Chapter 15

Synchronization

15.1 Introduction

Advanced multimedia systems are characterized by the integrated computer-controlled generation, storage, communication, manipulation and presentation of independent time-dependent and time-independent media [Ste93b, HS91]. The key issue which provides integration is the digital representation of any data and the synchronization of and between various kinds of media and data.

The word synchronization refers to time. *Synchronization* in multimedia systems refers to the temporal relations between media objects in the multimedia system. In a more general and widely used sense some authors use synchronization in multimedia systems as comprising content, spatial and temporal relations between media objects. We differentiate between time-dependent and time-independent media objects. A time-dependent media object is presented as a media stream. Temporal relations between consecutive units of the media stream exist. If the presentation durations of all units of a *time-dependent media object* are equal, it is called *continuous media object.* A video consists of a number of ordered frames; each of these frames has a fixed presentation duration. A *time-independent media object* is any kind of traditional media like text and images. The semantic of the respective content does not depend upon a presentation according to the time domain.

Synchronization between media objects comprises relations between time-dependent media objects and time-independent media objects. A daily example of synchronization between continuous media is the synchronization between the visual and acoustical information in television. In a multimedia system, the similar synchronization must be provided for audio and moving pictures. An example of temporal relations between time-dependent media and time-independent media is a slide show. The presentation of slides is synchronized with the commenting audio stream. To realize a slide show in a multimedia system, the presentation of graphics has to be synchronized with the appropriate units of an audio stream.

Synchronization is addressed and supported by many system components including the operating system, communication system, databases, documents and even often by applications. Hence, synchronization must be considered at several levels in a multimedia system.

The operating system and lower communication layers handle single media streams with the objective to avoid jitter at the presentation of the units of one media stream (e.g., [NS95c, DHH94, PZF94, OT93]). For example, users will be annoyed if an audio presentation is interrupted by pauses or if clicks result in short gaps in the presentation of the audio clip.

On top of this level is the run-time support for the synchronization of multiple media streams is located (e.g., [AH91b, CGCH92, AC91, IBM92b]). The objective at this level is to maintain the temporal relations between various streams. In particular the skew between the streams must be restricted. For example, users will be annoyed if they notice that the movement of the lips of a speaker does not correspond to the presented audio.

The next level holds the run-time support for the synchronization between time-dependent and time-independent media together with the handling of user interactions (e.g., [MHE93, Bla92, KG89, Lit93]). The objective is to start and stop the presentation of the time-independent media within a tolerable time interval, if some previously defined points of the presentation of a time-dependent media object are reached. The audience of a slide show is annoyed if a slide is presented before the audio comment introduces a new picture. A short delay after the start of the introducing comment is tolerable or even useful.

The *temporal relations* between the media objects must be specified. The relations may be specified implicitly during capturing of the media objects, if the goal of a presentation is to present the media in the same way as they were originally captured. This is the case of audio/video recording and playback.

The temporal relations may also be specified explicitly in the case of presentations that are composed of independently captured or otherwise created media objects (e.g., [BHLM92, BZ93b, IBM90]). In the slide show example, a presentation designer selects the appropriate slides, creates an audio object and defines the units of the audio presentation stream where the slides have to be presented. Also, the user interactivity may be part of a presentation and the temporal relations between media objects and user interactions must be specified. The tools that are used to specify the temporal relations are located on top of the previous levels.

In recent years, in nearly every multimedia workshop and conference, many synchronization-related contributions have been provided. Most of the contributions address only issues of one or a subset of the levels or regard synchronization only from a specific viewpoint and they are partly overlapping.

The objective of this chapter is to provide an integral view to the area of multimedia synchronization. Therefore, we focus on a consistent definition of synchronization-related terms, synchronization requirements, the synchronization specification, synchronization between media objects and the synchronization related to structuring of multimedia systems. An emphasis is also put on the synchronization in a distributed environment that introduces additional complexity but is very important regarding client/server architectures and future teleservices like access to information bases using an information highway.

Descriptions of low-level technical support for media synchronization, like EDF (Earliest Deadline First) or rate monotonic scheduling in the operating system, isochronous transport services and support for single media streams, are not part of this chapter.

In Section 15.2, the basic terms of synchronization are defined. Subsequently, in Section 15.3, the requirements for synchronization resulting from user perception of multimedia presentations are described. A synchronization reference model is

presented in Section 15.4 that allows the structuring of the levels of synchronization and the classification of existing synchronization systems. Section 15.5 provides an overview about synchronization specification methods. Some prominent and representative systems are presented and classified according to the synchronization reference model in Section 15.6. A summary and outlook are given in the last section.

15.2 Notion of Synchronization

15.2.1 Multimedia Systems

Several definitions for the terms multimedia application and multimedia systems are described in the literature. Three criteria for the classification of a system as a multimedia system can be distinguished: the number of media, the types of supported media and the degree of media integration.

The most simple criterion is the number of media used in an application. Using only this criterion, even a document processing application that supports text and graphics can be regarded as a multimedia system [HKN89]. This is not, however, our definition of multimedia (see Chapter 2, Section 2.3).

The types of supported media are an additional criterion [HS91]. In this case, we distinguish between time-dependent and time-independent media. A time-independent media object is usually presented using one presentation unit. An example is a bitmap graphic. Time-dependent media objects are presented by a sequence of presentation units. An example is a motion picture sequence without audio (i.e., a video sequence) presented frame after frame. Because the integration of time-dependent media objects is a new and essential aspect in information processing, some authors define a multimedia system as a system that supports the processing of more than one medium with at least one time-dependent medium [BB91a].

The degree of media integration is the third criterion [HS91]. In this case, integration means that the different types of media remain independent but can be processed and presented together.

Combining all three criteria, we propose the follow g *definition of a multimedia system*: a system or application that supports the integrated processing of several media types with at least one time-dependent medium.

Figure 15.1 classifies applications according to the three criteria. The arrows indicate the increasing degree of multimedia capability for each criterion.

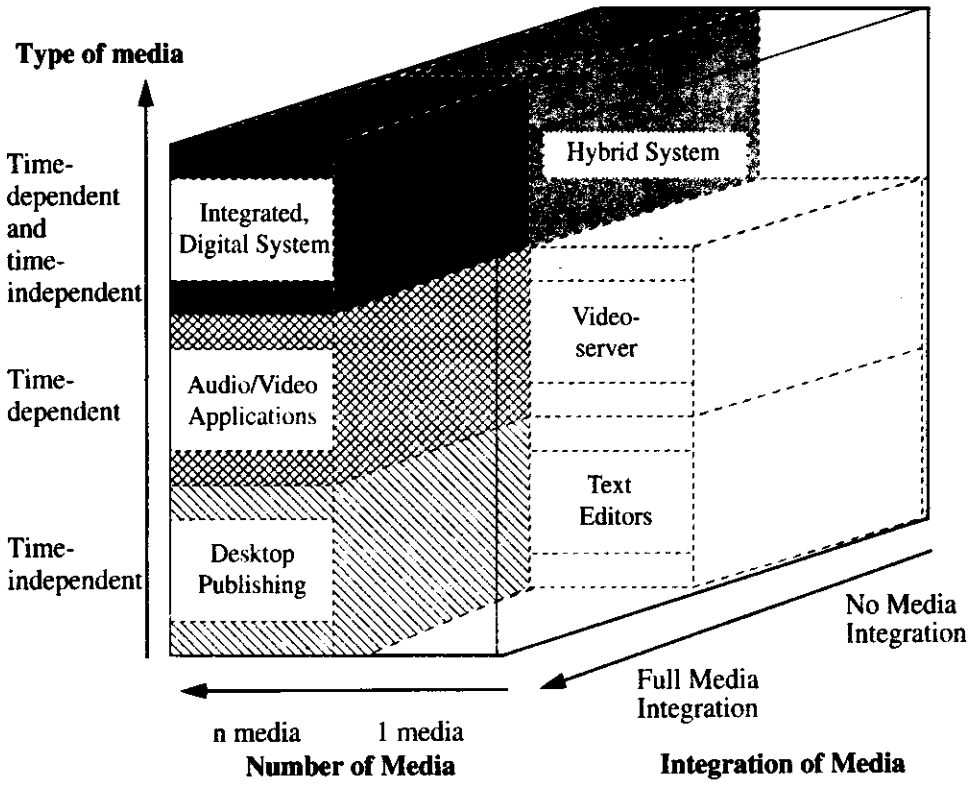


Figure 15.1: *Classification of media use in multimedia systems.*

Integrated digital systems can support all types of media, and due to digital processing, may provide a high degree of media integration. Systems that handle time-dependent analog media objects and time-independent digital media objects are called ***hybrid systems*** [SSSW89, HSA89]. The disadvantage of hybrid systems is that they are restricted with regard to the integration of time-dependent and time-independent media, because, for example, audio and video are stored on different

devices than time-independent media objects and multimedia workstations must comprise both types of devices. The same applies to the interconnection between workstations. Audio/Video-applications that implement functionalities of consumer devices, like video recorders, often do not support the integration of audio and video and time-independent media objects. In addition, they often do not support the separate handling of audio and video media objects. Single time-dependent media objects are often supported by audio and video servers. Traditional desktop-publishing systems are examples of integrated processing of time-independent media objects.

15.2.2 Basic Synchronization Issues

Integrated media processing is an important characteristic of a multimedia system. The main reasons for these integration demands are the inherent dependencies between the information coded in the media objects. These dependencies must be reflected in the integrated processing including storage, manipulation, communication, capturing and, in particular, the presentation of the media objects.

The word *synchronization* refers to time. In a more general and widely used sense, some authors use synchronization in multimedia systems as comprising content, spatial and temporal relations between media objects.

Content Relations

Content relations define a dependency of media objects from some data.

An example of a content relation is the dependency between a filled spreadsheet and a graphic that represents the data listed in the spreadsheet. In this case, the same data are represented in two different ways. Another example is two graphics that are based on the same data but show different interpretations of the data.

For integrated multimedia documents, it is useful to express these relations explicitly to enable an automated update of different views of the same data. In this case, only the data are edited and for the views, the kind of dependencies of the data and the presentation rules are defined. All views of the data are generated automatically and cannot be edited directly. An update of the data triggers an update of the

related views. This technique is, for example, used in database systems and may also be used for the different media of a multimedia system.

In general, the implementation of content relations in multimedia systems is based on the use of common data structures or object interfaces that are used to present objects using different media.

Spatial Relations

The *spatial relations* that are usually known as layout relationships define the space used for the presentation of a media object on an output device at a certain point of time in a multimedia presentation. If the output device is two-dimensional (e.g., monitor or paper), the layout specifies the two-dimensional area to be used.

In desktop-publishing applications, this is usually expressed using *layout frames.* A layout frame is placed and a content is assigned to this frame. The positioning of a layout frame in a document may be fixed to a position in a document, to a position on a page or it may be relative to the positioning of other frames.

The concept of frames can also be used to specify where the presentation units of a time-dependent media object are placed. For example, video frames may be positioned using layout frames. In window-oriented systems, a frame or group of frames may be represented by a window. A window may be resized, moved, iconified, etc. and gives the user additional manipulation freedom to adapt the presentation to the requirements.

Experimental three-dimensional output devices like holographic experiments and three-dimensional projection allow the user to create three-dimensional presentations. In usual window systems, the third dimension is only expressed in terms of overlapping windows. Stereo audio output devices also have layout that defines the positioning of an audio source in a presentation. This is, for example, used in audio and video conferences to give a participant the impression of a seat ordering [LPC90] that is related to the placement of pictures or videos of the other conference participants. This gives the user a more natural communication impression, makes it easier to follow a discussion and therefore, increases the user's acceptance.

Temporal Relations

Temporal relations define the temporal dependencies between media objects. They are of interest whenever time-dependent media objects exist.

An example of temporal relations is the relation between a video and an audio object that are recorded during a concert. If these objects are presented, the temporal relation during the presentations of the two media objects must correspond to the temporal relation at the recording moment.

These time relations are what we will understand to be synchronization in multimedia systems.

Comment

All three types of synchronization relations are important for integrated digital multimedia systems and are meanwhile subject to standardization efforts like MHEG [MHE93] and HyTime [Org92].

Content and spatial relations are well-known from publishing and integrated application systems with databases, spreadsheets, graphical tools and word processing systems. The key aspect in multimedia systems is the temporal relations derived from the integration of time-dependent media objects. Therefore, the rest of this chapter addresses temporal relations only.

15.2.3 Intra- and Inter-object Synchronization

We distinguish between time relations within the units of one time-dependent media object itself and time relations between media objects. This separation helps to clarify the mechanisms supporting both types of relations, which are often very different.

- *Intra-object synchronization*: intra-object synchronization refers to the time relation between various presentation units of one time-dependent media object. An example is the time relation between the single frames of a video

sequence. For a video with a rate of 25 frames per second, each of the frames must be displayed for 40 ms. Figure 15.2 shows this for a video sequence presenting a bouncing ball.

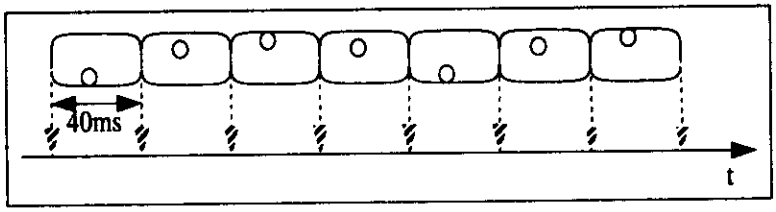


Figure 15.2: *Intra-object synchronization between frames of a video sequence showing a jumping ball.*

- *Inter-object synchronization*: inter-object synchronization refers to the synchronization between media objects. Figure 15.3 shows an example of the time relations of a multimedia synchronization that starts with an audio/video sequence, followed by several pictures and an animation that is commented by an audio sequence.

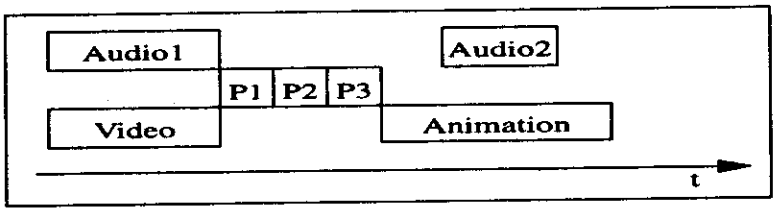


Figure 15.3: *Inter-object synchronization example that shows temporal relations in a multimedia presentation including audio, video, animation and picture objects.*

Time-dependent Presentation Units

Time-dependent media objects usually consist of a sequence of information units. Such information units are known as *Logical Data Units (LDUs)*.

In many cases, several granularity levels of LDUs in a media object exist. An example is the symphony "The bear" by Joseph Haydn (Figure 15.4). It consists of four

movements: vivace assai, allegretto, minuet and finale. Each of these movements is an independent, self-closed part of a composition. It consists of sequences of notes for different instruments. In a digital system, each note is a sequence of sample values. In the case of CD-quality with PCM coding without compression, a sample rate of 44100 Hz with two channels and 16-bit resolution per channel is used. On a CD, these sample values are combined to blocks of 1/75 s duration.

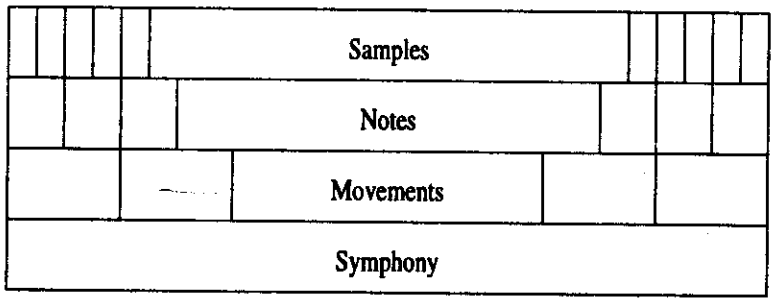


Figure 15.4: *The LDU hierarchy.*

The level of granularity used is application-dependent. It is possible to look at the whole symphony, the movements, the notes, the samples or the combined samples as LDUs. The selection of the LDUs depends on the operations that should be performed on a media object. For simple presentation operations like "play," the whole symphony or the movements are the useful LDUs. For applying instrument-based playing techniques, the notes as smallest description units in the musical area are the useful granularity. For the signal processing task, the operations are based on samples or blocks of samples.

Another example is an uncompressed video object that is divided into scenes and frames. The frames can be partitioned in areas of 16 × 16 pixels. Each pixel consists of luminance and chrominance values. All these units are candidates for LDU units.

In a video sequence coded in the MPEG [ISO93a] format, redundancies within subsequent frames may be used to reduce the amount of digital data used to represent the media object (inter-frame compression). In this case, a sequence of frames that are inter-frame compressed can be regarded as LDU.

The levels of granularity imply a hierarchical decomposition of media objects. Often

there are two kinds of hierarchies. The first is a content hierarchy that is implied by the content of the media object. This is the hierarchy of symphony, movement and notes in the symphony example. The second is the coding hierarchy based on the data encoding. For the symphony example, the hierarchy may be a media object representing a movement, that is divided into blocks of samples. The samples are the lowest level of the coding hierarchy.

In addition, LDUs can be classified into closed and open LDUs. *Closed LDUs* have a predictable duration. Examples are LDUs that are parts of stored media objects of continuous media like audio and video, or stored media objects with a fixed duration. The duration of *open LDUs* is not predictable before the execution of the presentation. Open LDUs typically represent input from a live source, for example, a camera or a microphone, or media objects that include a user interaction.

Classification of Logical Data Units

For digital video, often the frames are selected as LDUs. For example, for a video with 30 pictures per second, each LDU is a closed LDU with a duration of 1/30 s (Figure 15.5).

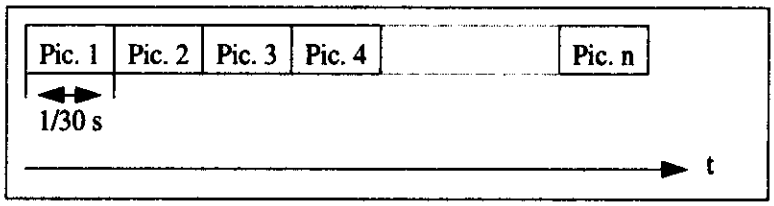


Figure 15.5: *Example of video LDUs.*

In the case of the basic physical unit being too small to handle, often LDUs are selected that block the samples into units of a fixed duration. A typical example is an audio stream where the physical unit duration is very small, therefore, LDUs are formed comprising 512 samples. In the example shown in Figure 15.6, one sample is coded with one Byte, and hence, each block contains 512 Bytes.

Captured media objects usually have a natural basic duration of an LDU. In computer-generated media objects, the duration of LDUs may be selected by the user. An

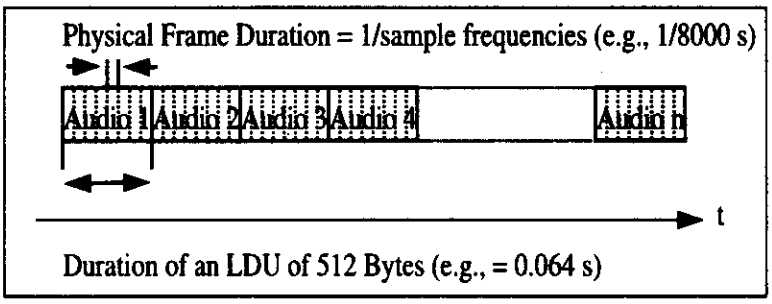


Figure 15.6: *Example of audio LDUs.*

example of these user-defined LDU durations is the frames of an animation sequence. For the presentation of a two-second animation sequence, 30 to 60 pictures may be generated depending on the necessary quality. Thus, the LDU duration depends on the selected picture rate (Figure 15.7).

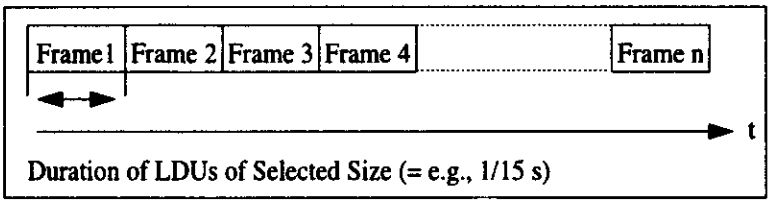


Figure 15.7: *LDU size selected by user.*

Streams are more complex when the LDUs vary in duration. An example is the recording of events at a graphical user interface to replay a user interaction. In this case, an LDU is an event with a duration lasting until the next event. The duration of LDUs depends on the user interaction and varies accordingly (Figure 15.8).

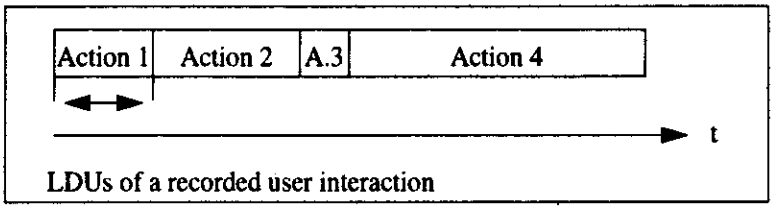


Figure 15.8: *LDUs of varying duration.*

	LDU Duration Defined During Capturing	LDU Duration Defined by the User
Fixed LDU Duration	Audio, Video	Animation, Timer
Variable, Unknown LDU Duration	Recorded Interaction	User Interaction

Table 15.1: *Types of LDUs.*

Open LDUs of unpredictable duration are given in the case that the LDU has no inherent duration. An example of an open LDU (i.e., an LDU with no inherent duration) is a user interaction in which the duration of the interaction is not known in advance (Figure 15.9).

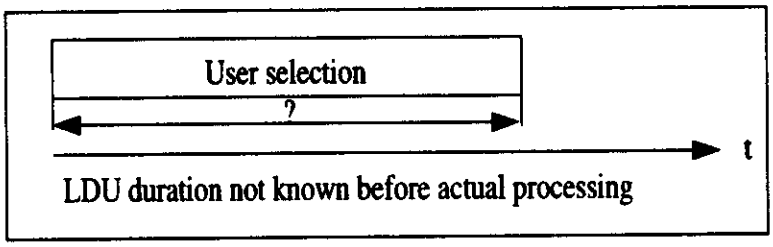


Figure 15.9: *An open LDU representing a user interaction.*

Timers can be regarded as streams of empty LDUs with a fixed duration (Figure 15.10).

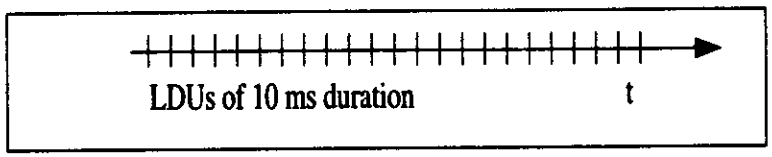


Figure 15.10: *LDUs of a timer.*

Table 15.1 gives an overview of the types of LDUs discussed above.

Further Examples

The following three examples show synchronization based on LDUs.

1. Lip synchronization demands tight coupling of audio and video streams. Synchronization can be specified by defining a maximal skew between the two media streams (Figure 15.11).

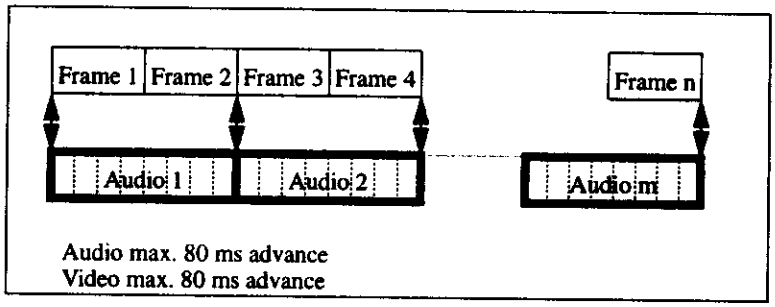


Figure 15.11: *LDU view of lip synchronization.*

2. A slide show with audio commentary demands that the change of slides be temporally related to the audio commentary (Figure 15.12).

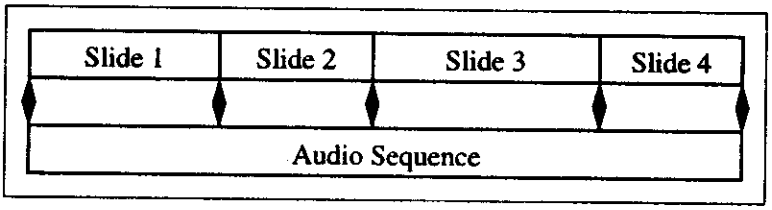


Figure 15.12: *LDU view of a slide show.*

3. The following example shown in Figure 15.13 will be used in Section 15.5 to demonstrate synchronization specification methods.

   A lip synchronized audio/video sequence (Audio1 and Video) is followed by the replay of a recorded user interaction (RI), a slide sequence (P1 - P3) and an animation (Animation), which is partially commented using an audio sequence (Audio2). Starting the animation presentation, a multiple choice question is

presented to the user (Interaction). If the user has made a selection, a final picture (P4) is shown.

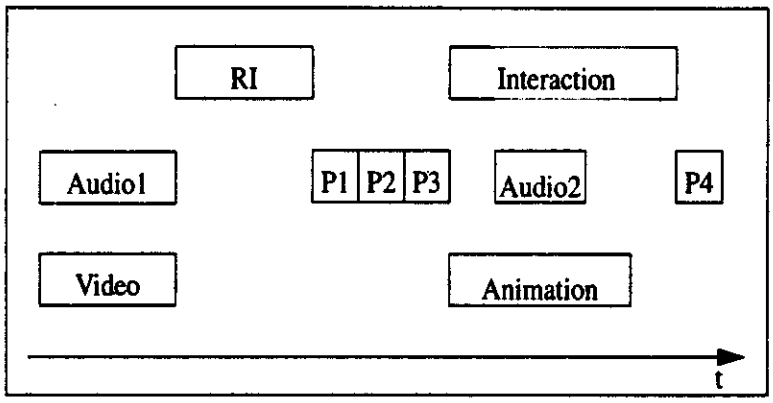


Figure 15.13: *Synchronization example.*

15.2.4 Live and Synthetic Synchronization

The live and synthetic synchronization distinction refers to the type of the determination of temporal relations. In the case of live synchronization, the goal of the synchronization is to exactly reproduce at a presentation the temporal relations as they existed during the capturing process. In the case of synthetic synchronization, the temporal relations are artificially specified [LG90b].

The following example shows aspects of *live synchronization*:

Two persons located at different sites of a company discuss a new product. Therefore, they use a video conference application for person-to-person discussion. In addition, they share a blackboard where they can display parts of the product and they can point with their mouse pointers to details of these parts and discuss some issues like: "This part is designed to ..."

This example covers two live synchronization aspects: video conference demands lip synchronization of the audio and video, and the movement of the mouse pointer must be synchronized to the corresponding explanation given in the video conference.

An example of *synthetic synchronization* is a learning environment of a city realized by the Bank Street College of Education, New York [Pre90]:

A learner may perform a virtual voyage (surrogate travel) to an ancient Mayan city. Using a joystick, the learner walks through the jungle and explores the Mayan ruins. At the same time, he hears the sounds from the jungle. He can also take a closer look at the nature in the environment and can "visit" a video museum to get further information.

In the case of synthetic synchronization, temporal relations have been assigned to media objects that were created independently of each other. The synthetic synchronization is often used in presentation and retrieval-based systems with stored data objects that are arranged to provide new combined multimedia objects. A media object may be part of several multimedia objects. For example, the same video clip about Germany may be part of a multimedia object that presents the countries of the European Union, as well as of a multimedia object that presents the countries qualified for the soccer world cup. Media objects of a multimedia object may be stored/located at different servers.

For synthetic synchronization, it is necessary to use a model for the specification and manipulation of temporal synchronization conditions and operations. Common examples [LG90a] of such operations are:

- Presenting media streams in parallel.
- Presenting media streams one after the other (serial).
- Presenting media stream independent of each other.

Live Synchronization

A typical application of live synchronization is conversational services. In the scope of a source/sink scenario, at the source, volatile data streams (i.e., data being captured from the environment) are created which are presented at the sink (Figure 15.14). The common context of several streams on the source site must be preserved at the sink. The source may be comprised of acoustic and optical sensors, as well

as media conversion units. The connection offers a data path between source and sink. The sink presents the units to the user. A source and sink may be located at different sites.

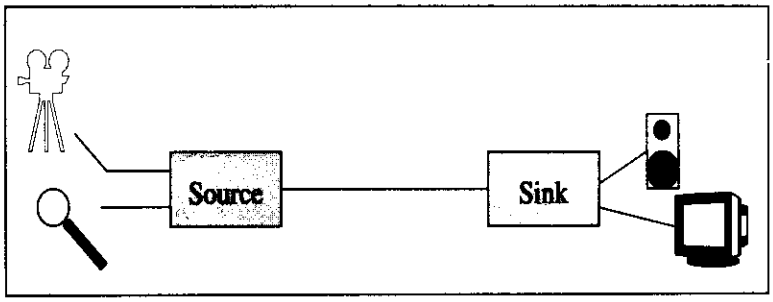


Figure 15.14: *Live synchronization without intermediate long-term storage.*

The goal of synchronization in such a scenario is to reproduce at the sink the signals in the same way as they appeared at the source. A possible manipulation by the sink is to adapt the presentation to the available resources. This may be, for example, a change of resolution or a lower frame rate. To reduce resource usage, it is preferable that such adaptations be already performed at the source, in particular if the source and sink are distributed and connected by a network.

Another type of live synchronization is shown in Figure 15.15 and includes storage which holds the encoded data. The presentation goal is the same as before, but the capturing and presentation are decoupled. In this case, it is possible to manipulate the presentation of the media. The presentation speed may be changed, and random access is possible (which is not possible in the scenario shown in Figure 15.14).

In summary, we must emphasize that the primary demand of live synchronization is to present data according to the temporal relations which existed during the capturing process of the media objects.

Synthetic Synchronization

The emphasis of synthetic synchronization is to support flexible synchronization relations between media. In synthetic synchronization, two phases can be distin-

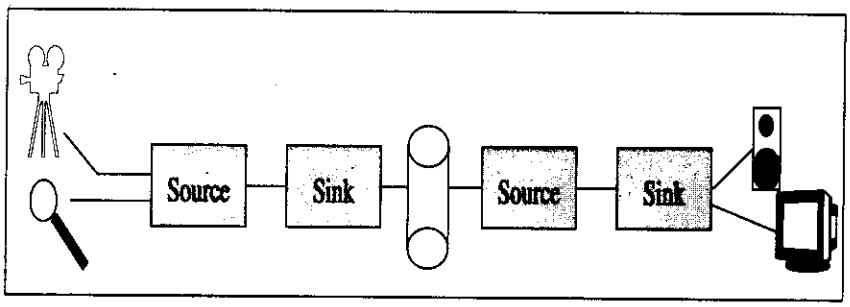


Figure 15.15: *Live synchronization with intermediate long-term storage and delayed presentation.*

guished:

- In the specification phase, temporal relations between the media objects are defined.
- In the presentation phase, a run-time system presents data in a synchronized mode.

The following example shows this for the creation of a multimedia presentation:

Four audio messages are recorded that relate to parts of an engine. An animation sequence shows a slow 360 degree rotation of the engine. With a software tool (e.g., a synchronization editor), time relations between the animation and matching audio sequences are defined. The media objects with the synchronization specification can be used by a presentation tool that executes the synchronized presentation.

In the specification phase of synthetic synchronization, the captured or created media objects are explicitly synchronized. Media objects that are stored in a live synchronization scenario can also be included in a synthetic synchronization playback.

Another variation of synthetic synchronization is the synchronization specification at run-time, such as:

In a railway time-table information system, a user specifies his demands. An automatically-generated audio sequence presents this information to the user. Dur-

ing the presentation, a video sequence is displayed that shows how to go to the departure gateways and how to proceed at the arrival station. The synchronization between the generated audio and video is performed at run-time.

15.2.5 Comment

In the case of live synchronization, the synchronization specification is implicitly defined during capturing. In the case of synthetic synchronization, the specification is done explicitly. If media objects are presented as delayed, presentation manipulations like changing the presentation speed and direction and direct access to a part of the object are possible. Adapting the presentation quality to the user demands or the capacity of the underlying system resources is possible in both cases.

User interaction in live synchronization includes only the interaction during capturing. Synthetic synchronization can include user interactions, for example, for navigation.

15.3 Presentation Requirements

For delivering multimedia data correctly at the user interface, synchronization is essential. It is not possible to provide an objective measurement for synchronization from the viewpoint of subjective human perception. As human perception varies from person to person, only heuristic criteria can determine whether a stream presentation is correct or not. In this section, results of some extensive experiments are presented that are related to human perception of synchronization between different media.

Presentation requirements comprise, for intra-object synchronization, the accuracy concerning delays in the presentation of LDUs and, for inter-object synchronization, the accuracy in the parallel presentation of media objects.

For intra-object synchronization, we try to avoid any jitter in consecutive LDUs. Whereas processes can wait for each other, using the method of blocking (e.g., in CSP – Calculus of Sequential Programming), a data stream of time-independent

LDUs can also be stopped.

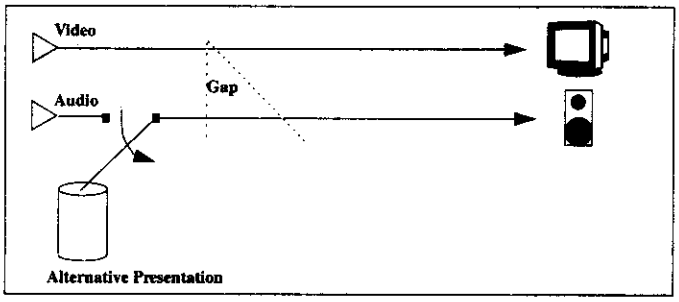


Figure 15.16: *The gap problem, restricted blocking.*

There is a different situation when looking at sequences of audio and moving pictures:

- What does the blocking of a stream of moving pictures mean for the connected output device?
- Should and can the last picture of a stream be shown during the blocking?
- Should, in the case of speech or music, a previous part be repeated during the blocking?
- How long can such a gap, as shown in Figure 15.16, exist?

This situation has become known as the *gap problem* [Ste90]. In the case of moving pictures, existing systems are solving the problem by simply switching the output device to dark or white, or by showing the last moving picture as a still picture. A practical solution must regard the factor time. It is significant, whether the duration of such a gap is a couple of milliseconds, a couple of seconds or even a couple of minutes. Only the actual application (and not the system) is able to select the best solution. Therefore, alternatives must be available that are selected independently of the expected blocking time. The concept of alternative presentations is indicated in Figure 15.16. In this example, it is shown that in the case that the gap between the late video and audio exceeds a predefined threshold, the audio presentation is switched to an alternative presentation. In the case that the gap is shorter, the

audio presentation may be stopped until the gap is closed. In general, in the case of blocking, alternative single pictures, sequences of pictures or audio signals can be presented, or simply previously used presentation units can be repeated. This method of process blocking, respectively streams of audio and video, is known as *restricted blocking.*

Restricted blocking uses as a means for resynchronization the repeated presentation of the last sample(s), or an alternative presentation. Another possibility is the *re-sampling* of a stream. The basic idea of re-sampling is to speed up or slow down streams for the purpose of synchronization. We distinguish off-line and on-line re-sampling. *Off-line re-sampling* is used after the capturing of media streams with independent devices. An example is a concert which is captured with two independent audio and video devices. If these devices, like many real-world devices, have insufficient accurate crystal clocks, the theoretic playback duration according to the sample rate of the stored audio and video sequences may differ. Before the execution of the presentation, it is possible to re-sample them to the same theoretic playback duration. *On-line re-sampling* is used during a presentation in the case that, at run-time, a gap between media streams occurs.

Methods for re-sampling are to re-define the playback rate, to duplicate, to interpolate or to skip samples or to re-calculate the whole sequence. The human perception of the re-sampling depends strongly on the media. Video sequences can be re-sampled by adding or deleting single frames in a stream, as is done in NTSC/PAL conversions. If the output device supports different playback rates, the playback rate can be directly adjusted.

Audio streams are more complex. A user will be annoyed by duplicated or deleted blocks of audio. Also, changes in the playback rate can easily be noticed by the user, especially in the case of music playback because the frequency is changing. The same is true for simple interpolation of samples. Algorithms exist that can stretch or widen an audio sequence without this frequency change, but they do not support real-time demands and are only suitable for off-line re-sampling.

For inter-object synchronization, more detailed results of studies in the lip synchronization and pointer synchronization [SE93] areas are described in the following to make clear the importance of user perception aspects for presentation accuracy. A

summary of requirements for other synchronization methods follows.

15.3.1 Lip Synchronization Requirements

Lip synchronization refers to the temporal relationship between an audio and video stream for the particular case of humans speaking. 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. Experiments at the IBM European Networking Center [SE93] measured skews that were perceived as "out of sync." In their experiments, users often mentioned that something was wrong with the synchronization, but this did not disturb their feeling for the quality of the presentation. Therefore, the experimenters additionally evaluated the tolerance of the users by asking if the data out of sink affected the quality of the presentation.

In discussions with experts that work with audio and video, the experimenters came to realize that generally, subjects responded to or remembered particular parts of the clips, therefore the experimenters observed a wide range of skews (up to 240 ms). A comparison and general usage of these values are somewhat doubtful because the environments from which they resulted were not comparable. In some cases, the experimenters encountered the "head view" displayed in front of some single color background on a high resolution professional monitor, whereas in others a "body view" in a video window at a resolution of 240 x 256 pixels was seen. To get accurate and good skew tolerance levels, the experimenters selected a speaker in a TV news environment in a head and shoulder shot (Figure 15.17). In this orientation, the viewer is not disturbed by background information and the viewer should be attracted by the gesture, eyes, and lip movement of the speaker.

Their study was performed in the news environment in which the experimenters recorded the presentation and then re-played it with artificially introduced skews created with professional editing equipment skewed at intervals of 40 ms, i.e., -120 ms, -80 ms, -40 ms, 0 ms, +40 ms, +80 ms, +120 ms. Steps of 40 ms were chosen for:

1. The difficulty of human perception to distinguish any lip synchronization skew with a higher resolution.

2. The capability of multimedia software and hardware devices to refresh motion video data every 33ms/40ms.

Figure 15.17: *Left: head view; middle: shoulder view; right: body view.*

Figure 15.18 provides an overview of the results. The vertical axis denotes the relative number of test candidates who detected a synchronization error, regardless of being able to determine if the audio was before or after the video. Their initial assumption was that the three curves related to the different views would be very different, but as shown in Figure 15.18, this is not the case.

A careful analysis provides us with information regarding the asymmetry, some periodic ripples and minor differences between the various views.

Left of the central axis, the graph relates to negative skew values where the video is ahead of the audio. On the right, the graph shows where the audio is ahead of the video. Day to day we often experience the situation where the motion of the lips is perceived a little before the audio is heard, due to the greater velocity of light than sound. This is indicated by the right-hand side of the curves being steeper than the left side.

The "body view" curve is broader than the "head view" curve, as at the former a small skew is easier to notice. The "head view" is also more asymmetric than the "body view," due to the fact that the further away we are situated, the less noticeable an error is.

At a fairly high skew the curves show some periodic ripples; this is more obvious in the case where audio is ahead of video. Some people obviously had difficulties in

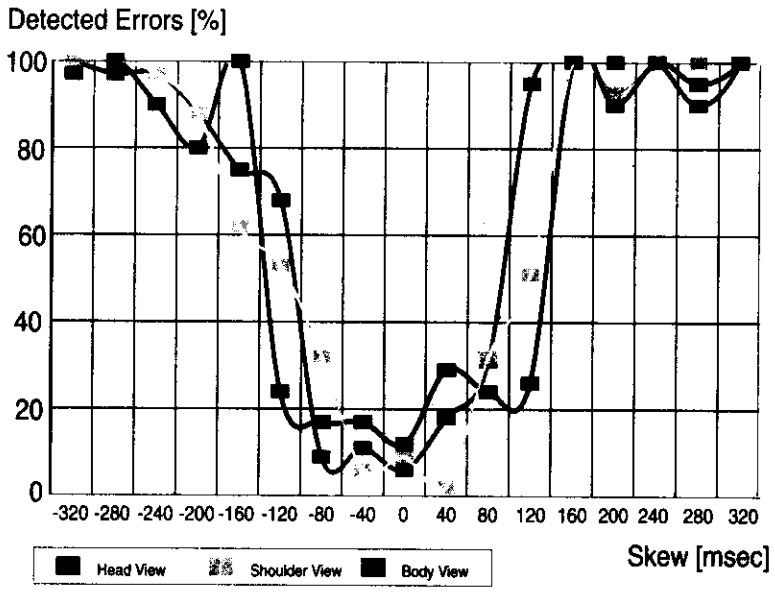


Figure 15.18: *Detection of synchronization errors with respect to the three different views. Left part: negative skew, video ahead of audio; right part: positive skew, video behind audio.*

identifying the synchronization error even with fairly high skew values. A careful analysis of this phenomenon is difficult due to the sample volume (few more than a 100), the media content to be synchronized and the human mind and mood. However, one plausible explanation could be: 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 and end of pauses, as well as whenever a change in tone was introduced (a point being emphasized). Errors in the middle of sentences were more difficult to notice. Also, we tended to concentrate more at the start of a conversation than once the subject was clear. A subsequent test containing video clips with skews according to these minima (without pauses and not showing the start, end and changes in tone) caused problems in identifying if there was indeed a synchronization error.

Figure 15.19 shows the following areas compiled according to the level of annoyance shown in Figure 15.20:

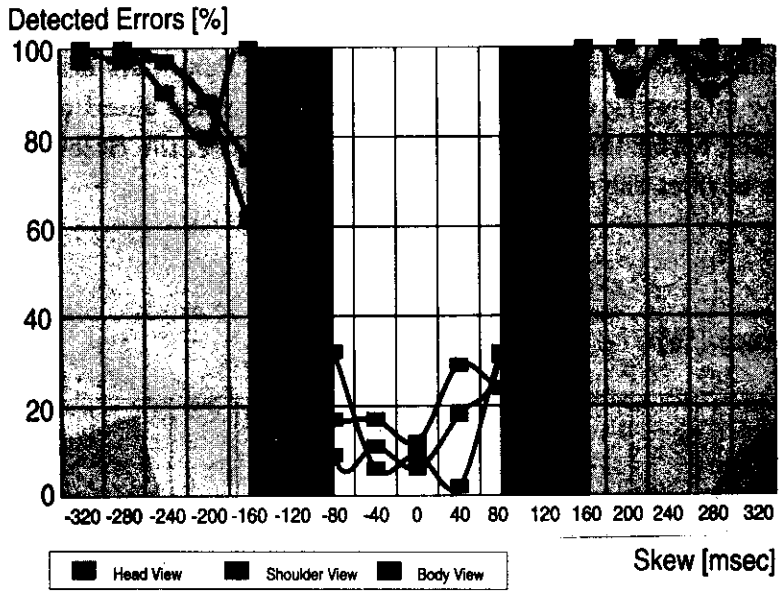


Figure 15.19: *Detection of synchronization errors.*

- The "in sync" region that spans a skew between -80 ms (audio behind video) and +80 ms (audio ahead of video). In this zone, most of the test candidates did not detect the synchronization error. Very few people said that if there was an error it affected their notion of the quality of the video. Additionally, some results indicated that the perfect "in sync" clip was "out of sync." Their conclusion is that lip synchronization can be tolerated within these limits.

- The "out of sync" areas span beyond a skew of -160 ms and +160 ms. Nearly everyone detected these errors and were dissatisfied with the clips. Data delivered with such a skew was 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 was ahead of video, the closer the speaker was, the easier errors were detected and described as disturbing. The same applied to the overall resolution, the better the resolution was, the more obvious the lip synchronization errors became.

- A second "transient" area, where video was ahead of audio, is characterized by a similar behavior as above as long as the skew values are near the in sync area. One interesting effect emerged, namely that video ahead of audio could be tolerated better than the opposing case. As above, the closer the speaker, the more obvious the skew.

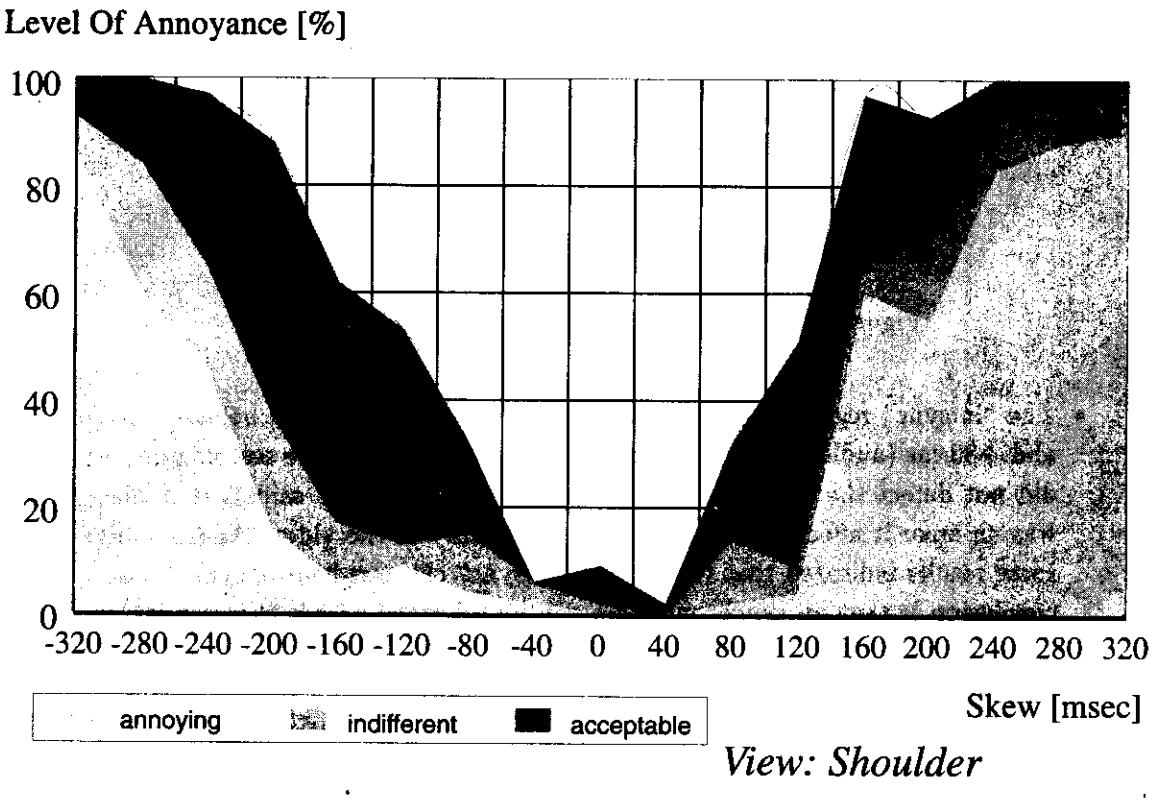


Figure 15.20: *Level of annoyance of audio/visual skew.*

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 fast light propagation compared to the acoustic wave propagation. The experimenters are just more used to this situation than the ones in the test.

15.3.2 Pointer Synchronization Requirements

In a Computer-Supported Co-operative Work (CSCW) environment, cameras and microphones are usually attached to the users' workstations. In the next experiment, the experimenters looked at a business report that contained some data with accompanying graphics. All participants had a window with these graphics on their desktop where a shared pointer was used in the discussion. Using this pointer, speakers pointed out individual elements of the graphics which may have been relevant to the discussion taking place. This obviously required synchronization of the audio and remote telepointer.

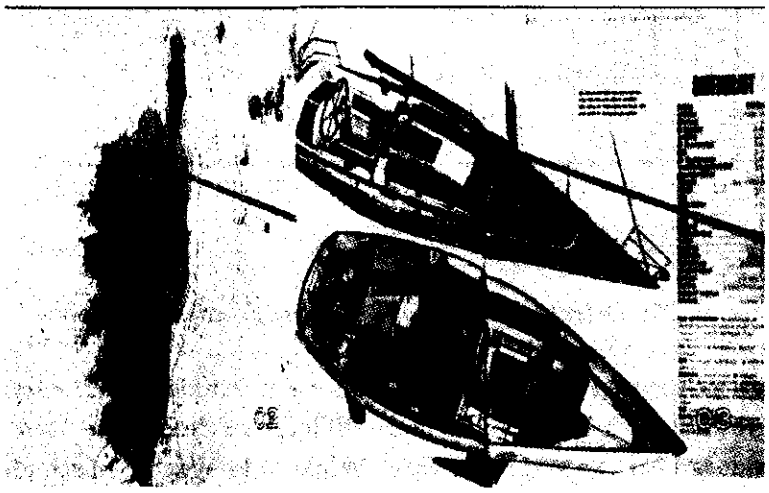


Figure 15.21: *Pointer synchronization experiment based on a map and technical sketch.*

The experimenters conducted two experiments:

- The first was to explain some technical parts of a sailing boat, while a pointer located the area under discussion (Figure 15.21, right side). The shorter the explanation, the more crucial the synchronization; therefore, the experimenters selected a fast-speaking person who used fairly short words.
- Additionally, the experimenters held a second experiment with the explanation of a traveling route on a map (Figure 15.21, left side). This involved the

continuous movement of the pointer.

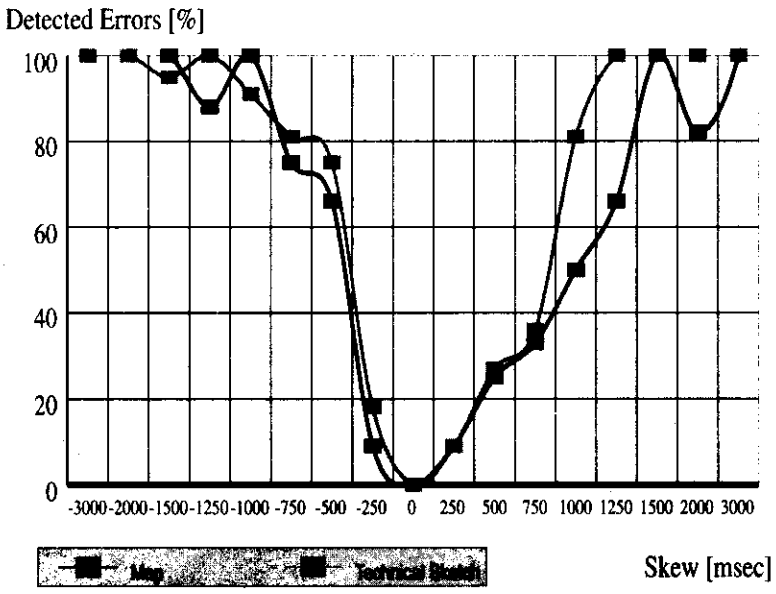


Figure 15.22: *Detection of the pointer synchronization errors.*

From the human perception point of view, pointer synchronization is very different from lip synchronization as it is much more difficult to detect the "out of sync" error at skew values near the error-free case. While a lip synchronization error is a matter of discussion for skews between 40 ms and 160 ms, for a pointer, the values lie between 250 ms and 1500 ms; Figure 15.22 shows some results.

Using the same judgement technique as in their first experiments, the "in sync" area related to audio ahead of pointing is 750 ms and for pointing ahead of audio it is 500 ms (Figure 15.22). This zone allows for a clear definition of the "in sync" behavior, regardless of the content.

The "out of sync" area spans a skew beyond -1000 ms and +1250 ms. At this point, the test candidates began to mention that the skew made the attempted synchronization worthless and became distracted unless the speaker slowed down or moved the pointer more slowly. From the user interface perspective, this is not acceptable. Quite clearly, the practice of pointing to one location on the technical

figure while discussing another is virtually impossible.

In the "transient" area, the experimenters found that many test candidates noticed the "out of sync" effect but it was not mentioned as annoying. This is certainly different from "lip sync" where the user was more sensitive to the skew and, without question, found it annoying.

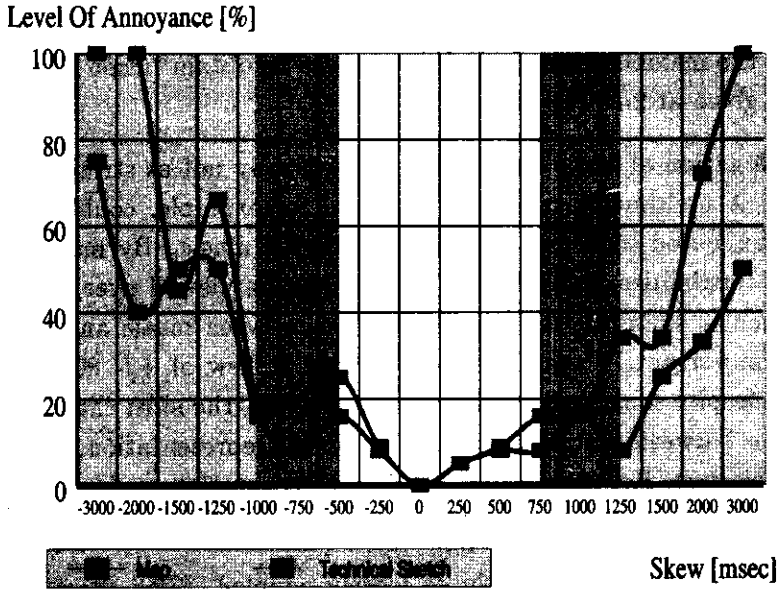


Figure 15.23: *Level of annoyance of the pointer synchronization errors.*

Figure 15.23 shows the number of people who disliked or were indifferent regarding the pointer synchronization error. It is worth mentioning that for several skew values, most of the test candidates detected the fault but did not object to such a skew, hence the broad "in sync" and "transient" areas.

15.3.3 Elementary Media Synchronization

Lip synchronization and pointer synchronization were investigated due to inconsistent results from available sources. The following summarizes other synchronization results to give a complete picture of synchronization requirements.

Since the beginning of digital audio, the *jitter* to be tolerated by dedicated hardware has been studied. Dannenberg provided some references and explanations of these studies. In [Ble78], the maximum allowable jitter for 16-bit quality audio in a sample period is 200 ps, which is the error equivalence to the magnitude of the LSB (Least-Significant Bit) of a full-level maximum-frequency 20-KHz signal. In [Sto72], some perception experiments recommended an allowable jitter in an audio sample period between 5 and 10 ns. Further perception experiments were carried out by [Lic51] and [Woo51], the maximum spacing of short clicks to obtain fusion into one continuous tone was given at 2ms (as cited by [RM80]).

The combination of audio and animation is usually not as stringent as lip synchronization. A multimedia course on dancing, for example, could show the dancing steps as animated sequences with accompanying music. By making use of the interactive capabilities, individual sequences can be viewed over and over again. In this particular example, the synchronization between music and animation is particularly important. Experience showed that a skew of +/- 80ms fulfills the user demands despite some possible jitter. Nevertheless, the most challenging issue is the correlation between a noisy event and its visual representation, e.g., the simulated crash of two 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 ms. [DS93] reports that the perceptible phase shift between two audio channels is 17 ms. This is based on a headphone listening experiment. Since a varying delay in one channel causes the apparent location of a sound's source to move, Dannenberg proposed to allow an audio sample skew between stereo channels within the boundaries of +/- 11 ms. 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 1 s arises which can be explained as follows [Dan93]: consider that it takes a second or so to advance a slide projector; however, people sometimes comment on the time it takes to change transparencies on an overhead projector, but rarely worry about automatic slide projectors.

A more elaborate 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, e.g., for tutoring purposes. [Dan93] 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 [Cly85, RM80, Ste87]).

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 some cases, the audio provides further acoustic information to the displayed or highlighted text in terms of "audio annotation." In an existing "music dictionary," an antique instrument is described and simultaneously played. An example of a stronger correlation is the playback of a historical speech, e.g., a speech of 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. The same applies to the teaching of a language where in a playback mode the spoken word is simultaneously highlighted. 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 in the order of magnitude of 500 ms. Therefore, the experimentally verified skew of 240 ms is affordable.

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

- In the overlay mode, the text is often an additional description to the displayed moving image sequence. For example, in a video of playing billiards, the image is used to denote the exact direction of the ball after the last stroke. The simultaneous presentation of the video and 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 nearby the topic of discussion is very useful. In such overlay schemes, this text must be synchronized to the video 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 for a certain time period to be correctly 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 and 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.

Consider the billiard ball example from before: a video shows the impact of two

billiard balls and the image of the actual "route" of one of the balls is shown by an animated sequence. Instead of a series of static images, the track of the second ball can be followed by an animation which displays the route of the ball across the table. In this example, any "out of sync" effect is immediately visible. 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 slightly different positions. An acceptable result can be achieved if every three subsequent frames the ball moves by its diameter. A smaller frame rate may result in the problem of continuity, as often seen in tennis matches on television. As each frame lasts about 40 ms and three subsequent frames are needed, an allowable skew of 120 ms would be acceptable. This is very tight synchronization, which was suitable for the examples the experimenters looked at. Other examples where video and animation are combined include computer-generated figures in films.

Multimedia systems also incorporate the real-time processing of control data. Telesurg is a good example where graphical information is displayed based on readings taken by probes or similar instruments. No overall timing demand can be stated as these issues highly depend on the application itself.

15.4 A Reference Model for Multimedia Synchronization

A reference model is needed to understand the various requirements for multimedia synchronization, identify and structure run-time mechanisms that support the execution of the synchronization, identify interfaces between run-time mechanisms and compare system solutions for multimedia synchronization systems.

To this end, we first describe existing classification and structuring methods. Then, a four-layer reference model is presented and used for the classification of multimedia synchronization systems in our case studies. As many multimedia synchronization mechanisms operate in a networked environment, we also discuss special synchronization issues in a distributed environment and their relation to the reference model.

15.4.1 Existing Classification Approaches

An overall classification was introduced by Little and Ghafoor [LG90b]. They identified a physical level, system level and human level, but gave no detailed description or classification criteria. Other classification schemes distinguish between intrastream (fine-grain) synchronization and interstream (coarse-grain) synchronization, or between live and synthetic synchronization [LG90b, SM92a].

The model of Gibbs, Breiteneder and Tschichritzis [GBT93] maps a synchronized multimedia object to an uninterpreted byte stream. The multimedia objects consist of derived media objects comprised of rearranged media sequences, e.g., scenes from a complete video. The parts of the media sequences are themselves part of an uninterpreted byte stream.

Ehley, Furth and Ilyas [EFI94] classify intermedia synchronization techniques that are used to control jitter between media streams according to the type and location of the synchronization control. They distinguish between a distributed control based on protocols, distribution based on servers and distribution on nodes without server structure. For local synchronization control, they distinguish control on several layers and the use of local servers.

These classification schemes seem to be orthogonal, and each one of them only captures some specific aspects. They do not fulfill the above stated requirements of the synchronization reference model.

An improved three-layer classification scheme has been proposed by Meyer, Effelsberg and Steinmetz [MES93]. The layers are: the media layer for intrastream synchronization of time-dependent media, the stream layer for interstream synchronization of media streams, the object layer for the presentation, including the presentation of time-independent media objects and the specification layer for authoring complex multistream multimedia applications. At each layer, typical objects and operations are identified. Each layer can be accessed directly by the application or indirectly through higher layers. This approach fulfills the demands of a reference model approach and we will enhance and interpret it appropriately in the following.

15.4.2 The Synchronization Reference Model

A four-layer synchronization reference model is shown in Figure 15.24. Each layer implements synchronization mechanisms which are provided by an appropriate interface. These interfaces can be used to specify and/or enforce the temporal relationships. Each interface defines services, i.e., offering the user a means to define his/her requirements. Each interface can be used by an application directly, or by the next higher layer to implement an interface. Higher layers offer higher programming and Quality of Service (QoS) abstractions.

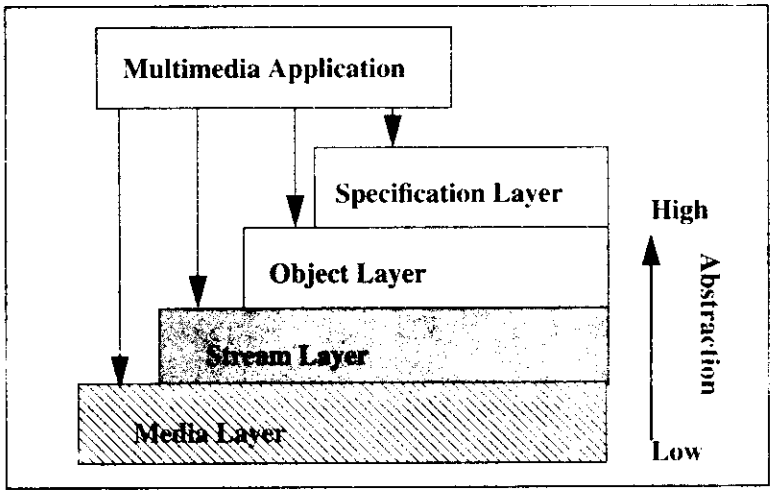


Figure 15.24: *Four-layer reference model.*

For each layer, typical objects and operations on these objects are described in the following. The semantics of the objects and operations are the main criteria for assigning them to one of the layers.

Detailed programming examples derived from a real interface provided by a real product, prototype or standard demonstrate how synchronization can be achieved through this layer. The scenario for the programming example is to display subtitles at predefined times during the playout of a digital movie.

Media Layer

At the *media layer*, an application operates on a single continuous media stream, which is treated as a sequence of LDUs.

The abstraction offered at this layer is a device-independent interface with operations like *read(devicehandle, LDU)* and *write(device-handle, LDU)*. Systems such as ActionMedia/II™'s audio-video kernel [IBM92a] or SunSPARC™'s audio device [TP91] provide the corresponding interfaces.

To set up a continuous media stream using the abstractions offered by the media layer, an application executes a process for each stream in the manner shown in the following example:

```
window = open(''Videodevice"); \\ Create a video output window
movie = open(''File"); \\ Open the video file
while (not eof(movie)) { \\ Loop
read(movie, &ldu); \\ Read LDU
if (ldu.time == 20) \\ Start the presentation
 print(''Subtitle 1"); \\ of the synchronized subtitles
else if (ldu.time == 26)
 print(''Subtitle 2");
write(window, ldu);} \\ Present LDU
close(window); \\ Close window
close(movie); \\ Close file
```

The process reads and writes LDUs in a loop as long as data are available. Synchronous playout of a subtitle is achieved by polling the timestamps of the LDUs to have a certain value.

Using this layer, the application itself is responsible for the intrastream synchronization by using flow-control mechanisms between a producing and a consuming device [RR93]. If multiple streams run in parallel, the sharing of resources may affect their real-time requirements. Usually, a resource reservation and management scheme allows for guaranteeing intrastream synchronization [VHN92]. The operating system schedules the corresponding process in real-time [MSS92]. In distributed

systems, the networking components are taken into account [AHS90, Fer91]. In the special case of lip synchronization, the interstream synchronization can be provided easily, where simultaneous audio and video frames are interleaved within the same LDU (e.g., ActionMedia-II's audio-video support system [IBM92a] and the MPEG data stream [ISO93a]). Finally, the synchronous playout of time-independent media objects and user interactions are tasks to be performed by the application.

Media layer implementations can be classified into simple implementations and implementations that provide access to interleaved media streams.

Stream Layer

The *stream layer* operates on continuous media streams, as well as on groups of media streams. In a group, all streams are presented in parallel by using mechanisms for interstream synchronization.

The abstraction offered by the stream layer is the notion of streams with timing parameters concerning the QoS for intrastream synchronization in a stream and interstream synchronization between streams of a group.

Continuous media is seen in the stream layer as a data flow with implicit time constraints; individual LDUs are not visible. The streams are executed in a Real-Time Environment (RTE), where all processing is constrained by well-defined time specifications [Her92]. On the other hand, the applications themselves that are using the stream layer services are executed in a Non Real-Time Environment (NRTE), where the processing of events is controlled by the operating system scheduling policies.

Typical operations invoked by an application to manage streams and groups from the NRTE are: *start(stream), stop(stream), create_group(list_of_streams), start(group)* and *stop(group)*. The interaction with time-independent media objects and user interactions is performed by the attachment of events to the continuous media streams (e.g., *setcuepoint(stream/group, at, event)*). Such an event is sent to the application whenever the stream reaches the specified point during playback. At this layer, the application is furthermore in charge of any time-independent media object and user

interaction processing. This leads to different application interfaces for continuous media and for time-independent media and user interactions.

The Sync/Stream Subsystem of IBM's MultiMedia Presentation Manager™(MMPM) for OS/2™provides a set of services which can be used to implement data streaming and synchronization. This subsystem, which can be understood as the RTE, is comprised of the Sync/Stream Manager and several stream handlers [IBM92b]. Stream handlers are responsible for controlling the continuous data flow in real-time. The Sync/Stream Manger provides a resource management and controls the registration and activities of all stream handlers.

The following programming example for the use of the stream layer uses the string command interface provided by MMPM.

```
open digitalvideo alias ex \\ Create video descriptor
load ex video.avs \\ Assign file to video descriptor
setcuepoint ex at 20 return 1 \\ Define event 1 for subtitle 1
setcuepoint ex at 26 return 2 \\ Define event 2 for subtitle 2
setcuepoint ex on \\ Activate cuepoint events
play ex \\ Start playing

switch readevent() { \\ Event handling
case 1: display(''Subtitle 1") \\ If event 1 show subtitle 1
case 2: display(''Subtitle 2") \\ If event 2 show subtitle 2
}
```

In MMPM/2™, interstream synchronization for synchronized playback of multiple streams within a group is achieved by a master/slave algorithm, where one stream (the master) controls the behavior of one or more subordinate streams (the slaves). The skip/pause algorithm introduced in [AH91a] gives a detailed discussion of the implementation of such a behavior. The synchronization mechanism in ACME [AH91b], as well as the Orchestration Service [CGCH92], support stream layer abstractions for distributed multimedia systems.

The stream layer abstraction was derived from the abstraction normally provided by the integration of analog media in the computer system. In the Muse and Pyg-

malion systems of MIT's Project Athena [HSA89] or in the DiME [SM92b] system, continuous media were routed over separated channels through the computer. The connected devices could be controlled by sending commands via the RS-232C interface to start and stop the media streams. In such systems, live synchronization between various continuous media streams is directly performed by the dedicated processing devices.

Stream layer implementations can be classified according to their support for distribution, to the types of guarantees that they provide and to the types of supported streams (analog and/or digital).

An application using the stream layer is responsible for starting, stopping and grouping the streams and for the definition of the required QoS in terms of timing parameters supported by the stream layer. It is also responsible for the synchronization with time-independent media objects.

Object Layer

The *object layer* operates on all types of media and hides the differences between discrete and continuous media.

The abstraction offered to the application is that of a complete, synchronized presentation. This layer takes a synchronization specification as input and is responsible for the correct schedule of the overall presentation. From our understanding, the abstractions are similar to the "object model" presented in [Ste90].

The task of this layer is to close the gap between the needs for the execution of a synchronized presentation and the stream-oriented services. The functions located at the object layer are to compute and execute complete presentation schedules that include the presentation of the non-continuous media objects and the calls to the stream layer. Further, the object layer is responsible for initiating preparation actions that are necessary for achieving a correctly synchronized presentation. The object layer does not handle the interstream and intrastream synchronization. For these purposes, it uses the services of the stream layer.

An example of interfacing this layer is an MHEG specification [MHE93]. The scope

of the MHEG standard is the coded representation of final form multimedia and hypermedia information objects. In the following, we give a rudimentary example of how our scenario might be coded in the MHEG standard (using a simple notation to demonstrate the essentials of our reference model):

```
Composite { \\ Composite object
start-up link \\ How to start the
\\  presentation

viewer start-up
viewer-list \\ Virtual views on
Viewer1: reference to Component1 \\ component objects
Viewer2: reference to Component2
Viewer3: reference to Component3
Component1 \\ Component objects
reference to content ''movie.avs" \\ of the composite
Component2
reference to content ''Subtitle1"
Component3
reference to content ''Subtitle2"
Link1 \\ Temporal relations
''when timestone status of Viewer1
becomes 20 then start Viewer2"
Link2
''when timestone status of Viewer1
becomes 26 then start Viewer3"
}
```

A possible implementation of the object layer is an MHEG run-time system, the *MHEG engine*. The MHEG engine evaluates the status of the objects and performs operations (actions) like prepare, run, stop or destroy on these objects. In the case of time-dependent media objects, the run operation may be mapped to the initiation of a media stream on the stream layer. In the case of a time-independent media object, this call directly demands the object to be presented. Prepare times are

necessary, for example, to allow the stream layer to build up a stream connection, or in the case of time-independent media objects, to prefetch the presentation, e.g., to adapt the picture color maps to the maps of the output device. The preparation is started by the prepare action.

Object layer implementations can be classified according to distribution capabilities and the type of presentation schedule computation. It can be distinguished whether the implementation calculates a schedule and, if it calculates one, whether the schedule is computed before the presentation or at run-time of the presentation. Concerning distribution, implementations may be local and may support distribution based on a server structure or full distribution without restriction.

The task of the application using the object layer is to provide a synchronization specification.

Specification Layer

The *specification layer* is an open layer. It does not offer an explicit interface. This layer contains applications and tools are located that allow to create synchronization specifications. Such tools are synchronization editors, multimedia document editors and authoring systems. Also located at the specification layer are tools for converting specifications to an object layer format. An example of such a conversion tool is a multimedia document formatter that produces an MHEG specification as proposed by Markey [Mar91a].

For example, the synchronization editor of the MODE system [BHLM92] may be used to specify the synchronization example. It offers a graphical interface to select the video and text objects to use, to preview the video, to select suitable points where the subtitles have to be shown, to specify the temporal relation of this point to the subtitle and to store the synchronization specification.

The specification layer is also responsible for mapping QoS requirements of the user level to the qualities offered at the object layer interface.

Synchronization specification methods can be classified into the following main categories:

- *Interval-based specifications,* which allow the specification of temporal relations between the time intervals of the presentations of media objects.
- *Axes-based specifications,* which relate presentation events to axes that are shared by the objects of the presentation.
- *Control flow-based specifications,* in which at given synchronization points, the flow of the presentations is synchronized.
- *Event-based specifications,* in which events in the presentation of media trigger presentation actions.

15.4.3 Synchronization in a Distributed Environment

Synchronization in a distributed environment is more complex than in a local environment. This is mainly caused by the distributed storage of synchronization information and the different locations of the media objects involved in the presentation. The communication between the storage and presentation site introduces additional delays and jitter. Often, we also encounter multi-party communication patterns.

Transport of the Synchronization Specification

At the sink node, the presentation component needs to have the synchronization specification at the moment an object is to be displayed. We distinguish between three main approaches for the delivery of the synchronization information to the sink:

- *Delivery of the complete synchronization information before the start of the presentation:* This approach is often used in the case of synthetic synchronization. Typically, the application at the sink node accesses the object layer interface with the specification or a reference to the specification as a parameter. The implementation of this approach is simple and it also allows easy handling in the case of several source nodes for the media objects. The

disadvantage is the delay caused by the transport of the synchronization specification before the presentation, especially if it is stored on another node. The transport of the synchronization specification is a duty of a component located at the object layer or above.

- *Use of an additional synchronization channel*: This approach, shown in Figure 15.25, is useful in the case of one source node only. It is used and is preferable in the case of live synchronization when all the synchronization information is not known in advance. No additional delays are caused by this method. A disadvantage is that an additional communication channel is needed that may cause errors due to delay or loss of synchronization specification units. It is often forgotten that the information on the synchronization channel must be decoded at the time the respective object is to be displayed, i.e., data communication at this channel must obey certain time behavior. Also, the case of multiple source nodes for synchronized media objects is difficult to handle. The synchronization channel must be handled by the object layer and possibly supported by the stream layer if the synchronization channel is to be defined as a stream.

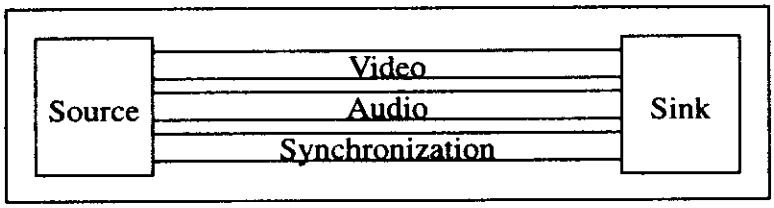


Figure 15.25: *Use of a separate synchronization channel.*

- *Multiplexed data streams*: The advantage of multiplexing data streams on one communication channel (Figure 15.26) is that the related synchronization information is delivered together with the media units. No additional synchronization channel is necessary and no additional delay is caused by this approach. An important problem regarding multiplexed media and synchronization information is the difficulty of selecting an appropriate QoS which matches the requirements of all involved medias, e.g., reliability is dominated by the most stringent media objects. This method is also difficult to use for multiple source nodes. It must be supported by the stream layer. The use

of multiplexed data streams may be implied by coding standards like MPEG. MPEG defines a bitstream that combines video, audio and the related synchronization information. Hence, this type of bitstream can be regarded as one medium on the stream layer and for the synchronization with other media, the other approaches can also be chosen.

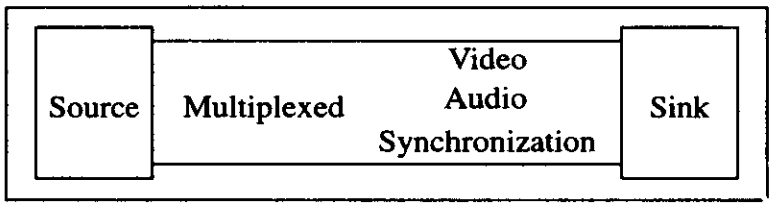


Figure 15.26: *Multiplexed media and synchronization channels.*

Location of Synchronization Operations

In some cases it is possible to synchronize media objects by combining the objects into a new media object. This approach may be used to reduce communication resource demands, as shown in Figure 15.27. In this case, an animation and two bitmaps that must overlay a video sequence are already merged at the source node to become a new video object to reduce bandwidth demands.

The mixing of objects, including time-independent media objects, must be supported by the object layer. The mixing of media streams, like mixing audio channels, must be supported by the stream layer.

Clock Synchronization

In distributed systems, the synchronization accuracy between the clocks of the source and sink nodes must be considered. Many synchronization schemes demand knowledge about the timing relations. This knowledge is the basis for global timer-based synchronization schemes, as well as for schemes that demand that operations on distributed nodes are timely and coordinated to ensure, on one hand, in-time deliv-

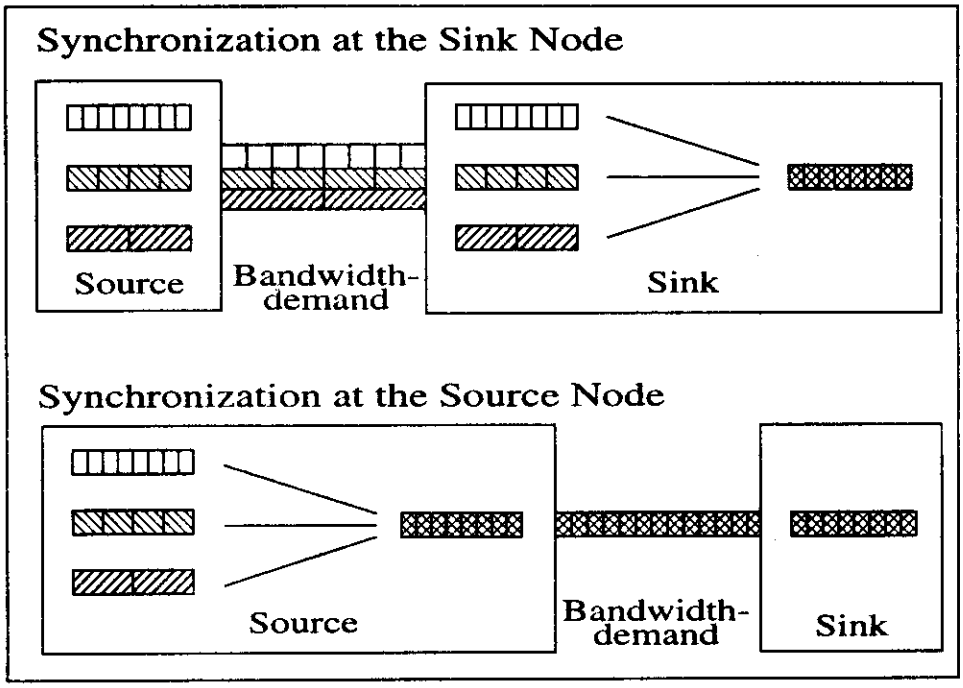


Figure 15.27: *Combining objects to reduce communication resource demands.*

ery and on the other hand, that operations are not performed too early to avoid a buffer overflow.

This problem is especially important for the synchronization in the case of multiple sources (Figure 15.28). If a synchronized audio-video presentation should start at time T_{av} at the sink node, the audio transmission of Source A must start at $T_a = T_{av} - N_{la} - O_a$, with N_{la} as the known net delay and O_a as the offset of the clock of node A with respect to the clock of the sink node. For source node B, the start time of the video transmission is $T_v = T_{av} - N_{lv} - O_v$.

The offsets O_a and O_v are not known. The resulting problem of delivery to the sink in time can be solved if the maximal possible values for O_a and O_v are known. It is possible to allocate buffer capacities at the sink and to start the transmission of the audio and video in advance to guarantee that the required media units are available. Because the necessary buffer capacity at the sink node depends on the

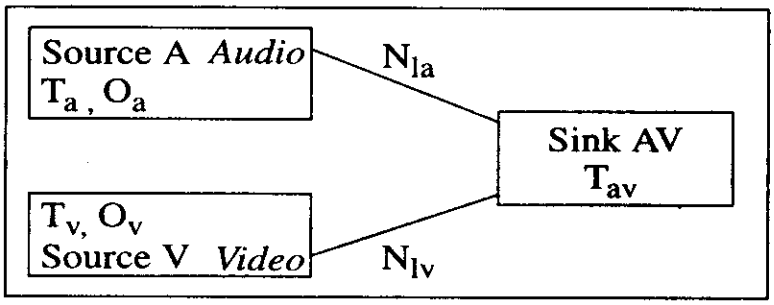


Figure 15.28: *Clock offsets in a distributed environment.*

possible offset, and we must assume limited buffer capacity, it is necessary to limit the maximal offset. This can be achieved with clock synchronization protocols like the Network Time Protocol [Mil91] that allows the synchronization of the clocks with an accuracy in the range of 10 ms. With the use of public broadcast timer signals, submillisecond accuracies are practical [Mil93a].

This accuracy is suitable for global timer synchronization and for distributed operation scheduling.

The in-time delivery of LDUs of a stream is a task of the stream layer that must handle the clock offsets. For in-time delivery of time-independent media objects, the object layer is responsible.

Multiple Communication Relations

Possible communication patterns are shown in Figure 15.29. Patterns with multiple sinks demand that at run-time, multicast and broadcast mechanisms be used to reduce resource requirements, in particular network resources. Also inefficient multiple executions of the same operation at different sinks should be avoided. The multicasting of streams is the task of the stream layer. Efficient planning of operation execution in the different communication patterns is a responsibility of the object layer.

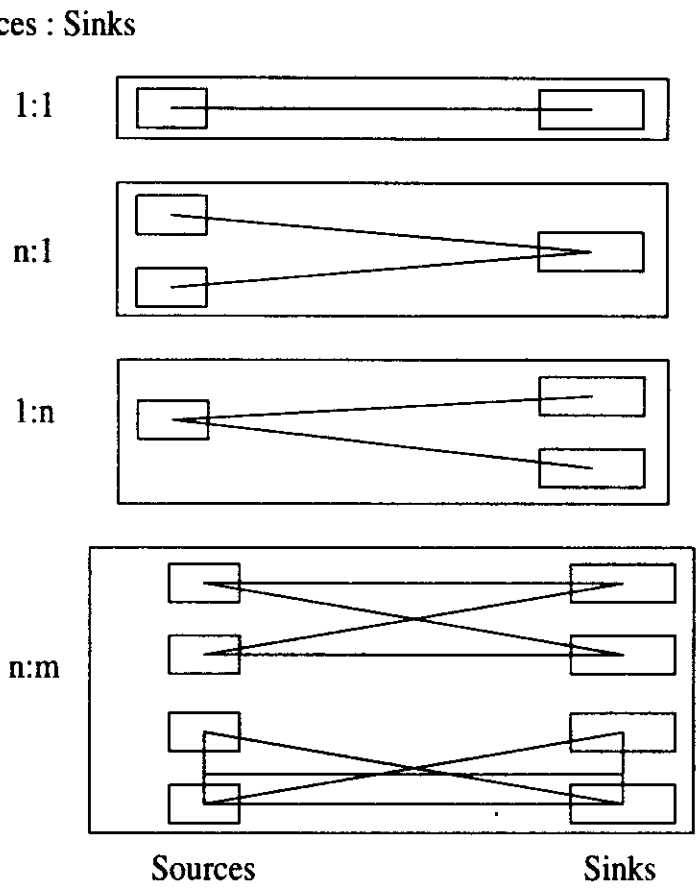


Figure 15.29: *Multiple communication relations.*

Multi-Step Synchronization

Synchronization in a distributed environment is typically a multi-step process. During all steps of the process, the synchronization must be maintained in a way that enables the sink to perform the final synchronization. The steps of the process are:

- Synchronization during object acquisition, e.g., during digitizing video frames.
- Synchronization of retrieval, e.g., synchronized access to frames of a stored video.

- Synchronization during delivery of the LDUs to the network, e.g., delivering the frames of a video to the transport service interface.
- Synchronization during the transport, e.g., by isochronous protocols.
- Synchronization at the sink, i.e., synchronized delivery to the output devices.
- Synchronization within the output device.

Manipulation of the Presentation

The support of functions like pause, forward and backward with different presentation speeds, direct access, stop and repeat is difficult in a distributed environment. The necessary information must be distributed in the environment. Objects that have already been prepared in advance for the presentation must be deleted. Network connections may be subject to change or must be rebuilt. Therefore, delays in the execution of these manipulation functions are difficult to avoid.

Consequences for Synchronization in a Distributed Environment

To achieve synchronization in a distributed environment, many decisions must be made. A first decision is the selection of the type of transport for the synchronization specification. In run-time, decisions must be taken concerning the location of the synchronization operations, handling of the offsets of the clocks and the handling of multicast and broadcast mechanisms. Especially, coherent planning of the steps in the synchronization process, together with the necessary operations on the objects, e.g., decompression, must be done. In addition, presentation manipulation operations demand additional re-planning at run-time.

In general, the execution of synchronized distributed presentations is a complex planning problem. The resulting plan is often known as a schedule.

15.4.4 Aggregate Characteristics of the Synchronization Reference Model

The reference model allows for the structuring and classifying of synchronization systems. The identification of the interfaces and layers enables one to combine existing solutions into complete systems. Table 15.2 provides an overview of the interface abstractions and tasks of all layers of our reference model. The classification of mechanisms and methods in the layers is summarized in Table 15.3.

15.5 Synchronization Specification

The synchronization specification of a multimedia object describes all *temporal dependencies* of the included objects in the multimedia object. It is produced using tools at the specification layer and is used at the interface to the object layer. Because the synchronization specification determines the whole presentation, it is a central issues in multimedia systems. In the following, requirements for synchronization specifications are described and specification methods are described and evaluated.

A synchronization specification should be comprised of

- Intra-object synchronization specifications for the media objects of the presentation.
- QoS descriptions for intra-object synchronization.
- Inter-object synchronization specifications for media objects of the presentation.
- QoS descriptions for inter-object synchronization.

The synchronization specification is part of the description of a multimedia object. In addition, for a multimedia object, it may be described in which presentation form, respectively in which alternative presentation forms, a media object should be presented. For example, a text could be presented as text on the screen or as

Layer	Interface Abstraction	Tasks
Specification	• The tools performing the tasks of this layer have interfaces; the layer itself has no upper interface	• Editing • Formatting • Mapping user-oriented QoS to the QoS abstraction at the object layer
Object	• Synchronization Specification • Objects that hide types of enclosed media • Media-oriented QoS (in terms of acceptable skew and jitter)	• Plan and coordinate presentation scheduling • Initiate presentation of time-dependent media objects by the stream layer • Initiate presentation of time-independent media objects • Initiate presentation preparation actions
Stream	• Streams and groups of streams • Guarantees for intrastream synchronization • Guarantees for interstream synchronization of streams in a group	• Resource reservation and scheduling of LDU processing
Media	• Device-independent access to LDUs • Guarantees for single LDU processing	• File and device access

Table 15.2: *Overview of the synchronization reference model layers.*

Layer	Classification Items
Specification	Synchronization specification method: • Interval-based synchronization • Axes-based synchronization • Control flow-based synchronization • Event-based synchronization
	Type of tool: • Textual specification tool • Graphical specification tool • Converter
Object	Type of distribution: • Local • Distributed, based on servers • Distributed without server usage
	Type of schedule computation: • No computation • Compile-time computation • Run-time computation
Stream	Type of distribution: • Local • Distributed
	Type of guarantees for stream QoS: • No guarantees for QoS, best effort • Guarantees for QoS by resource reservations
Media	Type of accessible data: • Single medium data • Interleaved, complex data

Table 15.3: *Classification of methods and mechanisms at the synchronization reference model layers.*

generated audio sequence. A specification may allow only one of these or a selection of the presentation form at run-time.

In the case of live synchronization, the temporal relations are implicitly defined during capturing. QoS requirements for single media are defined before starting the capture.

In the case of synthetic synchronization, the specification must be created explicitly. Several synthetic synchronization specification methods have been described in the literature. The most important are classified, surveyed and evaluated in the following sections.

15.5.1 Quality of Service

The necessary QoS depends on the media and application.

Quality of Service for a Media Object

The QoS specification for a media object includes the quality concerning single LDUs of a media object and the accuracy with which the temporal relations between the LDUs of this media object must be fulfilled if the media object is a time-dependent object.

Table 15.4 shows some QoS parameters for a media object. The white boxes contain qualities that are independent of temporal relations. The light shaded boxes contain timing related qualities that are under the limited influence of the presentation system because the quality depends on the quality selected during capture. Usually, only quality degradation via the presentation system is possible. The dark shaded boxes contain timing qualities which are potentially under full control of the presentation environment.

Media	Image (e.g., bitmap)	Video	Audio
Quality of Service	Color Depth	Color Depth	Lin. or log. sampling
	Resolution	Resolution	Sample Size
		Frame Rate	Sample Rate

Table 15.4: *Some QoS for the presentation of a media object.*

Quality of Service of Two Related Media Objects

Synchronization requirements can be expressed by a QoS specification. One QoS parameter can define the acceptable skew within the concerned data streams; namely, 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 audio and video parts of a film are stored as different entries in a database, lip synchronization according to the above-mentioned results should 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 subsequent playback. The stored data should be captured and recorded with no skew at all, i.e., "in sync." This is particularly applicable if the file is stored in an interleaved format. At the participant's site, the actual incoming audiovisual data is "in sync" according to the defined lip synchronization boundaries. Assuming the data arrive with a skew of +80 ms, and if audio and video LDUs are transmitted as a single multiplexed stream over the same transport connection, then it will be dis-

played as apparently "in-sync." Should the data be stored on the hard disk and presented simultaneously at a local workstation and to a remote spectator, then for correct delivery, 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 the outcome might be that by applying these boundaries twice, data are not "in sync." In general, any synchronized data which will be further processed should be synchronized according to a production-level quality, i.e., with no skew at all.

- The presentation requirements discussed in Section 15.3 identify *presentation-level synchronization.* This synchronization defines whatever is reasonable at the user interface. It does not take into account any further processing of the synchronized data; presentation-level synchronization focuses on the human perception of the synchronization. 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. The required QoS for synchronization is expressed as the allowed skew. The QoS values shown in Table 15.5 relate to presentation-level synchronization. Most of them result from exhaustive experiments and experiences, others are derived from literature as referenced. To their understanding, they serve as a general guideline for any QoS specification. During the lip and pointer synchronization experiments, we learned that many factors influenced these results. We understand this whole set of QoS parameters as a first-order result to serve as a general guideline. However, these values may be relaxed depending on the actual content.

Quality of Service of Multiple Related Media Objects

So far, media synchronization has been evaluated as the relationship between two kinds of media or separate data streams. This is the canonical foundation of all types of media synchronization. In practice, we often encounter more than two related media streams; a sophisticated multimedia application scenario incorporates the simultaneous handling of various sessions. An example is a video conference where a window displays the actual speaker and the audio emerges from an attached pair of speakers.

Media		Mode, Application	Quality of Service
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 Related to the Item to Which the Pointer Points	- 500 ms, + 750 ms[a]

a. Pointer prior to audio for 500 ms; audio prior to pointer for 750 ms.

Table 15.5: *Quality of Service for synchronization purposes.*

Video and audio data are related by lip synchronization demands. Audio and the telepointer are related by the pointer synchronization demands. The relationship of video data and the telepointer is then yielded by a simple combination. In this example, we will define the following skews:

```
max skew (video ahead_of audio) = 80 ms
max skew (audio ahead_of video) = 80 ms
max skew (audio ahead_of pointer) = 740 ms
max skew (pointer ahead_of audio) = 500 ms
```

leading to the skew

```
skew (video ahead_of pointer) =< 820 ms
skew (pointer ahead_of video) =< 580 ms
```

In general, these requirements can be derived easily by the accumulation of the canonical skew as shown in the above example. The information gathered by the aggregation of media is of interest for the user, as well as for the multimedia system which must provide service according to these values.

In some cases, too many specifications of a synchronization skew exist: for example, a language lesson that includes audio data in English and Spanish, as well as the related video sequences. The course builder enforces lip synchronization between video and audio regardless of the language (+-80 ms). Additionally, the sentences need to be synchronized to switch from one language to the other (we chose a figure of 400 ms for this case). As lip synchronization is more demanding than the synchronization between the languages, this would lead to the following skew specification:

```
1. max skew (video ahead_of audio_english) = 80 ms
2. max skew (audio_english ahead_of video) = 80 ms
3. max skew (video ahead_of audio_spanish) = 80 ms
4. max skew (audio_spanish ahead_of video) = 80 ms
5. max skew (audio_english ahead_of audio_spanish) = 400 ms
6. max skew (audio_spanish ahead_of audio_english) = 400 ms
```

This specification consists of a set of related requirements in all need to be fulfilled, i.e., we must find "the greatest common denominator." For each canonical form, the derived skews are computed as follows:

```
1+2+3+4:
max skew (audio_english ahead_of audio_spanish) = 160 ms
max skew (audio_spanish ahead_of audio_english) = 160 ms

1+2+5+6:
max skew (video ahead_of audio_spanish) = 480 ms
max skew (audio_spanish ahead_of video) = 480 ms

3+4+5+6:
max skew (video ahead_of audio_english) = 480 ms
max skew (audio_english ahead_of video) = 480 ms
```

In the second step, the most stringent set of all requirements are selected:

```
1. max skew (video ahead_of audio_english) = 80 ms
2. max skew (audio_english ahead_of video) = 80 ms
3. max skew (video ahead_of audio_spanish) = 80 ms
4. max skew (audio_spanish ahead_of video) = 80 ms
5. max skew (audio_english ahead_of audio_spanish) = 160 ms
6. max skew (audio_spanish ahead_of audio_english) = 160 ms
```

In the following step, any set of synchronization requirements can be chosen from the above derived calculations:

```
max skew (video ahead_of audio_english) = 80 ms
max skew (audio_english ahead_of video) = 80 ms
max skew (audio_english ahead_of audio_spanish) = 160 ms
max skew (audio_spanish ahead_of audio_english) = 160 ms
```

In summary, the above procedures allow us to solve two related problems:

- If the applications impose a set of related synchronization requirements on a multimedia system, we are now able to find the most stringent demands.
- If a set of individual synchronization requirements between various data streams is provided, we are now able to compute the required relationships between each individual pair of streams.

Both issues arise in non-trivial systems when estimating, computing or negotiating the QoS as it is outlined in the next section.

15.5.2 Multimedia Synchronization Specification Methods

For the complex specification of multiple object synchronization, including user interaction, sophisticated specification methods must be used. The following requirements should be fulfilled by such a specification method:

- The method shall support object consistency and maintenance of synchronization specifications. Media objects should be kept as one logical unit in the specification.
- The method should supply an abstraction of the contents of a media object that allows the specification of temporal relations that refer to a part of the media object, but on the other hand regard, the media object as one logical unit.
- All types of synchronization relations should be easily described.
- The integration of time-dependent, as well as time-independent, media objects must be supported.
- The definition of QoS requirements must be supported by the specification method. It should preferably be expressed in the method directly.
- Hierarchical levels of synchronization must be supported to enable the handling of large and complex synchronization scenarios.

In the following sections, specification methods are assessed according to the criteria described above.

15.5.3 Interval-based Specifications

In the interval-based synchronization specification, the presentation duration of an object is regarded as interval. Two time intervals may be synchronized in 13 different modes [All83, Ham72]. Some of these types are invertible like before and after. Figure 15.30 shows a reduced set of seven non-invertible types according to [LG90a]. A simple synchronization specification method for two media objects is to use these seven types.

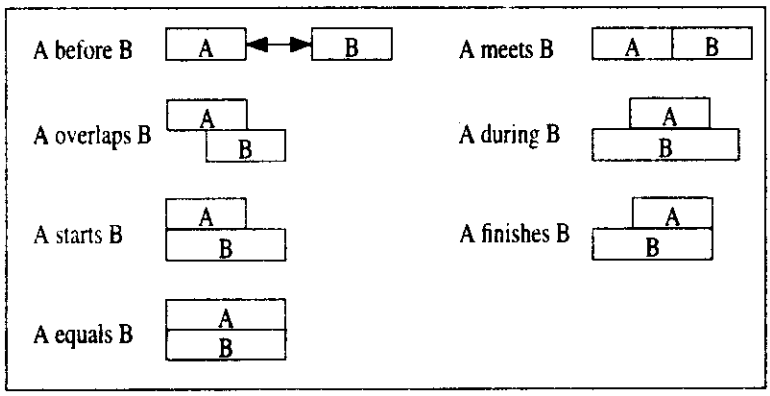
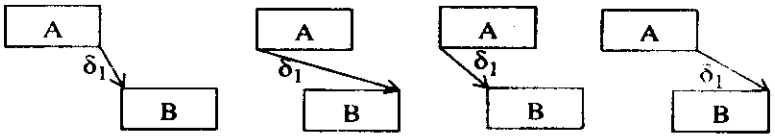


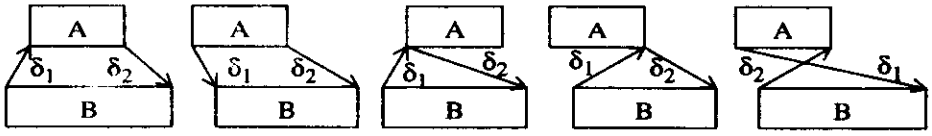
Figure 15.30: *Types of temporal relations between two objects.*

The enhanced interval-based model [WR94] is based on interval relations. The basic interval relations have already been shown in Figure 15.30. In the enhanced approach, 29 interval relations that are defined as disjunctions of the basic interval relations have been identified as relevant for multimedia presentations. To simplify the synchronization specification, ten operators have been defined that can handle these interval relations. These operations are shown in Figure 15.31. The duration of a presentation like A or B, as well as the delay d_i, are subsets of $+0$ because the duration of a presentation, as well as of a delay, may not be known in advance. In addition, the operations *beforeendof, delayed, startin, endin, cross* and *overlaps* d_i must not be 0.

Operations With One Delay Parameter



Operations With Two Delay Parameters



Operation With Three Delay Parameters

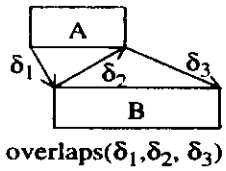


Figure 15.31: *Operations in the enhanced interval-based method.*

A slide show with slides $Slide_i$ $(1 \leq i \leq n)$ and an audio object *Audio* can be specified in this model by:

$$Slide_1 \ cobegin(0) \ Audio$$

$$Slide_i \ before(0) \ Slide_{i+1} \ (1 \leq i \leq n-1)$$

Lip synchronization between an audio object *Audio* and a video object *Video* is simply specified by:

$$Audio \ while(0,0) \ Video$$

The application example can be sketched as follows:

```
Audio1 while(0,0) Video
Audio1 before(0) RecordedInteraction
RecordedInteraction before(0) B1
P1 before(0) P2
P2 before(0) P3
P3 before(0) Interaction
P3 before(0) Animation
Animation while(2,5) Audio2
Interaction before(0) P4
```

This model allows the definition of a duration for time-dependent and time-independent media objects. This duration is used in the example to specify the duration of the presentation of the objects picture 1 to picture 3. The open duration of the user interaction can be specified by defining the duration as +0.

The advantage of this model is that it is easy to handle open LDUs, and therefore user interaction. It is possible to specify additional indeterministic temporal relations by defining intervals for durations and delays. Disjunction of operators can be used for specifications of presentation relations like not parallel. Therefore, it is a very flexible model that allows the specification of presentations with many run-time presentation variations.

The model does not include skew specifications. Despite the direct specification of time relations between media objects, it does not allow the specification of temporal relations directly between subunits of objects. Such relations must be defined indirectly by delay specifications, as shown in the while operation for the animation and audio in the application example, or by splitting the objects. The flexibility of specifiable presentations may lead to inconsistencies in run-time. For example, for two video objects A and B a *not* parallel relation has been defined. In run-time, A may be running and B may be coupled by a *before(0)* relation to the end of a user interaction. If this user interaction ends, video B must be started, but on the other hand, it may not be started because of the *not* parallel relation. It must be defined in the model how such inconsistencies must be handled in run-time or such potential inconsistencies must be detected before run-time and the specification must be rejected. Building of hierarchies is easily definable. The assessment of the enhanced

Advantages	Disadvantages
Logical objects can be kept	Complex specification
Good abstraction for media content	Additional specifications for skew QoS necessary
Easy integration of time-independent objects	Direct specification of time relations between media objects, but not for subunits of the media objects
Easy integration of interactive objects	Resolving of indeterminism at run-time may lead to inconsistencies
Specification of indeterministic temporal relations supported	

Table 15.6: *Assessment of the enhanced interval-based synchronization specification.*

interval-based method is summarized in Table 15.6.

15.5.4 Axes-based Synchronization

In an axes-based specification, the presentation events like the start and end of a presentation are mapped to axes that are shared by the objects of the presentation.

Synchronization Based on a Global Timer

For synchronization based on a *global timer*, all single-medium objects are attached to a time axis that represents an abstraction of real-time. This specification method is used, for example, in the Athena Muse project [HSA89], where synchronization is described by attaching all objects, independently of each other, to a time axis. Removing one object does not affect the synchronization of the other objects.

With modifications, this kind of specification is also used in the model of active

media [TGD91]. A world time is maintained, which is accessible to all objects. Each object can map this world time to its local time and moves along its local time axis. When the distortion between world time and local time exceeds a given limit, resynchronization with world time is required. A time axis mechanism is also used in QuickTime [DM92].

Synchronizing objects by means of a time axis allows a very good abstraction from the internal structure of single-medium objects and nested multimedia objects. Defining the beginning of a subtitle presentation relative to a scene in a video stream requires no knowledge of the related video frames. Since synchronization can only be defined based on fixed points of time, problems arise if objects include LDUs of unpredictable duration.

Moreover, synchronization based on one common global timer may not be sufficient for expressing the synchronization relations between different presentation streams. Depending on the coherence of these presentation streams, synchronization based on a common time axis might be either too strong or too weak. A possible solution is to define for each pair of media streams an additional QoS.

The use of the global timer demands that the media streams are able to synchronize themselves to the global timer. This may be difficult for audio streams because of the re-sampling problems. Therefore, the audio stream is often used as the global timer, but this still causes difficulties if several audio streams must be synchronized.

Figure 15.32 shows the specification of the application example. It can be seen that there is no natural possibility to handle the unpredictable duration of a user interaction.

The assessment of the time axis method is summarized in Table 15.7.

Synchronization Based on Virtual Axes

Virtual time axes, as used in the project Athena [HSA89] or the HyTime standard [Org92], are a generalization of the time axis approach. In this specification method, it is possible to specify coordinate systems with user-defined measurement units. A synchronization specification is performed according to these axes. It is also possible

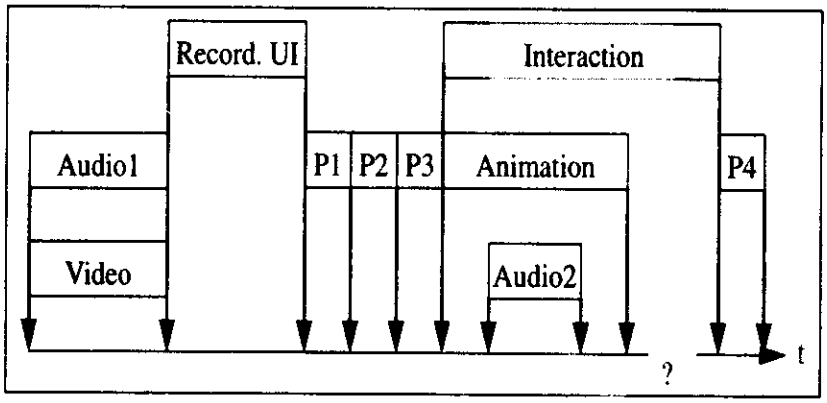


Figure 15.32: *Time axis synchronization specification example.*

to use several virtual axes to create a virtual coordinate space. An example is a music description by notes as shown in Figure 15.33. The tune frequency is defined by the position on the note lines. The sequence and duration is defined on the axis with the measurement unit beat.

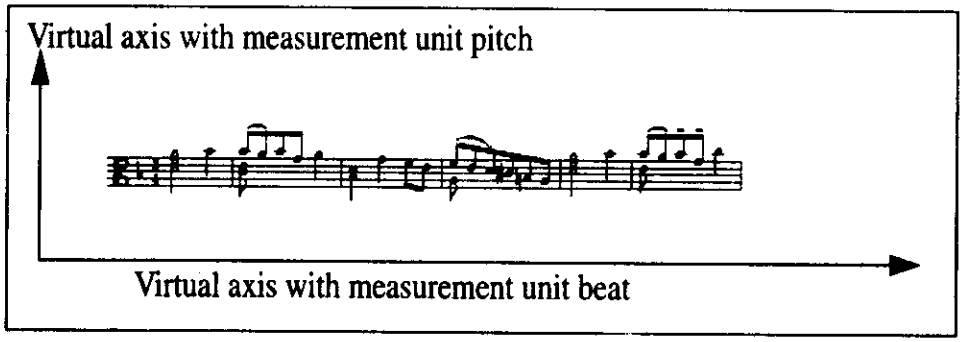


Figure 15.33: *Musical notes as an example of virtual axis.*

The mapping of the virtual axes to real axes is done in run-time. In the example shown in Figure 15.33, the pitch axis is mapped to the audio frequency and the beat axis is mapped to a timer.

The application example of Figure 15.13 can be realized in this approach by two time axes and an interaction axis (Figure 15.34). The latter should have interaction events as measurements units. The assessment of the virtual axes method is summarized

Advantage	Disadvantages
Easy to understand	Objects of unknown duration cannot be integrated, extensions to the model are required
Support of hierarchies easy to realize	Skew QoS must be specified indirectly by using the common time axis or additional QoS specifications must be given
Easy to maintain because of the mutual independence of objects	
Good abstraction for media contents	
Integration of time-independent objects is easy	

Table 15.7: *Assessment of the time axis synchronization specification.*

in Table 15.8.

15.5.5 Control Flow-based Specification

In *control flow-based specifications*, the flow of the concurrent presentation threads is synchronized in predefined points of the presentation.

Basic Hierarchical Specification

Hierarchical synchronization descriptions [Gro89, SS90] are based on two main synchronization operations: *serial synchronization* of actions and *parallel synchronization* of actions (Figure 15.35). In a hierarchical synchronization specification, multimedia objects are regarded as a tree consisting of nodes which denote serial or parallel presentation of the outgoing subtrees.

An action can be either atomic or compound. An atomic action handles the presen-

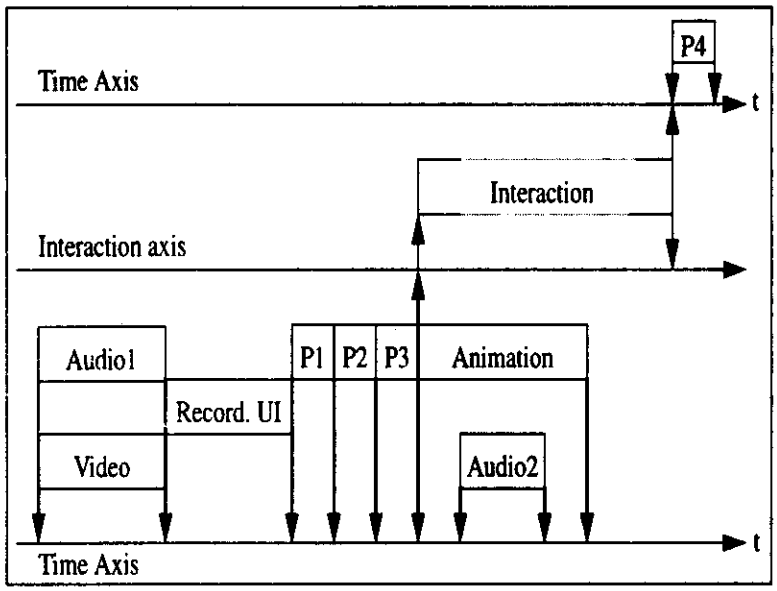


Figure 15.34: *Virtual time axis specification example.*

tation of either a single-media object, user input or delay. Compound actions are a combination of synchronization operators and atomic actions.

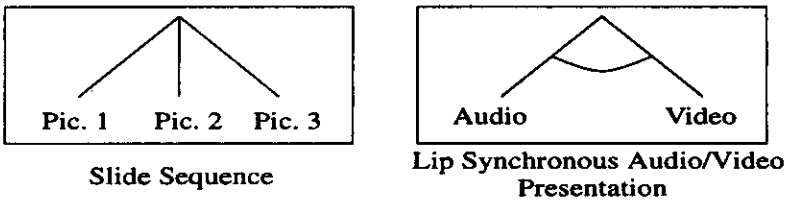


Figure 15.35: *Serial and parallel presentations.*

The introduction of a delay as a possible action [LG90a] allows the modeling of further synchronization behavior like delays in serial presentations and delayed presentations of objects in a parallel synchronization.

Hierarchical structures are easy to handle and widely used. Restrictions from the hierarchical structure arise from the fact that each action can only be synchronized at its beginning or end. This means, for example, that the presentation of subtitles at

Advantages	Disadvantages
Easy to understand	Skew QoS defined only indirectly or through additional specifications
Often specification is made according to the problem space possible	Specification may become complex with many axes
Good possibility for building hierarchies	Mapping of axes at run-time may be complex and time-consuming
Easy to maintain because objects are kept as units and mutually independent objects	
Good abstraction for media content	
Easy integration of time-independent media objects	
Interactive objects can be included using specialized axes	

Table 15.8: *Assessment of the virtual axis synchronization specification.*

parts of a video stream requires the video stream to be split into several consecutive components. This can be seen in Figure 15.36 for the synchronization specification of the animation and audio block in the example introduced in Section 15.2.3.3. The animation must be split into the parts Animation 1, Animation 2 and Animation 3 to be correctly synchronized with the audio block.

Accordingly, a synchronized multimedia object used as a component in another synchronization can no longer be regarded as an abstract unit if it has to be synchronized between the beginning and end of its presentation. That is to say, hierarchical structures do not support adequate abstraction for the internal structure of multimedia objects. In addition, there are synchronization conditions which cannot be represented using hierarchical structures. For example, the three objects shown in Figure 15.37 are presented in parallel, where any pair of objects is synchronized but always independently of the third object. To specify this synchronization, additional synchronization points must be used.

The assessment of the basic hierarchical method is summarized in Table 15.9.

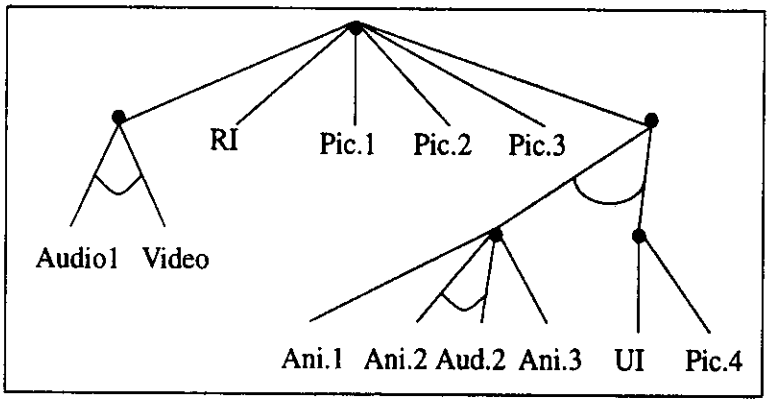


Figure 15.36: *Hierarchical specification example (RI = Recorded Interaction, Pic. = Picture, Aud. = Audio, Ani. = Animation, UI = User Interaction).*

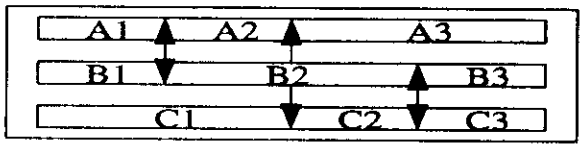


Figure 15.37: *Non-describable synchronization.*

Reference Points

In the case of synchronization via reference points [Ste90, BHLM92], time-dependent single-medium objects are regarded as sequences of closed LDUs. The start and stop times of the presentation of a media object, in addition to the start times of the subunits of time-dependent media objects, are called *reference points.* Synchronization between objects is defined by connecting reference points of media objects. A set of connected reference points is called a *synchronization point.* The presentation of the subunits that participate in the same synchronization point must be started or stopped when the synchronization point is reached. This approach to synchronization specifies temporal relations between objects without explicit reference to time.

Like synchronization based on a time axis, this description allows synchronization at any time during the presentation of an object; moreover, object presentations of

Advantages	Disadvantages
Easy to understand	Additional description of skew QoS necessary
Natural support of hierarchies	For the presentation of time-independent media objects, presentation durations must be added
Integration of interactive objects is easy	Splitting of media objects for synchronization purposes is necessary
	No adequate abstractions for media object contents
	Some synchronization scenarios cannot be described

Table 15.9: *Assessment of the basic hierarchical synchronization specification.*

unpredictable duration can be integrated easily. This type of specification is also very intuitive to use.

A drawback of reference point synchronization is that it requires mechanisms for detecting inconsistencies. In addition, synchronization based on reference points does not allow for specification of delays in a multimedia presentation. To solve this problem, Steinmetz [Ste90] proposes time specifications which specify explicit real-time-based delays. The inclusion of timers also solves this problem. The specification based on a global timer can be regarded as a subset of the reference point synchronization: a timer according to Figure 15.10 can be used as global timer and all objects refer only to this timer.

In a reference point synchronization specification, the coherence between data streams can be described by specifying a suitable set of synchronization points between the two data streams. A close lip synchronization with a maximal skew of +/- 80 ms can be realized by setting a synchronization point, for example, every second frame of a video (Figure 15.38). If no lip synchronization is required, it may be sufficient to set a synchronization point every 10 frames of the video. Therefore, the specification

of the skew QoS is directly integrated into this specification method.

An example of the synchronized integration of time-dependent and time-independent media objects is shown in Figure 15.39. Starting and stopping a slide presentation are initiated by reaching suitable LDUs in the audio presentation.

The application example can be completely specified with the reference point synchronization model shown in Figure 15.40.

Hierarchies in the reference point synchronization method can be created by regarding a set of synchronized objects as one object, with the start of the first object and end of the last object as reference points. Virtual reference points for this presentation can be specified and mapped to the reference points within the hierarchy. The semantic of this mapping can become complex in the case that objects of unknown duration are included in the hierarchy. The assessment of the reference point method is summarized in Figure 15.41.

Timed Petri Nets

Another type of specification is based on *petri nets* [LG92, LG91b] that are extended with duration specifications at various places, a kind of *timed petri net.*

The rules for a timed petri net are:

- A transition fires, if all input places contain a nonblocking token.
- If a transition fires, a token is removed from each input place and a token is added to each output place.
- A token that is added to a new place is blocked for the duration that is assigned to this place.

A slide show can be specified by assigning corresponding durations to the places (Figure 15.42).

For time-dependent media objects, each place in the petri net represents an LDU. Lip synchronization can be modeled on the basis of connecting appropriate LDUs

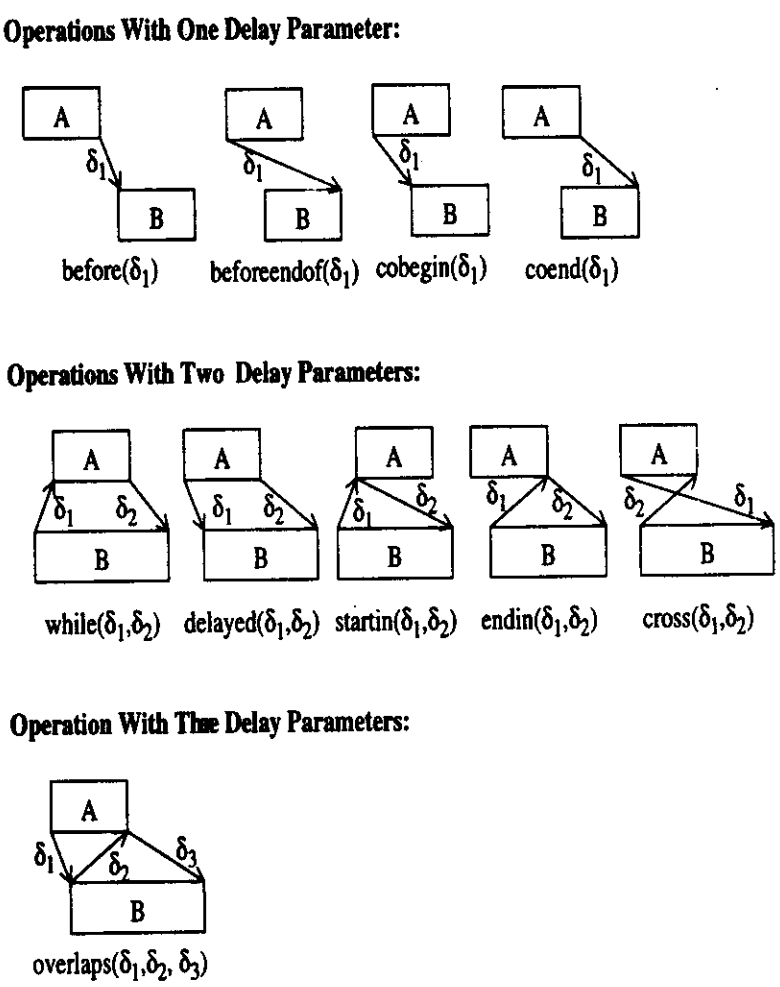


Figure 15.38: *Lip synchronization in the reference point synchronization model.*

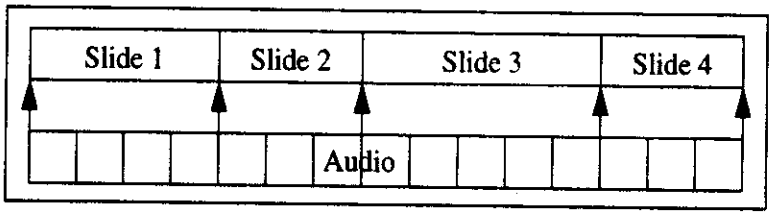


Figure 15.39: *Example of a slide show with an audio sequence in the reference point model.*

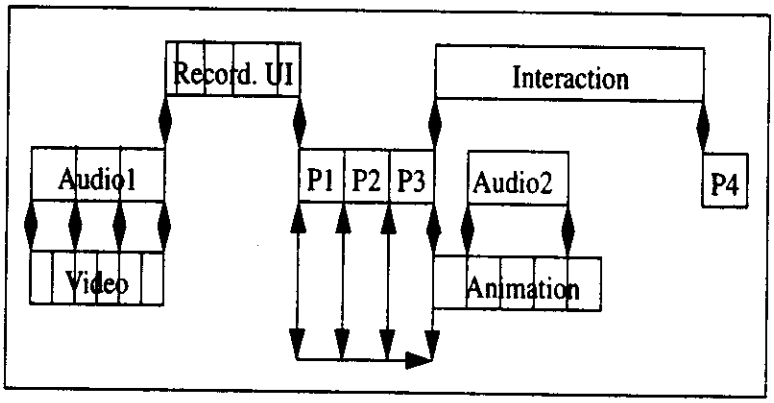


Figure 15.40: *Reference point synchronization specification example (with the integration of time-dependent and time-independent media objects, as well as closed and open LDUs).*

by transitions (Figure 15.43).

It is also possible to combine a set of consecutive LDUs to one place as long as no inter-object synchronization exists between these LDUs and others. A hierarchy can be constructed by creating subnets that are assigned to a place. The duration of the longest path in the subnet is assigned to the place (Figure 15.44).

The application example of Figure 15.13 can be modeled as shown in Figure 15.45. The subnets are not shown because they can be created by the straightforward use of the techniques described above.

Timed petri nets allow all kinds of synchronization specifications. The main drawbacks are the complex specifications and the insufficient abstraction of media object

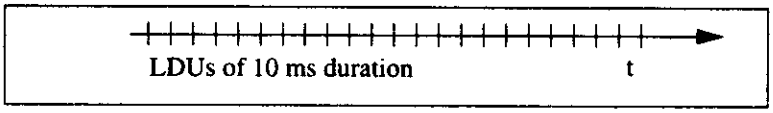


Figure 15.41: *Assessment of the reference point synchronization specification.*

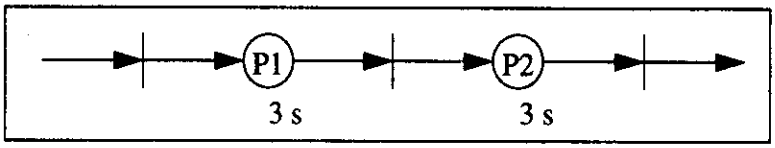


Figure 15.42: *Petri net specification of a slide show.*

content because, much like the hierarchical specification, the media objects must be split into subobjects. The assessment of the timed petri net method is summarized in Table 15.10.

15.5.6 Event-based Synchronization

In the case of event-based synchronization, presentation actions are initiated by synchronization events, e.g., as in HyTime and HyperODA [App89]. Typical presentation actions are:

- Start a presentation.
- Stop a presentation.
- Prepare a presentation.

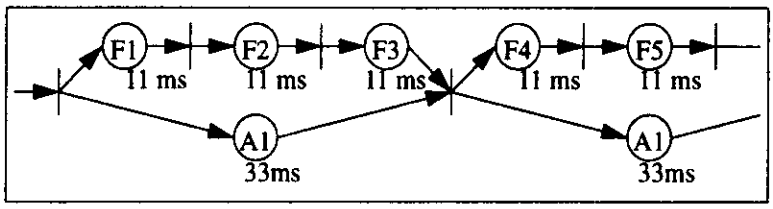


Figure 15.43: *Petri net lip synchronization.*

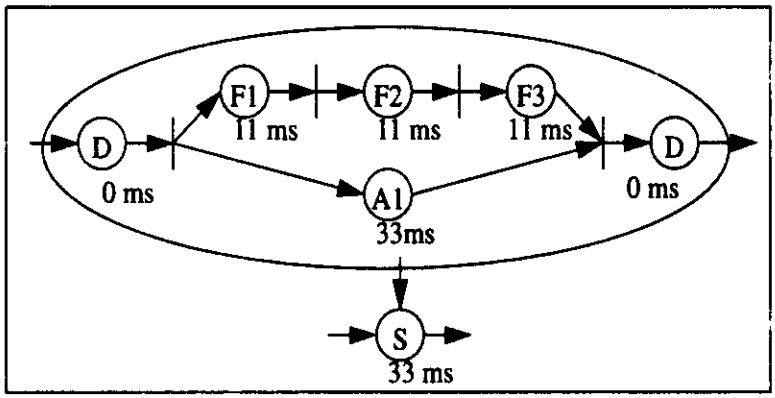


Figure 15.44: *Petri net hierarchy comprised of the synchronization of A1 and F1 to F3.*

The events that initiate presentation actions may be external (e.g., generated by a timer) or internal to the presentation generated by a time-dependent media object that reaches a specific LDU.

Table 15.11 sketches an event-based synchronization for parts of the application example.

This type of specification is easily extended to new synchronization types. The major drawback is that this type of specification is difficult to handle in the case of realistic scenarios. The user is lost in this state transition type of synchronization specification, hence creation and maintenance becomes difficult. The assessment of the event-based method is summarized in Table 15.12.

15.5.7 Scripts

A *script* in this context is a textual description of a synchronization scenario [IBM90, TGD91]. Elements of scripts are activities and subscripts. Often, scripts become full programming languages extended by timing operations. Scripts may rely on different specification methods.

A typical example is a script that is based on the basic hierarchical method and supports three main operations: serial presentation, parallel presentation and the

Advantages	Disadvantages
Hierarchies can be created	Difficult to handle
Easy integration of time-independent objects	Complex specification
Easy integration of interactive objects	Splitting of media objects
Integrated skew QoS	Insufficient abstraction of media object content

Table 15.10: *Assessment of the petri net synchronization specification.*

Event / **Action**	**Start**	**Audio1.stop**	**Timer1. ready**	**...**
Audio1	start			
Video	start			
Pic.1		start	stop	
Timer1		start(3)		
Pic.2			start	
...				

Table 15.11: *Event-based specification example.*

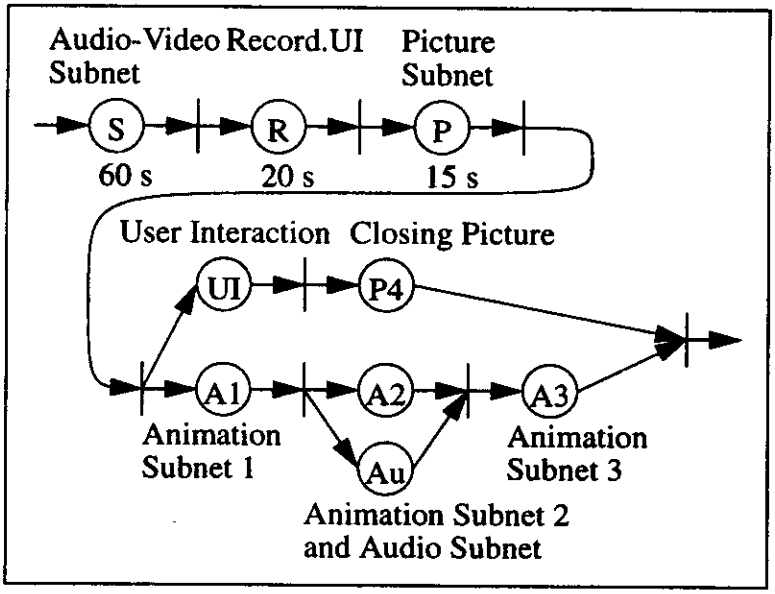


Figure 15.45: *Petri net specification example.*

repeated presentation of a media object.

The following example sketches a script for the application example from Figure 15.13. >> denotes a serial presentation, & denotes a parallel presentation and n denotes a presentation repeated n times. ([TGD91]):

```
activity DigAudio Audio(''video.au");
activity SMP Video(''video.smp");
activity XRecorder Recorder(''window.rec");
activity Picture Picture1(''picture1.jpeg");
activity Picture Picture2(''picture2.jpeg");
activity Picture Picture3(''picture3.jpeg");
activity Picture Picture4(''picture4.jpeg");
activity StartInteraction Selection;
activity DigAudio AniAudio(''animation.au");
activity RTAnima Animation(''animation.ani");
```

Advantages	Disadvantages
Easy integration of interactive objects	Difficult to handle
Easily extensible by new events	Complex specification
Flexible because any event can be specified	Hard to maintain
	Integration of time-dependent objects by using additional timers
	Separate descriptions of skew QoS necessary
	Difficult use of hierarchies

Table 15.12: *Assessment of the event-based synchronization specification.*

```
script Picture_sequence  3Pictures=  Picture1.Duration(5) >>
Picture2.Duration(5) >>
Picture3.Duration(5);

script Lipsynch AV = Audio & Video;
script AniComment AA = Animation & AniAudio.Translate(2);
script Multimedia Application_example {
AV >>
Record. UI >>
3Pictures >>
( (Selection >> Picture4) &
  AA )
```

Scripts are very powerful because they represent full programming environments. A disadvantage is that this method is more procedural than declarative. The declarative approach seems to be more easy for the user to handle. The assessment of the script method is summarized in Table 15.13.

Advantages	Disadvantages
Good support for hierarchies	Difficult to handle
Logical objects can be kept	Complex specification
Easy integration of time-independent objects	Implicit usage of common timers necessary
Easy integration of interactive objects	Special constructs for skew QoS necessary
Easily extensible by new synchronization constructs	
Flexible because programmable	

Table 15.13: *Assessment of the script synchronization specification.*

15.5.8 Comment

The presented synchronization specification methods have different specification capabilities and are different from the point of user-friendliness, but many of them just present different "views" of the same problem.

The different specification capabilities restrict the mapping between specifications of different methods to the common subset.

The selection of an appropriate specification method depends on the targeted application and on the existing environment. As the temporal behavior of multimedia objects is only one part of a presentation, we must keep in mind the context as it may be an audio/video editor or an MHEG presentation tool. The selected method must fit into the selected environment. There is no "best" or "worst" solution. For simple presentations without user interaction, the method based on a global timer seems to be appropriate. For complex structures with interaction, for example, the reference point model seems to be suitable.

In many cases, users will not directly specify the synchronization using a specific specification method. They will instead use a graphical authoring system that may

produce specifications based on different methods. Experience shows that usually one of these specification methods underlies the construction of the user interface and therefore indirectly the advantages and disadvantages of the method reflect themselves at the user interface. In addition, many authoring systems allow the author to step out of the high-level graphical representation and to specify a complex synchronization directly at the lowest synchronization specification level, e.g., the textual level provided by the underlying method, which is not the best way to proceed.

15.6 Case Studies

Some interesting approaches to multimedia synchronization are described in this section and classified according to the reference model presented previously. In particular, we analyze synchronization aspects in standards of multimedia information exchange and the respective run-time environments and prototype multimedia systems which comprise several layers of the synchronization reference model.

15.6.1 Synchronization in MHEG

The generic space in *MHEG* provides a virtual coordinate system that is used to specify the layout and relation of content objects in space and time according to the virtual axes-based specification method. The generic space has one time axis of infinite length measured in *Generic Time Units (GTUs)*. The MHEG run-time environment must map the GTUs to *Physical Time Units (PTUs)*. If no mapping is specified, the default is one GTU mapped to one millisecond. Three spatial axes (X=latitude, Y=longitude, Z=altitude) are used in the generic space. Each axis is of finite length in an interval of $[-32768, +32767]$. Units are *Generic Space Units (GSUs)*. Also, the MHEG engine must perform mapping from the virtual to the real coordinate space.

The presentation of content objects is based on the exchange of action objects sent to an object. Examples of actions are prepare to set the object in a presentable state, run to start the presentation and stop to end the presentation.

Action objects can be combined to form an action list. Parallel action lists are executed in parallel. Each list is composed of a delay followed by delayed sequential actions that are processed serially by the MHEG engine, as shown in Figure 15.46.

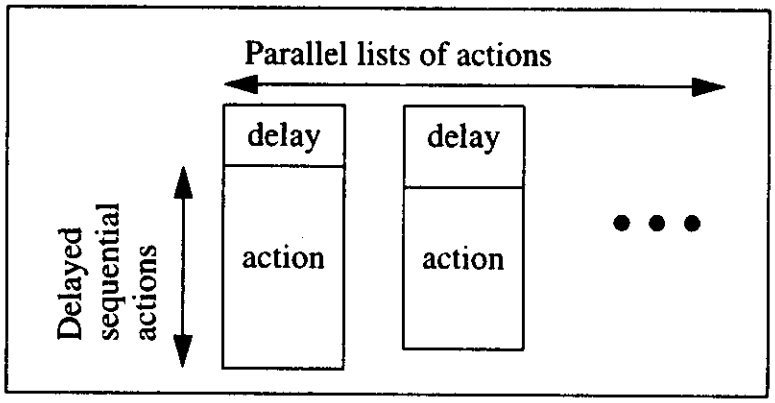


Figure 15.46: *Lists of actions.*

By using links it is possible to synchronize presentations based on events. Link conditions may be associated with an event. If the conditions associated with a link are fulfilled, the link is triggered and actions assigned to this link are performed. This is a type of event-based synchronization.

MHEG Engine

At the European Networking Center in Heidelberg, an MHEG engine [Gra94] has been developed. The MHEG engine is an implementation of the object layer. The architecture of the MHEG engine is shown in Figure 15.47.

The *Generic Presentation Services* of the engine provide abstractions from the presentation modules used to present the content objects. The *Audio/Video-Subsystem* is a stream layer implementation. This component is responsible for the presentation of the continuous media streams, e.g., audio/video streams. The *User Interface Services* provide the presentation of time-independent media, like text and graphics, and the processing of user interactions, e.g., buttons and forms.

The MHEG engine receives the MHEG objects from the application. The *Object*

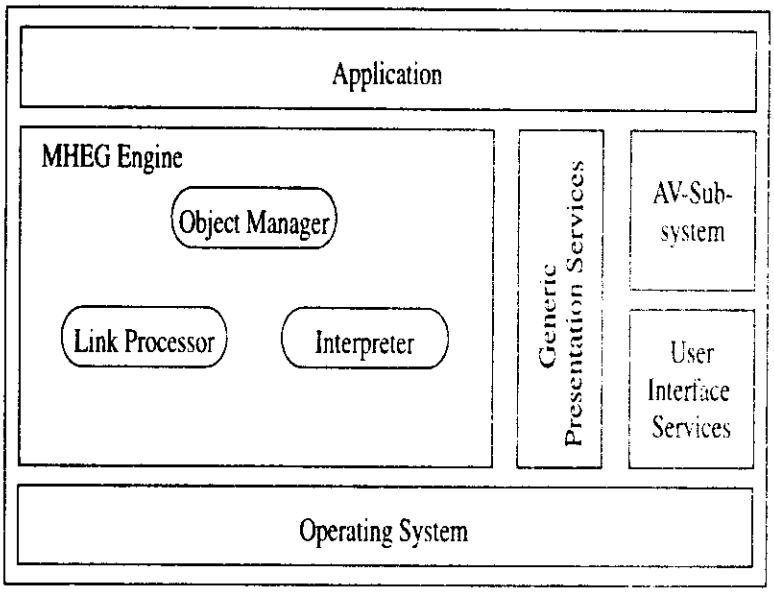


Figure 15.47: *Architecture of an MHEG engine.*

Manager manages these objects in the run-time environment. The *Interpreter* processes the action objects and events. It is responsible for initiating the preparation and presentation of the objects. The *Link Processor* monitors the states of objects and triggers links, if the trigger conditions of a link are fulfilled.

The run-time system communicates with the presentation services by events. The *User Interface Services* provide events that indicate user actions. The *Audio/Video-Subsystem* provides events about the status of the presentation streams, like end of the presentation of a stream or reaching a cuepoint in a stream.

Summary

MHEG is a standardized exchange format that is used as the exchange format at the object layer. The synchronization is based on the virtual axes- and event-based methods. An MHEG engine represents the object layer run-time environment. The object layer implementation of the described engine is based on media servers. The Audio/Video-Subsystem represents the stream layer. Figure 15.48 shows the relation

to the synchronization reference model.

Regarding distributed environments, the processing model of MHEG has the following drawback: the duration between the preparation and display action is coded in the MHEG object, but the duration depends on the run-time environment; therefore, this duration should be computed by the MHEG engine.

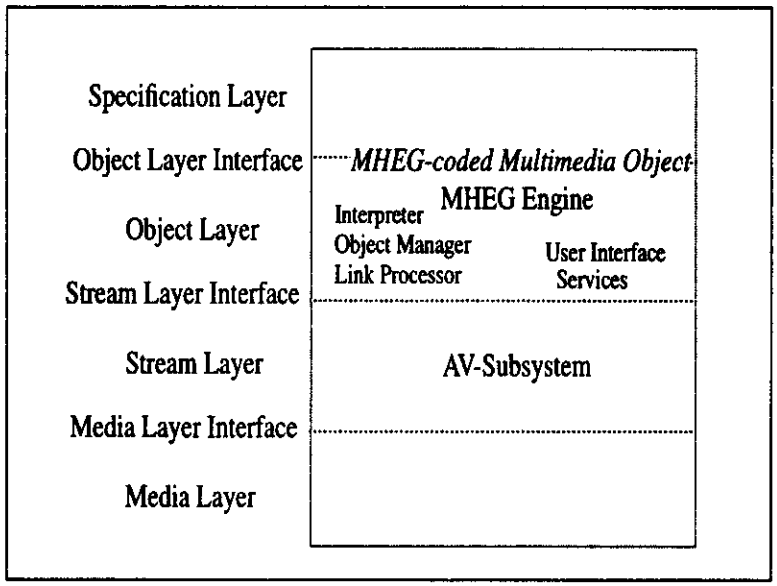


Figure 15.48: *Classification of MHEG and MHEG engine components according to the reference model.*

15.6.2 HyTime

HyTime (Hypermedia/Time-based Structuring Language) is an international standard (ISO/IEC 10744) [Org92] for the structured representation of hypermedia information. HyTime is an application of the *Standardized General Markup Language (SGML)* [Smi89] (see also Section 13.3).

SGML is designed for document exchange, whereby the document structure is of great importance, but the layout is a local matter. The logical structure is defined by

markup commands that are inserted in the text. The markups divide the text into SGML elements. For each SGML document, a *Data Type Definition (DTD)* exists which declares the element types of a document, the attributes of the elements and how the instances are hierarchically related. A typical use of SGML is the publishing industry where an author is responsible for the content and structure of the document, and the publisher is responsible for the layout. As the content of the document is not restricted by SGML, elements can be of type text, picture or other multimedia data.

HyTime defines how markup and DTDs can be used to describe the structure of hyperlinked time-based multimedia documents. HyTime does not define the format or encoding of elements. It provides the framework for defining the relationship between these elements.

HyTime supports addresses to identify a certain piece of information within an element, linking facilities to establish links between parts of elements and temporal and spatial alignment specifications to describe the relationships between media objects.

HyTime defines architectural forms that represent SGML element declaration templates with associated attributes. The semantic of these architectural forms is defined by HyTime. A HyTime application designer creates a HyTime-conforming DTD using the architectural forms he/she needs for the HyTime document. In the HyTime DTD each element type is associated with an architectural form by a special HyTime attribute.

The HyTime architectural forms are grouped into the following modules:

- The *Base Module* specifies the architectural forms that comprise the document.
- The *Measurement Module* is used to add dimensions, measurement and counting to the documents. Media objects in the document can be placed along with the dimensions.
- The *Location Address Module* provides the means to address locations in a document. The following addressing modes are supported:

  - *Name Space Addressing Schema:* Addressing to a name identifying a piece of information.
  - *Coordinate Location Schema:* Addressing by referring to an interval of a coordinate space if measuring along the coordinate space is possible. An example is to address to a part of an audio sequence.
  - *Semantic Location Schema:* Addressing by using application-specific constructs.

- The *Scheduling Module* places media objects into Finite Coordinate Spaces (FCSs). These spaces are collections of application-defined axes. To add measures to the axes, the measurement module is needed. HyTime does not know the dimension of its media objects. So-called events are used for the presentation of media objects. An event is an encapsulation of a media object and comprises the layout specification related to an FCS. The events can be placed absolutely or relatively to other events within the FCSs.

- The *Hyperlink Module* enables building link connections between media objects. Endpoints can be defined using the location address, measurements and scheduling modules.

- The *Rendition Module* is used to specify how the events of a source FCS, that typically provides a generic presentation description, are transformed to a target FCS that is used for a particular presentation. During the mapping, presentation-related modifications are executed, e.g., changing the color representation, projection of the dimensions from the source to the target FCS or scaling of the presentation.

HyTime Engine

The task of a HyTime engine is to take the output of an SGML parser, to recognize architectural forms and to perform the HyTime-specific and application-independent processing. Typical tasks of the HyTime engine are hyperlink resolution, object addressing, parsing of measures and schedules, and transformation of schedules and dimensions. The resulting information is then provided to the HyTime application.

The HyTime engine, HyOctane [BRRK94], developed at the University Massachusetts at Lowell, has the following architecture: an SGML parser takes as input the application data type definition that is used for the document and the HyTime document instance. It stores the document object's markups and contents, as well as the applications DTD in the SGML layer of a database. The HyTime engine takes as input the information stored in the SGML layer of the database. It identifies the architectural forms, resolves addresses from the location address module, handles the functions of the scheduling module and performs the mapping specified in the rendition module. It stores the information about elements of the document that are instances of architectural forms in the HyTime layer of the database. The application layer of the database stores the objects and their attributes, as defined by the DTD. An application presenter gets the information it needs for the presentation of the database content, including the links between objects and the presentation coordinates to use for the presentation, from the database.

Summary

HyTime is applicable to many application areas. It does not standardize content formats, encoding, document types or specific SGML DTDs. It provides a framework for addressing portions of hypermedia document contents and the definition of linking, alignment and synchronization. In the context of the synchronization reference model, a HyTime document, together with its DTD, can be used as input to the object layer. The synchronization is based on the virtual axes synchronization method. The SGML- and HyTime-related preprocessing is done by the HyTime engine in the object layer. The application presenter provides the other object layer and stream layer functionalities. Figure 15.49 shows the relation to the synchronization reference model.

Other classification possibilities are to regard the database as an object layer interface format or to use the database to generate an MHEG specification. In the latter case, the HyTime engine can be regarded as part of a format conversion tool.

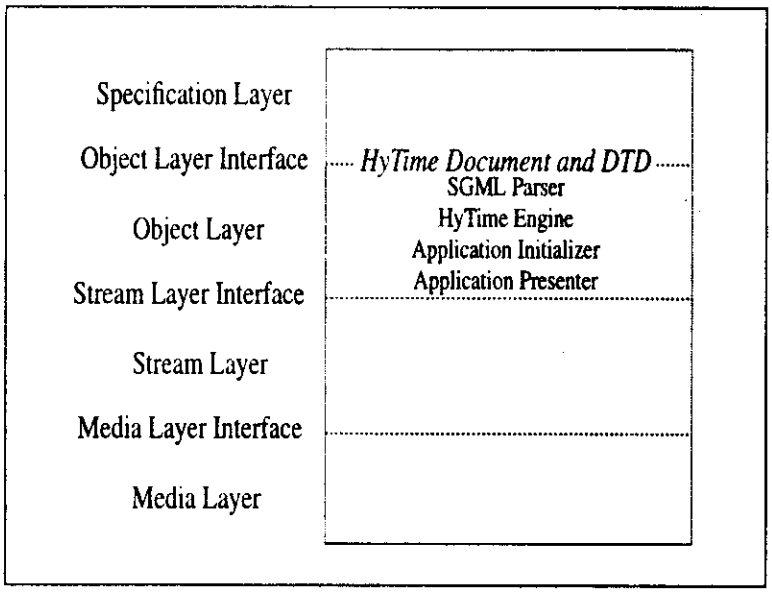


Figure 15.49: *Classification of HyTime and the HyTime engine according to the synchronization reference model.*

15.6.3 Firefly System

The objective of the approach of Buchanan and Zellweger [BZ93a, BZ93b] is to automatically generate consistent presentation schedules for interactive multimedia documents that comprise media objects of predictable behavior (like audio and video) and objects of unpredictable behavior (like user interactions). The generation algorithm is comprised of two phases. At the first phase, before execution of the presentation, high-level temporal specifications for a document are used to compute a presentation schedule, as far as possible without knowing the unpredictable durations. In the second phase during the presentation, the scheduling, depending on unpredictable durations, is incorporated.

The specification of the temporal constraints distinguishes media-level specifications that describe the temporal behavior of individual media objects and document-level specifications that describe the temporal behavior of a complete multimedia document, in particular the temporal relations between single media objects. Media

items are used for the media-level specification. They provide a reference to a media object and are used to describe the temporal behavior of this media object. A media item consists of:

- *Events,* which represent points in time at the presentation of a media object. They are comparable to a reference point.
- *Durations,* which specify the duration between two subsequent events in a media object. A duration is represented by a triple of values: *minDuration, optDuration* and *maxDuration.* If the three values are equal, the presentation duration is fixed. If they specify an interval, the presentation is adjustable. No values are assigned for an unpredictable duration.
- *Costs,* which can be used as measurement for the degree of degradation in the case of stretching the presentation toward the maximal duration, and respectively shrinking it toward the minimum duration.

A document-level specification consists of:

- *Media items,* which are involved in the presentation.
- *Temporal constraints,* which are used to describe explicit temporal relations between events in one or more media items. Temporal constraints are classified into temporal equalities that describe a fixed temporal relation between two events (e.g., same time, one event 10 s before the other), and temporal inequalities that describe a temporal relation without a specified time (e.g., one event before the other, one event at least 10 s and at last 20 s before the other).
- *Operations,* which can be associated with an event and include non-altering presentation-related operations, like increase volume of an audio presentation, and time-altering operations, like increase-playback-speed.
- *Duration and costs,* which can be described according to the media level. At the document level, this is used to describe a different behavior for several instances of one media item in a document.

- *Unpredictable event control*, which allows activation and deactivation of unpredictable events.

To support the development of temporal specifications, a graphical representation of the specification is supported. The synchronization specification method is a combination of reference point and interval-based synchronization. The scheduler for the presentation that is located at the object layer is divided into two parts: the compile-time scheduler and the run-time scheduler. The compile-time scheduler constructs a main schedule that controls the parts of a document that are predictable, and auxiliary schedules that control the parts of the document that depend on unpredictable events. It is an example of off-line schedule computation at the object layer.

The algorithm contains three parts:

- In the obtaining durations and costs step, the duration and costs for each media item are obtained. To do this, the media and document-level specifications for a media item are unified and time-altering operations are incorporated into the computation of the durations.

- In the finding connected components step, a union-find algorithm is used to find connected parts of a document. Two events are in the same connected component if they are related by a predictable duration or a temporal constraint. The connected components are called predictable, if there are no unpredictable events that trigger events of the component. Otherwise, they are called unpredictable.

- The assigning times to events step computes for each event in a connected component the time for that event with respect to the start time of the component. It uses a simplex algorithm with the durations and temporal constraints as constraints for the algorithm and the minimization of the costs as its objective function.

- In the creating commands step, the previous results are used to create the commands for the execution. A command includes a time when it must be

executed, the media item to process, an associated event, the list of unpredictable events to be activated or deactivated and the operations to be executed. All commands of the predictable components are integrated in the main schedule. For each unpredictable component, a separate auxiliary schedule is constructed. To improve performance for a continuous media object with units of fixed durations, only the start of the complete media object and events that refer to other media objects are considered, not every single event within the media item. It is assumed that this stream, like presentation scheduling, is done separately.

The run-time scheduler is an example of on-line schedule computation at the object layer and controls the document clock, the execution schedule and handles the unpredictable events. After the compile-time scheduler has produced the schedules, the run-time scheduler copies the main schedule into the execution schedule and starts the document clock. If the document clock reaches a time with an associated command, it initiates the command. If an activated unpredictable event occurs that triggers an unpredictable component, the run-time scheduler merges the corresponding schedule into the execution schedule taking the actual document time as start time for the first command in the schedule to merge. Because unpredictable components may be triggered several times, the run-time scheduler marks the instances in the execution schedule to be able to distinguish the commands for the different instances of an unpredictable schedule.

Summary

The Firefly system provides complete synchronization support. At the specification layer, an editor is provided. The temporal relations based on the reference point and interval-based specification methods are used at the object layer interface. The Scheduler provides off-line and on-line computation of presentation schedules at the object layer. The schedule of streams is only initiated at the object layer, the execution is located at the stream layer. Figure 15.50 shows the relation to the synchronization reference model.

The system provides well-organized scheduling planning and integration of unpre-

dictable durations. Currently, the system does not consider media preparation durations, presentation restrictions by insufficient or missing local resources or delays introduced by networks.

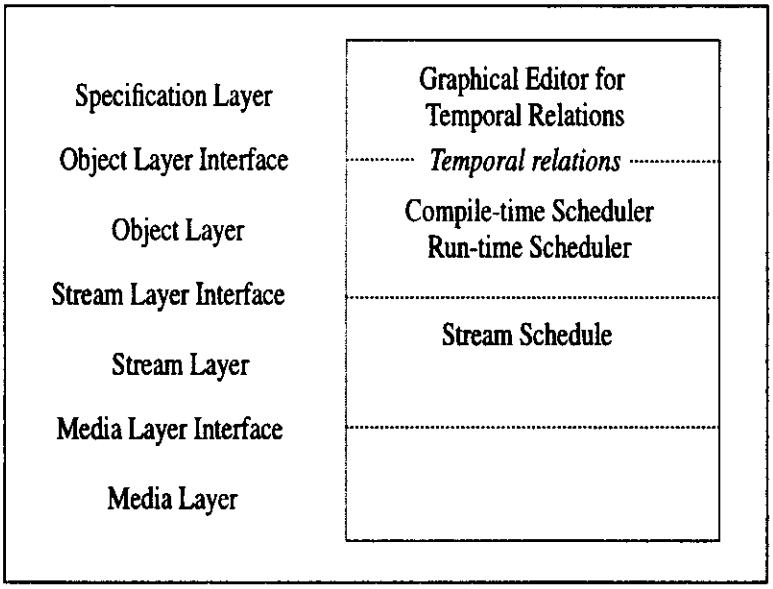


Figure 15.50: *Classification of the Firefly system according to the synchronization reference model.*

15.6.4 MODE

The *MODE (Multimedia Objects in a Distributed Environment)* system [Bla93], developed at the University of Karlsruhe, is a comprehensive approach to network transparent synchronization specification and scheduling in heterogeneous distributed systems. The heart of MODE is a distributed multimedia presentation service which shares a customized multimedia object model, synchronization specifications and QoS requirements with a given application. It also shares knowledge about networks and workstations with a given run-time environment. The distributed service uses all this information for synchronization scheduling when the presentation of a compound multimedia object is requested from the application.

Thereby, it adapts the QoS of the presentation to the available resources, taking into account a cost model and the QoS requirements given by the application.

The MODE system contains the following synchronization-related components:

- The *Synchronization Editor* at the specification layer, which is used to create synchronization and layout specifications for multimedia presentations.
- The MODE *Server Manager* at the object layer, which coordinates the execution of the presentation service calls. This includes the coordination of the creation of units of presentation (presentation objects) out of basic units of information (information objects) and the transport of objects in a distributed environment.
- The *Local Synchronizer*, which receives locally the presentation objects and initiates their local presentation according to a synchronization specification.
- The *Optimizer*, part of the MODE Server Manager, which performs the planning of the distributed synchronization and chooses presentation qualities and presentation forms depending on user demands, network and workstation capabilities and presentation performance.

Synchronization Model

In the MODE system, a synchronization model based on synchronization at reference points is used [BHLM92]. This model is extended to cover handling of time intervals, objects of unpredictable duration and conditions which may be raised by the underlying distributed heterogeneous environment.

A synchronization specification created with the Synchronization Editor and used by the Synchronizer is stored in textual form. The syntax of this specification is defined in the context-free grammar of the Synchronization Description Language. This way, a synchronization specification can be used by MODE components, independent of their implementation language and environment.

MODE distinguishes between dynamic basic objects and static basic objects. A presentation of a dynamic basic object is composed of a sequence of presentation

objects. This corresponds to a stream of LDUs. The index of each presentation object is called a reference point. The presentation of a static basic object, that may be a time-independent media object, as well as an interactive object, has only two reference points, the beginning and the end of the presentation. The description of a reference point, together with the corresponding basic object, is called a synchronization element, denoted in the form BasicObject.ReferencePoint. Two or more synchronization elements can be combined into a synchronization point. An entire inter-object synchronization is defined by the list of all synchronization points.

A presentation quality can be specified for each basic object. It is described by a set of attributes comprising an attribute name, preferred value and value domain that describes all possible values for this attribute.

Local Synchronizer

The Local Synchronizer performs synchronized presentations according to the synchronization model introduced above. This comprises both intra-object and inter-object synchronization. For intra-object synchronization, a presentation thread is created which manages the presentation of a dynamic basic object. Threads with different priorities may be used to implement priorities of basic objects. All presentations of static basic objects are managed by a single thread.

Synchronization is performed by a signaling mechanism. Each presentation thread reaching a synchronization point sends a corresponding signal to all other presentation threads involved in the synchronization point. Having received such a signal, other presentation threads may perform acceleration actions, if necessary. After the dispatch of all signals, the presentation thread waits until it receives signals from all other participating threads of the synchronization point; meanwhile, it may perform a waiting action.

Planning and Execution of the Distributed Presentation

Before starting any presentation, the Optimizer is invoked. The Optimizer uses a heuristic search algorithm taking the special conditions of a distributed environment

like multiple steps of the synchronization in a distributed environment, multiple communication patterns, buffering requirements and merging into account. It uses information about the network, like the available bandwidth, service qualities and available resources at the workstation, as well as information about the processing demands for media objects. This information is provided to the Optimizer by environment and application media descriptions [Bla92].

The planning result determines the achievable quality value for each presentation attribute according to both user demands and network and workstation resources. The result of the planning process is the MODE Flow Graph [Bla91b] that describes the times and nodes at which operations must be executed. The partitioned Flow Graph is delivered to the involved nodes and executed at run-time by the distributed MODE Server Manager.

Exceptions Caused by the Distributed Environment

The correct temporal execution of the plan depends on the underlying environment, if the workstations and network provide temporal guarantees for the execution of the operations. Therefore, MODE provides several guarantee levels. If the underlying distributed environment cannot give full guarantees, MODE considers the possible error conditions. Three types of actions are used to define a behavior in the case of exception conditions, which may be raised during a distributed synchronized presentation: 1) A waiting action can be carried out if a presentation of a dynamic basic object has reached a synchronization point and waits longer than a specified time at this synchronization point. Possible waiting actions are, for example, continuing presentation of the last presentation object ("freezing" a video, etc.), pausing or cancellation of the synchronization point. 2) When a presentation of a dynamic basic object has reached a synchronization point and waits for other objects to reach this point, acceleration actions represent an alternative to waiting actions. They move the delayed dynamic basic objects to this synchronization point in due time. Possible actions include temporarily increasing the presentation speed or skipping all objects in the presentation up to the synchronization point. 3) When a presentation object does not arrive in time, it is possible to skip the object and to present the next one.

Priorities may be used for basic objects to reflect their sensitivity to delays in their presentation. For example, audio objects will usually be assigned higher priorities than video objects because a user recognizes jitter in an audio stream earlier than jitter in a video stream. Presentations with higher priorities are preferred over objects with lower priorities in both presentation and synchronization.

Summary

MODE is a complete synchronization system especially designed to support synchronization in a distributed environment. MODE provides a Synchronization Tool at the specification layer. The output of the tool is used as reference point-based interface format between the specification and the object layer. The Optimizer is part of the object layer and performs an off-line computation of the presentation schedule before the start of the presentation. The MODE Server Manager and the Synchronizer are also part of the object layer. The threads generated by them for the handling of dynamic media objects are part of the stream layer. Figure 15.51 shows the relation to the synchronization reference model.

15.6.5 Multimedia Tele-orchestra

At the University of Ottawa, the Multimedia Communication Research Laboratory (MCRLab) of Prof. Nicolas D. Georganas has developed a multimedia synchronization system known as *multimedia tele-orchestra*. This system is comprised of a sophisticated specification schema, the Time Flow Graph (TFG) [LKG94], and an implementation of this synchronization in a distributed environment [KG89]. In contrast to many other specification methods, the TFG takes into account that temporal knowledge may often be relative, i.e., it cannot be described by exact time parameters. The authors call this a fuzzy scenario. In addition, the duration of presentation parts may be imprecise and not known in advance. Hence, neither the exact occurring time points nor the duration are required to specify synchronization.

The notion of intervals serves as a basis for the TFG. In [LKG94] it is shown that all temporal relationships between intervals can be represented with TFGs. This leads to a partial sequential ordering which is used by the actual processing of

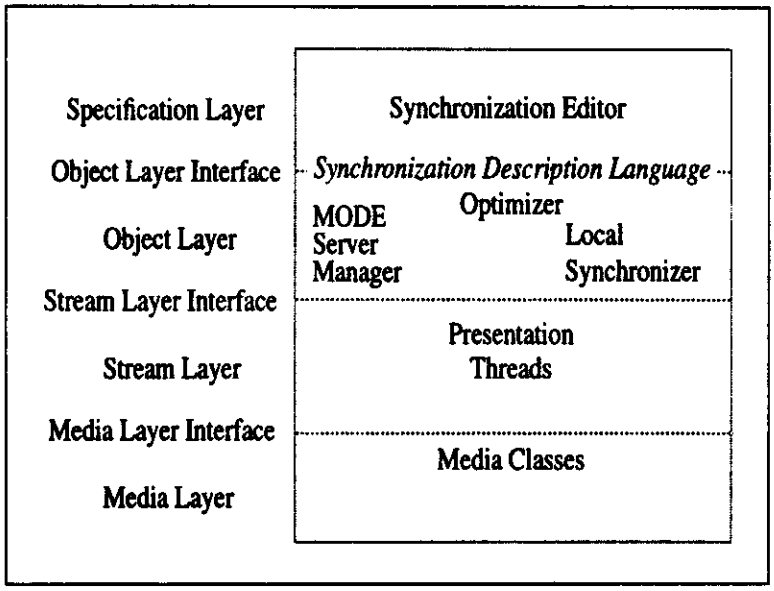


Figure 15.51: *Classification of the MODE system according to the synchronization reference model.*

synchronization at presentation time. With respect to our synchronization reference model, the TFG is an interval-based method located at the specification layer and also covers the interface between this layer and the object layer (see Figure 15.52).

Based on the TFG, a distributed multimedia synchronization schema was developed and became known as the *Synchronization Controller for Multimedia Communication (SCMC)* [KG89]. As a key feature, it takes into account that data may be originated by different sources located at different places. SCMC is targeted to run over ATM networks. However, the same algorithms can be used to operate on top of other multimedia-capable network configurations like Ethernet 10 Base-T, 100 Base-T and IsoEthernet.

In the tele-orchestration approach, a second component, the *Temporal Presentation Controller (TPC)*, is in charge of calculating a schedule with the earliest possible time to present objects at a remote computer. The result of the TPC, i.e., the respective schedule, is subsequently passed to the SCMC, which will actually control data processing to match the synchronization specification. In terms of the

synchronization reference model, the SCMC makes use of individual LDUs. It does not rely on a stream. The SCMC provides to its user the capability to provide synchronization between individual data streams. Hence, the SCMC is located at the media layer, as well as the stream layer. The TPC maps time constraints defined by the TFG onto SCMC primitives. The TPC calculates local schedules, whereas the SCMC unifies all local schedules to an actual implementation of the demanded synchronization. Hence, the TPC is located in the object layer according to Figure 15.52.

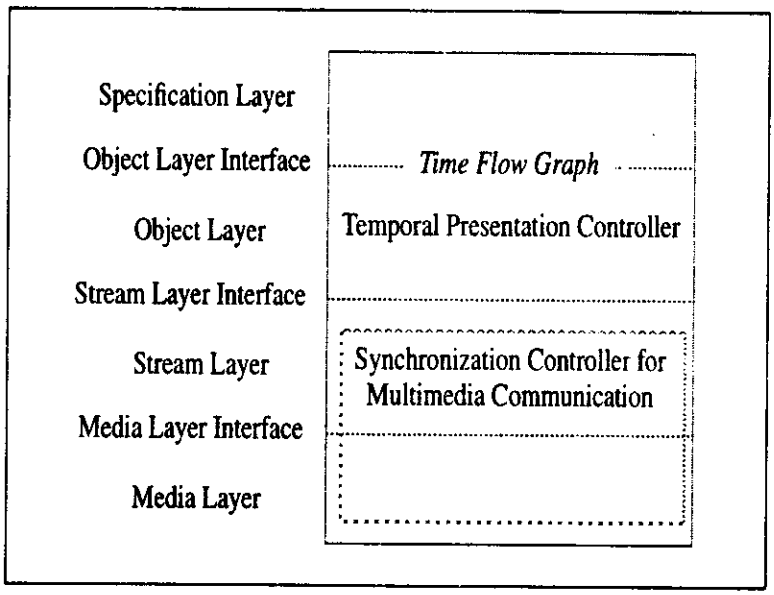


Figure 15.52: *Classification of the Tele-orchestra system according to the synchronization reference model.*

Summary

Tele-orchestra nicely covers the aspects of all layers of our synchronization reference model. Distribution is known and handled at the specification and stream layers. In [LLKG93], performance analysis results of this synchronization schema are presented.

15.6.6 Littles Framework

The main objective of this framework, currently integrated at the University of Boston [Lit93] to a multimedia information system, is to support the retrieval and delivery of multimedia data. This system is comprised of methods for synchronization specification, data representation, temporal access control and run-time intermedia synchronization. Especially, it provides mechanisms to overcome delays caused by storage, communication and computations on media objects. It also provides mechanisms for scalability and graceful degradation of multimedia services.

The specification of the synchronization is based on petri nets [LG90a] and global timer-based specifications that are mapped to the *Temporal-Interval-Base (TIB)* modeling approach. The temporal relations in this model include a start time for a data element, the duration of its presentation and the end time for it. Relative positioning is defined by delays between the start times of presentations.

Based on this specification, static and dynamic presentation scheduling is computed at the object layer. As an example of a simple planning algorithm in an environment with resource restriction, we present the static playout schedule computation algorithm [Lit92, LG92]. It assumes that the data elements are stored in a remote database. The data must be transported to the presentation workstation via a packet-switched network with restricted capacity. In a first step, the synchronization specification is used to compute the point of time for the start of the presentation (p_i) for each data unit. This is easily possible using the duration of the presentations (m_i). Using the start points of the presentations, it is necessary to compute the point of times to access the data units (q_i) from the database because they need a time (T_i) to be transported.

Let D_p be the constant propagation delay, D_t the delay proportional to the packet size (medium packet size / channel capacity) and D_v the variable load-dependent delay. Then, T_i is defined as $T_i = D_p + D_t + D_v$.

The following conditions must be fulfilled:

- $p_i \geq q_i + T_i$ (The data units must be available in time.)
- $q_{i-1} \leq q_i - T_{i-1} + D_p$ (Data should be accessed when previous sending of data

is finished.)

The following algorithm is used to compute q_i:

```
q[m] = p[m] - T[m] // Start with the last data unit.

for i = 0 to m-2
if q[m-i] < p[m-i-1] - Dp  // Collision
q[m-i-1] = q[m-1] - T[m-i-1] + Dp // Resolve collision
else
q[m-i-1] = p[m-i-1] - T[m-i-1] // No collision
end
end
```

Because static scheduling does not consider dynamic changes in the environment, as well as commands from the user that alter, for example, the presentation speed, dynamic scheduling is introduced. The dynamic scheduling approach is called *Limited A Priori (LAP) scheduling.* It performs the scheduling and reservation of resources only for a short period of time. The multimedia presentation is split into components of similar resource usage. For these components, the schedules are computed and statistical resource reservation is used. Subsequently, the session scheduler executes the presentation of the components. In the case of user-initiated presentation manipulation operations or of load changes, the schedules are recalculated.

To support interstream synchronization, skew control mechanisms are supported. They are based on dropping and duplicating data units, in the case that a queue representing the stream processing reaches low or high threshold values.

The petri net and time-line specifications in the specification layer are mapped to a TIB specification as an object layer interface format that is a type of interval-based synchronization. The off-line and on-line scheduling is located at the object layer. Additional skew control is provided at the stream layer.

Summary

The Littles framework represents a well-defined approach combining several layers. Its conception is concentrated on the retrieval of multimedia objects on one server and considers only a reduced set of distribution relevant parameters.

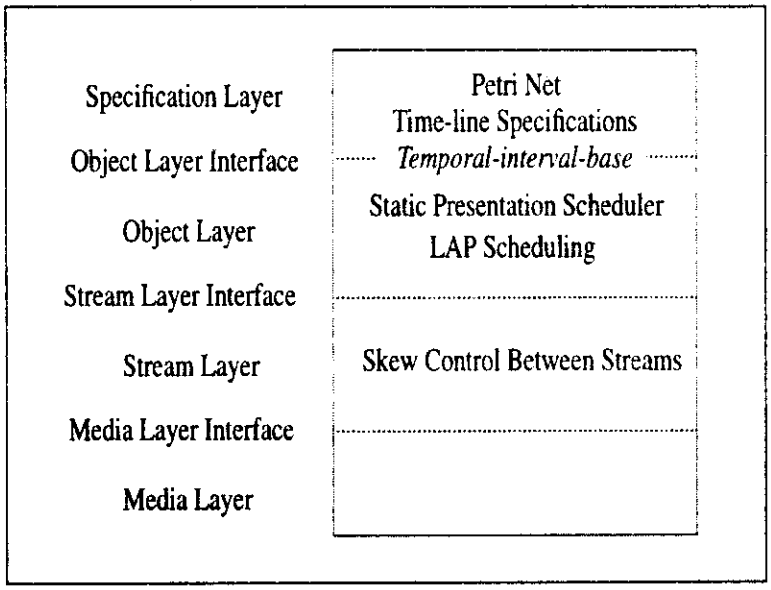


Figure 15.53: *Classification of Littles Framework according to the synchronization reference model.*

15.6.7 ACME

ACME (Abstractions for Continuous MEdia) [AH91b] is an I/O server for continuous data streams in the stream layer. The server controls a set of physical devices. Users can define logical devices as abstractions from physical devices. A stream path is build up by connecting input and output devices. The connection may be a real network connection. The stream consists of LDUs with an assigned time stamp.

A *Logical Time System (LTS)* synchronizes the I/O of logical devices. An LTS owns a clock that can be bound to the device which is most sensitive against delays, or it

may be driven by a specified connection. Each LDU will be processed by a logical device, if the time stamp matches the LTS clock.

A blocking caused by a connection may occur. In this case, the connection's input device is blocked and must buffer more and more units. The output device is starving, because it does not get enough LDUs. The blocking is resolved by skipping LDUs or by pausing the LTS in the case that a max skew value has been reached between time stamps of units and the LTS clock. The LTS is restarted if the time stamps of the logical data are close to the paused LTS clock and an additional amount of data for the start-up phase of the resynchronization was received.

ACME offers a programming interface and provides support for media streams at the stream layer only.

15.6.8 Further Synchronization-related Systems

Today, available multimedia extensions for operating systems, like Apple QuickTime [DM92], Microsoft Multimedia Extensions [Mic91a] and IBM Multimedia Presentation Manager/2 [IBM92b] contain synchronization mechanisms applied at the stream layer in the local domain. First, networked systems like the IBM Ultimedia Server cover some synchronization issues in a distributed environment.

The Orchestration Service [CGCH92] provides a stream-oriented interface for synchronized playout of continuous media in a distributed environment. Nicolai [Nic90], Little [LG91a], Escobar [EDP92], Shepherd [SS90], Ramanathan [RR93] and Anderson (as described in Section 6.7) have proposed techniques to control jitter among media streams in the stream layer. An evaluation and classification of these techniques is given in [EFI94].

Stefani, Harzad and Horn [SHH92] have proposed the use, at the object layer, of the synchronous language ESTEREL for the programming of multimedia synchronization. The language and its run-time environment provide support for fast event processing.

At the University of Geneva [TGD91], an object-oriented system with a global timer-based synchronization specification has been developed. At run-time, a global timer