INTERNATIONAL TELECOMMUNICATION UNION

TELECOMMUNICATION STANDARDIZATION SECTOR

STUDY PERIOD 1997 - 2000

COM 12-29-E
December 1997
Original: English

Questions: 11/12

Texte disponible seulement en
Text available only in
Texto disponible solamente en } E

STUDY GROUP 12 – CONTRIBUTION 29

SOURCE*: RAPPORTEUR QUESTION 11/12

TITLE: DRAFT NEW RECOMMENDATION ON MULTIMEDIA COMMUNICATION DELAY, SYNCHRONIZATION, AND FRAME RATE MEASUREMENT

_____

ABSTRACT

This draft Recommendation describes measurement for the time related aspects of Multimedia Communications Systems (with Audio, Video and Data channels). The methods cover capturing input and output media frame sequences with a common time scale, performing frame comparisons to determine active (non-repeated) output frames, and matching active output frames with unique input frames to determine transmission time and synchronization. The methods permit collection of delay, time skew, and frame inter-arrival time distributions which represent the desired parameters in their elemental forms.

This second draft of P.DEL adds sections on References, Definitions, Abbreviations, Data Channel Measurements, and Timer Stability and Synchronization. It has been modified to include additional video formats and to address comments.

* **Contact:** Mr. Arthur Webstser

Tel: +1 303 497 3567
Fax: +1 303 497 5323
Email: webster@its.bldrdoc.gov

P.DEL

DRAFT NEW RECOMMENDATION ON MULTIMEDIA COMMUNICATIONS DELAY, SYNCHRONIZATION, AND FRAME RATE MEASUREMENT

Introduction

An aspect of true Multimedia Communications Systems, that sets them apart from a mere collection of unrelated channels, is their ability to maintain a temporal relationship between the different media. This contribution specifies the parameters and measurement methods to assess relative synchronization between media channels, and two other key aspects of temporal quality. Transmission time, or delay through a channel, is critical when assessing a system's suitability for conversational and other interactive uses. Frame inter-arrival time and its reciprocal, frame rate, characterize a system's ability to deliver information continuously and consistently.

Today's Multimedia systems combine video, audio, and data channels to enhance communications. This contribution covers all these media. Video delay can vary widely over short sequences, audio and video sequences may be distorted during transmission, and data streams can have little or no structure and may contain bit errors. Although each media presents unique measurement challenges, the methods specified here meet and overcome them. The Mean Square Error based method expects and measures instantaneous video delay variations if present. The audio delay method accommodates channels where the original speech waveform is not preserved. There are also methods for data channels that take advantage of native structures and tolerate bit errors. All the methods allow test signals that are representative of the intended system applications.

1. Scope and Application

1.1 Scope

This Recommendation covers test methodologies for multimedia transmission systems utilizing digital transport facilities. It gives a set of measurement parameters to characterize the following aspects of system performance:

1) Active video frame inter-arrival time, which is the reciprocal of the elementary frame rate,
2) Visual channel transmission time, also called video delay,
3) Audio channel transmission time (or audio delay),
4) Data channel transmission time or delay (and frame inter-arrival time),
5) Temporal synchronization between channels.

The measurement scope is limited to cases where appropriate media input and output interfaces are present, or where these interfaces can be made available with optional test fixtures.

The following applications are beyond the scope of this Recommendation:

1) Measuring aspects of system performance other than delay, synchronization, and frame rate. Temporal measurements do not completely characterize the quality of a multimedia transmission system. For example, the reproduction quality of video frames from input to output is also of obvious importance to users. The optimization of such subjective performance for all quality parameters may take precedence over the optimization of the results of parametric measurements performed according to this Recommendation.
2) An unrestricted choice of useful and representative source content. The methods of measurement specified here require restrictions on their source signals for testing. Video source sequences with high motion activity often cause increased delay, decreased frame rate, and skewed audiovisual synchronization in some multimedia applications. Therefore, measurements should use test scenes which are realistic for the application of the multimedia system under evaluation. Other limitations are given in the sections for each measurement method.
3) Measuring the performance aspects of systems where the input and output interfaces are not accessible.
4) Limits for the parameters are beyond the scope at this time. This Recommendation only provides methods to measure these parameters without providing values for evaluation.

1.2 Application

1.2.1 User-to-User Channels

Ideally, the delay measurement would be conducted at the user interfaces, so as to characterize the entire user-to-user delay. The complete user-to-user channel begins and ends with user interface devices. For example, consider the visual channel with its camera and display components, as shown in Figure 1.

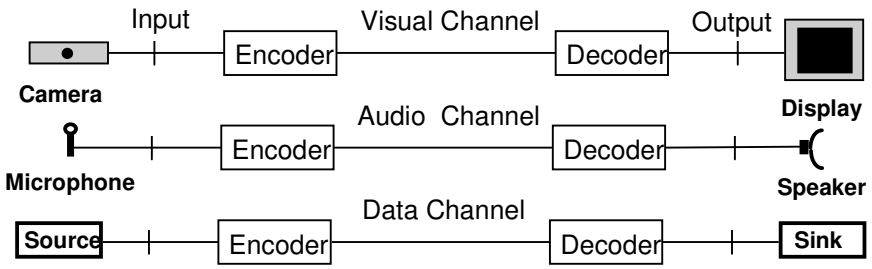


Figure 1 User-to-User Channels in a Multimedia System

Unfortunately, signals enter and leave this channel in the form of light, making the generation and collection of suitable signals for measurement a difficult task. To simplify the interconnection of measurement equipment with the channel, we specify the test channel between electrical interface connectors at the camera output and at the display input. This has the advantage of providing more physical and logical structure to the test interface. The additional delay contributed by a camera and display could be assessed separately (these delays are expected to be constrained within the sample/display interval, may be constant for displays, and are usually test-signal independent) and added to the measurements of variable delay made in accordance with this Recommendation.

We can identify similar input and output interfaces in audio user-to-user channels and data user-to-user channels.

1.2.2 Applicable Configurations

The following channel configurations are appropriate applications of this Recommendation. Each figure shows the necessary input and output interfaces.

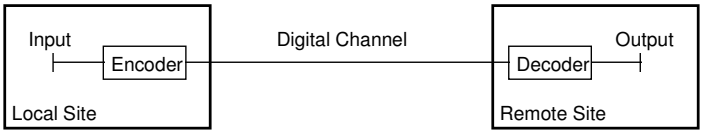


Figure 2 End-to-End Measurement[1]

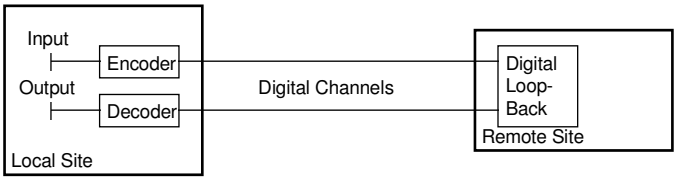


Figure 3 Remote Digital Loop-Back Measurement

---

[1] Note that if the digital channel contains processing components, caution must be used when interpreting the results to reflect interaction effects.

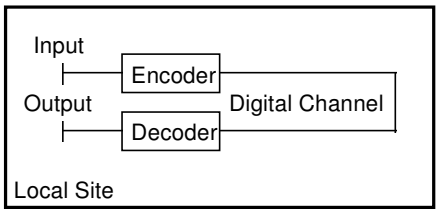


Figure 4 Local System Measurement

Figures 2, 3 and 4 show only an Encoder, Decoder, and Digital Channel for simplicity. The components that may comprise the media channel in these tests are not strictly limited.

Figure 5 shows a video channel measurement configuration with limited application. This would constitute a two-way delay measurement of two one-way systems and permit a single measurement device.

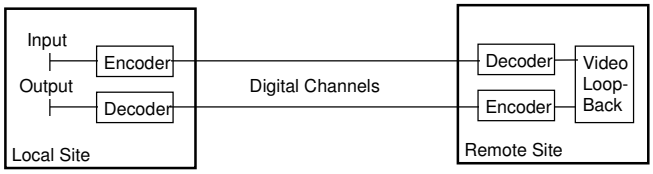


Figure 5 Remote Video Loop-Back Measurement

The video loop-back is not appropriate for cases where the digital channels have asymmetrical delay. For video teleconference systems encoding less than 30 frames per second, the coding of the forward path may influence the transmission delay of the return path, which would make loop-back testing inappropriate in this case. Under these circumstances, the two-way measurement would not reveal the desired one-way assessment.

Note: one-way assessment for symmetrical systems is simply one-half of the two-way delay.

1.2.3 Applicable Interfaces - Video

The work leading to the development of this Recommendation was primarily conducted using composite analog video signal interfaces. However, the design of video conference systems is rapidly changing from a collection of components (using the composite interface) to more integrated systems. Furthermore, the demands of high quality video production exceed the capabilities of the composite analog signal. It will be necessary to apply this Recommendation at new interfaces to keep pace with advancing technologies. Digital component interfaces, computer monitor RGB interfaces, and digital camera interfaces are likely candidates.

The measurement parameters defined here simply require the ability to supply video frames to the input and collect and compare video frames at the output of a visual channel. No technique demands a composite interface, per se.

To facilitate the measurements described in this Recommendation on fully integrated systems, optional interface access features will be needed to support testing. The complexity of individual systems may not be appreciably increased if nearly all of the additional functions required to implement these interfaces are contained in the optional sub-system. These interface features might be useful in other activities, such as fault isolation and manufacturing quality assurance, and should be desirable to manufacturers on this basis.

Applicable Interfaces - Audio

This specification is applicable at all standardized audio interfaces.

1.2.4 Applicable Interfaces - Data

This specification is applicable at all standardized data interfaces.

2. References

The following ITU-T Recommendations, and other references contain provisions which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; all users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published.

ITU-R Recommendation BT.601-5 Encoding Parameters of Digital Television for Studios.

CCIR Report 624-4 - Characteristics of television systems, 1990, Annex to Vol. XI-1

ITU-T Recommendation P.910, Subjective Video Quality Assessment Methods for Multimedia Applications, 1996.

ITU-T Recommendation P.920, Interactive Test Methods for Audiovisual Communications, 1996.

ITU-T Recommendation P.930, Principles of a Reference Impairment System for Video, 1996.

ITU-R Recommendation BT.500-6, Method for the Subjective Assessment of the Quality of Television Pictures, 1995.

ITU-T Recommendation P.861, Objective Quality Measurement of Telephone-band Speech Codecs, 1996.

ITU-T Recommendation P.80, Telephone Transmission Quality Subjective Opinion Tests, Methods for Subjective Determination of Transmission Quality, 1993.

ITU-T Recommendation P.84, Telephone Transmission Quality Subjective Opinion Tests, Subjective Listening Test Method for Evaluating Digital Circuit Multiplication and Packetized Voice Systems, 1993.

3. Terms and Definitions

This section contains the definitions for terms used throughout this Recommendation.

3.1 General Terms

3.1.1 Multimedia Communication System

A system that handles more than one media stream in a synchronized way from the user’s point of view. The system may allow interconnection of multiple parties, multiple connections, and the addition or deletion of resources and users within a single communication session.

3.1.2 Media Stream

A sequence of presentation units intended to convey some specific content.

Coding Hierarchy Levels

The nested units of signal representation into which a media stream can be decomposed.

3.1.3 Content Hierarchy Levels

The nested units of information into which a media stream can be decomposed.

3.1.4 Presentation Unit

The smallest convenient division of a media stream (defined by the measurement system) that conveys an independent, self-contained unit of content, from among the content hierarchy levels present in the stream.

3.1.5 Video Frame

A Presentation Unit of the visual channel. The lowest level of the content hierarchy in a video media stream, where differences between sequential units appear throughout the unit of presentation. The content hierarchy of this standard may re-use the terms in some coding hierarchies, if necessary. For 525-line and 625-line formats, a Video Frame is defined as one Field, where a Field is specified in CCIR Report 624-4.

3.1.6 Audio Frame

A Presentation Unit of the audio channel. A group of consecutive audio samples. The preferred number of samples in an Audio Frame depends on the audio sample rate, and is given in Section 5. These Audio Frames have no relationship to the frames designated by certain audio/speech codecs.

3.1.7 Data Frame

A Presentation Unit of the data channel. A group of consecutive data bits. The preferred number of bits in a Data Frame depends on the application for the data channel.

3.1.8 Digital Channel

A means for conveying information from one point to another in digital form. A digital channel may be implemented on a network composed of digital communications components.

3.1.9 Visual Channel

A means for delivering video frames from one point to another. A sequence of frames submitted to the channel input results in a similar (not necessarily identical) sequence of frames at the channel output. The visual channel may be comprised of the following components: video format conversion devices, encoders (compressors) and decoders (decompressors), rate smoothing buffers, multiplexors and demultiplexors, modulators and demodulators, transmission facilities, switches, multi-point conference units, and other components necessary to achieve the desired channel characteristics.

3.1.10 Audio Channel

A means for delivering audio signals from one point to another. An audio waveform submitted to the channel input results in a similar (not necessarily identical) waveform at the channel output. The audio channel may be comprised of the following components: encoders (compressors) and decoders (decompressors), buffers, multiplexors and demultiplexors, modulators and demodulators,

transmission facilities, switches, multi-point conference units, and other components necessary to achieve the desired channel characteristics.

3.1.11 Data Channel

A means for delivering data from one point to another. A sequence of data bits submitted to the channel input results in a similar (not necessarily identical) sequence of bits at the channel output. The data channel may be comprised of the following components: format conversion devices, encoders (compressors) and decoders (decompressors), buffers, stream segmentation and re-assembly devices, multiplexors and demultiplexors, modulators and demodulators, transmission facilities, switches, multi-point conference units, and other components necessary to achieve the desired channel characteristics.

3.1.12 Field Integrity

An attribute of a Visual Channel present when the content of odd (even) Fields in the 525-line or 625-line format source sequence is conveyed in the odd (even) fields at the output.

3.1.13 Repeated Video Frame

An output video frame that is indistinguishable from its preceding frame(s) in the sequence (when the corresponding input sequence frames possess distinguishable differences). A Repeated Frame is assumed to be generated at some intermediate point in the visual channel. Since Repeated Frames have not traversed the channel from input to output, they are not used in the compilation of the visual channel delay distribution. Repeated Frames also convey no new visual stimulus, and they are excluded from calculation of frame inter-arrival time (and subsequently elementary frame rate).

3.1.14 Non-Repeated Video Frame (Active Frame)

An output video frame that is distinguishable from its preceding frame(s) in the sequence (when the corresponding input sequence frames possess distinguishable differences). An Active Frame is assumed to have traversed the channel from input to output, and its delay may be included in the visual channel delay distribution. Since Active Frames convey new visual stimulus, they are the basis for calculation of frame inter-arrival time (and subsequently elementary frame rate). (Note: Any interpolated frames generated in a decoder will be interpreted as Active frames in this process.)

3.1.15 Frame Matching

The process of comparing one sequence of frames with another sequence of frames in order to determine the correspondence between frames in each sequence and the correspondence of individual frames. See Note 1.

Note 1: One means to test the correspondence between two video frames is to compare their digital representations on a pixel by pixel basis, and summarizing over all pixels as the mean-square of the differences (usually called Mean Square Error).

3.1.16 Repeated Video Frame Identification

The process of comparing each output video frame with its preceding frame(s) in sequence and quantifying the extent of correspondence between each pair. When the correspondence between a pair of frames is high (the only differences are attributable to the noise in the measurement), the pair is indistinguishable; and when the corresponding input sequence of frames possess distinguishable differences, then the current frame is categorized as a Repeated Frame. See Section 3.1.15, Note 1.

3.1.17 Active Video Frame Identification

The process of comparing each output video frame with its preceding frame(s) in sequence and quantifying the extent of correspondence between each pair. When there is limited correspondence between a pair of frames (such that the differences measured are distinguishable from the measurement noise), and the corresponding input sequence of frames possess distinguishable differences, then the current frame is categorized as a Active Frame. See Section 3.1.15, Note 1.

3.2 Framework of Measurable Parameters

This subclause gives the high-level definitions of the key measurement parameters. There are also method-specific definitions in the sections that follow.

3.2.1 Transmission Delay

The time a particular frame takes to traverse the transmission channel. This time is calculated by first recording the times that frames are placed onto the channel, then finding an output frame that has traversed the channel and noting its arrival time at the output. Next the output frame shall be uniquely matched with an input frame. The transmission delay is then equal to the arrival time minus the input time.

Note 2: When there is little or no activity in the channel, the methods described here will encounter difficulty in making valid measurements. However, this parameter becomes unimportant following the display of the first frame when all succeeding frames are identical (no activity or still video).

3.2.2 Media Stream Delay Distribution

The set of delays calculated for a sequence of output frames, expressed such that any variation between individual measurements is clearly illustrated. Classical summary statistics may also be supplied, as applicable.

3.2.3 Active Frame Inter-Arrival Time

The time between successive Active Frames at the output of the channel. This time is calculated by selecting an Active Frame (or Non-Repeated Frame that has traversed the channel) and noting its arrival time at the output. Then the most recent (previous) Active Frame shall be found and its arrival time is noted. The channel Active Frame Inter-arrival Time is then equal to the present frame's arrival time minus the previous arrival time. See Section 3.2.1, Note 2.

3.2.4 Active Frame Inter-Arrival Time Distribution

The set of inter-arrival times calculated for a sequence of active output frames, expressed such that any variation between individual measurements is clearly illustrated. Classical summary statistics may also be supplied, as applicable.

3.2.5 Elementary Frame Rate

The reciprocal of the Active Frame Inter-arrival Time for the present Active Frame. The Elementary Frame Rate is equal to 1 divided by the difference between the arrival times of the present and previous Active Frames.

3.2.6 Frame Rate Statistics

A set of statistics that are calculated for a sequence of active output frames, expressed such that any variability is clearly illustrated. When reporting summary statistics for frame rate, they shall be computed using the inter-arrival time distribution, and taking the reciprocal.

3.2.7 Elementary Frame Skipping Ratio

The ratio of input to output inter-arrival times (or elementary frame rates) for a pair of matching frames. A measure of the change between the input and output Active Frame rates.

4. Abbreviations

ANSI	American National Standards Institute
ATM	Asynchronous Transfer Mode
DC	Direct Current
DFT	Discrete Fourier Transform
EAV	End of Active Video
FFT	Fast Fourier Transform
fps	frames per second
GPS	Global Positioning System
IIR	Infinite Impulse Response
MSE	Mean Square Error
MTIE	Maximum Time Interval Error
PSD	Power Spectral Density
PSNR	Peak Signal-to-Noise Ratio
RGB	Red, Green, Blue
RMS	Root Mean Square
SAV	Start of Active Video
SMPTE	Society of Motion Picture and Television Engineers
TIE	Time Interval Error

5. Temporal Calculations for General Communication Media

This section gives a general model for calculation of the parameters in this specification and applies the model to each medium covered here. Figure 6 illustrates the general model. Two sequences of **presentation units**, P and P′, enter the channel input and leave the channel output interfaces. As the last part of each presentation unit passes the interface, the measurement system reads a timer, T, and associates the value T(n) with input presentation unit n, P(n) . At the output interface, the measurement system reads a timer, T′, and associates the value T′(m) with output presentation unit m, P′(m).

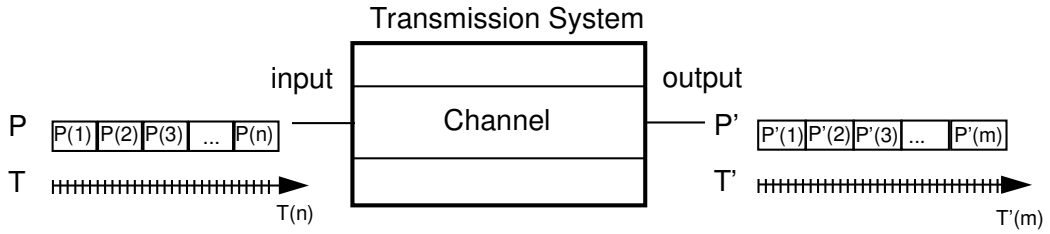


Figure 6 A General Model for Presentation Units

Note that timers T and T′ can be the same timer in some measurement configurations (e.g., local and loop back configurations as shown in Figure 3, Figure 4, and Figure 5). Otherwise (e.g., end-to-end configurations as in Figure 2), the timers shall be synchronized. Also, a single timer supplies near-simultaneous time stamps for all channels at a given interface, permitting direct calculations between and within channels.

The beginning and end of presentation units may not be coincidental, as Figure 6 shows. Time stamps shall be associated with the end of a presentation unit, to address both practical and perceptual matters. Beginning time stamps may be stored separately for additional measurements.

Note that P and P′ are illustrated as continuous streams above, but non-periodic arrivals are possible with many media. In fact, complete presentation units may take longer to exit the system than enter. This is one reason for time stamping the end of presentation units. Another is a practical consideration, the end of a presentation unit is often easier to anticipate than the beginning.

5.1 Single Channel Calculations

The Channel Delay for presentation unit P′(m), after determining that P′(m) is matched with P(n), is

$$t_P(m) = T'(m) - T(n)$$

Matching is usually trivial in a loss-less channel, but the methods defined here will deal with both distortion and complete loss of presentation units. There shall be differences between the successive presentation units at the input, or matching results will have ambiguity.

The Inter-Arrival Time for presentation unit P′(m) at the channel output is

$$b'_P(m) = T'(m) - T'(m-1)$$

where m-1 is the index of the previous presentation unit.

The Elementary Frame Rate for P′(m)at the channel output is

$$f'_P(m) = \frac{1}{T'(m) - T'(m-1)}$$

When systems routinely discard presentation units (as is the case with video on low bit-rate digital channels), the input and output frame rates will differ. The elementary Frame Skipping Ratio is

$$\frac{b'_P(m)}{b_P(n)} = \frac{f_P(n)}{f'_P(m)}, \text{ where } b_P(n) = T(n) - T(n-1), \text{ and } P'(m) \text{ matches } P(n)$$

Complete sets of individual measurements for delay and inter-arrival time, represented as

$$t_P = \{t_P(1), t_P(2), t_P(3), t_P(4), \ldots t_P(M)\}$$

and

$$b'_P = \{b'_P(2), b'_P(3), b'_P(4), \ldots b'_P(M)\}$$

may be collected for statistical and graphical analysis.

5.2 Media Frames

In the Visual Channel, the presentation units are Video Frames. Section 6 defines these units for each video interface as a single Field of the 525-line format, for example. V and V′ are the variables for the units of the input and output video streams.

In the Audio Channel, the presentation units are Audio Frames. Audio Frames are a group of digitized samples representing the audio stream (see Table 1). Section 0 defines these frames for audio interfaces. A and A′ are the variables for the units of the input and output audio streams. The recommended audio frame length is approximately the same duration as the associated **Video Frame,** if present (e.g. 16.66... ms for 525-line format).

Table 1 Preferred Audio Frame Length

samp/sec	preferred sample size, duration
8000	128, 16ms
16000	256, 16ms
32000	512, 16ms
44.1k	512, 11.61ms
48k	512, 10.66ms

In a Data Channel, the presentation units are Data Frames. Data Units are a unique test word, or a group of bits representing some presentation unit for the user application of the data stream. Section 8 defines these units for data interfaces. D and D′ are the variables for the units of the input and output data streams.

5.3 Synchronization Calculations

As an example of temporal calculation of synchronization parameters, consider the audio and video media streams where:

- A(m) and V(n) are associated, either by definition of the measurement device or the association is known a priori.
- V′(q) and V(n) are matched, through the methods described in this specification.
- A′(p) and A(m) are matched, through the methods described in this specification.

The time offset between associated audio and video frames at the input is

$$o_{AV}(m,n) = T_A(m) - T_V(n)$$

This parameter indicates the position of the audio frame with respect to the video frame in time.

The time offset between associated audio and video frames at the output is

$$o'_{AV}(p,q) = T'_A(p) - T'_V(q)$$

Note that $o'_{AV}(p,q)$ is the synchronization relationship perceived by a user at the channel output.

The time skew between associated audio and video frames at the output, introduced by the transmission system channel, is

$$s_{AV}(p,q) = o'_{AV}(p,q) - o_{AV}(m,n)$$

Following the convention of positive time delay, synchronization time lag is a positive value. If the second channel is leading , the parameter has a negative value.

For systems operating at low frame rates, time skew other than zero may provide perceptually more-accurate lip-sync. For systems with variable video delay, it is usually undesirable to vary the audio delay to maintain zero skew since the audio may become unintelligible through such manipulation.

If a transmission system is capable of multi-channel audio, these calculations are appropriate for assessing synchronization between audio channels.

6. Video Measurements

6.1 Collecting Video Frames for Measurements

This section describes the hierarchy of elements that can exist in a video sequence, and specifies the level of each video coding hierarchy that is the fundamental unit for comparisons. This Recommendation defines these fundamental units as **Video Frames**, re-defining the terms of the coding hierarchy as necessary. We first cover this topic at a conceptual level, to foster extension of the methods beyond the specific electrical interfaces dealt with later.

6.1.1 Description of Video Frames

Many granularity levels of information may be present in a video sequence. The information in video sequences can be divided as shown in Figure 7:

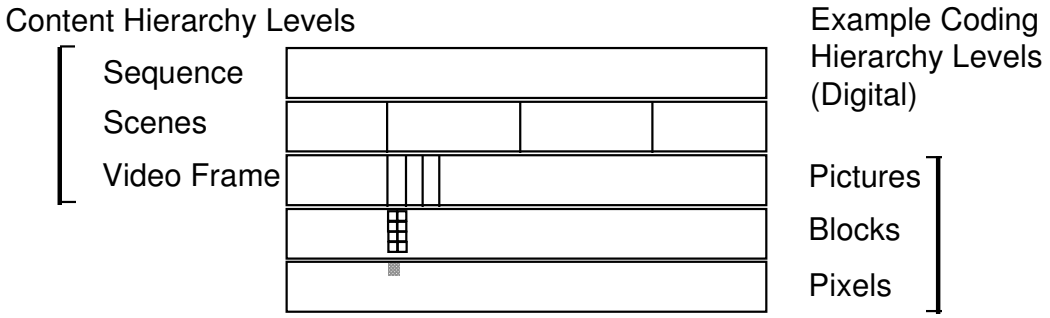


Figure 7 Example Video Sequence Hierarchy

Figure 7 shows a complete video sequence composed of four scenes with varied content. Each scene is composed of many pictures or frames that differ to convey motion or change. Pictures are the lowest autonomous level of this example Content Hierarchy in the temporal domain, in that the

smallest content changes take place at this level over time. This level is the fundamental presentation unit to the user. We define **Video Frames** at this level, where differences between sequential units appear throughout the unit of presentation.

Video Frames may also be at the highest level of the video coding hierarchy. Figure 7 shows an example digital coding hierarchy where this is true. However, if dependent coding between sets of pictures is allowed, then the content and coding hierarchies could overlap by more than one level. Video Frames would continue to be defined at the lowest content level.

6.1.2 Video Frames at Composite Interfaces

Figure 8 shows the video sequence hierarchy for 525-line and 625-line format signals.

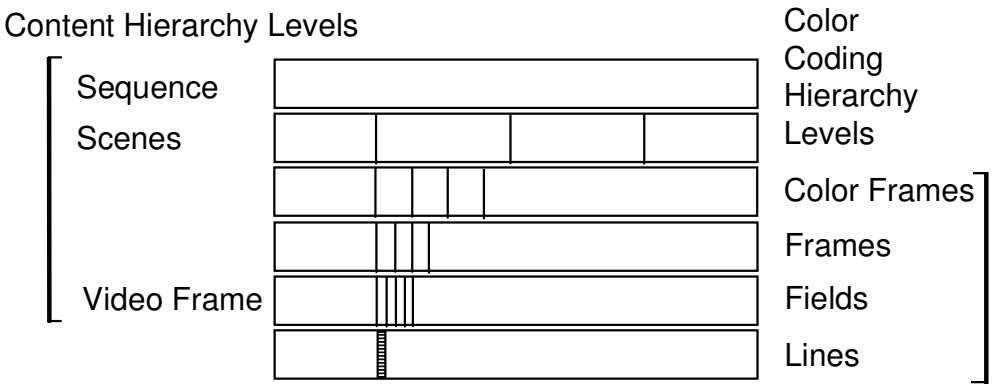


Figure 8 525-line and 625-line Video Sequence Hierarchy

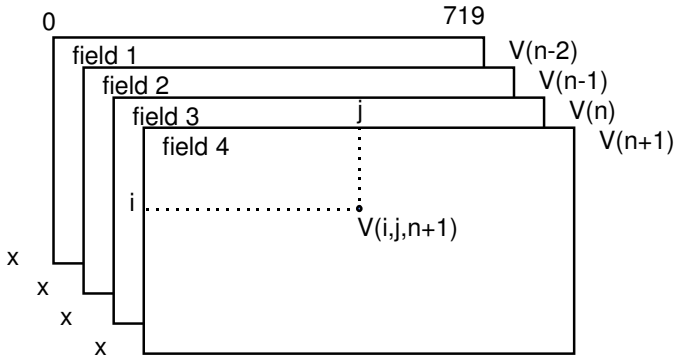
For 525-line and 625-line interlaced formats, a **Video Frame** is defined as one Field, where a Field is specified in CCIR Report 624 or in [10]. Successive fields may convey new information to the user throughout the unit of presentation, although not on every line of the presentation unit. Also, since the field rate is higher than the frame rate, this definition permits finer sampling of the video sequence.

This definition is also necessary to accommodate transmission systems that (due to system restrictions such as transmission bit rate) convey active video frames at rates less than the video interface frame rate and display each new active frame as soon as possible. This display update method can result in substantial differences between successive fields.

6.1.2.1 Analog to Digital Conversion

The methods of measurement described in subsequent sections require digitization of the analog composite signal. ITU-R Recommendation BT.601-5 provides a high-fidelity method to sample the analog active-line area of the 525-line or 625-line luminance signals. Experience has shown that the luminance signal supplies sufficient information to compare and match video frames for the measurements in this specification.

Figure 9 shows how the digitized samples corresponding to video frames may be organized.



Format	x =
525-line	241
625-line	286

Figure 9 Recommended Coordinates for Video Frame Digitization

V(n) is Video Frame n at time T(n), V(n-1) is the previous Video Frame, and V(n+1) is the next Video Frame. V(i,j,n) is the (i,j) luminance pixel in Video Frame n at time T(n).

Other coordinate systems are possible, but the numbering from upper left to lower right should be used to facilitate comparisons between different measurement implementations. Implementation of the Digital Component Interface in Section 6.1.3 will also facilitate these comparisons. As an example of other coordinate systems, specific line numbers can be assigned begining with vertical synchronization lines [10].

6.1.2.2 Time Stamp Assignment

With 525-line systems, the time, T(n), associated with Video Frame n shall be read immediately following the digitization of the last pixel in the frame, coordinate (242, 719), and before the next line begins (10.222 μs).

With 625-line systems, the time, T(n), associated with Video Frame n shall be read immediately following the digitization of the last pixel in the frame, coordinate (286, 719), and before the next line begins (10.666 μs).

6.1.2.3 Gain, Active Area, and Spatial Alignment

This specification requires a correction factor for gain (g), level offset (l), horizontal shift (h) and vertical shift (v) prior to measurement to ensure accuracy.

Bibliography [3] defines a method to measure Average Gain and Level Offset (in its clause 5.1); as well as manual and automated methods to measure Active Video Area (in its clause 5.3) and Active Video Shift (in its clause 5.4). This specification recommends these methods to ensure the quality of frame to frame comparison and frame matching methods. When measured and used, Active Video Area coordinates and Active Video Shift coordinates shall be specified.

If the transmission system re-sizes the input frames (i.e., expands or contracts the scale in the horizontal, vertical, or both directions), such that additional differences are present when comparing the input and output sequences, then it may be desirable to use a correction factor for the size

change (z) to minimize the differences prior to frame matching. When re-sizing is deliberate and extensive, it is essential to compensate for it. It is not known whether significant re-sizing is prevalent in multimedia communication systems, nor at what level the re-sizing contributes significant noise, but there is an opportunity to compensate for re-sizing and continue with the processes as specified here.

Measurement errors may result if these methods are not used. For example, if the channel has a video shift of 4 pixels, a camera pan in the same direction would result in incorrect matching frames with inaccurate delays. If left uncompensated, shifts and gain or level offsets would also increase the noise level and reduce the effectiveness of frame matching based on Mean Square Error.

There are circumstances where simplified methods yield accurate measurements, as shown in the contributions leading to the development of this specification (see Appendix A). The following section indicates where measurement simplifications are possible.

6.1.2.4 Recognized Options

A recognized option for this specification is to define a sub-frame area in terms of the x-y coordinates and make all comparisons and matching operations on this sub-frame. Alternatively, the sub-frame could be defined by the safe action area and/or safe title area in [7]. This allows measurements to exclude frame boundaries without determining the Active Video Area exactly, and improve the capture noise threshold by avoiding error at frame boundaries. If either of these options are adopted, the sub-frame coordinates shall be specified.

When defining a sub-frame area, the following items should be considered:

1. Avoid borders that do not contain picture.
2. Include still areas with both horizontal and vertical edges and a full range of pixel values when possible, to aid with determining spatial correction factors.
3. Include areas with motion, or delay measurements will produce ambiguous results.
4. The size of the sub-frame area in pixels is proportional to computation time. In addition, small areas may improve the frame matching operation when motion is localized, but shall be large enough to avoid ambiguity.

Another option is to use a lower horizontal sampling during digitization for storage. Experience has shown that horizontal resolution of 320 pixels in the digital active-line period can support sufficient quality computation for these measurements, and spatial alignment corrections may not be necessary. If this option is adopted, the horizontal pixels in the digital active-line period shall be specified.

When measuring systems where the picture rate is known to be less than 30 per second, it is permissible to collect every other Field to reduce capture storage requirements, noting the field selected (even or odd). One example system is ITU-T Rec. H.320 compatible terminals with video coding according to ITU-T Rec. H.261. If this option is adopted, the field selected shall be specified.

6.1.3 Video Frames at Digital Component Interfaces

When working with component video, we again define a Video Frame as one field of the 525-line or 625-line format.

6.1.3.1 Signal Organization

Video signals at the digital interfaces ([8] for parallel and [9] for serial) have already been converted from their analog form, and no additional digitization is required here. These interfaces multiplex 8 or 10 bit samples of the video components in the order (C_B, Y, C_R, Y, C_B, Y,…). Measurement systems may use only the Y samples.

Timing sequence information words are inserted in the bit stream to identify the digital active line. The active lines begin following the Start of Active Video (SAV) word and end prior to the End of Active Video (EAV) word, and contain 720 luminance samples. The recommended sample coordinates are the same as for the Composite interfaces.

6.1.3.2 Time Stamp Assignment

The same requirements (see subclause 6.1.2.2) for the Composite Interfaces apply here as well.

6.1.3.3 Gain, Active Area, and Spatial Alignment

The same requirements (subclause 6.1.2.3) and options (see subclause 6.1.2.4) for the Composite Interfaces apply here as well.

6.2 Mean Square Error Methods of Measurement for Video

This section gives the methods of measurement for a system employing a Mean Square Error approach. Implementations of this method shall be able to supply appropriate video frame sequences at the channel input. The method also requires capture and, if necessary, digitization of the luminance component of video frame sequences at the channel interfaces. Supply and capture shall be conducted in accordance with the provisions of Section 6.1 for the interfaces in use. Once the Active Frames and their matching frames in the input sequence are found, the desired temporal calculations shall be conducted as specified in Section 5.

The methods accommodate several special circumstances, including high quality transmission systems that maintain interlaced field integrity, and the use of source video sequences derived from the 3:2 pull-down process from 24 fps film.

6.2.1 General

Detecting Active Frames within a sequence of video frames, and finding Matching Frames between sequences requires a standard method of comparison. This method compares video frames on a pixel-by-pixel basis, and summarizes the difference between a pair of frames as the mean square error over all pixels of interest. Thus, for a pair of frames (one from the input sequence and one from the output sequence) the Mean Square Error (MSE) is

$$M[V'(m), V(n)] = \frac{1}{K_s} \sum_{j=Jmin}^{Jmax} \sum_{i=Imin}^{Imax} [V'(i,j,m) - V(i,j,n)]^2$$

where V′(i,j,m) is the value of pixel i,j in the output frame at time T′(m), and V(i,j,n) is the value of pixel i,j in the input frame at time T(n). K_s is the total number of pixels in the rectangular sub-frame of interest, given by

$$K_s=(I_{max}-I_{min}+1)\times (J_{max}-J_{min}+1).$$

Note that V′(i,j,m) has been corrected for any gain, level offset, horizontal shift, vertical shift, and spatial scaling (if necessary) between input and output (with corresponding correction factors g, l, h, v, and z):

$$V'(i,j,m) = \frac{V^*(x+v, y+h, m) - l}{g}$$

where V*(x,y,m) is the output pixel before application of the correction factors. If the output video must be re-sized to match the input, then

$$V'(i,j,m) = \frac{V^{**}(\hat{i}+v, \hat{j}+h, m) - l}{g}, \text{ where } V^{**}(m) = f(V^*(m), z)$$

and where $f(V^*(m), z)$ represents a re-sizing function.

For comparisons between adjacent frames within a sequence (e.g., to detect active frames at the output interface), V(i,j,n) becomes V′(i,j,m-1) in the equation for MSE, above.

MSE is an important factor in the calculation of Peak Signal-to-Noise Ratio (PSNR), similar to [3]:

$$PSNR = 20\log_{10}\left[\frac{V_{peak}}{\sqrt{M[V'(m), V(n)]}}\right] dB$$

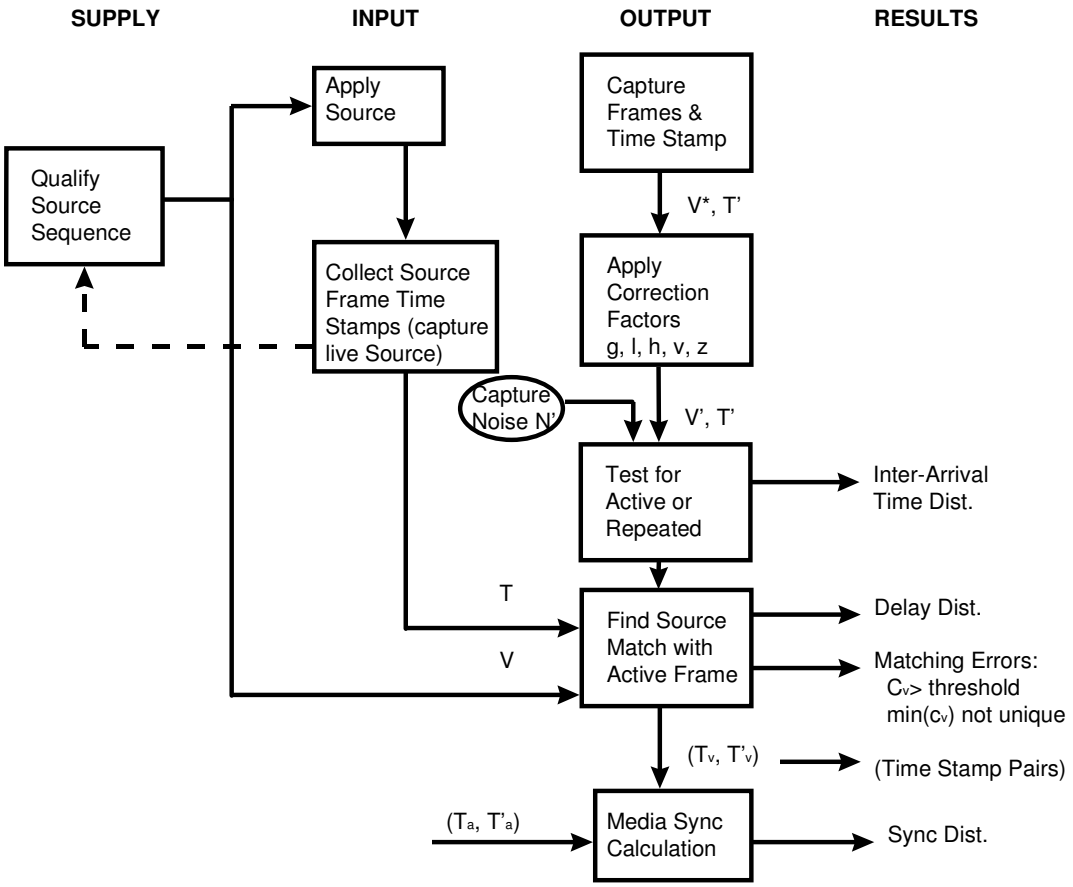


Figure 10 Flow Diagram for MSE-based Video Measurements

Figure 10 illustrates the high-level measurement process for the MSE method.

6.2.2 Calibrating the Minimum Distinguishable Difference Between Frames

This section specifies the method to determine the noise (or unwanted variation) in the digitization and storage processes that collect video frame sequences for comparison. This noise level is dependent on the specific options chosen (e.g., digitization format) and shall be known in order to make valid measurements.

The test conditions to calibrate the capture noise are as follows:

1. Apply a **still video** scene to the channel input. [2] defines still video as "video imagery that conveys no motion or change." It is important to maintain the same input video signal to noise ratio during calibration and measurement. One technique uses a source sequence composed of a single repeated frame from one or more of the motion video sequences intended for further testing. This does not reproduce the noise in the source sequence, and is only appropriate for cases where the capture noise is significantly higher than the source noise. For some test sequences and live video, it may be possible to divide the video frame spatially and define a still sub-frame (e.g., background) for this calibration and a motion sub-frame for other measurements. Still test signals (SMPTE Color Bars) have been used successfully as well (again, the source noise present in the still test signal shall be the same as in the testing sequence).
2. Capture (digitize and store) the corresponding sequence of frames at the channel output. 30 to 60 frames should be sufficient. When the channel employs digital compression, it shall be allowed to achieve a steady quality level on the still, thereby avoiding any **scene-cut response** that would corrupt the noise measurement.

In general, the capture noise at input will be different from the noise at the output. Some codecs filter out source noise to improve the signal for encoding.

For a 30 frame sequence, calculate the set of 30-1=29 adjacent frame MSE values,

M[V′(m), V′(m-1)].

The output capture noise level is the maximum MSE value of the set.

$$v'(m) = M[V'(m), V'(m-1)] \quad \text{for} \quad m = 2,3,\ldots 30$$

where $v'(m)$ is the MSE value for frame V′(m), and the capture noise, N′, is

$$N' = \max(v')$$

where v' is the set of MSE values for sequence V′. The variation within the set of MSE values should be small (<20%) owing to the averaging over many pixels for each value in the set. For the input sequence, we have N = max(v).

To allow for some margin between the capture noise level and a threshold for detecting Active Frames, we define output frames whose $v'(m)=M[V'(m),V'(m-1)] \leq 1.5\times N'$ to be Repeated Frames. For a source sequence, we define frames whose $v(m)=M[V(n),V(n-1)] \leq 1.5\times N$ to be **Indistinguishable frames**. There may be small differences between repeated or indistinguishable frames, but the measurement system cannot detect them reliably.

The selection and qualification of source sequences for testing shall take this threshold into account. One would not expect to detect Active Frames when frames in the source sequence are indistinguishable to the measurement device. This margin also fosters Active Frame detection with greater confidence.

6.2.3 Testing a Sequence for Distinguishable Differences

For a video sequence V, calculate the set of MSE values *v* and compare each member of the set with the threshold for indistinguishable frames (1.5×N). All frames V(n) whose M[V(n),V(n-1)] > 1.5×N possess distinguishable differences from their preceding frame. A channel under test shall be supplied with Input frames having distinguishable differences to test for Active frames and Repeated frames.

When considering source sequences for use with high quality transmission systems that preserve Field integrity, it is more appropriate to compare the current **Video Frame,** V(n), with V(n-2) to pair equivalent Fields and avoid comparison error from spatial offset between fields.

The following procedure gives conditional tests to ensure that a Video frame possesses distinguishable differences (for source sequences with interlaced fields).

1. Compute M[V(n),V(n-1)]
2. If result is ≤1.5N, declare frames indistinguishable, otherwise continue.
3. Compute M[V(n),V(n-2)]
4. If result is ≤1.5N, declare frames indistinguishable, otherwise continue.
5. Frame V(n) possesses distinguishable differences.

Further considerations on the subject of source or input scene characterization are discussed in section 6.2.6. Note that sequences created using a 3:2 pull-down process will fail this qualification, but can still be used under the provisions of section 6.2.7.

6.2.4 Categorizing Active frames and Repeated frames

For an output video sequence, V′, calculate the set of MSE values M[V′(m),V′(m-1)] and compare each value in the set with the threshold for indistinguishable frames (1.5×N′).

Note that many high quality transmission systems preserve Field Integrity, while also introducing minimal distortion. For these systems, it is also appropriate to compare the current **Video Frame,** V′(m), with V′(m-2) to pair equivalent Fields and avoid comparison error from spatial offset between fields. When testing at non-interlaced interfaces or using the recognized options for reduced capture rate and resolution, the comparison with V′(m-2) is probably unnecessary.

A frame V′(m) whose MSE results in M[V′(m),V′(m-1)] and M[V′(m),V′(m-2)] >1.5×N′, in response to an input sequence possessing distinguishable differences, has limited correspondence to either frame and shall be categorized as an **Active Frame**.

A frame V′(m) whose MSE results in M[V′(m),V′(m-1)] or M[V′(m),V′(m-2)] ≤1.5×N′, in response to an input sequence possessing distinguishable differences, has high correspondence to V′(m-1) or V′(m-2) and shall be categorized as a **Repeated Frame**.

6.2.5 Testing for Correspondence Between Frames (Matching Frames)

For Active frame m and an X frame input sequence, calculate the set of X MSE values, M[V′(m), V]. The input frame with the best correspondence is the one that produces the minimum MSE value in the set of

$$c_V(x) = M[V'(m), V(x)] \quad \text{for} \quad 1 \le x \le X$$

c_v is the set of MSE values for frame V′(m) in *comparison* with each frame in sequence V, and the input frame that best matches V′(m) is defined as

$$C_V = \min(c_v)$$

(the minimum error (MSE) represents the maximum correspondence or best match between frames).

A set of rules may improve the matching process and reduce ambiguity. There may be instances where a single Active frame corresponds closely to more than one input frame. Such cases should be minimized with matching criteria based on pixel comparison methods (MSE), but certain circumstances increase the likelihood of ambiguity. These are:

- Extreme spatial distortion due to low transmission bit rate, use of a low resolution digital frame format, etc.
- Source content - High motion (causing smearing or other distortion), repetitive motion, still intervals within a sequence.
- Low Active output frame rate leaves many source frames as possible matches.
- Use of frame interpolation can make matching more difficult.

Experience has shown that an implementation of this method can handle examples of these difficult circumstances.

As an aid to automating this method, the following rules may be helpful to resolve ambiguous matches:

1. Require one-to-one matching: no more than one Active Frame can match a given input frame. One possible explanation for a double match is that an Active Frame was falsely detected.
2. Enforce sequence: if V′(m) matches V(n), then the next Active Frame V′(m+2) shall match V(n+1), or V(n+2), or V(n+3), etc. V′(m+2) is not permitted to match V(n-1) or V(n), but such an event shall be flagged as a possible error.
3. Recognize minimum delay: No matches allowed resulting in delay less than t_{min}. Negative delay is precluded by $t_{min} \ge 0$.
4. Recognize no-match condition: Some Active frames may contain too much distortion to match with the transmitted sequence. Such frames shall be counted and reported, along with the no-match threshold used. Measurement system users shall determine the usual range of matching MSE values for the transmission system under test, and set this threshold above this range.
5. Diagnostics: The matching process could be repeated from the opposite end of the sequence to see if fewer ambiguous matches and no-matches occur.
6. Test next frame: If the next Active frame in the output sequence has a unique match in the transmitted sequence, use its match and enforce the rules above on the previous Active frame.

Select best at random: When ambiguity still exists following application of the above rules, random selection could be used. However, it is recommended that MSE be calculated with sufficient resolution to minimize such occurrences. Errors introduced in the distribution by a random process should cancel out over a sequence and some summary statistics would be unaffected. Such selections shall be counted and reported.

7. If the results using a specific scene tend to require extensive intervention and resolution using these rules, the measurement should be attempted using a different scene.

6.2.6 Source Sequence Qualification for the Mean Square Error Methods

The success of the MSE-based methods depend on the use of appropriate source sequences. As stated above, the source sequence shall possess frame-to-frame differences that are distinguishable to the measurement device, avoid repetitive motion, and avoid intervals of still video that will certainly cause matching ambiguity. When using a sub-frame area, the process shall adopt a similar area as a basis for qualification. The following procedure can determine suitable video clips for measurement:

1. Take the first video frame in the source sequence and compare it with all other frames in the sequence.
2. Record the count and position of all indistinguishable frames (as described in section 6.2.2).
3. Analysis: The interval between all indistinguishable frames shall be sufficient to resolve input-output alignment ambiguities using prior estimates of frame inter-arrival time and other information. For example, if it is known that transmission delay is <2 sec, indistinguishable frames may appear ≥2 sec apart.
4. Repeat the above steps with at least the first X frames (for example, X=60). We place emphasis on the early frames in the sequence because some comparison rule outcomes can be dependent on the prior results.

Also, note that:

Source sequences with similar adjacent frames (distinguishable, but by a small amount) will have greater opportunity to produce ambiguous matches with Active frames. A test of frame to frame differences will reveal the extent of these similarities, as described in section 6.2.3.

[1] specifies a set of source sequences that are appropriate for video teleconferencing/telephony system assessment. However, some sequences contain some still video, and those sections shall be avoided for delay measurement (the still sections are desirable for calibration). Other collections of scenes are available, and may be appropriate for other applications. Even live video is permitted, providing that the measurement system can capture and store both input and output sequences during a measurement, and that the input sequence is subsequently evaluated for suitability. All input sequences shall be evaluated using the specific measurement implementation according to the procedures given here and in section 6.2.3.

6.2.7 Considerations for use of 3:2 Pull-Down Source Sequences

There are several conditions which can result in duplicated frames (and hence, indistinguishable frames) in a source sequence:

1. The sequence content is still, conveying no motion or change.
2. A 3:2 pull-down process created the video sequence from film. Every fifth Field (in 525-line format) is a repetition of the previous corresponding field (even or odd).
3. After a 3:2 pull-down process, the video sequence is edited resulting in non-periodic Field repetitions.

Strictly speaking, the presence of any of these three conditions would cause the source sequence to fail the qualification tests. It is impossible to make either delay or frame rate measurements with a still sequence. However, some measurements are possible with the 3:2 pull-down cases.

There are two problems encountered when applying the MSE method of measurement with 3:2 pull-down sources:

- The process to match Active Frames with their source sequence counterparts is more complex when some source frames are repeated. If ambiguous matches occur, delay calculations are suspect for that frame. The ambiguity can be resolved by examining the source sequence for repeated frames.
- The actual frame rate may be higher than that calculated using the Active Frames alone. Some frames in the output sequence categorized as Repeated may have traversed the channel. These frames can be found through a supplementary matching process.

To make measurements with a 3:2 pull-down source sequence, it is necessary to create a record of Repeated Frames in the source sequence, as described in clause 6.2.4 (but applied to the source sequence). Following the Active/Repeated Frame categorization on the output sequence, we expect that the matching process will pair Active Frames with non-repeated source frames, and in some cases also with Repeated source frames. Rule 1 in clause 6.2.5 requires one-to-one frame matching, so we determine the better match by referring to the record of Repeated source frames and choosing the non-repeated source frame. Since original frames that result in Active Frames always come before their repeats, the matching process shall always follow the normal sequence from beginning to end (Rule 5 shall never be tried).

Once the matching process has been conducted on all Active Frames, a second matching process with the Repeated frames can begin. This pass considers both the Repeated source frames and any frames unmatched in the first pass, thereby avoiding most of the non-repeated source frames. If a unique match exists between Repeated frames in the source and output sequences (after matching the adjacent Active output frames) then the arrival time of the Repeated frame shall be included in the distribution of output inter-arrival times. Again, one-to-one frame matching shall be enforced, in case a Repeated source frame is duplicated by the transmission system. No temporal calculations will use Repeated output frames without unique matching source frames.

6.2.8 Factors That Influence Measurement Accuracy and Stability

In many video transmission systems, the decoder must supply Video Frames at its output according to a periodic display regimen (such as an analog composite interface). If the input and output display clocks are not synchronized, a buffer must be added in the decoder. When the decoder has a Video Frame ready for display, it must wait until the next output opportunity and thereby increase the overall system delay. We will call this decoder waiting interval the *output delay*.

The output delay is bounded by the interval between display updates. For transmission systems with composite interfaces that can update on Field boundaries, the maximum output delay is 16.7 ms. With updates on 525-line Frame boundaries, the maximum is 33 ms. The actual output delay will be some random value between 0 and the maximum.

When input and output waveform clocks have a small frequency offset, the output delay will vary over time. If the frequency offset is constant, the output delay will slew over its range at regular intervals. If the clocks are synchronized to independent Cesium beam controlled oscillators, the output delay will change <1 ms every 13,900 hours. If the clock accuracy is derived from independent quartz oscillators (with 2 ppm offset), the output delay will cycle through the entire 33 ms range every 4.58 hours.

Time stamps with sub-field resolution permit characterization of the output delay as an inextricable component of the overall transmission system delay.

6.3 In-Frame Time Code Methods of Measurement for Video

In some situations, it may be possible to insert visible symbols in the input video that can be used to uniquely identify each input frame. These symbols will be carried to the system output, and could be used in a simple way to measure the frame rate and delay. These methods, and restrictions on their use, are for further study.

7. Audio Measurements

7.1 Collecting Audio Frames for Measurements

7.1.1 Description of Audio Frames

An Audio Frame is a group of consecutive audio samples. The preferred number of samples in an Audio Frame depends on the audio sample rate, and is given in Section 5. Figure 11 below illustrates the position of Audio Frames within the audio Content and Coding Hierarchies (in this example, coding frames have shorter duration than Audio Frames).

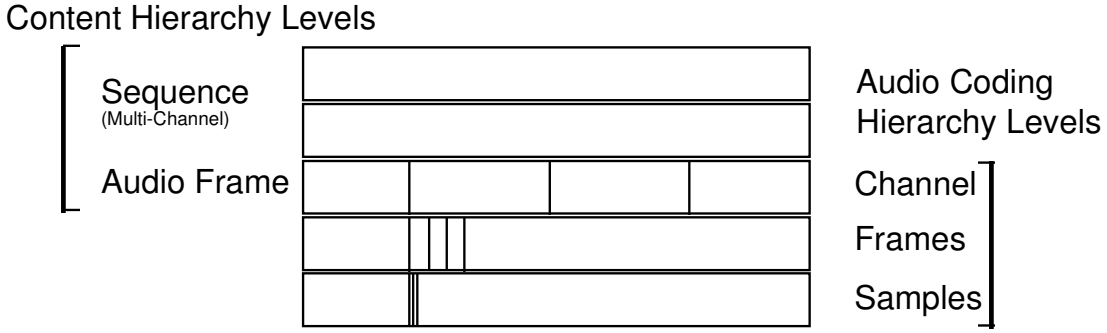


Figure 11 Audio Content and Coding Hierarchies

7.1.2 Analog to Digital Conversion

The methods of measurement described in subsequent sections require the digitization of the analog audio signal. The digitization process results in audio samples, which can then be grouped into Audio Frames. The audio sample rate is determined by the bandwidth required for the subsequent measurements. Using the Nyquist theorem, the sampling rate shall be at least twice this measurement bandwidth. For audio signals that are limited to speech, a sample rate of 8000 samples per second is sufficient. For other audio signals, higher sample rates may be required. The digitization process shall result in at least 8 bits of precision. Depending on the signal-to-noise ratio of the audio signal, additional precision up to 16 bits will often benefit the methods of measurement that follow. The digitization process shall include appropriate low-pass filtering to prevent aliasing, and shall be matched to the impedance and balance of the audio signals.

7.1.3 Time Stamp Assignment

The time, T(n), associated with the Audio Frame n shall be read immediately following the digitization of the last sample in Audio Frame n, and before the next sample is digitized. The time stamps for each sample within a frame may be computed from the Audio Frame time stamp, since the sample rate is known.

7.2 Delay Measurement for Audio

Many channels of potential interest are capable of delivering useable audio signals without preserving audio waveforms from input to output. Example audio coding specifications may be found in ITU-T Recommendations, G.728 (16kbps), G.729 (8kbps) CELP, and G.723 (6.4 and 5.3 kbps). This means that a robust delay measurement must not rely on audio waveforms alone. The measurement described here features a coarse stage that uses audio envelopes, and a fine stage that uses audio power spectral densities (PSD's). Audio envelopes and audio PSD's are approximately preserved by most channels. Figure 12 is a flow diagram for the measurement process.

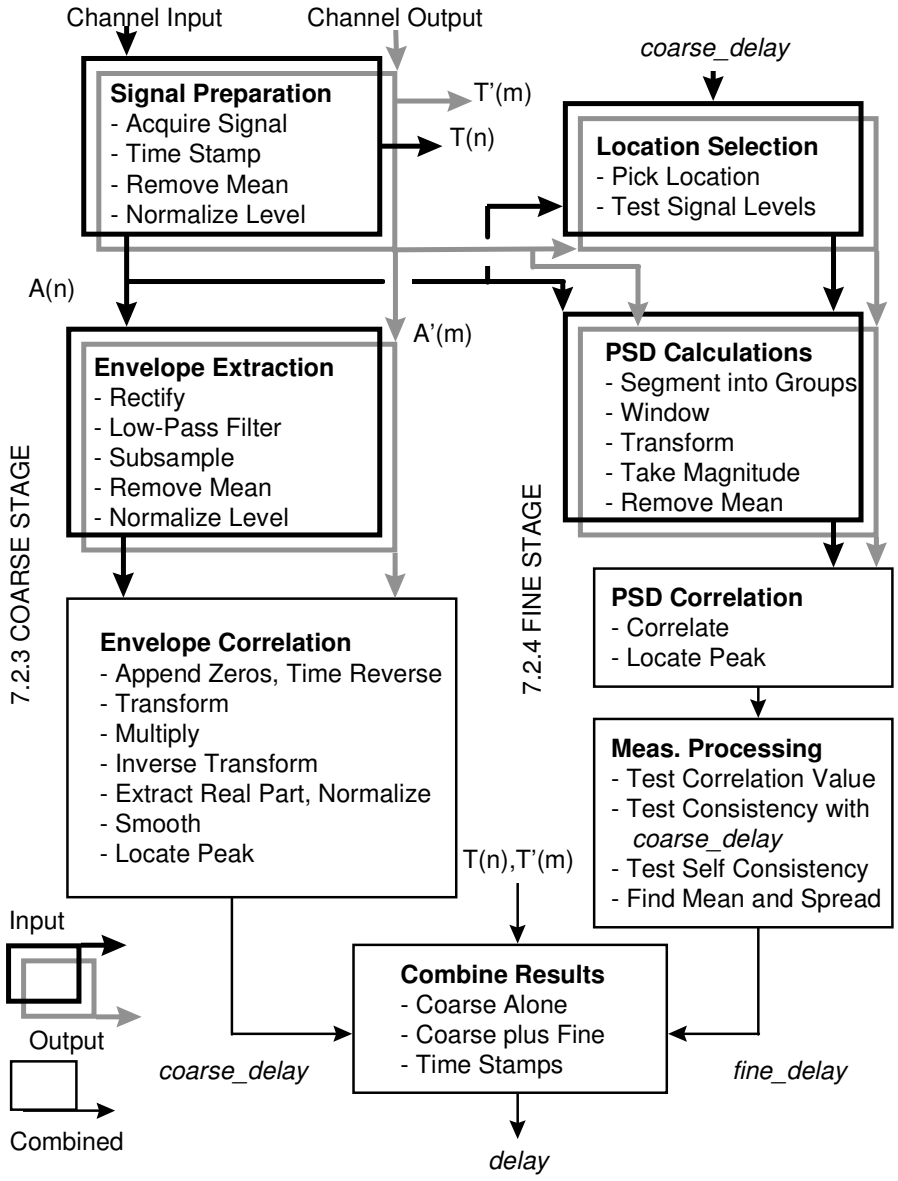


Figure 16 Flow Diagram for Audio Delay Measurement

The two stage process is efficient because the coarse stage searches a wide range of potential delay values, but at low resolution. If that same range were searched at high resolution, many more computations would be required. Once the coarse stage has finished its work, its low-resolution measurement can often be refined to a high resolution measurement by the fine stage that follows. The fine stage needs to search only a narrow range of potential delay values, consistent with the uncertainty of the coarse measurement. For some channels, audio PSD's are not adequately preserved and fine measurements are not possible. In other cases, multiple fine measurements will give inconsistent results. In these situations the coarse delay measurement, along with its inherent uncertainty, becomes the final delay measurement.

7.2.1 General

A series of Audio Frames shall be gathered from the channel input and the channel output before measurement can proceed. The use of more Audio Frames increases both the reliability and the complexity of the measurement. If a group of Audio Frames contains only the silence between words or phrases in a spoken conversation, or continuous tones, no reliable measurement will be possible, and additional Audio Frames shall be acquired before measurement can proceed. To detect an insufficient audio level condition, the RMS level of the audio samples in the group of Audio Frames acquired from the channel input shall be compared with the nominal RMS level of the channel input. This nominal level may be taken from channel input specifications or it may be measured. If the RMS level of the samples in the group of acquired channel input Audio Frames is more than 30 dB below the nominal channel input level, then additional Audio Frames shall be acquired before the measurement can proceed. Similarly, the RMS level of the audio samples in the group of Audio Frames acquired from the channel output shall be compared with the nominal RMS level of the channel output. If the RMS level of the samples in the group of acquired channel output Audio Frames is more than 30 dB below the nominal channel output level, then additional Audio Frames shall be acquired before the measurement can proceed.

For typical speech signals, the use of larger groups of Audio Frames reduces the possibility that the group contains only silence. Beyond this consideration, more Audio Frames bring more data to the measurement, and the measurement will be more reliable. For audio signals that are limited to speech, it is preferred that 256 Audio Frames be taken from the channel input and the channel output. The measurement will also work with 128 or 64 frames. When the sample rate is 8000 samples per second, and each frame contains 128 samples, these choices correspond to approximately 4 seconds, 2 seconds, or 1 second of speech signal respectively. The expected audio delay shall not be more than 25% of the duration of the speech signal used in this measurement. When 256 frames (4 seconds) of speech signal are used, delays up to 1 second may be measured. When 64 frames (1 second) are used, only 250 ms of delay shall be measured. The measurements are most efficiently computed when the number of frames acquired is a power of two.

7.2.2 Signal Preparation

When a measurement of audio delay is required, the Audio Frame most recently acquired from the channel input A(n), is concatenated with some number of previously acquired Audio Frames (e.g. A(n), A(n-1), ... A(n-255)), to form the most recent time-history of channel input samples. Similarly, the Audio Frame most recently acquired from the channel output A(m), is concatenated with the same number of previously acquired Audio Frames (e.g. A'(m), A'(m-1), ... A'(m-255)), to form the most recent time-history of channel output samples. As given in Section 5, the time difference between these two acquisition processes, is T'(m)-T(n). Positive values indicate that acquisition at the channel output happens after acquisition at the channel input.

The input samples are placed in an array called *ref*, which contains samples *ref(1), ref(2), ... ref(L1).* The output samples are placed in an identically sized array called *test*, which contains samples *test(1), test(2), ... test(L1).* The mean value of each array is then removed in order to eliminate any DC component in the digitized audio signals:

$$ref(i) = ref(i) - \tfrac{1}{L1} \cdot \sum_{j=1}^{L1} ref(j), \qquad test(i) = test(i) - \tfrac{1}{L1} \cdot \sum_{j=1}^{L1} test(j), \qquad 1 \le i \le L1.$$

Next, each array is normalized to a common RMS level:

$$ref(i) = ref(i) \cdot \left[\frac{1}{L1-1} \sum_{j=1}^{L1} ref(j)^2, \quad \right]^{-\frac{1}{2}}, \qquad test(i) = test(i) \cdot \left[\frac{1}{L1-1} \sum_{j=1}^{L1} test(j)^2, \quad \right]^{-\frac{1}{2}}, \qquad 1 \le i \le L1.$$

7.2.3 Coarse Stage

The measurement methodology starts with a coarse stage that extracts and cross-correlates audio envelopes. Audio envelopes are approximately preserved by most channels.

7.2.3.1 Envelope Extraction

Audio envelopes are calculated as follows. The digitized audio signals in *ref* and *test* are rectified by taking the absolute value of each sample. Because the original digitized audio signals in *ref* and *test* will be required by the fine stage, the rectified signals, and other subsequent intermediate results are stored in the temporary arrays *ref_temp* and *test_temp:*

$$ref\_temp(i) = |ref(i)|, \quad test\_temp(i) = |test(i)|, \quad 1 \le i \le L1.$$

The rectified signals are then low-pass filtered to create audio envelopes with a bandwidth of approximately 125 Hz. It is this low-pass filtering and subsequent subsampling that gives the coarse stage its reduced resolution and reduced computational load. The bandwidth reduction factor and the subsampling factor are both specified by the variable *B*. Appropriate values of *B* for some common audio sample rates are given in Table 2.

Table 2 Values of the bandwidth reduction factor, B.

Audio Sample Rate (samples/second)	B
8000	**32**
16,000	**64**
32,000	**128**
44,100	**176**
48,000	**192**

When the audio sample rate is 8000 samples per second, the bandwidth shall be reduced by a factor of *B*=32, from a nominal bandwidth of 4000 Hz to a nominal bandwidth of 125 Hz. The required bandwidth reduction can be adequately approximated using a seventh order, Infinite Impulse Response (IIR), low-pass Butterworth filter with a -3 dB point at 125 Hz. The direct form implementation is:

$$out(i) = \sum_{j=0}^{7} b_j \cdot in(i-j) - \sum_{j=1}^{7} a_j \cdot out(i-j), \quad 1 \le i \le L1,$$

$$where\ out(i) = in(i) = 0, \quad i \le 0.$$

Care shall be taken to eliminate any filter output samples that might contain a filter start-up transient. For the direct form implementation shown here, the start-up transient is limited to approximately 400 samples. The filter coefficients are given in Table 3.

Table 3 Coefficient values for 7th order, IIR, low-pass Butterworth filter.

j	a_j	b_j
0	1.00000000	0.00553833 x 10^{-7}
1	-6.55883158	0.03876830 x 10^{-7}
2	18.44954612	0.11630512 x 10^{-7}
3	-28.85178274	0.19384125 x 10^{-7}
4	27.08958968	0.19384206 x 10^{-7}
5	-15.27097592	0.11630465 x 10^{-7}
6	4.78557610	0.03876843 x 10^{-7}
7	-0.64312159	0.00553831 x 10^{-7}

Both the *ref _temp* and *test _temp* arrays are low-pass filtered using this filter. Next *ref _temp* and *test _temp* are subsampled by retaining only every B^{th} sample, resulting in a total of *L2* samples. For example, when B=32, samples 1, 33, 65, etc. would be retained. When 256 Audio Frames, each with 128 samples are used as input to the coarse stage, *L1*=32,768, and *L2*=*L1*/*B*=1024 samples result from the subsampling process. Both *ref _temp* and *test _temp* now contain audio envelopes. Finally, the audio envelopes in *ref _temp* and *test _temp* are normalized. The mean value of each array is removed, and each array is divided by its standard deviation to normalize each to a common RMS level.

$$ref\_temp(i) = ref\_temp(i) - \frac{1}{L2} \cdot \sum_{j=1}^{L2} ref\_temp(j),$$

$$test\_temp(i) = test\_temp(i) - \frac{1}{L2} \cdot \sum_{j=1}^{L2} test\_temp(j),$$

$$ref\_temp(i) = ref\_temp(i) \cdot \left[\frac{1}{L2-1} \sum_{j=1}^{L2} ref\_temp(j)^2 \right]^{-1/2},$$

$$test\_temp(i) = test\_temp(i) \cdot \left[\frac{1}{L2-1} \sum_{j=1}^{L2} test\_temp(j)^2 \right]^{-1/2}, 1 \le i \le L2 .$$

7.2.3.2 Envelope Cross-Correlation

The cross-correlation between the audio envelopes in *ref _temp* and *test _temp* is calculated by way of a circular convolution, which in turn is calculated by way of Discrete Fourier Transforms (DFT's) or Fast Fourier Transforms (FFT's). First, the array *ref _temp* is extended from length *L2* to length $2 \cdot L2$ by appending *L2* zeros. In the example above, *L2*=1024 zeros would be added to arrive at a final array size of 2048. Next the array *test _temp* is time-reversed. To do this in-place, samples 1 and *L2* of *test _temp* are exchanged, as are samples 2 and *L2-1*, samples 3 and *L2-2*, etc. When *L2* is even, the final exchange is between samples *L2*/2 and *L2*/2 + 1. When L2 is odd, the final exchange is between samples *L2*/2 - 1/2, and *L2*/2 + 3/2. After this time reversal, *test _temp* is extended from length *L2* to length $2 \cdot L2$ by appending *L2* zeros.

Now *ref _temp* and *test _temp* are transformed using DFT's or FFT's. When the array length, $2 \cdot L2$, is a power of two, FFT's can be used. If $2 \cdot L2$ is not a power of two, DFT's can be used. As an

alternative, the number of zeros appended in the previous step may be increased so that the array length is a power of two and FFT's may then be used. In any case, an in-place transformation algorithm may be used, resulting in transformed versions of *ref _temp* and *test _temp* overwriting the previous versions. The transformations result in complex numbers.

Next, the complex samples stored in *ref _temp* and *test _temp* are multiplied, sample by sample, and the complex results go into a new array called *cross_corr*, which has the same length as *ref _temp* and *test_temp*:

$$cross\_corr(i) = ref\_temp(i) \cdot test\_temp(i), \text{ for } i = 1 \text{ to } 2 \cdot L2 \ .$$

The array *cross_corr* is now Inverse Fast Fourier Transformed or Inverse Discrete Fourier Transformed, as dictated by its length. An in-place transformation may be used. In theory, the resulting contents of *cross_corr* would be real numbers. In practice, finite-precision calculations yield a small imaginary component. At this point, the real part of *cross_corr* is retained and the imaginary part is discarded. Next, each result in *cross_corr* is normalized:

$$cross\_corr(i) = cross\_corr(i) / (L2 - 1), \ 1 \le i \le 2 \cdot L2 \ .$$

Note that this normalization is required in order to get true cross-correlation values between -1 and +1, but it does not affect the smoothing or peak-finding steps that follow.

The array *cross_corr* holds the values of the cross-correlations between the speech envelopes in *ref_temp* and *test_temp* at every possible shift of those envelopes. These results are then smoothed with a symmetric, second-order, low-pass FIR filter, and stored in a smoothed cross-correlation array:

$$cross\_corr\_s(i) = .25 \cdot cross\_corr(i-1) + .5 \cdot cross\_corr(i) + .25 \cdot cross\_corr(i+1) \ , \ 2 \le i \le 2 \cdot L2 - 1,$$
$$cross\_corr\_s(i) = cross\_corr(i), \ i = 1, 2 \cdot L2.$$

After this smoothing, the largest value in *cross_corr_s* is taken as an indication of the coarse delay:

$$coarse\_delay = (L2 - j) \cdot B \ \text{ samples},$$
$$\text{where } cross\_corr\_s(j) > cross\_corr\_s(i), \ 1 \le i \le 2 \cdot L2, \ i \ne j.$$

The uncertainty in the value of *coarse_delay* at this point is taken to be ± *B* samples. If *cross_corr_s* does not contain a unique maximal value, then the measurement shall be made again using new audio samples.

7.2.4 Fine Stage

In many cases, the ±*B* sample uncertainty inherent in the coarse measurement of audio delay can be reduced by a fine stage of the delay measurement.

7.2.4.1 Location Selection

The fine stage is performed at *n1* locations in the acquired audio signal. When audio signals are limited to speech and 256 Audio Frames are used in the measurement of audio delay, the value of *n1* is 6. Other values of n1 may be more appropriate for other audio signals. At each location, a range of potential delay values from -3·*B* to 3·*B* samples is searched.

The locations where the fine stage is performed are randomly selected. At each location, $8 \cdot B$ samples are taken from the array *ref* and are stored in *ref_temp* and $2 \cdot B$ samples are taken from the array *test* and are stored in *test_temp*. The samples taken from *test* are offset by the measured coarse delay:

$ref\_temp(i) = ref(\text{loc}ation\text{-}4 \cdot B\text{-}1+i),\ 1 \le i \le 8 \cdot B$,

$test\_temp(i) = test(location+coarse\_delay\text{-}B\text{-}1+i),\ 1 \le i \le 2 \cdot B$,

where *location* is a uniformly distributed pseudo-random variable from the interval:

[max($4 \cdot B$ +1 , 1-coarse_delay+B) , min($L1$-$4 \cdot B$ +1 , $L1$-coarse_delay-B+1)].

The fine delay measurement will not work in silent regions or with steady tones. Two level tests for silent regions are conducted at each location to insure that the audio signal there is within 30 dB of the average audio signal level:

$$-30 \le 10 \cdot \log_{10}\left[\frac{1}{8 \cdot B-1}\sum_{i=1}^{8 \cdot B} ref\_temp(i)^2\right],\ -30 \le 10 \cdot \log_{10}\left[\frac{1}{2 \cdot B-1}\sum_{i=1}^{2 \cdot B} test\_temp(i)^2\right].$$

If either of the level tests is failed, then a new location shall be selected.

7.2.4.2 Power Spectral Density Calculations

The fine stage works by cross-correlating audio power spectral densities (PSD's) at each of the selected locations. The PSD's are calculated as follows. The $8 \cdot B$ samples in *ref_temp* are broken into groups of $2 \cdot B$ samples per group. There are $6 \cdot B+1$ such groups. Each group of samples is stored in an array called ref_temp_i:

$ref\_temp\_i(j) = ref\_temp(i+j\text{-}1)$, $1 \le i \le 6 \cdot B+1,\ 1 \le j \le 2 \cdot B$.

Each *ref_temp_i* array and the *test_temp* array is multiplied by a Hamming window, and then transformed to the frequency domain using a length $2 \cdot B$ DFT or FFT. These steps can be done in place:

$ref\_temp\_i(j) = ref\_temp\_i(j) \cdot \{.54 - .46 \cdot \cos(2\pi(j-1)/(2 \cdot B\text{-}1))\},\ 1 \le i \le 6 \cdot B+1,\ 1 \le j \le 2 \cdot B$,

$test\_temp(j) = test\_temp(j) \cdot \{.54 - .46 \cdot \cos(2\pi(j-1)/(2 \cdot B\text{-}1))\},\ 1 \le j \le 2 \cdot B$,

$ref\_temp\_i = \text{FFT}(ref\_temp\_i)$, $1 \le i \le 6 \cdot B+1$,

$test\_temp = \text{FFT}(test\_temp)$.

In the frequency domain, only the first B+1 complex samples in each array are unique, so only those samples are saved. The magnitude of each retained sample is taken, resulting in the square root of the power spectral density of each frame. These results are referred to as PSD's for simplicity.

$ref\_temp\_i(j) = |\, ref\_temp\_i(j)\,|$, $1 \le i \le 6 \cdot B+1,\ 1 \le j \le B+1$,

$test\_temp(j) = |\, test\_temp(j)\,|$, $1 \le j \le B+1$.

The mean value of each PSD is then removed:

$$ref\_temp\_i(j) = ref\_temp\_i(j) - \frac{1}{B+1} \cdot \sum_{j=1}^{B+1} ref\_temp\_i(j),\ 1 \le i \le 6 \cdot B+1,\ 1 \le j \le B+1,$$

$$test\_temp(j) = test\_temp(j) - \frac{1}{B+1} \cdot \sum_{j=1}^{B+1} test\_temp(j),\ 1 \le j \le B+1.$$

7.2.4.3 Power Spectral Density Cross-Correlation

A cross-correlation value is calculated between the PSD stored in the *test_temp* array and each of the 6·*B*+1 PSD's stored in the *ref_temp_i* arrays.

$$cross\_corr(i) = \frac{\left(\sum_{j=1}^{B+1} ref\_temp\_i(j) \cdot test\_temp(j) \right)}{\left(\sum_{j=1}^{B+1} ref\_temp\_i(j)^2 \right)^{\frac{1}{2}} \left(\sum_{j=1}^{B+1} test\_temp(j)^2 \right)^{\frac{1}{2}}}, \quad 1 \le i \le 6 \cdot B + 1.$$

The array *cross_corr* now holds the values of the cross-correlations between the reference and test PSD's at each time-domain shift. Note that the second term in the denominator of the equation for *cross_corr* is a normalizing constant that is required to get true cross-correlation values between -1 and +1. It does not have any impact on the peak-finding that follows, but does impact subsequent processing of the fine delay measurements. The largest value in *cross_corr* is taken as an indication of the fine delay:

$fine\_delay\_k = (3 \cdot B + 1) - j$ samples,

$corr\_k = cross\_corr(j), \; 1 \le k \le n1,$

where $cross\_corr(j) > cross\_corr(i), \; 1 \le i \le 6 \cdot B + 1., \; i \ne j.$

If *cross_corr* does not contain a unique maximal value, then the fine stage procedure shall be repeated at a new location. This entire fine stage, starting with the selection of a location, is repeated *n1* times, resulting in *n1* fine delay measurements stored in *fine_delay_1, fine_delay_2, ... fine_delay_n1,* and *n1* corresponding correlation values stored in *corr_1, corr_2, ... corr_n1*, respectively. Note that each of the fine delay estimates will fall between -3·*B* and 3·*B*, inclusive.

7.2.4.4 Fine Delay Measurement Processing

Once the *n1* fine delay measurements and corresponding cross-correlation values have been calculated, they are further processed to determine how they shall be used.

First, each of the *n1* correlation values are tested against a threshold:

$\sqrt{\frac{1}{2}} \le corr\_k \Rightarrow fine\_delay\_k$ is retained, $1 \le k \le n1.$

By this process, only fine delay measurements where at least half the PSD variance is accounted for are retained. The number of fine delay measurements that pass this test is *n2,* and the measurements are now renumbered as *fine_delay_1, fine_delay_2, ... fine_delay_n2.* If *n2* < *n1*/2, the fine stage will not produce a useful result. In this event, the value of *fine_delay* is set to "invalid" and the fine stage is terminated.

If *n2* ≥ *n1*/2, the fine stage continues and tests the remaining *n2* fine delay measurements for consistency with the coarse delay measurement. Since the uncertainty in the coarse delay measurement is ±*B* samples, and the coarse delay has been removed, only fine delay measurements between -*B* and *B* samples are retained:

$$|\, fine\_delay\_k \,| \le B \Rightarrow \; fine\_delay\_k \text{ is retained}, 1 \le k \le n2.$$

The number of fine delay measurements that pass this test is *n3,* and the measurements are now renumbered as *fine_delay_1, fine_delay_2, ... fine_delay_n3*. If *n3* < *n1*/2, the fine stage will not produce a useful result. In this event, the value of *fine_delay* is set to "invalid" and the fine stage is terminated.

If *n3* ≥ *n1*/2, the fine stage continues and tests for consistency among the remaining *n3* fine delay measurements. This test requires a search through all possible subsets of size *n3*, *n3*-1, on down to size *n1*/2. There is one possible subset of size *n3*, *n3*-1 possible subsets of size *n3*-1, *n3*·(*n3*-1)/2 possible subsets of size *n3*-2, and so forth. For each subset, the spread of the fine delay measurements is tested:

$$\max_i\{fine\_delay\_i\} - \min_i\{fine\_delay\_i\} \le \tfrac{B}{2}, \; fine\_delay\_i \in \text{current subset}.$$

The largest subset that passes this test is called the final subset. The fine stage fails to produce a useful result when:

- no subset passes this test, or
- there is not a single, largest subset that passes this test.

In either of these events, the value of *fine_delay* is set to "invalid" and the fine stage is terminated.

The number of fine delay measurements in the final subset is *n4*. These *n4* fine delay measurements are now renumbered as *fine_delay_1, fine_delay_2, … fine_delay_n4.* The mean value of these *n4* fine delay measurements is taken as the final fine delay measurement:

$$fine\_delay = \tfrac{1}{n4} \cdot \sum_{i=1}^{n4} fine\_delay\_i \; .$$

The spread of the *n4* measurements in the final sub-set is retained as a measure of uncertainty in the final fine delay measurement:

$$spread = \max_i\{fine\_delay\_i\} - \min_i\{fine\_delay\_i\}, \;\; fine\_delay\_i \in \text{final subset}.$$

6.2.8 Combining Coarse and Fine Stage Results

If the fine stage was not able to produce a useful fine delay measurement, then the fine stage will have set *fine_delay* to "invalid". In this case, the coarse measurement alone becomes the delay measurement. If the fine stage was able to produce a useful fine delay measurement, then the coarse measurement is augmented by that fine measurement and the uncertainty is reduced from that of the coarse measurement alone:

$$fine\_delay = "invalid" \Rightarrow delay = coarse\_delay \pm B \text{ samples},$$
$$fine\_delay \ne "invalid" \Rightarrow delay = coarse\_delay + fine\_delay \pm spread \text{ samples}.$$

These values of *delay* are correct only when the acquisition of audio samples from channel input and the channel output are simultaneous. After delay is converted to seconds, time stamps can be used to correct the delay measurement for non-simultaneous acquisition:

$$delay = delay / sample\_rate \text{ seconds,}$$

$$delay = delay + T'(m) - T(n) \text{ seconds.}$$

7. Combined Audio/Video Measurement Considerations

Using the concepts and methods defined in Sections 6 and 0, we can discuss some measurement issues for multiple channels.

7.1 Audio/Video Channel Activity and Synchronization Measurements

Both the video and audio methods of measurement require minimum signal levels in the channels under test in order to produce valid results. Each method has its own specific requirements. Video methods require distinguishable differences between the current and previous Video Frame, while audio methods require the RMS level of a group of Audio Frames to compare favourably with the nominal interface levels.

To be able to compare the audio and video delay measurements, the necessary activity conditions shall be present concurrently at each input, and valid measurements shall be accomplished in both channels (which relies on output activity). Otherwise, synchronization calculations are not possible and a different opportunity shall be sought.

7.2 Associating Individual Measurements

The synchronization calculations of Section 5 require an association between input frames and a frame matching process to yield pairs of time stamps for each channel tested. Calculation of the time offset between channels and the time error introduced by the transmission channel do not compute delay as an intermediate step. Strictly speaking, this prevents comparing delays measured at two different times to assess the synchronization of two channels.

However, audio channels represent a reasonable exception because they carry an isochronous media. When the measured audio delay variation is limited to the expected experimental error, then the average audio delay can be considered a representative constant value. This average delay can be compared with a video or data channel delay distribution to obtain a distribution of time skew between the channels. This allows comparison of audio and video measurements made singly, but under identical source (and other) conditions. This way, test devices that cannot make simultaneous multiple channel measurements may still supply useful information when conditions permit.

If the measured audio delay variation is beyond the expected experimental error, then single channel measurements may not be used.

8. Data Measurements

This section specifies the methods for delay measurements on data channels that are part of multimedia communications systems. Many data performance standards define delay or transmission time for specific communications protocols (such as ITU-T Recommendation X.25 and the various recommendations defining ATM).

8.1 Collecting Data Frames for Measurement

The Data Channels implemented in multimedia communication systems can vary widely in terms of their purposes and specific attributes. Rather than attempt to deal with all possibilities, this section gives the general methods applicable to Data Channels in two main categories (defined below).

8.1.1 Considerations for Defining Data Frames

This Recommendation deals with the Data Channel at a logical level, above the physical layer and its electrical interfaces. However, systems capable of the measurements described in this standard will have test facilities conforming to one or more electrical interface standards. Observations made at these electrical interfaces will be the basis for measurements.

This Recommendation refers to a sequence of Z consecutive bits as a Data Frame. Data Frame n is represented by D(n), and the first bit in D(n) is D(1,n). The length of a Data Frame may be determined by the application of the Data Channel.

We consider two possible configurations for user data:

1. Users submit information bits embedded within a standardized *structure*. These structures may be called packets, cells, or frames.
2. Users submit *unstructured* streams of bits. The multimedia communication system may perform its own segmentation on this bit stream.

When the user submits a structured bit stream, and that structure permits recognition of individual frames at each channel interface, then the native structure is considered the Data Frame for measurements of multimedia systems. A possible exception is when the native structure contains a large number of bits, and the structure insertion time is large compared with Audio Frames and Video Frames. In this case, it may be more efficient to treat the bit stream as unstructured. The ideal circumstance is equal insertion time for frames in all media, permitting a one-to-one correspondence.

When the Data Channel permits unstructured input bits, and it is possible for the measurement system to supply the bits, then a pseudo-random sequence generator may be used. This gives several advantages:

- The sequence can be generated easily at local and remote sites.
- The repeating length of the sequence can be chosen to avoid ambiguous matches.
- Data Frame length may be as small as a single multiple of the length of the linear feedback shift register, and may be chosen to closely match the length of other media frames.

When the Data Channel requires a large structure, the pseudo-random sequence generator may supply the information bits carried by the structure.

It may also be possible for the measurement system to collect the input bit stream from the Data Channel's usual source. In this case the Data Frame length will usually coincide with the native structure. When the data source is producing an idle pattern, successful measurements are highly unlikely.

8.1.2 Time Stamp Assignment

The time, $T_D(n)$, associated with Data Frame n shall be read immediately following the communication of the last bit in the frame across the interface and before the next bit is communicated across the interface.

Since frame insertion time is also a useful data transmission measure, additional time stamps may be associated with the first bit of a Data Frame, and shall be read before the next bit is communicated across the interface. Input insertion time may be different from the output insertion time in some systems. Further, insertion time may not be constant.

8.2 Delay Measurement for Data

This section gives two methods to match input and output Data Frames.

8.2.1 Matching Structured Data

Bibliography [4] discusses methods to determine correspondence among X.25 packets that are applicable to many forms of structured data. Usually the header bits communicate sufficient information, such as a sequence number, so that packets can be differentiated from one another. These embedded identifiers are a valid basis for matching Data Frames.

If the header information of a specific protocol is insufficient, then it may be possible to match additional identification within the user data field.

8.2.2 Matching Unstructured Bit Streams

Bibliography [5] defines correspondence between bits by comparing sequences of bits. Using the symbols of this standard, bits D′(m) correspond to an equal length input sequence if there exist integers n and d such that

$$D'(i,m)=D(i+d,n) \text{ for almost all integers } 1 \leq i \leq Z-d$$

$$\text{and } D'(i,m)=D(i-Z+d,n+1) \text{ for almost all integers } Z-d+1 \leq i \leq Z$$

where d is the positive integer offset (d<Z) that may exist between input, D(n), and output, D′(m), frame assignments.

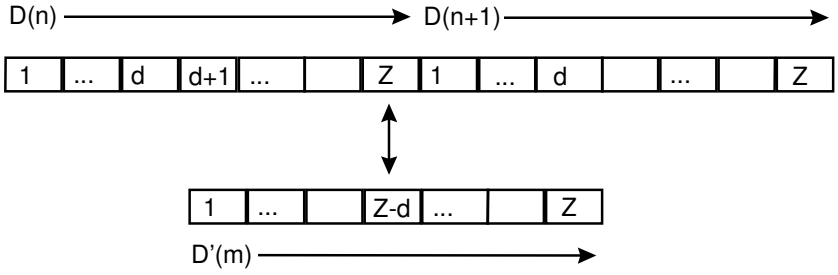


Figure 13 Correspondence Between Input and Output Data Frames

Figure 13 illustrates correspondence across input frame boundaries.

If the input or output Data Frames are re-aligned such that their bit offset is fixed to d=0, then the correspondence test simplifies to

$$D'(i,m)=D(i,n) \text{ for almost all integers } 1 \leq i \leq Z$$

If bits are communicated across the interfaces in a periodic manner, it is possible to calculate the time stamp for any bit in a Data Frame, making the time stamps available for the first or last bit in any new definition of D(n) that achieves d=0.

Allowing correspondence over *almost all* bits in a Data Frame permits successful matching in the presence of limited bit errors. When there are no bit errors, there shall be equality for all integers i in the range $1 \leq i \leq Z$.

9. Timer Stability and Synchronization Requirements

This section gives the minimum specifications for the internal timers or clocks that supply the time stamps for frames. There are two clock configurations to consider:

1. A single clock supplies the input and output time stamps (usually found in local and remote loop-back measurement applications). In this case, only the specifications for accuracy, stability, and resolution apply, since they fully characterize one clock's performance.
2. Two clocks, possibly in different (remote) locations at the input and output of the transmission system, supply the time stamps. This configuration applies in the end-to-end measurement application. All specifications of this section apply in this configuration.

9.1 Resolution

The minimum resolution of the time scale available for inclusion in time stamps is 0.1 μsecond (10^{-7} second). This is the intended internal storage resolution for measurements. Although this full resolution shall not be reported when internal clock accuracy does not support it, it allows developers a fine basis for clock stability/accuracy evaluation.

9.2 Accuracy and Stability (Allowable Time Interval Error)

The accuracy and stability of the internal clock time scale is fully constrained with a specification on Maximum Time Interval Error. Time Interval Error (TIE) is defined as the time variation of a given time clock's readings with respect to an ideal time scale over a particular observation period, S. Maximum Time Interval Error (MTIE) is the largest TIE for all possible measurement intervals within the observation period.

In practice, the transmission measurements conducted according to this standard will last 1 second or more. Therefore, the MTIE specification will begin at 0.01 seconds observation interval (smaller intervals are not specified).

In many applications of this standard, the clock(s) will be synchronized with a time reference signal, such as the Global Positioning System (GPS). In this case, the MTIE is described by the following equation

$$\text{mtie,ns} \leq 10^{-2} S + 150$$

and illustrated in Figure 14.

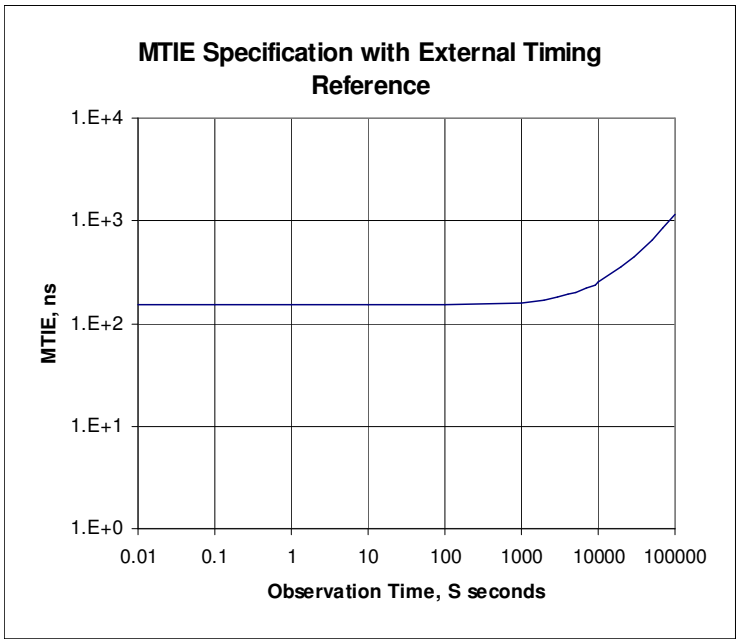


Figure 14 MTIE with Timing Reference

Other applications will use a single internal clock, or a remote clock that can be synchronized before measurement but then relies on its internal (a.k.a. holdover) accuracy to maintain time. This permits measurements for some limited time period when a primary timing source (e.g., GPS) is unavailable. For Type A internal clocks, the MTIE is constrained by

$$\text{mtie,ns} \le 10S + 150$$

and illustrated in Figure 15.

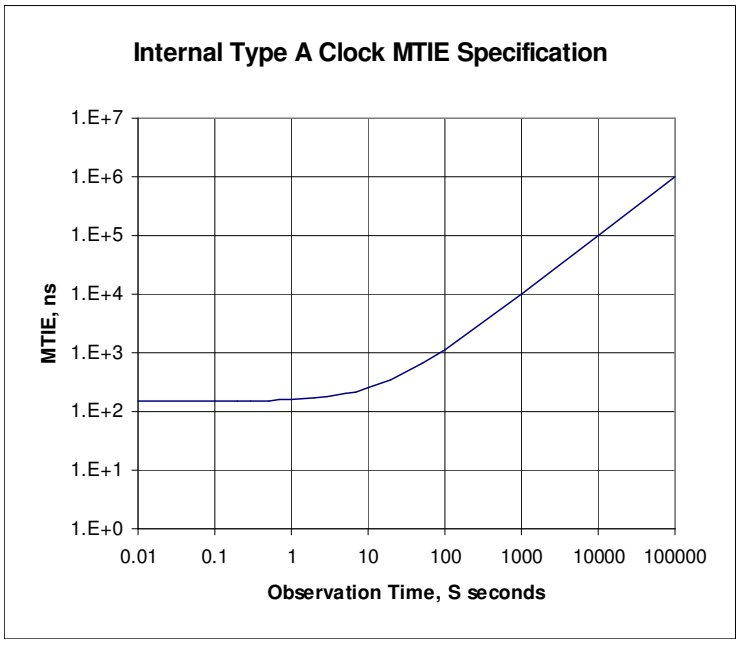


Figure 15 Internal Type A Clock MTIE Spec. (during measurements)

For Type B internal clocks, the MTIE is constrained by

$$mtie, ns \leq 138.9 \times S + 150$$

and illustrated in Figure 16.

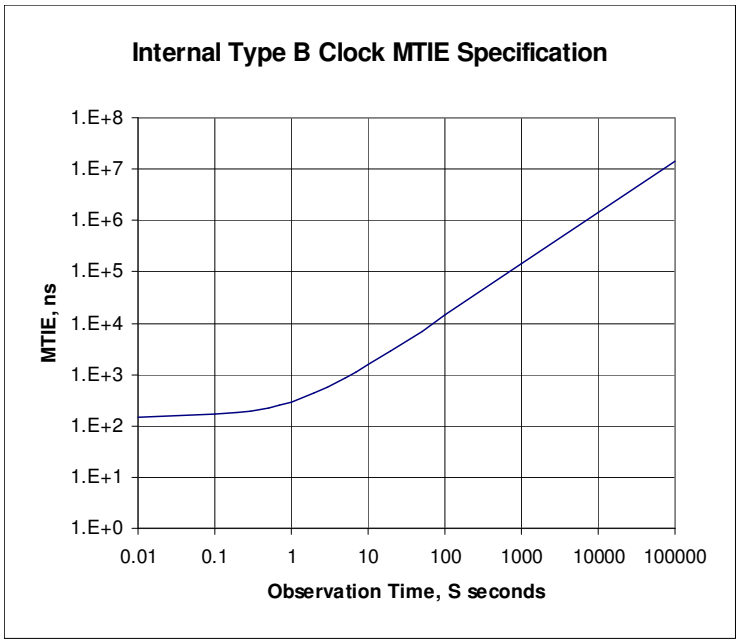


Figure 16 Internal Type B Clock MTIE Spec. (during measurements)

In all cases, measurement reports shall be accompanied by the maximum error (determined by the presence of a reference source), including the time elapsed since the reference was available, and the actual measurement interval.

9.3 Time Setting Error

If two or more clocks are used in a measurement, they shall be synchronized. When clocks are synchronized directly, or synchronized to some third reference clock, the maximum setting error will be ±0.075 µsecond (7.5×10^{-8} second).

APPENDIX I

Bibliography

[1] ANSI T1.801.01-1995 American National Standard for Telecommunications - Digital Transport of Video Teleconferencing/Video Telephony Signals - Video Test Scenes for Subjective and Objective Performance Assessment

[2] ANSI T1.801.02-1996 American National Standard for Telecommunications - Digital Transport of Video Teleconferencing/Video Telephony Signals - Performance Terms, Definitions, and Examples.

[3] ANSI T1.801.03-1996 American National Standard for Telecommunications - Digital Transport of One-Way Video Signals - Parameters for Objective Performance Assessment.

[4] ANSI T1.504a-1991 American National Standard for Telecommunications - Packet Switched Data Communication Service - Performance Measurement Methods.

[5] ANSI T1.517-1995 American National Standard for Telecommunications - Performance Parameters and Objectives for Integrated Services Digital Networks.

[6] ANSI T1.314-1991, Video Codec for Audiovisual Services at $p \times 64$ kbits.

[7] SMPTE RP 27.3-1989[3], Recommended Practice, Specifications for Safe Title Areas, Test Pattern for Television Systems.

[8] SMPTE 125M-1992, SMPTE Standard for Television - Component Video Signal 4:2:2 – Bit-Parallel Digital Interface. [3]

[9] SMPTE 259M-1993, SMPTE Standard for Television - 10-Bit 4:2:2 Component and $4f_{SC}$ NTSC Composite Signals – Bit-Parallel Digital Interface. [3]

[10] SMPTE 170M-1994, SMPTE Standard for Television - Composite Analog Video Signal - NTSC for Studio Applications[2].

[11] ITU-T Contribution COM 12-75-E, "Visual Channel Delay and Frame Rate Measurement - Initial Results with a Prototype System," International Telecommunication Union - Telecommunications Standardization Sector, Study Period 1993-1996, Study Group 12 White Contribution, source AT&T, March 1996. [2]

[12] A.N.Netravali and B.G.Haskell, Digital Pictures: Representation and Compression, Plenum Publishing Corporation, New York, NY, 1988.

---

[2] Available from the Society of Motion Picture and Television Engineers (SMPTE),595 West Hartsdale Ave. White Plains, NY, 10607.

APPENDIX II

Mathematical Symbol and Convention Key

V	a sequence of adjacent video frames at the channel input
V′	a sequence of adjacent video frames at the channel output
V′(m)	an output video frame at time T′(m)
V′(i,j,m)	luminance pixel (i,j) in output video frame m at time T′(m)
V*(i,j,m)	output luminance pixel before correction factors are applied
g	gain correction factor
l	level offset correction factor
h	horizontal shift correction factor
v	vertical shift correction factor
z	frame size correction factor
A	a sequence of adjacent audio units at the channel input
A′	a sequence of adjacent audio units at the channel output
D	a sequence of adjacent data units at the channel input
D′	a sequence of adjacent data units at the channel output
T	the ‘timer media’ stream at the input
T′	the ‘timer media’ stream at the output
T_P	the set of time stamps associated with general presentation units at the input
T'_P	the set of time stamps associated with output presentation units
$T'_P(m)$	the timer value (time stamp) associated with presentation unit m of a general media stream (at the output)
$T'_P(m-1)$	the timer value of the presentation unit preceding unit m
T_A	the set of time stamps associated with input audio frames
T'_A	the set of time stamps associated with output audio frames
$T'_A(m)$	the timer value (time stamp) associated with output audio frame m
$t_P(m)$	the channel delay for presentation unit m
t_P	a set of channel delays measured for a media stream
$b_P(m)$	the inter-arrival time for presentation unit m
b_P	a set of inter-arrival times measured for a media stream

$f_P(m)$ the elementary frame rate for presentation unit m

M[V′(m), V′(m-1)] The Mean Square Error (MSE) between two adjacent frames

M[V′(m), V(n)] The MSE between an output frame and an input frame

$K_s=(I_{max}-I_{min}+1)\times(J_{max}-J_{min}+1)$ the total pixels in the spatial sub-region for MSE

PSNR Peak Signal to Noise Ratio calculated with peak video level, V_{peak}

v′ the set of MSE values for adjacent output frames within a video sequence

c the set of MSE values comparing an output frame to the input sequence

c_v the set of MSE values comparing an output video frame to the input sequence

a′ the set of comparison values for adjacent output frames within an audio stream

d′ the set of comparison values for adjacent output frames within a data stream

N′ is the calibrated output capture noise

C_P the comparison value for the presentation unit that best matches a specific Active unit

$o_{AV}(m,n)$ time offset between associated audio and video frames at the input

$o'_{AV}(p,q)$ time offset between associated audio/video frames at the output

$s'_{AV}(m,n)$ time skew between assoc. audio/video frames at the output due to the transmission system/channel

Variables used in Section 7.

B: bandwidth reduction factor and subsampling factor

coarse_delay: delay as measured by coarse stage

cross_corr: temporary array, ultimately holds cross-correlation values

cross_corr(i): i^{th} element of cross_corr array

cross_corr_s: smoothed version of cross_corr in coarse stage

cross_corr_s(i): i^{th} element of cross_corr_s array

delay: final output of two-stage delay measurement

fine_delay: delay as measured by fine stage

fine_delay_k: k^{th} fine delay measurement

L1: number of audio samples input to measurement

L2: number of audio samples after subsampling

location: location where fine stage makes a measurement

n1: number of measurements made by fine stage

n2: number of fine stage measurements retained after first test

n3: number of fine stage measurements retained after second test

n4: number of fine stage measurements retained after third test

ref: array of audio samples from channel input

ref(i): i^{th} element of ref array

ref_temp: temporary storage array for channel input audio samples as they are processed

ref_temp(i): i^{th} element of ref_temp array

ref_temp_i : temporary storage array for channel input audio samples as they are processed

ref_temp_i(j): j^{th} element of ref_temp_i array

sample_rate: rate at which channel input and channel output are digitized

spread: spread in the final subset of fine delay measurements

test: array of audio samples from channel output

test(i): i^{th} element of test array

test_temp: temporary storage array for channel output audio samples as they are processed

test_temp(i): i^{th} element of test_temp array

Variables used in Section 9

Z length of a Data Frame in bits

D(n) input Data Frame n

D(i,n) bit i in input Data Frame n

d offset in bits between input and output Data Frames when determining correspondence

Variables used in Section 10

mtie, ns Maximum Time Interval Error (MTIE), given in nanoseconds

S Observation interval for MTIE measurements

_____