

AUDIO TO VIDEO SYNCHRONIZATION ERRORS, SOURCES, MEASUREMENT AND CORRECTION

J. Carl Cooper
Chris Smith
Mirko Vojnovic

ABSTRACT

This paper provides an overview of audio to video synchronization errors and their correction. Some of the more common sources of errors are described; the problems which the errors create and solutions to their measurement and correction are outlined.

Audio to video synchronization errors in video systems are usually quite subtle, and are often caused by the buildup of timing errors from several locations, any one of which will often go unnoticed.

TELEVISION SYSTEM AUDIO DELAYS

In the typical television system the video and audio are carried on separate paths, leading to differential processing delays. The audio processing path is typically fairly straight-forward in respect to the separate delays it generates, often being composed only of the distance from the sound source to the microphone, a preamp and a couple of relatively simple (in respect to delays) mixing boards.

The microphone distance from the sound source may add a significant delay component to the audio, for example as in a sporting event or outdoor environment. For a sporting event such as a tennis match, the delay might be in the order of 50 milliseconds or 1.5 NTSC frames. By contrast, for typical microphone placement on an announcer's lapel, the sound delay in reaching the microphone might be around 1 millisecond.

Fortunately, in many instances the viewer does not normally perceive the delay resulting from microphone placement as a problem. When the viewer is present in the same environment which is being televised, he receives the same or worse delayed sound with respect to the viewing of the image which creates the sound. A spectator in a sporting event will consciously or subconsciously perceive the sound as being slightly delayed with respect to the vision; this being a natural occurrence which the television viewer will subconsciously relate to his personal experience. Unfortunately, the addition of even small amounts of additional delay will cause the normal delayed sound sensation to become an annoying over delayed sensation.

The rest of the television sound channel, depending on the complexity of the television system, will likely range in the order of one or two milliseconds per processing device. The total additional delay rarely goes over a few tens of milliseconds.

VISION DELAYS

Figure 1 shows a typical television system from production studio to home receiver.

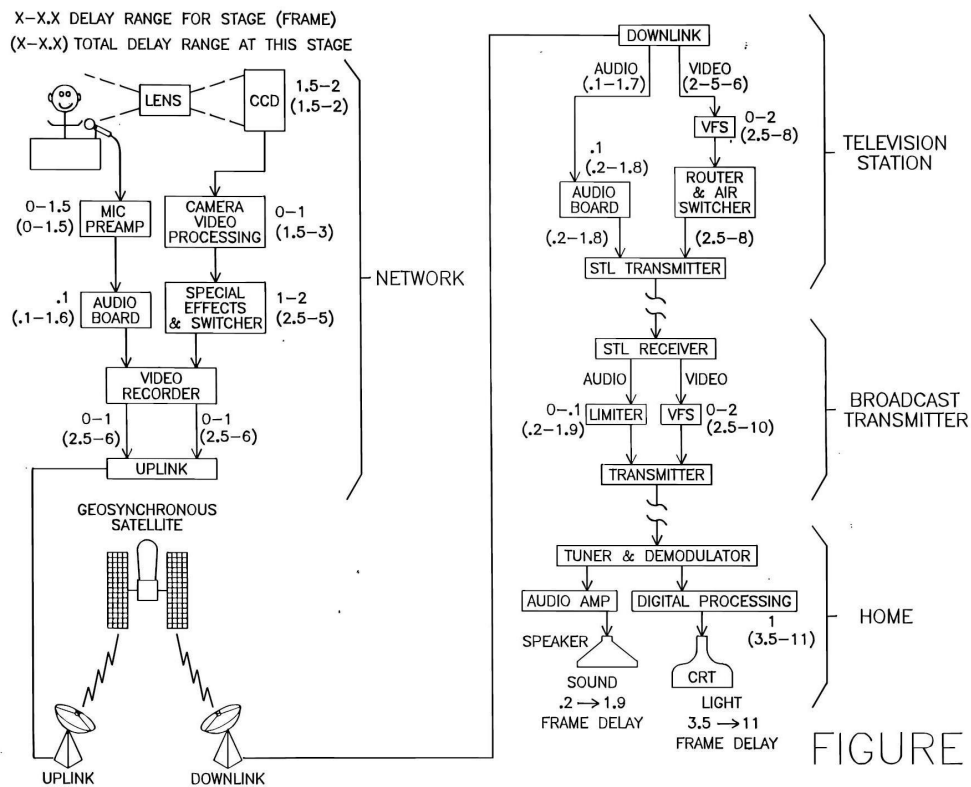


FIGURE 1

The diagram indicates the various delays for each of the several typical components of the video and audio signal paths, with the cumulative delay to that point shown in parenthesis. It is important to recognize that for even a simple system such as the one shown that audio advances of up to 11 frames can occur. It is also important to notice that much of the mismatch actually occurs after the television signal leaves the production environment, and thus can not be seen by the production engineering staff.

Virtually all cameras sold and in use today utilize CCD sensors. The visual portion of the television program which originates in front of the camera lens arrives at the light sensor in only a few nanoseconds at even the most extreme distances which can be associated with any sound. At the CCD sensor, the image light is integrated for some period of time, depending on the exposure setting. As indicated in **Figure 2**, for short exposures with shuttered CCDs, the time delay for converting the light energy into video signals will create an image delay of almost one frame.

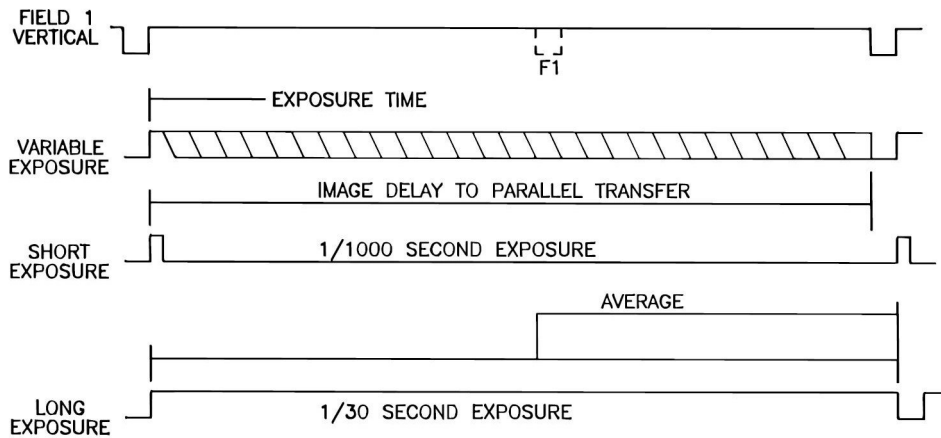


FIGURE 2

For maximum (1 frame) exposures, the corresponding image delay will be averaged over almost one frame. The improved temporal resolution of CCD cameras at short exposures improves the viewer's ability to accurately perceive motion, especially motion related to audio. In addition to the visual delay associated with the CCD integrating the image another delay occurs when the image is being shifted out of the CCD. **Figure 3** shows a simplified CCD sensor and shift register arrangement.

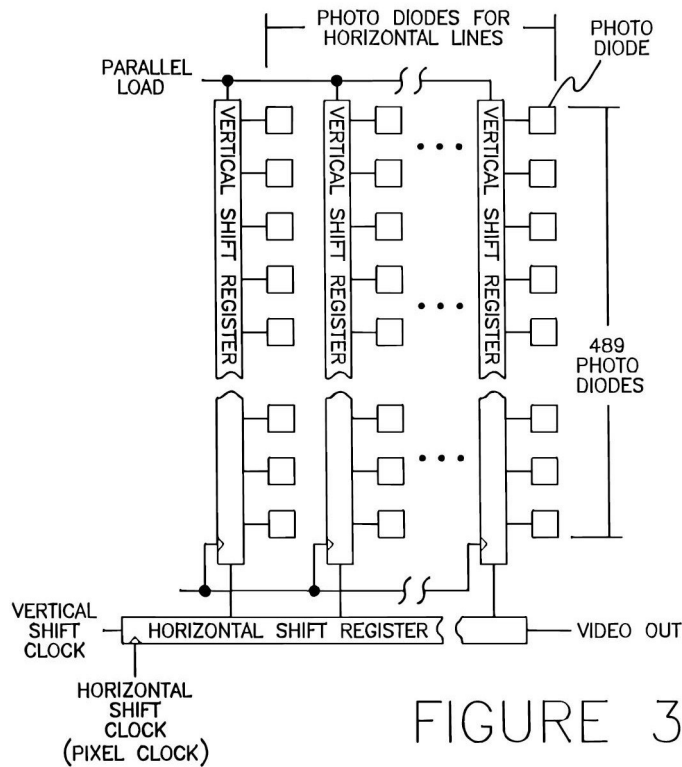


FIGURE 3

After vertical blanking, the charges which are accumulated in the sensor are transferred to the storage register where the charges they contain are clocked out line by line. Even though the entire image is taken in a "snapshot" in the CCD array, each pixel of the image receives a different delay due to the shifting mechanism. This shift out delay will be partially compensated by the subsequent scanning in the television receiver, if the receiver is of a scanning type. If the receiver is an LCD or other matrix type, the compensation will not necessarily be present.

VARIABLE TEMPORAL RESOLUTION IN THE CCD

It would be worthwhile to mention the effect that variable shutter speeds has on temporally sampling the image. At maximum exposure, that is a 1 frame shutter speed, the image is integrated over the entire frame. This long exposure blurs the image motion making it difficult for the viewer to accurately perceive the motion. With a fast shutter speed, the image is integrated over a relatively short time, for example 100~s for a 1/10,000 second exposure. The short exposure makes it much easier for the viewer to perceive motion because the video system has a greater temporal resolution that is the least amount of blurring. The increased temporal resolution also improves the viewer's ability to perceive the motion related to audio and video timing. In addition, one may recall from the psycho perceptual nature of television images that the response time of the viewer's eye is reduced for brighter objects. In other words, the ability to perceive motion is increased for bright objects. This is the same perceptual phenomena which causes bright areas of television displays to flicker when the darker portions do not, simply because the viewer's eye's response time is quicker for brighter objects. Assuming proper exposure levels, the edges of less blurred objects appear instantaneously brighter because during the shorter exposure time the light from any point or, edge on an object, is spread over relatively fewer CCD sensor diodes. The brighter and less blurred moving edges result in the viewer's improved ability to perceive higher temporal resolution. The reduced image smear on the CCD and increased viewer temporal perception at short exposures aggravates the correspondingly increased image delay time. Compared to tube cameras, CCD cameras make any audio to image mismatch easier for the viewer to consciously or subconsciously detect.

VIDEO PROCESSING DELAYS

In many production environments the video may also pass through several other delaying devices such as frame synchronizers, color correctors, noise reducers and a variety of nonlinear editing and image processing functions. Memory costs have declined allowing these devices to increase in complexity, including by use of frame based processing functions which add delays. As shown in **Figure 4**, video frame synchronizers are a common source of synchronization errors.

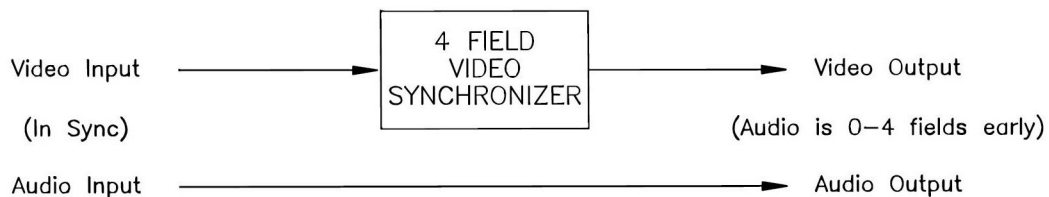


Fig. 4: Video Frame Synchronizer Causes Audio Sync Error

The correct audio and video sync is destroyed by the variable delay of the video synchronizer. Currently, 8 field frame synchronizers are in common use for high end applications. The video delays in a system may be continuously variable, such as shown for frame synchronizers, or may be switched in and out as the operator selects different modes of operation. The problem of variability of delay is especially true of many current noise reduction and color correction products where extra frames of delay are added for each additional selected function. The diagram of

Figure 1 shows how the use of cascaded switchers and video synchronizers can bring total system delay errors to over 10 frames.

HOME RECEIVERS

While there are a wide variety of home receivers available, the ones which are of largest concern are large screen types which often contain internal digital processing of the video. Due to the introduction of extremely low cost A-D converters many receiver manufacturers have found that it is possible to incorporate a considerable amount of field and frame based digital signal processing into their designs. These circuits can introduce yet another frame of image delay. At this time there are very few consumer TVs which provide any sort of compensating audio delay. As the cost of digital image processing circuitry continues to decrease, more and more frame based circuits will be found in consumer TVs. If the past is any indication, many products which are now found exclusively in production houses and TV stations will begin to appear as features of high end consumer products. All of these additional features may very well add additional frames of delay to the video signal before the viewer sees it.

VIEWER PERCEPTION PROBLEMS

The most obvious result of audio to video mismatch is visible "lip sync" errors. The frequency of these visible errors used to be relatively small but seems to be increasing significantly, and the need for correction is apparent. Additionally, when there is a large but only slightly discernible error, viewers are often not precisely aware that the problem exists.

Unfortunately, just because the viewer does not identify the problem, does not mean he is not affected by the problem. When the audio is advanced with respect to the video, the mistiming will cause a subconscious degradation of the program's entertainment quality as perceived by the viewer. The cause of this effect is believed to be the unnatural sound relationship which the television program presents. In our natural environment we are used to hearing audio slightly delayed with respect to video due to the slower speed of propagation of sound waves as compared to light. We are used to hearing the sound of a hammer after we see it hit, hearing a racquet striking after we see the ball hit and hearing a commercial actor after we see them talking. In television systems it is most often the video which is delayed, thus causing the sound to arrive at the viewer's ears before the visual sensation to which it corresponds.

Viewing a television program with advanced audio is very unnatural for the viewer, and therefore believed to cause subconscious stress to that viewer. It has been demonstrated in psychological tests at Stanford University¹ that viewers who watch television commercials having an audio advance "evaluate people on television more negatively (e.g. less interesting, more unpleasant, less influential, more agitated, less successful)" than the same commercials which were played with the audio in sync with the video. It was also discovered that this effect takes place with relatively small audio advances of 2.5 fields, where the existence of an audio problem could only be detected by very few average viewers.

It was also found in the Stanford tests that even when specifically asked, several of the test subjects which were negatively affected were completely unable to detect the delayed video. These test subjects were completely unaware that there were any audio-to-video synchronization problems at all, much less that their enjoyment of the program was being affected.

Watching a program where we hear action before we see it, or hearing commercial actors before we see them talking can be subconsciously annoying or stressful. Audio sync errors can cause the viewer to perceive a program or commercial less favorably than if the timing were correct, thus preventing the viewer from receiving the entertainment and messages intended by the advertisers. At current advertising rates, this becomes a potentially serious financial problem when the advertisers realize that they are not getting what they pay for due to audio sync errors.

SETTING PERFORMANCE STANDARDS

Several standards committees in various countries have addressed the problem and have set standards or guidelines for audio to video synchronization errors. For example, the Radio-communication Study Groups of The International Telecommunication Union states²:

"Given the operating practices employed in the United States and the requirement that a single picture and sound service may reach the consumer in different forms and via different paths, the list of preferred points should be as noted above and the tolerances required at each of the points should be the same (+1 field, -2 fields) with the understanding that these tolerances are absolute, are not accumulative, and apply to the overall system".

The International Telecommunication Union in the Draft New Recommendation [DOC. 11/59]3 reports that in tests in Australia that errors of and greater than +20 and -40 ms were "detectable" and errors of +40 and -160 ms were "subjectively annoying" (+ numbers indicate sound advanced with respect to video). While not deciding on a recommended specification, the draft recommendation did state:

A tighter tolerance on the range of values in the 10 studio and production paths would be required to allow this [partitioning of tolerances]. The situation might look something like this:

+20 ms ... -40 ms Overall tolerance
+10 ms ... -30 ms Production/presentation
+10 ms ... -10 ms Distribution/transmission
+2 ms ... -2 ms per codec

The ITU draft also reports that ABC, NBC and CBS have specified relative timing limits of +16 to -33 ms. The EIA/TIA-250-C standard⁴ calls for a +25 to -40 ms specification end to end for transmission facilities.

WHAT IS TO BE DONE

Clearly, television facilities need to be designed with audio to video synchronization problem in mind. Since it is impractical to remove all of the offending video delay mechanisms, the only remaining solution is to ensure that the program audio signal receives the same delay as the associated video signal.

Since the delay path of video signals is constantly changing as shown in **Figure 1**, it is unreasonable to expect that a single compensating audio delay located at the transmitter or elsewhere will solve the problem. The most practical solution to the problem is to measure the video delay at every delaying device and correct the corresponding audio at that point with an audio synchronizer (variable audio delay) such as shown in **Figure 5**.

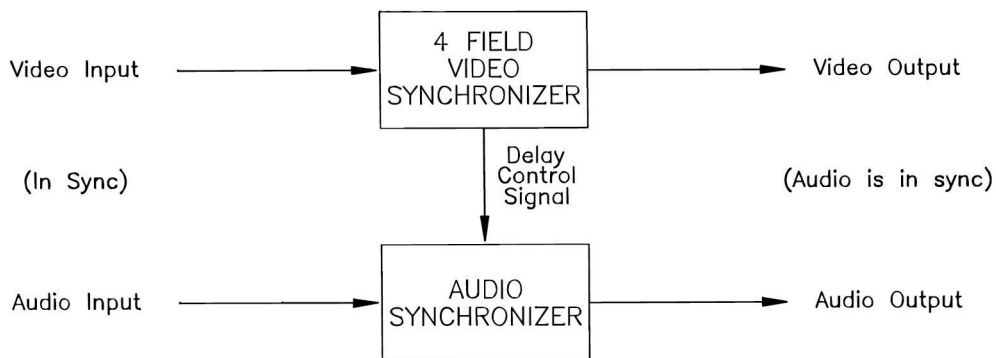


Fig. 5: Correcting The Audio Sync Error Automatically

Recently designed video synchronizers recognize this problem and have a delay control signal, known as a digital delay output (DDO) which provides a current delay value signal for use by a companion audio synchronizer. The audio synchronizer receives the DDO signal and automatically delays the audio signal by a corresponding amount thus keeping audio and video synchronized even though the total delay is changing.

VIDEO DELAY MEASUREMENT

Unfortunately, many other video delay devices such as noise reducers, color correctors, etc. do not have a DDO and until recently there has been no practical way to couple a companion audio delay to these devices.

Measurement of video delays is a fairly complex problem if you want to build a delay detector which can operate with a wide range of video equipment. By measuring the relative delay of input vertical sync and output vertical sync, it is possible to determine delays of up to one frame period. Many video devices, however, can have delays which range from zero to exactly one frame, or multiple frames. It is therefore difficult to determine if a delay is one frame, two frames or no frames when the input and output sync are near coincidence. Similarly, it is difficult to determine if the delay is $\frac{1}{4}$, $1\frac{1}{4}$, $2\frac{1}{4}$, etc. or to distinguish between any multiple frame delay ambiguities.

It would be possible to resolve multiple frame ambiguities by inserting a special code or signal in the video, for example: in the vertical blanking interval. By checking video to determine when the code goes in and later comes out of the video device, the delay can be measured. Unfortunately, this system will not operate with those many older devices which store only active video. For these older devices, the special code or signal would have to be inserted in active video which is generally unacceptable.

One manufacturer sells a device which can measure the relative audio to video delay by gating a test tone into the audio channel at the same time a specific video pattern is gated into the video path. At the end of the transmission path the relative delay of the two is measured. This system however cannot be used when the video or audio path is active, and since many video paths are constantly changing there is no guarantee that any measured value will remain correct after a few minutes, let alone throughout an entire program.

PIXEL INSTRUMENTS' APPROACH TO VIDEO DELAY MEASUREMENTS

LipTracker™ is highly sophisticated computer application program that utilizes cutting edge technology in machine vision and machine hearing. It is a non-invasive measurement tool for lip sync analysis which operates in the same way as a human observer by listening to the audio and looking at the video to measure the lip sync error. **LipTracker™** recognizes and compares selected sounds in the audio stream with the mouth shapes that create them in the video stream. Statistical analysis of these sounds and mouth shapes (called Mutual Events or MuEvs) produces a direct measurement of the lip sync error. This unique approach of analyzing real time video and audio does not require the insertion of cues, watermarks or codes into the program material. It connects to a system only as a node, and it is not within the main audio-video path. Therefore, **LipTracker™** can be used for in-service testing at any point in the transmission path.

DD2100 – This delay detector takes the unique approach of comparing frames of active video which are input to the device with frames of video which come out of the device. Key samples of input and output active video frames are taken and correlated in a high speed 20 bit DSP circuit. By correlating input and output frame samples, the relative delay is measured.

Figure 6 shows a conceptual block diagram of the Pixel Instruments DD2100 delay detector.

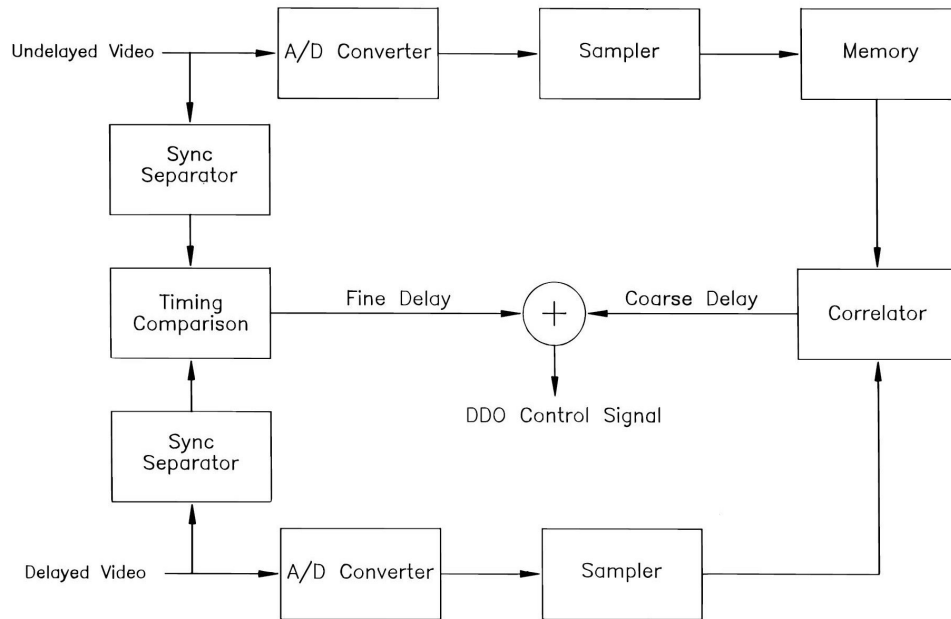


Fig. 6: Delay Detector Block Diagram

A given input frame is stored in memory and from that time onward all output frames are compared to the stored frame until an output frame which matches the stored frame is found. By counting the time which passes from the storing of the input frame until the matching output frame is detected, a very accurate delay measurement is obtained. In addition, the phase of input and output vertical sync is used to determine the fine delay. By adding the fine and coarse delays, a very accurate overall delay measurement is achieved to an accuracy of 102 μ s.

As shown in **Figure 7**, the DD2100 delay detector is very easy to add to an existing system, requiring only that input and output video be looped through the inputs.

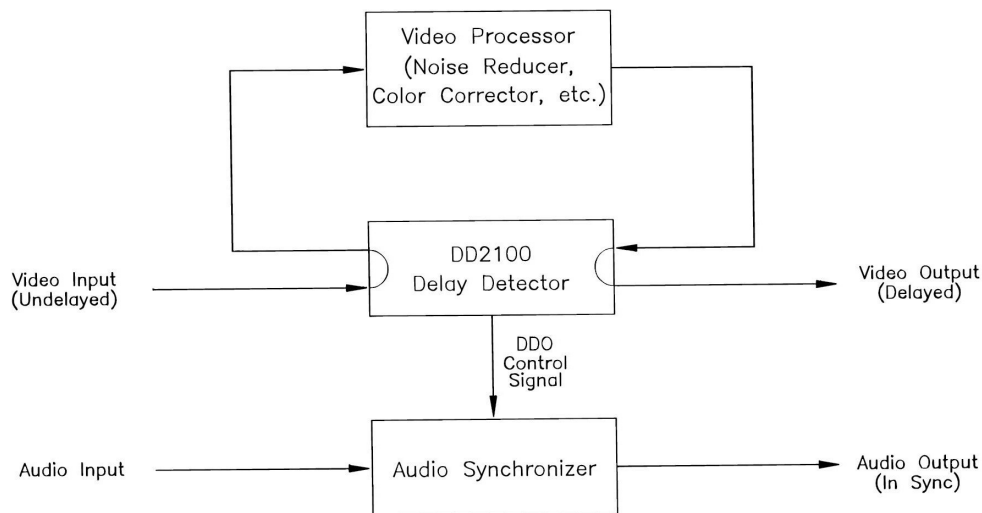


Fig. 7: Measuring Video Delays Directly

A front panel LCD display provides delay and update information as well as alarm messages in the event of problems. No modification to the video signal or the video device is required. The DD2100 detects delays up to 8.99 fields with 525 or 625 line signals and provides a DDO signal which is fed to the companion AD2100 or AD3100 audio synchronizer to make the appropriate audio corrections.

THE VARIABLE AUDIO DELAY

All of the currently available solutions to the audio to video synchronization problem use adjustable audio delays at some point in the system to delay the audio to match the delayed video. While the methods of determining the needed delay vary, the adjustable audio delay remains a key element. Several problems arise with these audio delays, centering mainly around the problem of making adjustments to the delay length which are imperceptible to the viewer. It should be remembered that all video synchronizers operate in a fashion which sooner or later either repeats or deletes a field or frame of video. This is known as a pointer crossing, and occurs when the video memory either fills or empties. At the pointer crossing, the video delay value takes a memory sized jump (usually a frame or two) which requires the audio delay to take on this new, greatly different value. Unfortunately, there are no audio frames which can be repeated or deleted, since the audio signal is continuous. Therefore, a continuously variable delay change is needed. Past variable audio delays operated to adjust the delay by jumping memory addresses. The corresponding pops and clicks were then masked either by limiting the jumps to periods of relative silence⁵ or by using two delay lines, jumping the length of the one not currently supplying audio, and fading from the delay line in use to the newly updated delay line⁶. These variable delay methods still generated a considerable number of undesirable artifacts for normal program audio.

The memory technology needed to obtain a continuously variable and a continuous audio signal is not trivial. As previously described, it is not possible to change the delay by simply jumping to a new memory read or write address as is done in the video synchronizer. To do so would cause an audio sample to be lost or repeated and a corresponding pop or click is created in the audio. These unwanted artifacts are impossible to hide.

The audio delay technology which is most successful in obtaining quality performance uses a memory which stores every audio sample which is taken by the A-D and reads every audio sample which is stored once and only once. In order to accomplish this task, the memory must have completely decoupled and asynchronous reading and writing functions, so that the reading rate can be faster or slower than the storing rate. By varying the reading rate with respect to the storing rate the delay time can be controlled, by causing the reading to catch up with the storing (to decrease delay) or to lag behind the storing (to increase the delay). This method works very well for making slowly changing delay adjustments.

Unfortunately if the reading is consistently faster or slower than the storing, a pitch change will occur in the audio, similar to when an audio tape is played off speed. To minimize the pitch change and make it unnoticeable to the viewer, it is necessary to limit the differential rate between memory writing and reading to keep the associated audio pitch change very small. Typically, limiting these changes to around .5% limits any pitch changes to amounts which the majority of viewers do not notice.

Unfortunately, with a small differential, the amount of time required to change large delay settings is correspondingly large, and the audio can be out of sync for several seconds, or even minutes after a pointer crossing.

It is possible to improve the rate of delay adjustment by changing the differential rate in response to the audio signal content, since larger ratios may be tolerated if no high frequency audio is present, or if there are periods of silence. Modulating the rate with the audio content does not provide a consistent significant improvement, and frequently is of no advantage if the program material has a musical background.

In order to both minimize perceptible pitch shifts during normal small delay changes, and to allow rapid delay change after pointer crossings, it is necessary that the audio synchronizer incorporate a pitch correction circuit. With pitch correction capability, it is possible to make rapid delay

changes with the pitch correction circuit removing corresponding audio pitch artifacts to a level where they go unnoticed by the viewer. One audio synchronizer that incorporates pitch correction is the Pixel Instruments AD3100.

The Pixel Instruments AD3100 provides a high performance pitch shifting capability in addition to the audio synchronizing functions. The AD 3100 can provide make a one minute delay change in 10 seconds with no perceptible pitch artifacts being generated for normal program audio with background music. In typical applications, a 4 frame pointer crossing change can be achieved in under 1/2 second.

In addition to achieving fast delay corrections for use in conjunction with pointer crossings of video synchronizers, the AD 3100 can also be used on line to compensate for instant delay changes in video production systems. For example the AD 3100 can be tied to a production switcher to automatically correct for the change in video delay when various video effects are switched in and out of the path. It can be used to compensate for color corrector and noise reducer delays, or for the constantly changing delay of MPEG compression and decompression channels. It can also be tied to the tally system to compensate for different camera to microphone distances, thus maintaining the psychological effect of receiving delayed sound at longer camera distances from the source.

SUMMARY

Audio to video synchronization has been proven to affect audience perception of the quality of programming and thus is a critical performance parameter for television facilities. System complexities make it impractical to totally prevent such problems and create a considerable number of associated technical challenges. The difficult problems of measuring video delays and correcting audio signals to keep proper synchronization are made easier by the use of video delay detectors and pitch correcting audio synchronizers.

- (1) Dr. Bryon Reeves & Dave Voelker, research report Effects of Audio-Video Asynchrony on viewer's Memory, Evaluation of Content and Detection Ability (1993)
- (2) International Telecommunication Union Document 10C/32-E, 11A/43-E, 11C/40E, CMTT-C/18-E 5 October 1993
- (3) International Telecommunication Union Document 11A/47-E, 13 October 1993
- (4) NAB Engineering Handbook, Television signal Transmission Standards (Washington, D.C.: National Association of Broadcasters), 621,
- (5) U.S. Patent 4,618,890
- (6) U.S. Patent 4,644,400