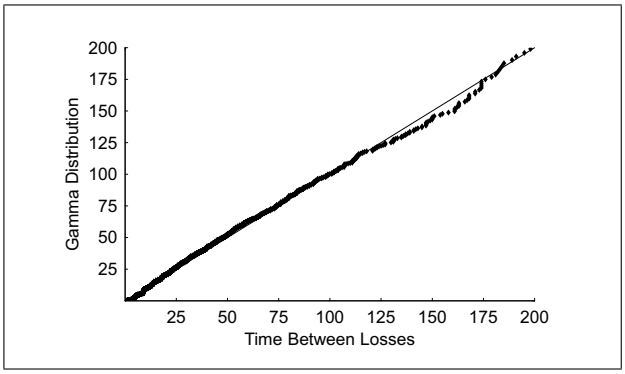


Figure 6. QQ-plot for the simulation of a DiffServ router with 3% average loss ratio versus Gamma distribution with the same mean and variance.

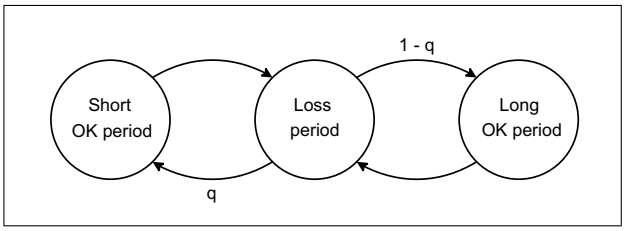


Figure 7. Three-state model representing the loss process for the best-effort simulations.

Table II. Gamma distribution parameters for the DiffServ packet loss process, given as packets between loss.

PLR	α	β
3%	0.688	44.942
5%	0.769	25.324
7%	0.881	16.048

The Burst Ratio (BurstR) [33, 30], which is a measure for the burstiness of a packet loss process, is shown in Table IV for all packet loss ratios. If BurstR is equal to 1, the loss process is by definition random. The burst ratios that were evaluated are rather small compared with the ones in [30].

4.3. Subjective Testing

The test items for the subjective test formed a balanced set of three audio clips, three packet loss ratios, two EC schemes, and two different packet loss models. All 36 test items were evaluated against CD-quality references. Both the stimuli and the listening test method are described below. The fact that no audio compression has been used, enables evaluation of the quality degradation caused by packet loss and subsequent processing, without having to consider other distortions as well.

Document downloaded by @DAEL, 2026-11-10 13:09:27 - personal use only

Table III. Gamma distribution parameters for the BE packet loss process. A: BE packet loss process parameters for 3%, q = 0.416. B: BE packet loss process parameters for 5%, q=0.338. C: BE packet loss process parameters for 7%, q = 0.485.

A	α	β
Loss: Packets between losses	0.624	27.144
Loss: Num. losses	1.272	9.237
Short ok: Num. packets	71.756	2.087
Long ok: Num. packets	20.744	12.115
B	α	β
Loss: Packets between losses	0.631	19.524
Loss: Num. losses	0.905	31.383
Short ok: Num. packets	70.759	1.741
Long ok: Num. packets	16.455	17.906
C	α	β
Loss: Packets between losses	0.564	18.030
Loss: Num. losses	0.653	73.147
Short ok: Num. packets	36.424	3.543
Long ok: Num. packets	15.446	19.628

Table IV. Burst ratios (BurstR) for the packet loss processes.

PLR	DiffServ	Best Effort
3%	1.054	1.128
5%	1.043	1.141
7%	1.019	1.186

Table V. Audio material.

Audio	Details & Source	Length
Castanets	EBU SQAM CD (track 27)	14s
Piano	Mozart, Turkish march	19s
Choir	EBU SQAM CD (track 48)	15s

4.3.1. Processing

Audio clips were exposed to packet loss ratios of 3, 5 and 7% distributed according to the two models developed from network simulation as described in Section 4.2. As each packet is set to contain only one single frame (Figure 3b) and no FEC scheme is employed, a packet loss equals the loss of one audio frame. Clips were divided into frames, with 1024 samples in each frame (23 ms at 44.1 kHz sampling frequency).

Frames missing due to packet loss were replaced with estimates from the EC algorithms. They included a very simple time domain frame repetition and a more computationally complex frequency domain interpolation scheme. The frame repetition algorithm simply inserted samples from the previous frame if a frame was lost. To avoid discontinuities there were transitions areas of 50 samples at the start and end where the inserted frame was faded in and out with a linear slope. The second method was a more complex interpolation algorithm. Amplitude and Phase Estimation (APES) is an adaptive filter bank method for spectrum estimation, and it was extended to handle gapped-data in [34]. In [35] Gapped-data APES (GAPES) was used to implement an EC scheme for MPEG-1 Layer 3 (MP3) encoded audio. In our implementation the GAPES algorithm is applied to each of the 1024 coefficients resulting from the MDCT in the AAC encoder, using data from three previous and five future frames. For each coefficients a maximum of six iterations were run.

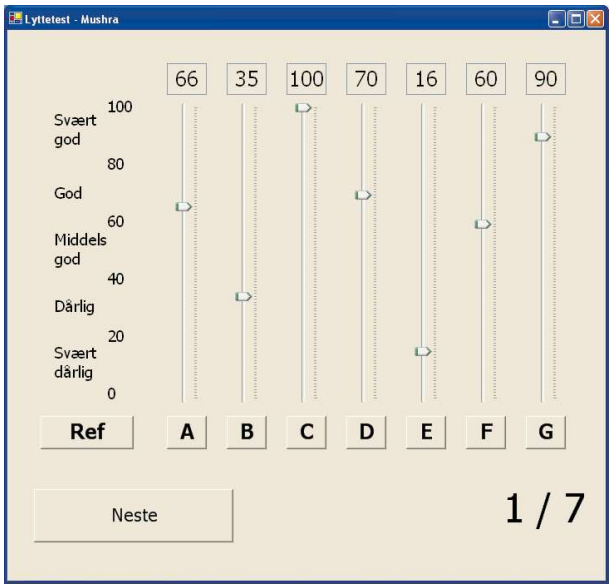


Figure 8. Computer interface. The shown interface is in Norwegian. Each screen presents 6 test items and a hidden reference to the subject.

4.3.2. Selection of Audio Clips

Audio test material included choir song, castanets, and piano. These particular clips are a subset of those used in [20], and were chosen because they were found to respond very differently to packet loss distortions. Clip lengths varied between 14 and 19 seconds. Clip details can be found in Table V.

4.3.3. Listening Test Methodology

Subjective testing was based on the procedures outlined in ITU-R Recommendation BS.1534 (MUSHRA) [36]. Due to the nature of the stimuli used in this test (drop-out errors), band-limited anchors were not employed. Our own software implementation was used (Figure 8).

The multi-stimulus methodology of MUSHRA enables subjects to instantly compare and rate several test items. Instead of just one, subjects are presented several test items and one hidden reference at a time (see Figure 8). It is better suited than ITU-R BS.1116 [37] for evaluation of medium and large impairments [38]. The MUSHRA grading scale runs from 0 to 100, where 0–20 means *bad*, 20–40 *poor*, 40–60 *fair*, 60–80 *good* and 80–100 *excellent audio quality*.

Document downloaded by @DAEL, 2026-11-10 13:09:27 - personal use only

4.3.4. Subjects and Procedures

26 non-expert listeners (6 female and 20 male) participated in the test. The subjects were university students between 19 and 26 years old, and they were paid to participate. The test was split up into three main session, and all listeners rated each of the 36 test items twice. The test items were ordered randomly for each subject, and presented to subjects 6 at a time, see Figure 8. I.e. in the first half of the test, subjects were presented with 6 'screens', 2 for each of the 3 audio clips. The sessions lasted approximately 30 minutes each, and in between sessions there were 10 minute breaks.

Before the actual test there was a short training phase where subjects were introduced to audio clips, distortions, test methodology, and the test software interface. Subjects received thorough instructions in writing before the start of the test, and were also asked to fill out a short questionnaire with simple questions about their hearing.

Test items were played out to subjects from local workstations using AKG K271 Studio headphones and EDIROL UA-25 USB sound cards. Subjects were able to adjust the playout-level during the training phase, but it was kept constant during the three main sessions. The test was performed in a sound insulated room designed for listening tests.

4.4. Subjective Test Results

Results of participants that reported hearing weaknesses were omitted. This applied to 2 out of 26 initial participants. Two main classes of post-screening methods are discussed in the MUSHRA recommendation [36]. One is the ability of subjects to make consistent repeated gradings. The other deals with inconsistencies of an individual grading compared with the sample mean of all subjects. All subjects performed satisfactorily with respect to both these criteria and consequently no subjects were omitted.

The figures in this section present sample means and 95% confidence intervals. Most of the presented confidence intervals are narrow. This is due to the high number of test subjects, test items, and hence data points.

4.4.1. ANOVA

ANalysis Of VAriance (ANOVA) was used for the statistical analysis of the output from the subjective test (See Table VI). A balanced ANOVA was used, with Subject as the only random factor. We investigated both main effects, which are effects of single factors on the audio quality averaged over all values of other factors, and interaction effects. Two factors interact if the effect of one depends on the value of the other.

As seen from the table; clip, EC and PLR are all significant main effects, but network architecture is not a significant factor in itself (due to the high P-value). However, architecture interacts significantly with both PLR and clip. Also, there are significant interaction effects between clip and both PLR and EC, which corresponds to the results in [20].

Table VI. ANOVA table.

Source	DF	SS	MS	F	P
Subject	23	288195	12530	62.98	< 0.001
Clip	2	20669	10335	51.95	< 0.001
PLR	2	131780	65890	331.20	< 0.001
EC	1	7252	7252	36.45	< 0.001
Architecture	1	380	380	1.91	0.167
Arch. x Clip	2	23027	11514	57.87	< 0.001
Arch. x PLR	2	8330	4165	20.94	< 0.001
Arch. x EC	1	4	4	0.02	0.889
Clip x PLR	4	5348	1337	6.72	< 0.001
Clip x EC	2	8905	4452	22.38	< 0.001
PLR x EC	2	127	64	0.32	0.726
Error	1685	335224	199		
Total	1727	829241			

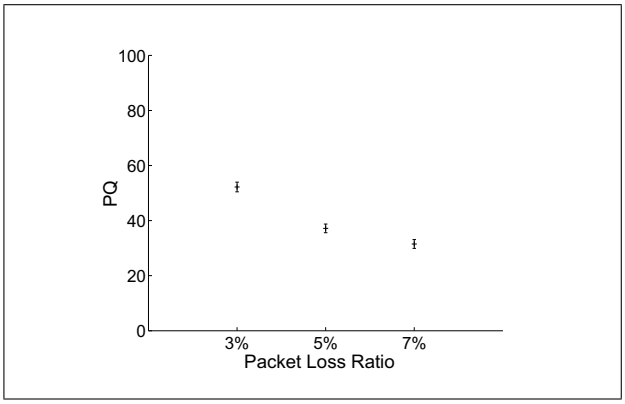


Figure 9. Perceived audio quality as a function of packet loss ratio. Sample means and 95 % confidence intervals are shown.

4.4.2. Audio Quality vs Packet Loss Ratio

Results from the subjective test were analyzed with respect to the packet loss ratio. As expected the perceived quality decreases with increasing packet loss ratio, and a steeper decrease is observed at low loss ratios (see Figure 9). This corresponds to the findings for music in [20] and also for speech [29].

The differences between the three packet loss ratios are all significant based on a Bonferroni test ($p < 0.001$ for all three comparisons). Note that the packet loss degradations are quite severe, even for 3% packet loss. In a BE network the packet loss ratio is hard to control. In a DiffServ architecture on the other hand, a traffic conditioner ensures that each traffic class conforms to its traffic profile. Traffic exceeding this profile may be given lower priority or dropped. In this way low loss ratios can be ensured for traffic classes following the defined traffic profile. Packet loss ratios well below 5% are then most realistic. Lower loss ratios than 3% are, however, not included in this work. Differences in loss patterns become increasingly hard to evaluate with a low number of losses, as this requires the clips to be very long.

It should also be noted that the parameters set in this work influence the perceived quality. For instance, the use

Document downloaded by @DAEL, 2026-11-10 13:09:27 - personal use only

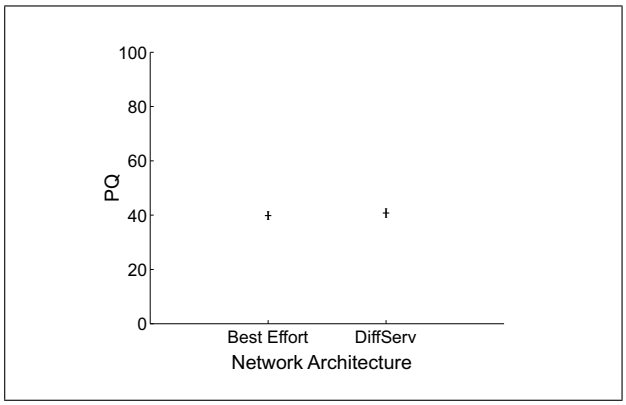


Figure 10. Sample means for the two network architectures, with 95 % confidence intervals.

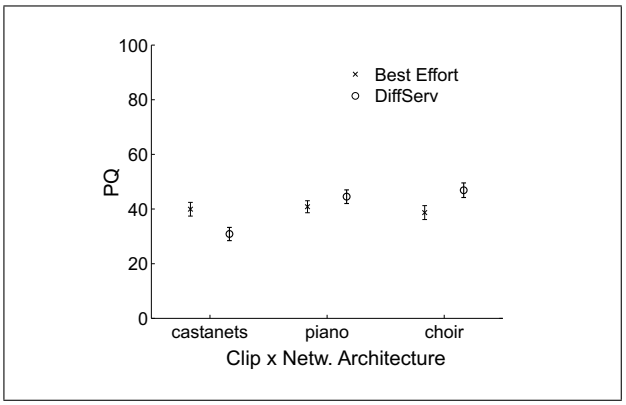


Figure 12. Interaction effects between network architecture and audio clip. Sample means and 95% confidence intervals.

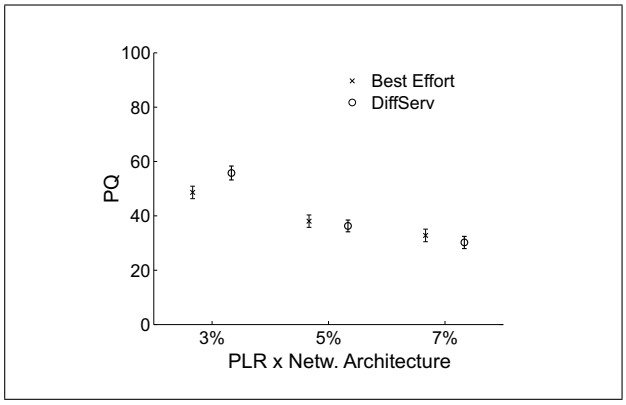


Figure 11. Interaction effects between network architecture and packet loss ratio. Sample means and 95 % confidence intervals are shown.

of shorter frames may have lead to smaller degradations in perceived quality as result of packet loss, at the price of more network overhead and reduced frequency resolution in the encoder. The choice of EC technique obviously also strongly influences the perceived quality, and smaller quality degradations may quite possibly be achieved using other algorithms than in this work. However, we believe that the effects of the various factors would still be relatively similar to those observed here. In the same way as for packet loss ratio, significant main effects of both audio clip and error concealment scheme were found. The clip "*castanets*" performed worse than the other clips. This clip has a very transient nature and was also found to be very sensitive to packet loss in [20]. Surprisingly, it was the faded frame repetition error concealment algorithm that performed best in the subjective test. The reason why our GAPES algorithm doesn't match those results found in [35] could be because they only consider rather high loss ratios (10%, 20%, and 30%).

4.4.3. Audio Quality vs Network Architecture

When averaging over all other factors, no statistically significant difference in audio quality was found between the two network architectures, see Figure 10. However, statistically significant interaction effects were found both between network architecture and packet loss ratio, and between network architecture and audio clip. The former is shown in Figure 11.

A post-hoc Bonferroni test showed no significant difference between the two network architectures for both the 5% loss ratio ($p = 0.11$) and the 7% loss ratio ($p = 0.06$). However, for the 3% loss ratio, the perceived quality of the DiffServ simulations is higher than for the corresponding BE cases ($p < 0.001$). Since the lower loss ratios are the most relevant ones, as argued in Section 4.4.2, the DiffServ architecture might offer an improvement over BE in realistic situations.

Interestingly, in [30] burstiness had more influence on the perceived quality (of speech), the higher the PLR. The results here indicate an opposite relationship. A possible explanation for this behaviour is that the audio quality becomes so distorted that the salience of the loss distribution is reduced.

Similarily, an interaction is observerd between network architecture and choice of audio clip, as can be seen in Figure 12. A post-hoc Bonferroni test nevertheless showed that for the BE architecture, there was no significant difference between any of the three audio clips ($p > 0.23$ for all three comparison). On the other hand, for the DiffServ simulations, the castanets clip gave significantly lower perceived quality than the two other clips ($p < 0.001$ for both comparisons), and there was no difference between piano and choir ($p = 0.13$). In addition, a Bonferroni test showed significant differences between the two architectures for all three clips: the DiffServ architecture performed better than the BE architecture for the piano and choir clips, but clearly worse than BE for the castanets clip ($p < 0.001$ for all three comparisons). This may simply be because it is hard to estimate missing frames in the castanets clip due to its transient nature, regardless of the loss distribution. The EC schemes then fail to produce satisfactory estimates of missing frames. However, since the BE losses occur more in bursts, there will also be "long" periods without any losses at all, which contribute positively to the overall quality rating of the clip. This may in fact be an additional reason why the DiffServ architecture is not found to be beneficial at higher loss ratios in this study.

Document downloaded by @DAEL, 2026-11-10 13:09:27 - personal use only

5. Conclusions

Simulations have been used to investigate the interaction between sender parameters and network performance. It was found that by reducing the size of, and time between, audio packets, the packet loss ratio of the flow can be reduced significantly as long as the audio flows constitute a small enough part of the total traffic. The average packet delay caused by the network is not affected by sender packetization, this delay is mostly dependent on the routers queue size.

The impact of packet loss on the perceived audio quality has also been evaluated for two different network architectures. Results from the subjective test show statistically significant main effects of both audio clip, error concealment scheme and packet loss ratio. Interaction effects between different network architectures and both audio clip and packet loss ratios were observed.

The packet loss process from the DiffServ network leads to a significantly higher subjective quality for a packet loss ratio of 3%, but no difference between the architectures is observed for the higher loss ratios. Also, for 2 out of the 3 test clips, the DiffServ packet loss process lead to a higher subjective quality when averaged over all packet loss ratios. Further investigations should be carried out into the relation between network distortions and perceived audio quality for interactive audio applications. Particularly the effects of audio frame size on perceived audio quality should be studied further with subjective tests.

Acknowledgement

The authors express their thanks to Dawn Behne, Dept. of Psychology, NTNU, for assistance with the statistical analysis.

References

[1] ITU-T Recommendation G.114: General Characteristics of International Telephone Connections and International Telephone Circuits: One-Way Transmission Time. Feb. 1996.

[2] C. Chafe, M. Gurevich: Network Time Delay and Ensemble Accuracy: Effects of Latency, Assymetry. Oct. 2004, Presented at the 117th AES Convention. Preprint 6208.

[3] S. Farner, A. Solvang, A. Sæbø, U. P. Svensson: Ensemble hand-clapping experiments under the influence of delay and various acoustic environments. Oct. 2006, Presented at the 121st AES Convention. Preprint 6905.

[4] I. T. Union: Packet Loss Distribution. 3rd FG IPTV meeting, http://www.itu.int/md/T05-FG.IPTV-C-0412/en, January 2007.

[5] J. E. Voldhaug, E. Hellerud, U. P. Svensson: How Much is Too Much? - On the Acceptability of Packet Loss Distorted Audio. 2007 IEEE International Conference on Signal Processing and Communications (ICSPC).

[6] E. Allamanche, R. Geiger, J. Herre, T.Sporer: MPEG-4 Low Delay Coding Based on the AAC Codec. Apr. 1999, Presented at the 106th AES Convention. Preprint 4929.

[7] M. Lutzky, G. Schuller, M. Gayer, U. Krämer, S. Wabnik: A Guideline to Audio Codec Delay. May 2004, Presented at the 116th AES Convention. Preprint 6062.

[8] APT-X: http://www.aptx.com. [Online].

[9] ISO/IEC 14496-3: Coding of Audio-Visual Objects – Part 3: Audio.

[10] Y. Wang, M. Vilermo: The Modified Discrete Cosine Transform: Its Implications for Audio Coding and Error Concealment. J. Audio Eng. Soc **51** (2003) 52–61.

[11] J. van der Meer, D. Mackie, V. Swaminathan, D. Singer, P. Gentric: RTP Payload Format for Transport of MPEG-4 Elementary Streams. RFC 3640, Nov. 2003.

[12] B. Grill, T. Hahn, D. Homm, K. Krauss, G. Ohler, W. Sörgel, C. Spitzner: Characteristics of Audio Streaming over IP Networks within the ISMA Standard. April 2002, Presented at the 112th AES Convention. Preprint 5514.

[13] H. Schulzrinne, S. Casner, R. Frederick, V. Jacobson: RTP: A Transport Protocol for Real-Time Applications. RFC 3550, Juli 2003.

[14] J. Postel: User Datagram Protocol. RFC 768, Aug. 1980.

[15] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, W. Weiss: An Architecture for Differentiated Services. IETF RFC 2475, Dec. 1998.

[16] S. Floyd, V. Jacobsen: Random Early Detection Gateways for Congestion Avoidance. IEEE/ACM Transactions on Networking **1** (Aug. 1993) 397–413.

[17] J. Heinanen, F. Baker, W. Weiss, J. Wroclowski: Assured Forwarding PHB Group. Internet RFC 2597, June 1999.

[18] J. K. Muppala, T. Bancherdvancih, A. Tyagi: VoIP Performance on Differentiated Services Enabled Network. Proc. of the IEEE International Conference on Networks, Sep. 2000, 419–423.

[19] X. Zhu, M. Chen, Y. Yu: Video Transmission Based on Diffserv over IP networks. Proc. of the International Conference on Communication Technology, ICCT, Apr. 2003, 1754–1757.

[20] J. E. Voldhaug, E. Hellerud, U. P. Svensson: Evaluation of Packet Loss Distortion in Audio Signals. May 2006, Presented at the 120th AES Convention. Preprint 6855.

[21] C. Perkins, O. Hodson, V. Hardman: A Survey of Packet Loss Recovery Techniques for Streaming Audio. IEEE Network **12** (Sep./Oct. 1998) 40–48.

[22] E. Hellerud, J. Voldhaug, U. Svensson: Perceptually Controlled Error Protection for Audio Streaming over IP Networks. Proc. of the International Conference on Digital Telecommunications (ICDT06), 2006.

[23] Y. Wang, A. Ahmaniemi, D. Isherwood, W. Huang: Content-based UEP: A new Scheme for Packet Loss Recovery in Music Streaming. Proc. of ACM Multimedia, Berkeley, CA, USA, Nov. 2-8, 2003, 2003, 412–421.

[24] J. Korhonen, Y. Wang, D. Isherwood: Toward Bandwidth-Efficient and Error-Robust Audio Streaming over Lossy Packet Networks. Multimedia Systems **10** (August 2005) 402–412.

[25] ITU-T Recommendation P.862: Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs. Feb. 2001.

[26] ITU-R Recommendation BS.1387: Method for Objective Measurements of Perceived Audio Quality (PEAQ). 2001.

[27] S. Möller: Assessment and Prediction of Speech Quality in Telecommunications. Kluwer Academic Publishers, USA-Boston, 2000.

[28] ITU-T Recommendation G.107: The E-Model, a Computational Model for Use in Transmission Planning. 2003.

Document downloaded by @DAEL, 2026-11-10 13:09:27 - personal use only

[29] A. Raake: Predicting Speech Quality under Random Packet Loss: Individual Impairment and Additivity with other Network Impairments. Acustica united with Acta Acustica **90** (Nov/Dec 2004) 1061–1083.

[30] A. Raake: Short- and Long-Term Packet Loss Behavior: Towards Speech Quality Prediction for Arbitrary Loss Distributions. Audio, Speech and Language Processing, IEEE Transactions on [see also Speech and Audio Processing, IEEE Transactions on] **14** (2006) 1957–1968.

[31] A. Clark: Modeling the Effects of Burst Packet Loss and Recency on Subjective Voice Quality. Internet Telephony Workshop (IPtel), New York, 2001.

[32] J-sim: http://www.j-sim.org. [Online].

[33] J. McGowan: Burst ratio: A measure of bursty loss on packet-based networks. U.S Patent 6 931 017, August 2005.

[34] E. Larsson, P. Stoica, J. Li: Amplitude Spectrum Estimation for Two-Dimensional Gapped Data. IEEE Transactions on Signal Processing **50** (June 2002) 1343–1354.

[35] H. Ofir, D. Malah: Packet Loss Concealment for Audio Streaming Based on the GAPES Algorithm. May 2005, Presented at the 118th AES Convention. Preprint 6334.

[36] ITU-R Recommendation BS.1534: Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems. 2003.

[37] ITU-R Recommendation BS.1116: Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems. 1997.

[38] G. A. Soulodre, M. C. Lavoie: Subjective Evaluation of Large and Small Impairments in Audio Codecs. Proc. of the 17th AES International Conference, Aug. 1999.

Document downloaded by @DAEL, 2026-11-10 13:09:27 - personal use only