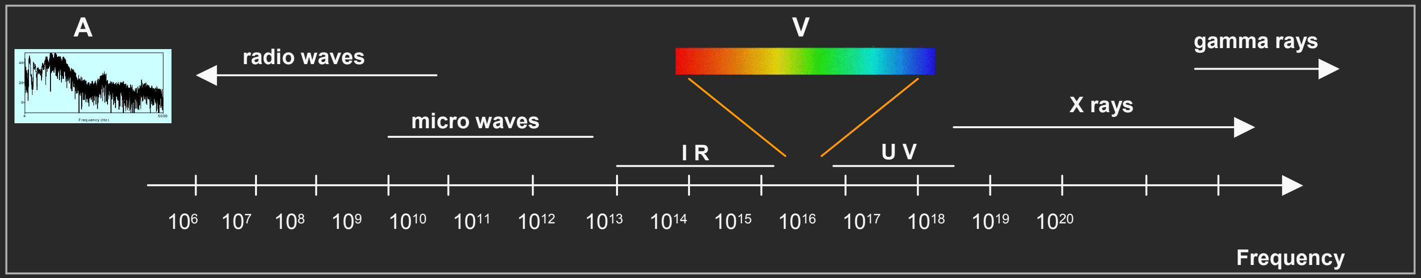
Electrophysiology of Auditory-Visual Speech Integration

A Forward Model of Auditory-Visual Speech

Virginie van Wassenhove
Ken W. Grant
David Poeppel

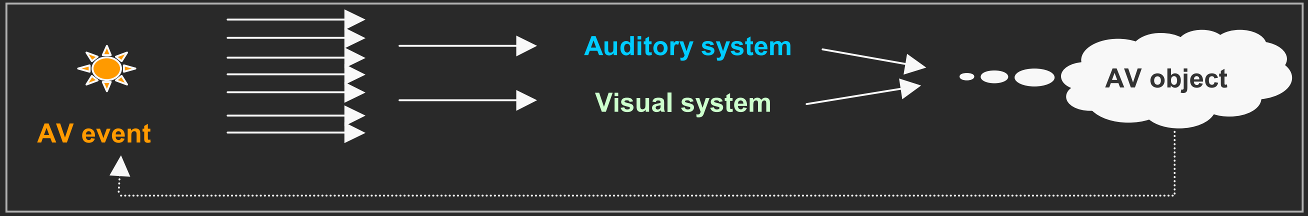
AVSP, St Jorioz 2003

Multisensory Perception

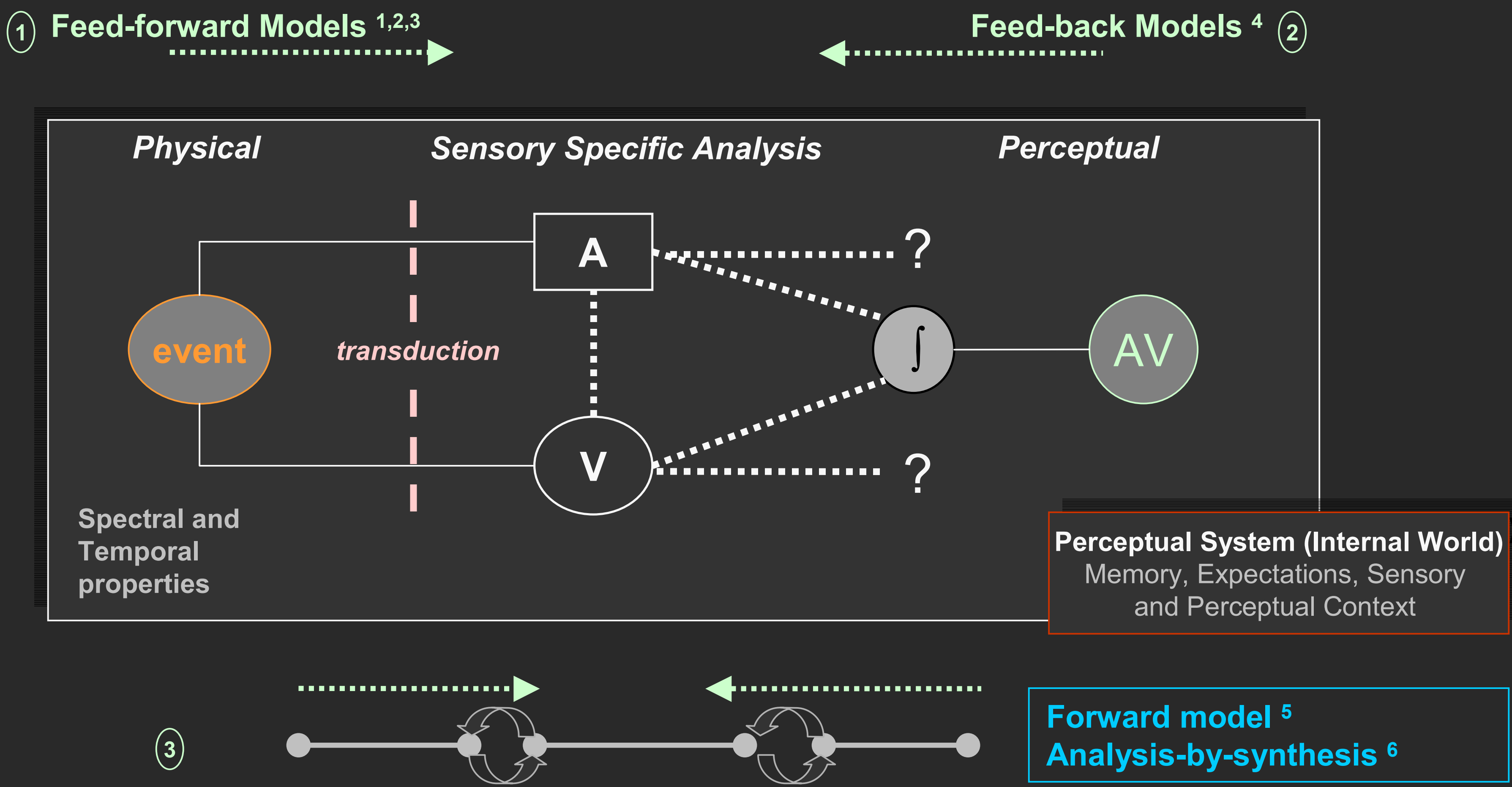


'Physical world' = continuous spectrum of electromagnetic energy

'Inner world' = discrete perceptual representations, sensory-specific or amodal ?



Multisensory Integration



1. Meredith A. (2002) Cognitive Brain Research 14, 31-40.
2. Liberman A.M.(1996) *Speech: a Special Code*. Cambridge, MA: MIT Press.
3. Massaro D.W. (1998) *Perceiving Talking Faces.* Cambridge, MA: MIT Press.
4. Calvert G. *et al.* (2001) Cerebral Cortex 11, 110-1123.
5. Wolpert D.M. *et al.* (1996) Science 269, 1880-1882.
6. Stevens K.N., Halle M. (1965) *in Models for the Perception of Speech and visual forms*. Cambridge, MA: MIT Press.

Space & Time - Universal sensory invariants?

**What information in the physical signals drive the integration across sensory systems?
Physical redundancy?**

→ Neural convergence & 'spatio-temporal coincidence principle' [1]

Co-occurrence of the stimuli in space (location) and time drive the integrative properties of multisensory neurons, or 'supra-additivity'.

→ AV speech [2,3]

Correlation of lip movements and acoustic amplitude envelope has been proposed to cue the integration process (low frequency range ~3-5 Hz).

- **Is 4Hz scale information a *sufficient* constraint for AV speech integration?**
- **Are multisensory sites of integration the seat of perceptual emergence?**
- **Integration : When? How? Where?**

1. Stein B.E. & Meredith A.M. (1993) *The Merging of the Senses*. Cambridge, MA: MIT Press
2. Grant K.W. & Seit P.F. (2000) J. Acoust. Soc. Am. 108, 1197-1208
3. Grant K.W. & Greenberg S. (2001) AVSP Proceedings.

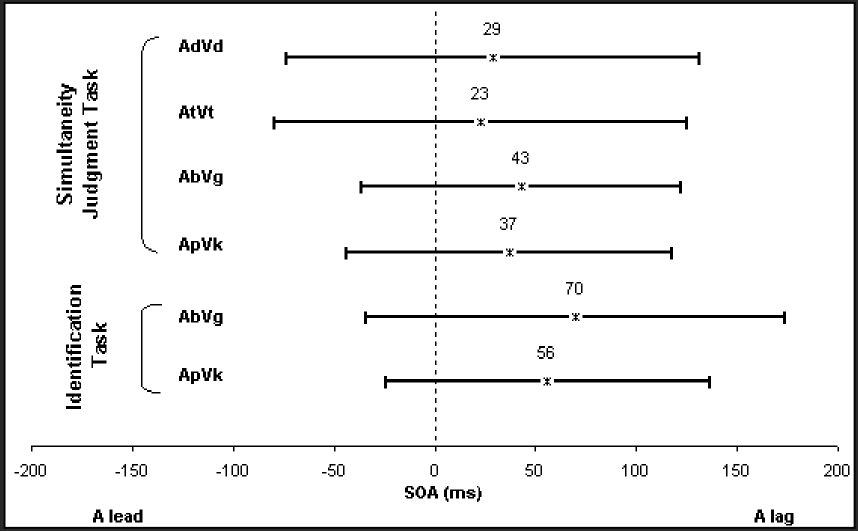
Timing properties of AV speech

- Movements of the articulators naturally precede the auditory speech output.

  Auditory and visual <u>onsets</u> are not 'simultaneous'.

- AV speech integration tolerates signals desynchronization of <u>~250ms</u> and visual leads are less detrimental to integration than auditory leads[1,2].

  Phonetic categorization is processed on a shorter time scale than visemic categorization *<u>i.e.</u>* need for a finer grain scale interaction.



▪ Neural integrative time windows of ~20 and ~200ms have been proposed to mediate featural and perceptual unit formation, respectively[3].

▪ In particular, a fine-grained temporal resolution is necessary for phonetic processing (e.g. voice-onset time and place-of-articulation) while a coarser integrative time window may underlie syllabicity[4].

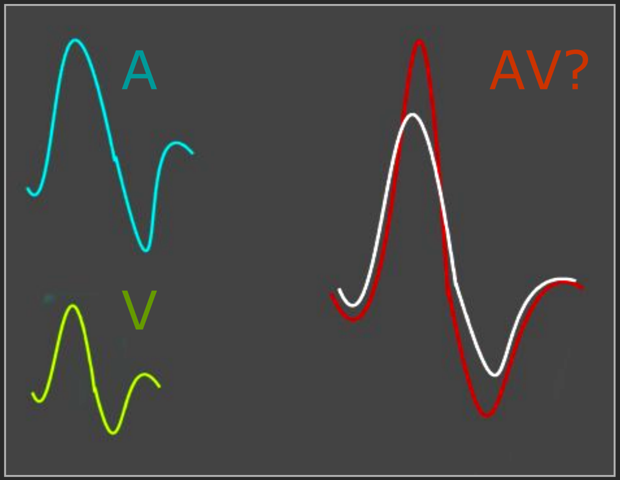
1. van Wassenhove *et al. (2001), Annual Meeting of the Society for Neuroscience.*
2. Conrey B. & Pisoni D.B. (2003) AVSP
3. Poeppel D. (2003) Speech Communication, 41(1) 245-255.
4. Arai T. and Greenberg S. (1997) Eurospeech Proceedings, 1011-1014.

Working Hypotheses

1. **Multisensory <u>supra-additivity</u> was expected early on (~40-90ms) (with reference to non-speech data [1,2])**

2. **Based upon previous fMRI findings[3], we predicted an <u>enhanced amplitude of the auditory event related potentials N1/P2</u> (~100-200ms post auditory onset)**

3. **Incongruent speech was predicted to yield a less enhanced response than congruent AV speech on the basis of spatio-temporal coincidence principle[4] and violation of acoustic amplitude envelope[5].**



Question: Can we find cortical activity that systematically correlate with perceptual changes?

1. Giard M.-H. & Peronnet (1999) Journal of Cognitive Neuroscience, 11(5), 473-490.
2. Calvert *et al.* (1997) Science, 276(5312), 593-596.
3. Stein & Meredith (1993) *The Merging of the Senses. Cambridge, MA:* MIT Press
4. Grant K.W. & Greenberg S. (2001) AVSP Proceedings
5. Sams M. & Aulanko R. (1991) Neuroscience Letters, 127, 141-147.
6. Colin *et al.* (2002) Clinical Neurophysiology, 113, 495-506.
7. Lebib *et al.* (2003) Neuroscience Letters, 341, 185-188.

Experimental Design

Natural Stimuli
audio alone A
video alone V
congruent AV

/ka/, /pa/, /ta/

incongruent [1] McGurk fusion A_pV_k

Task 3AFC [ka] [pa] [ta]

Instructions

Unimodal "Identify (A) what you hear or (V) what the person is articulating"

Bimodal "Identify what you hear while looking at the face " (conversational setting)

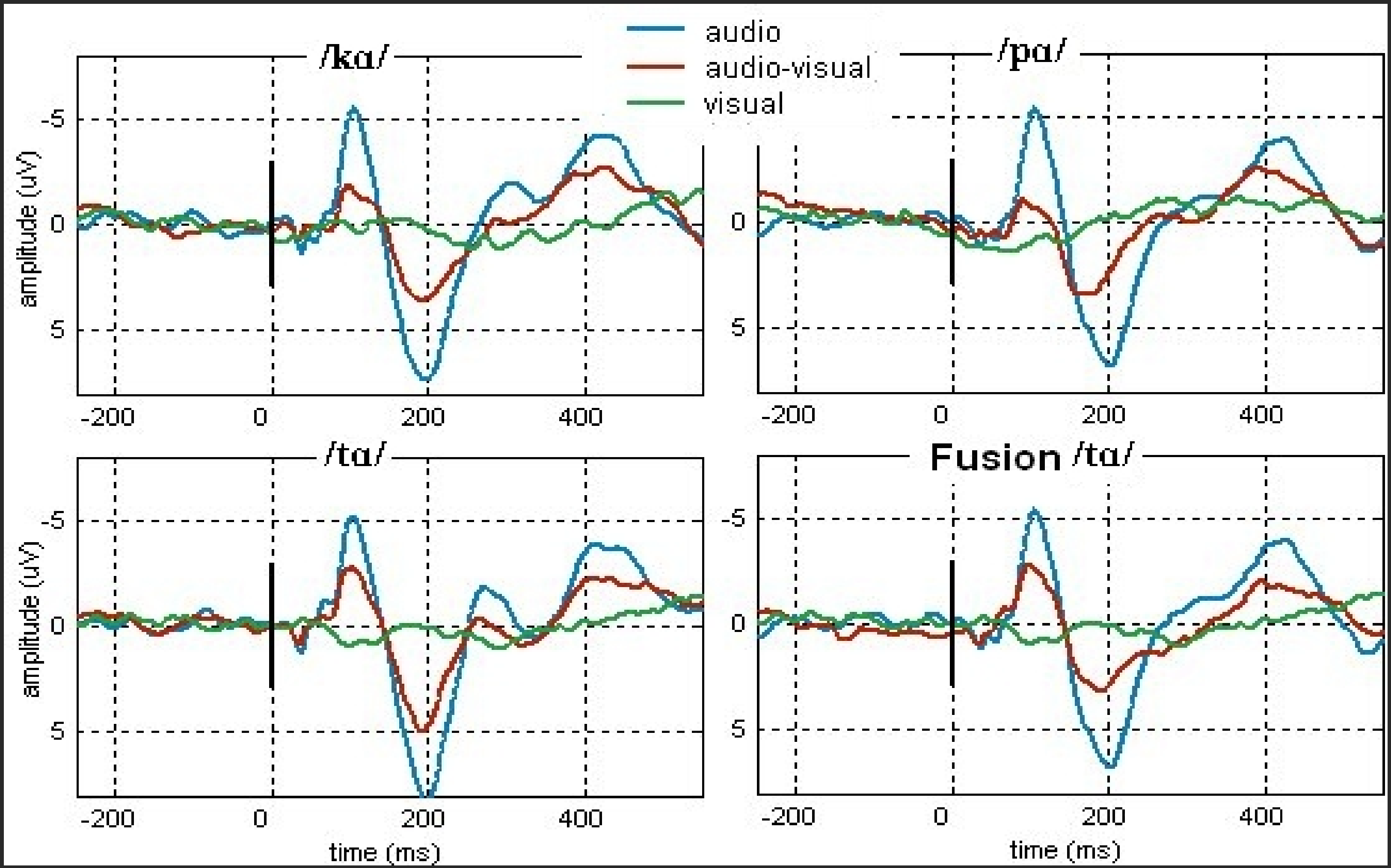
Visual attention "Identify what you see while ignoring the sound"

**No strategy was stated to the participant (i.e. participants were never asked to lip-read nor advised to)*

- **Experiment 1 Block design**
  unimodal (A,V) intermixed – bimodal (AV) separate blocks (n=16)
- **Experiment 2 Pseudo-random design**
  unimodal and bimodal intermixed (A, V, AV) (n=10)
- **Experiment 3 Visual attention** in incongruent speech
  (n=10, also took part in Experiment 1)

1. McGurk & McDonald (1976) Nature, 264, 746-748.

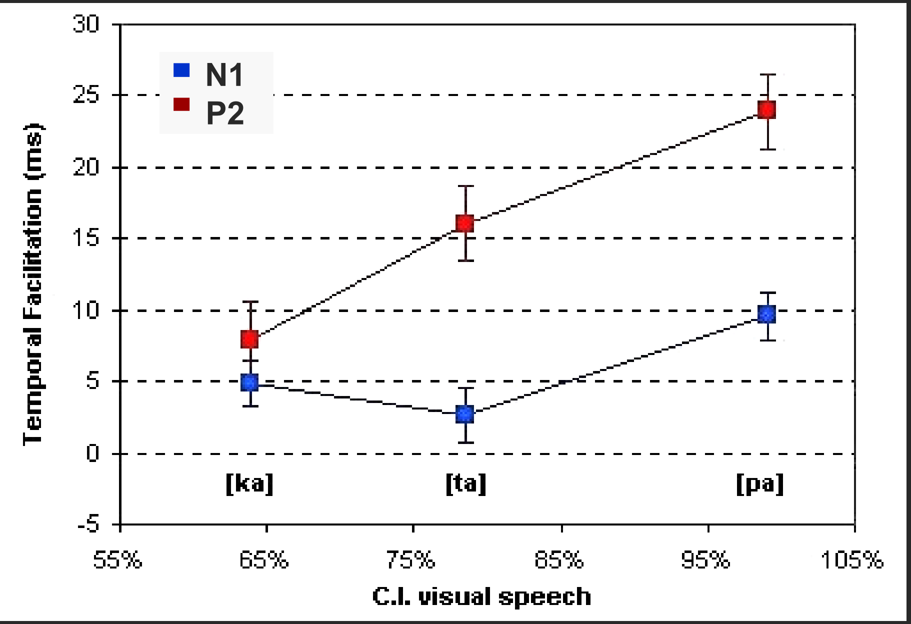
Results – Experiment 1 (n=16)



Visual speech modulates auditory ERPs early on

Temporal Facilitation – Short time scale (~20-50ms)

The rate of correct identification in visual alone condition predicts the degree of temporal facilitation of the N1/P2 complex.



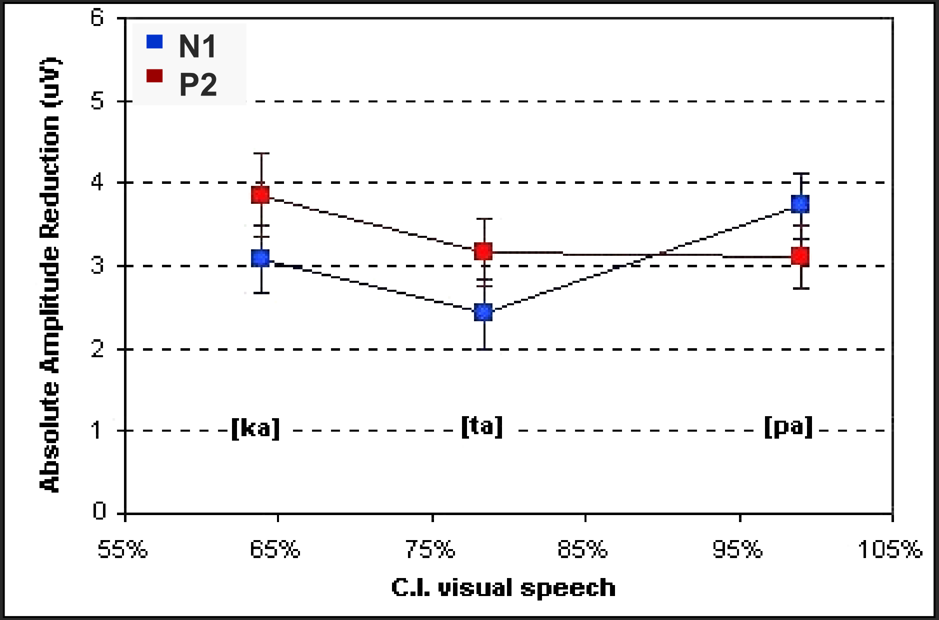
Temporal facilitation (ms) of the N1 and P2 peak latencies as a function of correct identification in visual alone condition.

Latency of N1 (and P2) in AV conditions was subtracted from the latency of N1 (and P2) in their respective A condition. A positive value indicates that AV is faster than A.

Visual speech modulates auditory ERPs

Amplitude reduction - Long time scale (~250ms)

Contrary to the temporal facilitation of the N1/P2 , the amplitude reduction did not depend upon visual ambiguity and was similar for all tokens. The amplitude decrease was observed over the entire N1/P2 complex (up to ~350ms) but not before (i.e. we did not observe a P50 amplitude decrease for these stimuli).



Amplitude decrease (uV) of the N1 and P2 peak amplitude as a function of correct identification in visual alone condition.

Amplitude of N1 (and P2) in AV conditions was subtracted from the latency of N1 (and P2) in their respective A condition. A positive value indicates that AV is smaller than A.

Why no supra-additivity?

Auditory specific event-related potentials not *a priori* originating from multisensory neurons.

Recent data suggest a distributed network[1,2]:

- **intersensory suppression** of unisensory cortical sites
- **enhancement in multisensory** subcortical and cortical sites

AV speech (vowels) has been shown to lead to early suppressed responses at 50ms post-auditory onset[3] (but we did not replicate this observation at P1).

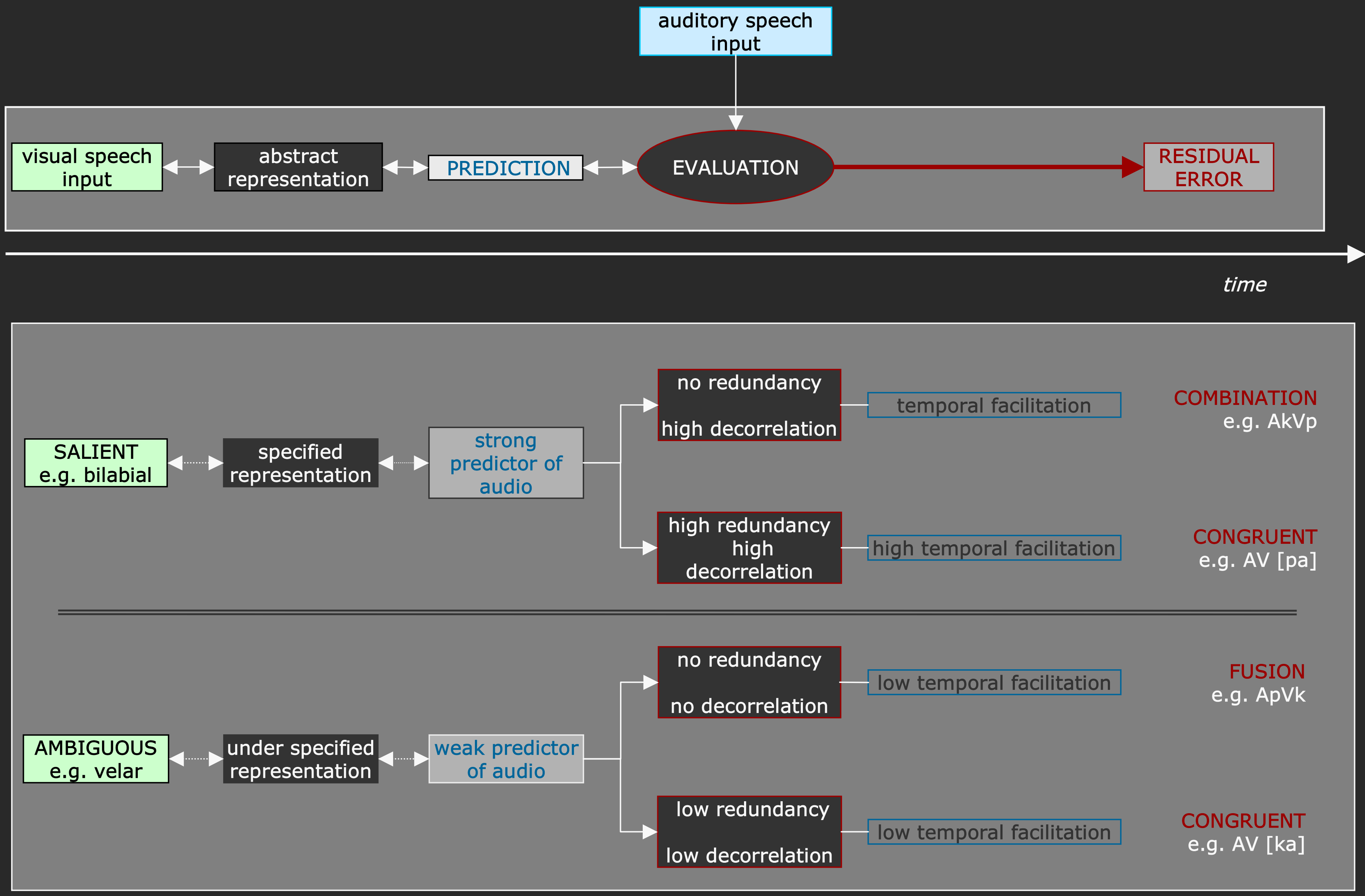
Implications for models of AV speech perception (and multisensory perception)

Early dependency of sensory-specific neural processing.

Intermediary ***abstract representation*** (i.e. amodal) needs to be postulated to account for electrophysiological data.

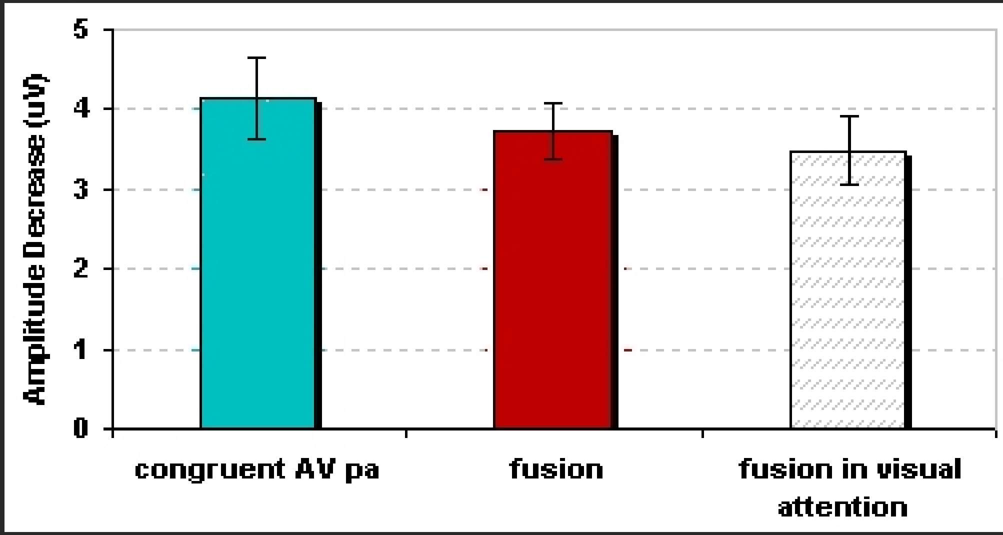
1. Laurienti *et al.* (2001) Journal of Cognitive Neuroscience, 14(3), 420-429.
2. Bushara *et al* (2003) Nature Neuroscience, 6(2), 190-195.
3. Lebib *et al.* (2003) Neuroscience letters, 341, 185-188.

Forward model AV speech perception



Intersensory bias and incongruent Speech (1)

Predictions:
(1) Similar amplitude reduction in congruent and incongruent conditions.
(2) Little-to-no temporal facilitation of audio /pa/ dubbed onto visual /ka/.

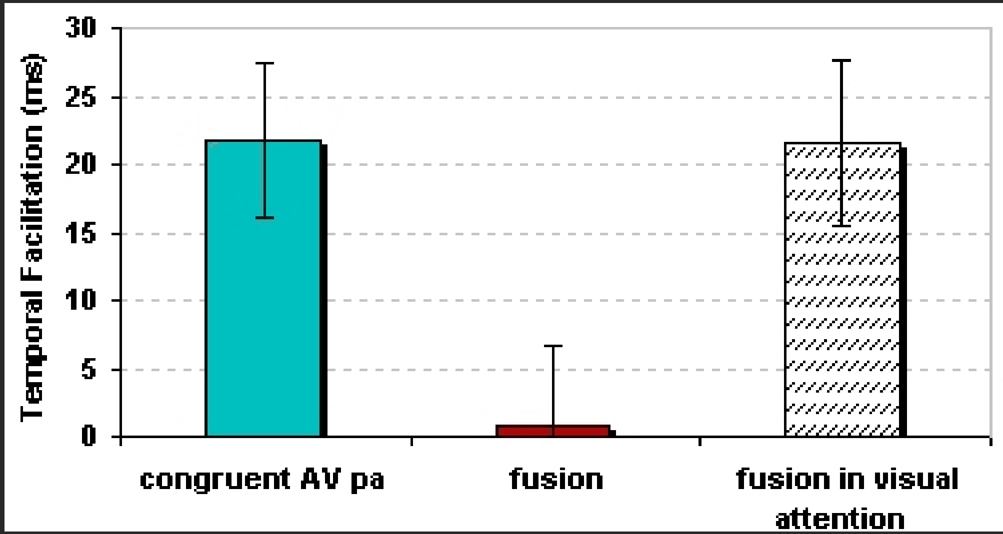


As predicted, no significant amplitude variations was found across AV stimuli, regardless of

(i) attended modality,
(ii) AV incongruency and,
(iii) stimulus identity (as observed in experiments 1 and 2)

Intersensory bias and incongruent Speech (2)

Predictions:
(1) Similar amplitude reduction in congruent and incongruent conditions.
(2) Little-to-no temporal facilitation of audio /pa/ dubbed onto visual /ka/.



In <u>experiment 1 and 2</u>, little-to-no temporal facilitation was observed in fusion (red) as compared to congruent AV /pa/ (blue).

In <u>experiment 3</u>, the temporal facilitation was recovered in fusion (gray) despite the AV incongruency and the low predictive value of visual /ka/.

These results suggest that the *weight of the predictor* at the evaluation stage depends upon attended modality, in agreement with the notion that in conflicting multisensory situations, the non-attended modality (here auditory) is increasingly biased with directing attention to the other modality (here visual).

1. Welch R.B. & Warren D.H. (1980) Psychological Bulletin, 88, 638-667.

Conclusions

1. **The more salient the visual speech is, the faster the auditory speech is processed (~10-30ms of temporal facilitation).**

2. **AV speech engages in a bimodal mode of processing, marked by a deactivation of the auditory cortex spreading over ~250ms, independently of (i) the identity of the speech stimuli, (ii) their congruency and (iii) attended modality.**
   (Further experiments are needed to specify the origin of this deactivation (e.g. threshold variation, reduced number of neurons,…)

3. **A forward model of AV speech is proposed, that integrates the idea of analysis-by-synthesis and neural predictive coding as primary computational strategies.**

4. **No sensory-specific supra-additivity was found.**

Thank You! ☺