Compensating for Low Frame Rates

Hendrik Knoche
University College London
Gower St
London WC1E 6BT, UK
h.knoche@cs.ucl.ac.uk

Hermann de Meer
University of Passau
Innstr. 33
94032 Passau, Germany
demeer@fmi.uni-passau.de

David Kirsh
University of California, San Diego
9500 Gilman Dr
La Jolla, CA 92093-0515, USA
kirsh@ucsd.edu

ABSTRACT
Experiments were conducted to investigate the interdependency of frame rates (30, 15, 10 fps) and audio-visual skew (from +163 to -233 ms[1]). Noised nonsense words like 'abagava' were presented to 20 participants who were asked to identify the middle consonant. At low frame rates (10 fps) consonant perception was impaired when audio ran ahead of video content (skew of -113 to -233ms). When audio lagged video, performance improved monotonically to a maximum at +167ms, where performance equaled 30fps in synch. The results suggest that frame rate and skew are not orthogonal parameters but must both be taken into consideration for AV-delivery. The findings do not support the current notion that 10 fps videos do not adequately capture visual content for speech perception. Participants were able to integrate the given bi-modal information as well as the 30 fps condition if the audio channel was subjected to an additional 167ms delay.

Author Keywords
Audio-Visual Integration, Frame Rates, Skew, Speech Perception

ACM Classification Keywords
H.5.1 Multimedia Information Systems: Audio, Evaluation, Video; H.5.2 User Interfaces: Voice I/O

INTRODUCTION
In telecommunication, video displaying the face of the speaker has proven to enhance communication under 'noisy' conditions, e.g., in mobile or poor audio scenarios, and for non-native speakers. Speech perception is superior in the presence of additional visual information, whether the materials presented are sentences [1, 2], meaningful words [3], or nonsense syllables [4], [5]. The generality of the findings supports the idea that vision contributes to speech perception regardless of lexical status or sentential context. Moreover, studies have indicated that normal-hearing participants make use of lip-reading under adverse listening conditions [6, 7, 8].

More is not necessarily better, however. The benefits of visual cues can only be reaped if video and audio are played out in a synchronized fashion. Moreover, McGurk found that discrepant acoustic and visual information may lead to perceived sounds differing from both inputs (e.g., a visual 'ba' dubbed with an acoustic 'ga' can be fused to a 'da') [10].

Consequently, poorly synchronized presentations due to technical imperfections or induced by low frame rates might not only render the supporting video useless but could produce errors that wouldn't happen with audio only. To ensure congruent inter-sensory presentations, the amount of asynchrony (skew) between audio and video has to be kept within bounds and the video must neither omit significant cues nor present discrepant visual information. This translates to requirements for both frame rates and audio-visual skew. Two guidelines for audio-visual communication can be readily derived from the existence of the McGurk effect. First, the synchronization of the audio and video stream has to be ‘tight enough’ so that no discrepant acoustic and visual information are presented to the user. Second, the chosen frame rate must ‘adequately’ capture the significant moments in the message, e.g., closed lips moments. The studies that have investigated these bounds are presented in the following two sections.

Lowered video frame rates have been a popular countermeasure to reduce bandwidth requirements. They also simplify the display of audio-visual content on mobile devices. However, several studies [1], [15], [16] have shown that low frame rates - in some cases as high as 10 fps - under-sample the visual data and omit valuable cues e.g., closed lips occurrences, beneficial for the recognition of labial consonants like ‘b’ and ‘p’.

Despite the somewhat obvious dilemma that low frame rates present data continuously going out of synch, the

CHI 2005, April 2–7, 2005, Portland, Oregon, USA.

ACM 1-59593-002-7/05/0004.

[1] Throughout this paper, positive skews will indicate audio lagging the video while negative skews will signify the opposite.

relationship between frame rates and skew has not yet been studied. Rather, skew and frame rate have been considered to be orthogonal parameters and have been implemented likewise in current synchronization schemes.

In Section 2 we give an overview of audio-visual integration. Section 3 describes the details of the experiment. Results of the experiment are discussed in Section 4 and our conclusions are given in Section 5.

AUDIO-VISUAL INTEGRATION

Temporal Constraints on AV Integration

McGrath and Summerfield explored the lip-reading performance of normal-hearing adults as a function of skew (0, 20, 40, 80, 160 ms) [12]. The results implied that skews of up to about 40 ms do not materially affect audio-visual speech understanding.

Pandey et al. studied the effect of skew (6 steps from 0 to 300 ms) on speech perception with an audio signal degraded by a masking noise (SNR of 0 and -10 dB) [13]. Since skews of up to 80 ms did not affect the result scores they followed McGrath and Summerfield's hypothesis that skew is not significantly disruptive for phonemic identification in connected speech but becomes important at a syllabic level. Skews of up to 120 ms were projected to be acceptable if the information provided by the audio signal was fairly high.

Massaro et al. examined effects of various skews (0, ±67, ±167, ±267, ±533 ms) for the correct hearing of syllables [14]. Audio-visual integration wasn't found to be disrupted with skews up to the range of ±150 ms but was clearly impaired at approximately half a second.

Effects of Frame Rates on AV Integration

Frowein et al. examined the effect of transmitting 64 kbit/s video-telephony of varying temporal resolution on the lip-reading ability of individuals with hearing loss [15]. They recommended a temporal resolution of 15fps when 'speech readability' is an important aspect of video-telephony.

A study conducted by Vitkovitch and Barber assessed participants' ability to shadow[2] verbal messages when they could both hear and see (at 8.3, 12.5, 16.7, 25fps) the speaker in comparison to an audio-only baseline [16]. The presence of the visual image of the relevant speaker generally improved performance compared to the baseline condition. Performance was impaired for both 8.3 and 12.5fps.

Nakazono conducted studies with different frame rates and one with frame rates paired with audio lagging the video [1]. The first study concentrated on the impact of different frame rates (30, 10, 5, 2fps) on the McGurk effect. The results showed that degrading the frame rate decreased the incidence of mishearing for discrepant stimuli or, in other words, the McGurk effect. Nakazono concluded that the contribution of the visual to speech perception was degraded by a lower frame rate. In another study still pictures of normal Japanese speech were inspected to determine a lower limit of frame rate from the view point of hearing assistance. A frame rate of 10fps was considered to be sufficient since 66% of labial consonants were successfully captured with respect to frames displaying closed lips.

The findings of Vitkovitch et al. and Nakazono suggest that 10 fps videos are under-sampling the visual content and omit cues that are present in 15 and 30 fps videos.

THE EXPERIMENT

Participants

The participants, 9 female 11 male, were all American, native English speakers between 19 and 33 years old. The participants reported having normal hearing and normal or corrected-to-normal vision.

Stimulus Material and Preparation

Participants had to identify the middle consonant of four-syllable nonsense words. The 64 base stimuli covered all permutations of the consonants 'b', 'd', 'v', 'g', interleaved by the vowel 'a', with each stimulus beginning and ending in 'a', e.g., 'abadava'. This structure of the stimuli assured that even under skews of 200 ms there would still be dynamic visual input that would contend for integration. Considering an average syllable length of 250 ms, the acoustic ‘b’ of a short stimuli as ‘aba’ wouldn’t have any counterpart in the visual domain if skews were as big as 250 ms.

For the audio-visual and the audio only stimuli, the audio track(s) were extracted from the video. The audio was dubbed with 11dB white noise to make the words harder to recognize in order to avoid cut-off effects in task performance. For the skewed audio-visual stimuli respective amounts of silence (40-200 ms) were inserted at the beginning or end of the stimulus and the same amount was deleted at the opposite end of the stimulus. The 30 fps videos were combined with nine skews (0, ±80, ±120, ±160, ±200 ms). The lower frame rate videos (15 and 10 fps) were generated at 9 different skews (-233, -193, -153, -113, -33, +47, +87, +127, +167ms).

There were also two audio-only conditions one with a blank screen and the second showing a still image of the speaker. This resulted in a total of 29 conditions and 29*64=1856 different stimuli.

The speaker was a woman speaking unaccented American English at a normal rate (3-4 syllables/sec). The obtained syllables had approximately equal vowel duration and volume. She started and ended every stimulus with closed lips.

Stimuli were recorded using a Sony camcorder TR700 for both audio and video. The videos were captured with a

[2] The message is presented at the same time with an irrelevant background message.

Miro DC30plus on a Windows NT 4.0 System. Video-editing was carried out with Adobe Premiere software. Alterations to the audio were made using Cool Edit Pro.

Apparatus

The experiment was set up in a soundproof chamber. The stimuli were presented through a Panasonic ct-1381 TV monitor (13" viewable diagonal), which obtained its audio and video signal from the aforementioned video card in a dedicated Windows NT 4.0 machine The distance between participants and the screen was about 1.2m (viewing distance to picture-height ratio of 6). The volume of the audio was set to a reasonable level.

Answers were recorded with a repurposed IBM computer-keyboard with click that only had five keys - one key for each of the four respective consonants 'b', 'd', 'v', 'g', and 'o' for answers other than the former four. The keyboard was positioned at the monitor to reduce to a minimum head movement between the screen and keyboard.

Procedure

An introduction explaining the course of the experiment was given on the screen. Participants were told that it would help to watch the speaker's lips before pressing the key that corresponded to the consonant that they had perceived. Following the instructions, 4 of the 64 original videos were presented, i.e., without noise. To ensure that the participants had understood the instructions and actually concentrated on the second consonant each consonant was given once with no repetition of that consonant in the stimulus word. Then the experiment commenced in a subject-driven manner. After each stimulus a black masking frame was shown. The next stimulus was played within one to two seconds after an answer had been received. The total experiment covered 8 blocks of 45 stimuli that were interleaved by one-minute pauses. The randomization of the stimuli assured that all participants saw all consonants in all of the (9*3+2=29) conditions equally often.

RESULTS AND ANALYSIS

The results for each subject in each of the (3*9+2=29) configurations were averaged. The audio-only conditions were taken as the baseline performance. All the presented averages of the participants' average scores in Figure 2 and Figure 3 are given with 0.05 confidence intervals.

As shown in Figure 1 participants' performance was different for lower frame rates.

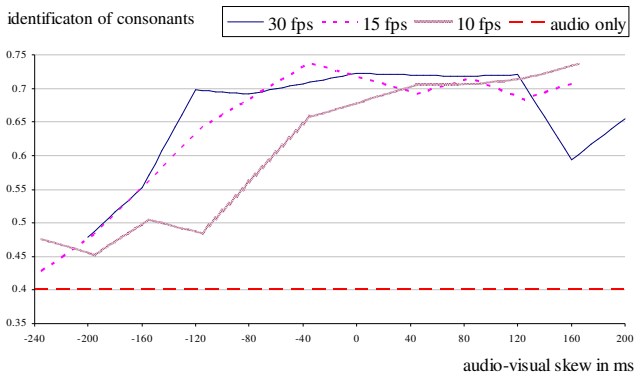

Figure 1: Frame rates in comparison

Apart from the unexpected value at 200 ms, the pattern of the 30 fps curve is consistent with former findings of [11, 12, 13]. This deviation might be due to the structure of the stimuli.

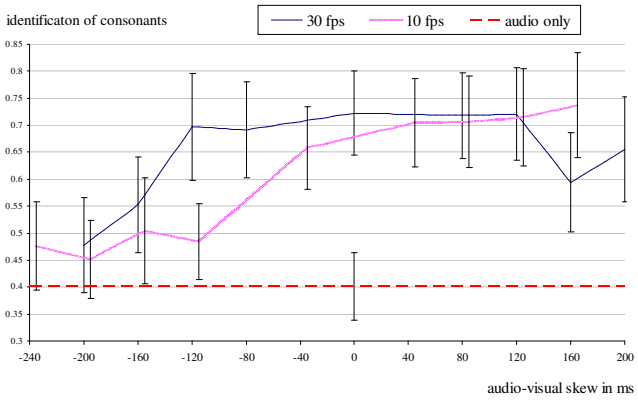

Figure 2: Performance for 10 fps in comparison to 30 fps

In Figure 2 we can see that the performance in consonant perception for 10 fps at -113 ms was significantly different from the performance of the 30 fps video at -120ms and -80ms. At +167 ms 10 fps videos scored comparably to 30 fps video within ±120 ms. We conclude that no visual cues are lost but that in order to make use of them they have to be presented at a different time. It also seems that the exact moment of closed lips might not be necessary for the audio-visual integration for labial consonants but that a sudden reversal of lip movement with almost closed lips might be enough of a cue.

The changes in performance for 15 fps depicted in Figure 3 are not as drastic as for 10 fps. Nevertheless, we can observe that at -113ms the value for consonant perception at 15 fps is already less than for 30 fps at -120ms. On the opposite end, the 15 fps graph shows a better result at positive skews (+167ms) as 30 fps.

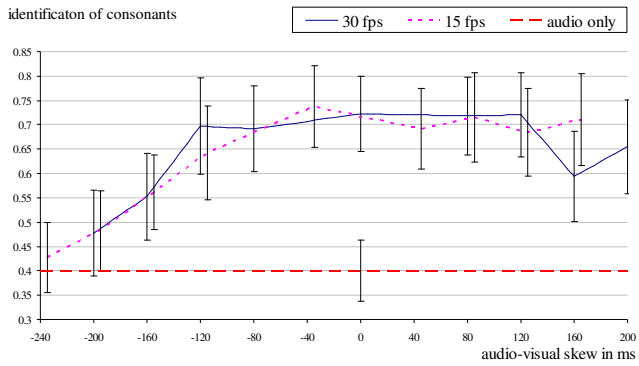

Figure 3: Consonant identification at 15 and 30 fps

CONCLUSIONS

We have studied how frame rates interact with AV-skew and found that the lower the frame rate the more sensitive comprehension is to audio skew. On the one hand, positive skews improve comprehension; on the other, negative skews degrade comprehension. More precisely, at frame rates of 10 and 15 fps, audio lags of between +127ms and +167ms result in better perception of consonants in a noisy environment of 11db white noise than at full 30 fps. At low fps video frames are exposed long enough to close the gap between audio and video so that subjects can compensate for the lag. The case with audio leads is that, e.g., at -120ms AV integration of 10fps video is 30% less than that of 30fps. This is because at 10fps audio is leading each video frame between –120 and –220ms, which pushes the AV integration limit of humans (commonly thought to be around ±120ms) by presenting possibly increasingly conflicting visual information and simultaneously withholding more up-to-date information from visual integration.

A consequence is that the temporal window of AV integration has to be changed to accommodate frame rates. Previously, it has been assumed that audio and video are considered to be "in synch" anywhere between ±120 ms. This remains true for 30fps but does not hold for lower rates. We suggest a more general rule: the window of AV integration is to be determined by adding a *frame rate corrector* to the two bounds. The corrector is obtained by subtracting 33ms from the presentation time of a frame. In the case of 10fps 100ms – 33ms = +66ms or, for 15 fps this is 66ms – 33ms = 33ms. Thus, for a 10 fps video the window for AV integration runs from –54ms to +186ms.

The significance of these findings for multimedia delivery is twofold: 1) Video at 10fps can be effective in noisy environments if audio is delayed between +120 and +170ms. Low frame rates work as well as high frame rates under these conditions. Indeed, if you are stuck with an audio lag of 170ms or higher it is advisable to switch from 30fps to a lower frame rate. 2) The lower the frame rate the more sensitive comprehension is to audio leads. If you are stuck with low frame rates (10fps) it is important to let audio lag video by at least +40ms.

Whether this temporal burden can be placed upon the end-to-end delay – telecommunication's Achilles heel - has to be dealt with in future research.

REFERENCES

1. Nakazono, K. Frame rate as a QoS parameter and its influence on Speech Perception, Multimedia Systems 6, 1998
2. Reisberg, A., and Lubker, J. Prosody and speech-reading, STL-Quarterly Progress and Status Report, Dept. of Linguistics, Univ. of Stockholm, Sweden, 1978
3. Campbell, R., and Dodd, Hearing by Eye, Quarterly Journal of Experimental Psychology, 32, 1980
4. Binnie, C., A., Montgomery, A. A., and Jackson, P. L. Auditory and visual contributions to the perception of consonants, J. of Speech and Hearing Res., 17 (4), 1974
5. Smeele, M. T., Sittig A. C. The contribution of vision to speech perception', Proc. Human Factors in Telecom., Torino, 1990
6. O'Neill, J. Contributions of the Visual Components of Oral Symbols to Speech Comprehension, J. of Speech and Hearing Disorders, 19, 1954
7. Neely, H. Effects of Visual Factors on the Intelligibility of Speech, Journal of the Acoustical Society of America, 28, 1956
8. McCormick, B. Audio-Visual Discrimination of Speech, Clinical Otolaryngology, 4, 1979
9. Sumby, W., and Pollack, I. Visual contributions to speech intelligibility in noise, J. of the Acoustical Society of America, 26, 1954
10. McGurk, H., and MacDonald, J. Hearing lips and seeing voices, Nature, Vol. 264, (no.5588), 1976
11. Steinmetz, R. Human perception of jitter and media synchronization, IEEE JSAC, Vol.14, (no.1), Jan., 1996
12. McGrath, M., and Summerfield, Q. Intermodal timing relations and audio-visual speech recognition by normal hearing adults, J. of Ac. Soc. of America 77 (2), 1985
13. Pandey, P. C., Kunov, H., Abel, S. M. Disruptive effects of auditory signal delay on speech perception with lipreading, J. of Auditory Research, Jan. 26 (1), 1986
14. Massaro, D. W., Cohen, M. M., and Smeele, P. M. T. Perception of asynchronous and conflicting visual and auditory speech, J. Ac. Soc. of America, 100 (3), 1996
15. Frowein, H., Smoorenburg, G., Pyters, L., and Schinkel, D. Improved speech recognition through video-telephony: Experiments with the hard of hearing', IEEE JSAC, 9, 1991
16. Vitkovitch, M., Barber, P. Effect of Video Frame Rate on Subjects' ability to Shadow One of Two competing verbal Passages, J. of Speech and Hearing Res., Vol. 37, 1994