November 2022

Precise Latency Calculation for Audio-Video Synchronization

Jason Chihhao Lee

Peggy Pei Chi Kao

Hung-Jen Yu

Hung Ren Liang

James Chen Chao Huang

See next page for additional authors

Inventor(s)
Jason Chihhao Lee, Peggy Pei Chi Kao, Hung-Jen Yu, Hung Ren Liang, James Chen Chao Huang, Jabez Hsu Chung, Hao-Wei Lee, Eric Huang, and Lin Chi Lin

Precise Latency Calculation for Audio-Video Synchronization

ABSTRACT

Synchronization between the audio and video tracks in recording equipment is usually achieved using an audio-first approach. In this approach, the timestamp of a target video frame is compared to the timestamp of the sound emitted during that frame, timestamps being counted in units of video frames. Videos have a relatively low sampling rate, e.g., a 60 frame-per-sec video has frames separated by 16.67 milliseconds. Thus, the measurement of audio-video asynchrony is imprecise.

This disclosure describes video-first techniques for audio-video synchronization. A target video frame is captured, and its timestamp is mapped to the audio track. The audio track has millisecond-level time resolution due to high audio-sampling rates. Using the audio track, the timestamp of the sound (pulse) emitted during the target video frame is determined to millisecond accuracy. Timestamps of the target video frame and of the audio pulse are differenced to obtain a high-precision estimate of audio-video asynchrony.

KEYWORDS

- Audio-video synchronization
- Audio-video asynchrony
- Audio-video synchronization test
- De-synchronization
- Time reference
- Audio latency
- Video latency

BACKGROUND

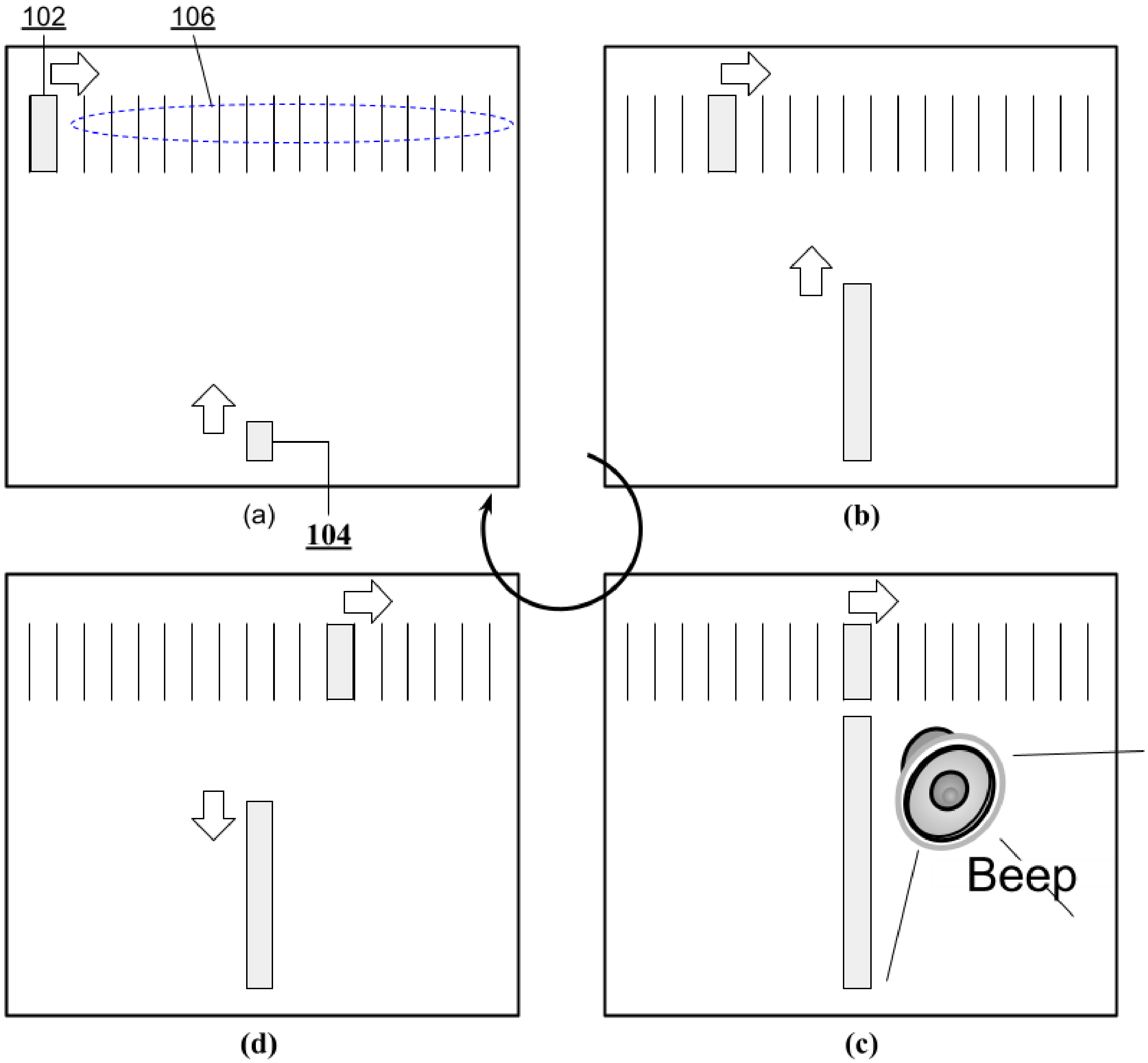

Fig. 1: A test movie used to achieve audio-video synchronization

Audio-video synchronization (AV-sync) is critical to user experience in applications such as streaming, gaming, video recording and playback, etc. AV-sync is usually tested using a standard movie with an audio track, an example of which is shown in Fig. 1. A first bar (102) moves along a horizontal axis. A second bar (104) grows or shrinks along a vertical axis. When the two bars meet (Fig. 1c), a beep sound is emitted. The meeting of the two bars can be compared to the clapping of hands. The movie loops endlessly. The time-ticks along the horizontal axis (106) are separated by a frame length. For example, in a 60 frame-per-second (fps) video, the time-ticks are 1000÷60=16.67 milliseconds apart.

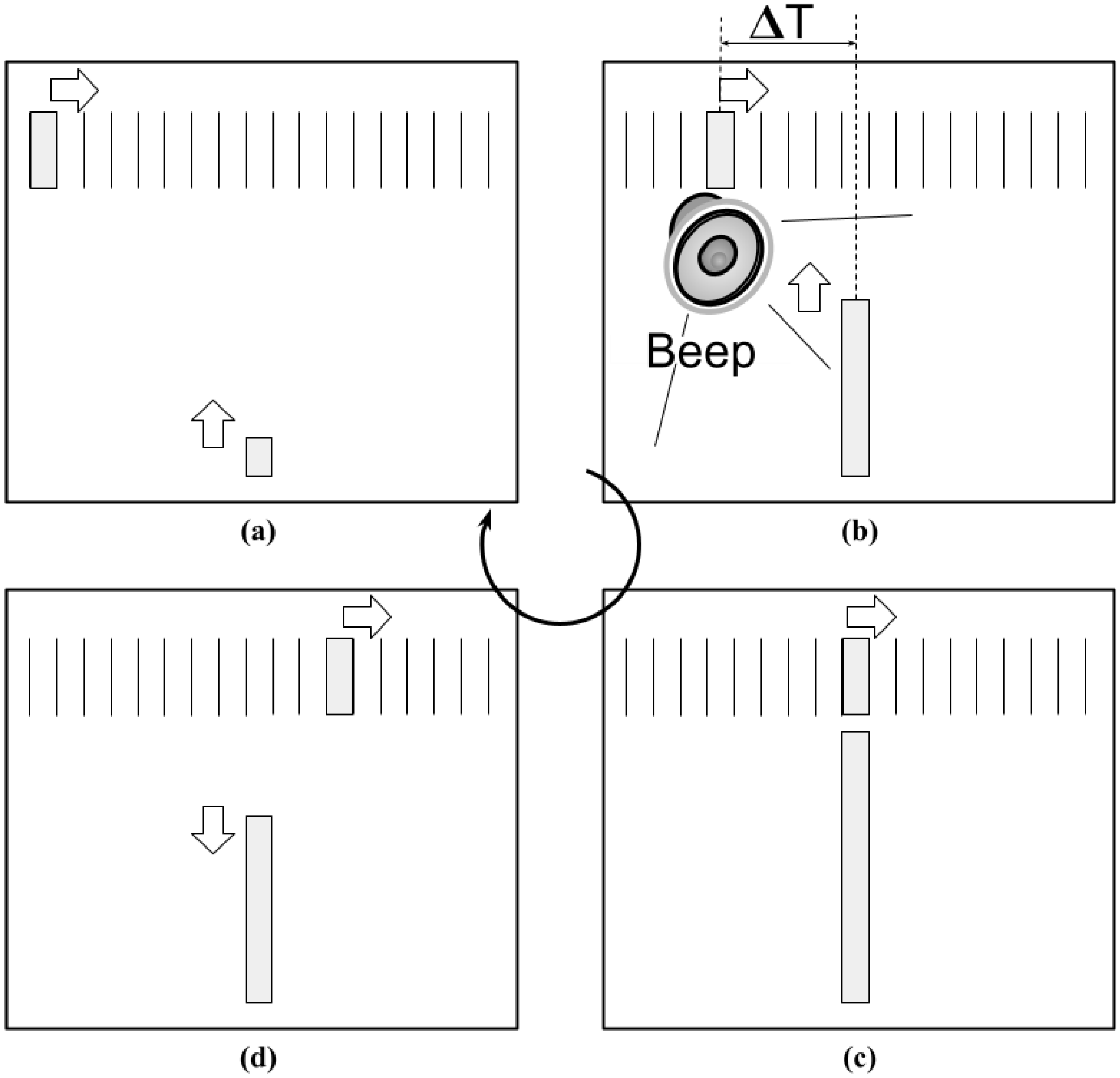

Fig. 2: An example of asynchrony between the audio and video feeds

To test for AV-sync, e.g., in a video recording and playback application, the movie of Fig. 1 is recorded using the recording equipment under test. When the recorded movie is played back, if the beep is emitted exactly when the bars meet, then the recording equipment has audio and video synchronized. If, on the other hand, the beep is heard on the playback before the bars meet (Fig. 2b) or after the bars meet, then the recording equipment has audio and video out of synchrony. The amount of asynchrony can be measured as the time difference **Δ**T between the onset of the beep and the point of intersection of the bars. An equivalent delay is introduced into the audio (or video) path of the recording equipment to compensate for the audio-video asynchrony. The measurement of the time difference **Δ**T typically uses the onset of the beep as zero-time - it is *audio-first*.

The accuracy in measuring **Δ**T is limited by the fps parameter of the test video, since the time base is the video track. In the above example, the test video has 60-fps, so that **Δ**T can be measured at an accuracy no better than 16.67 msec. The accuracy of the conventional AV-sync test is often insufficient and leads to issues such as lack of experimental repeatability and apparent unstable results across a battery of tests.

DESCRIPTION

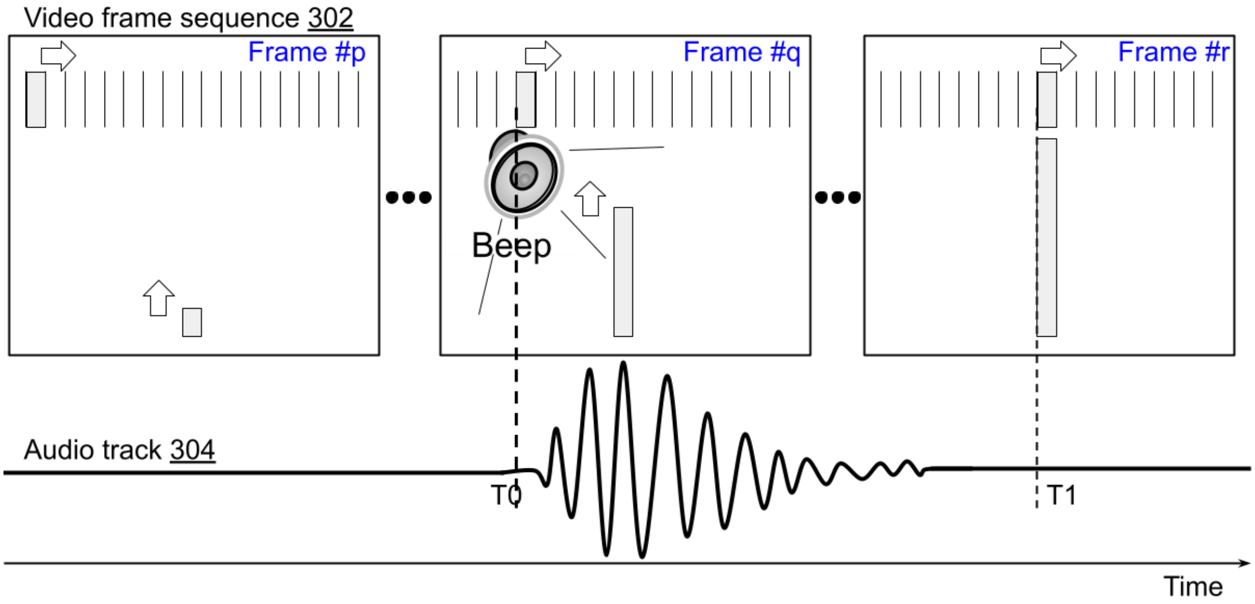

Fig. 3: Measuring audio-video asynchrony to high precision

In contrast to audio-first techniques of AV-sync, this disclosure describes AV-sync techniques that are *video-first*. As illustrated in Fig. 3, the test movie and audio track, as recorded by the recording equipment under test, are played back in a movie editor that can parallelly display the sequence of video frames (302) and the audio track (304). The recording equipment under test suffers from audio-video asynchrony, such that the beep sound occurs at frame #q, prior to and asynchronous with the meeting of the vertical and horizontal bars, which occurs at frame #r. The parallel audio track shows the waveform of the beep. Because the audio waveform is sampled at 64 kilo-samples/second (or at other common audio sampling rates such as 8000, 16000, 22050, 32000, 44100, 48000, 88200, 96000, etc. samples per second) , the accuracy with which the start of the audio waveform can be located is very high, e.g., within ±1/64,000 seconds, which is less than one millisecond. To measure the magnitude of the audio-video asynchrony, the following can be executed.

The frame at which the horizontal and vertical bars meet is determined (Frame #r). The corresponding time, T1, of the frame is noted on the audio track, which is of millisecond granularity. The movie is rolled back to the start of the beep. Due to the high audio-sampling rates, the start-time T0 of the beep can be determined with very high accuracy, e.g., ± 1 millisecond. The audio-video asynchrony is determined to be $\boldsymbol{\Delta}T=T1-T0$. This audio-video asynchrony is determined to an accuracy of 1 millisecond, which is much better than conventional audio-first techniques that are limited by the accuracy of the frame rate (16.67 milliseconds for a 60-fps video).

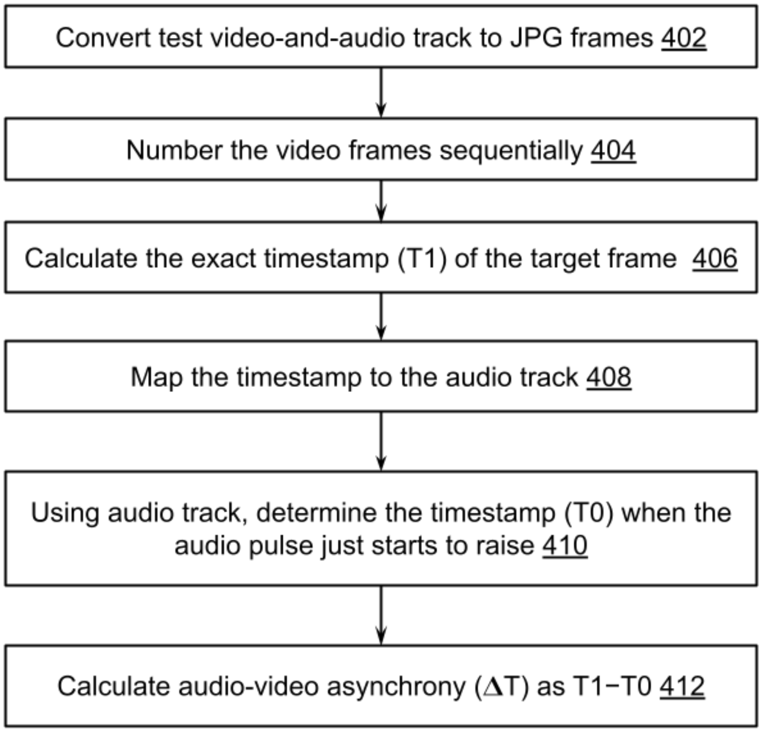
Convert test video-and-audio track to JPG frames 402
Number the video frames sequentially 404
Calculate the exact timestamp (T1) of the target frame 406
Map the timestamp to the audio track 408
Using audio track, determine the timestamp (T0) when the audio pulse just starts to raise 410
Calculate audio-video asynchrony (ΔT) as T1−T0 412

Fig. 4: Workflow for determining audio-video asynchrony to high precision

Fig. 4 illustrates an example workflow to determine audio-video asynchrony to high precision using a video-first approach. For every sequential search through a test video, the following is executed. The test video is converted to images (402), e.g., JPG frames, and the frames are sequentially numbered (404), e.g., as 1.jpg, 2.jpg, 3.jpg ... n.jpg. A target image - the image in which the horizontal and vertical bars meet, is determined based on the area-pixel RGB delta. The exact timestamp T1 of the target image is determined (406) in milliseconds, e.g., using the formula

$$T1 = \text{target image frame index} \times 1000 \div \text{fps}.$$

For example, if the target image file name is 100.jpg and the frames per second (fps) is 60, then the exact timestamp T1 is 100×1000 ÷ 60 = 1666 milliseconds.

The timestamp T1 is mapped to the audio track (408). Using the audio track, a timestamp T0 is determined (410) when the audio pulse just starts to rise, e.g., above a predefined threshold. As explained earlier, the timestamp T0 can be determined to very high, e.g., millisecond, accuracy. The audio-video asynchrony is calculated (412) as $\Delta T = T1 - T0$.

Alternatively, as mentioned earlier, audio-video asynchrony can be determined in an audio-first approach, e.g., by detecting and comparing the duration between timestamps of peak audio power; by mapping the timestamps to the video track; and by capturing image frame gaps that synchronize closely with the audio beep. Effectively, audio pulses are detected using a predefined threshold over a mono audio track and timestamps are extracted to the closest video frames.

A difference between the audio-first and the video-first approaches is that the audio-first approach operates by repeatedly swapping adjacent audio pulses whose power is above a predefined threshold and maps the timestamp of the pulses onto the video track, while the video-first approach repeatedly finds the target frame and maps its corresponding timestamp onto the audio track. Effectively, the video-first approach uses the audio track as a time base and calculates the latency between the timestamp of the target video frame (mapped to the audio track) and the start of the audio pulse. The audio-first approach does the reverse - it uses the timestamp of the target frame on the video track as the base and calculates the latency between that timestamp and that timestamp of the audio pulse mapped to the video track, which is of much coarser time-granularity.

While in the worst-case scenario, both audio-first and video-first approaches have the same time complexity, e.g., O(n), on average the video-first approach delivers quicker and more accurate results. The accuracy of the video-first approach can be traced to the higher density of samples on the audio track due to the higher audio sampling rate relative to the video frame rate.

CONCLUSION

This disclosure describes video-first techniques for audio-video synchronization. A target video frame is captured, and its timestamp is mapped to the audio track. The audio track has millisecond-level time resolution due to high audio-sampling rates. Using the audio track, the timestamp of the sound (pulse) emitted during the target video frame is determined to millisecond accuracy. Timestamps of the target video frame and of the audio pulse are differenced to obtain a high-precision estimate of audio-video asynchrony.

REFERENCES

[1] Lei, Yong, and Ke Yang. "Systems and methods for detecting audio-video synchronization using timestamps." U.S. Patent 10,158,927, issued December 18, 2018.

[2] Thenmozhi, A., and P. Kannan. "Performance analysis of audio and video synchronization using spreaded code delay measurement technique." *ICTACT Journal on Image & Video Processing* 9, no. 1 (2018).

[3] Russell, Michael E., and Arnold Sheynman. "Synchronizing remote audio with fixed video." U.S. Patent Publication No. US 2009/0091655 A1, published April 9, 2009.

[4] "AVSynchronizedLayer | Apple Developer Documentation" available online at https://developer.apple.com/documentation/avfoundation/avsynchronizedlayer accessed Sep. 26, 2022.

[5] "Audio Video Sync Test 60 FPS" available online at https://www.youtube.com/watch?v=TjAa0wOe5k4 accessed Sep. 26, 2022.