Electrophysiological profile of auditory-visual speech: ERP study

Virginie van Wassenhove[1], Ken W. Grant[2], David Poeppel[1,3]

[1]Neuroscience and Cognitive Science graduate program, Department of Biology, Cognitive Neuroscience of Language Laboratory, University of Maryland, College Park, MD 20742

[2]Walter Reed Army Medical Center, Army Audiology and Speech Center Washington, DC 20307

[3]Department of Linguistics, University of Maryland, College Park, MD 20742

Introduction

Auditory-visual (AV) speech perception constitutes one of the most complex instances of multisensory integration and remains one of the most challenging issues for existing speech theories. The classic principle of 'spatio-temporal coincidence' (STC) described by Stein and Meredith [1] conditions the 'supra-additive' property of multisensory neurons, which respond to multisensory events with a greater firing rate than would be expected by summation of the neurons' responses to the same stimuli presented unimodally. By analogy to the STC principle, the coherence of lip area and acoustic signal amplitude envelope in bimodal speech emerges as a possible source of inter-sensory correlation [2].

Electrophysiological recordings have suggested early supra-additive effects originating from multisensory areas and sensory-specific cortices in response to non-meaningful AV events such as paired tones/circles (e.g. [3]). Similar conclusions were reached with bimodal speech using fMRI techniques [4a], and lip-reading information was suggested to access auditory cortices [4b]. An electrophysiological account of bimodal speech has not yet been reported.

Bimodal speech needs to be distinguished from classically tested multisensory stimuli in two major ways. Firstly, auditory and visual speech stimuli (phonemes and visemes, respectively) presented unimodally are perceptually categorizable. Acoustic speech signals provide three major phonetic features (voicing, place and manner of articulation), which lead to full phonemic representation and further phonological categorization. Visemic representation is dominated by place of articulation and does not provide voicing information (e.g. /ba/ and /pa/ share the same 'bilabial' visemic class). Consequently, in bimodal speech, information provided by auditory and visual inputs differs in quality and in quantity.

Secondly, the visual speech signal usually precedes the auditory signal by few hundreds of milliseconds during which preparatory articulatory movements occur. AV speech events are therefore inherently asynchronous –i.e. their onsets do not precisely co-occur, visual kinematics leading audio onset. It follows that in bimodal speech, visual information may initiate perceptual categorization earlier than its resulting auditory production, essentially informing on place of articulation.

How do visual speech inputs modify auditory speech processing?

In light of classic multisensory interactions and suggested access of visual speech inputs to auditory cortices,

(1) **a supra-additive effect of auditory evoked potentials (i.e. AV>A+V) is predicted when visual inputs are available.**

(2) **spatio-temporal incongruency between acoustics and kinematics may prevent effective AV integration and lead to non supra-additive effects.**

To test these predictions and establish an electrophysiological profile of simple bimodal speech, congruent and incongruent AV syllables were used[1].

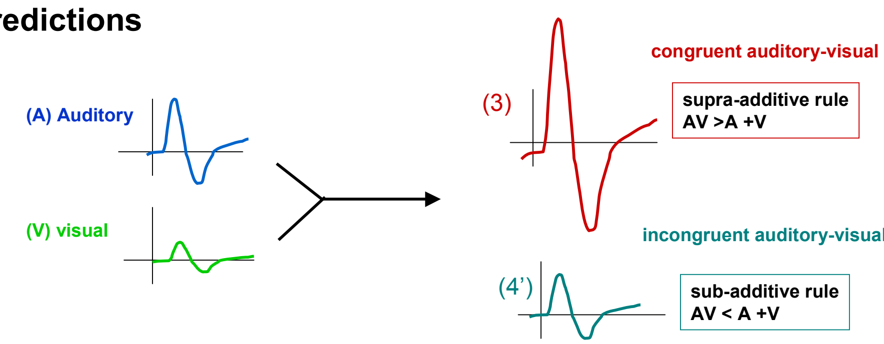
[1] Incongruent Speech:

McGurk Fusion: **audio /pa/** dubbed onto **visual /ka/** leads to **fusion percept /ta/**

McGurk Combination:**audio /ka/** dubbed onto **visual /pa/** leads to **combination percepts** (pka, kapa, paka,etc.)

[5] McGurk and McDonald, 1976

Predictions



(A) Classic auditory evoked response N1/P2.

(V) If visual information access auditory cortices, visual alone should lead to significant activity over auditory cortices.

(AV) According to the supra-additive rule of multisensory integration, combined auditory-visual information should lead to supra-additive enhancement of auditory evoked responses.

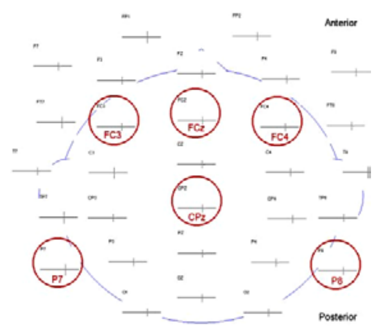
(AV) If the spatio-temporal coincidence rule is violated, the AV integrative process is non-optimal.

3 alternative-force choice (3AFC) identification experiments were conducted while participants were recorded under EEG.

Experiment 1 (n=16) – **BLOCK DESIGN**
Intermixed unimodal audio and visual syllables /ka/,/pa/,/ta/ were tested separately from AV congruent (/ka/,/pa/,/ta/) and incongruent (audio /pa/ dubbed onto visual /ka/) syllables.

Experiment 2 (n=10) – **PSEUDO-RANDOM DESIGN**
The same unimodal and bimodal speech syllables were tested intermixed in a pseudo-random design with different participants.

Experiment 3 (n=10) – **VISUAL ATTENTION**
McGurk fusion (audio /pa/ dubbed onto visual /ka/) and combination (audio /ka/ dubbed onto visual /pa/) AV speech were submitted to identification while subject focused onto visual input.



EEG Recordings Parameters

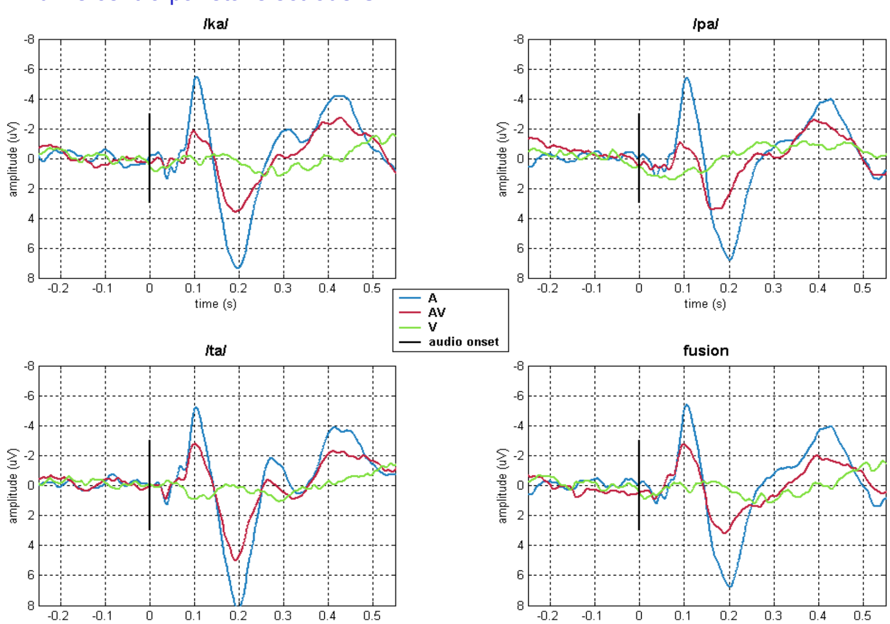
- 32 channels
- AC recording, sampling rate 1kHz
- A/D filter: 1Hz -100Hz
- Gain setting: 1000
- All reported data: ocular artifact reduction, threshold +/-100 µV (~20% rejection), 400ms pre-stimulus baseline correction and BPF 1-55Hz, double-pass Butterworth filter, 24dB cut-off each pass.

Bootstrapping Method

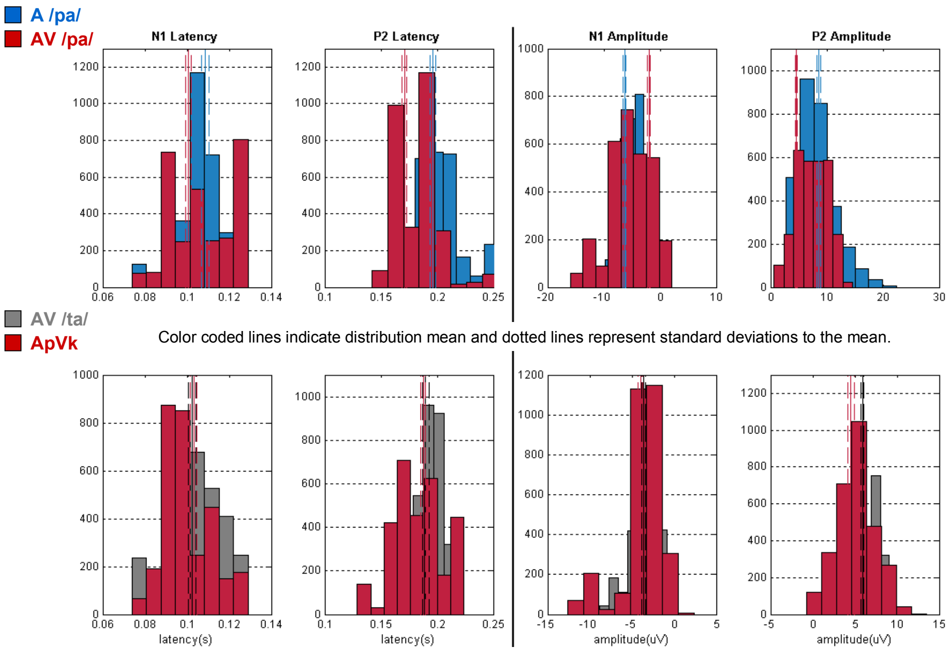
- Principle: 100 original recordings / stimulus / participant are resampled with 300 bootstraps
- Purpose: population estimator of ERPs parameters
- Benefit: individual and group distributions of ERPs amplitude, latency (including dispersion values)
- Applied to 6 channels: CPz, FCz, FC3, FC4, P7, P8

Block Design – Auditory Evoked Potentials (n=16)

midline centro-parietal electrode CPz

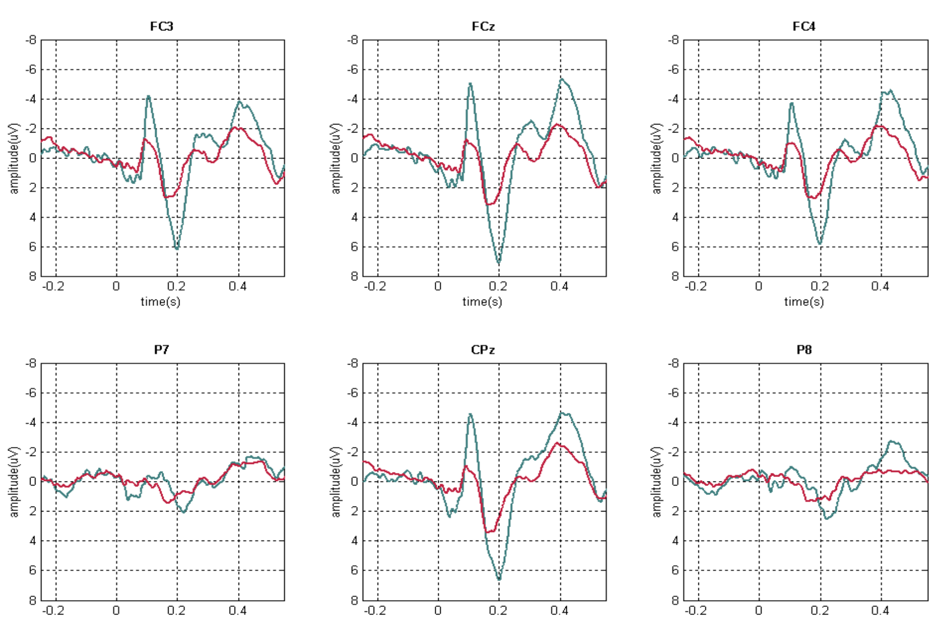


Bootstraps N1P2 latency/amplitude distributions (n=16) CPz



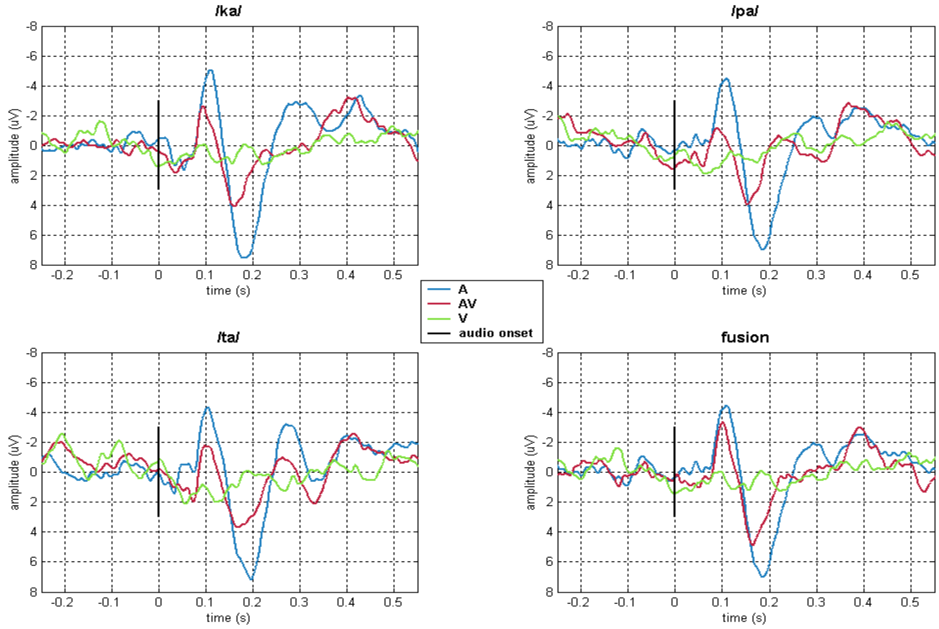
Linearity and Supra-Additivity e.g. /pa/

A+V
AV

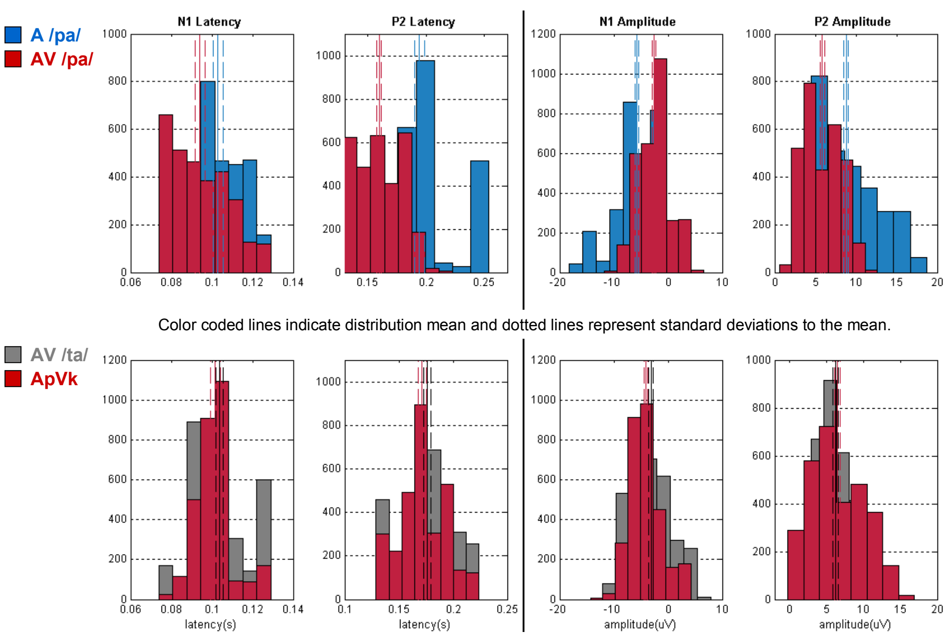


Pseudo-random design – Auditory Evoked Potentials (n=10)

CPz

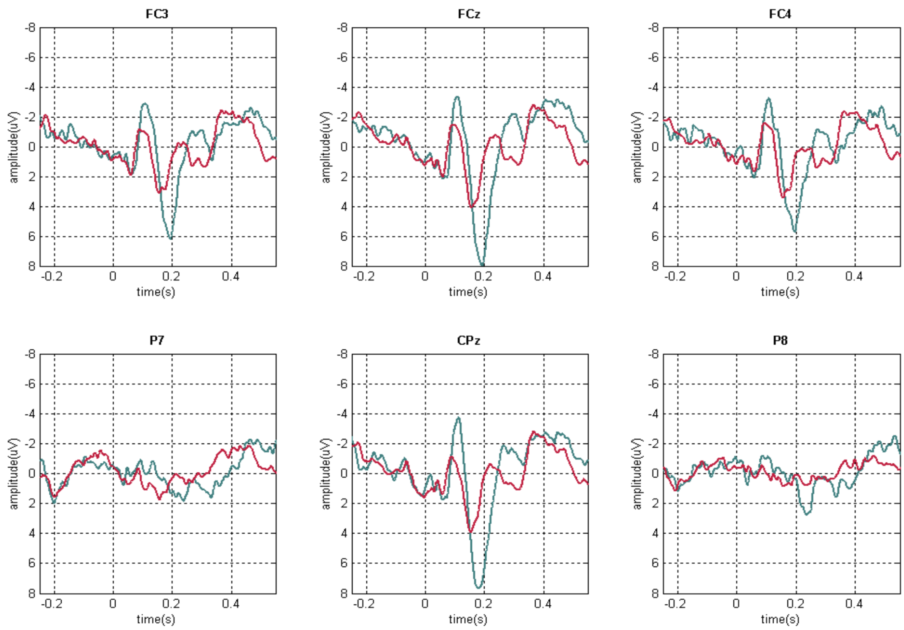


Bootstraps N1P2 latency/amplitude distributions (n=10) CPz

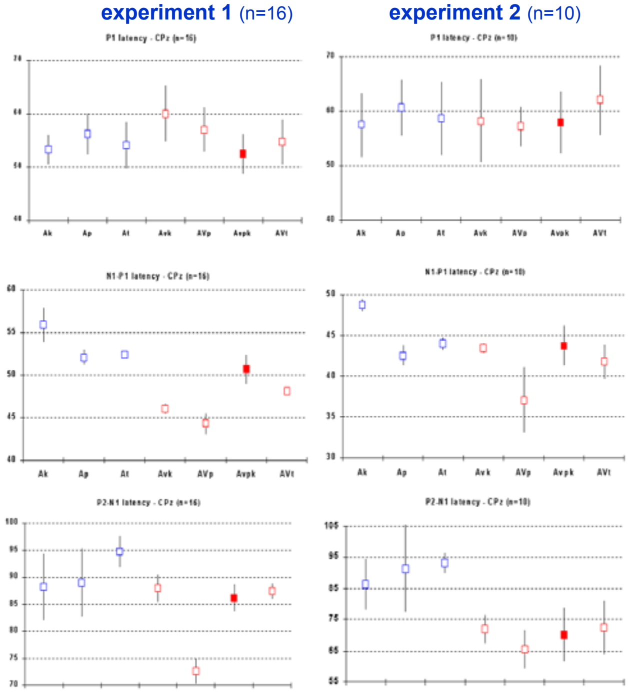


Linearity test and supra-additivity e.g. /pa/

A+V
AV



experiment 1 (n=16) / experiment 2 (n=10)



Relative Latency Facilitation CPz

No difference in P1 latency, great variability.

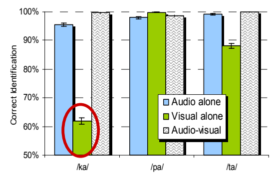
~10ms gained between P1 and N1 in AV condition

(1) except for real and illusory /ta/, and in (2) /ta/ and /ka/ shows the same latency difference suggesting that prior visual input does not allow clear disambiguation.

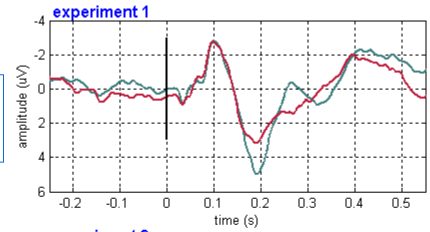
~20ms gained between N1 and P2 in AV condition

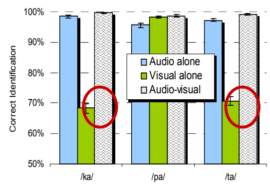
(1) Constant gain since P1-N1 and (2) clear differentiation during the N1-P2 transition

Real AV/ta/ Illusory /ta/ (audio /pa/ dubbed onto visual /ka/)

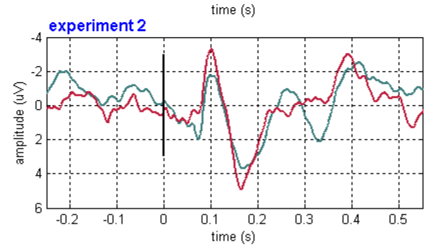


When visual /ta/ is non ambiguous, the illusory /ta/ does not differ from a real /ta/ until to ~200ms.





When visual /ta/ is ambiguous, a significant amplitude difference is observed as early as P1 The trend reverses towards higher amplitude for illusory /ta/ (N1 and P2).



Early auditory processes are modulated by the degree of visual ambiguity and suggest that,

(i) The integration process is initiated prior to or about 100ms post auditory onset,

(ii) Auditory feature extraction may be regulated by the informational content of visual inputs.

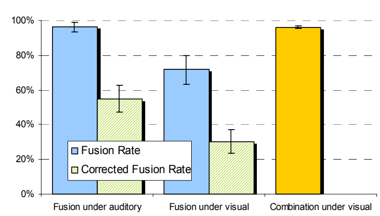
While perceptually equivalent to a real /ta/, the illusory /ta/ does not elicit a P2b. Later processing stages (such as lexical access) may be affected by the representativeness of the incongruent AV speech.

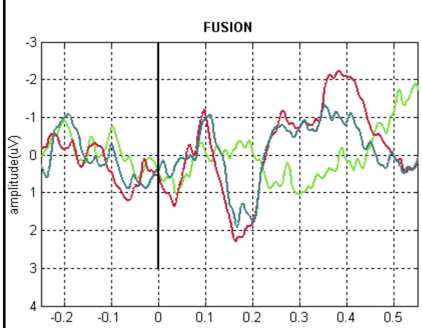
Visual saliency and Auditory-Visual integration

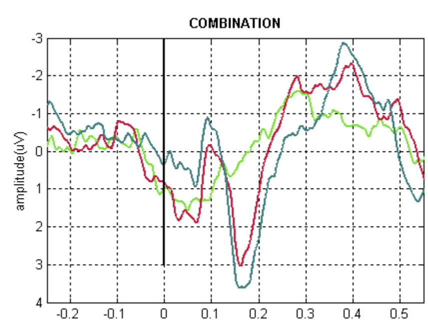
Visual identification of /pa/ in McGurk combination is not affected by incongruent auditory input.

In contrast, in McGurk fusion, illusory percepts remain high.

Recordings show a slow positivity starting ~ 100ms prior to auditory onset when visual inputs are non-ambiguous







ApVk visual attention
V /ka/
ApVk block design

AkVp visual attention
V /pa/
AV /pa/ block design

Summary of Results

- Experiment 1 showed an amplitude decrease of the N1/P2 auditory complex for all AV speech conditions. The amplitude reduction was accompanied by a latency shift of the P1/N1/P2 complex (~20ms) suggesting faster processing of bimodal speech as compared to audio alone condition.

  No significant supra-additive effect was found.

  The illusory McGurk percept /ta/ induced in the incongruent speech condition significantly differed from a congruent /ta/ in the 250-350ms range.

- Experiment 2 replicated the amplitude reduction of experiment 1. However, as visual /ka/ and /ta/ were more easily confused than in experiment 1, the latency shift for bimodal /ka/ and /ta/ was observed later on and shows a robust 20ms shift during the N1/P2 transition.

- Experiment 3 tested the effect of attending the visual modality in two conditions of incongruent speech. The amplitude reduction of the auditory N1/P2 complex was larger than in experiments 1 and 2, suggesting that attentional effects can not entirely account for results observed in previous experiments.

  The robust /pa/ percept in the combination condition was accompanied by a slow positive deflection ~100ms prior to auditory onset. This deflection was absent in the fusion condition and point out to the importance of visual saliency in the two AV stimuli tested.

Conclusions

In bimodal speech, visual kinematics usually precede auditory onset. This natural chronology of events hypothetically enables earlier processing of visual inputs, which may in turn affect the attentional resources allocated to processing auditory inputs as observed in the amplitude reduction.

However, the robust latency shift of the N1/P2 complex suggests that visual information instead facilitates the processing of auditory speech information. This latency facilitation effect indicates that visual kinematics allow participants to assess a degree of expectancy of auditory inputs. Furthermore, the saliency of visual kinematics tends to correlate with the modulatory effect of the auditory evoked-related potentials: the clearer the visual input, the more robust the auditory latency facilitation (e.g. /pa/).

AV speech integration is here shown to occur as early as N1 in agreement with early AV speech models of integration. It is however apparent that later processing stages (~250-350ms) are affected by the representativeness of the bimodal stimulus as shown in incongruent speech.

Taken together these results suggest that,

(i) auditory and visual speech information interact early on in the speech pathway,

(ii) visual saliency drives the strength of AV electrophysiological facilitation, and

(iii) the type of AV interaction (fusion vs. combination) in incongruent speech is contingent upon the degree of visual ambiguity.

Acknowledgments

- Supported by Grant NIH DC 05660 to DP
- Special thanks to Dr. Jonathan Z. Simon for his helpful discussions on signal analysis methods.

References

[1] Stein BE, Meredith AM (1993) The merging of the senses. Cambridge, MA, MIT Press.

[2] Grant KW, Greenberg S (2001) Speech intelligibility derived from asynchronous processing of auditory-visual information. Proceedings of the Workshop on Audio-Visual speech processing (AVSP).

[3] Giard MH, and Peronnet F (1999) Auditory -visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study. Journal of Cognitive Neuroscience 11:5, pp.473-490.

[4a] Calvert GA, Brammer MJ, Bullmore ET, Campbell R, Iversen SD, Anthony SD (1999) Response amplification in sensory-specific cortices during cross-modal binding. Neuroreport, 10: 2619-2623.

[4b] Calvert GA, Bullmore ET, Brammer MJ (1997) Activation of auditory cortex during silent lipreading Science, 276: 593-596.

[5] McGurk H, McDonald J (1976) Hearing lips and seeing voices. Nature 264: 746-747.