[BS97a] Reinhard Bertram, Ralf Steinmetz: Scalability of Audio Quality for Networked Multimedia Environments, Proceedings of ICMCS'97, Ottawa, Canada, 1997, pp. 294 - 301

Scalability of Audio Quality for Networked Multimedia Environments

Reinhard Bertram, Ralf Steinmetz
Darmstadt University of Technology
Department of Electrical Engineering and Information Technology
Industrial Process and System Communications
Merckstr. 25, D-64283 Darmstadt, Germany
{Reinhard.Bertram,Ralf.Steinmetz}@KOM.tu-darmstadt.de

Abstract

As a result of listening tests performed with users of multimedia applications we present constraints for a dynamic adaptation (scaling) of audio data streams to the available system and network resources. For the speech medium any severe deformation of frequency characteristics should be avoided, while shortening and expanding of speech pauses is suitable to cope with varying delays. For the music medium the best scaling algorithm is a smooth fade between quality levels. Scaling resulting in lower quality should be performed in multiple steps, while in scaling towards higher quality any approach can be chosen. It appears that six levels of quality are sufficient for scaling of music from high quality (CD) to very low quality (cellular phone) and vice versa. In our proposal of a Dynamic QoS Centered Architecture, where the bitrate and scheduling of associated media streams are adapted dynamically, the highest possible QoS for the user has to be taken into account.

1. Introduction

The integration of continuous media (e.g. audio and video) into existing computer environments involves the processing of time-dependent data. In networked multimedia systems various entities typically cooperate in order to provide real-time guarantees for the data presented at the user interface using resource management. These systems provide mechanisms for streams with guaranteed or statistical Quality of Service (QoS) (see, e.g., [10],[15],[17]), and coordinate multiple media streams. Most of the involved mechanisms are developed for a completely error-free presentation of continuous-media data at the user interface. In today's networked environments resources are finite and we still encounter many data paths over networks using communication protocols which are not capable of providing a guaranteed real-time service. In such setups it is a key issue to decide which data items must be presented at the user interface and which may be discarded due to timing constraints.

The QoS defines the properties of media streams. In this paper we distinguish between two layers of QoS: network QoS and user QoS. The former describes, for instance, parameters like bandwidth, end-to-end delay, inter/intra stream synchronization, ordered delivery of data and error recovery. The latter describes requirements for the perception of multimedia data at the user interface.

Multimedia applications typically negotiate a desired QoS during the connection setup phase. The negotiated QoS may be lower than the QoS the network can provide at a specific moment, for example when a connection is established during a peak load period. Without reserved data paths loss of data is likely to occur. If the network or the system is unable to provide the desired QoS, today's multimedia applications often set up an end-to-end connection and transfer the data with a best-effort approach. In both cases it is desirable to adapt a media stream to the available resources.

2. Motivation and related work

Most often an adaptation of media streams to the available bandwidth (scaling) is performed as follows: Whenever the network becomes congested, the data rate of a media stream under control of a resource management system is decreased drastically. When more bandwidth is available the data rate is increased gradually until the desired value is reached or congestion occurs again. Using software codecs, multimedia applications on workstations (including PCs) are often capable of adjusting media quality and data rates over a wide range. This capability allows for dynamic scaling mechanisms, which means to continuously retrieve the state of the network and adjust the network QoS accordingly [1],[2]. The subject of scaling and filtering has most often been discussed with a focus on video data (see, e.g. [3],[4],[16]). The bitrate of a video data stream can be varied by means of many individual

parameters like the frame rate, the video spatial resolution, and the DCT parameter quantization.

One possible implementation of a dynamic QoS control system can be found in [8]. However, most of the proposed scaling mechanisms still do not account for user QoS. Varying video quality may annoy the user of multimedia applications more than a slightly lower quality that remains constant for a longer period of time. It is the purpose of the work outlined in this paper to find out the constraints for the scaling of multimedia data based on the human perception of quality changes.

We focus on the scaling of audio data, which needs to be kept under limited constraints, because "the ear is surprisingly sensitive to sound variations lasting only a few milliseconds. The eye, in contrast, does not notice changes in light level that last only a few milliseconds" [14]. The real-time nature of audio (and video) data streams and the variety of incompatible encoding schemes used for different types of audio makes it hard to find suitable dynamic scaling mechanisms. Most audio encoding algorithms are optimized for certain types of audio signals and some discrete bitrates. Hence, for a significant change of bitrates (for example in the range from cellular phone audio to CD audio) or a change of the audio type the encoding must be changed as well. A rough comparison of some standardized speech encoding schemes is given in Table 1.

Encoding scheme	Bitrate [kbit/s]	Quality [MOS]*	Delay [ms]
PCM: ITU-R Rec. G.711	64	4.10	0.125
ADPCM: ITU-T Rec. G.726	32	3.85	0.125
LD-CELP: ITU-T Rec. G.728	16	3.61 - 3.85	0.625
CS-ACELP: ITU-T Rec. G.729	8	3.92	15
GSM FR: RPE-LTP	13	3.50	20
IS-54: Mobile USA	8	3.54	20
JDC HR: Mobile Japan	3.45	2.82	48

Table 1: Comparison of some standardized speech encoding schemes [11]
* MOS: Mean Opinion Scale, see Table 2

Human perception of aural data varies from person to person. It is based on different notions of good sound quality, musical skills, age, hearing defects and many other issues. Therefore, any experiment to define user QoS must take into account these variations.

The remainder of this text is organized in 4 sections: Section 3 gives a brief overview of subjective assessment techniques. Our experiments and the respective results are presented in Section 4. Implications on scaling mechanisms based on our results follow in Section 5. A glimpse of our implementation of these scaling mechanisms can be found in the proposed Dynamic QoS Centered Architecture outlined in Section 6.

3. Subjective assessment of impairments in audio systems

The main differences between common subjective assessment methods result from the addressed environment of these tests: broadcast or telecommunications. While telecommunication originated tests primarily address speech transmission problems (like listening effort), broadcast originated tests primarily address small impairments caused by audio transmission. Most of these methods focus on *small* impairments. The required effort in time, manpower and expenses during the subjective assessment of impairments in audio systems lead to the development of "objective perceptual audio assessment" methods. Evaluation of proposed objective methods currently takes place within ITU-R working group 10-4. It turns out that objective methods are well suited for the evaluation of impairments near masked thresholds, but not applicable for severely distorted signals [6].

Simplified methods for the subjective assessment of impairments in audio systems were proposed by, e.g., [12]. Those methods still require high effort and are mainly applicable for the development of new audio codecs. They allow a comparison of new codecs to a standard version for which previous listening tests were performed using the methods proposed in the recommended standard [5].

"Conversation-opinion tests" are primarily designed to test telephone equipment in a two-way conversation under laboratory conditions and well-defined background noise, (see, e.g. [7] Annex A). It would be challenging to set up such a test environment for multimedia teleconferences.

"Listening tests" comprise three different test set-ups. In the first method the subjects (listeners) have to rate the presented samples by absolute grades (without any given reference sample). In the second the subjects are asked to do a (most often) pairwise comparison of given samples. The third form uses a "double-blind triple-stimulus with hidden reference" technique, which is a combination of the former two. Examples of the three methods are:

- Absolute Category Rating (ACR): [7] Annex B
  Originating from the former CCITT this test is intended to rate telephony equipment. The speech material is presented without references and the subjects are asked to judge each sample according to a non-continuous listening effort scale or a loudness preference scale, both of which are applicable to multimedia environments.
- Degradation Category Rating (DCR): [7] Annex D
  The stimuli are presented to the listeners by pairs (A-B) or repeated pairs (A-B-A-B) where A is the quality reference sample and B the same sample processed by the system under evaluation. The purpose of the reference sample is to anchor each judgement by the listeners. Some "null pairs" (A-A), at least one for each speaker, are included to check the quality of anchoring. The lis-

teners are instructed to rate the conditions according to a five point non-continuous degradation category scale.

- Reference method [5]:
  The "double-blind triple-stimulus with hidden reference" method has been found to be especially sensitive, to be stable and to permit accurate detection of small impairments. Therefore it should be used for tests of systems generating small impairments.
  In the preferred and most sensitive form of this method, one subject at a time is involved and the selection of one of three stimuli ("A", "B", "C") is at the discretion of this subject. The known reference is always available as stimulus "A". The hidden reference and the object under test are simultaneously available, but are "randomly" assigned to "B" and "C", depending on the trial. In each trial the subjects are asked to rate the perceived difference (if any) between "B" and "A" on the one hand and the difference between "C" and "A" on the other hand using the five-grade scale shown below. Two grades must therefore be given on each trial, one for "B" and one for "C". At least one grade of "5.0" is expected to be given on each trial. The presented continuous scale (Table 2) uses the scores only as anchor points.

Score	Category
5.0	Imperceptible
4.0	Perceptible but not annoying
3.0	Slightly annoying
2.0	Annoying
1.0	Very annoying

Table 2: Five-Grade impairment scale [5] (MOS)

All of the test methods presented above foresee a statistical pre- and post-screening of the subjects to test their reliability. The test setup is typically a laboratory with professional audio equipment and well-defined acoustic characteristics including background noise level.

[9] lists categories of impairments which may occur with digital coding or transmission techniques. Some of these are applicable only to high quality audio while others are subject to be investigated in our context of multimedia applications. For example, our tests were performed with forced impairments from the following categories: quantization defect, deformation of frequency characteristic, extra sound, missing sound. Hence, we did not assess the impairment of sound quality by single artifacts but a degradation of overall sound quality in discrete large steps.

For severely impaired audio no reference assessment method exists, so we decided to merge some of the proposed methods without requiring the well defined listening conditions. Our tests were intended to find constraints on the perception of *noticeable* changes in audio quality, which are not going to be masked by higher background noise or slightly different listening conditions for individual subjects. Changes in audio quality may result from changing audio encoding algorithms, but not necessarily result in a change of the *perceived* audio quality.

The measurement of an aural impression still should take into account the listening environment, the program material (the excerpts) as well as the experience and expectation of the subjects.

4. Experiments and results

As a first step we defined the target application domain to be networked multimedia applications running on workstations and mobile equipment utilized by users with typically no specific musical skills.

Considering the target environment we decided to use only two different excerpts for the investigation: Contemporary music and speech. The specific excerpts (see Table 3) were selected by a selection panel in which one of the authors and one of the later subjects participated.

description	source
Contemporary Jazz/Pop	OPCD-6003
Male German Speech	DG 435 460-2

Table 3: Program excerpts used during the tests

To define the specific set of parameters to be investigated we ran a set of preliminary experiments where many parameters were varied. These preliminary experiments showed:

- The tests need to take place in a low to medium noise environment in order to maintain subject concentration during the sessions.
- The total session time should not exceed 1hour for untrained subjects.
- Six levels of quality are sufficient in the range from CD to mobile cellular phone-(GSM in Europe) like quality for the purpose of our investigations.
- It is important to select people with different aural capabilities (musical background/knowledge).
- It is better to keep the sound quality constant for a minimum time of ten seconds to allow for an adaptation of the ear. Shorter times seem to confuse the subjects, while longer periods unnecessarily lengthen the time for each experiment.
- One of the major influencing factors related to the grades (given by the subjects) is the playback device used for the test. In a typical multimedia application environment one cannot expect to find studio reference equipment, so we used different devices mixed randomly between the subjects under test. A list of the specific devices can be found in Table 5.

Quality	HighPassFilter	LowPassFilter	Filter Level	Quantization	Limiter	NoiseReduction
Q1; CD	none	none	0.0 dB	none	none	none
Q2; MM-PC	210 Hz 18 dB	4800 Hz 18 dB	+3.4 dB	none	none	none
Q3; Telephone	300 Hz 18 dB	3400 Hz 18 dB	+5.2 dB	none	none	none
Q4; Radio	500 Hz 18 dB	2000 Hz 18 dB	+8.4 dB	none	none	none
Q5; Mini-Radio	560 Hz 36 dB	1800 Hz 36 dB	+12.4 dB	none	none	none
Q6 approximately GSM-like	830 Hz 36 dB	1200 Hz 36 dB	+18.0 dB	8bit dithering: type1 shaping: ultra	-0.1 dB threshold -10dB	broadband –18 dB; att: 5 ms rel: 2 ms; smoothing 80% High shelving: 1200 Hz -6 dB

Table 4: Levels of audio quality used during the final tests

device	description	environment
cassette recorder	medium quality children's cassette player	low to medium surround noise; office
high quality Multimedia PC device	Sun Ultra1 with audio card connected to a HiFi amplifier and speakers	low to medium surround noise; office
HiFi	typical home HiFi equipment	low surround noise

Table 5: Playback Devices used during the tests

- For the purpose of possible later enhancements it is better to use a continuous scale only, as the five-grade non-continuous scale showed no significant difference.
- Different hearing positions for individual subjects show no significant influence, especially when listening to severely impaired audio data.

We have chosen the lowest level of quality to be cellular phone-like (GSM), because in an integration of mobile phones into networked multimedia systems, audio data originating from mobile phones will be coded in such a low quality. Each session consisted of samples we created in a professional recording studio by applying various digital filters to different sections of the given excerpt. For a list of the quality levels used during the experiments, see Table 4.

Prior to the beginning of the experiment we introduced the sessions by explaining the purpose of the tests and the scale (Table 2), which was translated to German as in [12]. We also instructed the subjects to concentrate on the *change in quality* as opposed to the quality itself. Most often multiple subjects participated in a listening session. The loudness of the reproduction was fixed for a complete session and could not be changed by the subjects. The samples for each session were played back sequentially with a 5 sec gap in between.

We did not check the reliability of the subjects, as a post screening of all the subjects using a t-test method as proposed by [5] is only useful when the members of the listening panel are expert listeners. The results are presented using *diffgrades* which are the grades rated by the subjects minus the maximum grade of 5.0 (as proposed by [5]). Hence, 0.0 in the plots corresponds to 5.0 in the impairment scale. The mean diffgrades and the range in which the subjects rated is plotted for each trial in the session. For the jazz/pop sessions no. 3 to no. 5 a plot of the sequences in the time domain and the applied changes in quality can be found in the lower part of the result plots for each of the respective results. The different levels of sound quality refer to the levels listed in Table 4.

4.1 Session 1: Impairment of speech quality

Session 1 was performed using a modified hidden reference method. The subjects were asked to use two scales for each trial – a listening effort scale ([7] Annex B) and a continuous impairment scale (Table 2) for audio quality. The stimuli were presented to the listeners by pair (A-B) where A was the sample under test and B was the quality reference sample. Each sample (male German speech) had a total playing time of 140 sec. Sample A contained multiple severe impairments using quality levels listed in Table 4. No quantization effects were added, but an additional quality level of pitch shifting by 2/3 octave down was introduced to hide the speaker identity. These quality levels switched randomly between words, while no switch was made in the middle of single talk spurts. The subjects were asked to answer an additional question related to how many speakers they believed they listened to. As this effect is only obtainable when no reference was heard before, we could use only one trial to address this specific problem.

The results (shown in Fig. 1 and Fig. 2) prove that speaker identity (mainly based on the frequency characteristics) is a very important factor in terms of speech quality. More than half of the subjects were not able to determine that there was only one speaker and not multiple speakers.

Figure 1: Quality grades session 1

Any quality change resulting in a deformation of frequency characteristic should be avoided when scaling speech.

The absolute diffgrades are not comparable to the other sessions, because in this trial the complete range of quality levels was not known to the subjects.

4.2 Session 2: Speech pause variation

Session 2 used a modified hidden reference method. The stimuli (male German speech) were presented to the listeners on repeated pairs (A-B-A-B) where A contained a variation of speech pauses and B was the quality reference sample. The variation of speech pauses consisted of a randomly chosen mixture of dilation and shortening adding to a total time of the sum of the original pauses. The audio quality level was kept constant. We found it very difficult to determine a fixed acceptable length for a pause, a relative factor is more suitable than using fixed lengths. The maximum pause shortening was a total removal of the pause, while the maximum pause dilation was chosen to be 4.0 times the pause of the original for pauses during the normal media flow. The absolute shortening and dilation of pauses should not exceed 2 seconds. In our preliminary test experience, the factor of 4.0 was indicated as the maximum acceptable dilation for an experienced listener, so untrained listeners, like the subjects which formed the basis of these tests, would tolerate them as well. The same scales as in session 1 were used.

The variation of speech pauses proved as an alternative approach to scale speech in the time domain. In this context it is interesting that the first (degraded) trial of this session received the highest average rate but also the lowest minimal rates (see leftmost area in Fig. 3). This happened due to the fact that some of the subjects *expected* a quality impairment without actually hearing one. The grades of the test of this session show that only a few subjects noticed a variation of the speech pauses (see Fig. 3 and Fig. 4).

Figure 2: Listening effort grades session 1

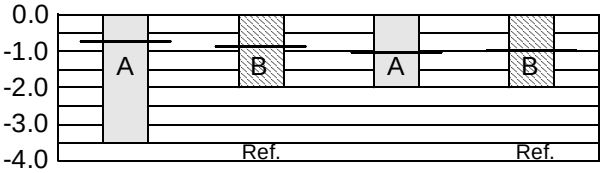

Figure 3: Quality grades session 2

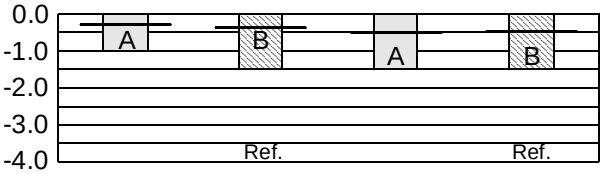

Figure 4: Listening effort grades session 2

Although we used a minimum pause shortening factor of 0.0 we recommend that speech pause shortening smaller than by a factor of 0.25 should be avoided. For speech pause dilation we recommend prolonging the original pause no more than by a factor of 4.0 and never exceeding 2 seconds as the absolute duration.

4.3 Session 3: Direction of music quality scaling

For the jazz/pop sessions we used a modified ACR method. In the context of judging *changes in quality* and because it is unsuitable to define a reference quality change, an ACR method is the only possible test method.

The subjects were instructed to rate each complete sequence as a unit and not each change in quality by itself on a continuous scale.

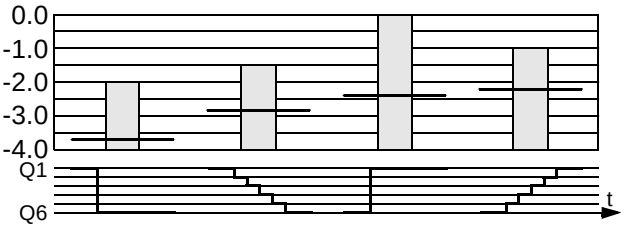

Figure 5: Results session 3

Session 3 consisted of four 80sec sequences where the subjects had to rate each of them individually. We tried to check for differences between an increase and a decrease in quality. All candidates rated an increase in quality better than a decrease (see Fig. 5). This is independent of the kind of increase (multiple-step or single-step). The situation is different when the quality is decreased: A multiple-step approach is preferred over a single-step approach.

4.4 Session 4: Variation of music scaling method

The main question behind the fourth session relates to the usefulness of multiple steps and a comparison with today's alternative (single-step) and the smooth fading approach when the quality is decreased and increased again. The session consisted of three 140sec sequences. Here the subjects had (first) to rank each sequence after

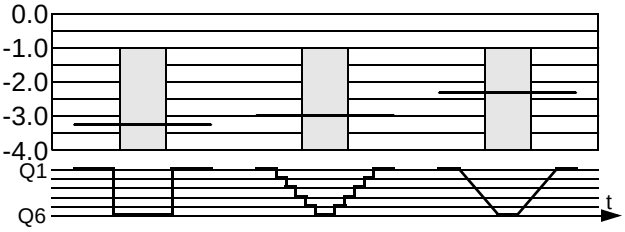

Figure 6: Results from session 4

playback of all three sequences and (second) to rate them individually using the impairment scale. This was used to anchor the judgements for this session relative to each other.

Fading turned out to be the best alternative in any case (see Fig. 6). Approximately half of the subjects preferred changes in quality via the multiple-step over the single-step approach, while the other half graded vice versa. From discussions with the subjects we derived the following: The first group prefers a smoother transition, while the second feels disturbed by many changes in quality.

4.5 Session 5: Discrete levels of music quality

The final jazz/pop session was intended to verify our decision of using six quality levels and to answer questions about preferences in the direction of jumps in quality. The start level of a single jump was lowered from CD quality to cellular phone-like quality and back again. The height of the jumps was increased until either the minimum or the maximum quality level was reached. This adds to a total of 30 samples for this session – each with a duration of 35 sec. The subjects were asked to rate every single change in quality as soon as the sample was played back. The results for decreasing quality are plotted in Fig. 7 and for increasing quality in Fig. 8.

In general all candidates rated larger changes in quality worse than changes by smaller steps and decreases in quality were rated worse than increases. This indicates that the subjects not only rated the changes in quality but also the absolute quality level. For the increase of quality there is no significant difference whether a transition occurs between 2 or more steps. This result is interesting, because session 3 showed that an increase in quality by the multiple-step and the single-step approach are similar in terms of subjective quality perception. Our interpretation is that smaller steps are preferable if the quality is expected to change often, while a large jump is preferred when the higher level of quality is expected to stay constant for a longer time.

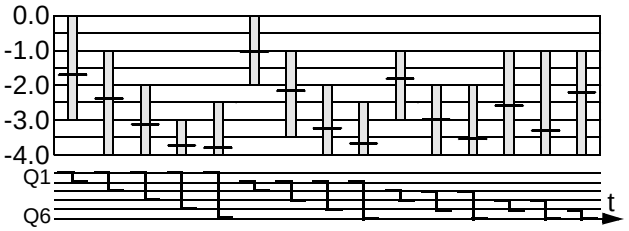

Figure 7: Results for session 5 (downwards)

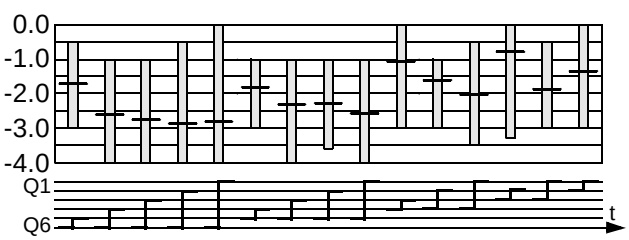

Figure 8: Results for session 5 (upwards)

5. Implications

Analyzing the results one can clearly state that they were influenced by the expectations of the listeners and the actual equipment used. Most candidates who listened to the portable device gave better grades than those who listened to high quality equipment. It was revealed that the subjects tend to accept a variation of quality up to 3 levels below the level of quality of which the playback equipment is capable. System and network QoS should be restricted to this lower limit, i.e. during the connection setup phase a minimum QoS should be reserved and adaptation should only lead to media quality above this lower limit.

In most cases the subjects still noticed the difference of quality of adjunct steps. However, the change is sufficient to have such a step defined as another level of quality. This is true for music using the full-range audio spectrum, while for speech any changes in the quality level, which result in a deformation of frequency characteristic should be avoided. The results show that an increase in the number of steps makes quality changes during the transition from one step to an adjunct step less noticeable.

User QoS may be raised if audio coding is scalable to a higher extent than it is today, such that application and system software may control the audio data rate dynamically. It would also be helpful to have smooth transitions defined for changing the data rate. This would allow activation of fading capabilities and thereby enhance the audio user interface.

The comparison of increase and decrease in quality showed that a multimedia scaling system is, in general, free to choose the most suitable way and timing to adapt the QoS of the respective media stream whenever quality is going to be enhanced. Given the results from session 3 and 4 this is true as long as a decrease in quality is not expected to occur less than 10 sec after the last increase; a situation where a single-step approach is preferable. However, any decrease in quality should be performed in smaller multiple steps. Applying this result to today's Internet technology is of no advantage for the communication and operating systems as the reaction to any congestion would only be a smooth decrease of the load. The TCP protocol uses the opposite approach with its well-known slow-start and fast-stop behavior, which makes it (for other reasons too) unfavorable for the transfer of audio data.

The fading approach results in higher grades compared to a multiple-step approach of changing quality; so this scheme should be used whenever scaling of audio data is

needed. The same is expected for video data. Most of today's audio encoding schemes are optimized for specific bitrates; new encoding schemes, which allow scaling over a wide range need to be developed. Further research is certainly needed.

Sophisticated codecs exist for speech, hence it seems unnecessary to use a modification of frequency characteristics to reduce the bitrate of speech data. A field for further study is to recognize the content of an audio stream and to adapt the encoding accordingly.

A better user QoS might also be achieved by delaying audio streams containing speech to a certain extent until more data can be presented at the user interface. The maximum delay is determined by constraints resulting from our experiments (inter stream synchronization and the type of data transmitted). This allows the development of implementations to cope with varying network delays. A first snapshot of how such an implementation may look is shown in the next section. For reasons of inter stream synchronization it is often a bad approach to add extra delays, but in cases where no such constraints are violated it would add an alternative to catch up varying network delays.

6. Dynamic QoS Centered Architecture

Our results of human perception related to the changes of QoS for audio described above allow the development of a scaling range with certain constraints. Without the development of new audio codecs capable of wide-range bitrate adaptation the proposed approach to change the bitrate of audio-streams containing music over a wide range is to switch the audio encoding scheme on-the-fly. A Dynamic QoS Centered Architecture needs therefore codecs processing audio streams which are self-contained; i.e., they need to make use of headers in which the type of encoding and the actual parameters for reproduction are listed.

Fig. 9 shows an example of one audio track (with 2 shown segments of information) and a second track with one media segment. The playback shall occur according to a well defined schedule. In this case at a certain point in time an over-utilization of resources may occur leading to errors at presentation time.

So far scaling is understood as the possibility to reduce or expand the capacity resources for a certain point in time. Scaling in terms of time normally refers to decrease the

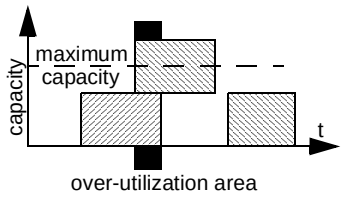

Figure 9: Example of resource over-utilization

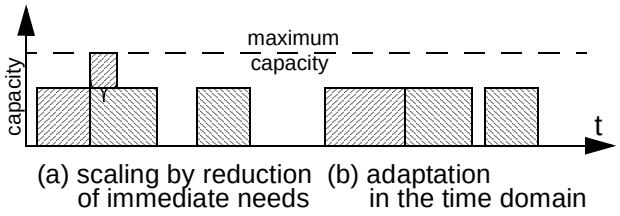

Figure 10:Reaction to resource over-utilization

granularity of data in the time domain. For example video data can be scaled by reducing the frame rate.

Fig. 10a shows how traditional scaling occurs according to the needs and possibilities of the system environment; i.e. the bandwidth may be reduced.

Content-based scaling with the ability of scaling in the bandwidth-utilization and time domain leads to a new possibility of reducing network load. Given that the content is known to each instance (from sender to receiver), scaling is possible at the point an over-utilization of resources is recognized with only little loss of data because there is no need to ask the sender (in the form of e.g. choke packets) to slow down. With the exploitation of shortening and dilation of silence (speech) intervals, i.e. the possibility to shift data in the time domain, the resource management may additionally cope with time-dilation as a new dimension for scalability. Fig. 10b shows an example where by shifting of a speech segment of the first track in the time domain the overall schedule can be met without any further decrease of QoS.

The exploitation of such an adaptation is no problem when the scheduler knows the occurrence of talk spurts and silence intervals in advance. In this case all behavior is predictable and can be pre-computed.

The situation is different when an application allows for interactivity and provides mechanisms to change a schedule on-the-fly. At the moment of interaction the data to be displayed is selected and exactly at this point in time the schedule can be computed for the path recently selected.

In a scenario with live media data, the system can in general neither predict the advent of a silence interval nor determine the duration ahead. However, by the introduction of an artificial delay (see the parameter threshold in the algorithm on the following page) and a detection of a silence interval of some reasonable minimal duration (see parameter some-time-back) we can make use of our scaling approach in the time domain.

This algorithm has to be performed by a scaling manager instance which controls the "regulator" (or "stream adaptor") as shown in Fig. 11. The regulator delivers the scaled data to the communication, manipulation and/or presentation components which are restricted in their total capacity. It is an issue of further study how these architectural components should be best placed in an overall sys-

tem (instead of just having them located near the sender of the data stream). The following algorithm shows the operation in more detail:

```
DEFINE:
  threshold          = minimal value which allows to activate the
                       algorithms to change the duration
                       e.g., 1ms would not make sense as any
                       computation would not lead to significant
                       improvements of QoS
                       e.g., 20ms would be a suitable lower value to
                       activate the QoS scheduling algorithm
  some-time-back     = meaningful minimal duration to activate the
                       algorithms in a predictive mode
                       e.g. 5ms silence may lead to a dilation of 20ms
                       which may be helpful for QoS scheduling
COMPUTE:
  add-on-delay       = compute (possible artificial additional
                       end-to-end delay to be introduced based on
                       what the QoS specifies and agreed upon
                       before)
  IF add-on-delay > threshold
    insert add-on-delay
    DO for-ever
      pass-data UNTIL silence-detected (now)
      determine-schedule-opportunity    (delay-mode &
                                         skip-ahead-mode)
      IF time-jump-to-be-done
        SKIP-in-time-domain (delay-or-skip-ahead-value)
    END DO for-ever
  ELSE
    DO for-ever
      pass-data UNTIL silence-detected (some-time-back)
      determine-schedule-opportunity (delay-mode)
      IF time-jump-to-be-done
        SKIP-in-time-domain (delay-value)
    END DO for-ever
  END IF add-on-delay > threshold
```

7. Conclusion

We understand this set of conducted experiments as a means of leading to new ideas of what, when and how to scale audio data in networked multimedia systems. In particular multimedia tools and services need to be enhanced by components to handle scaling of different media in different ways in order to achieve best quality for the end user instead of only avoiding congestion or achieving maximum throughput. We are currently integrating these algorithms into our proposed Dynamic QoS Centered Architecture.

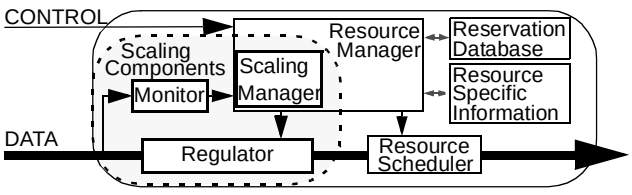

Figure 11: Scaling integrated with a resource management system like [15]

Acknowledgments

We want to acknowledge all of our test subjects for their patience and excellent comments. In particular, Lars Wolf, Martin Karsten and Jennifer Wilson provided many detailed comments. Thanks to SchokoPro, Wiesbaden, Germany, we were able to record and edit all the excerpts on high end professional audio equipment. This work is sponsored in part by Volkswagen-Stiftung, D-30519 Hannover, Germany.

References

[1] I. Busse, B. Deffner and H. Schulzrinne: Dynamic QoS control of multimedia applications based on RTP, Computer Communications, Vol. 19, No. 1, 1996, pp. 49 – 58

[2] J.-C. Bolot, T. Turletti and I. Wakemann: Scalable feedback control for multicast video distribution in the Internet, Proceedings of SIGCOMM, London, UK, 1994, pp. 58 – 67

[3] L. Delgrossi, C. Halstrick, D. Hehmann, R.G. Herrtwich, O. Krone, J. Sandvoss, C. Vogt: Media Scaling with HeiTS, ACM Multimedia Systems, Vol. 2, No. 4, 1994, pp. 172 – 180

[4] C. Gonzales, E. Viscito: Flexible Scalable Digital Video Coding, Signal Processing: Image Communication, Vol. 5, No. 1-2, February 1993, pp. 5 – 20

[5] ITU-R: Methods for the Subjective Assessment of Small Impairments in Audio Systems including Multichannel Sound Systems, ITU-R, Recommendation BS.1116, 1994

[6] ITU-R: Comparison of Output Parameters Used in the Perceptual Quality Evaluation of Audio Signals, ITU-R, Doc 10-4/15-E, 1995

[7] ITU-T: Methods for Subjective Determination of Transmission Quality, ITU-T, Recommendation P.80, 1993

[8] K. Kawakuni, H. Tokuda: Dynamic QoS control Based on the QoS-Ticket Model, Proceedings of ICMCS 96, Hiroshima, pp. 78 – 85

[9] ISO/MPEG: Report on the MPEG/Audio Multichannel Formal Subjective Listening Tests, ISO/IEC JTC1/SC29/WG11 NO685 MPEG94/063, 1994, page 5

[10] K. Nahrstedt and R. Steinmetz: Resource Management in Networked Multimedia Systems, IEEE Computer, Vol. 28, No. 5, 1995, pp. 52 – 63

[11] G. Schröder: Sprachcodierung für digitale Übertragungssysteme mit niedriger Kanalkapazität, Technical Report of Deutsche Telekom AG, Highlights aus der Forschung, 1996, pp. 6 – 11

[12] W. H. Schmidt, E. Steffen, U. Wüstenhagen: Report on Simplified Methods for the Subjective Assessment of Small Impairments in Audio Systems, Technical Report of Deutsche Telekom AG, FZ 144, Berlin, 1995

[13] R. Steinmetz and K. Nahrstedt: Multimedia: Computing, Communications & Applications, Prentice Hall, 1995

[14] A. S. Tanenbaum: Computer Networks, Prentice Hall, 1996

[15] L. C. Wolf: Resource Management for Distributed Multimedia Systems, Kluwer, 1996

[16]L. C. Wolf, R. G. Herrtwich and L. Delgrossi: Filtering Multimedia Data in Reservation-based Networks, Kommunikation in Verteilten Systemen, (Ed.) K. Franke, U. Hübner. W. Kalfa; Springer Verlag, 1995, pp. 101 – 112

[17]L. Zhang, S. Deering, D. Estrin, S. Shenker, and D. Zappala: RSVP: A new Resource ReSerVation Protocol, IEEE Network, Vol. 7, No. 5, 1993, pp. 8 – 18