TIMING OF AUDITORY- VISUAL INTEGRATION IN THE MCGURK EFFECT

V. Van Wassenhove*1 ; K.W.Grant 2 ; D.Poeppel 1

1.Dept Biol, Univ Maryland CNL Lab, College Park, MD, USA

2. Army Audiology and Speech Center, Walter Reed Army Hospital, Washington, DC, USA

email :vvw@glue.umd.edu

INTRODUCTION

In the following set of experiments, we used the McGurk illusion first reported by McGurk and McDonald (1976) to examine multisensory integration. In its "fusion" component, the illusion emerges when a participant is presented with an auditory bilabial (e.g. /ba/) dubbed onto a visual velar (e.g. articulatory movement /ga/). Under these conditions participants consistently report hearing an alveolar /da/ or /Da/, virtual percept resulting from the AV fusion.

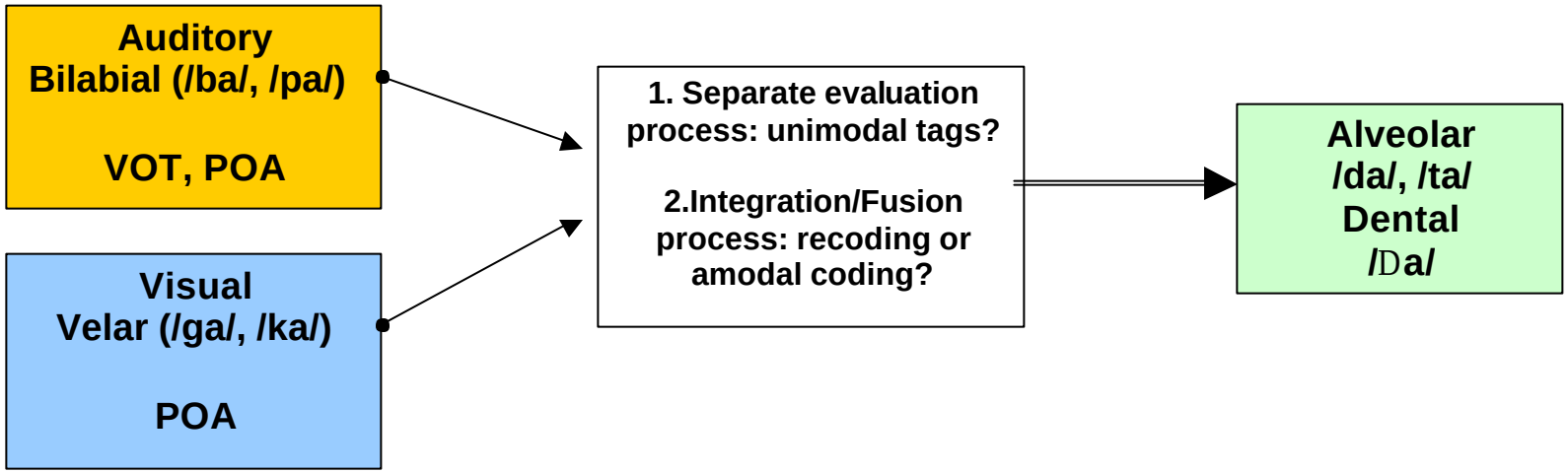
The McGurk paradigm is particularly helpful in quantifying AV integration as a function of the degree of illusion shown by the participant. This paradigm addresses essential questions regarding the cortical code used in the multisensory processing of speech and on the timing requirements for the integration of multimodal speech processing involving different initial sensory pathways.

A fundamental challenge of multisensory information binding lies in the compatibility of codes used by sensory channels. We assume that the sources of information originating from a common event (e.g. AV utterance) must share a minimum of compatible cues in order for bimodal information to merge or interact at the integration stage. As a general rule, objective information is primarily governed by spatial and temporal congruency –that is we tend to subjectively consider a multisensory event as a perceptual unit when signals are in close temporal and spatial proximity.

Spatial disparities have little to no effect on the McGurk illusion, suggesting that spatial mapping of auditory and visual sources -mediated subcortically by the superior colliculus- does not play a major role in AV speech identification (Jones & Munhall, 1997). On the other hand, large timing discrepancies between sensory modalities should intuitively reduce the probability for the information to be bound as a single event.

Sources of information in bimodal speech divide as follows: the **place of articulation** (**POA**) is primarily provided by the visual modality ("visemes") but is also present in the auditory signal (F2/F3 formants transitions). **Voicing** (**VOT**) is entirely provided by the auditory signal. Visual kinematics are the main cue for AV integration (Rosenblum & Saldanã, 1996) and are essentially confined within a short time period corresponding to the consonantal release or POA. Similarly, auditory formant transitions are essential to AV integration (Green and Norrix,

1997). Visual kinematics and F2/F3 rapid frequency shifts at the onset of voicing are thus necessary information for integration in the McGurk illusion.



The following set of experiments focuses on the effect of the temporal asynchrony of these rapid and temporally well-delimited AV signals on the integration process. Temporally misaligned AV McGurk pairs were submitted to (i) identification and (ii) simultaneity tasks. Contrary to previous reported studies (Massaro *et al.*, 1996, Munhall *et al.,* 1996) a temporal window of about 250ms was found within which AV integration reaches an optimal level. Larger time discrepancies reduced the contribution of visual information bias (including about 500ms of auditory lead and auditory lag).

TASK

2 fusion tokens

1- Auditory /ba/ dubbed onto Visual /ga/ A_bV_g
2- Auditory /pa/ dubbed onto Visual /ka/ A_pV_k

29 Timing discrepancies

[-467 ms: +467ms] in increments of 33.33ms ([-14 f: +14 f] frame increment)

2 tasks

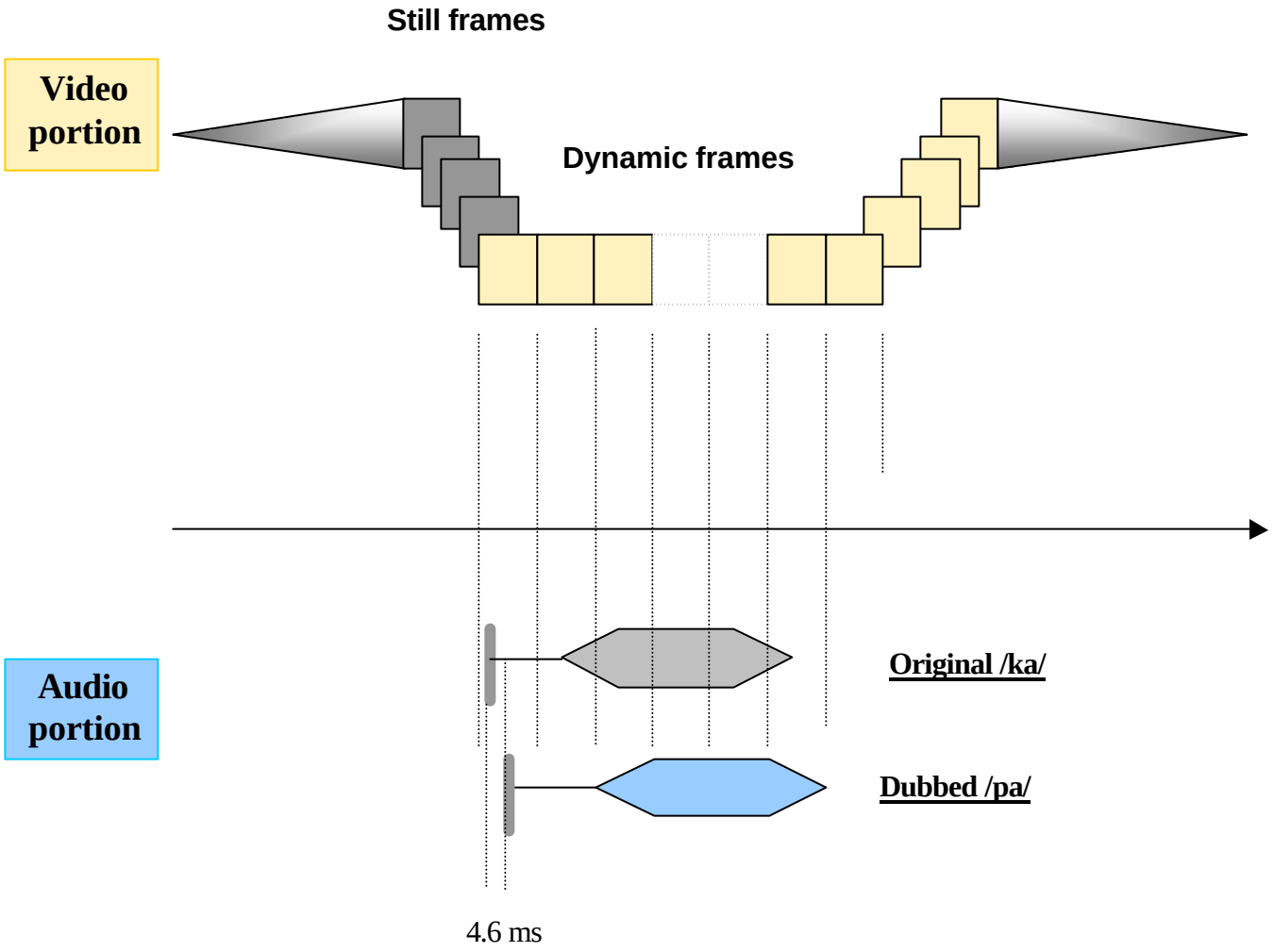
1- <u>Identification Task (3AFC)</u>

		A driven	*Fusion response*	*V driven*
A_bV_g	→	**/ba/**	**/da/ or /∆a/**	**/ga/**
A_pV_k	→	**/pa/**	**/ta/**	**/ka/**

<u>2- Temporal Judgment Task (2AFC)</u>

A_bV_g, A_dV_d
A_pV_k, A_tV_t → **Simultaneous or Successive**

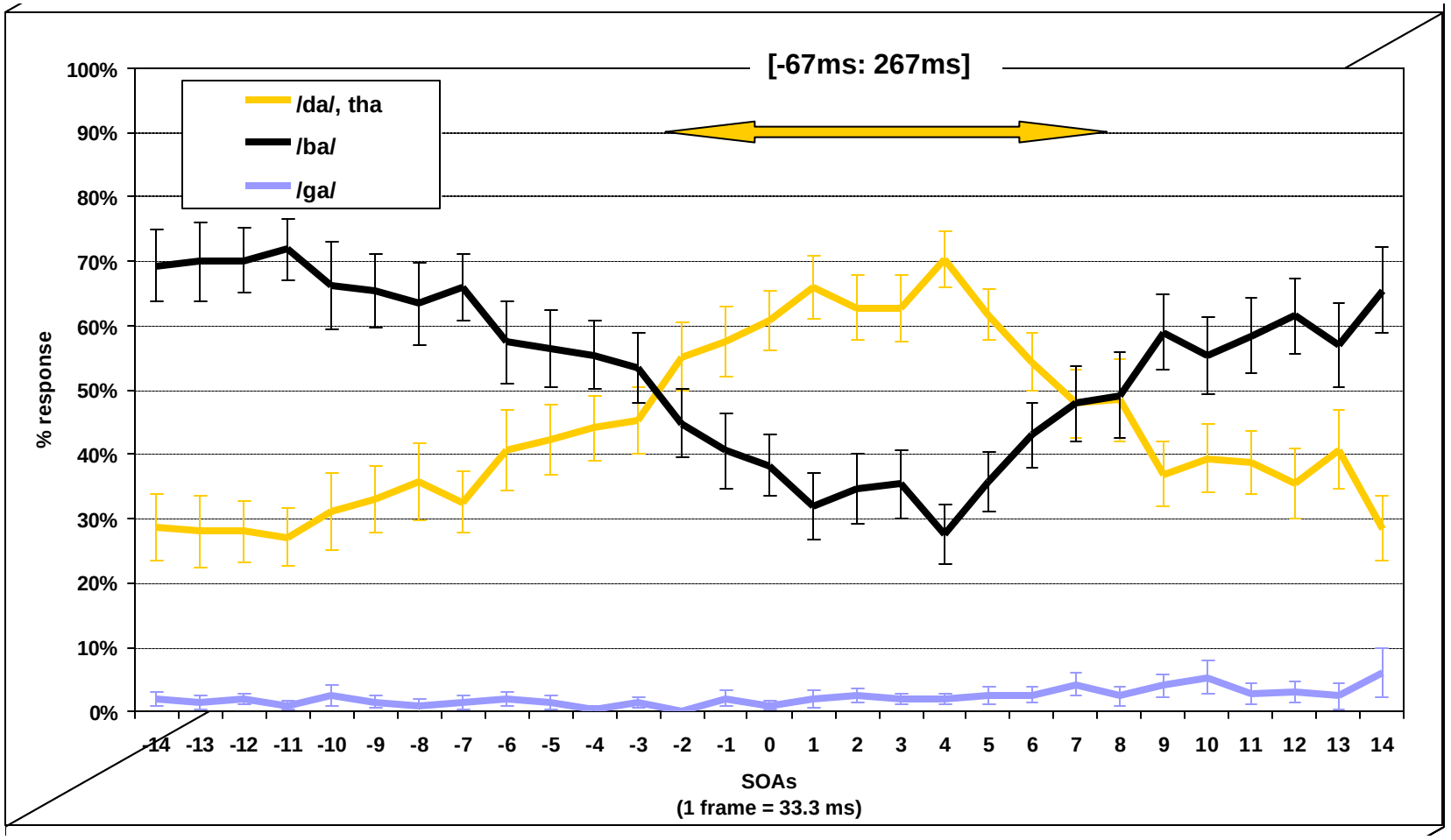
DUBBING PROCESS



STIMULI ALIGNMENT

	SOA (ms)			
Frame SOA	A_bV_g	AV /da/	A_pV_k	AV /ta/
-14	-462.4	-467	-462.4	-467
-13	-427.9	-433	-428.4	-433
-12	-394.4	-400	-395.4	-400
-11	-361.4	-367	-362.4	-367
-10	-327.9	-333	-328.4	-333
-9	-294.4	-300	-295.4	-300
-8	-261.4	-267	-262.4	-267
-7	-227.9	-233	-228.4	-233
-6	-194.4	-200	-195.4	-200
-5	-161.4	-167	-162.4	-167
-4	-127.9	-133	-128.4	-133
-3	-94.4	-100	-95.4	-100
-2	-61.4	-67	-62.4	-67
-1	-28.1	-33	-28.4	-33
0	5.6	0	4.6	0
+1	38.6	33	37.6	33
+2	72.1	67	71.6	67
+3	106.1	100	104.6	100
+4	138.6	133	137.6	133
+5	172.1	167	171.6	167
+6	206.1	200	204.6	200
+7	238.6	233	237.6	233
+8	272.1	267	271.6	267
+9	306.1	300	304.6	300
+10	339.1	333	337.6	333
+11	372.1	367	371.6	367
+12	406.1	400	404.6	400
+13	439.1	433	437.6	433
+14	472.1	467	471.6	467

IDENTIFICATION TASK A_bV_g



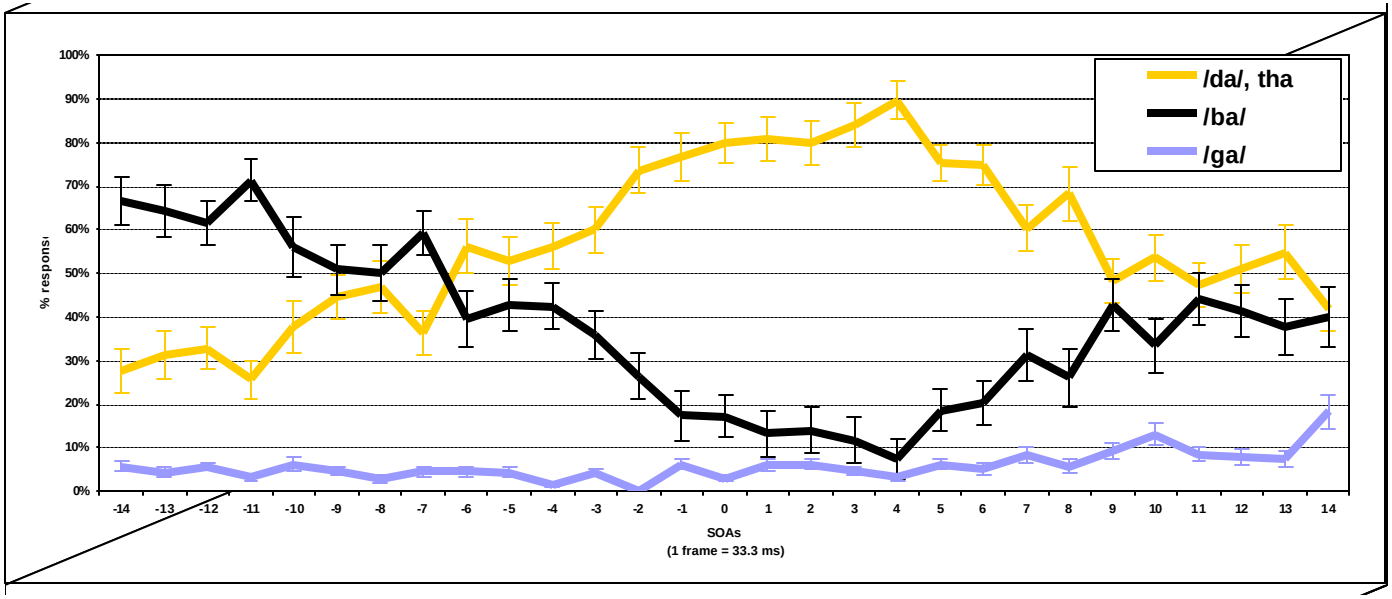
Significant influence of SOAs on response type ($F(1,28)= 6.39$, $P<0.0001$)
Plateau [-67ms: +267ms] ($F(1,10)=1.937$, $P=0.416$)

A_bV_g Group A (N=7)

% Fusion >70% within TWI

Maximum fusion response plateau:
[-67ms: +267ms]

(F(1,10)=1.623, p=.1193)

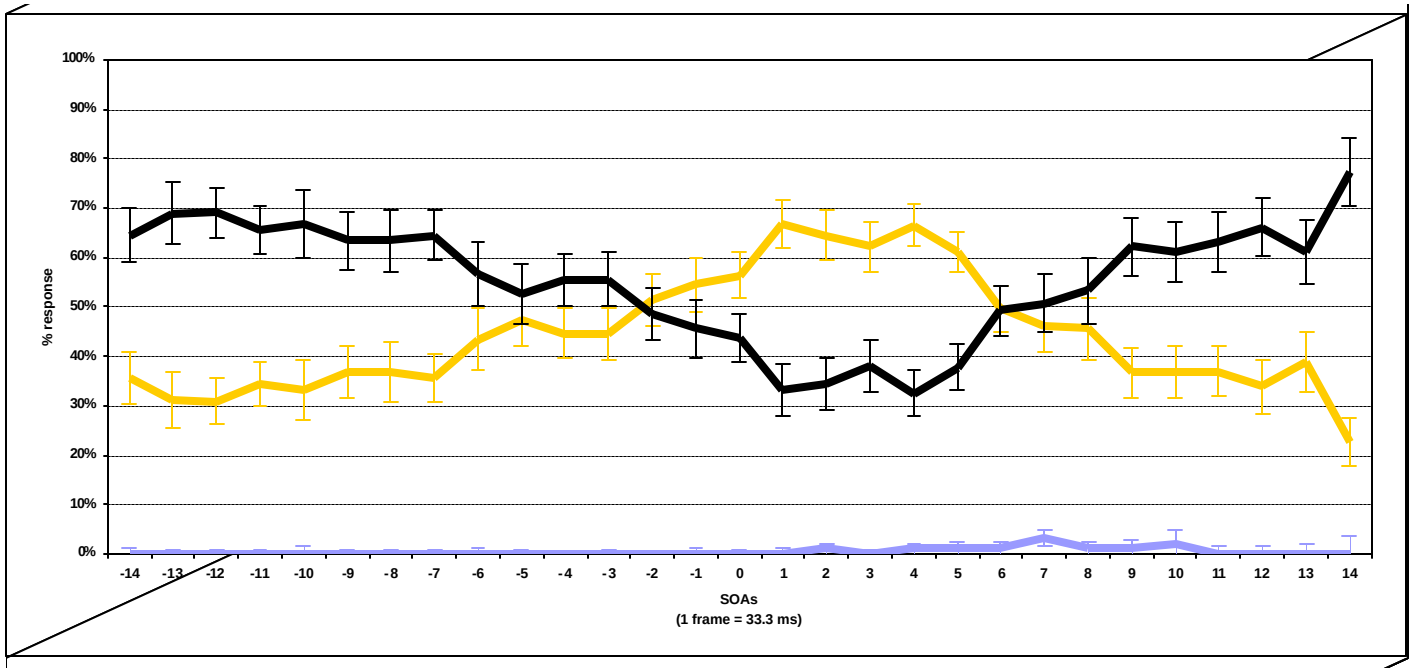


A_bV_g Group B (N=9)

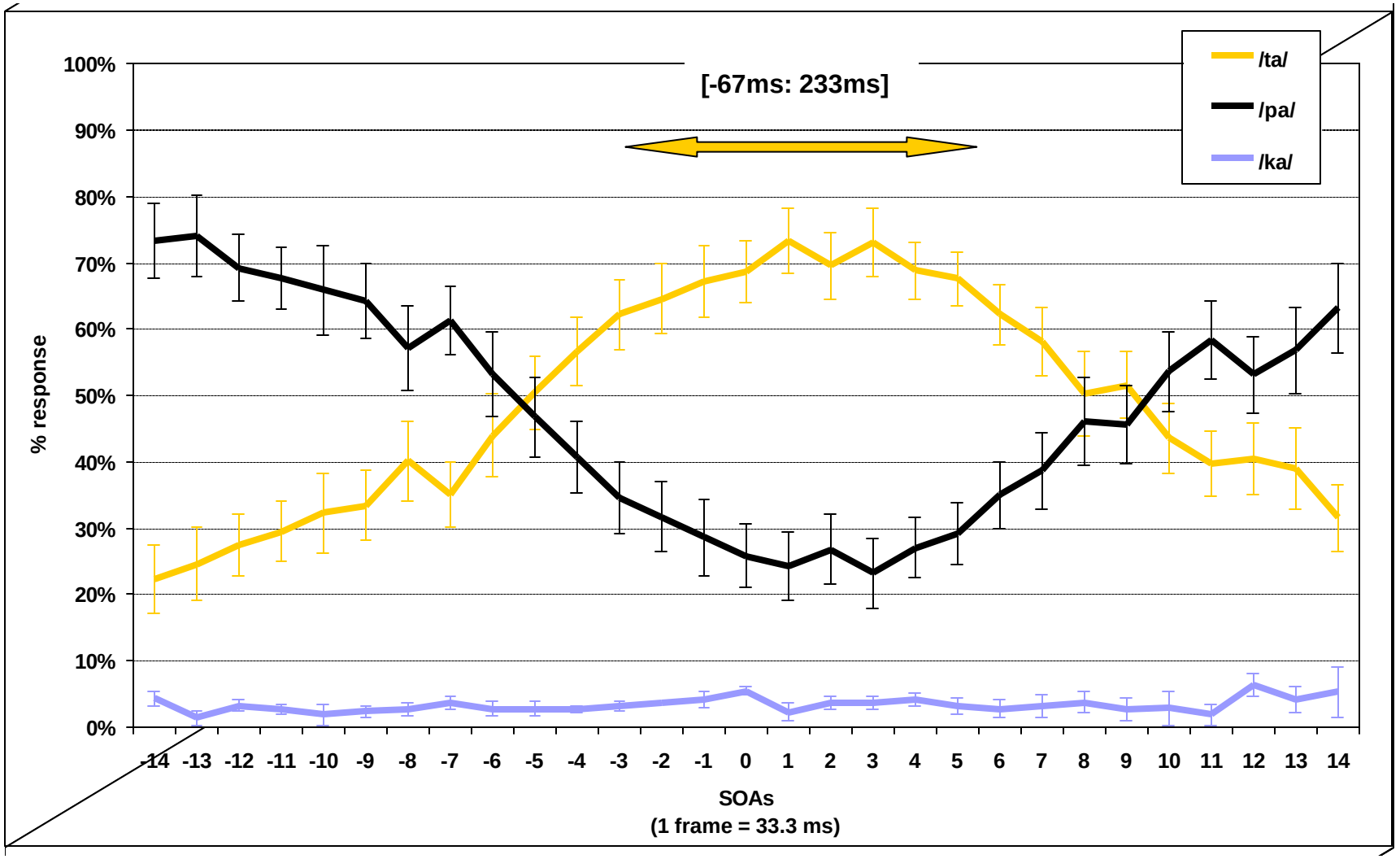
40%< % Fusion <70% within TWI

Maximum fusion response plateau:
[-67ms: +267ms]

(F(1,10)=1.527, p=.143)



IDENTIFICATION TASK A_pV_k



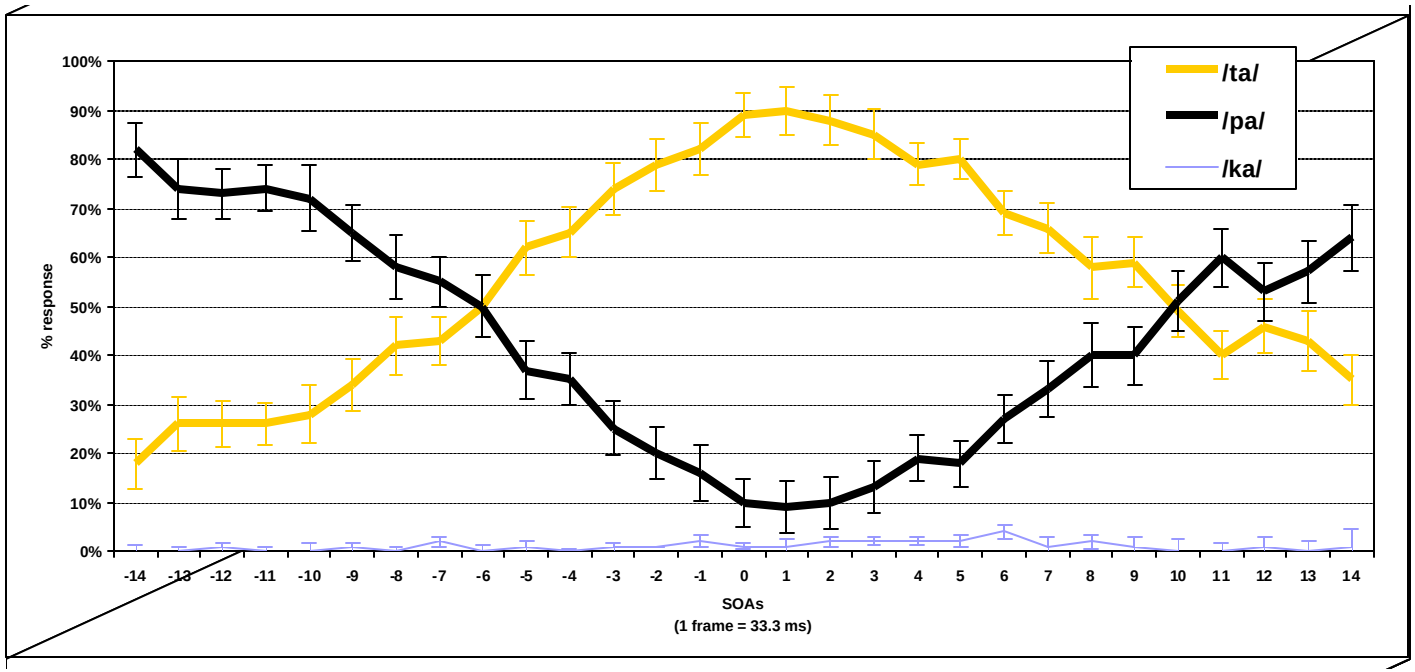
Significant influence of SOAs on response type ($F(1,28) = 16.8$, $P<0.0001$)
Plateau [-67ms: +233ms] ($F(1,10)=1.23$, $P=0.27$)

A_pV_k Group A (N=10)

% Fusion >70% within TWI

Maximum fusion response plateau: [-67ms: +167ms]

(F(7,1)= 1.68, p= 0.1266)

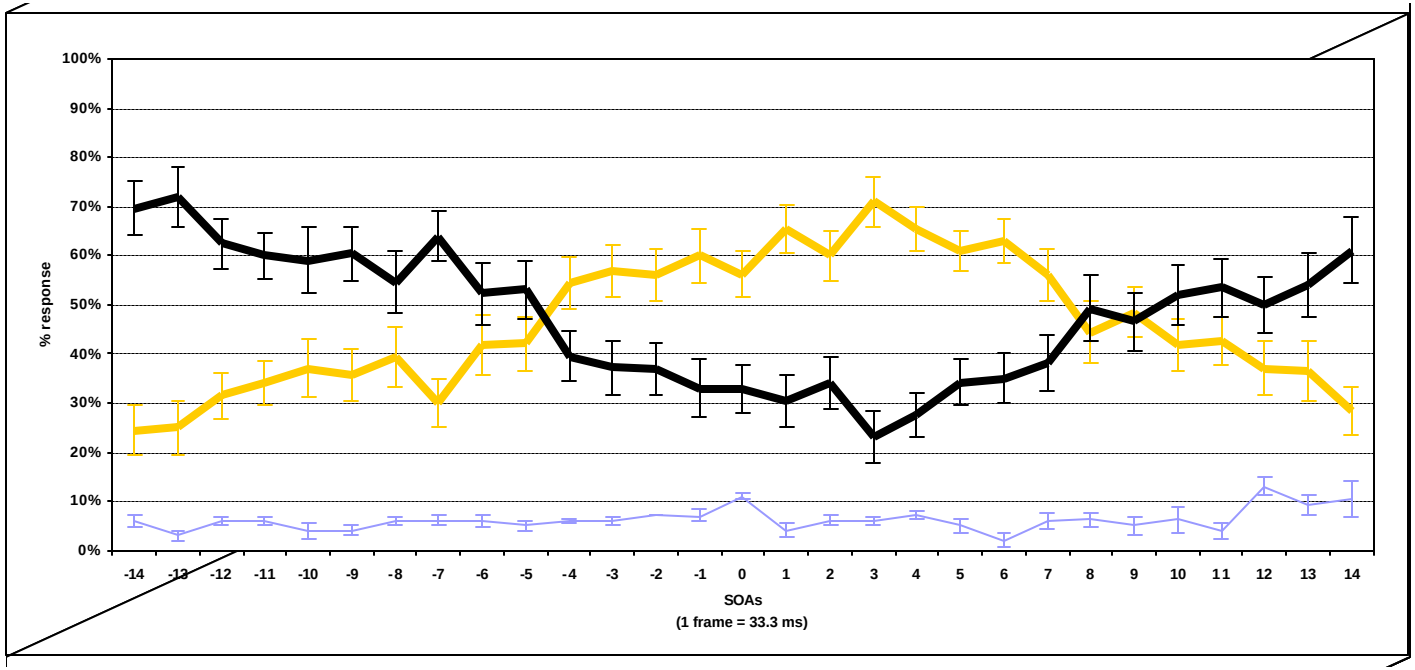


A_pV_k Group B (N=10)

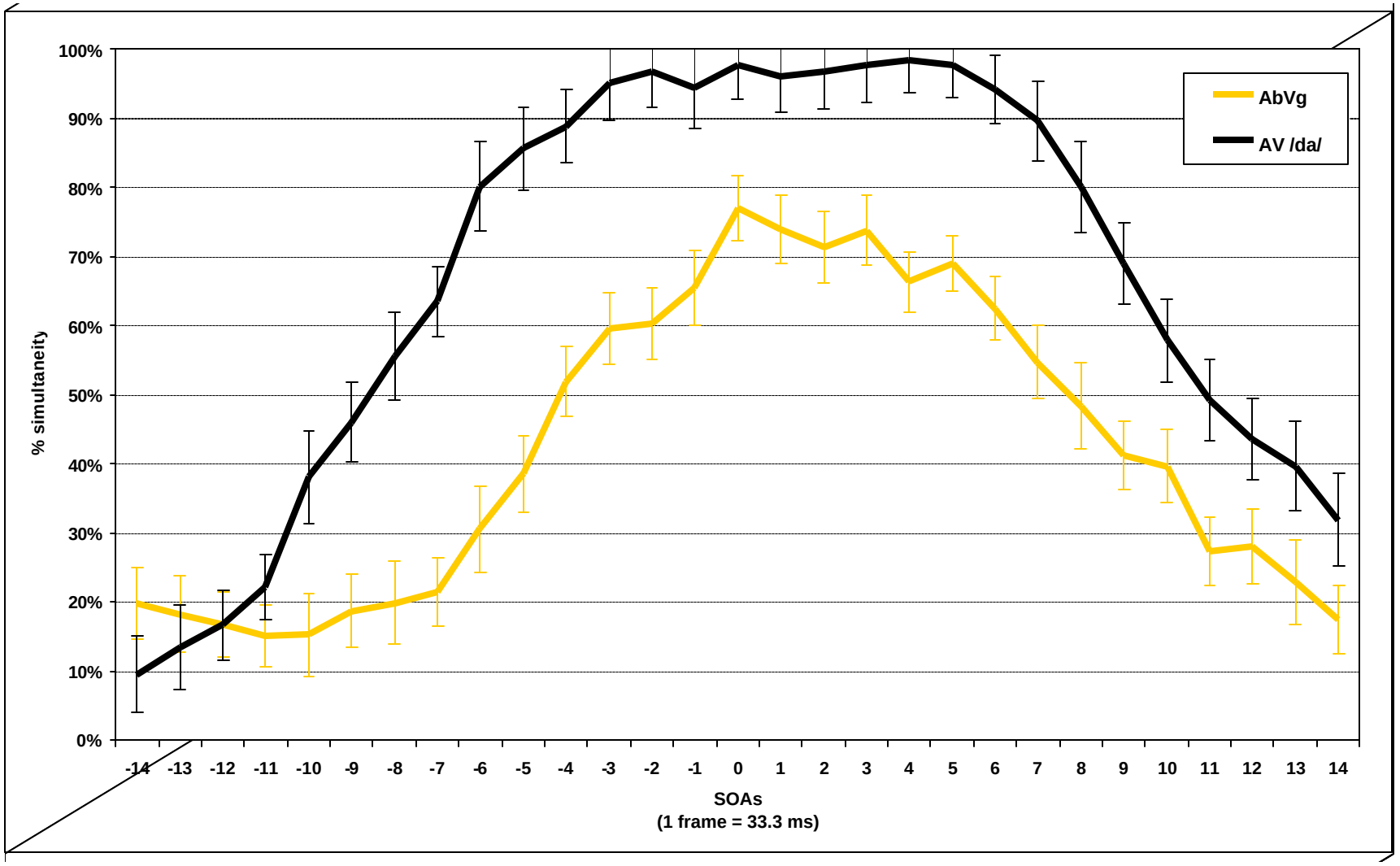
<40 % Fusion <70% within TWI

Maximum fusion response plateau: [-133ms: +233ms]

(F(11,1) = 1.48, p=0.14)



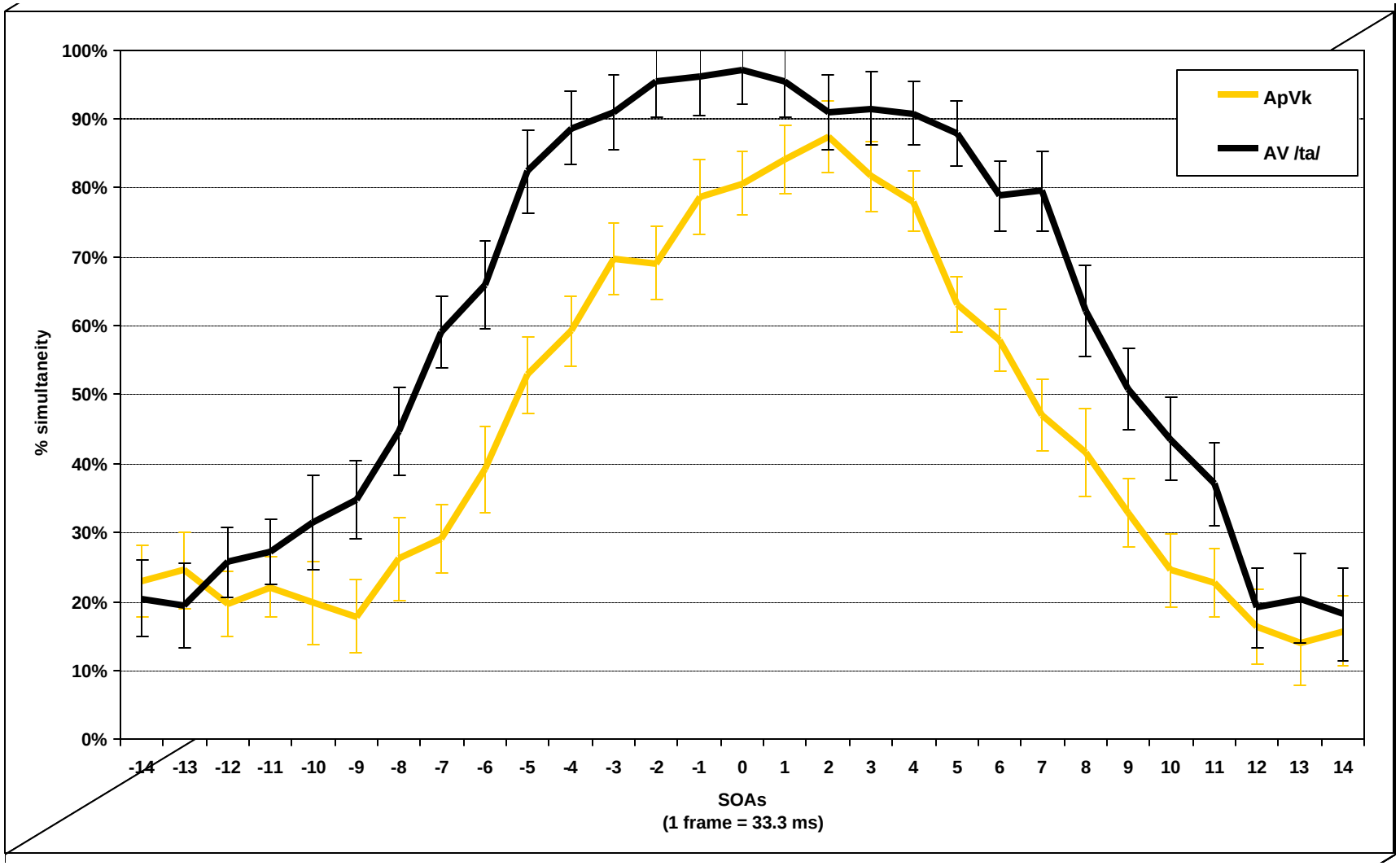
SIMULTANEITY JUDGMENT TASK A_bV_g vs. A_dV_d



Significant difference of simultaneity judgment between A_bV_g and A_dV_d across SOAs ($p<0.0001$)

Plateau A_bV_g [-100ms: +200ms]; Plateau A_dV_d [-100ms: +233ms]

SIMULTANEITY JUDGMENT TASK A_pV_k vs. A_tV_t



Significant difference of simultaneity judgment between A_pV_k and A_tV_t across SOAs ($p<0.0001$)

Plateau A_pV_k [-100ms: +133ms]; Plateau A_tV_t [-200ms: +200ms]

Discussion

- The observed **temporal window of integration** (TWI ~ 250 ms) suggests the existence of intrinsic temporal constraints on the AV integration system. Outside the TWI, visual influence remains surprisingly high (30% fusion), which indicates that AV integration is still possible, in agreement with an early evaluation of AV speech information prior to the integration stage.

- The observed **asymmetry of the TWI** indicates that multimodal inputs follow a different processing time course prior to the integration stage. Because the quality of information content carried on by the two modalities differs, the integration process might not take effect until the visual information has become sufficient to interfere with the initial auditory estimate (that is not before 60 to 100ms of auditory lead). Conversely, the evaluation of visual input limited to visemes tolerates a longer auditory delay to disambiguate the decision process (up to 270ms of auditory lag).

- The **window of subjective simultaneity** shows that larger asynchronies for AV congruent utterances are better tolerated than for McGurk tokens. Interestingly, participants' proportion of fusion reponses remains independent of their proportion of simultaneity judgment. One hypothesis is that the binding of AV features disrupts the temporal judgment, which implies the separation and reevaluation of input signals.

Conclusions

- **Within a temporal window approximating 250ms, stimulus onset asynchronies had no effect on the magnitude of the illusion or on the perception of subjective simultaneity**. These results question the temporal course of the AV evaluation process prior to the integration stage. The Fuzzy Logical Model of Perception (FLMP) described by Massaro proposes a continuous evaluation of auditory and visual information, which is not accounted for by neurophysiological constraints of temporal coding.

- Numerous studies argue for the existence of a temporal window of integration in the auditory system approximating 250 ms. Our results show the existence of a similar window in AV integration thus suggesting that:

1) **Temporal windows could be a general rule of perceptual binding at the cortical level, in one or more modalities.**

2) **The AV integration in speech might be constrained by temporal processing of the auditory cortices.** This argument is consistent with the activity of auditory cortices recorded in silent lipreading conditions (Calvert *et al.,* 1998) and AV speech (Sams *et al.*, 1991).

References

- Calvert G., Brammer M, and Iversen S.D. Cross-modal identification (1998). Trends in Cognitive Sciences 2(7), 247-253.
- Jones J.A. and Munhall K. The effects of separating auditory and visual sources on audiovisual integration of speech. Canadian Acoustics 25(4), 13-19. 1997.
- McGurk H. and MacDonald J. Hearing lips and seeing voices. Nature 264, 746-747. 1976.
- Massaro D.W., Cohen M.M., and Smeele P.M. Perception of asynchronous and conflicting visual and auditory speech. Journal of the Acoustical Society of America 100(3), 1777-1786. 1996.
- Munhall K., Gribble P., Sacco L., and Ward M. Temporal constraints on the McGurk effect. Perception and Psychophysics 58(3), 351-362. 1996.
- Sams M. and Aulanko R. Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. Neuroscience Letters 127, 141-147. 1991.
- Rosenblum L.D. and Saldaña H.M. An audiovisual test of kinematic primitives for visual speech perception. Journal of Experimental Psychology: Human Perception and Performance 22(2), 318-331. 1996.

Support

NIDCD DC 0463801