Multimodal Interaction and Perception

Armin Kohlrausch
TU/e
Philips Research Europe

TU/e Technische Universiteit Eindhoven University of Technology

Where innovation starts

Overview of presentation

- Explanation/definition of terms
- Examples of multimodal interaction
- Example of multimodal perception: Interaction between audio-visual stimuli

Explanation of the terms

- Multimodal interaction: Interacting with a system using more than one (dominant) modality
  - Multimodal interfaces
- What defines a modality in this term?
  - The various sensory systems/channels (visual, auditory, haptic as the most dominant ones)
  - In less formal terms, a modality is a path of communication between the human and the computer (this would then also include brain-computer interfaces)
- This area is strongly embedded in HCI research, Human Computer Interaction

Explanation of the terms II

- Interaction implies receiving information from the system (output modalities) and providing input to the system (input modalities)
- Input and output modalities do not have to be the same
  - Classical computer interface: Output via display (visual), input via keyboard/mouse (haptic)
  - Speech is an example of using the same modality for input (requires automatic speech recognition) and output (requires speech synthesis)
- Be careful: The term interaction is also used with a different meaning

Meaning of the terms III

- Multimodal perception (I prefer multisensory perception): The study of combined stimulation of more than one sensory system.
  - In daily life, this is the default situation for our perception
- "Interaction" in this context refers to interactions between modalities: Does a stimulus provided in modality 1 affect the percept in modality 2
  - Perceptual illusions, good example for audio-visual interaction is the ventriloquism effect, or the McGurk effect in speech

Reasons to study multimodal interaction

- **Multimodal interaction has the promise to increase usability**
- Weakness of one modality can be compensated for by another modality
  - Weakness can be a general property, e.g., you don't have eyes at your back
  - Or it is defined by the context (use of a display in sunlight)
- Can we increase the bandwidth (in a technical sense) of the interaction by using modalities in parallel?

Examples

- If it is unclear where your focus of attention is, a sound is a much better alarming stimulus than a light flash (e.g. in control rooms)
- If your hands are busy (for a doctor during a medical intervention; for yourself in the kitchen), speech as input modality has great advantages
- If you attend a winterschool, and you expect an important phone call, vibration is a good modality for signaling

Examples II

- Regarding sensory bandwidth effects
- Positive effect: It is known that speech perception can be improved by matched visual and acoustic representation (at least for difficult acoustic situations)
- Negative effect: It is known that having a phone conversation during driving (requires visual attention) leads to increased reaction times (also for hands-free conditions)

Examples III

- A special case of specific contexts of use are interfaces designed for people with reduced perceptual and motor abilities
  - often referred to as accessibility
  - Sensory substitution
- Speech and braille output for blind users
- See with your Ears!
- Image to sound transformation system, in particular for the blind; by Peter Meijer
- http://www.seeingwithsound.com/
- winvoice.htm

Example from research project at TUe

- SATIN: Sound And Tangible Interfaces for Novel product design (http://www.satin-project.eu/)
- The main objective of the project is to develop a new generation of multimodal and multisensory interfaces, supporting free-hand interaction with virtual shapes
- Based on fusion of force feedback, sound and vision for representing global and local properties of shape and material
- Our contribution: Use data sonification (e.g., of curvature values) to represent object shape properties which cannot easily be perceived visually or haptically
- Research question: Are end-users capable of interpreting the sound and perceive it as an intuitive addition to the other information channels

Example of multisensory perception: Interaction in audio-visual stimuli

- Spatial disparity
- Temporal and rhythm disparity

Spatial disparity between audio and video
(ventriloquism)

- Research question: What do we perceive when audio and video are presented from different positions (directions)?
- Application areas: Audio-visual reproduction systems, video conferencing systems

From Stein and Meredith: The merging of the senses (1993)

Figure 1.1 The ventriloquism effect. The ventriloquist "throws his voice" by minimizing his own movements so that the only visual cues the audience can associate with speech come from the dummy. This says less about the ventriloquist's skill than about how strong visual-auditory intersensory biases are in the audience.

Spatial disparity II

- Basic observation: The perceived direction of the sound is influenced through the simultaneous presentation of the visual stimulus
- Capturing effect for horizontal angle differences of 30 to 40 degrees
- For elevation differences: up to 55 degrees
- Audio and video have to be related (that's the art of the ventriloquist)

Discussion

- Possible explanation: Visual localization is more accurate than auditory localization. In the case of discrepant information, more weight is given to the visual stimulus.
- In the case of temporal variations, more weight is given to the audio signal (see next topic/demo).

Discussion (II)

- Supported by recent data (Alais and Burr, The Ventriloquist Effect Results from Near-Optimal Bimodal Integration Current Biology, Vol 14, 257-262, 3 February 2004)
- Alais and Burr found clear evidence for visual dominance when the visual stimulus was sharply focused, and for auditory dominance when the visual stimulus was blurred. In both cases, the observed percepts were very close to those predicted by the optimal combination rule.

Example: Auditory dominance

Illusions

What you see is what you hear

Vision is believed to dominate our multisensory perception of the world. Here we overturn this established view by showing that auditory information can qualitatively alter the perception of an unambiguous visual stimulus to create a striking visual illusion. Our findings indicate that visual perception can be manipulated by other sensory modalities.

Ladan Shams*, Yukiyasu Kamitani*, Shinsuke Shimojo*†

**California Institute of Technology, Division of Biology, MC 139-74, Pasadena, California 91125, USA e-mail: ladan@caltech.edu*

†NTT Communication Science Laboratories, Human and Information Science Laboratory, Atsugi, Kanagawa 243-0198, Japan

788 NATURE | VOL 408 | 14 DECEMBER 2000 | www.nature.com

Example: Auditory dominance

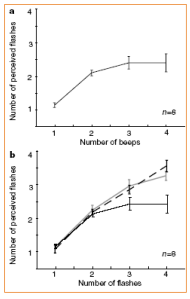


Figure 1 Illusory flashing. **a,** Perceived number of visual flashes by eight observers plotted as a function of the number of auditory beeps for a single flash. The number of perceived flashes did not increase linearly with the third and fourth beeps because they fell outside the optimal window of audiovisual integration, as revealed by our next experiment.

b, Perceived number of flashes by eight observers plotted as a function of the actual number of flashes presented for trials with no sound (dashed line), and trials with single beeps corresponding to catch trials (grey line). Observers performed the task very well in the absence of sound (dashed line). The results of the catch trials (grey line) confirm that the observers' responses were not determined by their auditory percepts. The curve in **a** (for a single flash) is superimposed for comparison.

Demo (a more magical version than a double flash, a quintuple rabbit)

- Kamitani, Y. & Shimojo, S. (2001) Sound-induced visual "rabbit". Journal of Vision (abstract)
- WEB source:
  http://www.cns.atr.jp/~kmtn/audiovisualRabbit/index.html

Demo visual rabbit

The relation with film design

Research result

Therefore, taken together, research now shows that auditory stimuli can dominate both the perceived rate and time-of-occurrence of associated visual stimuli, while visual stimuli dominate the more spatial aspects of multisensory perception

Spence, Squire, 2003

Application in audio-visual media

In the course of audio-viewing a sound film, the spectator does not note these different speeds of cognition as such, because added value intervenes. Why, for example, don't the myriad rapid visual movements in kung fu or special effects movies create a confusing impression? The answer is that they are 'spotted' by rapid auditory punctuation, in the form of whistles, shouts, bangs, and tinkling that mark certain moments and leave a strong audiovisual memory.

CHION, MICHEL (1994): Audio-Vision. Sound on Screen. New York, Chichester: Columbia University Press.

Effects of asynchrony between auditory and visual stimuli

- Most of you will have experienced badly synchronized videos
- Asynchrony is particularly obvious in speech, e.g. in interviews recorded under difficult circumstances (news items on TV)
- There, the broadcasting industry has defined limits for tolerable AV delays

Asynchrony and quality degradation (Rihs, 1995)

- Material: 50 s video (Talk show with several speakers). AV delays: -200 ms to + 200 ms
- Quality judgments on 5 point impairment scale, 18 subjects

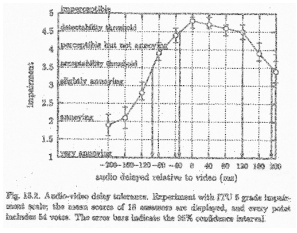
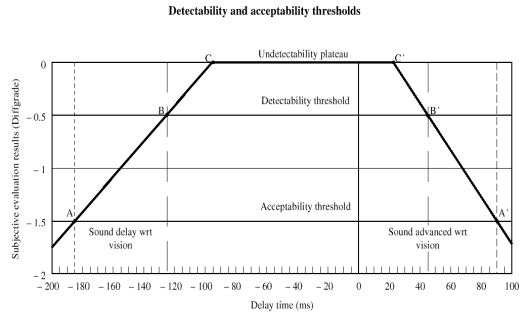


Fig. 13.2. Audio-video delay tolerance. Experiment with ITU 5 grade impairment scale; the mean scores of 18 assessors are displayed, and every point includes 54 votes. The error bars indicate the 95% confidence interval.

ITU recommendation (1998): Relative timing of Sound and Vision for Broadcasting



Take-home message

- Clear asymmetry in the sensitivity for audio and for video delays
  - We are perceptually more sensitive for asynchronies when the audio component is leading, compared to leading video
- Rules of multisensory perception (tolerances for asynchrony and spatial disparity, sensory substitution, multisensory integration) form necessary (and helpful) parameters for system specification, e.g. for multimodal interfaces
- When testing quality of multimedia systems, possible interaction effects have to be taken into account
- Never test video in isolation, i.e. without sound!

Asynchrony judgments for distant stimuli

- Question: Does the (a)synchrony judgment change if a stimulus is presented from a distance?
- Example: Stimulus 15 m. distance: Travel time for sound is ca. 45 ms, for light 0 ms.
  - We now have: Physical AV delay at the source: T_s
  - Physical AV delay at the human head T_h = T_s + 45 ms
- We know for usual stimuli the relation between perceived synchrony and physical synchrony: PSE ~ 50 ms
- When judging asynchrony for the above stimulus
  - Is PSE= 50 ms relative to T_s?
  - Or is it 50 ms relative to T_h ?
- In other words: Do we correct (cognitively) for the extra physical delay when we perceive distant stimuli?

Results for distant stimuli (I)

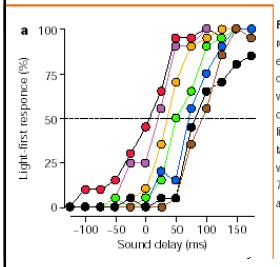
- Rudloff (1997): He compared quality judgments for two video scenes:
  - Object at short (5 m) distance
  - Object at distant (30 m) position (in the recording, the sound had about 100 ms delay)
- Quality judgments as a function of extra delay differed
  - For short distance, maximum quality at extra delay of 80 ms
  - For large distance, maximum quality at 0 ms extra delay
- Conclusion: No (complete) compensation for physical delay in stimulus

Results for distant stimuli (II)

- Stone et al. (2001): AV synchrony judgments for simple stimuli
  - Short distance (0.5 m)
  - Larger distance (3.5 m)
- The difference in sound travel time between the two conditions is 11 ms
- The PSE values obtained in the two conditions differed for 3 of the 5 subjects
- When 11 ms were subtracted (the difference between T_h and T_s), no significant difference remained.
- Conclusion by the authors: Observers *do not discount the effects of distance* when making judgments of simultaneity (what counts is the temporal relation at the head of the observer)

Results for distant stimuli (III)

- Sugita and Suzuki (2003) in NATURE
  - Short noise burst, presented over headphone (with HRTF for frontal incidence)
  - LED at distances between 1 and 50 m (real distance), intensity of the light source was increased with the square of the distance
- Subject did temporal order judgments
- Indicated times of sound delay correspond to times at the observers head (T_h)



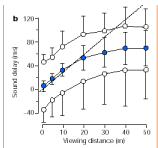


Figure 1 Synchrony in audiovisual perception. **a**, Representative results from one observer. The percentage of light-first response for each viewing distance is plotted against sound delay (stimulus-onset asynchrony). Different colours represent results for different viewing distances (red, pink, yellow, green, blue, brown and black correspond to 1, 5, 10, 20, 30, 40 and 50 m, respectively). Dashed line indicates the 50% point, which corresponds to subjective simultaneity. **b**, Points of subjective equality (filled circles) plotted against viewing distance. Hollow circles, plots of 25% (bottom curve) and 75% (top) of light-first response indicate the threshold for detecting asynchrony. Dashed line represents the real sound-arrival time.

Conclusion by Sugita and Suzuki

Our findings indicate that auditory and visual inputs are coordinated not because the brain has a wide temporal window for auditory integration, as was previously thought, but because the brain actively changes the temporal location of the window depending on the distance of the visible sound source.

We have shown ***that the brain takes sound velocity into account*** when integrating audiovisual information. The brain can therefore integrate audiovisual information over a wide range of temporal gaps, and correctly match sound and visual sources.

The latest data on this issue

- **Lewald and Guski (2004):**
- **Similar to the previous experiment**
- **Free field (outdoor).**
  - **Five loudspeakers at distances between 1 and 50 ms**
  - **White LED at each loudspeaker**
- **Simuli: 5 short bursts (noise, light) with a rate 1/sec**
- **NO compensation for decrease in intensity with distance!**
- **Delay times are given as timing at the source T_s**

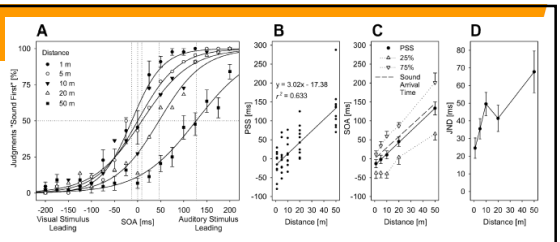


Fig. 1. Influence of distance on crossmodal temporal-order judgments. (A) Mean psychometric functions for all subjects. The percentage of 'sound first' responses is plotted as a function of the stimulus-onset asynchrony (SOA). For each distance (1, 5, 10, 20, and 50 m), data are fitted to a sigmoid equation. Dotted lines indicate the determination of the point of subjective simultaneity (PSS), which is defined as the SOA at which responses are 50%. Error bars (± SE) are shown only for distances of 1 and 50 m. (B) Individual PSSs from 11 subjects and the resulting regression line, plotted as a function of distance. (C) Mean PSS (closed symbols, ± SE) and the resulting regression line for all subjects, plotted as a function of distance. The dashed line indicates sound arrival times. Open symbols and dotted lines indicate the points of the psychometric functions at which the percentages of 'sound first' judgments were 25 and 75% (means for all subjects, ± SE). The half of the interval between these values is defined as the just noticeable difference (JND). (D) Mean JND (± SE), derived from the individual psychometric functions of all subjects, as a function of distance.

Conclusion by Lewald and Guski

Thus, in conclusion, the present data *clearly refute the hypothesis of a temporal compensation for sound-transmission delays,* **but rather support the view that our experience of perceptual simultaneity of auditory-visual events is based on neural processes of temporal integration.**

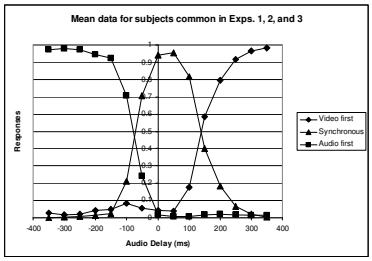
Discussion

- **Clear asymmetry in the sensitivity for audio and for video delays**
- **Typical explanation: Asymmetry reflects difference in arrival times of A and V for distant sources (a distance of 10 m corresponds to 30 ms sound propagation time)**
- **But: The asymmetry exists already for babies of 2 months of age**
- **PSE corresponds approximately to the difference in transduction time between peripheral sensors and first integrating neurons in the Superior Colliculus**
- **There might be a methodological issue in using the TOJ paradigm**
- **Need of more data studying the influence of object distance**

Braille display

- **A refreshable Braille display or Braille terminal is an electro-mechanical device for displaying Braille characters, usually by means of raising dots through holes in a flat surface.**
- **Those displays are commercially available (but not cheap), several 1000 EURO**

Results for disk falling down



PSE: +36 ms. Synchrony range: -76 ms …+148 ms