Determining the Amount of Audio-Video Synchronization Errors Perceptible to the Average End-User

Audrey C. Younkin and Philip J. Corriveau

***Abstract*—The Media and Acoustics Perception Lab (MAPL) designed a study to determine the minimum amount of audio-visual synchronization (a/v sync) errors that can be detected by end-users. Lip synchronization is the most noticeable a/v sync error, and was used as the testing stimuli to determine the perceptual threshold of audio leading errors. The results of the experiment determined that the average audio leading threshold for a/v sync detection was 185.19 ms, with a standard deviation of 42.32 ms. This threshold determination of lip sync error (with audio leading) will be widely used for validation and verification infrastructures across the industry. By implementing an objective pass/fail value into software, the system or network under test is held against criteria which were derived from a scientific subjective test.**

***Index Terms*—Absolute detection threshold, audio/video synchronization, lip synchronization.**

I. Introduction

AUDIO-VIDEO synchronization (a/v sync) is one of the largest problems that video teleconferencing applications face. Different amounts of delay in the signal processing in both the audio and video channels might occur independently from each other, which require the signals from the two channels to be realigned.

Various standards have been proposed by different organizations or labs regarding the minimum amount of audio-video synchronization error that is acceptable for end-users' experience. According to the ATSC Implementation Subcommittee, "The sound program should never lead the video program by more than 15 milliseconds, and should never lag the video program by more than 45 milliseconds" [1]. This a/v sync standard is based on the notion that light travels faster than sound, which makes people more tolerant of audio lagging than audio leading the video programs. Reports on linear acoustics has indicated that experts (film editors) can detect a/v sync errors as short as $pm1/2$ film frame (about ± 20 ms) [2].

Lip-sync error has been observed and discussed intensively in the literature and is the most common type of a/v sync error [3]. Lip-sync error occurs when the sound is heard earlier than the movement of lips is seen. Therefore the present study focuses on the detection threshold of lip-sync errors by using clips with single person speaking. Since people are more aware of the audio leading than lagging case, the detection threshold under audio leading condition is assumed to be lower than the audio lagging case.

A staircase design is a method that is used to measure absolute threshold. Georg von Bekesy introduced the staircase method in 1960 in his study of auditory perception [4]. In the staircase method, stimulus intensity is progressively increased (ascending limits) until the participant reports seeing the stimulus or detecting a certain property of the stimulus. At this point, the intensity value is recorded and the stimulus intensity is then progressively reduced (descending limits), until the participant reports not seeing the stimulus or the specified property of the stimulus. The threshold is determined by the average of several of these reversal points.

The staircase method is a quick way of determining threshold; however, two kinds of errors can occur; the errors of habituation and the errors of anticipation. The errors of habituation occur when participants develop a habit of responding to a stimulus. For example, in ascending staircases, the participant may report seeing the stimulus property three steps past the threshold every time, thus giving a false threshold point. The error of anticipation occurs when participants report seeing the stimulus before the threshold, which usually happens in descending staircase. In the present study, two simultaneous staircases are randomly interleaved to minimize such errors.

The goal of this paper is to establish the detection threshold of lip-sync errors. Stimuli in the study include video clips of single person talking while the amount of synchronization error varies based on the participant's response and the staircase method. Even though the threshold may vary from person to person, an overall estimate of the mean threshold will provide valid information regarding to the amount of sync error that a person can normally detect, therefore providing an empirical bases for setting product standard with respect to sync errors.

II. Method of Introducing A-V Sync Errors

A. Stimuli Generation

The audio channel led the video channel by various amounts of time. The stimuli were created by shifting the audio channel relative to the video channel, generating a constant a/v sync error for the entire duration of the video clip.

Technically, the audio stream was cut at the beginning and the same amount of silence was added to the end of the audio

Manuscript received October 15, 2007; revised April 22, 2008. Published August 20, 2008 (projected).

The authors are with Intel, Hillsboro, OR 97124 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBC.2008.2002102

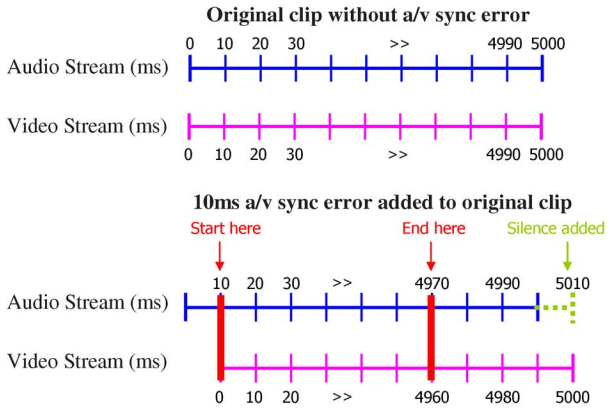

Fig. 1. Visualization on how an a/v sync error is created.

stream. Then after that we can cut the same amount of video and audio at the end so that there is no silence in the clip.

Using 10 ms of audio leading as an example, assume that one frame of video stream is 40 ms and the original video clip is 5 seconds long. The first 10 ms of audio was cut and the same amount of silence (10 ms) was added to the end of the audio stream. The modified video clip played 4960 ms instead of the original 5000 ms (i.e. 5 s). The ending point in audio stream was at 4970 ms of the original stream and the ending point of video was at 4960 ms of the original video stream [see Fig. 1].

When generating the stimuli with certain a/v sync error, the inherent a/v sync error due to the computer system and the playback software was considered. The reading from [5] showed that there was 10 ms video leading that was inherent to the computer system and video playback software [6].

This amount of inherent a/v sync error was considered when creating the video clips for the experiment and adjustment was made to ensure that the amount of a/v sync error associated with a video clip was what was shown in the video clips when played back by the system.

The original audio files were parsed from the original video clips by using [7] and then cut and padded in Matlab (Matlab R2006a, the MAthworks, Inc). The last few frames of video were cut in VirtualDub and the audio and video streams were then combined in VirtualDub. See Appendix A for the more details for creating the video clips.

The stimuli were generated by various amounts of audio leading time (from 0 to 400 ms with an interval of 10 ms). The amount of error created in the clips was confirmed by the AVsync tool developed by Royce Fernald at Intel Corporation (AVSync 0.1). The various amounts of error were presented to the participants.

The length of each video clip was held constant at 9.58 seconds. For all the video clips based on the same original video clip, the video stream was kept the same while the audio stream was shifted by various amounts in time.

B. Equipment

The CPU loading, disk traffic, network traffic, system memory, and driver's performance did not cause any artifacts in a real-time playback. The following systems were used in the evaluation:

- Desk-Top computer with CRT
- Graphic Chips: NVIDIA GeForce 6800 Ultra
- Screen size: 20 inch
- Resolution: 1024 × 768
- Refresh rate: 75 Hz
- CPU: 3.47 G Hz
- Memory: 1.00 GB
- Hard drive: 149 GB
- SynCheck:The SynCheck test files and the SynCheck hardware tool (SyncheckTM 2005, 2006) were used to estimate the inherent a/v sync error due to the computer system and playback software.
- CRC slider box:CRC slider box (CRC: Communication Research Centre in Canada) was used to record participants' response and let participants control the presentation of the video clips.

C. Software

The testing procedure and stimuli presentation were controlled by a program written in Matlab (R2006a, the MAthworks, Inc) using the Psychophysics Toolbox extensions [9], [10] and CRC slider Toolbox (Communications Research Centre, Canada).

[6]was used to play the video clips.

III. Experimental Design

A. Procedure

Before the experiments, participants were instructed about the procedure and the task required of them during the testing session. The task was to detect whether the audio and video were in sync or not. Participants responded by pressing a "Yes" or "No" button on the CRC slider box.

Four practice trials (with a/v sync error of 0 ms and 300 ms) were used for participants to become familiar with the setting and the task.

The first part of the main study used two video clips with various amounts of a/v sync errors to get an estimation of the threshold. The estimated threshold was used for setting the initial error amount in the second part of the main study. The second part of the main study used 22 video clips with various a/v sync error. A detection threshold was derived for each participant based on the staircase method.

The Staircase method was used for estimating an absolute detection threshold in the two parts of the main study. A classical 3-down-1-up staircase rule was followed. In other words, the error amount decreased by one step after three correct responses (i.e. "No" responses in this study) and increased one step after one wrong response (i.e. "Yes" responses). In the first part of the main study, one staircase with 20 ms step size was used. The starting value of the staircase was 200 ms based on pilot testing results. Two video clips were shown alternatively with certain amount of a/v sync error based on the staircase. The staircase terminated after two reversals. The reversal was defined as when the response changed from "Yes" to "No" or from

TABLE I

Effect	Level of Factor	Level of Factor	N	Threshold Mean	Threshold Std.Dev.	Threshold Std.Err	Threshold -95.00%	Threshold +95.00%
Total			45	185.19	42.32	6.31	172.48	197.90
Gender	M		29	179.41	36.40	6.76	165.57	193.26
Gender	F		16	195.66	50.97	12.74	168.50	222.81
Age	20-29		11	190.00	44.67	13.47	159.99	220.01
Age	30-39		16	182.44	35.82	8.96	163.35	201.53
Age	40-49		12	193.04	53.00	15.30	159.37	226.72
Age	50-59		6	168.00	34.05	13.90	132.27	203.73
Gender*Age	M	20-29	7	171.43	20.37	7.70	152.59	190.27
Gender*Age	M	30-39	9	177.00	40.60	13.53	145.79	208.21
Gender*Age	M	40-49	11	183.41	43.18	13.02	154.40	212.42
Gender*Age	M	50-59	2	196.25	39.24	27.75	-156.35	548.85
Gender*Age	F	20-29	4	222.50	60.07	30.04	126.91	318.09
Gender*Age	F	30-39	7	189.43	30.16	11.40	161.54	217.32
Gender*Age	F	40-49	1	299.00				
Gender*Age	F	50-59	4	153.88	24.91	12.46	114.23	193.52

"No" to "Yes", the error amount in the video that corresponded to the changed response was recorded as the reversal value. In the first part of the study, the average of the error amounts at the two reversals was used as an initial estimation of a/v sync error detection threshold. In the second part of the main study, two staircases were randomly interleaved. A step size of 10 ms was used. The starting values for the two staircases were three steps (30 ms) higher and lower than the estimated threshold based on the results from the first part of the main study. The step size in the second part of the study (i.e. 10 ms) was used to get a finer estimate of the detection threshold. Each staircase terminated after 10 reversals. The detection threshold was the average of the 20 reversal values from the two staircases.

B. Testing Conditions

The testing took place in a semi-anechoic chamber, located in Hillsboro, OR at Intel Corporation. One participant underwent the study at a time. Participants were provided with Plantronics Headsets to equalize listening playback among different users.

C. Participant Instructions

Participants were given a set of instructions describing the test set up. The participants were informed that the purpose of this study was to collect subjective perception of audio-video synchronization. Four examples were provided to familiarize the participants to the test and help disseminate any overt concerns with the experimental set up. Participants were encouraged to ask questions during this practice session.

D. Data Analysis

The a/v sync error amount at the 20 reversal points in the two staircases for each participant were collected. The average of the 20 data points was considered the detection threshold for that participant. The average of all the participants' thresholds was considered as a representative threshold for setting product standard.

IV. Results

A. Data Screening

The data were collected from August 30 to September 22, 2006. A total of 48 participants partook in the testing. Their age, gender, and responses for each trial were recorded. Detection thresholds were calculated based on the reversal results.

Based on the mean and standard deviation of the thresholds, all the data outside of 2 standard deviations from the mean were considered as outliers and were not included for further analysis. Three thresholds were removed accordingly. The following analysis includes a data set of 45 participants.

B. Descriptive Statistics

The overall average threshold was determined to be 185.19 ms and the standard deviation was 42.32 ms. The descriptive statistics were calculated to show the average differences in the thresholds for the categorical predictors of gender and age. The data of 16 females and 29 males were included in the study. The mean thresholds for female and male participants were 195.66 and 179.41 ms respectively. Age was binned into four categories: 20–29, 30–39, 40–49, and 50–59 with thresholds of 190.00, 183.43, 193.04, and 168.00 respectively (see Table I).

A univariate test of significance was run to determine if there was any statistical significant difference between the means with respect to the categorical predictors of gender, age, and the interaction between gender and age. No significant difference was revealed.

V. Discussion

Utilizing a staircase methodology to attain the absolute threshold for audio leading errors revealed an average boundary of a/v sync detection to be 185.19 ms. This threshold can be used to set landing zones for future Intel products.

The data collected had considerably higher threshold levels when compared to recent literature (ATSC), 185.19 ms and 45 ms respectively. However a direct comparison does not lend

to include differences in methodology, processing, or specific conditions upheld by the MAPL organization. The ATSC audio lead value of 45 ms was derived from end-to-end DTV audio-video production, distribution and broadcasting systems with emphasis on latency imposed by compression/decompression components.

This specific study was aimed at lip-sync detection in the context of a single speaker. Lip movement varied from each of the six speakers, which represents more realistic content than that of professional broadcasters. However, some participants commented that some speakers were harder to tell whether there was sync error than other speakers, due to lip variability. With this factor in mind, the threshold from the present study on one hand was limited by the video clips that were used; on the other hand it was not limited to the professional broadcasters.

With focus on the temporal relationship between audio and video signals, the actual physical position of the person (i.e. where the image of the person was shown) and the perceived position of the sound source based on the sound heard by the two ears, were not controlled. It would have been a more realistic experience for the participant if the two positions were kept at the same point and the threshold might have decreased since participants can estimate the time that sound travels to the ear by judging the distance between him/her and the sound source. However, in most viewing conditions end-users experience in daily life, the two positions are rarely at the same place and there is still a need for the audio and video signals to be synchronized.

VI. Conclusion

The overarching goal of the a/v sync related products is to enable engineers the ability to deliver audio and video in proper synchronization to the end-user. By determining the absolute threshold and setting landing zones, quantifiable levels can be achieved. The data collected here can be used for understanding the workloads on proof of concept platforms, knowing the platform capabilities for all enterprise platforms and determining roadmap usages: what a platform can do over time across generations of servers and client systems.

Appendix

All the created video streams have the same length without any silence in the audio channel. In other words, for the same original clip, only the audio channel is shifted forward by a certain amount, all the video channels are the same (i.e. the same length and same content for each clip).

Steps that are taken to generate the video clips with certain amount of a/v sync error.

1 Cut a video clip to a certain length (e.g. 10 s) using VirtualDub, the video clips used in the testing will be shorter (13 frames less) than the original length after the following steps.
2 Check the AV sync of the computer system by using the video clips from SynCheck and the hardware of SynCheck, to determine the amount of AV sync error inherent to the computer system and this amount of error is incorporated when creating the video clips with certain amount of AV sync errors. For example, the amount of error found from SynCheck is video leading 10 ms.
3 Save the .wav file from the original video clips in VirtualDub (Menu "File" then "Save WAV").
4 Cut a certain amount α (e.g. 23.36 ms) of the audio at the beginning of the audio file according to the AV sync error that is found in step 2. This amount corresponds to the time of one frame of video (e.g. 33.36 ms) minus the amount of video leading time that is found in step 2 (e.g. $33.36 - 10 = 23.36$ ms).
5 Cut another amount β (e.g. 30 ms) of the audio stream at the beginning of the audio stream. This amount corresponds to the real audio leading time you want to create.
6 Pad silence (with the total cut amount of the audio stream in step 4 and 5, $\alpha + \beta = 53.36$ ms) to the end of the audio file to keep the audio the same length as the original audio stream.
7 Cut the last 13 frames of audio (the frame length is based on the length of one frame of video clip) since we do not want any silence in the result video clips. The number of frames is determined by the length of the maximum error we want to create in the clips and also the length of the inherent a/v sync error due to the system. In this particular case, the maximum error amount in the audio leading is 400 ms and the systems' error is 10 ms, so we need to cut 13 frames to get rid of any silence. A Matlab program is used to do steps 4, 5, 6 and 7.
8 Cut the first frame and the last 12 frames of the original video clip in VirtualDub. Create a job file for one clip and then run the job for all the other clips.
9 Combine the cut video clips with the corresponding .wav files in VirtualDub. Save a job file for combining one clip with one amount of error. Make job files for all the clips and all the amounts of errors.

Acknowledgment

The authors would like to thank Rina Doherty for her effort in experimental design and statistical analysis. The authors would also like to thank Eric Salskov for the laboratory and device set up. Thanks also go out to Baoxia Liu for facilitating the subjective assessment.

References

[1] "ATSC Implementation Subcommittee Finding: Relative Timing of Sound and Vision for Broadcast Operations," (Doc. IS-191, 26 June, 2003), 2003.
[2] "Audio and Video Synchronization: Defining the problem and implementing solutions (rev 1)," Linear Acoustic Inc. [Online]. Available: www.LinearAcoustic.com.
[3] J. C. Cooper, A Short Tutorial on Lip Sync Errors, the Sources and Solutions Pixel Instruments Corp. [Online]. Available: http://www.pixelinstruments.tv/5ProfesArticles/Lip%20Sync%20Errors%20-%20A%20Short%20Tutorial.pdf
[4] J. G. Snodgrass, "Psychophysics," Experimental Sensory Psychology, 1975, pp. 17–67.
[5] "SynCheck (SynCheckTM,)," [Online]. Available: http://www.SynCheck.com
[6] "VLC media player 0.8.4a," [Online]. Available: http://www.videolan.org
[7] "VirtualDub (VirtualDub-MPEG2 1.6.10," [Online]. Available: http://www.virtualdub.org/, Copyright 1998–2005
[8] International Telecommunication Union, Methodology for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems ITU-R Recommendation BS.1116-1, 1997.

[9] D. H. Brainard, "The Psychophysics Toolbox," Spatial Vision 10, 1997, pp. 433–436.
[10] D. G. Pelli, "The VideoToolbox Software for Visual Psychophysics: Transforming numbers into movies," Spatial Vision 10, 1997, pp. 437–442.

Audrey C. Younkin is a perceptually focused human factors engineer with Intel's Channel Platform Group. She received a B.S. degree from the University of Portland. Her current research incorporates human perception of video and audio. Her email address is audrey.c.younkin at intel.com.

Philip J. Corriveau graduated with a Bachelors of Science Honors degree in Psychology from Carleton University in Ottawa, Ontario. He then worked for the Communications Research Centre (CRC), a Canadian Government Research and Development facility in Ottawa, doing applied subjective assessment work for HDTV standardization. In May 2001 Mr. Corriveau joined Intel Corporation in Hillsboro, Oregon. He now manages a multi-disciplined team investigating all aspects of user experience. His e-mail is philip.corriveau at intel.com.