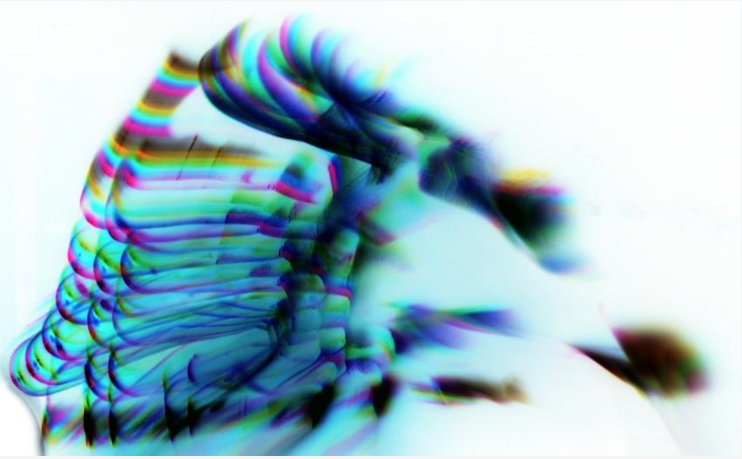
7ICOM

International Conference on Multimodality

June 2014

Desynchronized speech-gesture signals still get the message across



Caro Kirchhof
Bielefeld University
Germany
ckirchhof@uni-bielefeld.de

Asynchrony of multimodal signals in real life

- thunder & lightning
- dubbing
- subtitles in movies or video games
- delays in online streaming or on Skype/facetime

Asynchrony of multimodal signals in research

thunder & lightning	psychophysics
dubbing	phonetics & psycholinguistics
subtitles in movies or video games	psycholinguistics
delays in online streaming or on Skype/facetime	phonetics & psycholinguistics

Perception of asynchrony – audiovisual integration (AVI)

- thunder & lightning
- dubbing
- subtitles in movies or video games
- delays in online streaming or on Skype/facetime

- cause & effect
- irritating to inacceptable
- distracting to confusing
- irritating to inacceptable

Asynchrony: speech-lips vs. speech-gesture

- McGurk effect:
  - "fused percepts" (McGurk 1976)
- temporal window of AVI:
  - lips up to 500ms before speech (Massaro et al. 1996)
  - speech up to 30 ms before lips (van Wassenhove et al. 2007)

- little research (yet)
- synchrony is essential to production (e.g. McNeill 2005)
- visual 160-360 ms before speech acceptable (Habets et al. 2011)

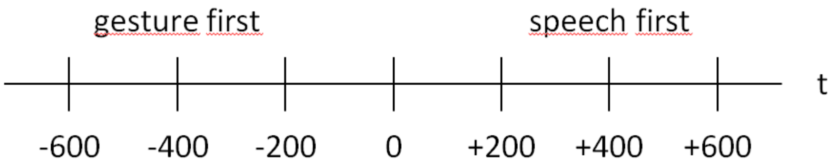
Do multimodal messages get the message across when the channels are not in synchrony?

speech + lips = yes (within a small temporal window)

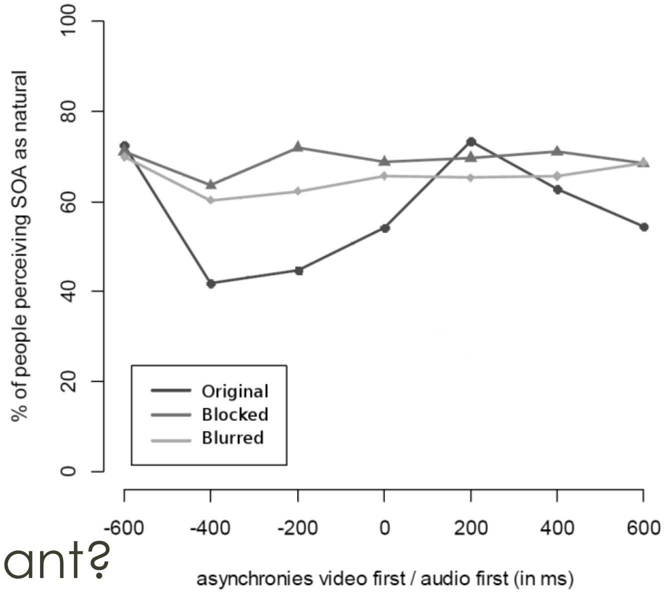
speech + gestures = ?

Study 1:
Perceptual judgment study

- 24 clips of natural speech
- AV-desynchronization:

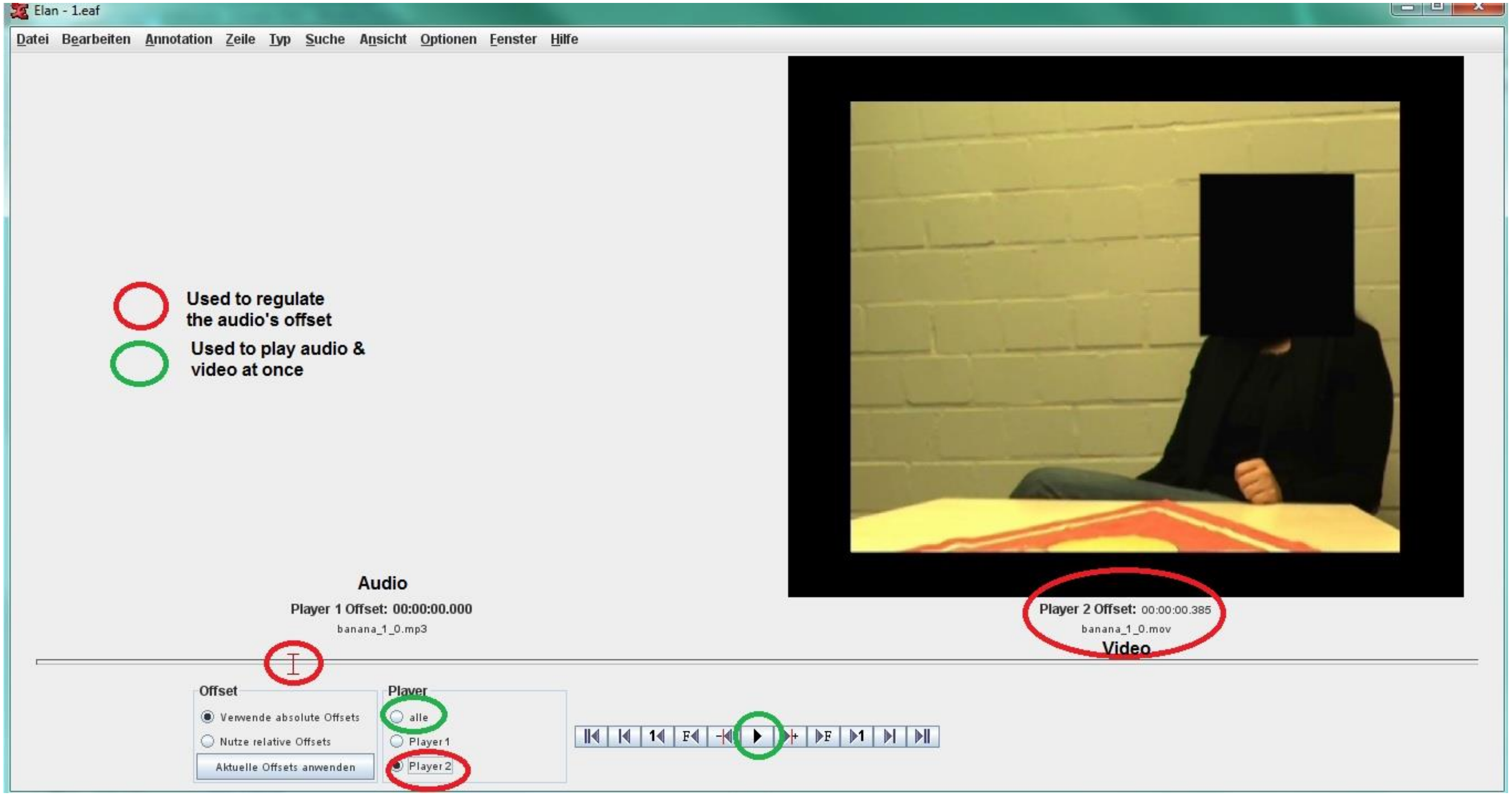


- conditions: head visible/obscured/invisible
- 618 participants
- results:
  - visible: within known AVI window
  - obscured/invisible:
    >60% of people **accepted**
    -600 to +600ms
    for head-obscured conditions (p<.05)
- Is speech-gesture synchrony less relevant?



But: Do the windows ***accepted*** differ from those ***reproduced***?

Studies 2 & 3:
User-specified synchronization



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

○ Player 1

● Player 2

Study 2

- 18 stimuli:
  - 15 iconic gestures from Study 1 w/ blob with
  - 5 pseudorandomized initial asynchronies (277-1034ms)
  - Baseline: 3 “physical events” (hammer & snap) w/ 902ms video advance
- a slider-interface (ELAN)
- 20 participants (mean age 25, 6 male)
- ➢ 300 manipulated stimuli

Study 2 - results

physical events

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

- range:
  (video first)
  -978 ms to +442 ms
  (audio first)

- mean: +14 ms (stddev. 246)

gestures

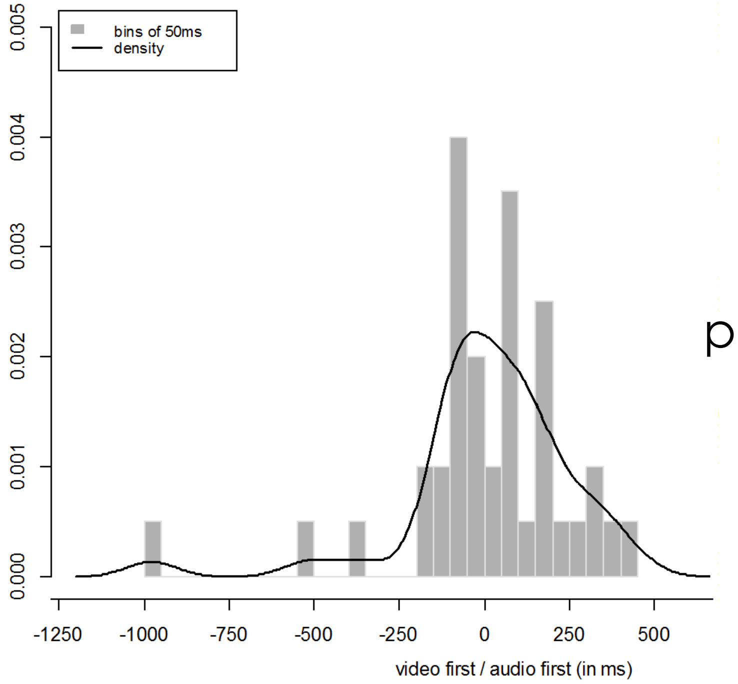
- audio first: 155/300
- video first: 153/300

- range:
  (gesture first)
  -1778 ms to +754 ms
  (speech first)

- mean: -72 ms (stddev. 422)

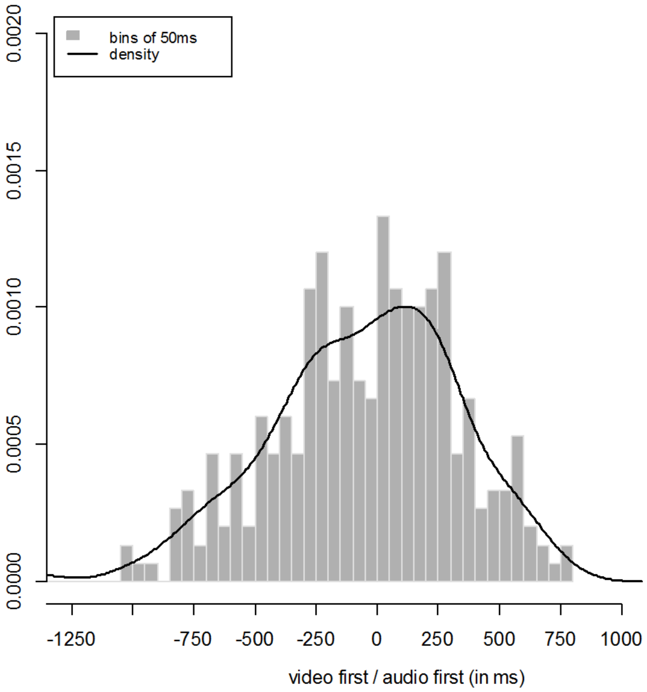
Study 2 - results

physical



vs. at $p<.05$[1]

gestures



[1] right-tailed t-test

Study 3 – follow-up to study 2

- 19 stimuli:
  - gestures from Study 1 w/ blob:
    - 6 iconic, 4 deictic, 3 emblematic
    - with 5 pseudorandomized initial asynchronies (277-1034ms)
  - 6 "physical events" (book, clap, glass, keyboard, knock, champagne)
    - with 902ms video advance

- 23 participants (mean age 25, 12 male)

➢ 437 manipulated stimuli

Study 2+3 - results

physical events

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

- range:
  (video first)
  -978 ms to +672 ms
  (audio first)

- mean: +86 (stddev. 214.4)

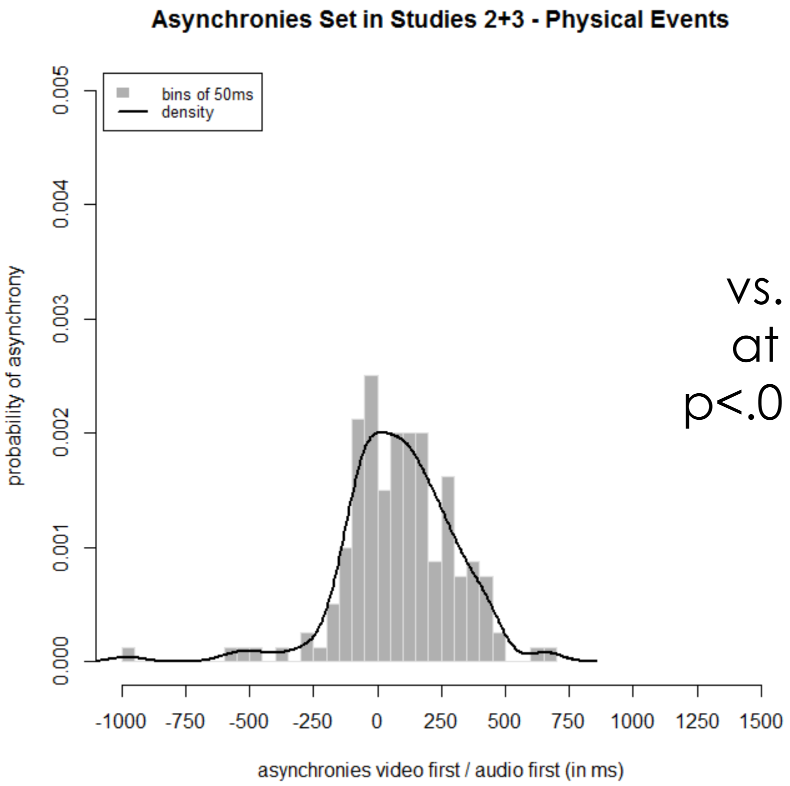
gestures

- audio first: 155/300
- video first: 153/300

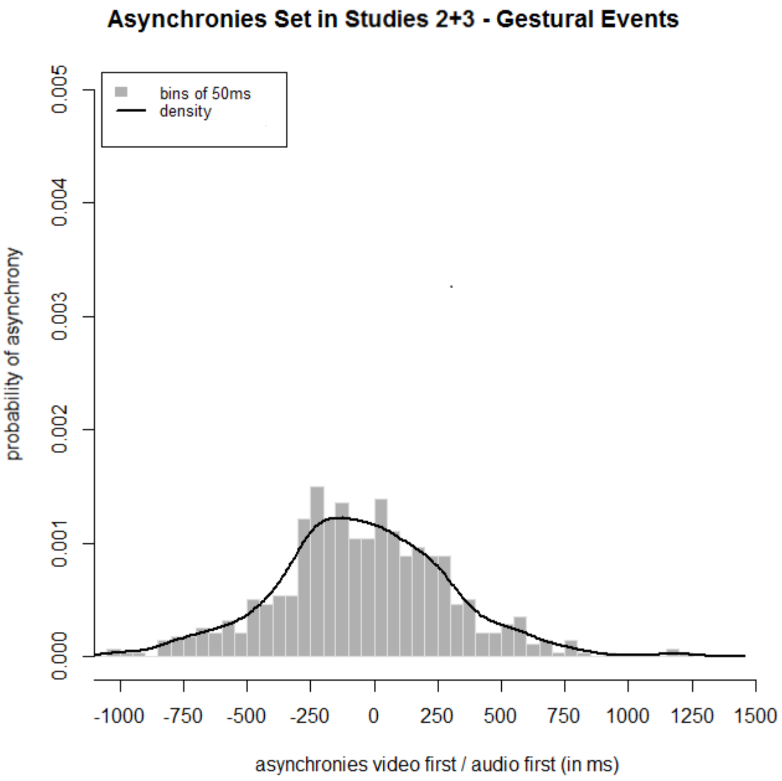
- range:
  (gesture first)
  -1908 ms to +1216 ms
  (speech first)

- mean: -54.5 (stddev. 370.7)

Study 2+3 - results

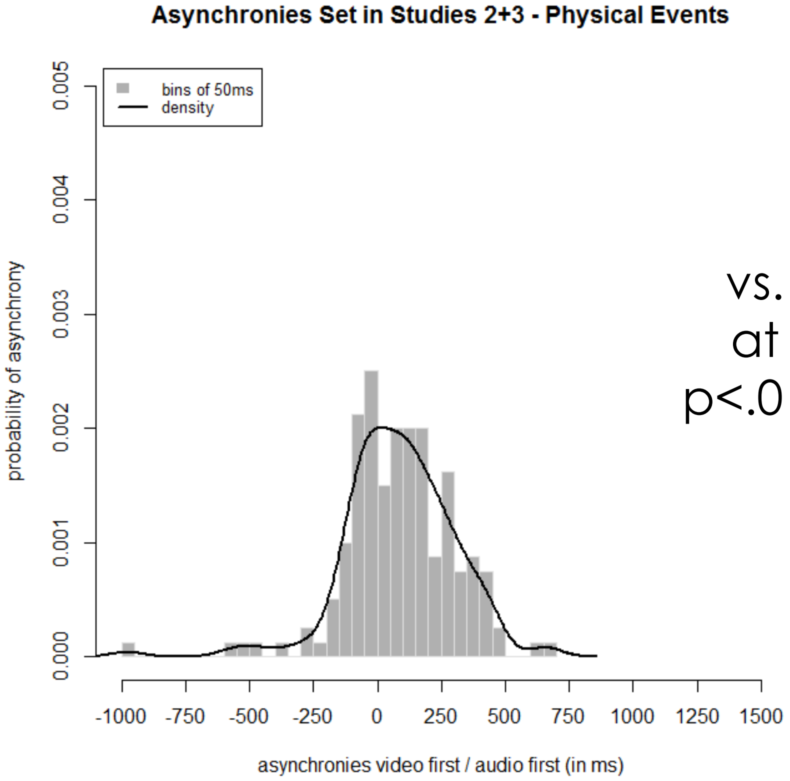


vs.
at
p<.01[1]

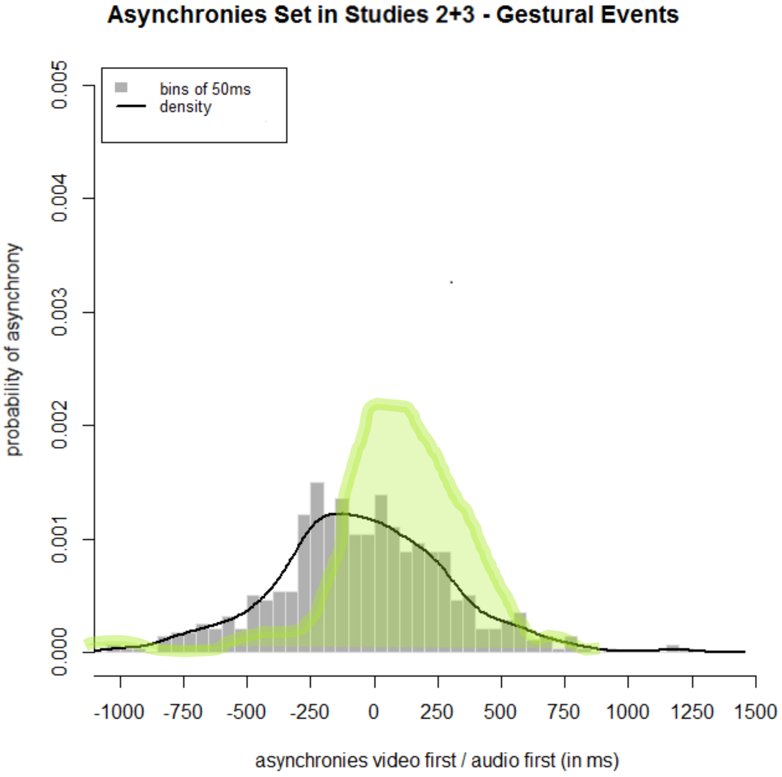


[1] right-tailed t-test

Study 2+3 - results

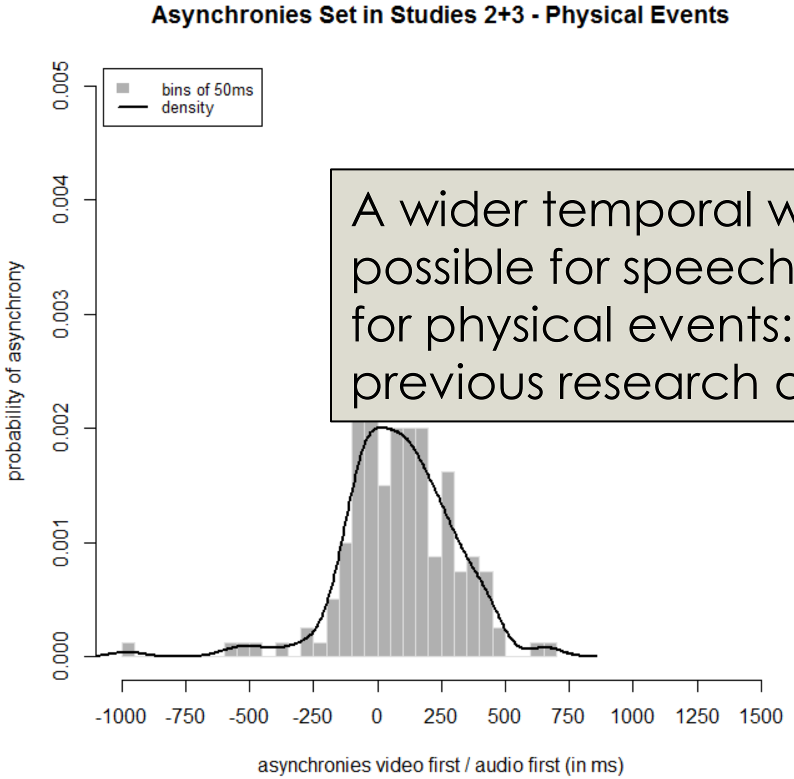


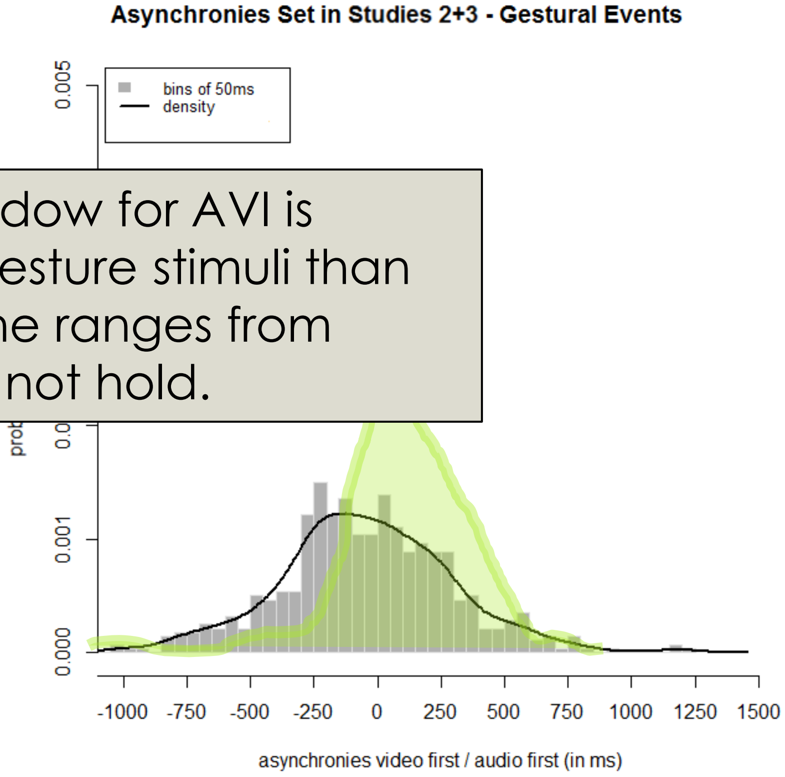
vs.
at
p<.01[1]



[1] right-tailed t-test

Study 2+3 - results


Asynchronies Set in Studies 2+3 - Physical Events


Asynchronies Set in Studies 2+3 - Gestural Events

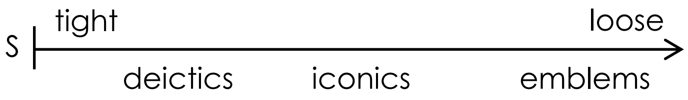
A wider temporal window for AVI is possible for speech-gesture stimuli than for physical events: The ranges from previous research do not hold.

Continua of Speech-Gesture Production & Perception

Speech-Gesture Synchrony in *Production*

After Kendon:

(McNeill 2005, pp. 7 ff.)



S

tight

loose

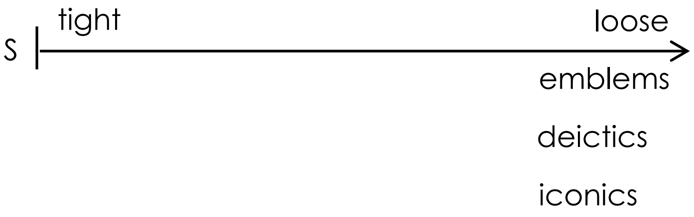
deictics

iconics

emblems

Speech-Gesture Synchrony in *Perception*

Hypothesis:



S

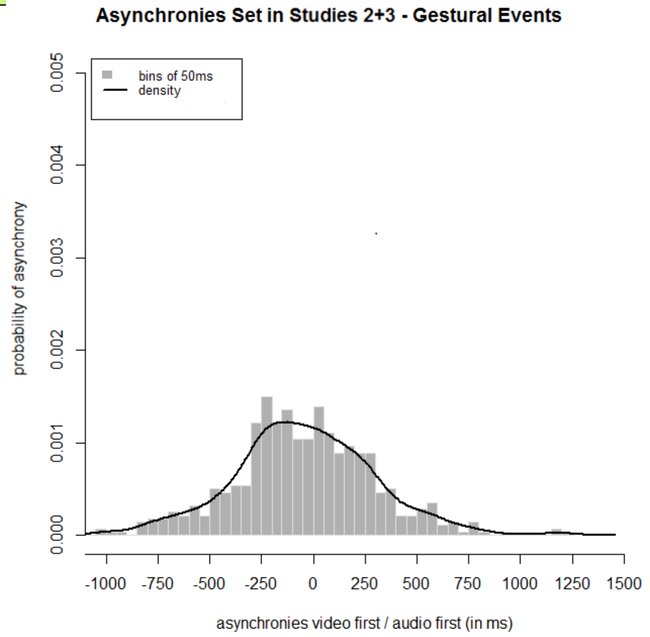
tight

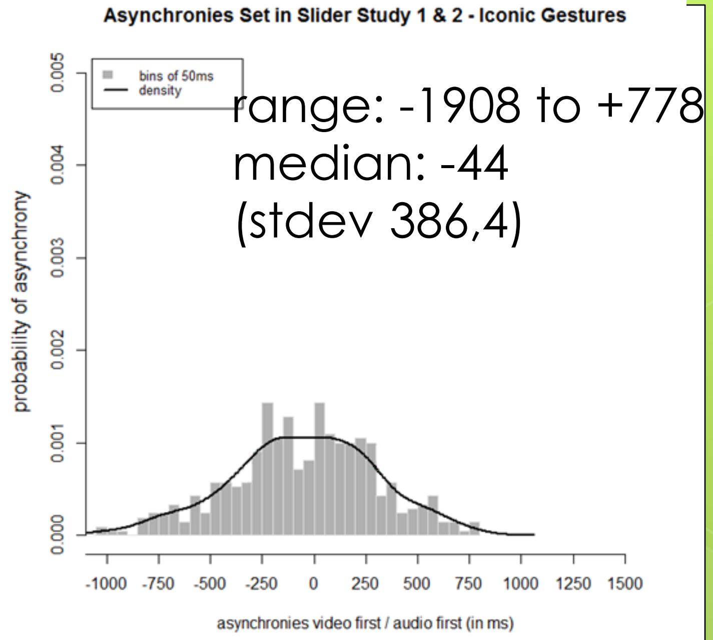
loose

emblems

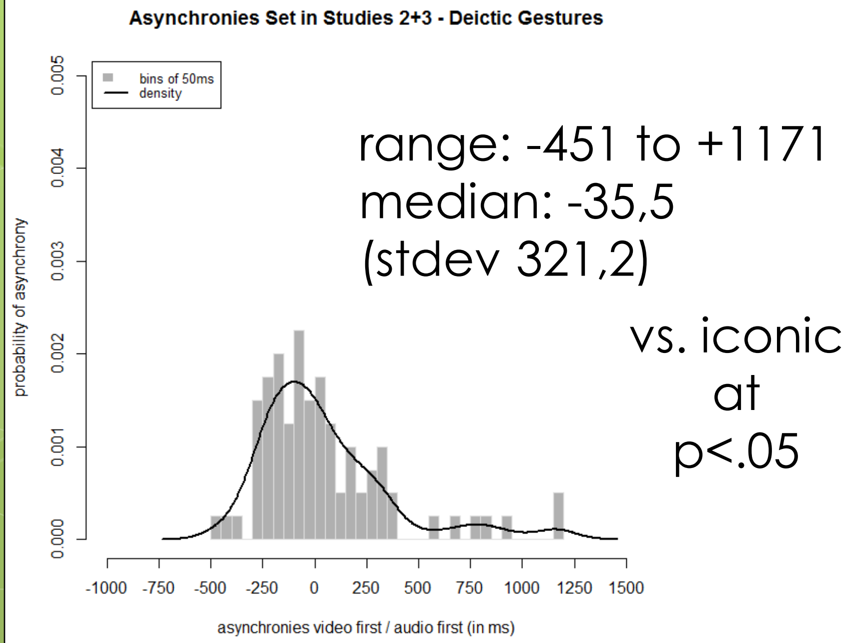
deictics

iconics


Asynchronies Set in Studies 2+3 - Gestural Events

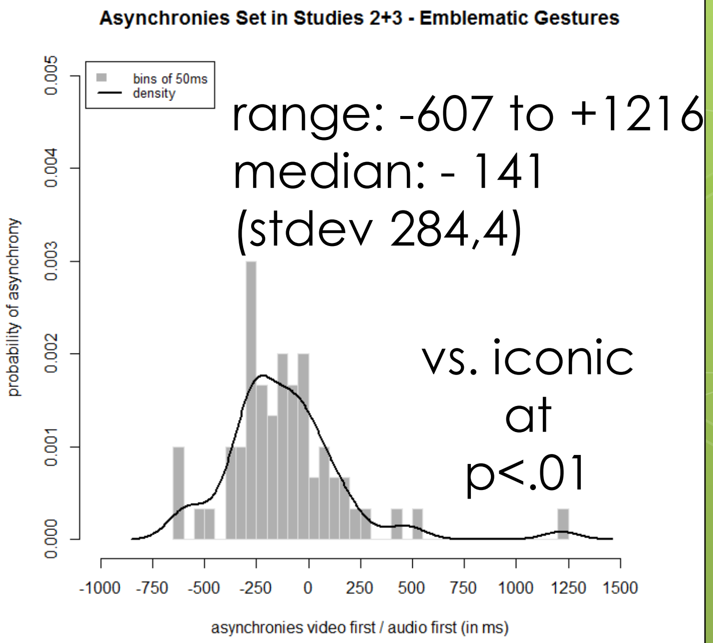

Asynchronies Set in Slider Study 1 & 2 - Iconic Gestures

range: -1908 to +778
median: -44
(stdev 386,4)


Asynchronies Set in Studies 2+3 - Deictic Gestures

range: -451 to +1171
median: -35,5
(stdev 321,2)

vs. iconic
at
p<.05


Asynchronies Set in Studies 2+3 - Emblematic Gestures

range: -607 to +1216
median: - 141
(stdev 284,4)

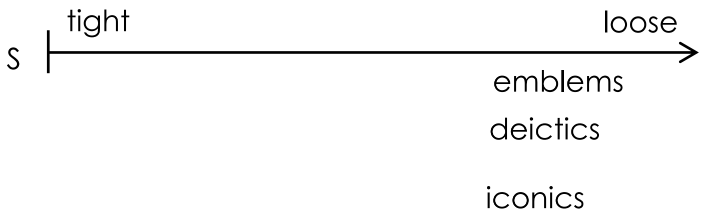
vs. iconic
at
p<.01

Different gestures, different synchrony ties

- *iconics*: wider, flatter tolerance
- *deictics*: preferred start before speech, still looser than physical events
- *emblems*: even more preferred before speech

Speech-Gesture Synchrony in Perception

hypothesis:



S

tight

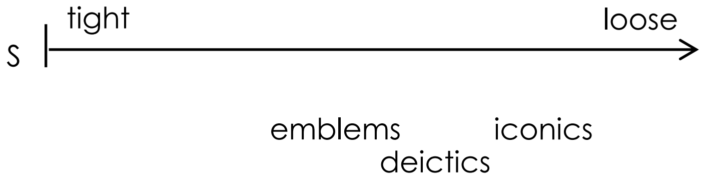
loose

emblems

deictics

iconics

study:



S

tight

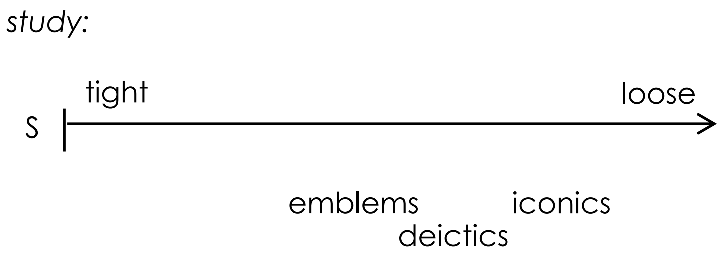
loose

emblems

iconics

deictics

Speech-Gesture Synchrony in *Perception*


study:

S

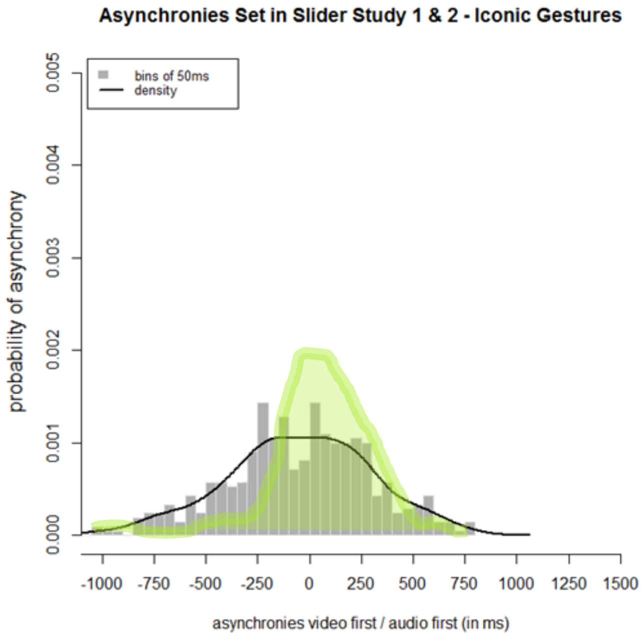
tight

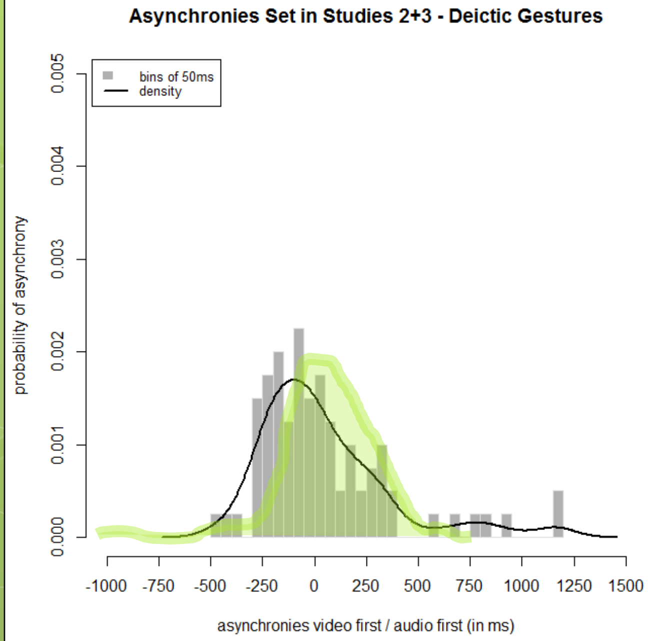
loose

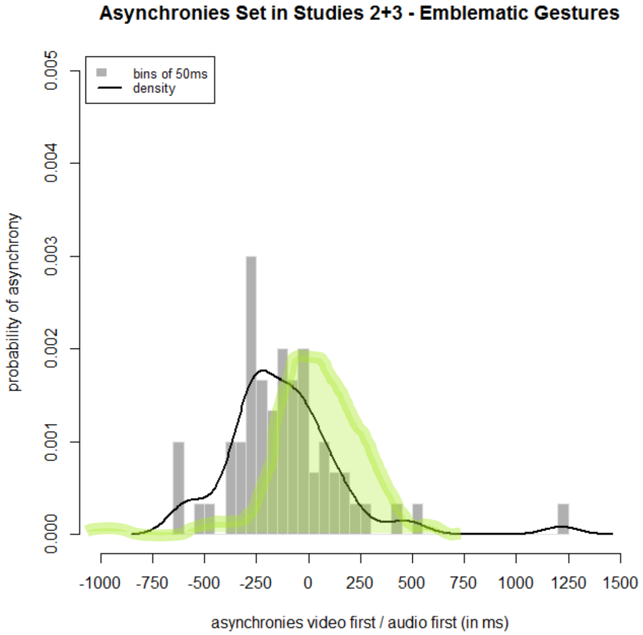
emblems

iconics

deictics


Asynchronies Set in Slider Study 1 & 2 - Iconic Gestures


Asynchronies Set in Studies 2+3 - Deictic Gestures


Asynchronies Set in Studies 2+3 - Emblematic Gestures

Findings

1. Speech-gesture synchrony is tighter in production than necessary for perception.
2. Synchronization for emblems is similarly critical as for deictics.
3. Synchronization for deictics & emblems is more critical than for iconics.

Do multimodal messages get the message across when the channels are not in synchrony?

speech + lips = yes (within a small temporal window)

speech + gestures = yes (within larger temporal windows)

Questions or comments?

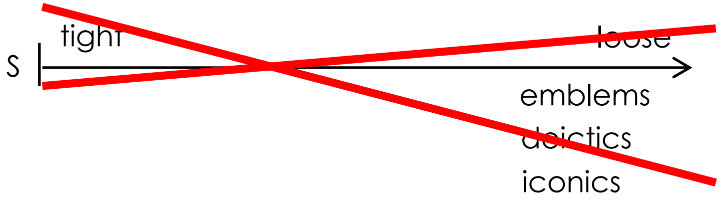
Speak now or contact me later:

ckirchhof@uni-bielefeld.de

Discussion

- The hypothesis that ***gestures in general*** *need only be synchronized loosely with speech for perception* has been **falsified**.

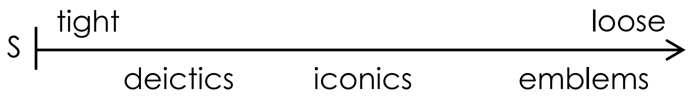
Speech-Gesture Synchrony in *Perception*



- Explanation:
  - **Deictic** gestures **correspond** to deictic POS to which they are semantically/temporally bound. Their phases are short, the temporal window for AVI is small.
  - **Emblematic** gestures are **redundant** to certain POS to which they are semantically/temporally bound. Their phases are short, the temporal window for AVI is slightly larger.
  - **Iconic** gestures **complement** utterances. They do not target specific POS. Their phases are flexible in duration, the temporal window for AVI is only bound by the duration of the utterance.

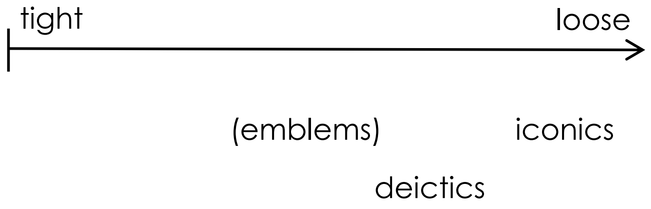
Alternative Hypothesis

- In **production**, the **gesture stroke** is synchronized with the speech it corresponds to semantically (cf. *Kendon Continuum*, McNeill 2005, pp. 7 ff.):



S
tight
loose
deictics
iconics
emblems

- For **perception**, the **duration of the gesture phrase** is synchronized with the speech it corresponds to semantically.



tight
loose
(emblems)
iconics
deictics

Sources

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

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.

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.

Kirchhof, C. (2012). On the audiovisual integration of speech and gesture. *Presented at the ISGS 2012*, 24-27 July 2012, Lund, Sweden.

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.

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.