

SECOND GENERATION AUDIO TO VIDEO SYNCHRONIZATION

ABSTRACT

This paper presents a tutorial on second generation audio to video synchronization error correction for television systems. Some of the more common sources of errors are described; the problems which the errors create and solutions to the error causes are outlined.

VIEWER PERCEPTION PROBLEMS

The most obvious result of audio to video mismatch is visible "lip sync" errors. This problem certainly can and does happen in today's systems, with the frequency of occurrence becoming a significant concern to advertisers and station management. The mistiming of audio and video will always cause a subconscious degradation of the program's entertainment quality as perceived by the home viewer when the audio is advanced with respect to the video. 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. For example, we are used to hearing a racquet striking after we see the ball hit and hearing a commercial actor after we see them talking. In today's television systems however, it is the video which is delayed thus causing the sound to arrive at the viewer's ears before the corresponding visual sensation.

Viewing a television program with advanced audio is unnatural for the viewer, and is believed to cause subconscious stress. Psychological tests at Stanford University ⁽¹⁾ demonstrate that viewers who watch television commercials having advanced audio "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 where the mere existence of an audio problem was detected by very few average viewers.

In addition to the negative perception of the commercials in the presence of advanced audio, there was also evidence this caused the test subject's memory of the negative aspects of the commercial to be remembered longer than normal. The worst possible scenario takes place, the viewer perceives the advanced audio commercial in a bad light, and also remembers it longer than a commercial which is properly presented. Obviously, such problems can cause a great deal of concern for television advertisers.

CCD CAMERA GENERATED VISION DELAYS

Audio to video synchronization errors are becoming more troublesome as television technology progresses. The wide use of cameras having CCD sensors is aggravating this synchronization problem. All CCD sensors have an inherent visual delay mechanism. Depending on the sensor type, the visual delay may be several fields for newer camera types. In particular, the liberal use of digital frame store based image processing in newer cameras is creating previously unknown vision delays of several fields, with a four field delay not being uncommon.

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, tending to blur any motion in the image and making it difficult for

the viewer to distinguish precisely such events as lip movement. This blurring was normal with tube based cameras which were continuously exposed to light.

With a fast shutter speeds of CCDs, the image is integrated over a relatively short time, for example 100As for a 1/10,000 second exposure. In television systems, the frame rate (assuming a frame rate CCD exposure) is equivalent to the sampling rate in sampling theory. The exposure time is equivalent to aperture time. The ratio of exposure time to frame rate is the aperture ratio. It is known from sampling theory that the aperture ratio effect on frequency response, which in this case is the ability to accurately convey motion. For short exposures, the ability to convey motion to the viewer increases dramatically. The shorter exposure time gives brighter and less blurred moving edges which result in the viewer's improved ability to perceive motion. The CCD camera induced improved motion perception aggravates the corresponding increased image delay time, and makes any audio to image mismatch easier for the viewer to consciously or subconsciously detect.

VIDEO PROCESSING DELAYS

Video signals are often passed through a special effects generators, color correctors, noise reducers, frame synchronizers and a variety of other editing and image processing functions. As memory costs continue to decline, these devices increase in complexity, and many incorporate frame based processing functions which add delays which are switched in and out. Unlike the past where video delays slowly drifted due to differing sync generator phases, the video delay in many of today's systems take instant jumps of one or more frames, as editors and other operators select different processing modes. This situation is especially true of many current noise reduction and color correction products where extra frames of delay are added for each additional selected function. This instant change of delay length poses special challenges for the corresponding audio synchronizer which must keep up with these instant large changes in video delay.

SETTING PERFORMANCE STANDARDS

Several standards committees have set standards or guidelines for audio to video synchronization errors. The Radiocommunication Study Groups of The International

⁽²⁾ Telecommunication Union states

"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 (+1field, -2 fields) with the understanding that these tolerances are absolute, are not accumulative, and apply to the overall system".

⁽³⁾
The International Telecommunication Union in the Draft New Recommendation [DOC. 11/59] reports that errors of and greater than +20 and -40 ms are detectable and errors of +40 and -160 ms are "subjectively annoying" (+ numbers indicate sound advanced with respect to video). The draft recommendation states:

A tighter tolerance on the range of values in the 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

EIA/TIA-250-C standards call for a +25 to -40 ms specification end to end for transmission facilities. Given the inherent video delays in CCD cameras, very little additional delay can be tolerated in the rest of the system.

MEASURING THE VIDEO DELAY

Clearly, television facilities need to be designed with audio synchronization in mind. It is impractical to remove the offending video delays, so the only remaining solution is to ensure that the program audio receives the same delay as the associated video.

Part of the solution is to measure the video delay at each significant delaying device so that a corresponding audio delay can be inserted at that point. Several video synchronizer manufacturers have a digital delay output (DDO) which provide a current video delay value signal for use by a companion audio synchronizer. Additionally, video delay detectors are available for devices which do not provide DDO signals. The audio synchronizer receives the DDO signal and automatically delays the audio signal by a corresponding amount.

Delay detectors for video devices without DDOs operate by storing a given input video frame and comparing all output frames to the stored frame. By counting the number of frames which pass until the previously input frame is output, the video delay is obtained. These devices are easy to add to an existing system, requiring only that input and output video be looped through their inputs. They provide a DDO signal which may be utilized by a companion audio synchronizer to make appropriate corrections.

THE SECOND GENERATION AUDIO SYNCHRONIZER

It should be noted that all currently viable solutions to the audio to video synchronization problem utilize adjustable audio delays at some point in the system to delay the audio to match the delayed video. The adjustable audio delay remains a key element in system designs, and second generation synchronizers are challenged with the problem of making adjustments to the delay length which are imperceptible to the viewer.

As video delay values take jumps of one or more frames, the audio delay is required to take on the new, greatly different delay value without disrupting the audio. old style audio delays often operated by dropping or repeating audio samples, and relied on slowly changing video delays to operate properly. The occasional sample manipulation usually went unnoticed by the home viewer. When faced with instant delay jumps of a frame or more, these old devices required several seconds or even minutes to attain new delay values, with the sample manipulation creating noticeable distortion the whole time. Consequently, the audio would be both out of sync and noticeably degraded for the duration of the time to make the change. In systems where large jumps in delay are frequently made, this is unacceptable performance.

In order to overcome the problems inherent with sample manipulation, and more importantly to preserve the integrity of AES/EBU digital audio, it is necessary to have 1:1 correspondence between input and output samples in the audio synchronizer.

The audio delay memory must store every audio sample which is taken by the A-D, or received on the digital input, and read every stored audio sample once and only once. In order to accomplish this task, the memory must have completely decoupled and asynchronous reading and writing, 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). In digital systems, this must be performed with the obviously inconsistent requirement of maintaining the clock rate at the correct frequency.

Varying the reading rate with respect to the storing rate creates an annoying pitch change artifact, and requires re-clocking audio to maintain the proper output clock rate for digital audio.

In theory, to make the pitch change resulting from the memory read rate change indistinguishable to the viewer, it is necessary to limit the differential rate between memory storing and reading to keep the associated audio pitch change very small. Unfortunately, if the differential rate between memory storing and reading is small, the amount of time required to change delay settings is correspondingly large.

It would be possible to modulate the relative reading 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 however, and frequently is of no advantage for any program material having a musical background.

In order to minimize perceptible pitch shifts during delay changes, to facilitate rapid large delay changes and to maintain proper clock frequencies for correction of AES/EBU digital audio, it is necessary that the audio delay incorporate a pitch correction circuit. With pitch correction, it is possible to make rapid delay changes and maintain proper output clock frequency with the pitch correction circuit removing corresponding audio pitch artifacts so they are unnoticed by the viewer.

One commercial product which incorporates pitch correction is the AD-3100 manufactured by Pixel Instruments Corp. of Los Gatos, CA. This device has selectable analog and AES/EBU digital inputs and simultaneous analog and digital outputs. It receives a DDO signal from a video instrument and adjusts the reading rate of the internal memory to increase or decrease the delay while at the same time providing digital signal processing pitch correction to maintain both proper pitch and output sample rate. In this device, multiple frame delay changes can be made in a matter of milliseconds without introducing artifacts or losing proper synchronization.

(1) Dr. Byron 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 10OC/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,