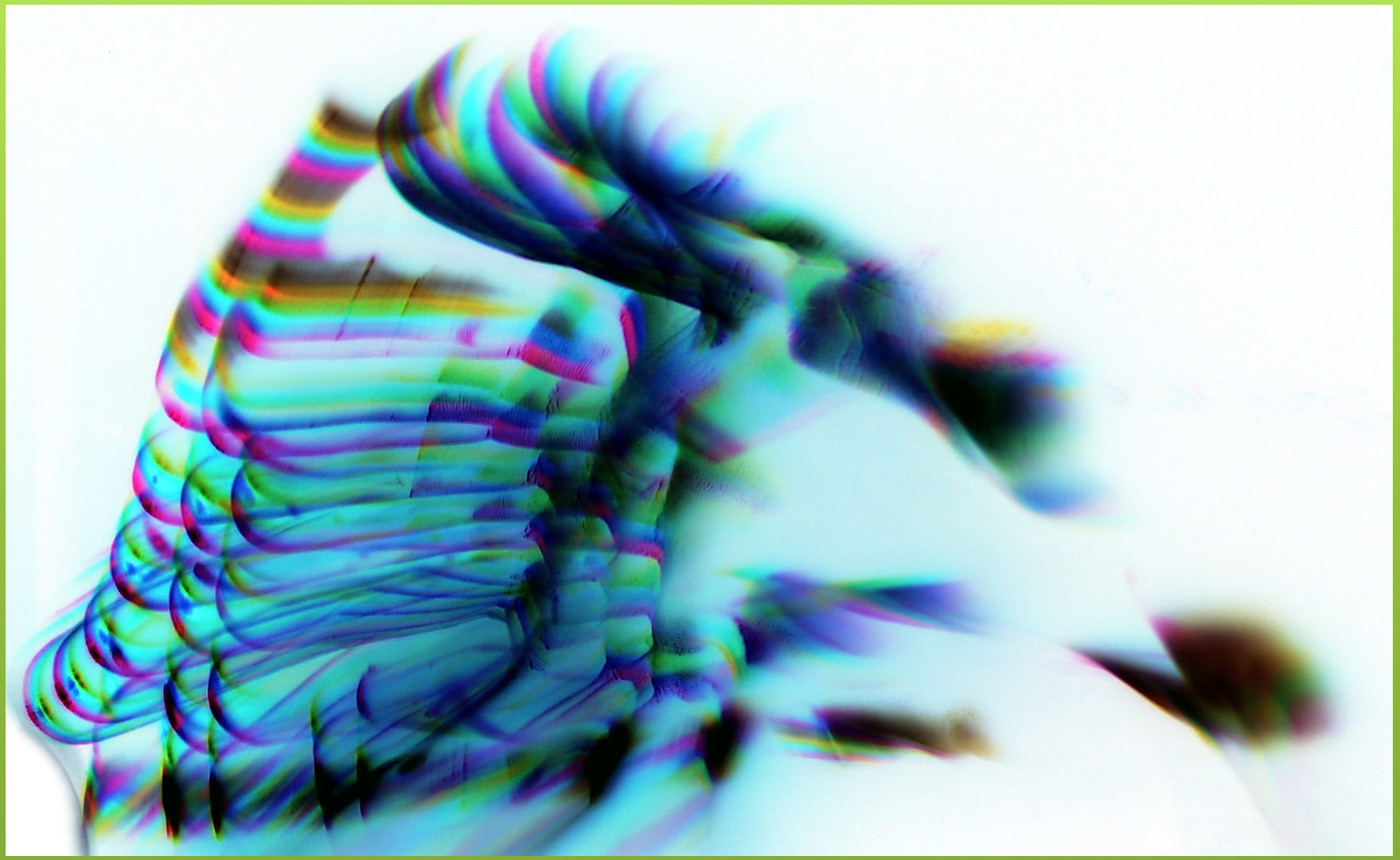
On the audiovisual integration of speech and gesture



Caro Kirchhof & Jan P. de Ruiter Bielefeld University ISGS5 Lund 2012

Overview

- background
  - motivation
  - Audiovisual Integration (AVI) of speech & gesture
- Study 1 – an online survey of perceptual judgment
- Study 2 – a user-specified synchronization experiment
- implications for theories on gesture and speech processing

Motivation

Some things we know about speech & gesture:

- semantic affiliation
  (e.g., Kendon 1972, 2004; McNeill 1985, 2005)
- temporal synchrony in production
  (e.g., Kendon 1980, 2004; McNeill 1985, 2005)
- listeners perceive co-speech gestures
  (e.g. Alibali et al. 2001; Holler et al. 2009)

What we don't know:

- How important is it **for the listener** that speech and gesture are synchronized?

Psychophysics of speech perception

- light travels faster than sound
- perception of audio-visual synchrony varies
  (e.g. Fujisaki & Nishida 2005; Nishida 2006)
- speech-lip asynchrony is perceived as unnatural
  (e.g. Vatakis et al. 2008; Feyereisen 2007)

Gesture & AVI

- gesture is perceived during discourse/ attracts attention
  (e.g. Gullberg & Holmqvist 2006)
- gestures 160 ms earlier than speech are integrated
  (Habets et al. 2011; Özyürek et al. 2007)

Summary so far

- Habets et al. (2011):
  - semantic congruency influences AVI
  - audio delay between 160ms and 360ms acceptable

- Psychophysics research on auditory delay:
  - 200ms: “asymmetric bimodal integration window“ (van Wassenhove et al. 2007)
  - 250ms: “boundary of AV integration” (Massaro et al. 1996)
  - 500ms: “significant breakdown“ in perceptual alignment” (Massaro et al. 1996)

Open Questions

- What about naturally co-occurring speech & gesture?
- Do we align speech & gesture in perception as in production?
- How large is the AVI-window in which speech and gesture are still recognized as co-expressive?
- What happens when **speech** comes first?
- Are there differences between perceptual judgment and preference?

Perceptual Judgment vs. Preference

- Study 1
  - online survey
  - 7 levels of speech-gesture asynchrony
  - 3 types of head-visibility
  - measured acceptability using 4-point Likert scale
- Study 2
  - 15 speech-gesture stimuli out of sync
  - 3 physical events out of sync
  - users requested to resynchronize stimuli using ELAN slider interface

Study 1 – Perceptual Judgment Guiding Questions:

- What is the acceptable range of speech-gesture asynchrony?
- Does the AVI break down when gesture precedes speech more than 200ms?
- Does AVI work when speech precedes gesture?

Material

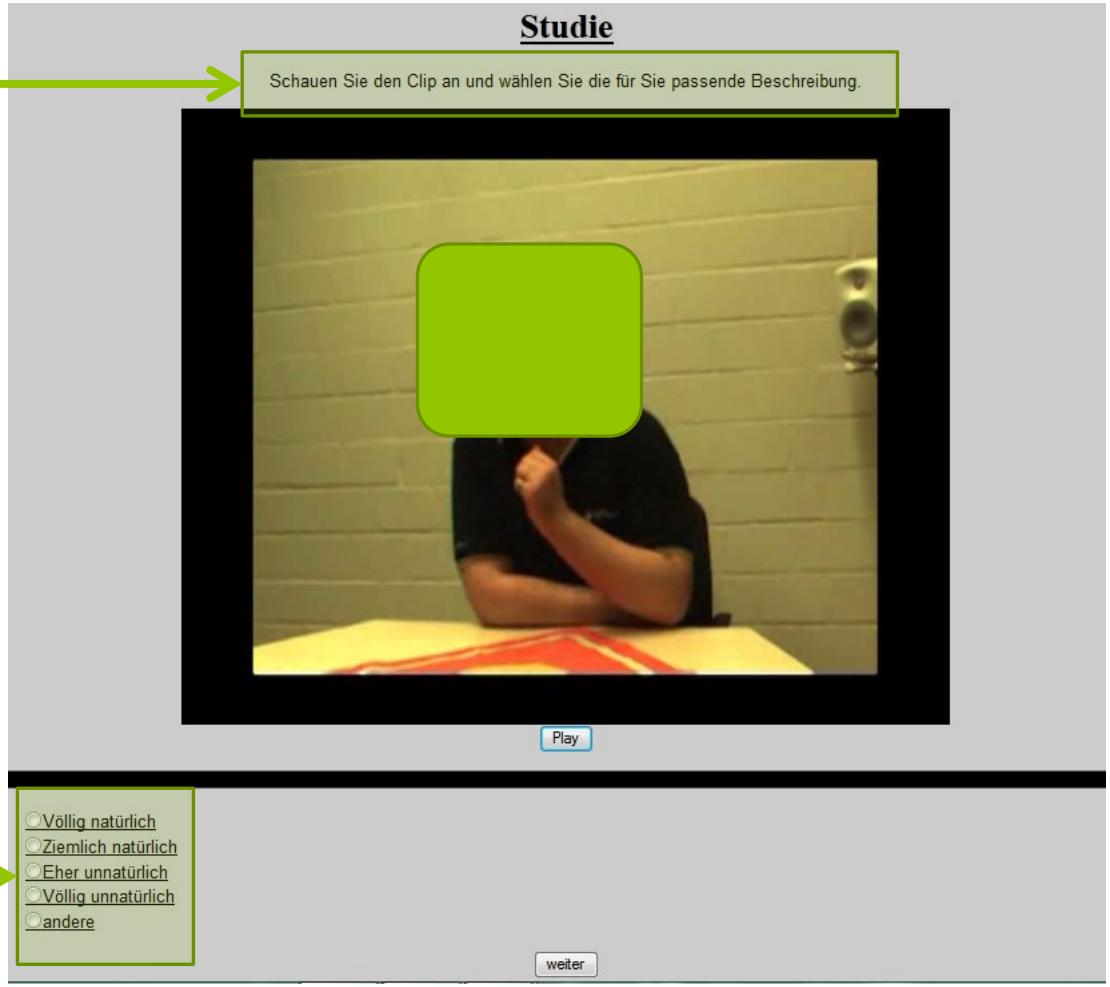
- 24 clips from naturalistic cartoon narrations:
  - one utterance long
  - accompanied by “large“ iconic gestures
  - original / head blurred / head blobbed (separate studies)
- AV-desynchronization:

  gesture first speech first
  - asynchronies of -600 -400 -200 **0** +200 +400 +600
- 168 stimuli to be rated for perceived naturalness (4-point Likert scale)

Design – Online Interface (blob)

Watch the clip and select the description most suitable to you.

- fully natural
- somewhat natural
- somewhat unnatural
- fully unnatural
- (other)



An example: Sylvester the Cat

200ms gesture advance

600ms audio advance

Subjects

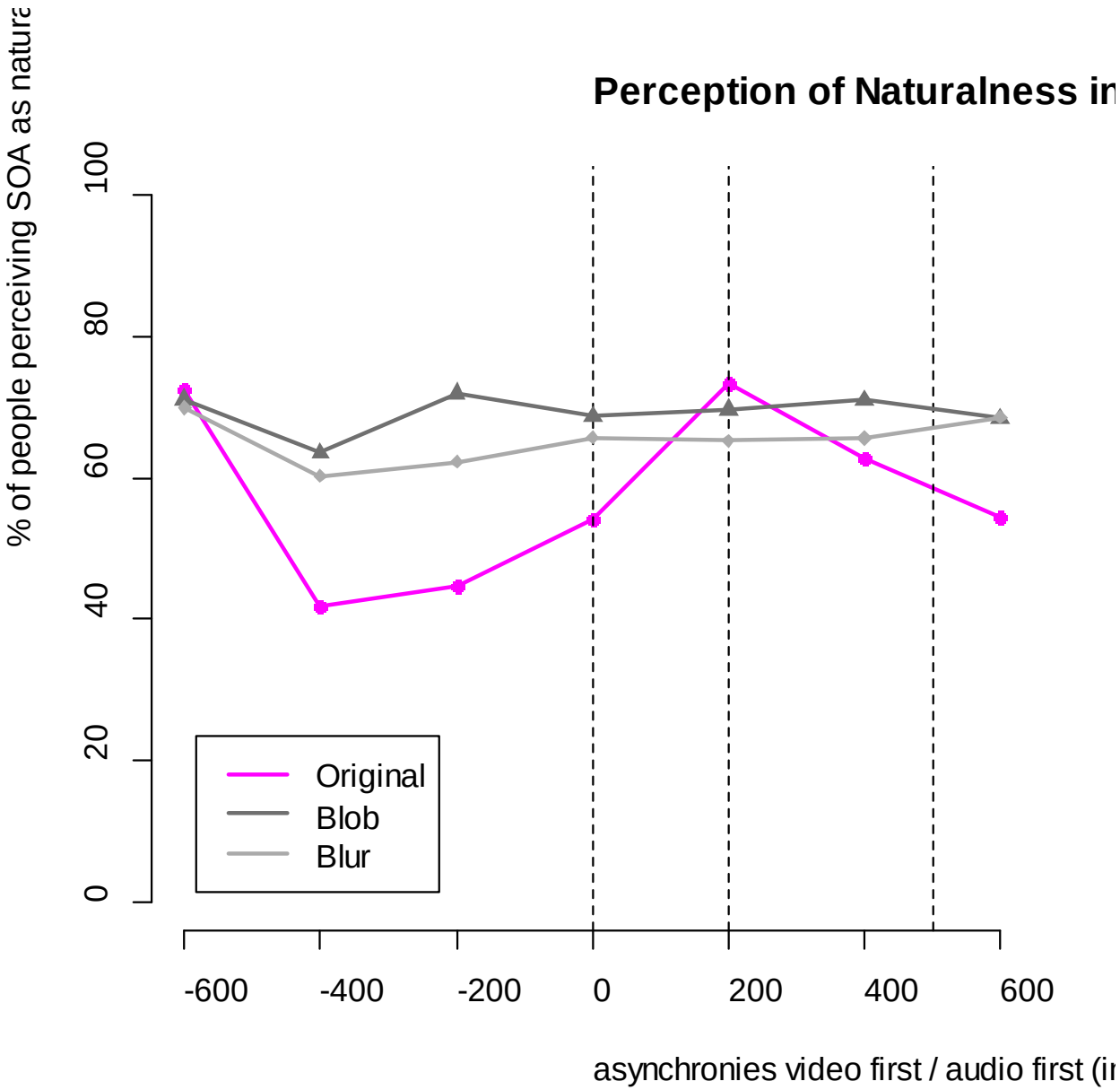
- all native speakers of German
- original:
  - 146 people age 16-73 (mean: 26)
  - 41 male, 115 female
- blurred faces:
  - 135 people age 15-67 (mean: 23)
  - 42 male, 93 female
- blocked heads:
  - 337 people age 17-67 (mean: 23)
  - 85 male, 252 female

Results

(percentages for „fully natural“ and „somewhat natural“ combined)

Gesture/speech first (ms)	Original	Blur	blob
-600	72,4	71,1	69,9
-400	41,7	63,5	60,2
-200	44,7	72	62,3
0	54	68,7	65,7
+200	73,5	69,6	65,2
+400	62,6	71	65,6
+600	54,4	68,4	68,6

- gesture advance of 600ms seems very acceptable
- “favorite” asynchrony varies across conditions
- acceptability ↔ head obscurity


Perception of Naturalness in
% of people perceiving SOA as natura
0
20
40
60
80
100
-600
-400
-200
0
200
400
600
asynchronies video first / audio first (i
Original
Blob
Blur

Partial Replication Study (in lab)

- Design:
  - 3x5 stimuli
  - gesture 600 ms before speech, 0 asynchrony, speech 200 ms before gesture
  - selection of most natural stimulus out of 3
  - original, blurred, blobbed
- Results:
  - lips visible: [-600]: 0%, [0]: 50%, [+200]: 50%
  - head obscured: random (approx. 33% each)

Discussion

- original lip-synchrony results largely replicated (in head-visible condition)
- for head-obscured conditions
  - >60% of people accepted -600 to +600ms

- Conclusion:

We need the speech to be synchronized with the lips, but not with the gestures.

But...

- Online studies may have low validity due to motivational factors.
- The maximal extent of the AVI-window for speech and gesture is still unclear.

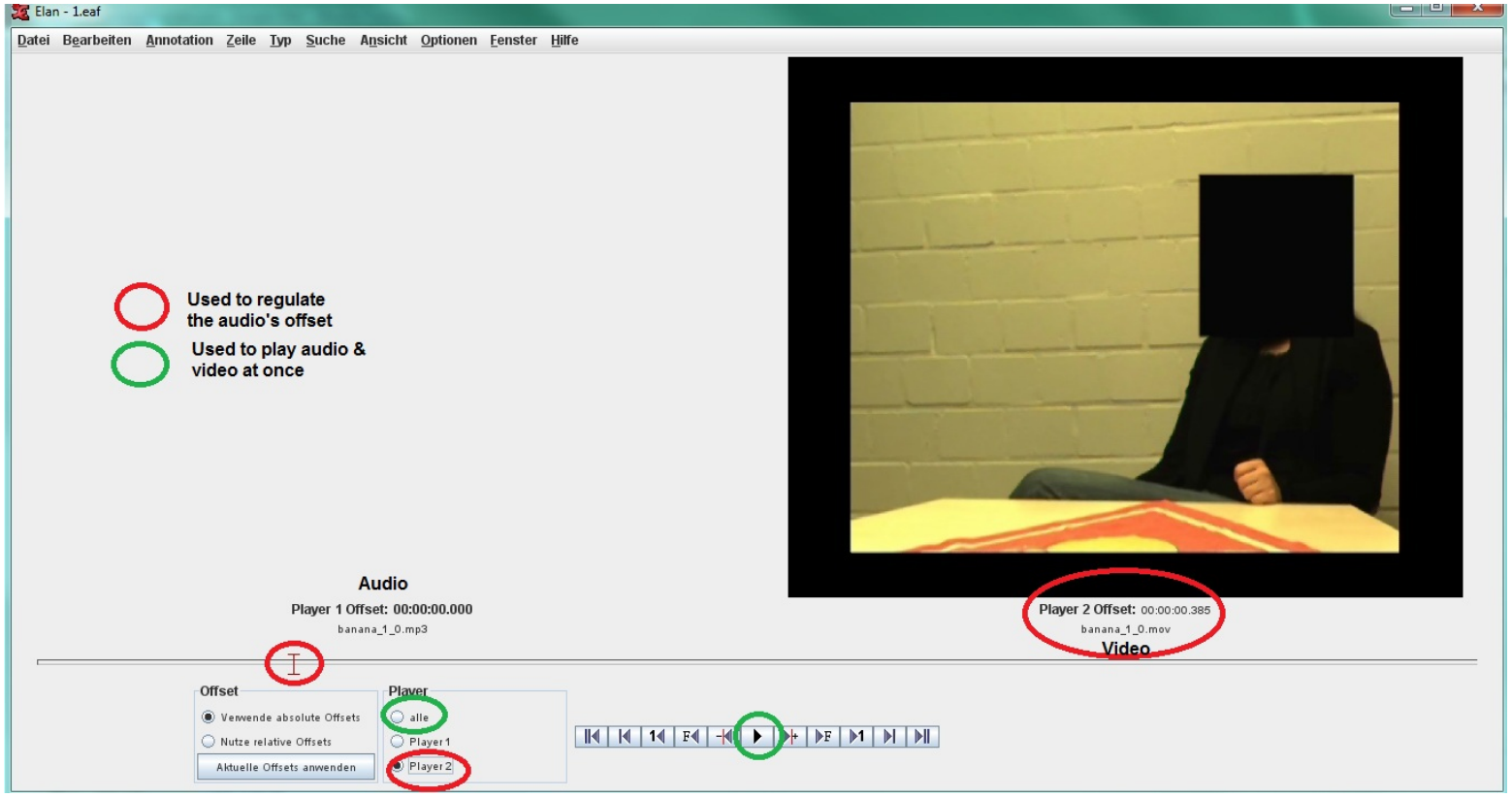
Study 2 – User-Specified Synchronization

- Will people produce the same range of asynchronies as in the perceptual judgment study?
- Or, will they choose a more restricted window?

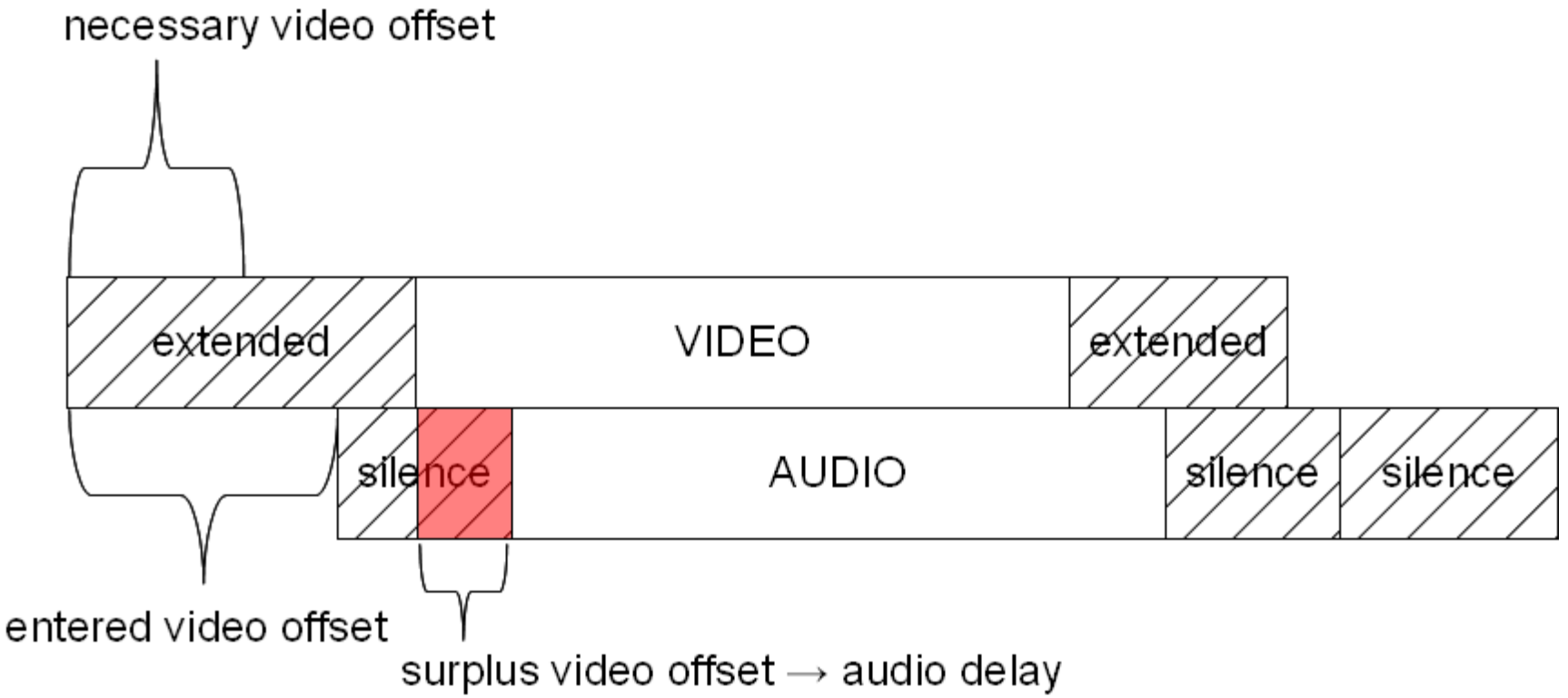
Design

- 18 stimuli:
  - 15 iconic gestures from Study 1 w/ blob with
  - 5 pseudorandom initial asynchronies
  - Baseline: 3 “physical events” (Hammer, Ball, Snap) w/ 902ms video advance
- a slider-interface (ELAN)
- 20 participants
- 300 manipulated stimuli

Interface


Elan - 1.eaf
Datei Bearbeiten Annotation Zeile Typ Suche Ansicht Optionen Fenster Hilfe
Used to regulate the audio's offset
Used to play audio & video at once
Audio
Player 1 Offset: 00:00:00.000
banana_1_0.mp3
Player 2 Offset: 00:00:00.385
banana_1_0.mov
Video
Offset
Verwende absolute Offsets
Nutze relative Offsets
Aktuelle Offsets anwenden
Player
alle
Player1
Player2

Example Video Offset for Slider



Subjects

- 14 female, 6 male
- mean age 25
- German mother tongue
- university students
- 2 left-, 18 right-handed

Results – Physical Events

- snap & hammer stimuli:

- audio first: 21/40
- video first: 19/40

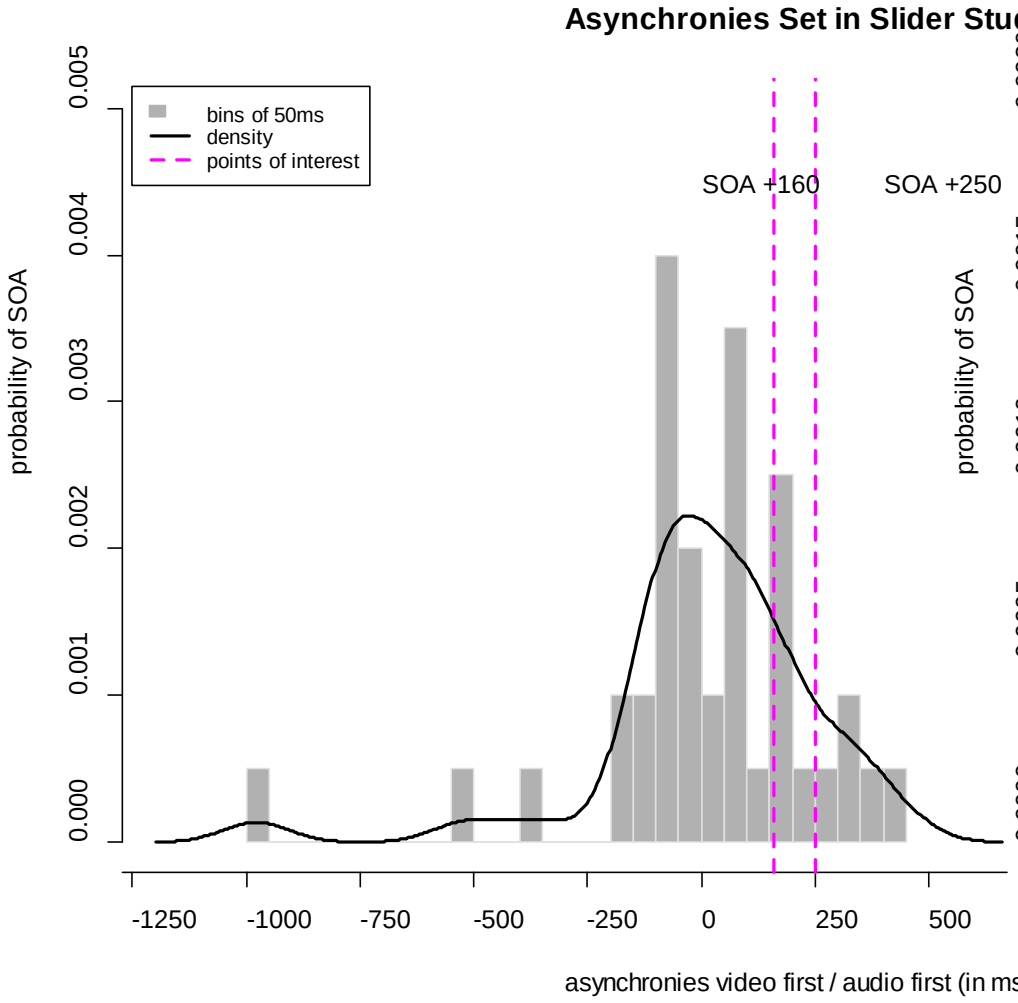
- SOA range:|
  -978ms (gesture first) to +442ms (speech first)
- SOA mean: +14 ms (stddev 246)

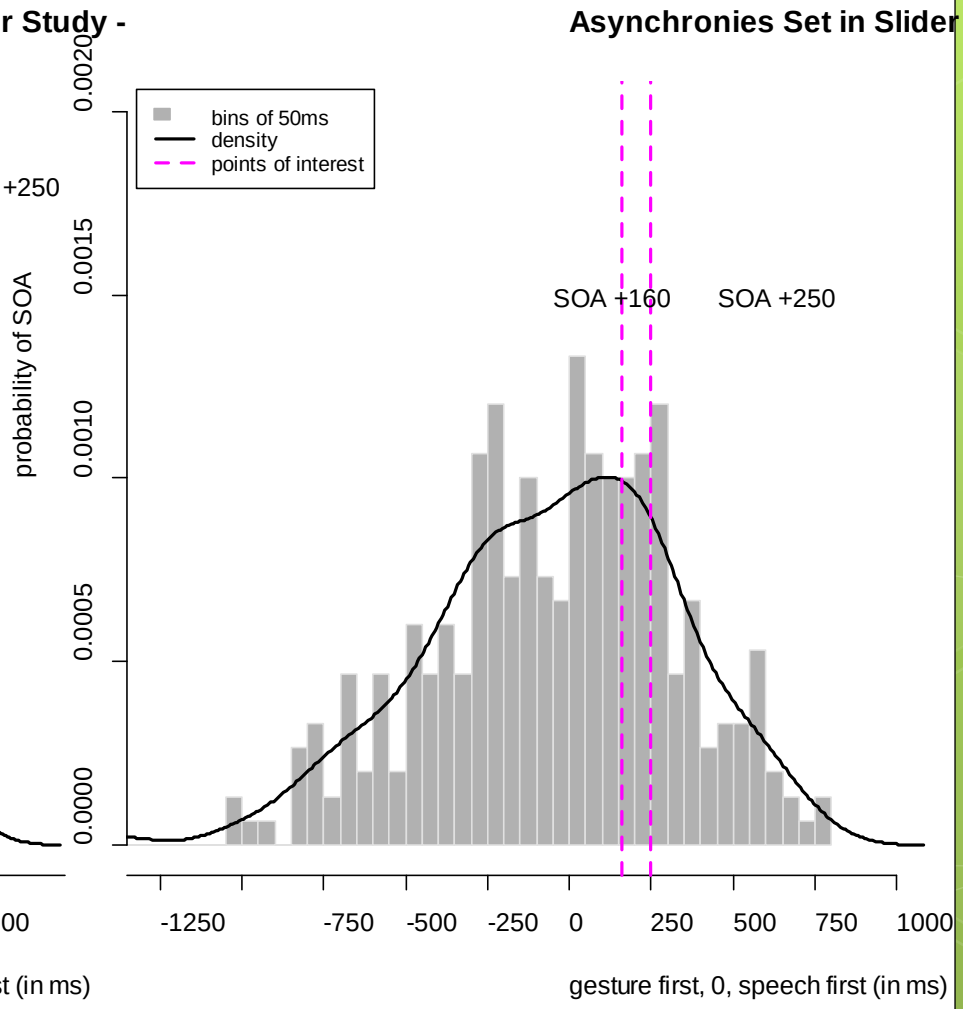
- ping pong ball: taken out of results due to bad video quality

Results – Gesture Stimuli

- audio first: 155/300
- video first: 153/300
- Range:
  -1778 ms (gesture first) to +754 ms (speech first)
- Mean: -72 ms (stddev. 422)

Distribution of Asynchronies





Summary

- the AVI window for physical events is close to the expected value:
  - Massaro et al. (1996): audio delay of 250ms to 500ms
  - Our study: audio delay or advance of ≈200ms
- the AVI window for speech and gesture
  - is larger than for physical events
  - shows audio advance and delay
  - is larger than expected (ca. -600 to +600 ms)

Implications for theories on gesture and speech processing

- the GP is temporally very flexible in perception
- allows for higher tolerance in modeling gestures in virtual agents and robots
- gesture-speech synchrony might be a consequence of the **production** system, but not be essential for **comprehension**

Questions? Comments?

Or contact me:
ckirchhof@uni-bielefeld.de

Some sources

De Ruiter, J. (2000). The production of gesture and speech. In McNeill, D. (Ed.), *Language and Gesture* (pp. 284-311). Cambridge, UK: CUP.

Gullberg, M., & Kita, S. (2009). Attention to speech-accompanying gestures: Eye movements and information uptake. *Journal of Nonverbal Behavior*, 33(4), 251-277.

Habets, B., Kita, S., Shao, .Z, Özyürek, A., & Hagoort, P. (2011). The role of synchrony and ambiguity in speech-gesture integration during comprehension. *Journal of Cognitive Neuroscience*, 23(8), 1845-54.

Holler, J, Shovelton, HK, Beattie, GW. (2009). Do iconic hand gestures really contribute to the communication of semantic information in a face-to-face context? *Journal of Nonverbal Behavior*, 33, 73-88.

Kendon, A. (2004). Gesture: Visible Action as Utterance. Cambridge, UK: CUP.

Kirchhof, C. (2011). So What's Your Affiliation With Gesture? *Proceedings of GeSpIn*, 5-7 Sep 2011, Bielefeld, Germany.

Massaro, D.W., Cohen, M.M.,& Smeele, P.M.T. (1996). Perception of Asynchronous and Conflicting Visual and Auditory Speech. *Journal of the Acoustical Society of America, 100,* 1777-1786.

Mc Neill, D. (2005). Gesture and thought. Chicago, IL: University of Chicago Press.

McNeill, D. (in press). How Language Began: Gesture and Speech in Human Evolution (Approaches to the Evolution of Language). New York, NY: CUP.

Özyürek, A., Willems, R. M., Kita, S., & Hagoort, P. (2007). On-line integration of semantic information from speech and gesture: Insights from event-related brain potentials. *Journal of Cognitive Neuroscience*, 19(4), 605-616.

Van Wassenhove V., Grant K. W., & Poeppel D. (2007). Temporal window of integration in auditory–visual speech perception. *Neuropsychologia*, 45, 598–607.

Vatakis, A., Navarra, J., Soto-Faraco, S., & Spence, C. (2008). Audiovisual temporal adaptation of speech: temporal order versus simultaneity judgments. *Experimental Brain Research*, 185(3), 521-9.