January 2023

Audio-video Synchronization with Arbitrary, Non-periodic Video Sources

Hao-Wei Lee

Jason Chihhao Lee

Hung-Jen Yu

Hung Ren Liang

James Chen Chao Huang

See next page for additional authors

Inventor(s)
Hao-Wei Lee, Jason Chihhao Lee, Hung-Jen Yu, Hung Ren Liang, James Chen Chao Huang, Jabez Hsu Chung, Eric Huang, and Yi-Chin Shih

Audio-video Synchronization with Arbitrary, Non-periodic Video Sources

ABSTRACT

The latency between audio and video streams of a device is usually measured using stock test videos. Although the use of stock test videos eases analysis, the test video differs materially from real-world videos, which tend to be far more diverse in content and encoding schemes, resulting in laborious experimental setup and inaccurate synchronization. This disclosure describes techniques to measure the latency between the audio and video streams of a given device using arbitrary, real-world, audio-visual footage (test video). Characteristic video and audio frames and their differences in timestamps (characteristic durations) are identified within the test video. The test video is played by the device-under-test while being recorded by a high-precision video camera. Characteristic durations of the recorded footage are determined. The differences in characteristic durations between the test and the recorded videos are statistically analyzed to determine the AV asynchrony of the device-under-test.

KEYWORDS

- Audio-video synchronization
- Audio-video asynchrony
- Synchronization test
- De-synchronization
- Time reference
- Audio latency
- Video latency
- Latency measurement
- Test video
- Scale-invariant feature transform (SIFT)
- Random sample consensus (RANSAC)

BACKGROUND

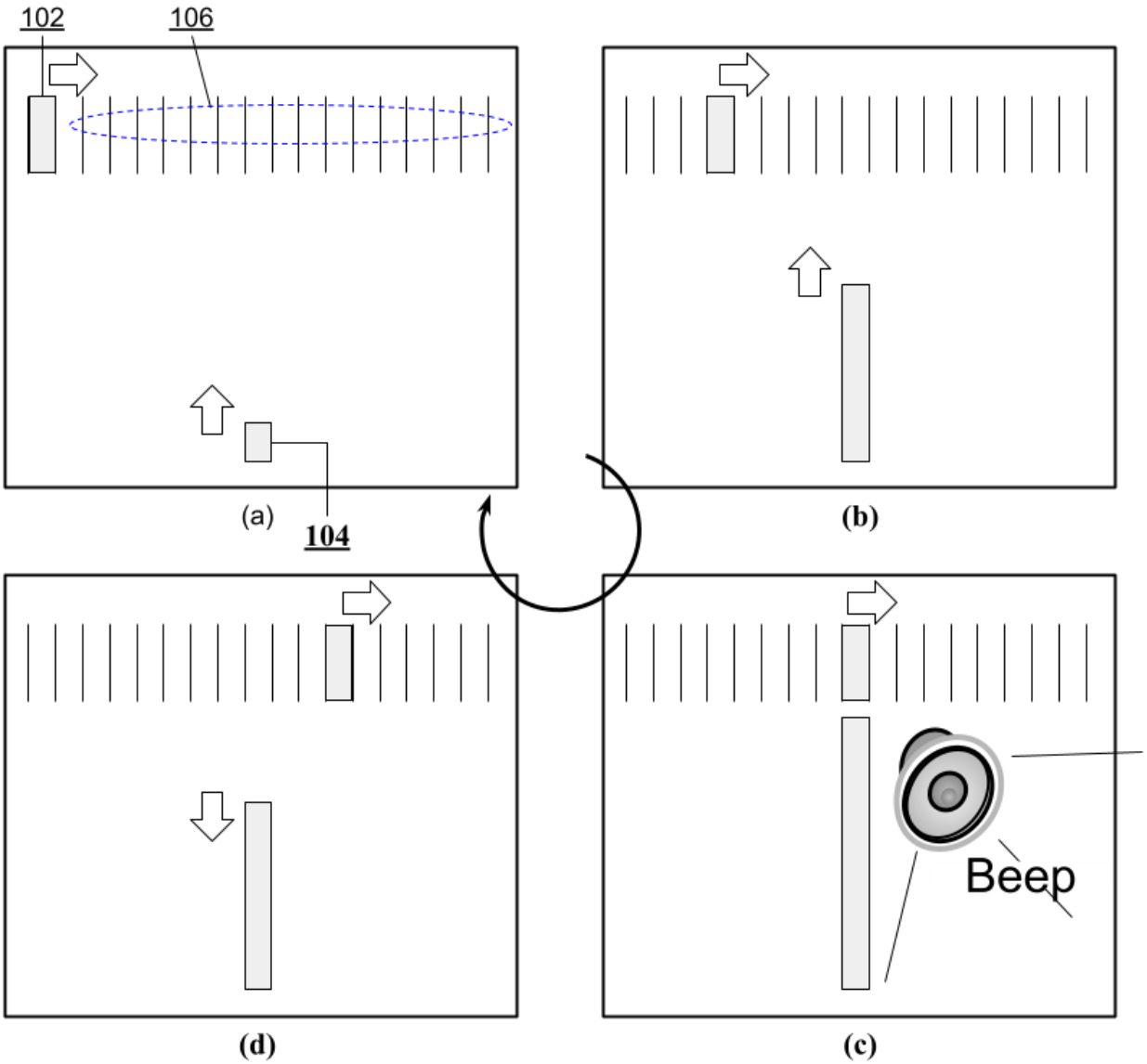

Fig. 1: A test movie used to achieve audio-video synchronization

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.

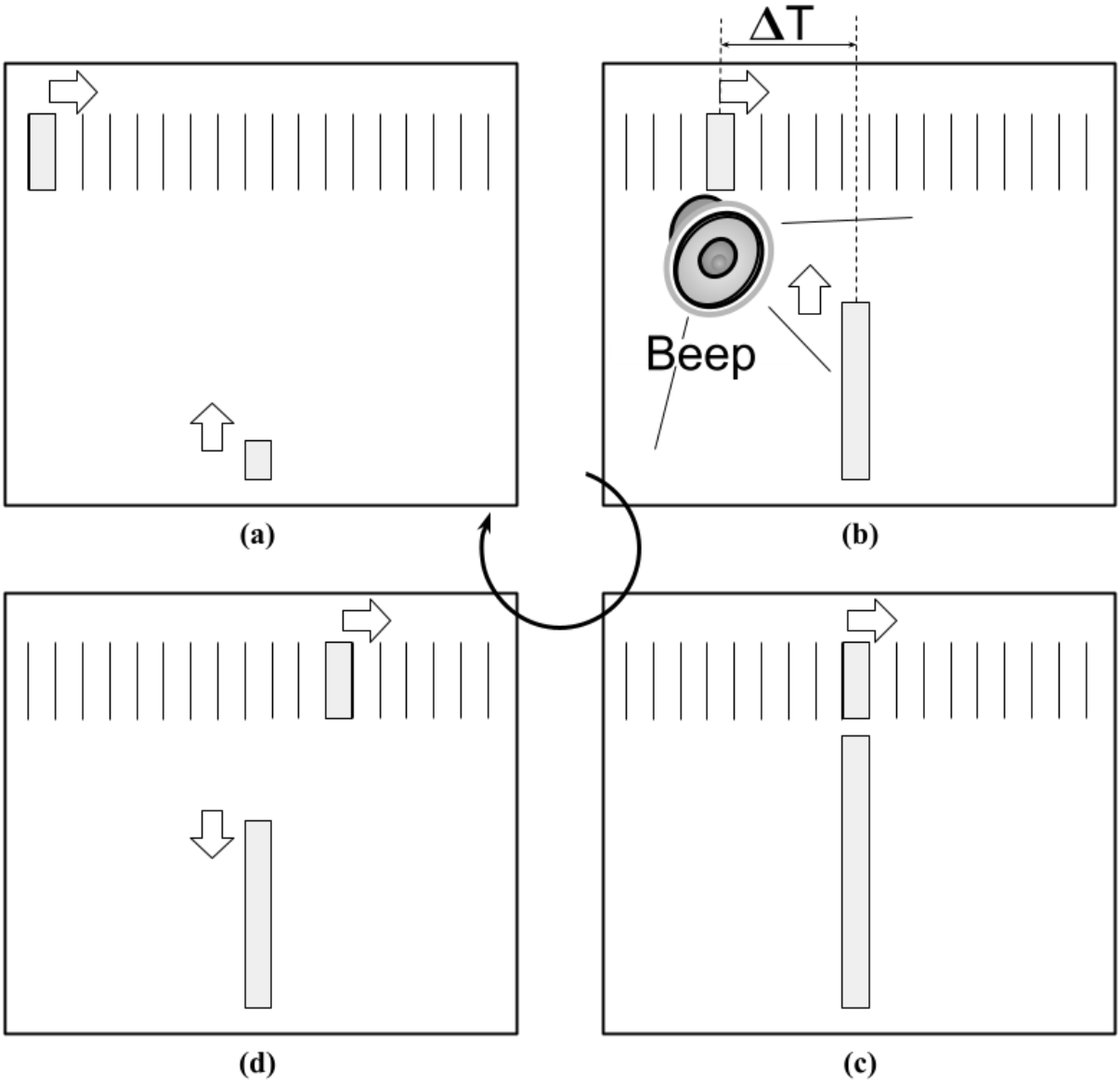

Fig. 2: An example of asynchrony between the audio and video feeds

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.

Although the use of stock test or source videos (as in Fig. 1) eases analysis, the test video differs materially from real-world videos, which tend to be far more diverse in content. Furthermore, there can be substantial diversity in encoding schemes, e.g., MP4, flv, mov, avi, mpeg, wmv, swf (for online video); mov, avi, mkv (HDTV); mpeg-2 (DVD/Blu-ray); WebM, HTML5 (websites); etc.

Moreover, the use of a predefined test video source assumes that AV asynchrony within the source is zero; indeed, the experimenter must ensure zero AV asynchrony prior to automating AV-sync measurement. The precise measurement of latency between audio and video streams has several limitations and the accuracy of the test is contingent on the assumptions holding true during test execution. AV-sync measurement today, which uses stock test videos, lacks robustness and can easily be broken by ordinary human experimenter errors, e.g., the use of a slightly different test video source, an unexpected encoding scheme, etc.

In an audio-first approach [1], AV asynchrony is measured as the time difference between audio pulse locations and their corresponding video frames (generally the frame that appears at the center of the ticker bar, as in Fig. 1c, also known as the zero-position frame). However, imperfections are not uncommon in test video files. In particular, the audio pulse in some test video files is found not to correspond to the zero-position frame. Such imperfection in the test video, known as *intrinsic latency*, optimally should not exist, but in practice does sometimes exist, with the experimenter often unaware or unsure of its existence.

Since light and sound travel at substantially different speeds, even a test video with zero intrinsic latency reaches an observer at a non-zero distance from the audio-visual source with some asynchrony between the video and audio tracks. In most practical instances, the distance between the screen and the observer is small enough that AV asynchrony arising out of the differential speeds of light and sound can be ignored due to near imperceptibility by the human brain.

An audio-visual signal $s(t)$, which is a function of time, comprises a video signal s_v and an audio signal s_a. The video signal, which is physically a time-varying image on the screen, is a function of the x and y coordinates on the screen as well as of time, whereas the audio signal, which is physically an air-pressure wave, is a function of time alone. Mathematically,

$$s(t) = \{s_v(x,y,t), s_a(t)\}.$$

Another way of stating the above is that the human visual response is a two-dimensional function of time while the auditory response is a one-dimensional function of time. Upon sampling the real-world, analog, audio-visual signal $s(t)$, the audio-visual signal can be stored or transmitted in the form of digital files or signals. Given a visual sampling frequency f_{sv} and an audio sampling frequency f_{sa}, the digital video signal is mathematically represented as

$$s[t_n] = \{s_v[x, y, t_n], s_a[t_n]\},$$

where t_n is the n^{th} sampling instant, e.g., $t_n = n/f_s$, with $f_s = \max(f_{sv}, f_{sa})$.

Modern audio-video codecs generate playback adaptively, e.g., when the codec finds one signal (e.g., audio signal) ahead of the other (e.g., video signal) or vice-versa, it slows down the faster signal or skips frames on the slower signal to align the audio and video streams in time. Therefore, a proper measurement of AV-sync latency is statistical in nature, e.g., numerous latency measurements are to be taken and statistically characterized.

DESCRIPTION

This disclosure describes techniques to measure the latency, or asynchrony, between the audio and the video streams of a given device-under-test using arbitrary, real-world, audio-visual footage. Pre-test, test, and post-test phases are defined. In the pre-test phase, also known as the feature-finding phase, the arbitrary audio-visual footage (hereafter referred to as the test video) is analyzed to locate characteristic video frames, e.g., frames computationally easy to recognize, and characteristic audio frames, e.g., audio clips of impulsive RMS power close to characteristic video frames.

An example of a characteristic video frame followed by a characteristic audio frame is the image of a gun firing followed by the audio clip of the gunshot. Another example of a characteristic video frame is the light of distant fireworks followed by their sound, or lightning followed by thunder. Yet another example is a new year countdown (‘5, 4, 3, 2, 1, 0’) followed by loud, festive cheering. It should be clear that a delay, e.g., a period of silence, may exist between a characteristic video frame and the sound generated by the event in the characteristic video frame. For example, due to the distance of the fireworks from the camera, the image and the sound of the fireworks can be separated by tens of milliseconds. Lightning and thunder can be separated by a silence of as much as a few seconds. A period of tens of milliseconds may exist between the last numeral of a new year countdown and the loud, festive cheering.

In any case, the frame number (time instant) of a characteristic video frame and the duration of the corresponding audio characteristic, referred to as characteristic duration, are noted. As noted earlier, the characteristic duration can include an intermediate zone of silence followed by a sharp jump in audio energy. In the test phase, the test video is played by the device-under-test while being recorded by a video camera with zero (or nearly so) intrinsic

latency. Footage captured by the zero-latency video camera is analyzed to locate the characteristic video and audio frames. In the post-test phase, the differences in characteristic durations between the original test video and the video recorded by the zero-latency camera are statistically analyzed to determine the AV asynchrony of the device-under-test. The pre-test, test, and post-test stages are described in greater detail below.

Pre-test (feature-finding) phase

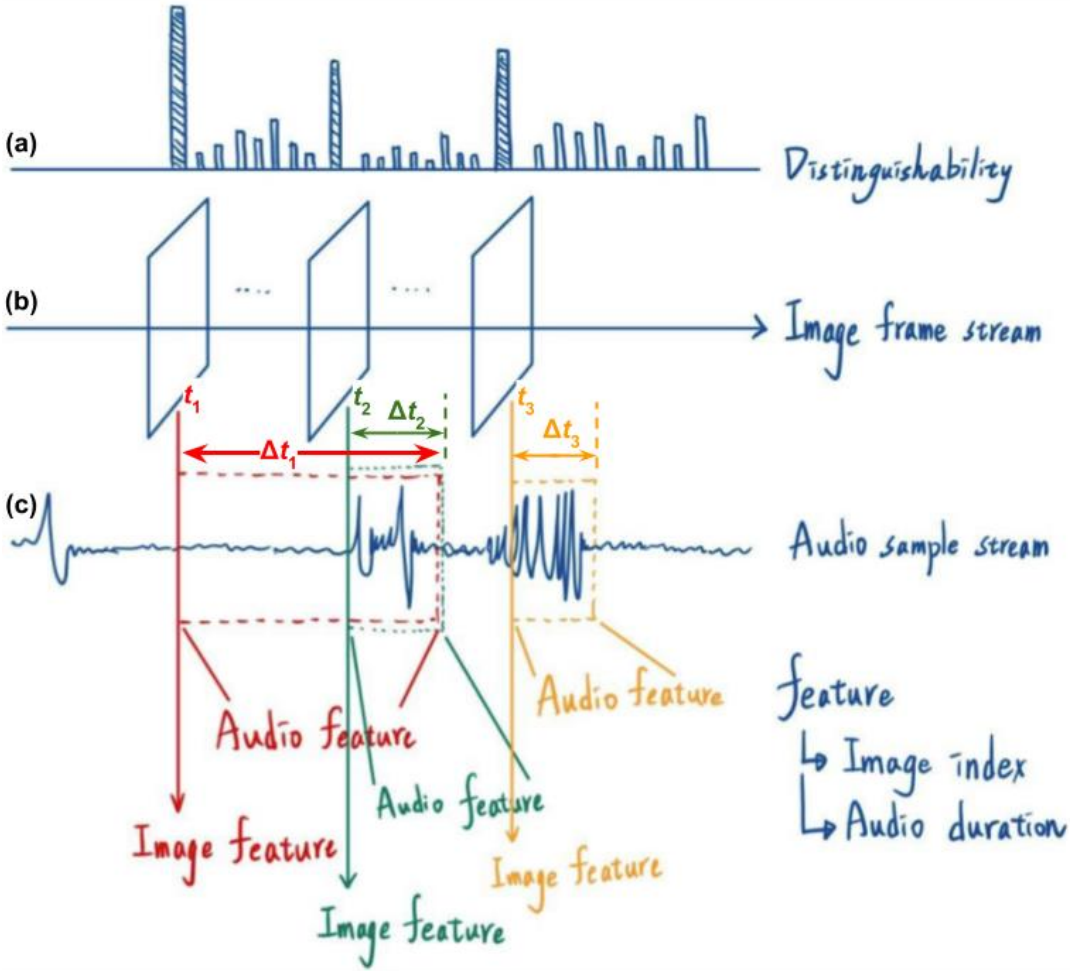

Fig. 3: Audio-video synchronization with an arbitrary non-periodic video source (pre-test phase). (a) Image-frame sequence by distinguishability; (b) Video (image-frame) stream; (c) Audio stream

In the pre-test phase, illustrated in Fig. 3, an arbitrary, non-periodic, audio-visual footage (the test video) is analyzed to locate characteristic video frames, e.g., frames computationally

easy to recognize, and characteristic audio frames, e.g., audio clips of impulsive RMS power close to characteristic video frames. An example of a characteristic video frame followed by a characteristic audio frame is the image of a gun firing followed by the audio clip of the gunshot, or lightning followed by thunder.

Fig. 3(a) illustrates a time sequence of the distinguishability of the images in the video stream. Images in the video can be measured for distinguishability using feature detectors, e.g., SIFT (scale-invariant feature transform). In Fig. 3(b), which is the video stream $s_v(x, y, t)$, a feature detector, e.g., SIFT, is used to identify video frames at times t_1 (red), t_2 (green), and t_3 (yellow) that have particularly distinguishable features. These are referred to as characteristic video frames and are easy to recognize computationally. As mentioned earlier, there can be an intermediate zone of silence between a characteristic video frame (e.g., lighting) and the sound generated (thunder) by the event depicted in the characteristic video frame. For example, such is the case for the red (t_1, $\mathbf{\Delta}t_1$) image-audio feature pair in Fig. 3, where the audio trace shows a period of silence just after the image feature before jumping sharply in energy.

The audio sequence ($s_a(t)$, Fig. 3c) is analyzed to find audio clips near the characteristic video frames of appropriate duration. These are referred to as characteristic audio durations. The audio clips can be selected by the lengths of their root mean square (RMS) power curves, e.g., a characteristic audio feature can be an impulse in the audio waveform. The RMS power of an audio clip is selected to be strong enough to avoid multiple matches on silent sections. For example, the characteristic video frame at t_1 has a corresponding characteristic audio duration of $\mathbf{\Delta}t_1$; the characteristic video frame at t_2 has a corresponding characteristic audio duration of $\mathbf{\Delta}t_2$; the characteristic video frame at t_3 has a corresponding characteristic audio duration of $\mathbf{\Delta}t_3$; etc.

Example

An audio-visual stream has a video frame rate of 25 frames per second and an audio sampling rate of 16 kHz. A characteristic video is found at the 56th video frame. The characteristic audio duration following the characteristic video frame has 480 samples. Here,

f_{sv} = 25;

f_{sa} = 16,000;

$t_1 = 56/f_{sv}$ = 2.24 seconds;

$\mathbf{\Delta}t_1 = 480/f_{sa}$ = 0.03 seconds;

so that the characteristic video frame appears at 2.24 seconds and the characteristic duration is the following 0.03 seconds, e.g., the interval [2.24, 2.27] seconds.

Test phase

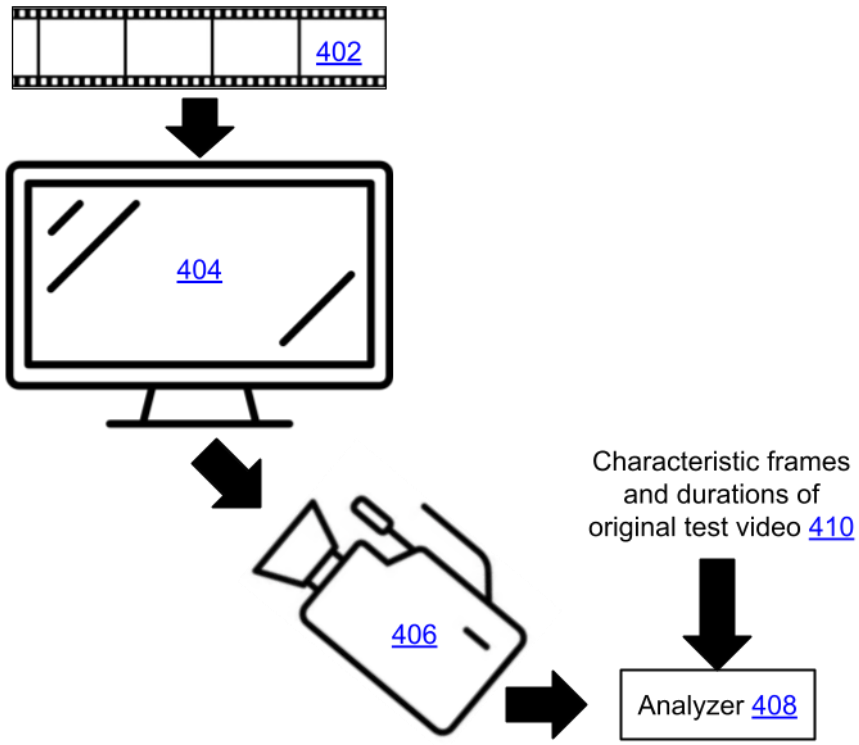

Fig. 4: Audio-video synchronization with an arbitrary non-periodic video source (test phase)

In the test phase, illustrated in Fig. 4, the test video (402), as obtained from the pre-test phase, e.g., with characteristic frames and durations, is fed to a device-under-test (404), e.g., a device with a screen capable of playing video. The AV asynchrony of the device-under-test is to be studied and characterized. A zero-latency (or nearly so) video camera (406) captures the video as played back by the device-under-test. The captured video is sent to an analyzer (408), that determines the characteristic frames and durations of the video as played back by the device-under-test from the characteristic frames and durations (410) of the original test video (402).

The analyzer can determine the characteristic frames in the video captured (recorded) by the zero-latency camera by matching the recorded video with the characteristic frames of the original test video using matching procedures such as those based on Euclidean distance. Matching the timestamps where the video frames of the captured video resemble the characteristic frames of the original test video, characteristic video timestamps (t_1', t_2', t_3', …) are determined. Similarly, by matching the audio stream of the recorded video with the characteristic audio clips of the original test video (e.g., using sliding window correlation), characteristic audio durations ($\mathbf{\Delta}t_1'$, $\mathbf{\Delta}t_2'$, $\mathbf{\Delta}t_3'$, …) of the recorded audio are determined. The differences ($\mathbf{\Delta}t_1-\mathbf{\Delta}t_1'$, $\mathbf{\Delta}t_2-\mathbf{\Delta}t_2'$, $\mathbf{\Delta}t_3-\mathbf{\Delta}t_3'$, …) in characteristic durations between the original test video and the recorded video are determined. If the device-under-test had no AV asynchrony, then the differences in characteristic durations are a zero vector.

Example

An audio-visual stream with a video frame rate of 25 frames per second and an audio sampling rate of 16 kHz has a characteristic video frame at the 56th video frame and a characteristic audio duration of 480 samples following the characteristic video frame. A zero-latency camera records a video of the audio-visual stream played by a device-under test and finds

a video frame-match at the 81st frame of the recorded video, and an audio frame-match *starts* at the 52,800th sample and lasts for 480 samples. It is required to determine the AV asynchrony of the device-under-test.

In this case, f_{sv} = 25 and f_{sa} = 16,000, such that the original test video has its characteristic video frame at $t_1 = 56/f_{sv} = 2.24$ seconds, and its characteristic audio duration is $\mathbf{\Delta}t_1 = 480/f_{sa} = 0.03$ seconds = 30 milliseconds. As for the video captured by the zero-latency camera, the characteristic video frame appears at $t_1' = 81/f_{sv} = 3.24$ seconds. The characteristic audio frame starts at the 52,800th audio sample, i.e., 3.3 seconds, and ends 480 samples = 30 milliseconds later. That is, the video as played by the DUT has a delay between its characteristic video and audio frame of 3.3−3.24 = 0.06 seconds = 60 milliseconds. The AV asynchrony of the device-under-test is thus determined to be 60 ms. Specifically, the audio stream of the device-under-test is 60 ms behind its video stream, or its video stream is 60 ms ahead of its audio stream.

Post-test phase

In the post-test phase, the characteristic durations of the video as played by the device-under-test and as recorded by the zero-latency camera are statistically analyzed to evaluate the AV asynchrony of the device under test. The results can be compared with devices known to have excellent AV synchrony (golden devices).

If the recording angle and position are fixed, it can be possible to determine characteristic video and audio frames of the test video without the use of image or audio matching procedures such as SIFT or RANSAC (random sample consensus). Characteristic frames can be directly located by finding the brightest video frames and by matching using a Euclidean-distance

criterion, since a constant affine projection between the source test-video and the recorded video can be expected.

Some advantages of the described techniques include:

- Applicable to virtually any video source, not merely stock test videos.
- No need for the experimenter to account for the intrinsic latency of a test video.
- Freedom from source criteria: Latency measurement is achieved regardless of types of devices (e.g., mobiles, tablets), operating systems, content, etc.
- **Diversity:** Testing can be accomplished under various sample scenarios, including online video, HDTV, DVD, Blu-ray, multimedia on websites, etc.
- **Precision:** An increased precision arises out of a reduction in engineering difficulties and experimental complexity. Tests can be run under time and space constraints, while runtime is optimized to boost overall test performance and precision.
- **Portability:** The techniques can run on fully automated test infrastructures.

In this manner, by selecting test videos from the universe of real-world videos, which have a vast diversity in content, platform, operating system, and encoding schemes, accurate AV-sync that works on real-world videos can be achieved. The AV-sync test and measurement procedure gains robustness and is more forgiving of human errors. The number of assumptions and requirements to be abided by the human experimenter is reduced for a cleaner, swifter, less effortful, more accurate, and robust experimental procedure.

CONCLUSION

This disclosure describes techniques to measure the latency between the audio and the video streams of a given device using arbitrary, real-world, audio-visual footage (test video). Characteristic video and audio frames and their differences in timestamps (characteristic

durations) are identified within the test video. The test video is played by the device-under-test while being recorded by a high-precision video camera. Characteristic durations of the recorded footage are determined. The differences in characteristic durations between the test and the recorded videos are statistically analyzed to determine the AV asynchrony of the device-under-test.

REFERENCES

[1] Lee, Jason Chihhao; Kao, Peggy Pei Chi; Yu, Hung-Jen; Liang, Hung Ren; Huang, James Chen Chao; Chung, Jabez Hsu; Lee, Hao-Wei; Huang, Eric; and Lin, Lin Chi, "Precise latency calculation for audio-video synchronization," Technical Disclosure Commons, (November 06, 2022) https://www.tdcommons.org/dpubs_series/5485

[2] A. Al-Nuaimi, Anas, Burak Cizmeci, Florian Schweiger, Roman Katz, Sinan Taifour, Eckehard Steinbach, and Michael Fahrmair. "ConCor+: Robust and confident video synchronization using consensus-based Cross-correlation," *2012 IEEE 14th International Workshop on Multimedia Signal Processing (MMSP)*, pp. 83-88. IEEE, 2012.

[3] "Scale-invariant feature transform," available online at https://en.m.wikipedia.org/wiki/Scale-invariant_feature_transform accessed Dec. 12, 2022.

[4] "Random sample consensus" https://en.m.wikipedia.org/wiki/Random_sample_consensus accessed Dec. 12, 2022.

[5] "Apple TV audio/video sync issues" available online at https://discussions.apple.com/thread/253040954 accessed Dec. 12, 2022.