0[Stei94e] *Ralf Steinmetz;* **Human Perception of Audio-Visual Skew;** Kapitel im Buch zu ausgewählten Beiträgen des Dagstuhl-Seminars "Hochgeschwindigkeitsnetze und Kommunikationssysteme" 1994, Herausgeber: W.Effelsberg, A.Danthine, O.Spaniol, Kluwer-Verlag.

HUMAN PERCEPTION OF AUDIO-VISUAL SKEW

Ralf Steinmetz

IBM European Networking Center
Vangerowstraße 18 • 69115 Heidelberg • Germany

ABSTRACT

Multimedia synchronization comprises the definition and the establishment of temporal relationships among audio, video, and other data. The presentation of 'in sync' data streams by computers is essential to achieve a natural impression. If data is 'out of sync', the human perception tends to identify the presentation as artificial, strange, or annoying. Therefore, the goal of any multimedia system is to present all data without perceptible synchronization errors. The achievement of this goal requires a detailed knowledge of the synchronization requirements at the user interface. This paper presents the results of a series of experiments about human media perception of skew mainly related to audio and video data. It leads to a first guideline for the definition of a synchronization quality of service. The results show that a skew between related data streams may still let data appear 'in sync'. It also turned out that the notion of a synchronization error highly depends on the types of media.

1 INTRODUCTION

We understand multimedia according to [1] [2]; a multimedia system is characterized by the integrated computer-controlled generation, manipulation, presentation, storage, and communication of independent discrete and continuous media. The digital representation of any data and the synchronization between various kinds of media and data are the key issues for integration. Multimedia synchronization is needed to ensure a temporal ordering of events in a multimedia system.

At a first glance this ordering applies to single data streams: a stream consists of consecutive logical data units (LDUs). In the case of an audio stream, LDUs may be individual samples transferred together

from a source to one or more sinks. A video LDU typically corresponds to a single video frame and consecutive LDUs have to be presented at the sink with the same temporal relationship as they were captured at the source leading to intrastream synchronization.

The temporal ordering also applies to related data streams. The most often discussed relationship is the simultaneous playback of audio and video with 'lip synchronization'. Both kinds of media must be 'in sync', otherwise the viewer would not be satisfied with the presentation. In general an interstream synchronization involves relationships between all kind of media including pointers, graphics/images, animation, text, audio, and video. In the following, 'synchronization' always means interstream synchronization.

For delivering multimedia data correctly at the user interface, synchronization is essential. Unlike other notions of correctness, it is not possible to provide an objective measurement for synchronization. As human perception varies from person to person, only heuristic criteria can determine whether a stream presentation is correct or not. This paper presents our results of some extensive experiments related to human perception of synchronization between different media.

To reach the goal of an error-free data delivery, audio, video, and other data are often multiplexed (i.e. physically combined in one data unit) and, hence, synchronized at the source and demultiplexed just before presentation at the sink. Multiplexing is not always possible and wanted, e.g., because multimedia data needs to go through different routes in a computing system. The separate handling of previously related data leads to time lags between the media streams. These lags have to be adjusted at the sink for 'in sync' presentation.

Some work on how to implement multimedia synchronization was done in related projects [3] [4] [5] [6] [7] [8]. Work has also been devoted to define synchronization requirements [9] [10] [11] [12] [13]. It is often reported that audio can be played up to 120 ms ahead of video and in the reverse situation video can be displayed 240 ms ahead of audio. Both temporal skews will sometimes be noticed, but can easily be tolerated without any inconvenience by the user [14]. Some authors report a skew of +/-16 ms [15] or no skew at all to be acceptable.

Implementing our own synchronization mechanisms, we were unable to draw the right conclusions from these reports - their statements were contradictory. There was a lack of an in-depth analysis of synchroniza-

tion between the various kind of media and, in particular, for lip synchronization. We decided to conduct our own study and to explore these fundamental issues to obtain results that allow us to quantify the quality of service requirements for multimedia synchronization.

The remainder of this text is organized into six sections. Section 2 outlines the main results of lip synchronization experiments, the notion of the 'quality of synchronization' is elaborated in Section 3. Section 4 describes the test strategy, how the results were achieved including influencing factors. Remaining types of media synchronization are discussed in Section 5. Section 6 provides a comprehensive overview of all types of media skew in terms of the required quality of service parameters.

2 EXPERIMENTAL SET-UP

'Lip synchronization' denotes the temporal relationship between an audio and a video stream where speakers are shown while they say something. The time difference between related audio and video LDUs is known as the 'skew'. Streams which are perfectly 'in sync' have no skew, i.e., 0 ms. We conducted experiments and measured which skews were perceived as 'out of sync' for audio and video data. In our experiments, users often mentioned that something is wrong with the synchronization, but this did not disturb their feeling for the quality of the presentation. Therefore, we additionally evaluated the tolerance of the users by asking if the data out of sink affects the quality of the presentation.

In several discussions with experts working with audio and video, we noticed that most of the personal experiences were derived from very specific situations. As an immediate consequence we have been confronted with a wide range and tolerance levels up to 240 ms. A comparison and a general usage of these values is doubtful because the environments from which they resulted were incomparable. In some cases we encountered the 'head view' displayed in front of some single color background on a high resolution professional monitor. In another set-up a 'body view' was displayed in a video window at a resolution of 240*256 pixels in the middle of some dancing people. In order to get the most accurate and stringent affordable skew tolerance levels, we selected a speaker in a TV news environment as a 'talking head' (see Figure 2). In this scenario, the viewer is not disturbed by background information. The user is attracted by the gestures, eyes, and lip

movement of the speaker. We selected a speaker who makes use of gestures and articulates very accurately.

We recorded the presentation and then played it back in our experiments with artificially introduced skew that was adjusted according to the frame rate, i.e., n times 40 ms (derived from the European TV standard) which was introduced by professional video editing equipment. We conducted some experiments with a higher resolution time scale by cutting the material with the help of a computer where it was possible to introduce a smaller delay in the audio stream. It turned out that there was no need for any test with higher granularity than 40 ms.

Figure 1 Left: Head View, Middle: Shoulder View, Right: Body View[1]

We expected a relationship between the detectable skew and the actual size of the head displayed at the monitor. As shown in Figure 1 we selected three different views of the speaker. At the very close 'head view' the head completely fills the screen, the 'shoulder view' shows the head as well as the shoulders while the third 'body view' captures the whole person sitting in a room.

Lip synchronization usually applies to speech as an acoustic signal related to its visual representation of the speaker. We expand this notion to cover the correlation between noise and its visual appearance, e.g., clapping. For the latter, our experiments included a person working with a hammer and some nails. The most exhaustive study, however, was performed in the news environment.

Figure 2[2] provides an overview of the main results. The vertical axis denotes the relative amount of test candidates who detected a synchronization error, regardless of being able to determine if audio was before or after the video. As one might expect, if the skew is relatively small

1. This is just an outline of the different views, the quality of the original clips is TV-like.
2. In all figures, negative skew denotes 'video ahead of audio', while positive skew means 'video being behind audio'.

most of the people did not notice it; large skews became obvious. However, our initial assumption was that the three curves related to the different views would be very different. However as shown in Figure 2 this is not the ease.

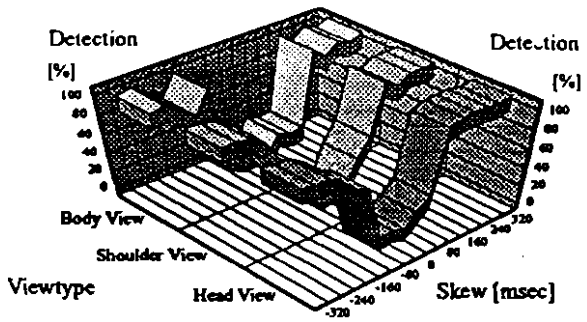

Figure 2 Detection of synchronization errors with respect to the three different views.

Figure 3 shows these curves in detail. A careful analysis provides us with information regarding the asymmetry, some periodic ripples and minor differences between the various views.

The left side of the figure relates to negative skew values, where video is ahead of audio. In our daily life, we experience this situation whenever we talk to some distant located person. All three curves are, in general, flat in this region. Since we are not accustomed to hearing speech ahead of the related visual impression, the right side of the curves turns out to be steeper.

The 'body view' curve is broader than the 'head view' curve, at the 'head view' a small skew was easier to notice. This was more difficult in the 'body view'. The 'head view' is also more asymmetric than the 'body view'. Basically, the further away we are situated, the less noticeable the error is.

At a fairly high skew, the curves show some periodic ripples. This is more obvious in the case of audio being ahead of video. It means that some people had difficulties in identifying the synchronization error even with fairly high skew values. A careful analysis of this phenomenon lead to the following explanation; At the relative minima, the speech signal was closely related to the movement of the lips which tends to be quasi periodic. Errors were easy to notice at the start, at the end, at the borders of pauses, and whenever changing drastically the

mood (e.g., from an explanation style to a sudden aggressive comment). Errors were more difficult to notice in the middle of sentences. A subsequent test containing video clips with skews according to these minima (without pauses and not showing the start, the end, and changes of mood) caused problems in identifying if there was indeed a synchronization error.

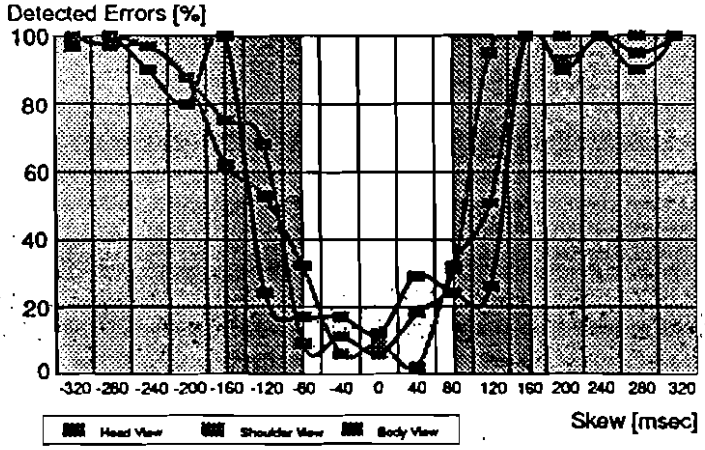

Figure 3 Areas related to the detection of synchronization errors

The main results of about 100 test participants are captured in the different areas of Figure 3:

- The **'in sync' area** spans a skew of between -80 ms (audio after video) and +80 ms (audio ahead of video). In this zone most of the users did not detect the synchronization error. Very few mention that if there is an error it does affect their notion of quality video. Additionally, we had some results where test candidates mentioned that the perfect 'in sync' clip (skew = 0ms) is 'out of sync'. Therefore, we introduced a range of uncertainty in the graph which captures these types of inconsistencies. We came to realize that lip synchronization tolerates the above mentioned skew, this result applies to any type of lip synchronization.
- The **'out of sync' areas** span beyond a skew of -160 ms and +160 ms. Nearly everybody detected these errors and was dissatisfied with the clips. Data delivered with such a skew is in general not acceptable. Additionally, often a distraction occurred; the viewer/listener became more attracted by this 'out of sync' effect than by the content itself.
- In the **'transient' area** where *audio is ahead of video*, the closer

the speaker is, the easier errors are detected and reported as disturbing. The same applies to the overall resolution; the better the resolution is, the more obvious the lip synchronization errors became.

- A second **'transient' area** where *video is ahead of audio* is characterized by a similar behavior as the other transient area as long as the skew values are near the in sync area. The closer the speaker is, the more obvious the skew is. Apart from this effect we noticed that video ahead of audio can easier be tolerated than the vice versa.

This asymmetry is very plausible: In a conversation where two people are located 20 m apart, the visual impression will always be about 60 ms ahead of the acoustics due to the faster light propagation compared to the acoustic wave propagation. We are just more used to this situation than to the previous one.

Concerning the different areas, we got similar results with the noise and video experiment (hammer with nails) although the transient areas are more narrow. In this experiment, the type of view had a negligeable influence. The presentation of some violinist in a concert and a choir did not show more stringent skew demands than the speaker being synchronized.

A comparison between sets of experiments ran in English and German showed no difference. There might be, however, a problem inherent with the test candidates: as Germans are used to watch synchronized films and movies, they could be less sensitive to synchronization errors than, e.g. Americans. However some minor experiments with English, Spanish, Italian, French and Swedish presented always by native speakers to native speakers verified that the specific language has almost no influence on the results.

We did not find any variation between groups of participants with different habits regarding the amount of TV and films usually watched.

3 QUALITY OF SKEW VALUES

Figure 3 outlines the perception of synchronization errors. More important than just to notice the error is the effect of such an 'out of sync' video clip on the human perception. If in an extreme case all peo-

ple tend to like audio data to be, e.g., 40 ms ahead of video, we should take it into account. Therefore the test candidates were asked to qualify a detected synchronization error in terms of being acceptable, indifferent, or annoying. From these answers we derived the 'level of annoyance' which quantifies the quality of synchronization.

Figure 4 shows by which degree a skew was believed to be acceptable or intolerable. We used the 'talking head' experiment and depict here the 'shoulder view' as it is a compromise between the 'head' and the 'body view'.

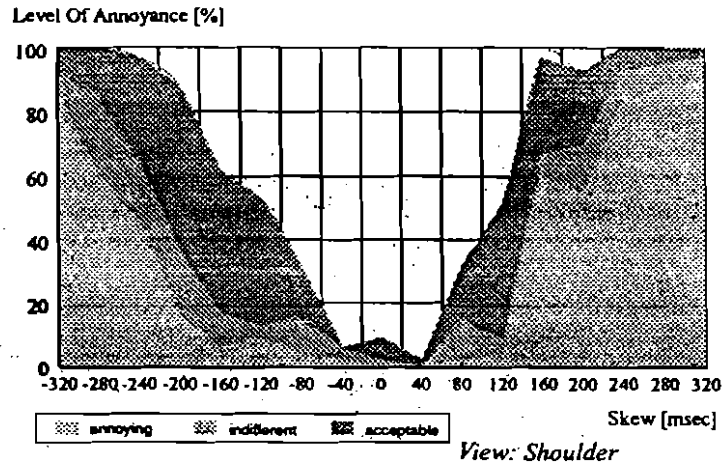

Figure 4 Level of annoyance at shoulder view

The dark grey areas relate to all test candidates who would accept to listen to and watch video with this synchronization error. In a small follow-on experiment we selected a few test candidates who would tolerate such a skew and presented them a whole movie with this synchronization error. We chose a skew of -160 ms (video ahead of audio). They did not complain and very soon concentrated on the content instead of being attracted by looking for some synchronization problem. The curve at the bottom of the dark grey area shows an obvious asymmetry which occurs due to the more natural acceptance of visual perception being ahead of related acoustic impression.

The light grey area relates to all people who really dislike this skew and were distracted by it. It also contains the asymmetry discussed above. During the evaluation phase of this study on synchronization, we introduced a skew of +80 ms and -80 ms into two whole movies. These movies were shown to a few candidates who mentioned that

such a skew is annoying. It turned out after a short discussion that, if we introduced this artificially (or if we cheated), they did not object at all. The same experiment with a skew of -240 ms or +160 ms would lead to a real distraction from the content and to a severe feeling of annoyance.

This evaluation of the level of annoyance provides a furth :r argument for allowing the skew of lip synchronization to take values between -80 ms and +80 ms as mentioned in the former section.

4 TEST STRATEGY

For each person, the lip synchronization test took approximately 45 minutes. The experiment was intentionally carried out with the same audio and video over and over again. This led to some concentration problems during the whole test, which were alleviated by introducing breaks.

We always ran all tests related to one view in one session. Then, the second and the third view were shown in their sessions. The order of the sessions had no effect. Individual samples, each having a different skew, were shown randomly.

Initial experiments showed that a total length of about 30s with a small subsequent break is sufficient for getting the users impression. All experiments with longer video clips did not provide any additional or different results. With some test candidates, who were more experienced with video technology and synchronization issues, 5s turned out to be sufficient. Nevertheless we decided to use 30s for each sample.

The background of all scenes was static (i.e., not moving) and out of focus in order to keep the distraction to a minimum. We focussed on the detection of synchronization errors in the most challenging set-ups, this allowed the determination of skew values independently from the actual content of the video and audio data. In these experiments the viewer should never have been distracted by the background.

The same consideration, i.e. background vs. foreground, can be applied to the audio data. The voice of the speaker can be mixed with some background noise or music. In order to differentiate between foreground and background, the volume of the speaker should be at least twice the volume of the background audio. In contrast to the video

analogy discussed in the previous paragraph, any background audio did not influence our results.

The group of people was selected according to an equal distribution of sex and ages. To have a representative distribution we did not take into account habits (like the time spent for watching TV) and the social status or any other characteristics of the test candidates.

It would have been very interesting if - before presenting each sample - the candidates were not aware of the fact that we were looking for synchronization issues. As soon as the test candidates noticed the first time a synchronization fault, they would not have been allowed to continue the experiment with further skews. This would have led to results for casual unexperienced users. As a matter of fact, we started to run the experiment in this way with very few people. It turned out that lip synchronization is not detected so easily leading to a broader range of the 'in sync' zone.

In order to provide results for building multimedia systems for all types of users, we have to make the assumption that a user can also make frequent use of such a system and interact for a longer time with the application. Therefore, the results of users being aware of possible synchronization faults provide the correct basis.

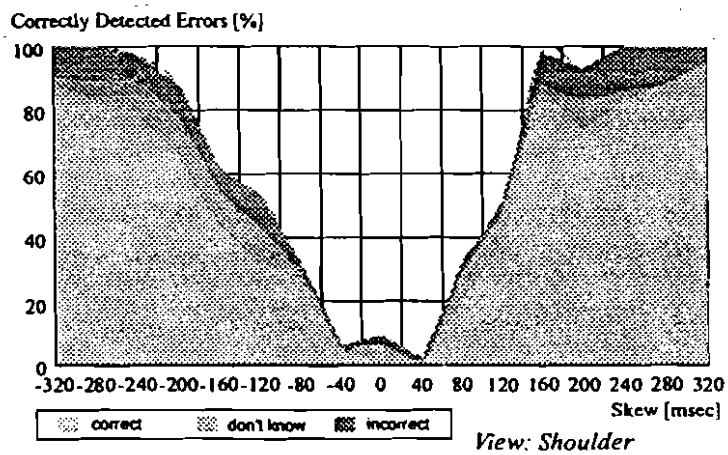

Figure 5 Correct detection of the perceived synchronization error

For the purpose of double checking, the candidates were asked to define exactly which type of synchronization error they noticed. It is easier just to detect that something is wrong than to precisely state if

audio is ahead of video or vice versa. Figure 5 summarizes the results of correct perception of the skew in the 'shoulder view' scenario.

Near the error-free synchronization value (at 0 ms) it was difficult to determine the type of skew, as soon as the values ranged beyond -40 ms or +40 ms almost everybody provided correct answers.

5 MEDIA DEPENDENCY OF SKEW

Lip synchronization was investigated by us because of the contradictory results available from other sources. In the following, we will summarize other synchronization results which we found provided less diverse statements. We do this to arrive at a more complete picture of synchronization requirements.

Since the beginnings of digital *audio* the ('affordable') *jitter* and the jitter to be tolerated by dedicated hardware has been studied. In discussions with Dannenberg, he provided us some references and the following explanations of these studies: In [17] the maximum allowable jitter in a sample period for at 16 bit quality audio is mentioned to be 200ps, this is explained as the error equivalence to the magnitude of the LSB (least significant bit) of a full-level maximum-frequency 20KHz signal. In [18] some perception experiments recommend an allowable jitter in an audio sample period between 5 and 10 ns. Further perception experiments were carried out by [19] and [20], the maximum spacing of short clicks to obtain fusion into a single percept was mentioned to be 2ms (as cited by [21]).

In a computer-supported cooperative work (CSCW) environment, cameras and microphones are attached to the users' workstations. The audio and video data of one participant is simultaneously presented at the other remote workstation(s), e.g., we assumed the issue of the discussion is a business report including data related to some graphics. All participants have a window with these graphics on their desktop where a shared telepointer is used in the discussion. With the pointer, speakers point to individual elements of the graphics which they are referring to while speaking simultaneously. This requires synchronization of *audio and the remote telepointer*. Using the same margins as in our lip synchronization experiments, the 'in sync' area related to audio ahead of pointing is 750 ms and for pointing ahead of audio it is 500 ms [25]. In most of the daily occurring discussions using a telepointer

these results can be relaxed. This zone allows for a clear definition of the 'in sync' behavior regardless of the content.

The combination of *audio and animation* is usually not as stringent as lip synchronization. A multimedia course on dancing, e.g., comprises the dancing steps as animation with the respective music. By making use of the interactive capabilities, individual sequences can be viewed and listened to over and over again. In this example the synchronization between music and animation is particularly important. Experience shows that a skew of +/- 80 ms fulfills all user demands even though some jitter may occur. Nevertheless, the most challenging issue is the correlation between a noisy event and its visual representation, e.g. a simulated crash of cars. Here we encounter the same constraints as for lip synchronization, +/- 80 ms.

Two audio tracks can be tightly or loosely coupled, the effect of related audio streams depends heavily on the content:

- A stereo signal usually contains information about the location of the sources of audio and is *tightly coupled*. The correct processing of this information by the human brain can only be accomplished if the phases of the acoustic signals are delivered correctly. This demands for a skew less than the distance between consecutive samples leading to the order of magnitude of 20 µs. [22] reports that the perceptible phase shift between two audio channels is 17µs. This is based on a headphone listening experiment. Since a varying delay in one channel causes the sound's source location apparently to move, Dannenberg proposed to allow an audio sample skew between stereo channels within the boundaries of +/- 11µs. This is derived from the observation that a one-sample offset at a sample rate of 44kHz can be heard.
- *Loosely coupled* audio channels are a speaker and, e.g., some background music. In such scenarios we experience an affordable skew of 500 ms. The most stringent loosely coupled configuration has been the playback of a dialogue where the audio data of the participants originate from different sources. The experienced acceptable skew was 120 ms.

The combination of *audio with images* has its initial application in slide shows. By intuition a skew of about 1s arises which can be explained as follows [15]: Consider that it takes a second or so to advance a slide projector. People sometimes comment on the time it

takes to change transparencies on an overhead projector, but rarely worry about automatic slide projectors.

A more elaborated analysis leads to the time constraints equivalent to those of pointer synchronization. The affordable skew decreases as soon as we encounter *music* played in correlation *with notes* for, e.g., tutoring purposes. [15] points out that here an accuracy of 5 ms is required: Current practice in music synthesizers allows delays ranging up to 5 ms, but jitter is less than total delay. A 2 ms number refers to the synchronization between the onset times of two nominally simultaneous notes or the timing accuracy of notes in sequence, see also [23] [21] [24].

The synchronized presentation of *audio with some text* is usually known as audio annotation in documents or, e.g., part of an acoustic encyclopedia. In an existing 'music dictionary', an antique instrument is described and is played simultaneously. An example for a stronger correlation is the playback of a historical speech of, e.g. J.F. Kennedy with simultaneous translation into German text. This text is displayed in a separate window and must relate closely to the actual acoustic signals. Karaoke systems are another good example of necessary audio and text synchronization.

For this type of media synchronization the affordable skew can be derived from the duration of the pronunciation of short words which last approximately 500 ms. Therefore the experimentally verified skew of 240 ms is acceptable

The synchronization of *video and text* or *video and image* occurs in two different fashions:

- In the *overlay mode*, the text is often an additional description to the displayed moving image sequence. In a video of playing billiard, the image is used to denote the exact way of the ball after the last stroke. The simultaneous presentation of the video and the overlaid image is important for the correct human perception of this synchronized data. The same applies to a text which is displayed in conjunction with the related video images: Instead of having the subtitles always located at the bottom, it is possible to place text close to the respective topic of discussion. This would cause an additional editing effort at the production phase and may not be for the general use of all types of movies but, for tutoring purposes some short text near by the topic of discussion is very useful. In such overlay schemes, this text must be synchronized to the video in order to assure that it is placed at the

correct position. The accurate skew value can be derived from the minimal required time. A single word should appear on the screen in order to be perceived by the viewer: 1 s is certainly such a limit. If the media producer wants to make use of the flash effect, then such a word should be on the screen for at least 500 ms. Therefore, regardless of the content of the video data we encounter 240 ms to be absolutely sufficient.

- In the second mode *no overlay* occurs, skew is less serious. Imagine some architectural drawings of medieval houses being displayed in correlation with a video of these building: While the video is showing today's appearance, the image presents the floor plan in a separate window. The human perception of even simple images requires at least 1 s, we can verify this value with an experiment with slides: the successive projector of non-correlated images requires about 1 s, as the interval between the display of a slide and the next one in order to catch some of the essential visual information of the slide. A synchronization with a skew of 500 ms (half of this mentioned 1 s value) between the video and the image or the video and text is sufficient for this type of application.

Sometimes *video* is combined *with animation* as there may be a film where some actors become animated pictures. But, for the following short reasoning of synchronization between video and animation let us go back to the example of a video showing the stroke of a billiard ball and the image of the actual 'route' of this ball. Instead of the static image, the track of the ball can be followed by an animation which displays this route at the time the ball is moving on the table. In this example any 'out of sync' effect is immediately visible. In order for humans to be able to watch the ball with the perception of a moving picture, this ball must be visible in several consecutive adjacent video frames at a slightly different position: a good result can be achieved, if in every 3 subsequent video frames, the ball moves by the distance of it's diameter. Less frames will result in the problem of visibility of what occurs, e.g., in tennis, and it may lead to difficulties with the notion of continuity. Derived from this number of 3 subsequent frames, we allow the equivalent skew of 120 ms to occur. This is very tight synchronization, and we have not found any practical requirement which cannot be handled with this value of the affordable skew.

Multimedia systems also incorporate the real-time processing of *control data* and the presentation of this data using various media. A tight timing requirement occurs if the person has to react to this displayed

data. No overall timing demand can be stated as these issues highly depend on the application itself.

6 QUALITY OF SERVICE

The control of synchronization in distributed multimedia systems requires a knowledge of the temporal relationship between media streams. The result of this study is of service to this management component. Synchronization requirements can be expressed by a quality of service (QoS) definition. This QoS parameter defines the acceptable skew within the involved data streams, it defines the affordable synchronization boundaries. The notion of QoS is well established in communication systems; in the context of multimedia, it also applies to local systems. If the video data is to be presented simultaneously to some audio and both are stored as different files or as different entries in a database, lip synchronization according to the above mentioned results has to be taken into account.

In this context we want to introduce the notion of *presentation* and *production level synchronization*:

- *Production level synchronization* refers to the QoS to be guaranteed prior to the presentation of the data at the user interface. It typically involves the recording of synchronized data for a subsequent playback. The stored data should be captured and recorded with no skew at all, i.e. totally "in sync". This is of particular interest if the file is stored in an interleaved format applying multiplexing techniques. Imagine a participant of an audio-video conference who additionally records this audiovisual data to be played back later for a remote spectator. At the conference participant's site, the actual incoming audiovisual data is 'in sync' according to the defined lip synchronization boundaries. Let the data arrive with a skew of +80 ms and let audio and video LDUs be transmitted as a single multiplexed stream over the same transport connection. It is displayed to the user and directly stored on the harddisk (still having this skew). Later on, this data is presented simultaneously at a local workstation and to the remote spectator. For a correct data to be deliverable, the QoS should be specified as being between -160 ms and 0 ms. At the remote viewer's station - without this additional knowledge of the actual skew - it might turn out that by applying these boundaries twice, data is not 'in sync'. In general, any syn-

chronized data which will be further processed should be synchronized according to a production level quality, i.e. with no skew at all.

- The whole set of experiments discussed in this report identifies *presentation level synchronization*, it focuses on the human perception of synchronization and defines whatever is reasonable at the user interface. As shown in the above paragraph, by recording the actual skew as part of the control information, the required QoS for synchronization can be easily computed. Therefore, in advanced systems, data may also be recorded 'out of sync' leading to an 'in sync' presentation.

Media		Mode, Application	QoS
video	animation	correlated	+/- 120 ms
	audio	lip synchronization	+/- 80 ms
	image	overlay	+/- 240 ms
		non overlay	+/-500 ms
	text	overlay	+/- 240 ms
		non overlay	+/-500 ms
audio	animation	event correlation (e.g. dancing)	+/- 80 ms
	audio	tightly coupled (stereo)	+/- 11 µs
		loosely coupled (dialogue mode with various participants)	+/- 120 ms
		loosely coupled (e.g. background music)	+/- 500 ms
	image	tightly coupled (e.g. music with notes)	+/- 5 ms
		loosely coupled (e.g. slide show)	+/- 500 ms
	text	text annotation	+/- 240 ms
	pointer	audio relates to showed item	-500 ms, + 750 ms[a]

Table 1 Quality of Service for synchronization purposes

a. pointer ahead of audio for 500 ms, pointer behind audio for 750 ms[25]

The required QoS for synchronization is expressed as the allowed skew. The QoS values shown in Table 1 relate to presentation level synchronization. Most of them result from exhaustive experiments and experiences, others are derived from the literature as referenced in the paper. To our understanding, they serve as a general guideline for any QoS specification. During the lip synchronization experiment we learned that there are many factors such as the distance of a speaker which to some extend influence these result. We understand that this whole set of QoS parameters as first order result to serve as a general guidance. These values may be relaxed using the knowledge on the actual content.

7 OUTLOOK

Synchronization QoS parameters allow the builders of distributed multimedia and communication systems to make use of the affordable tolerances. In our Heidelberg multimedia system, the HeiRAT component [26] is in charge of the resource management. HeiRAT accepts QoS requests from the applications and serves for this QoS demands as interface to the whole distributed system. It makes use of the flow specification of the ST-II multimedia internetwork protocol to negotiate them among the whole set of involved system components [27]. It provides a QoS calculation by optimizing one QoS parameter dependent on the resource characteristics. Subsequently resources are reserved according to the QoS guarantees. At the actual data transfer phase, resources are scheduled according to the provided guarantees.

Synchronization is a crucial issue of multimedia systems. In local systems it is often easy to provide because there are sufficient resources or it is a single user configuration. In networked systems we encounter a plethora of concurrent processes making use of the same scarce resources. A skew between media easily arises.

This paper provides a set of quality of service values for synchronization. It is a feather in our cap to reach results for wide range of media synchronization with extensive user interface experiments. As the next step, we currently run experiments related to the affordable jitter in continuous media presentations. The enforcement of which remains to be a different item which already has been addressed by several systems. The presentation of audio and video, according to some logical time system is one of the possible solutions.

ACKNOWLEDGMENTS

First of all I would like to acknowledge the enthusiastic work done by Clemens Engler: We spent hours and nights of controversial discussions on the expected results, the influencing factors and the design of the experiments. Clemens Engler also carried out most of the experimental work. Wieland Holfeider helped in producing the basic video material, and we like to acknowledge the patience and accuracy of all our test candidates. Roger Dannenberg, CMU Pittsburgh, provided many valuable hints concerning jitter of audio samples and synchronization related to music. Ralf Guido Herrtwich substantially commented the final version of the paper. Thank you.

REFERENCES

[1] Ralf Steinmetz, Ralf Guido Herrtwich: *Integrated Distributed Multimedia-Systems*, Informatik Spektrum, Springer Verlag, vol.14, no.5, October 1991, pp.280-282.

[2] Ralf Steinmetz, *Multimedia-Technology: Fundamentals (in German: "Multimedia Technologie: Einführung und Grundlagen")*, Springer-Verlag, September 1993.

[3] David P. Anderson, George Homsy: *Synchronization Policies and Mechanisms in a Continuous Media I/O Server*, International Computer Science Institute, Technical Report no. 91-003, Berkeley, 1991.

[4] Gerold Blakowski, Jens Huebel, Ulrike Langrehr, Max Muhlhaeuser: *Tools Support for the Synchronization and Presentation of Distributed Multimedia*, computer communications, vol. 15, no. 10, December 1992.

[5] L.Li, A. Karmouch, N.D. Georganas, *Synchonization in Real Time Multimedia Data Delivery*, IEEE ICC'92, Chicago, USA, June 1992. pp. 322.1.

[6] L.Li, L. Lamont, A. Karmouch, N.D. Georganas: *A Distributed Synchronization Control Scheme in A Group-oriented Conferencing Systems*, Proceedings of the second international conference, Broadband Islands, Athens, Greece, June 15-16, 1993

[7] Doug Shepherd, Michael Salmony: *Extending OSI to Support Synchronisation Required by Multimedia Applications*, Computer Communications, vol.13, no.7, September 1990, pp. 399-406.

[8] Ralf Steinmetz: *Multimedia Synchronization Techniques: Experiences Based on Different System Structures*, IEEE Multimedia Workshop '92, Monterey, April 1992.

[9] Thomas D.C. Little, Arif Ghafoor: *Network Considerations for Distributed Multimedia Objects Composition and Communication*, IEEE Network Magazine, vol. 4 no. 6, November 1990, pp. 32-49.

[10] Thomas D.C. Little, A. Ghafoor: *Synchronization and Storage Models for Multimedia Objects*, IEEE Journal on Selected Areas in Communication, vol. 8, no. 3, Apr. 1990, pp. 413-427.

[11] Cosmos Nicolaou: *An Architecture for Real-Time Multimedia Communication Systems*, IEEE Journal on Selected Areas in Communication, vol. 8, no. 3, April 1990, pp. 391-400.

[12] Kaliappa Ravindran: *Real-time Synchronization of Multimedia Datastreams in High Speed Packet switching Networks*, in Workshop on Multimedia Information Systems (MMIS '92), IEEE Communications Society, Tempe, AZ, February 1992.

[13] Ralf Steinmetz: *Synchronization Properties in Multimedia Systems*, IEEE Journal on Selected Areas in Communication, vol. 8, no. 3, April 1990, pp. 401-412.

[14] Alan Murphy: *Lip Synchronization*, personal communication on a concerning set of experiments, 1990.

[15] Roger Dannenberg: *Personal communication on sound effects and video synchronization and on music play back and visualization of the corresponding strokes*, 1993.

[16] Harald Rau, personal communication, 1993.

[17] Barry Blesser: *Digitization of Audio: A Comprehensive Examination of Theory, Implementation, and Current Practice*, Journal of the Audio Engineering Society, JAES 26(10), October 1978, pp. 739-771.

[18] T. Stockham: *A/D and D/A Converters: Their Effect on Digital Audio Fidelity,* in Digital Signal Processing, L. Rabiner and C. Rader, (Eds.), IEEE Press, NY 1972.

[19] J.C.R. Licklider: *Basic correlates of the auditory stimulus,* in S. S. Stevens, ed. Handbook of Experimental Psychology, Wiley, 1951.

[20] H. Woodrow: *Time Perception,* in S. S. Stevens, (Ed.) Handbook of Experimental Psychology, Wiley, 1951.

[21] Dean Rubine, Paul McAvinney: *Programmable Finger-tracking Instrument Controllers,* Computer Music Journal, vol. 14, no. 1, Spring 1980, pp. 26-41.

[22] Roger Dannenberg, Richard Stern: *Experiments Concerning the Allowable Skew of Two Audio Channels Operating in the Stereo Mode,* personal communication, 1993.

[23] M. Clynes: *Secrets of Life in Music: Musicality Realized by Computer* in Proceedings of the 1984 International Computer Music Conference, San Francisco, International Computer Music Association, 1985.

[24] M. Stewart: *The Feel Factor: Music with Soul,* Electronic Musician, vol. 3, no. 10, pp. 55-66, 1987.

[25] Ralf Steinmetz, Clemens Engler: *Human Perception of Media Synchronization,* IBM Technical Report 43.9310, IBM European Networking Center, Heidelberg, 1993.

[26] Carsten Vogt, Ralf Guido Herrtwich, Ramesh Nagarajan: HeiRat: *The Heidelberg Resource Administration Technique Design Philosophy and Goals,* IBM Technical Report 43.9307, IBM European Networking Center, Heidelberg 1992.

[27] Luca Delgrossi, Ralf Guido Herrtwich, Frank Oliver Hoffmann: *An Implementation of ST-II for the Heidelberg Transport System,* IBM Technical Report 43.9303, IBM European Networking Center, Heidelberg, 1993