ISCA Archive
http://www.isca-speech.org/archive

Auditory-Visual Speech Processing
(AVSP) 2011
Volterra, Italy
September 1-2, 2011

On the Audiovisual Asynchrony of Speech

László Czap

Department of Automation and Communication-Technology,
University of Miskolc, Miskolc, Hungary
czap@mazsola.iit.uni-miskolc.hu

Abstract

The temporal synchrony of auditory and visual signals is known to affect the perception of audiovisual speech. Several papers have discussed the asymmetry of acoustic and visual timing cues. These results are usually based on subjective intelligibility tests and the reason is remained obscure. It is not clear that the observation is perception or production origin. In this paper the effect of audio-visual asynchrony is studied in an automatic bimodal speech recognition task, eliminating the perception expertise of observers. Results are utilized to improve naturalness of audiovisual speech synthesis.

Index Terms: speechreading, multimodality, audiovisual asymmetry

1. Introduction

Human speech perception is highly influenced by vision. Watching the speaker's mouth movements can significantly improve intelligibility, both for normal listeners in noisy environments and especially for the hearing impaired. When perceiving audiovisual speech, subjects tolerate visual leading asynchronies, but are very sensitive to auditory leading asynchronies that are less likely to occur in natural speech. It might be explained by human evolution in a world where sound travels more slowly than light. Taking into account the 330 m/s sound velocity, less than 11 meters distance means one frame acoustic signal delay of NTSC video (13.2 meters for PAL TV system). With sound amplifiers and video projectors we can get practice in comprehension of audiovisual speech, time shifted with a number of frames. This suggests a perception reason, as in natural environment the sound can be late to the vision when the listener is further away from the speaker.

On the other hand, it is a well known practical observation that articulation movement precedes the speech production, so facial movements always occur before acoustic speech signal. It implies an argument for a production cause.

2. Related works

Early studies of McGrath and Summerfield investigating of audiovisual integration and asynchrony have focused on conditions where the visual signal precedes the audio signal [1]. Audio-visual identification of sentences was measured as a function of audio delay in untrained observers with normal hearing. The soundtrack was replaced by rectangular pulses originally synchronized to the closing of the talker's vocal folds and then subjected to delay. It was concluded that most observers, whether good lipreaders or not, possess insufficient sensitivity to intermodal timing cues in audio-visual speech for them to be used analogously to voice onset time in auditory speech perception. The results of experiments imply that delays of up to about 40 ms introduced by signal-processing algorithms in aids to lipreading should not materially affect audio-visual speech understanding. Grant and Greenberg [2] demonstrate that the ability to integrate auditory and visual information under conditions of bimodal asynchrony is highly dependent on whether the audio signal leads or lags the video. When the audio leads, audiovisual integration declines steeply for even small onset asynchronies, similar in pattern to intelligibility under conditions of spectral asynchrony for acoustic-only signals. When the video leads the audio, A/V integration remains relatively stable for onset asynchronies as long as 200 ms, similar in pattern to those observed for a variety of speech materials.

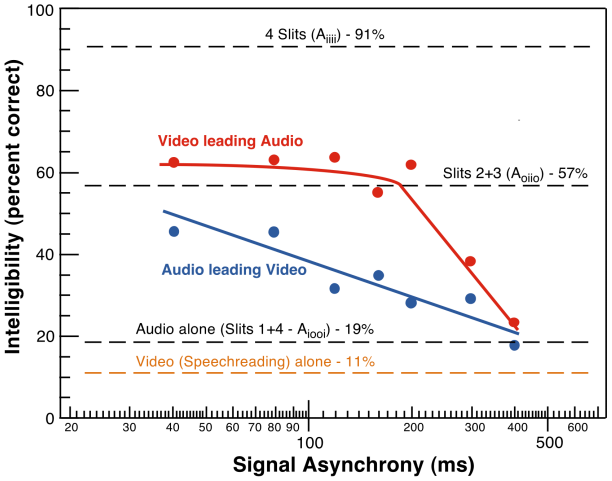

Figure 1. *Average intelligibility (for 9 subjects) associated with audio-visual speech recognition as a function of bimodal signal asynchrony. The audio-leading-video conditions are marked in blue, the video-leading audio conditions shown in red. Baseline audio-only conditions are marked in black, dashed lines, and the video-alone condition is shown in orange.[2]*

In their contribution Grant and Greenberg seek the explanation for the asymmetry in perception as bimodal integration pertains to the level of analytical abstraction associated with audio- and video-alone signals. When the audio signal leads the video, the time constants germane to speech processing may be relatively short, on the order of 40-120 ms, and pertains principally to the articulatory dimensions of voicing and manner, which are largely derivable from the acoustic signal alone. When the video signal leads the audio, the time constant for integration is likely to be longer, as the visual modality contains information particularly pertinent to place-of-articulation cues evolving over syllabic intervals of roughly 200 ms. The modality that leads in the asynchronous conditions may thus determine the level of abstraction (and

hence the time constant) over which bimodal processing proceeds.

Other interpretation of the same results can be seen in Figure 2.

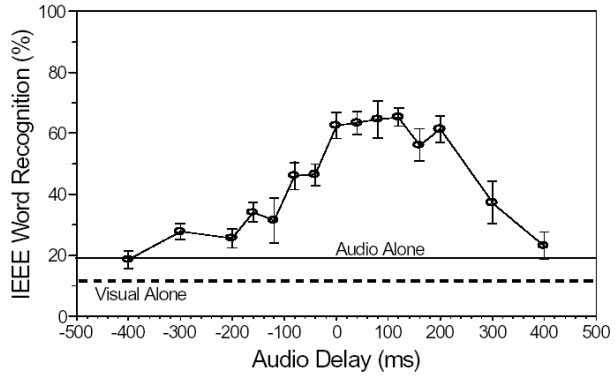

Figure 2. *Average auditory-visual intelligibility of IEEE sentences as a function of audio-video asynchrony. Note the substantial plateau region between -50 ms audio lead to 200 ms audio delay where intelligibility scores are high relative to the audio-alone or video-alone conditions. [3]*

There are reasonable grounds for suspecting that the audiovisual asymmetry originates in perception:

Grant et al. [3] demonstrate that the asynchrony across modality is true regardless of whether the subject's task is to recognize and identify words or syllables, or to simply discriminate which of two auditory-visual speech inputs is synchronized. This suggests that as soon as auditory-visual asynchrony is detected, the ability to integrate the two sources of information declines. They note that this is a fairly unusual outcome in psychological tests where the limits of sensitivity to a particular stimulus property (as measured by detection and discrimination) coincide with its use in higher-order decisions (e.g., recognition and identification).

Brungart et al. [4] state that one of the components that might influence the dynamics of AV speech integration is the amount of time needed for the sensory system to process the audio and visual portions of the multimodal stimulus. Neurophysiological evidence from the Macaque suggests that visual feedback information reaches the auditory cortex roughly 50 ms poststimulus, while auditory feedforward information reaches the auditory cortex roughly 11 ms poststimulus. Thus, one might expect sensory information to reach the auditory cortex simultaneously when the visual signal leads the audio signal by roughly 40-50 ms.

By Levitin et al. [5] one of the oldest questions in experimental psychology concerns perception of simultaneous events, particularly when input arrives through different sensory channels (sight / sound or touch / sound). How far apart in time must two events be to be perceived as sequential? Their paper reports preliminary data from a cross-modal simultaneity task designed for ecological validity. Results indicate a smaller threshold for successiveness than that found in previous experiments, which used more artificial tasks. They consider their findings relevant to theories of time, order and perception.

Hay-McCutcheon et al. [6] found from their first examination of the audiovisual integration skills of individuals who use cochlear implants suggest that aging has a greater effect on the detection of AV asynchronous speech than a severe-to-profound hearing loss that has been partially corrected through the use of a cochlear implant. Additionally, the temporal width of the AV asynchrony function was not correlated with speech perception skills for hearing-impaired individuals who use cochlear implants. However, when exploring the relationship between AV asynchrony detection and speech perception skills, the results suggest that middle-aged and elderly individuals might process auditory and visual speech cues differently in a range of word and sentence perception tasks.

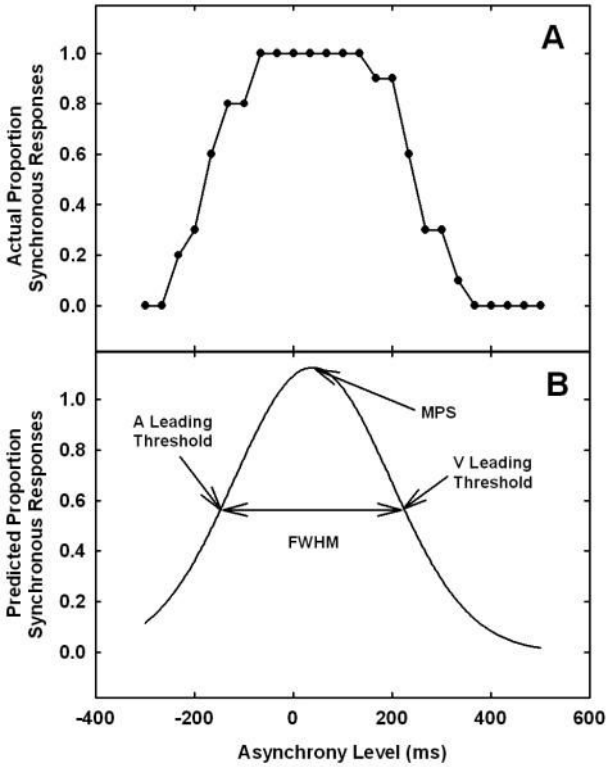

Figure 3. *An individual AV asynchrony function. Panel A displays the observed function and Panel B shows the Gaussian curve fitted to the observed function. The proportion of synchronous responses is shown as a function of the asynchrony level. The A-leading threshold is the asynchrony level for the y value at 50% of the distance from the minimum to the maximum of the auditory leading portion of the curve. The V-leading threshold is the asynchrony level for the y value at 50% of the distance from the maximum to the minimum of the visual leading portion of the curve. The FWHM is the value of the asynchrony width of the half-maxima of the function. Also displayed is the mean point of synchrony (MPS).[6]*

3. Method

Audiovisual recognition of sentences was measured as a function of the delay of auditory and visual modality in an automatic HMM based recognition task.

3.1. The database

The audiovisual database consists of full-face frontal video and audio of 536 words and phrases of a single speaker. The basic element of recognition is the *diphone*. 163 sort of diphones can be found in the database, at least five times each. 35 long sentences of connected digits and date serves testing purposes, containing 1243 diphones.

3.2. Visual preprocessing

The interlaced scan of PAL video refers to the common method for "painting" a video image on an electronic display screen by scanning or displaying each line or row of pixels. This technique uses two fields to create a frame. One field contains all the odd lines in the image, the other contains all the even lines of the image. A PAL-based television display scans 50 fields every second (25 odd and 25 even). The two sets of 25 fields work together to create a full frame every 1/25th of a second, resulting in a display of 25 frames per second, but with a new half frame every 1/50th of a second. Decomposing each frame to fields, we can get an image of the speaker in every 20 milliseconds. Lip contour geometric features of width (a), height (b) of inner lips and the average brightness of oral cavity (k), had driven by image processing represent the articulation characteristics.

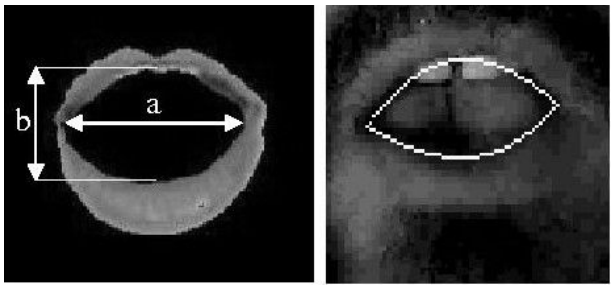

Figure 4. *Features of visual preprocessing. a: width, b: height of inner lips, k: average intensity of oral cavity*

3.3. Acoustic feature extraction

It is self-evident to choose the sound time shift equally with the video field interval. The 22050 Hz sampling rate yields 441 samples during a 20 ms space of field time. An audio frame consists of 512 samples, so the sound intervals have 16% overlaps. Commonly used MFCC feature vectors of the 512 samples represent the acoustic signal.

3.4. The experiment

Audiovisual recognition of sentences was measured as a function of the delay between the acoustic and visual signal. The time step of audiovisual delay is also 20 ms from -400 ms to 400 ms. The Acoustic and visual features are combined by early integration, the traditional concatenation of the audio and visual features as the joint audiovisual feature vector, and an HMM based recognition system is trained on the basis of them.

4. Results

Diphone recognition results indicate that performance declines monotonically and asymmetrically with delay.

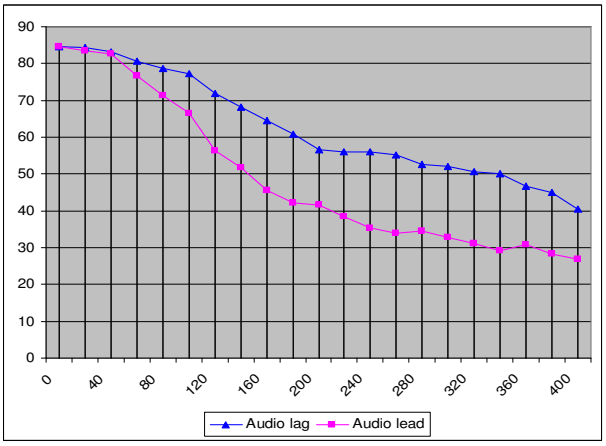

Figure 5. *Declining of recognition rate. Function of audio lead and lag (ms).*

Figure 6. illustrates the same results in an other interpretation showing the audiovisual asymmetry obviously.

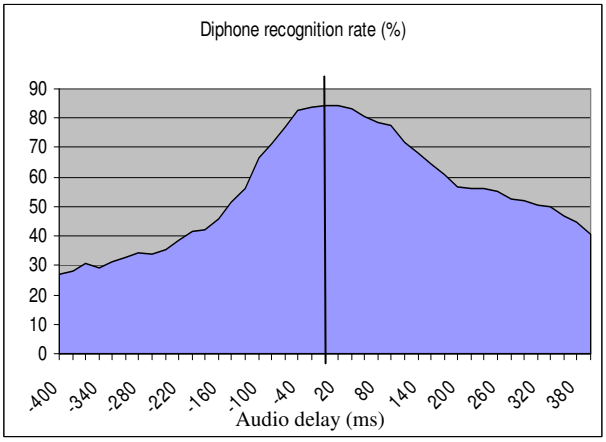

Figure 6. *Diphone recognition rate. (Function of audio delay. Negative values mean audio leading.)*

Based on the audiovisual asymmetry observed in the diphone recognition study significant improvements have been achieved in our audiovisual speech synthesis system. The developments focused on the following aspects:

- Pre-articulation. Prior to an utterance, a silent period is inserted – imitating breathing by opening the mouth – then the first dominant viseme is moved from the neutral starting position.
- Realizing the temporal asynchrony effect. A filtering and smoothing algorithm has been developed for adaptation to the tempo of either the synthesized or natural speech.

Pre-articulation: Prior to an utterance, there is an approximately 300 ms silence period is inserted designed to imitate breathing through the mouth – then the first dominant viseme is moved from the neutral starting position. Because of this pre-articulatory movement, the sound is produced as in natural speech.

Adapting to the tempo of speech and filtering: During the synchronization to human or synthesized speech, we have encountered different speech tempos. When speech is slow, viseme features approach their nominal value, while fast speech is articulated with less precision in natural speech. For flexible features, the round off is stronger in fast speech. A

median filter is applied for interpolation of flexible features: the values of neighbouring frames are sorted and the median is chosen. A feature is formed by the following steps:

- linear interpolation among values of dominant and flexible features, neglecting uncertain ones,
- median filtering is performed when flexible features are juxtaposed,
- values are then filtered by the weighted sum of the two previous frames, the actual and the next one.

The weights of the filter are fixed, so knowledge of speech tempo is not needed. The smoothing filter refines the movements and reduces the peaks during fast speech. By considering the two previous frames, the timing asymmetry of articulation is approximated, this way takes filtering function the phenomenon into account. Other improvements – as inserting a permanent phase into long vowels and synchronising phases of a viseme to a phoneme at several points – refine the articulation.

Figure 7. depicts the effect of median filtering and smoothing. In this example, the slow speech has twice as many frames as the fast one. The horizontal axis shows the number of frames, while vertical values represent the amplitude of the feature.

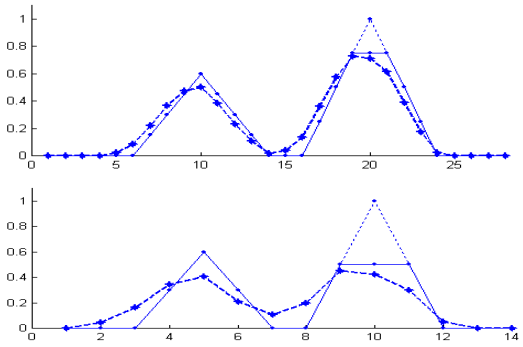

Figure 7. The interpolation of dominant (first peak) and flexible (second peak) features for slow (above) and fast (below) speech after linear interpolation (…), median filtering (___) and smoothing (---). Smoothing puts the visual lead into effect.

5. Discussion

Results coincide with those of Feldhoffer et al. [8] by studying mutual information of audiovisual features. They have analysed the fine temporal structure relations of acoustic and visual features to improve speech to facial animation conversion system. Mutual information of acoustic (MFCPCA) and visual (FacePCA) features has been calculated with different time shifts. Their result shows that the movement of feature points on the face of professional lip-speakers can precede the changes of acoustic parameters even by 100 milliseconds. Considering the measured time-shifts in synchrony in system design, the quality of speech driven animations can be improved.

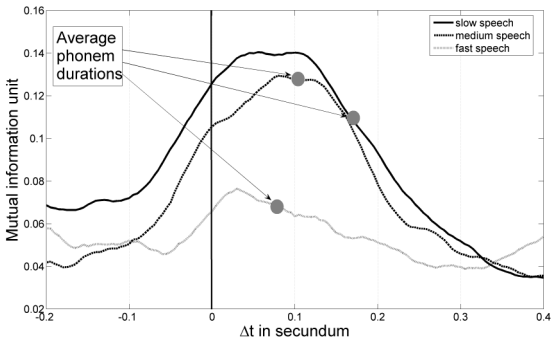

Figure 8. *Shifted FacePCA1 and MFCPCA mutual information. Positive t means voice lag [8].*

Speech tempo related effects also have been reported in this study as Brungart et al did it in [4].

6. Conclusion

Audiovisual asynchrony findings reported here are similar to those previously contributed based on audiovisual perception experiences. Eliminating the perception skills of subjects makes it possible to conclude that the asynchrony between acoustic and visual modality is presumably arisen from speech production. Contrary to the findings of perception based contributions, these results are free from audiovisual expertise and experience of listeners. The high correlation with the perception based results suggests production origin of those observations as well.

7. Acknowledgement

This research was carried out as part of the TAMOP-4.2.1.B-10/2/KONV-2010-0001 project with support by the European Union, co-financed by the European Social Fund.

8. References

[1] McGrath, M, Summerfield, Q., “Intermodal timing relations and audio-visual speech recognition by normal-hearing adults”, Journal of the Acoustical Society of America, 678–685, 1985

[2] Grant, K. W., and Greenberg, S., "Speech intelligibility derived from asynchronous processing of auditory-visual information", AVSP 2001, 132-137, Scheelsminde, Denmark, September 7-9, 2001.

[3] Grant, K.W., van Wassenhove, V., Poeppel, D., “Discrimination of Auditory-Visual Synchrony” AVSP 2003, 31-35, Jorioz, France, 2003

[4] Brungart, D. S., Iyer, N, Simpson, B. D, van Wassenhove, V., “The effects of temporal asynchrony on the intelligibility of accelerated speech” AVSP 2008, Moreton Island, 19-24, Australia, 2008

[5] Levitin, D. J., Mathews, M. V., and MacLean, K., “The Perception of Cross-Modal Simultaneity” International Journal of Computing Anticipatory Systems, 323-329, Belgium, 1999

[6] Hay-McCutcheon, M. J., Pisoni, D. B., and Hunt, K. K.: “Audiovisual Asynchrony Detection and Speech Perception in Hearing-Impaired Listeners with Cochlear Implants: A Preliminary Analysis”, Int J. Audiol. 48(6): 321–333, 2009,

[7] Grant, K. W., Greenberg, S., Poeppel D, van Wassenhove, V., “Effects of spectro-temporal asynchrony in auditory and auditory-visual speech processing” Seminars in Hearing, 3:241–255, 2004

[8] Feldhoffer, G., Bárdi, T., Takács, G., Tihanyi, A., “Temporal Asymmetry in Relations of Acoustic and Visual Features of Speech”, 15th European Signal Processing Conf., 2341-2345, Poznan, Poland, September 2007