Perceptually Motivated Guidelines for Voice Synchronization in Film

ELIZABETH J. CARTER*, Carnegie Mellon University
LAVANYA SHARAN*, Disney Research Pittsburgh
LAURA C. TRUTOIU, Carnegie Mellon University
IAIN MATTHEWS, Disney Research Pittsburgh
JESSICA K. HODGINS, Carnegie Mellon University and Disney Research Pittsburgh

We consume video content in a multitude of ways, including in movie theaters, on television, on DVDs and Blu-rays, online, on smart phones, and on portable media players. For quality control purposes, it is important to have a uniform viewing experience across these various platforms. In this work, we focus on voice synchronization, an aspect of video quality that is strongly affected by current post-production and transmission practices. We examined the synchronization of an actor's voice and lip movements in two distinct scenarios. First, we simulated the temporal mismatch between the audio and video tracks that can occur during dubbing or during broadcast. Next, we recreated the pitch changes that result from conversions between formats with different frame rates. We show, for the first time, that these audio visual mismatches affect viewer enjoyment. When temporal synchronization is noticeably absent, there is a decrease in the perceived performance quality and the perceived emotional intensity of a performance. For pitch changes, we find that higher pitch voices are not preferred, especially for male actors. Based on our findings, we advise that mismatched audio and video signals negatively affect viewer experience.

Categories and Subject Descriptors: I.3.3 [**Computer Graphics**]: Picture/Image Generation—*Display Algorithms*; I.3.6 [**Computer Graphics**]: Methodology and Techniques—*Standards*; H.5.5 [**Information Interfaces and Presentation**]: Sound and Music Computing—*Methodologies and techniques*

General Terms: Documentation, Languages

Additional Key Words and Phrases: multisensory perception and integration, human perception and performance, auditory: perceptual research, visual psychophysics

ACM Reference Format:

Carter, E.J., Sharan, L., Trutiou, L., Matthews, I., Hodgins, J.K. 2010. Perceptually Motivated Guidelines for Voice Synchronization in Film ACM Trans. Appl. Percept. 7, 4, Article 23 (July 2010), 12 pages.
DOI = 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

1. INTRODUCTION

The globalization of movie and television markets has meant that content is now produced and distributed rapidly across borders, in a variety of languages and media. Occasionally, this dissemination results in

*Equal contribution. Author's addresses: E. J. Carter, L. Trutiou, and J. K. Hodgins, Robotics Institute, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213; L. Sharan and I. Matthews, Disney Research, Pittsburgh, 4615 Forbes Ave, Ste. 420, Pittsburgh, PA 15213.

DOI 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

Fig. 1. Screen shots of actors performing emotional scripts. From left to right: disgust, fear, happiness and sadness.

unwanted side effects, such as the desynchronization of audio and video tracks in time (e.g., during dubbing) or changes in pitch (e.g., when naively converting from 24 frames per second in film to 25 in PAL). In both cases, an actor's voice may not end up matching his or her physical appearance. This mismatch is a violation of 'synchresis' [Chion 1994], the melding together of auditory and visual information. The effects of these two types of violations on viewer enjoyment and perceived performance quality have not yet been thoroughly studied.

Some of the timing disparities arise during the post-production process of dubbing. Actors come back to the studio after the film has been shot to rerecord their lines for the final soundtrack, a process called automatic dialogue replacement (ADR). The challenge of aligning the new audio recording and with the original video recordings is non-trivial. Some studios engage specialists for the alignment procedure. These individuals rely largely on industry guidelines that lack scientific underpinnings. Axioms used by professionals in the industry include, "...vocal sync issues usually warrant rejection when exceeding one frame out of sync" and "...generally reject any hard effects more than four frames out, in either direction" [Phillips 2010].

Once the film is released, other difficulties arise. Eventually, films become available for viewing on television, a transition that can cause audiovisual mismatches. Motion pictures are filmed at 24 frames per second (fps). However, televisions typically have display rates of 25 (PAL) or 30 (NTSC) fps, or a multiple of one of those numbers (50, 60, 100 or 120). In many cases, the video and corresponding audio are simply sped up when going from 24 fps to 25 fps systems, such that the first frame of second 2 becomes the last frame of second 1, and so on. The resulting four percent increase in speed results in higher pitched sounds unless corrected, which does not always occur [Spill 2010; Phillips 2010]. The reverse problem can happen if material recorded at 25 fps is converted to Blu-ray format, which can operate at 24 fps. Slowing the material results in lower pitched audio. These transitions in frame rate may also introduce an audiovisual mismatch because a person's voice at a different pitch might no longer match the expectations generated by his or her overall appearance, particularly if the viewer usually watches films by that actor in a movie theater that plays them at the original speed. The industry standard for pitch changes is "...pitch-up is generally found acceptable... but pitch-downs are less acceptable, particularly when female voices are lowered" [Phillips 2010].

In addition to pitch changes, timing mismatches can occur in the transition to television, even in films that were carefully matched before release. Over-the-air analogue broadcasts can be marred by timing misalignment when the two signals use separate carriers [Summerfield 1992]. The introduction of cable television reduced the frequency of this problem originally, but the recent proliferation of digital set-top cable boxes has occasionally reintroduced problems with lag [Mason and Salmon 2008]. Mismatches, particularly audio delays, can also occur when audio processors are used to connect a television with separate speakers for sur-

round sound signal processing [Mason and Salmon 2008]. Thus, both timing and pitch problems can occur in the transition to television.

When streaming movies and television shows over the internet, live or on demand, audiovisual mismatches often occur in many of the available media formats. In a study of various streaming technologies, viewers reported both timing misalignments and speed (and therefore, pitch) changes, even though the survey did not specifically request this information, indicating that it is highly salient [Huang et al. 2007]. These reports further support our hypothesis that both types of mismatch are irritating to viewers.

We sought to recharacterize viewers' tolerance for audio and video track mismatches using today's technology, as the introduction of high-definition recording and viewing potentially has rendered prior research obsolete. Additionally, we employed a new set of metrics: viewers' opinions of performance quality, emotional intensity, and preference. Thus, we could determine whether mismatches are merely irritating to the viewer or actually damaging to the quality of the message being conveyed.

We recorded high-definition video at 1080p29.97 with audio for four actors while they read emotional and neutral statements. In Experiment 1A, participants had to determine whether the audio and visual information for each clip they viewed was synchronous or asynchronous and which track (audio or video) came first. They detected differences when the audio led by 63.38 ms and when video led by 235.31 ms. This is consistent with previous research despite improvements in viewing technology. For Experiment 1B, participants rated performance quality, emotion, and intensity for clips that were in sync or out of sync to varying degrees. Clips that were noticeably out of sync had lower performance quality and intensity ratings than those that were perceived as synchronous. In Experiment 2, participants selected their preferred clip from pairs that included videos played at varying speeds. They reliably preferred clips that were played at the original recording speed to those that were played faster with a higher pitch. Additionally, they preferred the lower pitch clips for male voices. Altogether, these results suggest that the industry should ensure that audio and video signals match on both aspects, timing and pitch, for the best audience experience.

2. RELATED WORK

It is well known that when speech occurs in noisy environments, intelligibility increases when it is heard and seen simultaneously [Sumby and Pollack 1954]. This finding suggests that the two sensory streams, auditory and visual, are united into a single, multimodal percept. Moreover, having both types of information assists in the discrimination between consonants that are easily confused when presented in just one modality [Summerfield 1987]. Further support comes from the McGurk-MacDonald effect, which occurs when an audio recording of a syllable is played simultaneously with a video recording of another and results in a single percept of a third, intermediate syllable [McGurk and MacDonald 1976].

Previous explorations of the limitations of multimodal integration have suggested time windows in which auditory and visual streams are perceived as one. Faced with the task of moving the audio and video tracks relative to each other until asynchrony was detected, participants adjusted the sound to lead the video by up to 150 ms or to lag behind the video by up to 250 ms [Dixon and Spitz 1980]. Munhall and colleagues reported that the McGurk-MacDonald effect could occur for single syllables up to a difference of 180 ms when the audio track lagged the video track [Munhall et al. 1996]. A more recent report found this effect up to 170 ms with an audio lag, but only up to 30 ms when the audio track preceded the video track [van Wassenhove et al. 2007].

Sakamoto and colleagues expanded the audio track of Japanese recordings such that the onset of the audio and video clips were the same, but the audio was lengthened by up to 400 ms [Sakamoto et al. 2007]. For words, length increases up to 400 ms were tolerated; however, this alteration was only acceptable for full sentences when the increase was at or below 200 ms. A similar threshold for sentences was found for asynchronous on- and offsets. In a forced-choice paradigm, participants saw two videos of actors reading sentences and selected the one in which the audio and video tracks were mismatched [Grant et al. 2003]. The

threshold for detection of asynchrony was 45 ms audio lead and 200 ms audio delay for the four participants. Similar limits of 40 ms of audio lead or 160-200 ms of audio lag were found to increase intelligibility over audio or video streams played alone [Grant and Greenberg 2001].

The prior research has not accounted for the changes in technology that have resulted in different viewing experiences for audiences. For example, many viewers now receive programming in high definition and watch it on LCD screens. Preliminary research has begun to explore the differences between old and new viewing arrangements, but no results have been reported yet [Mason and Salmon 2008]. Moreover, no one has looked at the effects of asynchrony on the audience's enjoyment of films, including their ability to properly interpret emotional performances and gauge acting quality. These factors are important because asynchrony can have unexpected effects. For example, Tinwell and Grimshaw found that the perceived eeriness of a virtual character was correlated with the lack of synchronization between the sound and lip movements [Tinwell and Grimshaw 2009]. In addition, mismatches between the visual and auditory displays of emotion can result in an 'emotional McGurk' effect by creating the perception of a third, different emotion [Abelin 2007]. These findings emphasize the importance of assessing the subjective effects of timing mismatches as well as the objective thresholds for perception.

In addition to audiovisual mismatches based on timing, problems can occur with mismatches between voice and overall appearance for actors. Viewers can match voices with photographs of speakers [Lass and Harvey 1976]. Mismatches between voice and appearance might be problematic when the pitch of the voice is changed, such as when a film is shown at a faster frame rate on television. Typically, editors follow a rule of thumb by which they are willing to speed a recording up from 24 to 25 fps, but they would be unwilling to slow a recording from 25 to 24 fps. This tendency is based on a belief, mentioned earlier, that higher-pitched voices are more acceptable to audiences than lower-pitched voices [Phillips 2010]; however, we know of no published data to support this view.

3. EXPERIMENTAL METHODS

A database of film recordings was created and used for our experiments. A subset of this database was selected for each study as described below.

3.1 General Methods

We recruited participants using fliers posted around the university campus. Participants gave informed consent for each experiment and were monetarily compensated for their time. All experimental procedures took less than 60 minutes to complete and were approved by the Institutional Review Board. Unless otherwise noted, a new set of naive participants was recruited for each experiment.

Experiments took place in a quiet room on campus and participants were tested individually. Video clips were displayed on 30-inch LCD display (Apple Cinema HD, 60 fps). In order to direct attention to the faces of our actors, we inserted vertical mattes in the video clips while preserving the original aspect ratio and resolution. Participants were seated approximately 0.6 m from the display (33.4 x 25.4 degrees visual angle). They were fitted with headphones (Sennheiser EH 350) and allowed to adjust the volume until they reported that they could clearly hear the stimuli. All trials were self-initiated, so participants pressed a key to play each clip. Before every experiment, there was a brief practice session to familiarize participants with the procedures.

Experiments were written in MATLAB using the Psychophysics Toolbox (PTB-3) [Brainard 1997; Pelli 1997]. Clips were encoded at source with the XDCAM EX 1080p29.97 codec and system time stamps were used to verify the display frame rate of 30 fps. All clips were edited using Final Cut Pro (Apple, Inc.).

3.2 Creation of the Stimuli

Four actors were recruited from the local community to record the stimuli for this study. Each actor was seated in front of a green background (see Figure 1) and recorded with a Sony PMW-EX3 camera and an Audio-Technica AT8537 lavalier microphone. They were asked to read sentences aloud from a teleprompter placed in front of the camera. These sentences were drawn from two sources: i) a set of original, short, emotional vignettes and ii) a phonetically balanced corpus of emotionally neutral sentences [Garofolo et al. 1993; Theobald et al. 2003]. For the emotional segments, the actors were shown scene notes about the vignettes before they were asked to perform their lines. For example, we presented this background to the actor:

> Disgust: Jane stumbles out of her friend's kitchen, still reeling from having opened a month-old tupperware container of spoiled food.

before he or she read aloud,

> Ohhhhh! The SMELL! It's EVERYWHERE!

We recorded 60 emotional vignettes, each one multiple sentences in length. These vignettes covered a range of emotional scenarios - a distraught parent searching for a child (fear), a job applicant getting an offer (happiness) or unexpected guests at the door (surprise). From within each long vignette, statements ranging from 3 to 35 words in length and containing full sentences were selected.

For the emotionally neutral stimuli, actors were asked to keep a neutral expression while reading aloud. Each actor recorded approximately 200 sentences from the corpus, which ranged from 4 to 18 words in length. The sentences were selected to maximize the number of unique visemic contexts (using a standard phoneme-to-viseme clustering) given the fixed (200) number of sentences [Garofolo et al. 1993; Theobald et al. 2003]. These recordings of emotionally neutral sentences serve as a control for the emotional vignettes in the experiments that follow.

3.3 Selection of the Stimuli

We performed a preliminary study in order to determine which of our 60 emotional stimuli were suited for our purposes, based on the accuracy of emotion identification by viewers and similarity to the sentences available from our neutral speech recordings.

3.3.1 *Selection of Emotional Sentences.* We selected clips from 33 emotional vignettes that were approximately the same length in syllables as the sentences from the neutral speech corpus and contained one of the six common emotions (anger, fear, disgust, happiness, sadness, and surprise) [Ekman et al. 1972]. Eight individuals (5 females and 3 males, age 26-34 years, average age 28.73) who were fluent in English and possessed normal hearing and vision participated in this experiment. The setup was identical to that described in Section 3.1.

The participants viewed 33 emotional clips for all four actors for a total of 132 clips. After each clip ended, they saw the following question screens, one at a time:

> 1. Select the EMOTION that the actor portrayed. (Press 1 = Anger, 2 = Fear, 3 = Disgust, 4 = Happiness, 5 = Sadness, 6 = Surprise, 7 = Don't know)
> 2. Select the INTENSITY of the emotion that was portrayed. (Press 1 = Neutral, 2 = Not very intense, 3 = Medium intensity, 4 = Quite intense, 5 = Very intense)
> 3. Rate the PORTRAYAL of the emotion (i.e., acting). (Press 1 = Very bad, 2 = Somewhat bad, 3 = Mediocre, 4 = Somewhat good, 5 = Very good).

Participants were allowed as much time as needed to answer all questions, but they only were allowed to view each clip once. The presentation order of clips was counterbalanced between participants.

Based on the results of this experiment, we selected 19 clips for which participants were able to identify the emotion correctly ($> 90\%$ correct based on the emotion labels in the original scripts). Of these 19 clips, there were 4 that were angry, 3 disgusted, 2 fearful, 5 happy, and 5 sad. The intensity and portrayal ratings for the clips (averaged across participants) ranged from 2.65 to 4.35 and 3.15 to 3.9 respectively.

3.3.2 *Matching Emotional and Neutral Sentences.* We analyzed the 19 emotional clips that fit our previous criteria and all 200 of the recorded neutral sentences. From these collections, we created two sets of clips that matched on the following characteristics: number of words, number of syllables, number of bilabial consonants (i.e., b, m, p), number of labiodentals (i.e., f, v), number of interdentals (i.e., th), and number of alveolar consonants (i.e., t, d, s, z, n, l, sh) (all non-significant; $p > 0.05$). We chose to match the two stimulus categories for these types of consonants because they are formed with the lips, making it relatively easy to tell if the audio and visual timing is off. Thus, any differences that might be found in the results for the emotional and neutral sentences could be attributed to the presence or absence of emotion rather than physical characteristics of the sentences themselves.

We performed t-tests to pick the sentences in each set. We were able to match 13 emotional segments out of the 19 (3 angry, 3 disgusted, 2 fearful, 3 happy, and 2 sad) to 13 neutral speech sentences out of the 200 in the corpus without significant differences in the above-mentioned characteristics ($p > 0.05$). This final selection of 26 clips for each actor ($26 \times 4 = 104$ clips) was used in the experiments that follow.

4. EXPERIMENT 1: THE EFFECTS OF AUDIOVISUAL ASYNCHRONY ON THE VIEWING EXPERIENCE

In this study, we determined whether thresholds for viewers using updated technologies were the same as those described previously and examined the effects of presenting clips that had mismatches above and below threshold on performance and intensity ratings.

4.1 Experiment 1A: Thresholds for the Detection of Asynchrony

We used a method of constant stimuli design to measure thresholds for detecting audio visual asynchrony. Four interleaved staircases, two for the audio lag threshold and two for audio lead (half descending and half ascending with a step size of one frame offset) were run.

4.1.1 *Experimental methods.* Of the 26 total clips (13 emotional and 13 neutral), we selected six (three emotional and three neutral) to measure thresholds for detecting audio visual synchrony. We edited these six clips for every actor such that the audio track ranged from being -7 to +13 frames out of sync with the video track to produce 6 clips $\times$ 4 actors $\times$ 21 offsets = 504 clips. The ends of the audio track were trimmed to match those of the video track to avoid black frames in the stimuli. In all cases, the spoken words fell completely within the final trimmed clips.

On each trial, a participant was presented one of these 504 clips and asked:

> Tell us about the SYNCHRONY of the audio and video tracks. Press 1 = In sync, 2 = Out of sync, audio first, 3 = Out of sync, video first.

For each temporal offset, the choice of video clip was randomized across actors and scripts. Thresholds were selected to be the temporal offsets in a staircase when participants changed their responses from in sync to out of sync (or vice versa).

4.1.2 *Participants.* Eight participants (3 females and 5 males, ages 18-30 years, average age 22.63), all of whom were fluent in English and possessed normal vision and hearing, completed this experiment. Each participant completed 10 sets of four interleaved staircases.

4.1.3 *Results.* The left graph in Figure 2 shows the distribution of measured thresholds for all participants and all sets. Our findings (mean audio lag threshold = 235.31 ms, mean audio lead threshold = -63.38 ms)

are within the range of variation reported in the audio visual speech literature. Even though our setup (HD quality videos and display) is different from previous studies, the perceptual thresholds we find are very similar.

Next, we analyzed the accuracy of our participants at identifying the synchronization (zero offset vs. non-zero offset) and direction (positive offset vs. negative offset). Participants were reliably better at distinguishing synchronization (accuracy = 45.36%) than distinguishing both synchronization and direction (accuracy = 32.87%; paired, two-tailed t-test, $t(7) = 4.36$, $p = 0.003$). The right graph in Figure 2 presents this result in the form of distributions of out-of-sync responses. The considerable overlap between 'audio lead' and 'audio lag' responses indicates that determining the direction of offset in an asynchronous clip is hard, confirming previous research.

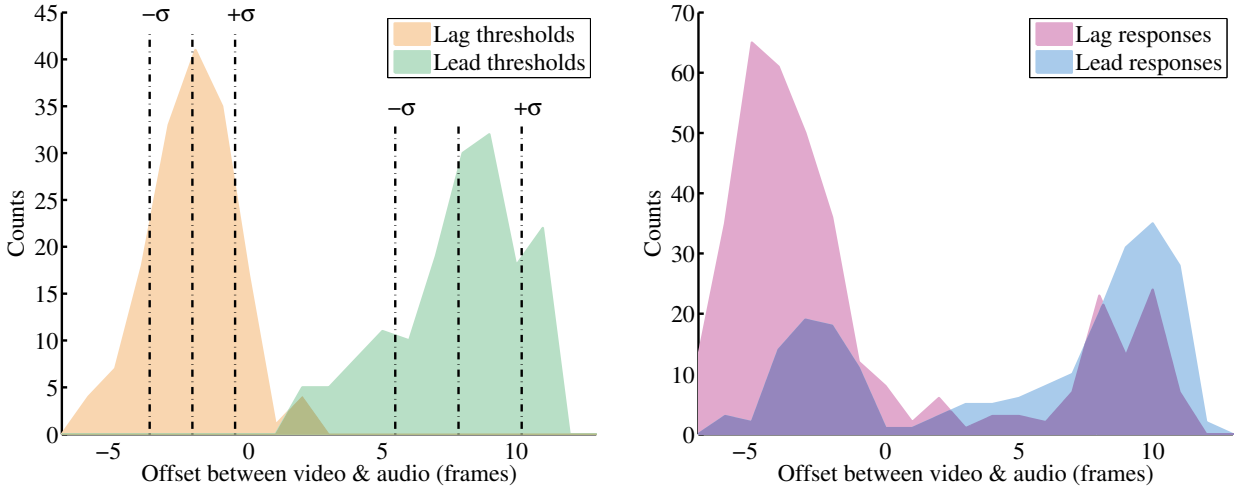

Fig. 2. Results of Experiment 1A. (Left) The distribution of thresholds, audio lag on the right and audio lead on the left, summed over all participants. The mean thresholds are -2.11 and +7.84 frames (@30 fps, -63.38 ms & 235.31 ms). The dotted vertical bars indicate the mean, mean - standard deviation and mean + standard deviation points for each distribution. (Right) The distribution of 'audio lead' (pink) and 'audio lag' (blue) responses summed over all participants. The bimodality of the distributions suggests that participants are better at judging in sync vs. out of sync than audio lag vs. audio lead.

4.2 Experiment 1B: Temporal Offsets and the Viewing Experience

For this study, we used the thresholds measured on our setup to examine the viewing experience under varying levels of audio-visual asynchrony.

4.2.1 *Experimental methods.* We selected 20 clips (10 emotional and 10 neutral) that were different from the ones in Experiment 1A. We edited these clips for every actor in the same way as before such that the audio track began at -5, -1, 0, +5 and +11 frame offset relative to the video track. The offsets were chosen based on the results of Experiment 1A. We chose two offsets (-5 and +11) that produced noticeably out-of-sync clips and two that were more subtle (-1 and +5) and fell below the threshold from Experiment 1A. Thus, we produced 20 clips × 4 actors × 5 offsets = 400 clips. These clips were divided into two non-overlapping sets of 200 clips each. Each set contained ten sentences, five emotional and five neutral, that were balanced in the manner described in Section 3.3.2. Each participant viewed only one of these sets, in the interest of time. Similar to the emotional clip selection procedure, they watched each video clip once and answered the following questions as they appeared afterwards:

1. Rate the PERFORMANCE of the actor. Press 1 = Very bad, 2 = Somewhat bad, 3 = Mediocre, 4 = Somewhat good, 5 = Very good.
2. Was the sentence NEUTRAL or EMOTIONAL. Press 1 = Neutral, 2 = Emotional.
3. Rate the INTENSITY of the performance. 1 = NEUTRAL SENTENCE, 2 = Not very intense, 3 = Medium intensity, 4 = Quite intense, 5 = Extremely intense.

The purpose of the second question was to help participants shape their responses to the third question. As half of our clips were emotionally neutral, we wanted participants to decide if a sentence conveyed emotion or not, before assigning an intensity rating. If the response to the second question was 'neutral', participants were asked to indicate an intensity rating of 1. If the response to the second question was 'emotional', we expected ratings in the range of 2 to 5. The presentation order of clips was randomized for each participant.

4.2.2 *Participants.* Twenty participants (10 females and 10 males, ages 19-40 years, average age 25.85) completed this experiment, all of whom were fluent in English and possessed normal vision and hearing. Half of the participants saw the first set of 200 clips, and the rest viewed the other set. One of the participants took part in Experiment 1A as well.

4.2.3 *Results.* First, we examined the results for performance and intensity ratings for the neutral and emotional sets together. A two-way, repeated-measures ANOVA for performance ratings revealed a significant main effect of clip type ($F(1,19) = 19.86$, $p = 0.0003$), but not for offset condition ($F(4, 16) = 1.52$, $p = 0.24$). There was no significant corpus $\times$ offset condition interaction ($F(4, 76) = 1.45$, $p = 0.26$). For the intensity ratings, there was a trend towards a significant difference for the offset condition in the two-way, repeated measures ANOVA ($F(4, 16) = 2.85$, $p = 0.06$).

Next, we considered each set of clips separately. For the neutral clips, a one-way, repeated-measures ANOVA found no significant main effect of offset condition on performance ratings ($F(4, 16) = 1.34$, $p = 0.30$). However, there was a trend towards lower performance scores for asynchronous clips when the the data for the three offset conditions perceived as synchronous (-1, 0, and +5 offsets) were combined and compared to the two perceptibly out-of-sync conditions (-5 and +11 offsets) ($F(1, 19) = 4.26$, $p = 0.05$). Additionally, there was no significant main effect of offset condition on intensity ratings ($F(4, 16) = 2.52$, $p = 0.08$). There was also no significant effect on intensity when perceptibly in-sync and out-of-sync conditions were combined for comparison ($F(1, 19) = 1.95$, $p = 0.18$). This was expected as intensity ratings should have been selected as '1 = Neutral' for all of these clips.

To determine whether dramatic, emotional displays were particularly affected by synchrony perception, we examined the results for the emotional clips alone. The one-way, repeated-measures ANOVA for performance ratings revealed no overall effect ($F(4, 16) = 1.71$, $p = 0.20$). However, when the combined data for the three seemingly synchronous (-1, 0, and +5 offsets) versus two asynchronous conditions (-5 and +11 offsets) were compared, the clips perceived as in sync received higher performance ratings ($F(1, 19) = 5.65$, $p = 0.03$). The pattern of results for intensity ratings was similar, with no significant main effect of offset condition (($F(4, 16) = 1.39$, $p = 0.28$), but the combined data showed significantly lower ratings for those perceived as out of sync relative to those perceived as synchronous ($F(1, 19) = 4.52$, $p = 0.047$).

These results suggest that synchronization has an effect on viewer's experience of film clips, particularly for emotional sentences. Noticeably out-of-sync clips received lower performance quality and intensity ratings for emotional statements (see Figure 3). Also, there was a trend for lower performance ratings for asynchronous neutral statements.

5. EXPERIMENT 2: THE EFFECTS OF PITCH SHIFTS ON THE VIEWING EXPERIENCE

In this study, we created clips to recreate the effects of speed changes that would occur when transitioning (without sound correction) from recording in 24 fps to playing at 25 fps and going from 25 to 24 fps and

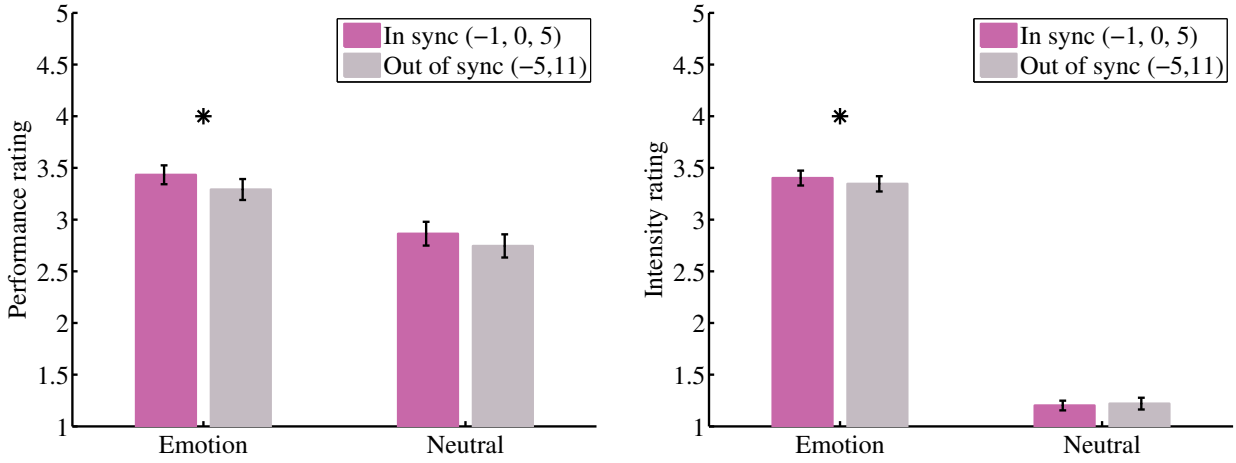

Fig. 3. Results of Experiment 1B. The responses averaged across participants are graphed here for performance ratings (left) and intensity ratings (right), both of which ranged from one to five. Error bars indicate 1 standard error of the mean for the averaged data of 20 participants. For the neutral sentences, participants were instructed to respond with an intensity rating of one.

examined viewers' preferences.

5.1 Experimental Methods

The full set of 26 matched neutral and emotional segments was used. The original clips (recorded and displayed at 30 fps) were modified in Final Cut Pro to play at 104% speed (similar to going from 24 to 25 fps) and 96% speed (similar to the 25 to 24 fps conversion). They were then exported at 30 fps. The pitch of the voice was not held constant during these speed changes; thus, the movies that played at 104% had higher pitched voices (HP condition) and the 96% had lower voices (LP condition) than the original clips (Original condition). All three types of clips were used in the experiment.

For each trial, video clips were presented in pairs. After the participant initiated a trial, one clip played, followed by a one-second pause, and then the second clip played. A screen was then displayed asking, "Which movie did you prefer? 1 = First; 2 = Second." The participant then keyed in an answer and initiated the next trial. All participants were instructed to select an answer at random if they had no preference.

Each pair of clips belonged to one of three sets: HP vs. Original, LP vs. Original, or HP vs. LP. Members of a pair were pitch-shifted versions of the same clip thus participants always compared clips containing the same actor speaking the same words. All possible pairs were generated for each sentence (4 actors × 26 sentences × 3 pairs =312 clips). The 312 clips were divided into four non-overlapping sets such each set contained all 26 sentences spoken by only one of the four actors. The assignments of actors to sentences was counterbalanced across the sets. Each experimental session involved viewing two of these sets, for a total of 156 pairs. Therefore, for any sentence, each participant viewed only two of the four actors performing it to ensure that the experiment did not exceed one hour. Each set was presented to eight participants. The presentation order of the clips in a pair and of clip pairs overall was randomized for each participant.

5.1.1 *Participants.* Sixteen participants (6 females and 10 males, ages 20-33 years, average age 26.75) completed this experiment. All were fluent speakers of English and had normal hearing and vision. Two of them participated in Experiment 1 as well, but they did Experiment 2 first. This order prevented them from knowing the original voices of the actors when doing Experiment 2.

5.2 Results

A one-sample, two-tailed t-test showed that the participants had no preference when Original clips were paired with LP clips, $t(15) = 0.40$, $p = 0.69$. This result suggests that they did not dislike the slowed movies relative to the correct speed, and movies could be transformed from 25 to 24 fps without negative consequence. For the HP vs. Original comparison, participants selected the Original clips significantly above chance, $t(15) = 2.40$, $p < 0.03$). Contrary to conventional wisdom, these data indicate that viewers might not like movies that have been sped up from 24 to 25 fps as much as those played at original speed. Finally, the LP vs. HP comparison showed a trend towards significant preference for LP stimuli above chance, $t(15) = 1.90$, $p = 0.08$.

Next, the trials were combined to examine whether high or low pitch was consistently preferred for each gender. For example, the higher pitched of the two options in each trial could be the Original clips in the LP vs. Original comparison as well as the HP clips when HP was paired with Original and LP clips. The participants showed no significant pitch preference in either direction for female actors ($t(15) = 0.18$, $p = 0.86$), but the lower pitch was consistently preferred for the male actors ($t(15) = 3.02$, $p = 0.01$).

Our results suggest that, contrary to industry practice [Phillips 2010], going from 25 to 24 fps has fewer perceptual consequences than going from 24 to 25 fps in the absence of pitch correction.

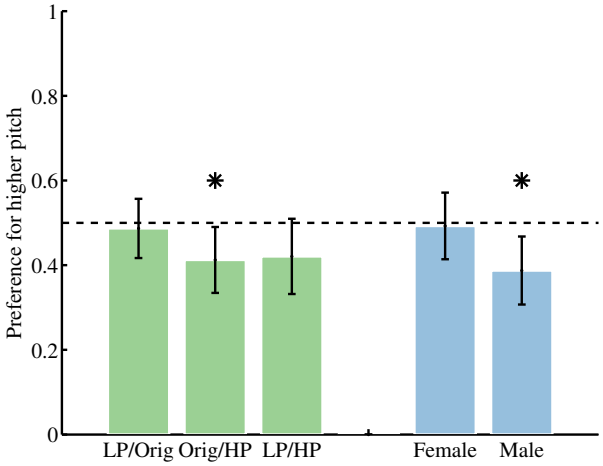

Fig. 4. Results of Experiment 2. The fraction of times the higher pitch in a pair was selected, is presented here in two ways. The first, on the left, shows the preference for higher pitch as a function of the type of comparison: LP vs. Original, Original vs. HP and LP vs. HP. The second, on the right, shows the preference as a function of the gender of the actors in the clips. The bars shown here indicate 95% confidence intervals for the averaged data of 16 participants. The dotted horizontal line indicates a preference of 0.5 (i.e., no preference).

6. DISCUSSION AND FUTURE WORK

Our findings suggest that media producers and distributors should take great care to ensure matching auditory and visual tracks for their products. We determined that the thresholds for audiovisual mismatch remain roughly the same despite increases in technology quality. However, the qualitative effects of timing mismatches on the audience had not previously been described beyond anecdotal reports of irritation [Summerfield 1992]. In fact, viewers' assessments of emotional performances were affected by noticeably out-of-sync playback such that ratings of both performance quality and emotional intensity decreased. During a

longer film, these factors could result in a decrease in emotional engagement, disrupting the dramatic experience. The industry should work to reduce synchronization problems in theaters, on television, and online, particularly for dramatic programming.

Additionally, we found that the common industry wisdom regarding changes in playback speeds that alter pitch is incorrect. Contrary to the belief that shifting material recorded at 24 fps to 25 fps has no audience effect, viewers in fact prefer the original speed and pitch. This result suggests that they do not like higher pitched voices with the same face, even for actors whom they have not heard in real life or at the correct pitch in a movie theater. On the other hand, there is no preference either way when material is slightly slowed and lower in pitch relative to the original. This finding, too, goes against current practice. Production companies should attempt to avoid these upshifts either by recording material at the same speed at which playback will occur or using a pitch shifter for the audio track to remove the one semitone difference.

In this work, we used clips in which the audio and video were recorded simultaneously, but that is not always the case. In the future, we plan to look at the effects of adding a voice track recorded at a later time, for example in ADR post-production, on the viewer experience. Additionally, movies and television programs in many markets are dubbed from the original language in which they were recorded into another language. The mismatch between the timing of mouth movements and sound could have similar results as the asynchronous presentation of the same source material, such that viewers may experience lower perceptions of performance quality and intensity when watching it in the new language. Studios attempt to prevent pitch mismatches during foreign language dubbing by hiring actors with voices that reasonably align with the appearance of the actors onscreen and often stick with a single voice actor for all films performed by a specific foreign actor [Tuckman et al. 2006]. We intend to address the effects of these procedures in foreign language dubbing in future research as well as potential remedies for timing mismatches, such as having the voiceover actor overact during emotional scenes to counteract intensity and performance problems.

ACKNOWLEDGMENTS

We would like to thank Greg Phillips (Disney ABC Television Group), Holger Spill (an independent motion pictures and film professional), and Ali Israr (Disney Research, Pittsburgh) for their advice and input. Additionally, we are grateful to everyone involved in creating the Disney face database: Barry-John Theobald, Moshe Mahler, Justin Macey, Shira Mahler, Darren Cosker, J. M. Conrad, Jeffrey Cohn, Rocky Smith, our FACS coders, and our actors. Finally, we thank all our participants for their time and effort. This study was funded in part by NSF Grant 0811450 and a Microsoft Research grant awarded to JKH.

REFERENCES

Abelin, A. 2007. Emotional McGurk effect - an experiment. In 7^{th} *International Conference on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems.*

Brainard, D. H. 1997. The psychophysics toolbox. *Spatial Vision 10*, 433–436.

Chion, M. 1994. *Audio-Vision: Sound on Screen.* Columbia University Press.

Dixon, N. F. and Spitz, L. 1980. The detection of auditory visual desynchrony. *Perception 9*, 719–721.

Ekman, P., Friesen, W. V., and Ellsworth, P. 1972. *Emotion in the human face: Guidelines for research and an integration of findings.* Pergamon Press.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, F., Pallett, D., and Dahlgren, N. 1993. Darpa timit acoustic-phonetic continuous speech corpus. Published on CD-ROM: NIST Speech Disc 1-1.1, Oct. 1990 NISTIR 4930, National Institute of Standards and Technology, Gaithersburg, MD.

Grant, K. W. and Greenberg, S. 2001. Speech intelligibility derived from asynchronous processing of auditory-visual information. In *Proceedings of Auditory-Visual Speech Processing 2001.*

Grant, K. W., van Wassenhove, V., and Poeppel, D. 2003. Discriminatory of auditory-visual synchrony. In *Proceedings of Auditory-Visual Speech Processing 2003.*

Huang, E., Sisk, J., Kirk, T., Coryell, G., and Stewart, J. 2007. http://www.iupui.edu/~nmstream/live/introduction.php.

Lass, N. J. and Harvey, L. A. 1976. An investigation of speaker photograph identification. *Journal of Acoustical Society of America 59*, 1232–1236.

Mason, A. and Salmon, R. 2008. Factors affecting perception of audio-video synchronization in television. In *Proceedings of the Audio Engineering Society Convention*. Audio Engineering Society.

McGurk, H. and MacDonald, J. 1976. Hearing lips and seeing voices. *Nature 264*, 746–748.

Munhall, K. G., Gribble, P., Sacco, L., and Ward, M. 1996. Temporal constraints on the McGurk effect. *Perception and Psychophysics 58*, 351–362.

Pelli, D. G. 1997. The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision 10*, 437–442.

Phillips, G. 2010. Personal communication.

Sakamoto, S., Tanaka, A., Tsumura, K., and Suzuki, Y. 2007. Effect of speed difference between time-expanded speech and talker's moving image on word or sentence intelligibility. In *Proceedings of Auditory-Visual Speech Processing 2007*.

Spill, H. 2010. Personal communication.

Sumby, W. H. and Pollack, I. 1954. Visual contribution to speech intelligibility in noise. *Journal of Acoustical Society of America 26*, 212–215.

Summerfield, Q. 1987. *Some Preliminaries to a Comprehensive Account of Audio-visual Speech Perception*. Lawrence Erlbaum Associates, London.

Summerfield, Q. 1992. Lip reading and audio-visual speech perception. *Philosophical Transactions: Biological Sciences 335,* 1273 (Aug.), 71–78.

Theobald, B.-J., Bangham, A., Matthews, I., and Cawley, G. 2003. Evaluation of a talking head based on appearance models. In 7^{th} *International Conference on Audio Visual Speech Processing*.

Tinwell, A. and Grimshaw, M. 2009. Survival horror games: An uncanny modality. In *Thinking After Dark*.

Tuckman, J., Chrisafis, A., Hooper, J., Watts, J., Smee, J., and Ramesh, R. 2006. http://www.guardian.co.uk/film/2006/nov/03/3.

van Wassenhove, V., Grant, K. W., and Poeppel, D. 2007. Temporal window of integration in bimodal speech. *Neuropsychologica 45,* 3, 598–607.

Received June 2010; revised June 2010; accepted June 2010