Extremely economical: How key frames affect consonant perception under different audio-visual skews

H. Knoche[a], H. de Meer[b], D. Kirsh[c]

[a] *Department of Computer Science, University College London, London WC1E 6BT, UK*
[b] *Department of Computer Science, University of Passau, 94032 Passau, Germany*
[c] *Department of Cognitive Science, University California San Diego, La Jolla, CA 92093, USA*

Abstract

In audio-visual telecommunication, low video frame rates represent a popular method for saving on bandwidth requirements. When key frames displayed the extremes of lip movements we found that participants performed comparably to standard displays at 30 frames per second. Experiments were conducted to compare the effectiveness of a small number of algorithmically chosen key frames - typically 7 to 8 frames per second (fps) - to 30fps displays where audio and video were out of synch by as much as 233ms. Noised non-sense words like 'abagava' were presented to 20 participants who were asked to identify the middle consonant. The results indicate that key frame displays are as effective as 30fps when audio lags video by 87 to 167ms. Despite the low temporal resolution and varying exposure lengths, participants were able to integrate the given bi-modal information as well as the 30fps condition if the audio channel lagged the video by 87ms. The latter is recognized as being within the region of optimal audio-visual (AV) integration.

Keywords: Video Communication, Multimedia, Human Factors, Speech Perception, Synchronization

1. 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. More is not necessarily better, however. The benefits of the provided visual cues can only be reaped if video and audio are played out in a synchronized fashion. McGurk found that discrepant acoustic and visual information may lead to perceived sounds differing from both input (e.g., a visual 'ba' dubbed with an acoustic 'ga' can be fused to a 'da'). Consequently, poorly synchronized presentations due to technical imperfections or induced by low frame rates might not only render the supporting video useless but, far worse, could produce errors that would not 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 capturing/presentation must neither omit significant cues nor present discrepant visual information.

One of the major deterrents from audiovisual communication is the additional cost of transmitting video. Lowered video frame rates have been a popular countermeasure to cut down on these expenditures. However, frame rates as high as 10fps might under-sample the given data and omit valuable cues, e.g., closed lips occurrences, beneficial for the recognition of labial consonants like ‘b’ and ‘p’ [1]. Furthermore, users complain about the lack of synchronization between audible and visual content when confronted with low frame rates [2].

Typically, lower frame rates have constantly elongated intervals between captured frames and likewise for their corresponding presentation times. We want to explore another resource-saving alternative that displays only 'essential' so-called key frames for different amounts of time and test it under different audio-visual skews. Thereby, we hope to find how to optimally utilize humans' natural capability of audio-visual integration.

Most commonly, the term key frame is used to identify a video frame that carries semantically important information. In this paper we want to consider as key frames those frames that capture the most meaningful parts of a video sequence, i.e., those that display the lips of the speaker in extreme positions (closed or open). For example, in the stimulus 'abababa' there are 7 extreme lip positions (9 assuming closed lips before and after the utterance).

In Section 2 we give an overview of audio-visual integration. Section 3 describes the key frame algorithm that was simulated and tested in the experiment in Section 4. The results of the experiment are discussed in Section 5 and the conclusions are given in Section 6.

2. Audio-visual integration

Speech perception is superior in the presence of additional visual information, whether the materials presented are sentences [1], [5], meaningful words [6], or nonsense syllables [7], [8]. The generality of the findings across languages 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 [9], [10], [11]. Sumby and Pollack reported that the relative contribution of the visual information is independent of the signal-to-noise ratio (SNR), but the absolute contribution can be more profitably exploited at low SNR [12].

2.1 McGurk effect

When audio and video information are discrepant, participants perceive a different sound than that which is actually uttered. The perceived, or misheard, sound is sometimes the same as the mouth motion and sometimes a different sound that is fused by acoustic and visual stimulus. This is referred to as the McGurk effect [13]. McGurk and McDonald observed that lip movements for [ga] (acoustic [ba]) are frequently perceived as [da], while those for [ka] (acoustic [pa]) are sometimes perceived as [ta]. [pa] and [ba] are often confused with one another.

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.

2.2 Temporal Constraints on AV Integration

Dixon and Spitz reported a mean value of 250ms for the minimum detectable skew during the presentation of connected speech [14]. Steinmetz conducted similar experiments in which participants had to judge if audio and video were in synch [15]. He reported that skews from -80ms to 120ms were not perceived by participants. Koenig studied the effect of skew (15, 30, 60, .., 1920ms) on speech understanding using low-pass filtered words and sentences on one subject [16]. Performance in both cases was not affected until the skew exceeded 240ms.

McGrath and Summerfield explored the lip-reading performance of sentences of normal-hearing adults as a function of skew (0, 20, 40, 80, 160ms) [18]. The audio track was replaced by rectangular pulses, originally synchronized to the closing of the talker's vocal folds and then subjected to skew. Performance was significantly decreased for all participants at 160ms. The performance of a subgroup of better lip-readers was reduced with increased skew from 0 to 80ms. In their second experiment normal hearing observers were asked to determine whether a 120-Hz complex tone started before or after the opening of a pair of lip-like Lissajou figures. The evaluation of the second experiment suggested that speech-like stimuli do not resolve skews between -80 and 140ms. The results of both experiments implied that skews of up to about 40ms (introduced by signal-processing algorithms) do not materially affect audio-visual speech understanding.

Pandey et al. studied the effect of skew (6 steps from 0 to 300ms) on speech perception with an audio signal degraded by a masking noise (SNR of 0 and -10dB) [19]. The test material consisted of sentences with or without a context picture representing one of the key words in each sentence. The disruptive effect of the skew was found to be a function of both the context and the SNR. Since skews up to 80ms 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 only at a syllabic level. Skews of up to 120ms were projected to be acceptable if the information provided by the audio signal was fairly high.

Massaro et al. examined effects of various skews for the correct hearing of syllables. They presented combined synthetic and natural speech audio syllables with synthetic video speech syllables at various skews (0, ±67, ±167, ±267, ±533ms). The synthetic visual CV (consonant-vowel) syllables 'ða', 'ba', 'da', 'va' were permutated over the same acoustic syllables [20]. Audio-visual integration was not found to be disrupted with skews up to the range of ±150ms but was clearly impaired at around half a second.

2.3 Effects of Frame Rates on AV Integration

Frowein et al. examined the effect of transmitting 64kbit/s video-telephony of varying temporal resolution on the lip-reading ability of individuals with hearing loss [21]. Participants had to report sentences that were presented against background speech-spectrum noise. Increasing temporal resolution from 5 to 15fps improved speech reception scores, although beyond 15fps no further improvement was observed. Frowein et al. 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 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 [22]. 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 several 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.

Knoche et al. looked at the interaction of frame rates and audio-visual skews. They studied the performance of consonant perception (b, d, g, v) when videos was presented at 10, 15 and 30fps with audio-visual skews from -233ms to 200ms. The study found that speech perception at the lowest frame rate was more impaired when audio lead video by 120ms or more. With video leading audio up to 167ms consonant perception at lower frame rates was at least as good as or better than at 30fps [23].

3. Key frames

Generally, key frames denote semantically important frames. In many coding schemes this coincides with intra-coded frames, which do not rely on the existence of other frames to display their content. In this paper we choose to regard frames as important if they are indispensable for the performance of speech perception. Out of a stream of a regular 30fps video we want to pick a minimal number of frames while still maintaining optimal communication, i.e., vowel perceptio, in this case. Unlike constant frame rates that sample at equidistant instances in time, our approach allows for key frames at irregular intervals. The resulting sequence consists of key frames that are to be displayed for variable amounts of time.

We considered two major facts in the design of our key frame algorithm. By choosing frames where lips are in extreme positions (closed or open, see below), we hoped to optimize vowel perception according to the findings from AV integration. Second, we set 6fps as our lower visual refresh bound considering that users avoid frame rates below 5fps [2].

3.1 Algorithmic Selection of Key Frames

For our algorithm we assumed the existence of a lip movement tracker providing the distance, d, between the speaker’s lips. For each frame we would compute Δ_n the difference between d_n - the lip distance of the n-th frame – and d_{n-1}. In order to prevent oscillations in the algorithm, differences below a certain threshold were set to zero. Negative differences occur when lips are closing whereas opening lips yield positive differences. If the obtained Δ_n has an inverted sign compared to that of frame n-1 then the current frame starts a reversed lip movement. By choosing the frame n-1 right before this reversal, we hoped to pick local minima and maxima in lip positions. For the videos in the experiment described in Section 4 we manually simulated the key frame algorithm and the aforementioned threshold was applied on the basis of common sense.

In short, the algorithm filters key frames from a 30fps stream and presents each of them until the occurrence of the next key frame according to the following two rules:

1. If we have not selected any of the past five frames we pick the current frame as a key frame.

2. If the direction of lip-movement has been reversed with the current frame, we select the preceding frame as a key frame.

Choosing the preceding frame according to rule 2 incurs an audio-visual skew of -33ms (video leading audio) for those stretches where direction reversals occur. This approach does not provide an upper bound for frames per second below 30fps but will usually bring us below 10fps. Considering a speech frequency of 4-5 syllables per second and a maximum of 3 key frames for the first and 2 for each additional syllable, we are looking at 8-10fps at most.

The picking algorithm is demonstrated in the video sequence in Figure 1 displaying the ‘abab’ part of the stimulus word ‘ababava’. The top row depicts the original 30fps sequence with the frames 1, 5, 9, and 11 cut out and used as the key frames in the key frame sequence in the bottom row. Frame 1 was chosen due to rule 1. Frames 5, 9, and 11 were selected according to rule 2 and were subjected to the 33ms delay in the resulting stream. The light pictures in the key frame row signify the presentation time of the key frames. We emulated the key frame algorithm by creating copies of the key frames and filling up the 30fps grid with them until the next key frame was chosen.

4. The Experiment

4.1 Participants

The participants, 9 female and 11 male were between 19 and 33 years old. They were all university students, American and native English speakers. The participants reported having normal hearing, and normal or corrected-to-normal vision. Deficits in hearing were controlled for with audio-only stimuli during the experiment.

4.2 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'. The structure of the stimuli was motivated by shortcomings of former studies in which under skewed conditions, the acoustic information did not have any corresponding visual input to contend for integration. Considering an average syllable length of 250ms, the audio of the ‘b’ in ‘aba’ does not have any counterpart in the visual domain if skews are as big as 250ms.

For the skewed stimuli, the audio track(s) were extracted from the video. Then respective amounts of silence (40-200ms) were inserted at the beginning or end of the stimulus and the same amount was deleted at the other end of the stimulus. After that, the audio track was dubbed with 11dB white noise to make the words harder to recognize in order to avoid ceiling effects in task performance. Each of the 64 videos was generated in both frame rates (30fps and key frames) paired with the 9 different skews (0, ±80, ±120, ±160, ±200ms) and an audio-only condition. This resulted in a total of 64*2*9+64=1216 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

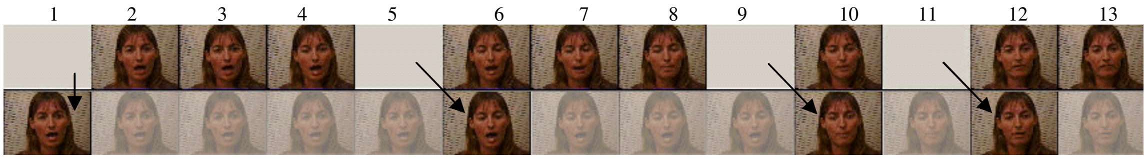

Figure 1: Example of a key frame selection process (bottom row) from a 30fps video (top row)

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

4.3 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 1.2 meter (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 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 others than the former. The keyboard was positioned right at the monitor to reduce head movements between the screen and input to a minimum.

4.4 Procedure

Participants were tested individually. 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 and to press the key that corresponded to the consonant that they perceived. Following the instructions, 4 of the 64 videos were presented without noise. Each consonant was given once with no repetition of that letter in the stimulus word, such that it could be controlled that the participants had understood the instructions and actually concentrated on the second consonant. Then the experiment commenced.

The experiment was subject-driven. 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 30 stimuli that were interleaved by one-minute pauses. The randomization of the stimuli assured that all participants saw all consonants in all of the conditions equally often.

5. Results and analysis

The vowel identification results for each subject in each configuration were averaged. The audio-only conditions were taken as a baseline performance indicator. The graph depicting the key frame videos in Figure 2 already reflects the introduced skew of the key frame algorithm, i.e., has been shifted by -33ms. Figure 2 depicts the results for both labial and non-labial consonants taken together with 0.05 confidence intervals.

On average participants scored better than chance (25%) in the audio-only condition (40%, not across all consonants, though). The results for the 30fps generally replicated former findings of [16], [18], and [19] with the exception of the increase at +200ms that might be an artefact due to the number of consonants used in the stimuli. As we can see for the key frame videos at +47, +87, +127, and +167ms performance is not significantly different than the best performance of the 30fps videos in the region between 0 and +120ms. However, the identification performance for the key frame algorithm is far more sensitive to audio leading video.

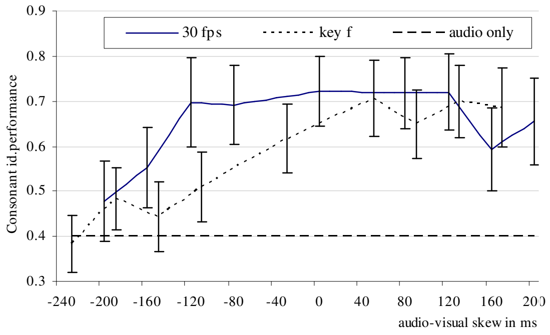

Figure 2: Consonant identification performance

One possible explanation for this could be that in 30 fps video participants have access to visual cues much earlier (for example the downward movement of lips) whereas in key frame videos participants only experienced discrete states of open and closed lips.

6. Conclusions

We have studied a key frame algorithm that used a minimal number of frames and was tailored towards optimal performance in labial consonant perception by capturing extreme positions of lips. The temporal window where audio-visual integration performs best was different for the key frame videos compared to the 30fps control. In general, consonant perception for the key frame videos was more sensitive to negative skew, i.e., when audio lead video. Despite the somewhat jerky movements the audio-visual integration did not break down and consonant perception for key frame videos with audio skewed between +47 and +167ms was at least as good as or better than videos at 30fps. We expect these findings to be relevant across languages.

Benefits from this study can be reaped even without an actual implementation of the key frame algorithm. If we had the means to detect the lip movements, frames that contained extreme lip positions could be tagged differently than other frames when sent across a network to differentiate between two quality of service (QoS) classes. For example, a higher loss rate could be tolerated for the non-key frame class. The knowledge can also be applied in minimal audio-visual communication scenarios where models of the partners are rendered locally instead of transferring the real pictures. The results of this study suggest aligning the audio content 40 to 160ms after the corresponding video content.

REFERENCES

[1] Nakazono, K. Frame rate as a QoS parameter and its influence on Speech Perception, Multimedia Systems 6, 1998

[2] Pappas, T., Hinds, R. On Video and Audio Integration for Conferencing, Proceedings SPIE, Vol. 2411, The International Society for Optical Engineering, Bellingham, WA, 1995

[3] Reisberg, D., and McLean, J., Goldfield, A. Easy to hear but hard to understand: A lipreading advantage with intact auditory stimuli', in [4]

[4] Dodd, B., and Campbell, R. Hearing by Eye: The psychology of lip reading, Lawrence Earlbaum Assoc., London, 1987

[5] Reisberg, A., and Lubker, J. Prosody and speech-reading, STL-Quarterly Progress and Status Report, Department of Linguistics, University of Stockholm, Sweden, 1978

[6] Campbell, R., and Dodd, Hearing by Eye, Quarterly Journal of Experimental Psychology, 32, 1980

[7] Binnie, C., A., Montgomery, A. A., and Jackson, P. L. Auditory and visual contributions to the perception of consonants, Journal of Speech and Hearing Research, 17 (4), 1974

[8] Smeele, M. T., Sittig A. C. The contribution of vision to speech perception', in Proceedings of the 13th International Symposium on Human Factors in Telecommunications, Torino, 1990

[9] O'Neill, J. Contributions of the Visual Components of Oral Symbols to Speech Comprehension, Journal of Speech and Hearing Disorders, 19, 1954

[10] Neely, H. Effects of Visual Factors on the Intelligibility of Speech, Journal of the Acoustical Society of America, 28, 1956

[11] McCormick, B. Audio-Visual Discrimination of Speech, Clinical Otolaryngology, 4, 1979

[12] Sumby, W., and Pollack, I. Visual contributions to speech intelligibility in noise, Journal of the Acoustical Society of America, 26, 1954

[13] McGurk, H., and MacDonald, J. Hearing lips and seeing voices, Nature, Vol. 264, (no.5588), 1976

[14] Hirsh, I. J., and Sherrick, C. E. Perceived order in different sense modalities, Journal of Experimental Psychology, 62, 1961

[15] Dixon, N., and Spitz, L. The Detection of Auditory Visual Desynchrony, Perception, 9, 1980

[16] Steinmetz, R. Human perception of jitter and media synchronization, IEEE Journal on Selected Areas in Communications, Vol.14, (no.1), Jan., 1996

[17] Koenig, E. Electroacoustic Equipment for Determining the Effect of the Time Factor on Directional Hearing and Fusion Phenomena, Int'l Audiol., 3, 1964

[18] McGrath, M., and Summerfield, Q. Intermodal timing relations and audio-visual speech recognition by normal hearing adults, Journal of the Acoustical Society of America 77 (2), February, 1985

[19] Pandey, P. C., Kunov, H., Abel, S. M. Disruptive effects of auditory signal delay on speech perception with lipreading, Journal of Auditory Research, Jan. 26 (1), 1986

[20] Massaro, D. W., Cohen, M. M., and Smeele, P. M. T. Perception of asynchronous and conflicting visual and auditory speech, Journal Acoustical Society of America, 100 (3), Sept., 1996

[21] Frowein, H., Smoorenburg, G., Pyters, L., and Schinkel, D. Improved speech recognition through videotelephony: Experiments with the hard of hearing, IEEE Journal on Selected Areas in Communications, 9, 1991

[22] Vitkovitch, M., Barber, P. Effect of Video Frame Rate on Subjects' ability to Shadow One of Two competing verbal Passages, Journal of Speech and Hearing Research, Volume 37, October, 1994

[23] Knoche H., de Meer, H., Kirsh, D. Compensating for Low Frame Rates, Extended abstracts of CHI 2005, 4-7 April, Portland, OR, USA, 2005