Audio-visual ~~fusion~~ binding in speech perception

Jean-Luc Schwartz, GIPSA-Lab, UMR 5216
CNRS & Université de Grenoble

The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013 Grant Agreement no. 339152)

http://florence.apln-blog.fr

A typical multi-speaker scene is made of a mixture of sounds and sights
(Cocktail party effect)

This resulted in separate developments in Auditory Scene Analysis (ASA)
and Auditory-Visual Speech Perception (AVSP)

ASA and AVSP should be integrated inside AVSA
(***Audio-Visual (speech) Scene Analysis***)
(Berthommier & Schwartz, 1998-2015)

1. Audiovisual Speech perception without scene analysis?

2. AVSSA: experimental data about streaming and chunking

3. Possible theoretical, neural and computational bases for AVSSA

1. Audiovisual Speech perception without scene analysis?

2. AVSSA: experimental data about streaming and chunking

3. Possible theoretical, neural and computational bases for AVSSA

The role of multisensory interactions reevaluated in neurosciences

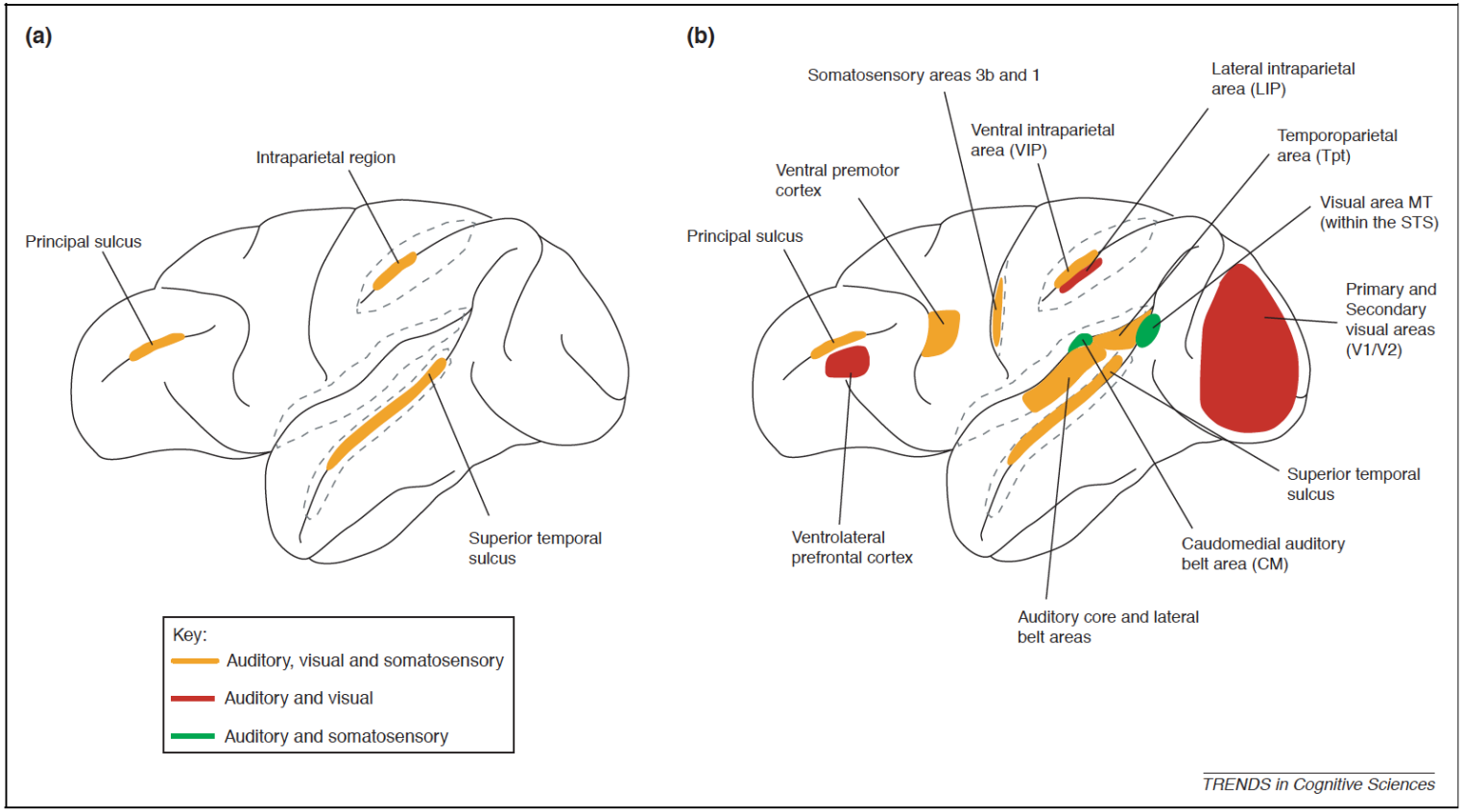
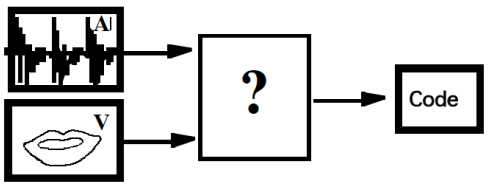


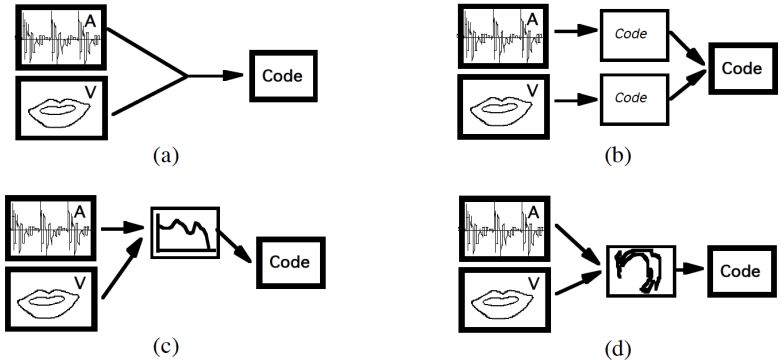
Figure 1. (a) Traditional scheme of the cortical anatomy of multisensory areas in the primate brain. **(b)** Modern scheme of the cortical anatomy of multisensory areas. Colored areas represent regions where there have been anatomical and/or electrophysiological data demonstrating multisensory interactions. In V1 and V2, the multisensory interactions seem to be restricted to the representation of the peripheral visual field. Dashed gray outlines represent opened sulci.

Ghazanfar & Schroeder, TICS 2006

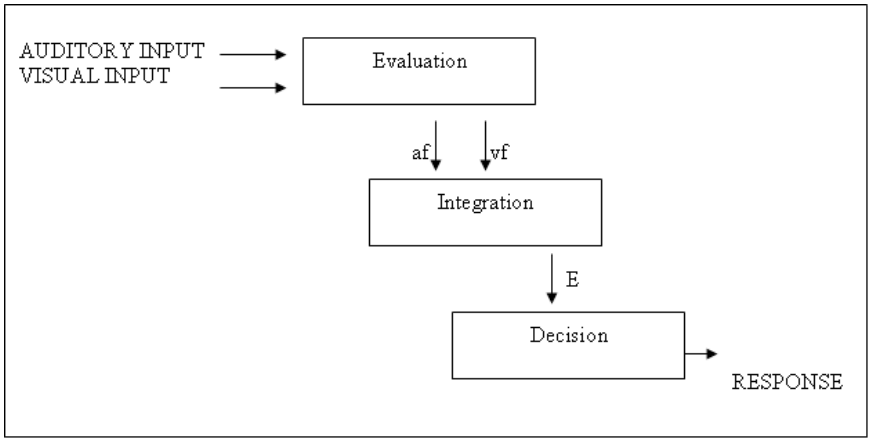
"In recent years the field of multisensory research has expanded and altered radically with the realization that multisensory influences are much more pervasive than classical views assumed and may even affect brain regions, neural responses, and judgments traditionally considered modality specific" (Driver & Noesselt, Neuron 2008)

Classical AVSP models





Summerfield, 1987
Schwartz et al., 1998



FLMP - Massaro, 1987, 1989, 1998

The "standard model" $P_{AV} = F(I_A, I_V)$
The local inputs entirely predict the output

AVSP “psychophysics”

Variations of the McGurk effect with:

- Language (Spanish, German (Duran, 1995), Italian (Bovo et al., 2009), Dutch (de Gelder et al., 1995), Hungarian (Grassegger, 1995), French (Cathiard et al., 2001, Colin et al. 2002) vs. Japanese, Chinese (Sekiyama & Tohkura, 1991, 1993), (Sekiyama, 1997), (Hisanaga et al, 2009)
- Age (from development to aging, Sekiyama & Burnhal, 2008; Sekiyama et al., 2014) and subject (Schwartz, 2010)
- Audiovisual incongruence in speaker’s identity (Green et al, 1991)
- Audiovisual incongruence in time (temporal integration window) (Munhall et al, 1996, van Wassenhove et al, 2007) or space (Jones & Munhall, 1997, Bertelson et al, 1994, Colin et al, 2001)
- Rate of articulation (Colin & Radeau, 2003, Munhall et al, 1996)
- Noise (Sekiyama & Tohkura, 1991, Colin et al, 2004, MacDonald et al, 1999).

Towards a change of paradigm?

From the "standard model"

$P_{AV} = F(I_A, I_V)$ (FLMP)

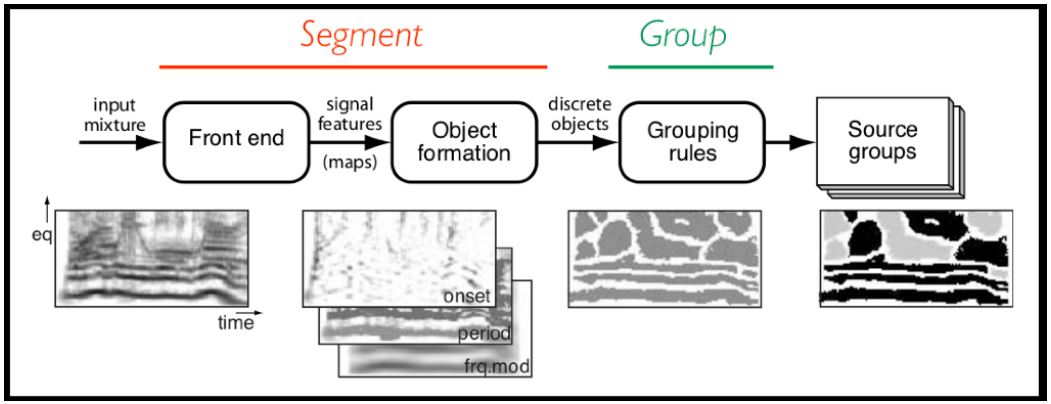
BMS

to a composite model

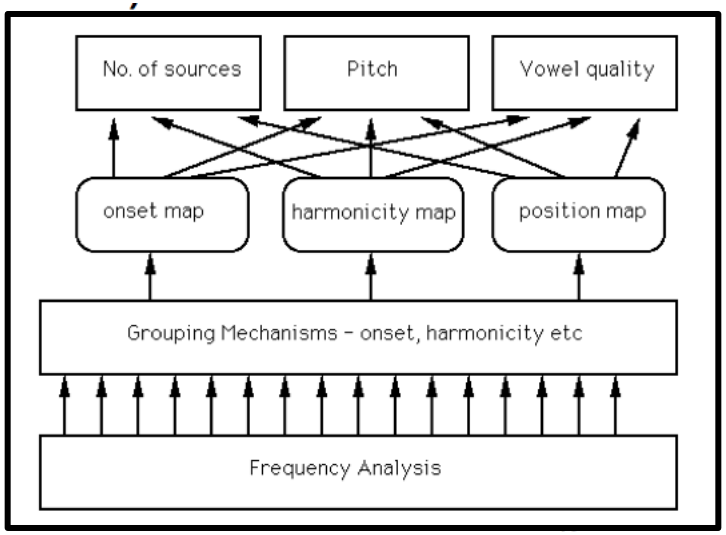
$P_{AV} = F(I_A, I_V, \text{Subj}, \text{Lang}, \text{Noise}, \text{Attention}, \text{Binding}?)$ (WFLMP)

The ASA/CASA revolution in audition

Bregman, A.S. (1990). Auditory Scene Analysis: the perceptual organization of sound. Cambridge, Mass, Bradford Books, MIT Press.



Brown 1992



Darwin 1996

Analyze -> Extract features -> categorize

Analyze -> Extract features and primitives -> Group bundles of features -> categorize

Primitives (bottom-up) and Schemas (top-down) / Attentional factors

Do we need an AVSA/CAVSA revolution in audiovisual speech perception?

Can we include in audiovisual speech perception a preliminary binding/streaming stage organizing the scene and extracting the adequate auditory and visual cues to be bound and processed together?

Can this shed light on the “psychophysics of audiovisual speech perception”?

Can this significantly change our conceptions of audiovisual speech processing in the human brain?

1. Audiovisual Speech perception without scene analysis?

2. AVSSA: experimental data about streaming and chunking

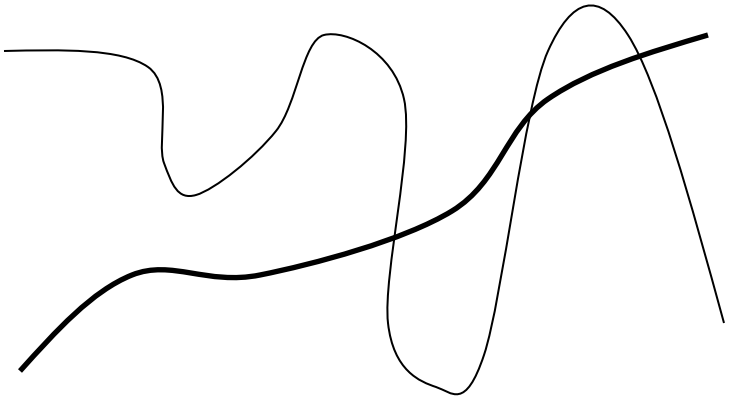
3. Possible theoretical, neural and computational bases for AVSSA

AudioVisual Speech Scene Analysis

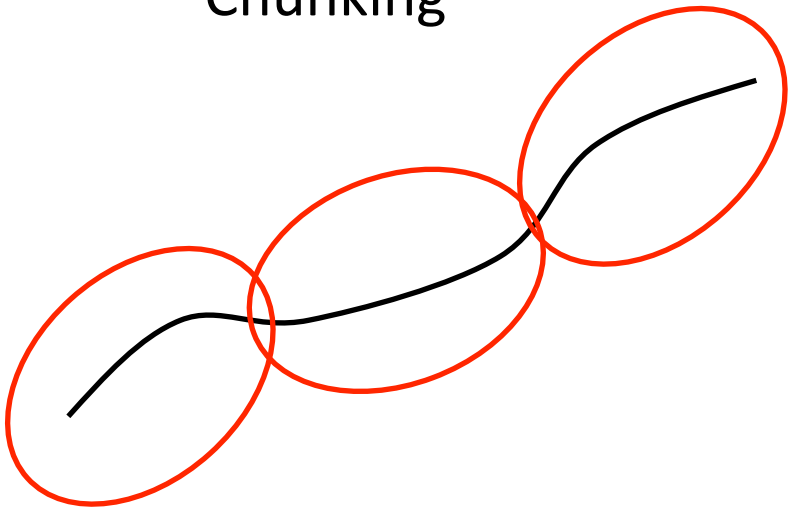
Audiovisual primitives, audiovisual schemas

Binding = Streaming + Chunking

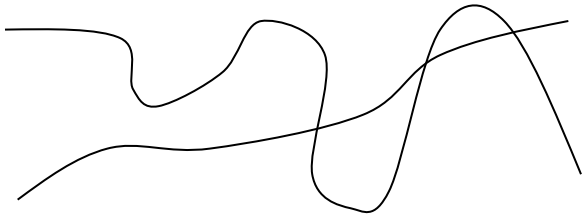
Streaming



Chunking

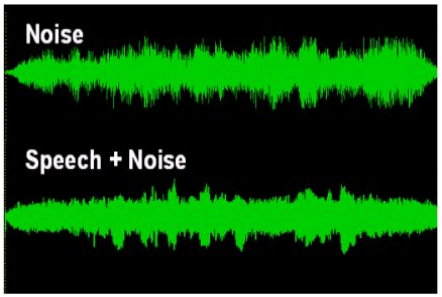


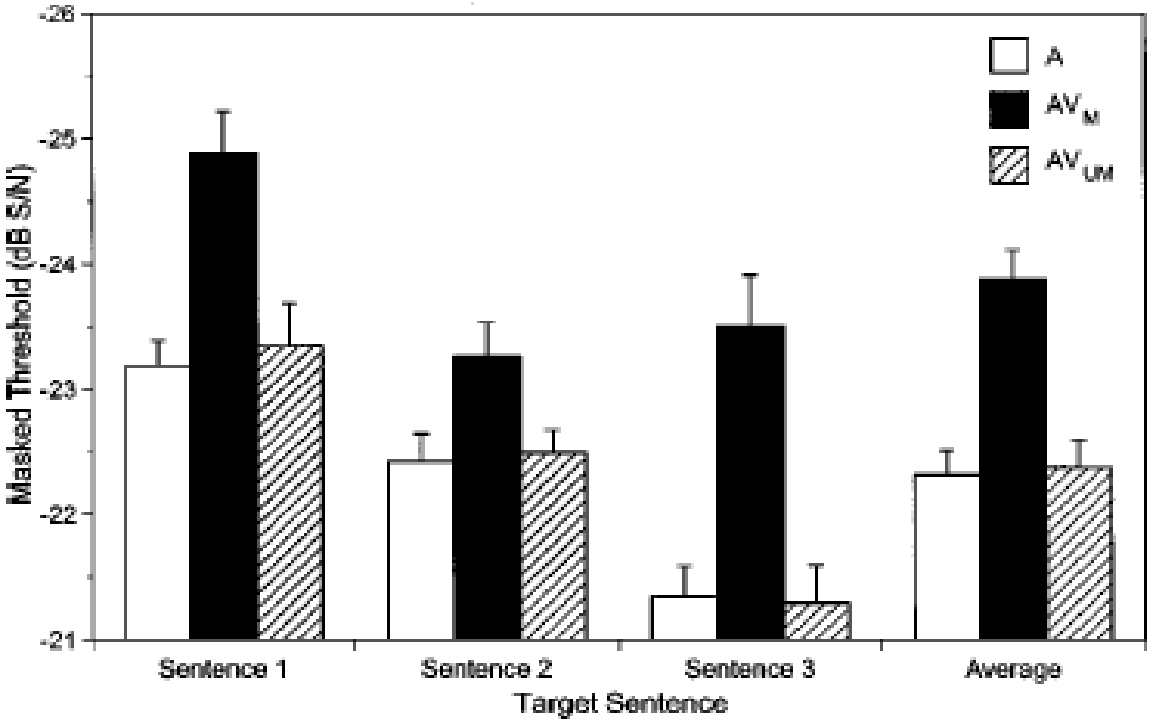
AV Streaming



The audiovisual speech detection advantage

Grant and Seitz, 2000. JASA 108, 1197-1208.
Kim & Davis, 2003, 204; Bernstein et al., 2004

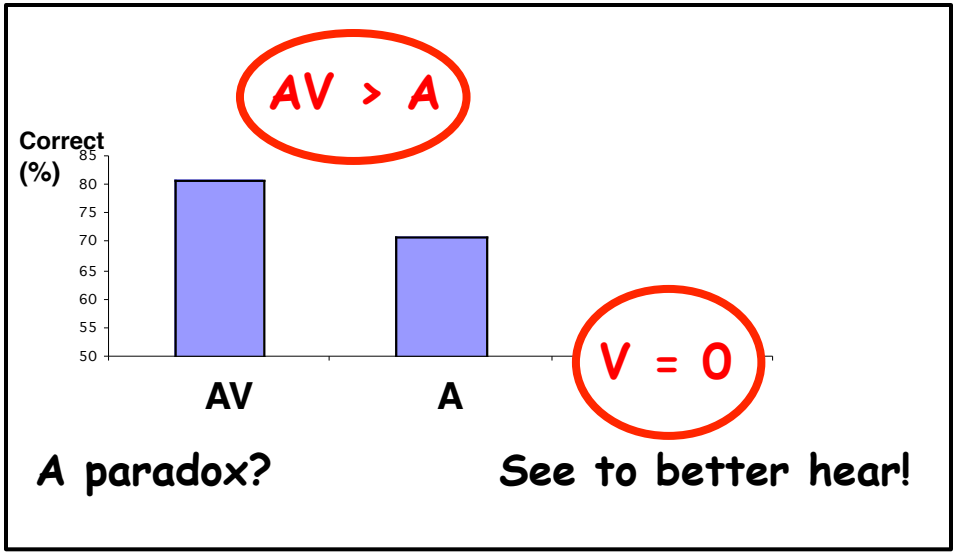


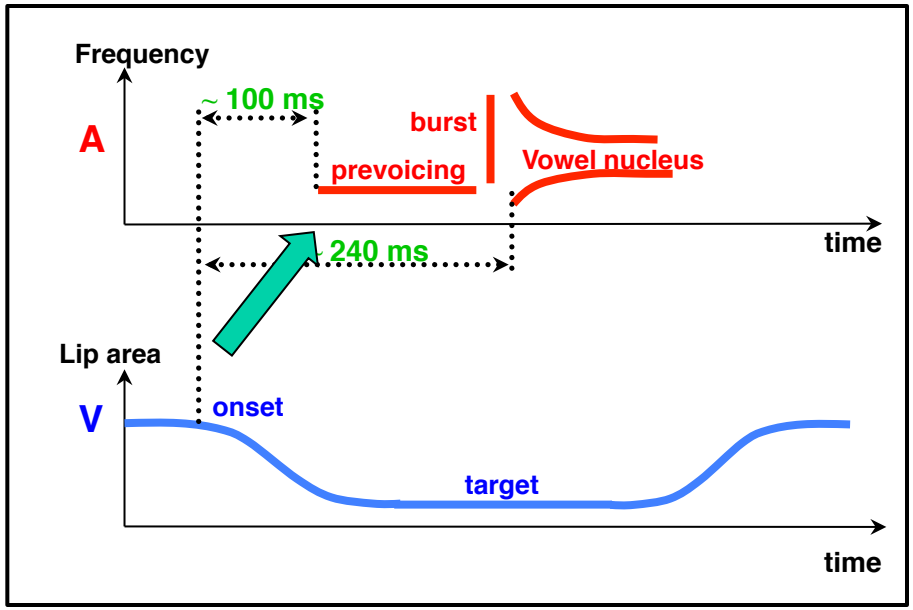


Interpretation : AV binding based on comodulations

Binding for detection (predicting and testing predictions)
Is it used also for comprehension?

Seing to hear better (Schwartz, Berthommier, & Savariaux, 2004. *Cognition,* 93, B69–B78)
Experimental trick: Study the intelligibility in noise
of visually similar speech utterances (visemes) - [u tu ku du gu]

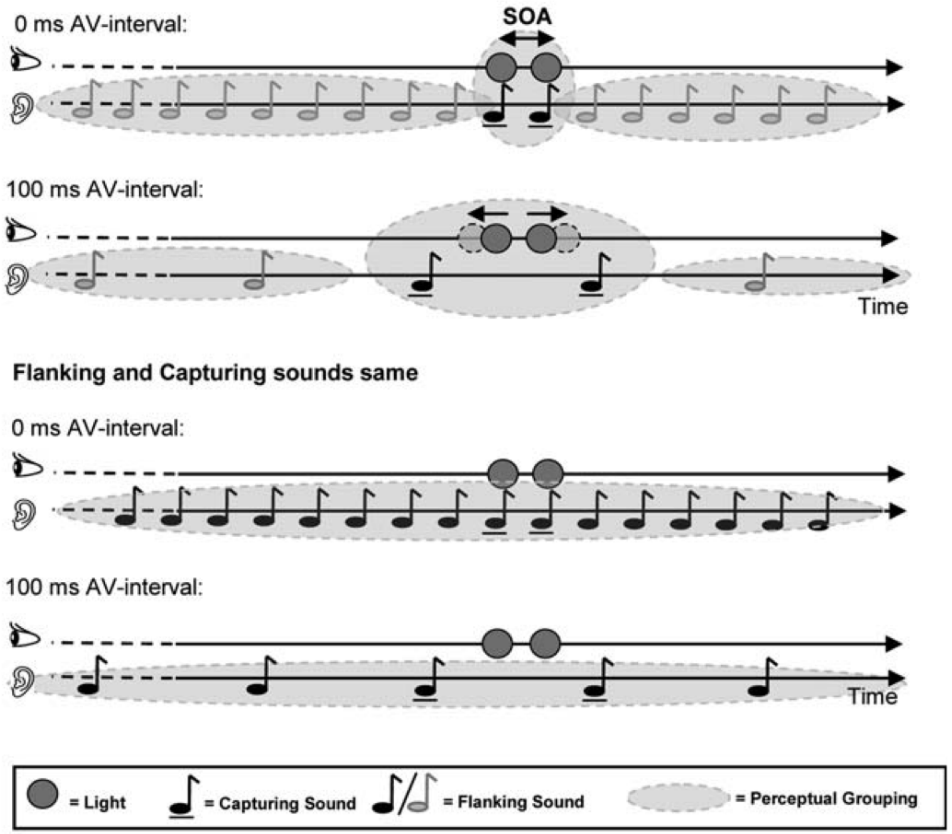




Does multisensory binding precede unisensory processing?

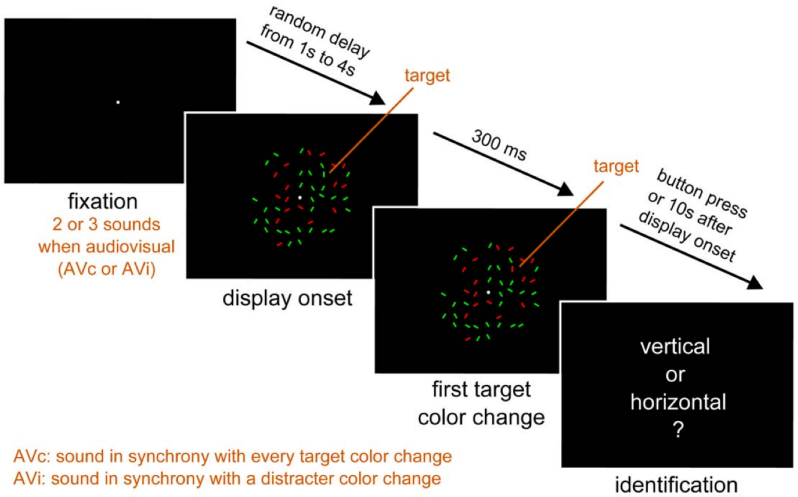
NO

Keetels et al., 2007



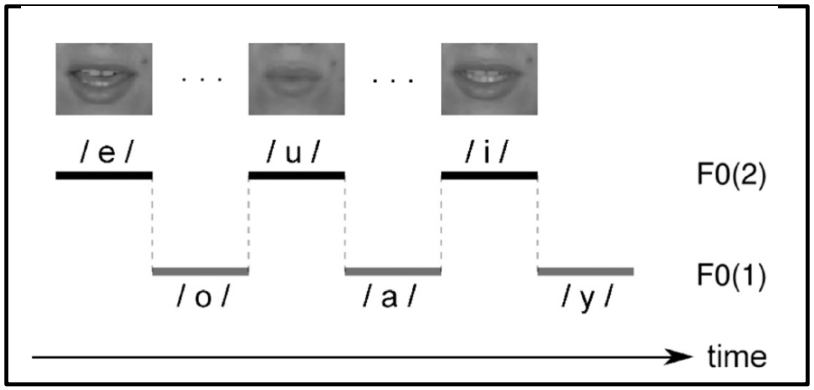
YES

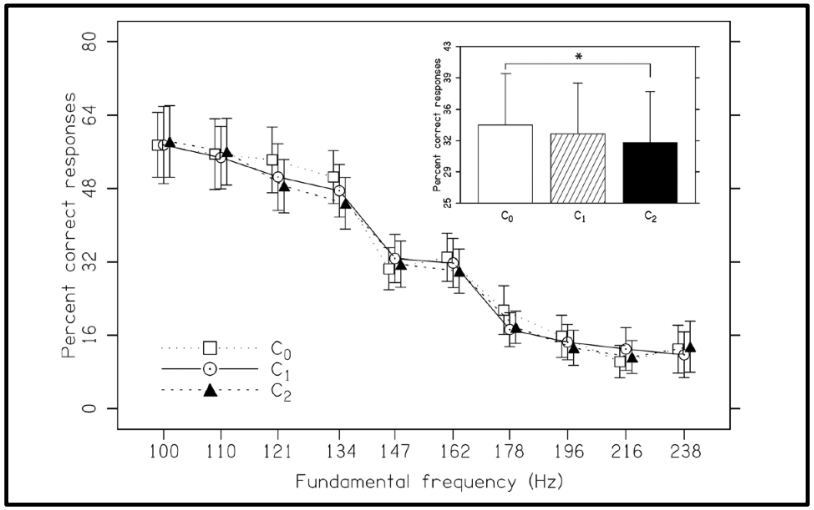
Kosem & van Wassenhove 2012



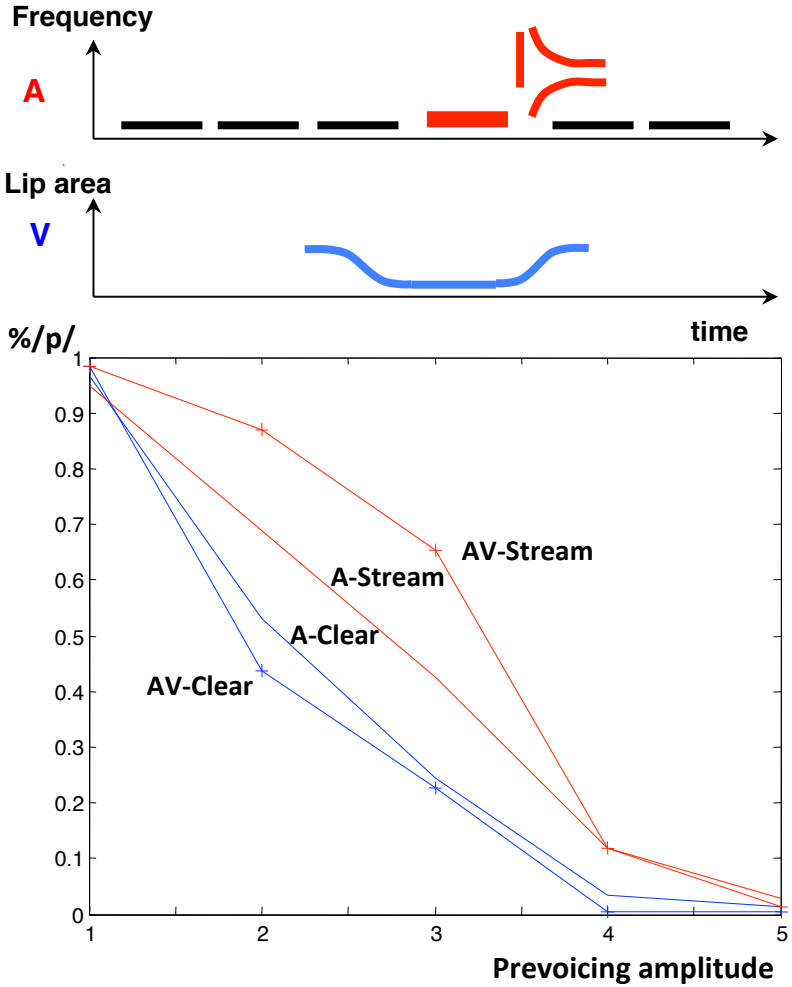
Audiovisual streaming: two (partly convincing) examples in speech perception

The effect of lip-reading on primary stream segregation – Devergie et al., 2011, JASA



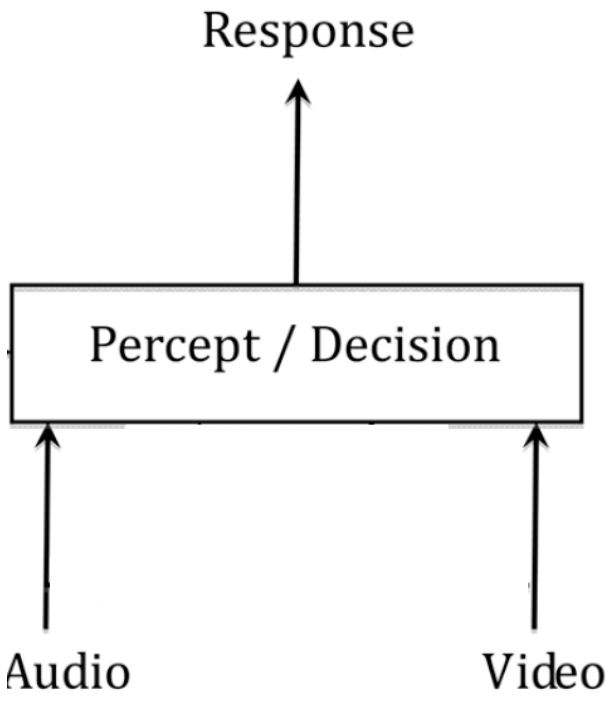


AV streaming in voicing perception – Berthommier & Schwartz AVSP 2011



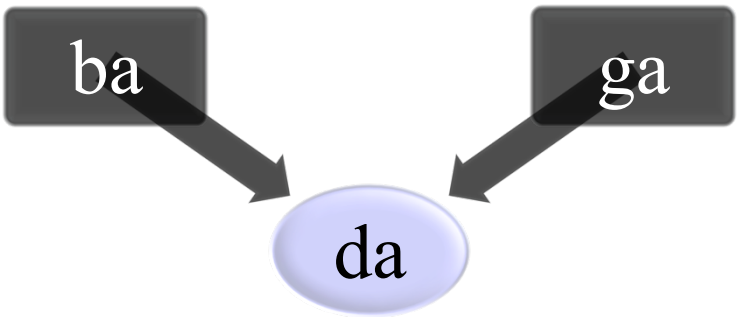
The two-stage AVSSA model

Berthommier, 2004, Speech Comm.



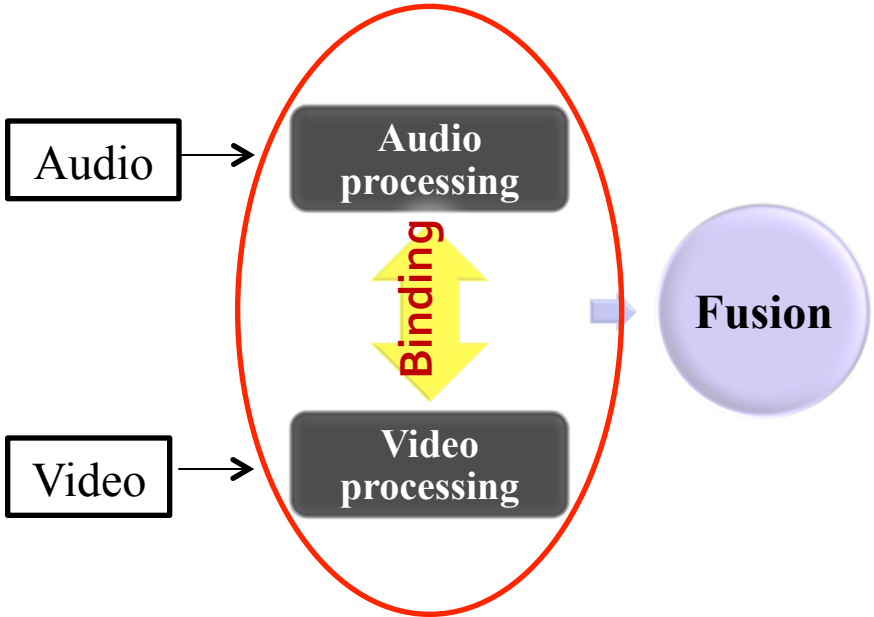
If there is binding, there should be unbinding!
-> conditional fusion in the McGurk Paradigm

Nahorna, Berthommier & Schwartz, 2012, JASA

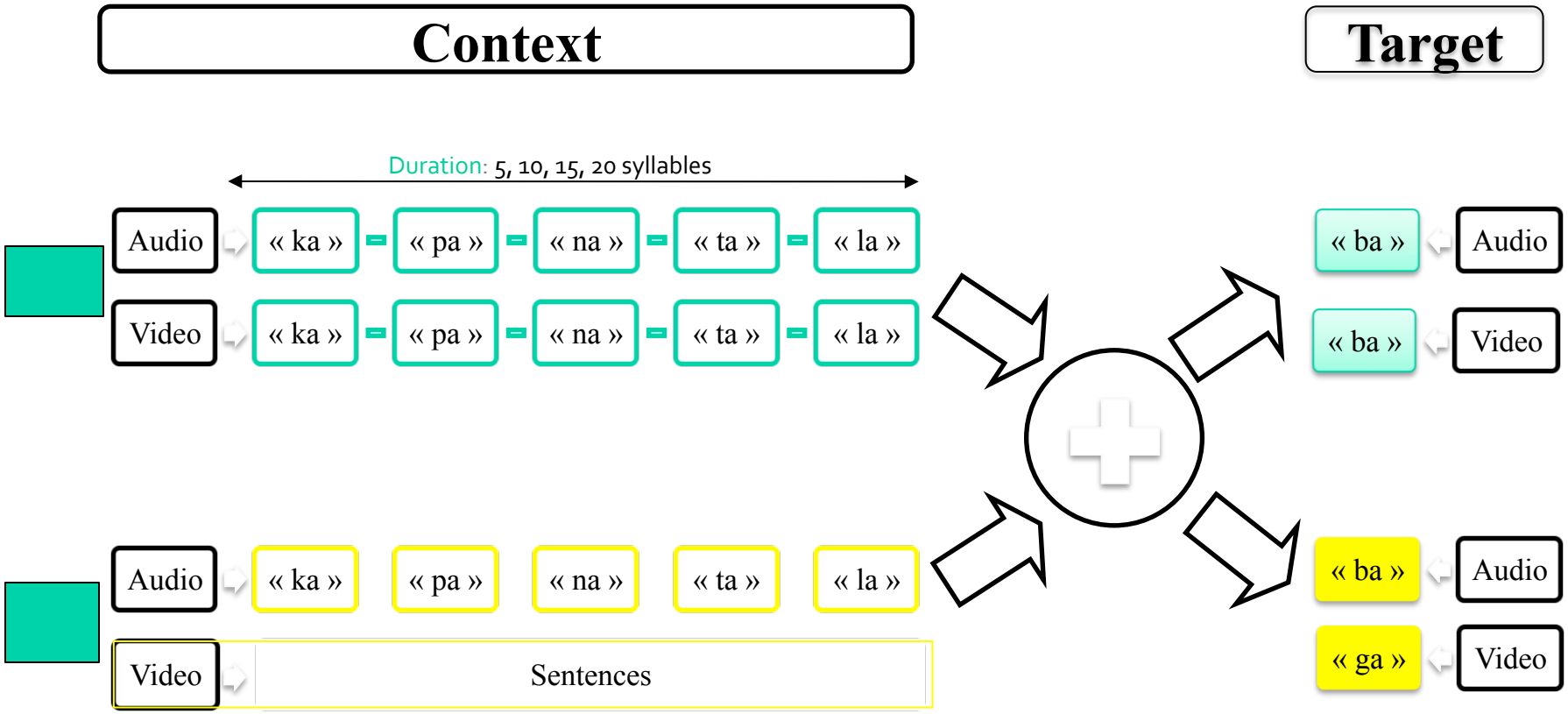


Is fusion automatic?

You can remove/decrease the McGurk effect (conditional binding/unbinding)

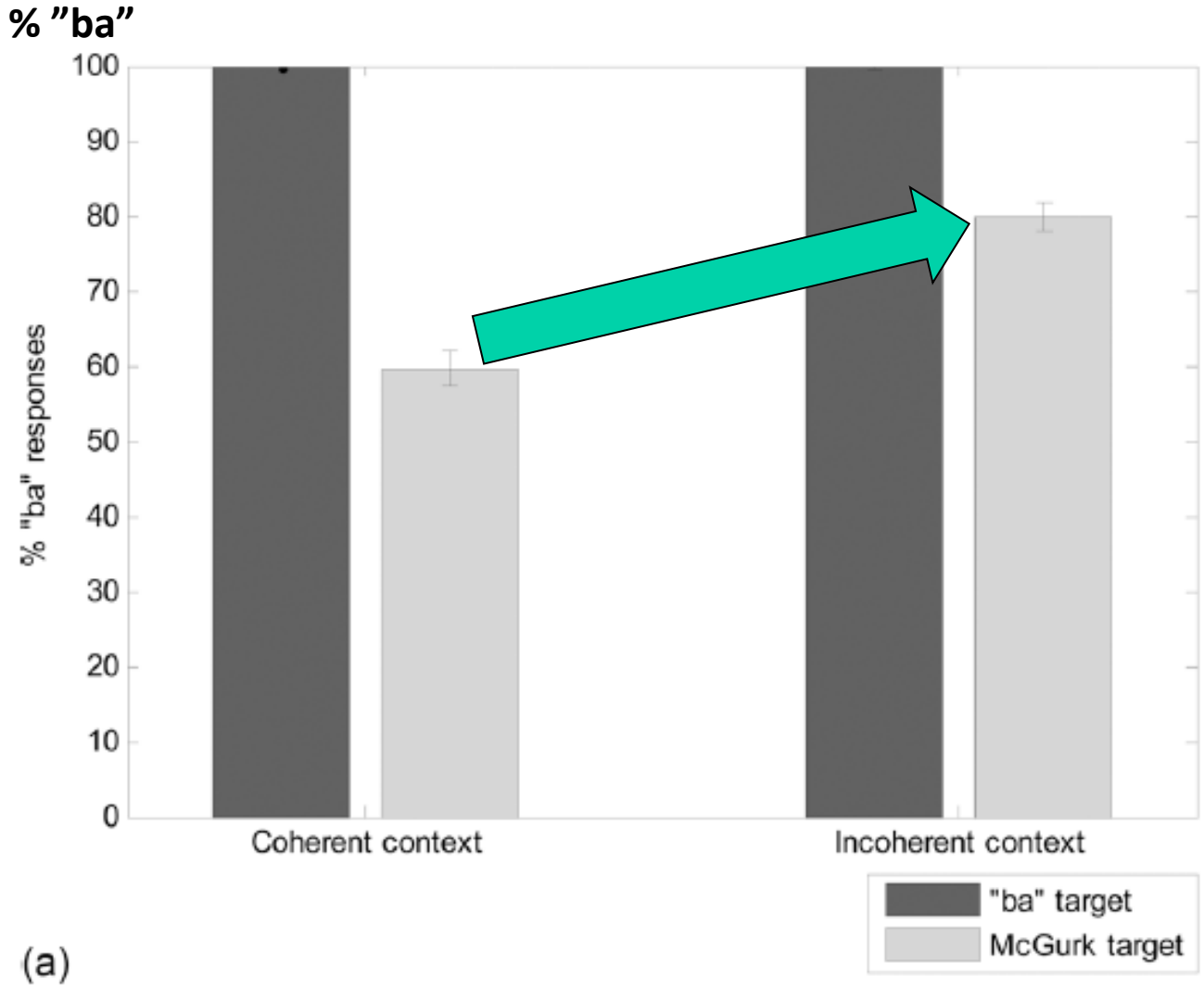


Experimental Paradigm



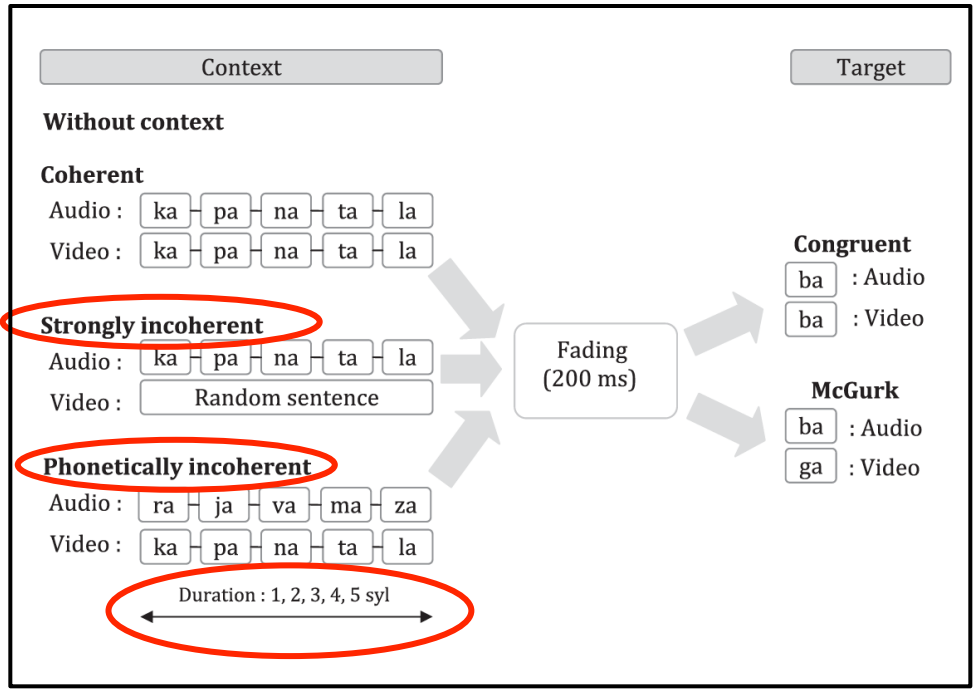
1. Evidence for unbinding

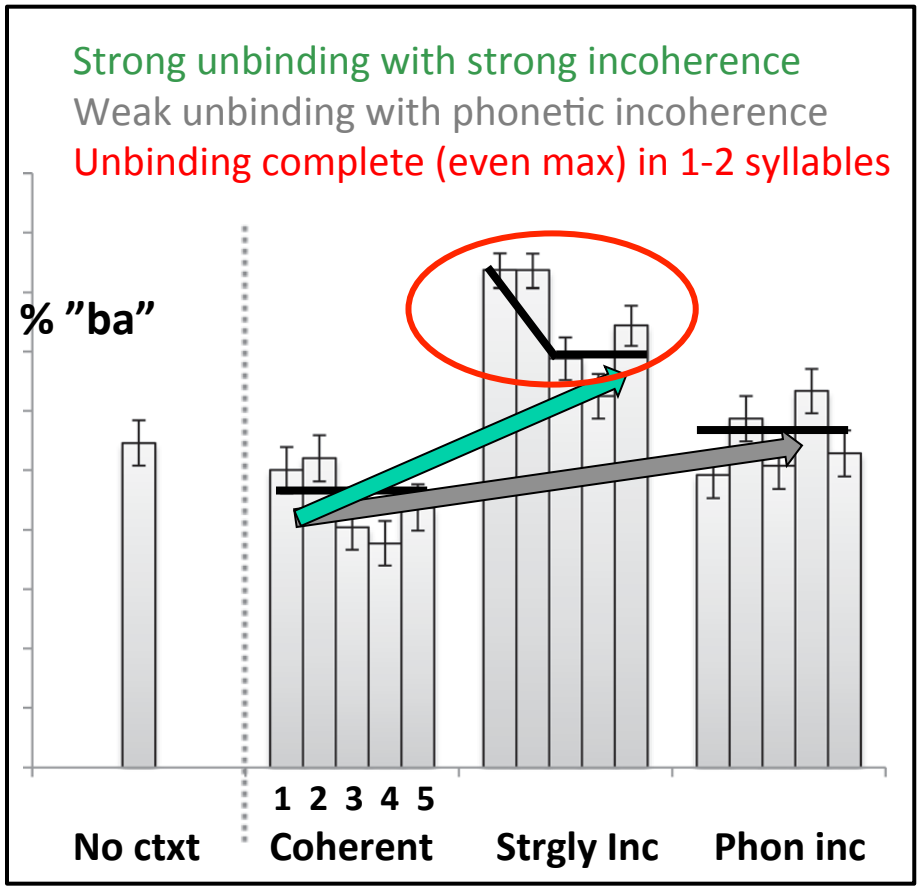
Nahorna, Berthommier & Schwartz, 2012, JASA



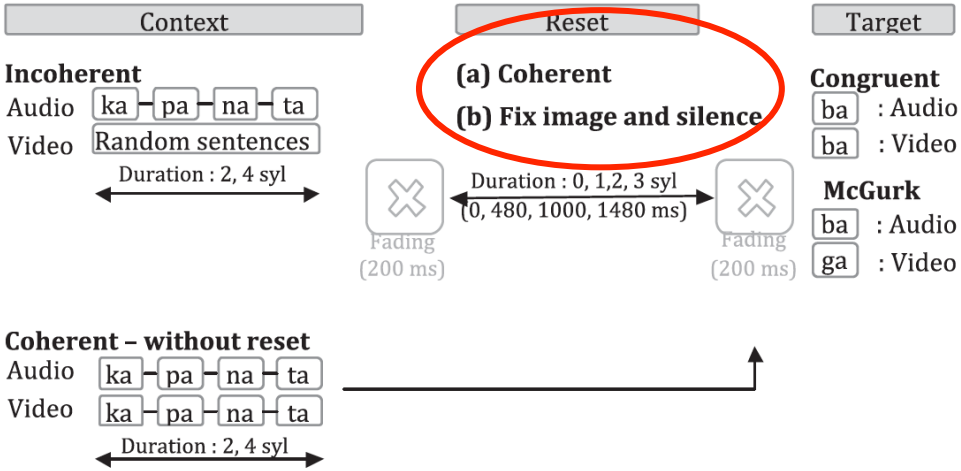
2. Dynamics of unbinding

Nahorna, Berthommier & Schwartz, 2015, JASA





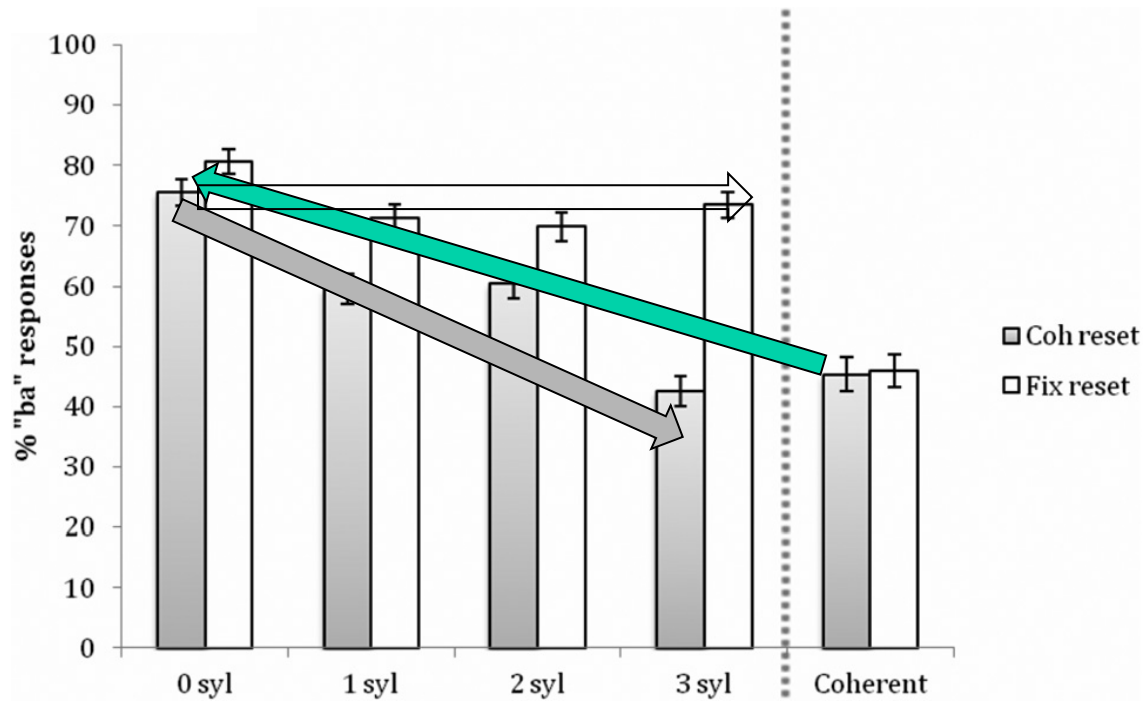
3. Dynamics of rebinding



Unbinding

Rebinding with coherent reset

Freezing with silent reset



4. AV binding in noise and competing sources

Ganesh, Berthommier & Schwartz, AVSP 2013, ISH 2015, in prep.

Incoherent context plus noise

Incoherence decreases the role of the visual input (decreases the MG effect)

Audio noise increases the role of the visual input (increases the MG effect)

This confirms the two-stage model

And sheds light on an old question: why does McGurk increases when there is audio noise? Because audio intelligibility decreases or because visual weight increases? Now we know (better …)

Two competing audiovisual streams in context (sentences and syllables)

Binding allows stream selection

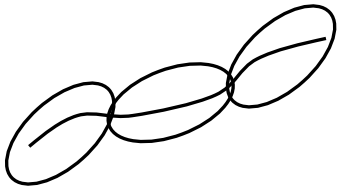
If the selected stream has a high AV coherence, the MG effect is larger than if it has a low AV coherence

Attention intervenes in the selection process and hence in the McGurk output

5. AV binding in seniors

See poster by Ganesh Chandrashekara, poster session 2, this afternoon

AV Chunking

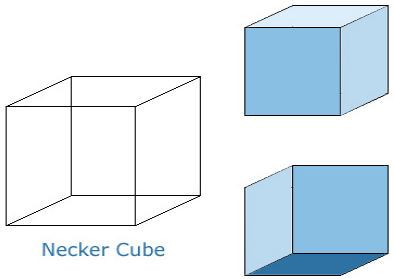


Chunking features in a speech stream: the natural CV comodulation unit
Larger natural units? the Multistability paradigm

Sato, Basirat & Schwartz, 2007, *Percept. Psychophys*
Basirat, Schwartz & Sato, 2012, *Phil. Trans. B*.

- A paradigm for studying perceptual organization (decision, attention, consciousness, …)

- From visual multistability to audition and speech



(From http://www.optical-illusion-pictures.com/)

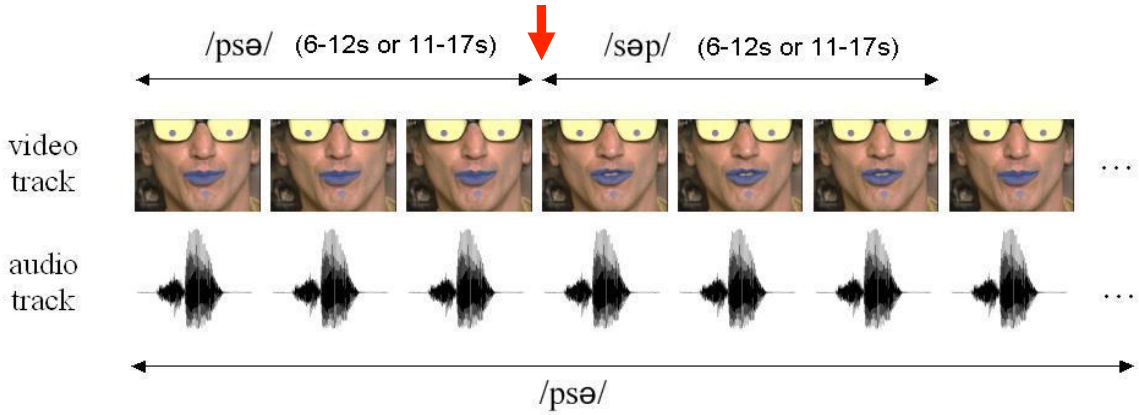
Rapid repetition of
"life" → "fly"

(Verbal Transformations,
Warren and Gregory, 1958)

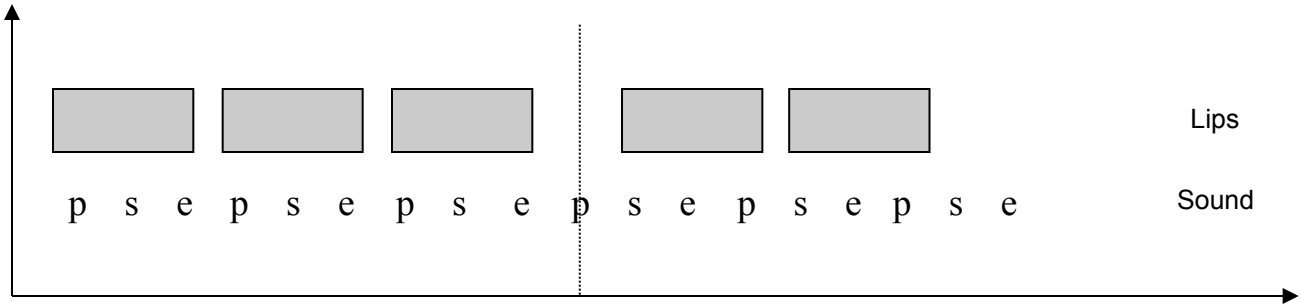
- Multistability in speech (Verbal Transformations) as a perceptuo-motor process

1. Verbal Transformations are multisensory

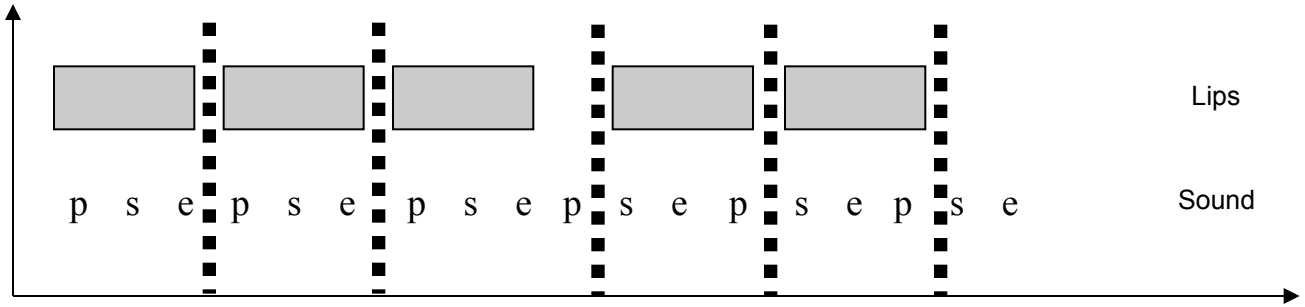
- **Experiment I: Vision penetrates the chunking process**
  - Audiovisual “pse” and “sep”



Stable audio + Video alternation between congruent and incongruent stimuli



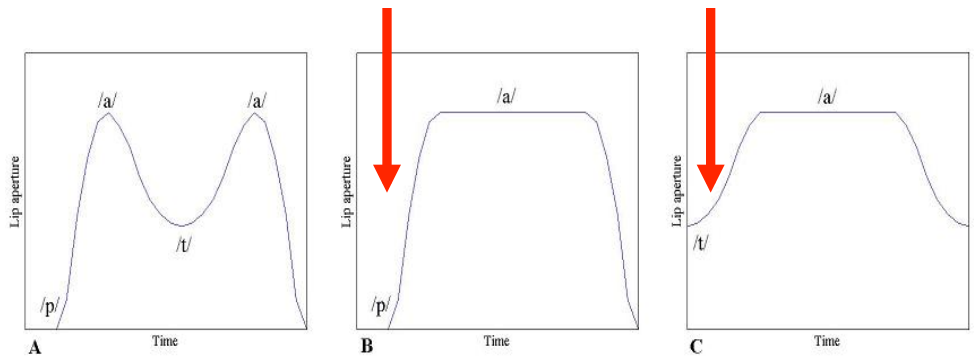
- Results:
  - In incongruent AV: strong visual influence
  - Synchronous transformations were congruent with switches in video track



2. AV chunking in verbal transformations

- **Experiment II: Vision penetrates the chunking process** *(continuing)*
  - Stimuli: /pata/ and /tapa/ in audio, AV, AVpa, AVta

Hypothesis: lip opening as a "bootstrap"

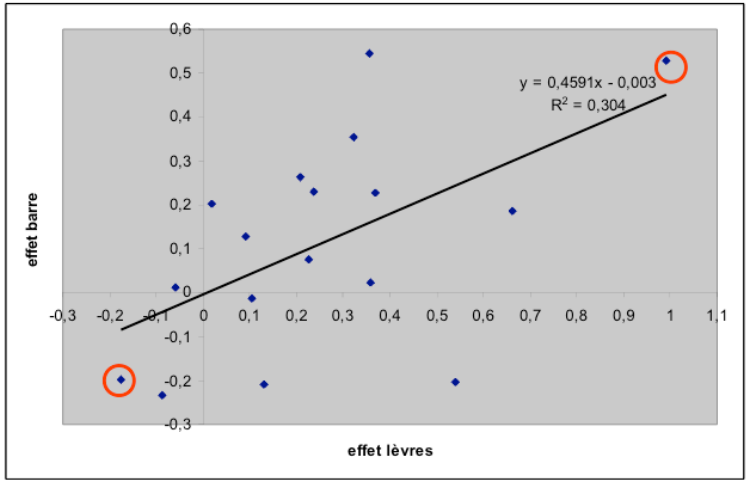


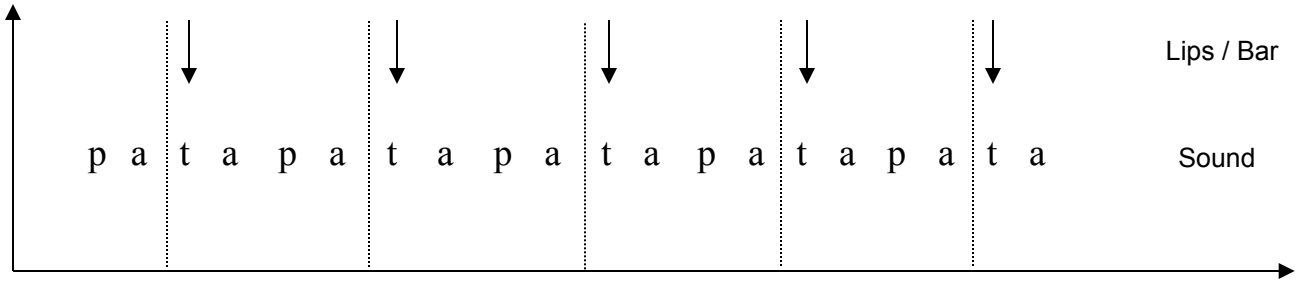
Results:

- Percept /pata/ more stable in AVpa than in Avta
- Percept /tapa/ more stable in AVta than in Avpa

The effect seems speech specific (Basirat et al., 2012):

Less effect with bars than with true lips

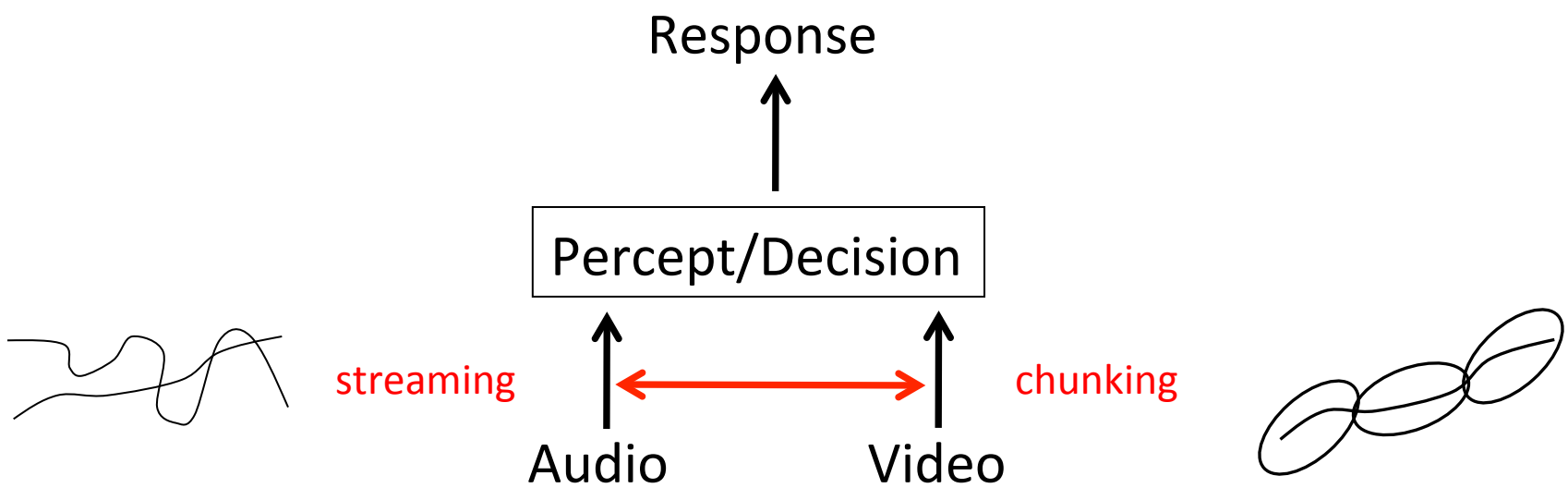




3. AV chunking and speech segmentation

See poster by Antje Strauss,
poster session 3,
Sunday afternoon

Conclusion: early audio-visual binding,
achieving audiovisual speech scene analysis,
including both audiovisual streaming
and audiovisual chunking (segmentation)



1. Audiovisual Speech perception without scene analysis?

2. AVSSA: experimental data about streaming and chunking

3. Possible theoretical, neural and computational bases for AVSSA

What are the underlying mechanisms for AVSSA?

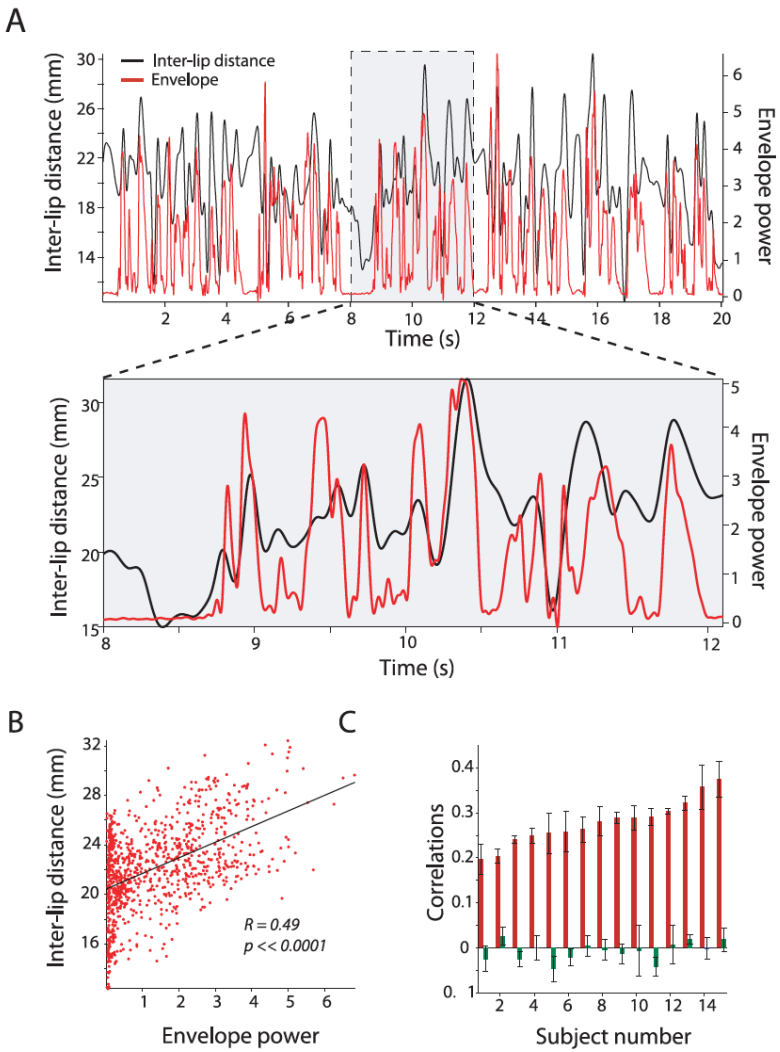
1. AVSSA building blocks
   - Primitives / Schemas ?
   - Gestalt / Common fate ?

2. Underlying neural processes ?

3. Computational mechanisms ?

1. AVSSA building blocks

Primitives based in comodulations in time (the basic Gestalt)



Yehia et al, 1998
Barker & Berthommier, 1999
Grant & Seitz, 2000
Chandrasekaran et al, 2009

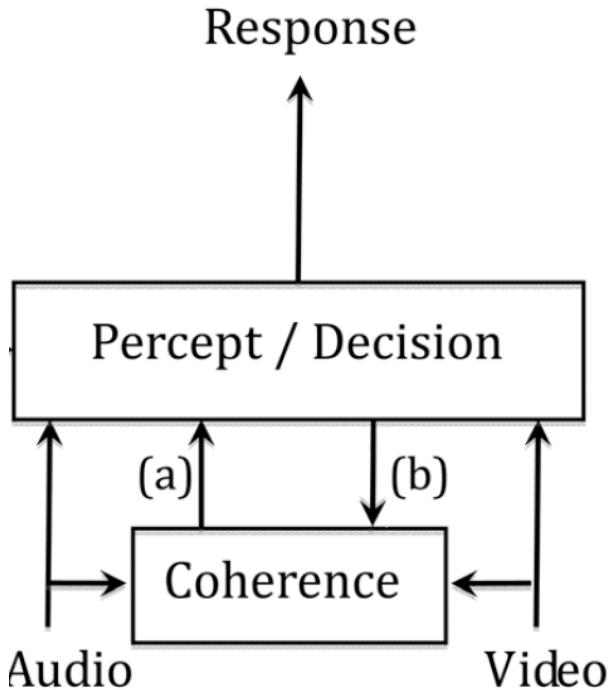
The common fate for audiovisual coherence should be driven by motor processes

Motor primitives or motor schemas?

Learned or innate audiovisual coherence?

Learned of innate perceptuo-motor resonance mechanisms?

(see the innate active intermodal mapping AIM model by Meltzoff)



Co-modulation in time primitives

Motor primitives/schemas

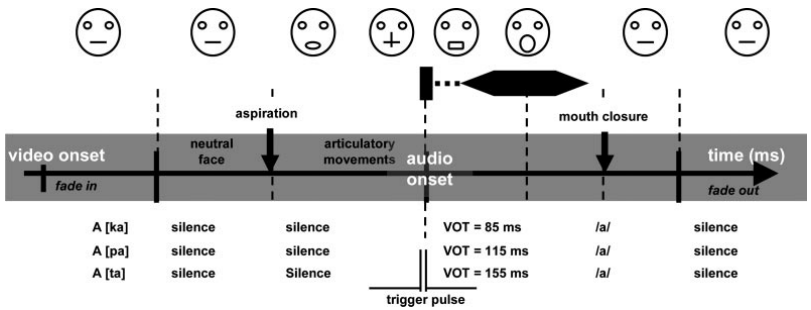
Linguistic schemas

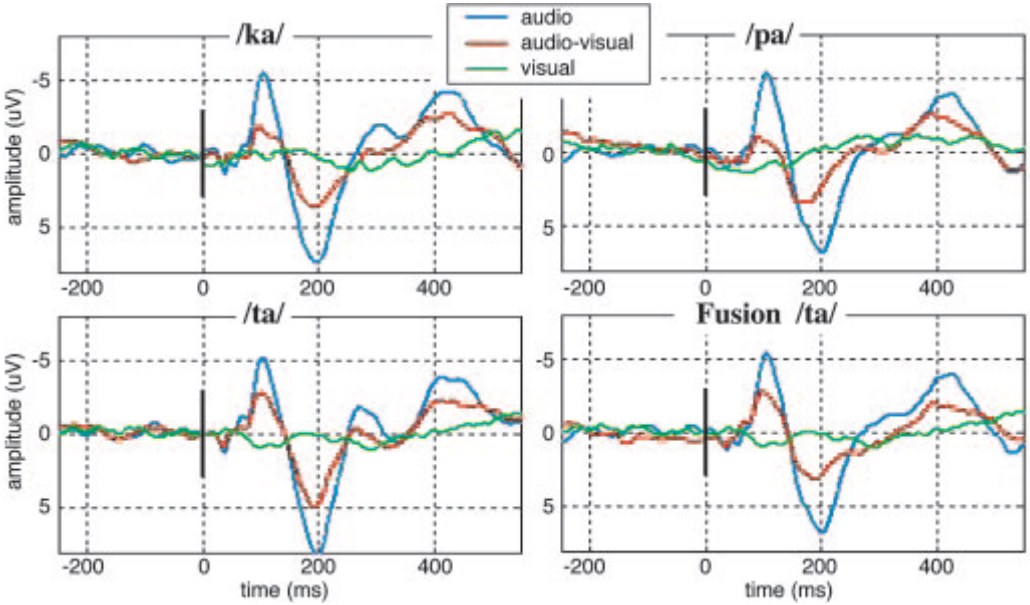
2. Underlying neural processes

Predictive coding

Multiplex coding

Predictive coding and the N1-P2 AV effect





Vision "predicts" audition

Though this does NOT require that V precedes A
(Schwartz & Savariaux 2014)

The prediction effect is a binding effect … which can be blocked by unbinding
(Ganesh et al. 2014)

Besle et al. 2004, **Van Wassenhove et al. 2005,** Arnal et al., 2009

Multiplex coding and the phase-resetting mechanisms

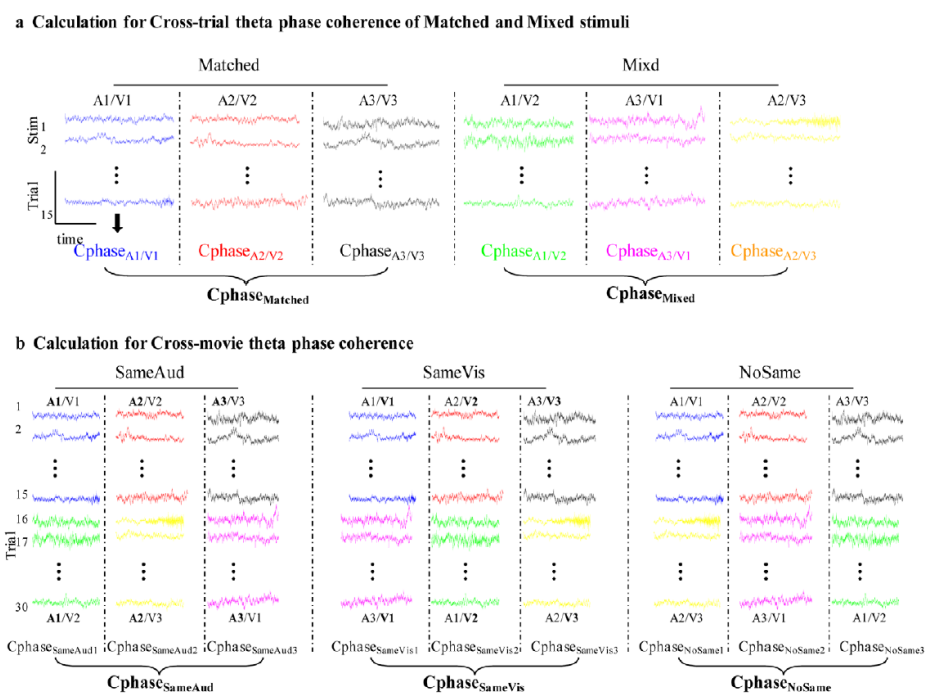
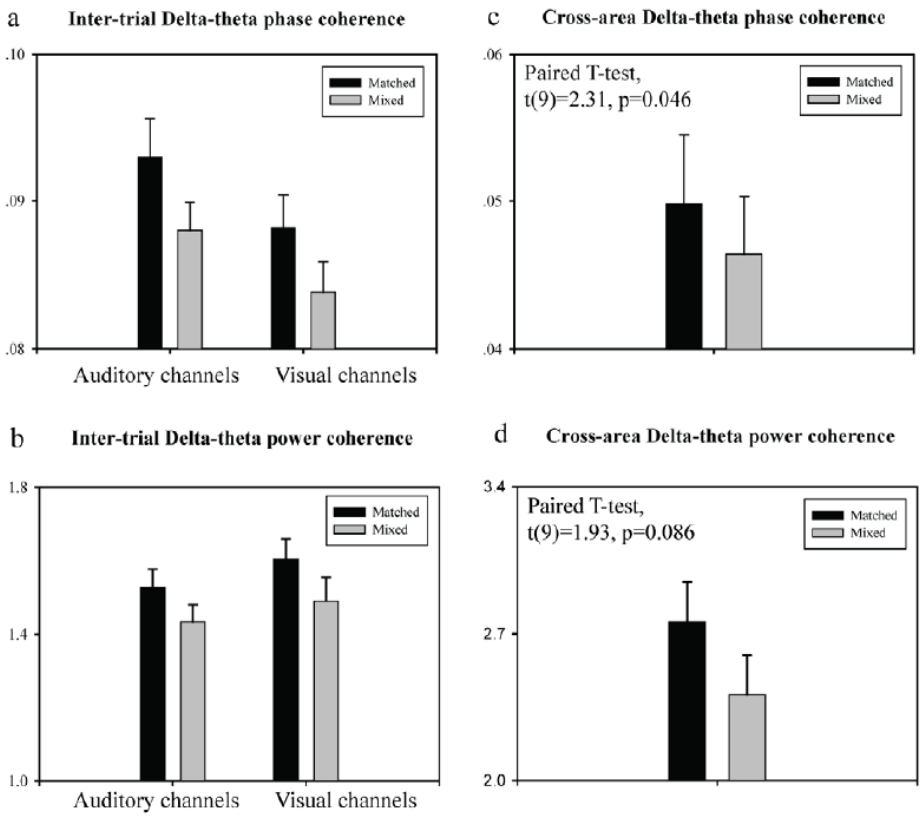


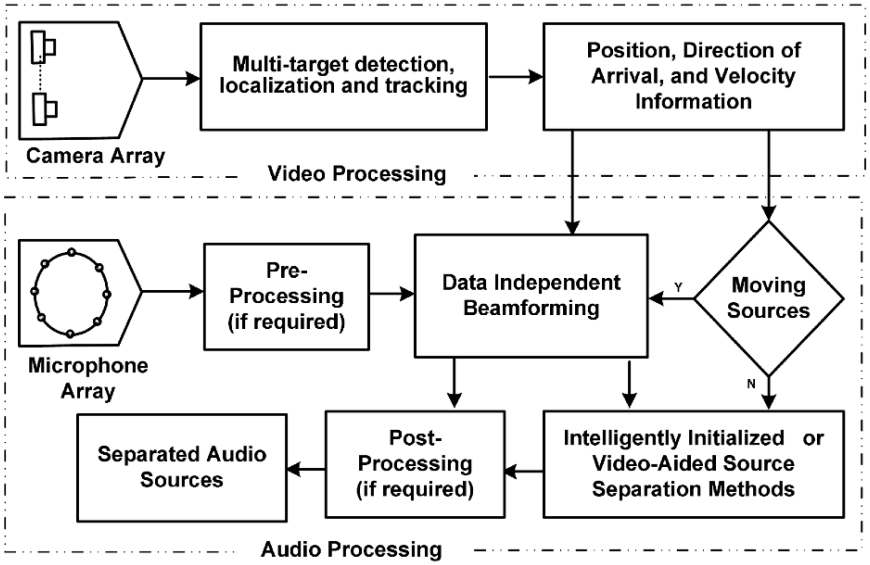
Figure 1. Schematic illustrating experimental design and phase coherence analysis (for a single MEG channel). The colors represent single-trial responses to each of the six audiovisual streams. The coherence analyses are performed on each of the 157 MEG channels separately. (a) The cross-trial phase coherence is calculated on all 15 trials of the same stimulus condition (same color) and compared to a mixture of trials (see Methods) to get the phase-based and power-based movie discrimination ability (see Figure 2ab). (b) Cross-movie phase coherence is calculated by combining response trials across two movie stimuli (two different colors in each column), where one dimension is matched in auditory (SameAud), visual (SameVis), or neither modality input (NoSame). See more equation details in Methods.
doi:10.1371/journal.pbio.1000445.g001



Luo & Poeppel, 2010

3. Computational mechanisms

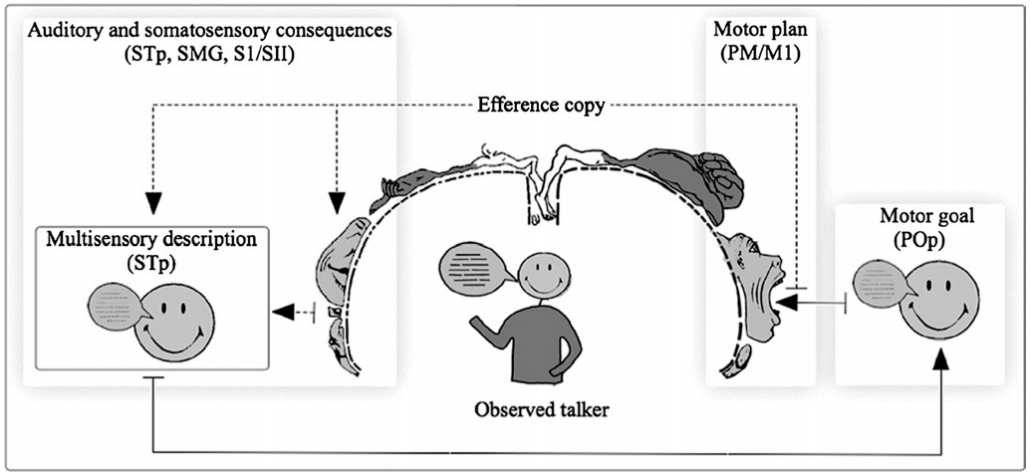
Audiovisual speech source separation



Girin et al., 2001
Sodoyer et al., 2004
Rivet et al.,2007, 2014

Fig. 2. Block diagram of visual scene analysis based method for speech enhancement. Video localization is based on face and head detection. A video tracker is implemented for tracking of multiple humans and based on the MCMC-PF. The output of the video processing is position, direction of arrival, and/or velocity information. On the basis of the visual scene the pre-processed audio mixtures are separated either by a data independent beamformer or intelligently initialized video-aided source separation method. Finally, post-processing is applied to enhance the separated audio sources.

Perceptuo-motor speech perception models

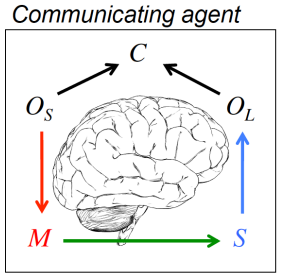


Skipper et al., 2007

The *COSMO* model (Communicating Objects using Sensori-Motor Operations)

A Bayesian model of a communicating agent

$$P(C\ O_L\ S\ M\ O_S) = \underbrace{P(O_S)}_{\text{prior}} \times \underbrace{P(M \mid O_S)}_{\text{motor repertoire}} \times \underbrace{P(S \mid M)}_{\text{forward model}} \times \underbrace{P(O_L \mid S)}_{\text{auditory classifier}} \times \underbrace{P(C \mid O_S\ O_L)}_{\text{communication success}}$$



Moulin-Frier et al., 2012, 2015
Laurent et al., submitted

Ready for the AVSSA revolution?

Thanks for your attention!