

# Temporal factors in the electrophysiological markers of audiovisual speech integration

*Michael Pilling, Sharon Thomas*

MRC Institute of Hearing Research, Science Road, University Park,  
Nottingham, NG7 2RB. United Kingdom

[m.pilling@ihr.mrc.ac.uk](mailto:m.pilling@ihr.mrc.ac.uk), [s.thomas@ihr.mrc.ac.uk](mailto:s.thomas@ihr.mrc.ac.uk)

## Abstract

Recent research had shown that concurrent visual speech modulates the cortical event-related potential N1/P2 to auditory speech. Audiovisually presented speech results in an N1-P2 that is reduced in peak amplitude and with shorter peak latencies than unimodal auditory speech [11]. This effect on the N1/P2 is consistent with a model in which visual speech integrates with auditory speech at an early processing stage in the auditory cortex by suppressing auditory cortical activity. We examined the effects of audiovisual temporal synchrony in producing modulations in the N1/P2. With the visual stream presented in synchrony with the auditory stream our results replicated the basic findings of reduced peak amplitudes in the N1/P2 compared to a unimodal auditory condition. With the visual stream temporally mismatched with the auditory stream (so that the auditory speech signal was presented 200 ms before its recorded position) the recorded N1/P2 was similar to unimodal auditory speech. The results are discussed in terms of Wassenhove's 'analysis-by-synthesis model' of audiovisual integration.

## 1. Introduction

### 1.1. Background

Auditory speech perception is generally found to be enhanced by the presence of visual speech information. This is particularly true when the auditory signal is impoverished [1-3], though advantages are found even under quiet listening conditions [4]. One of the most powerful demonstrations of the effect of visual speech on auditory speech processing is the McGurk illusion [5]. When incongruous visual speech is presented with the auditory speech signal, for example a visual articulation of /ga/ with an auditory /ba/, observers report hearing /da/, a cross-modal fusion of the two signals.

The McGurk illusion suggests that the visible aspects of speech are in some way integrated with the auditory signal. Models disagree in the locus of this integration process. Some view integration as a relatively late process, occurring after the auditory modality has categorized phonemic features of the speech signal e.g. [6]; others propose that visual speech integrates with auditory speech early on before categorization has occurred e.g. [7]. Some behavioral evidence is consistent with the early integration model. Visual speech can influence the reported perception of sub-phonetic features [8] and seems to occur before attentional selection [9]. However, behavioral experiments tap all stages of information processing and therefore only provide indirect evidence of the locus of interaction effects. The event related potential (ERP)

technique is, by contrast, well suited to examining the time-course of audiovisual interactions because of its impressive temporal resolution. The studies conducted to-date have shown evidence consistent with the early integration model. It has been shown that audiovisual (AV) speech presentation results in a decrease in the amplitude of the N1/P2 auditory brain response [10-12]. The N1/P2 is thought to reflect the operation of early cortical mechanisms processing the initial physical attributes of an auditory stimulus [13, 14]. The decrease in amplitudes does not result from a simple linear summation of N1-P2 activity with visually generated responses: activity in the AV condition was also reduced compared with the aggregate unimodal responses. The amplitude decrease in the N1/P2 response seems to be a reliable electrophysiological marker of the early integration of visual with auditory speech information. Klucharev et al. [13], however, found amplitude reduction in the N1/P2 even with perceptibly mismatched AV speech. Here audiovisual pairings were of vowels that were congruent (e.g. visual /i/ with auditory /i/), or incongruent (e.g. visual /o/ with auditory /i/) with one another. Incongruent vowels did not produce a McGurk-type integrated percept: visual information was perceived as being in clear conflict with the auditory speech signal. However, compared to the unimodal condition suppressed amplitudes in the N1/P2 were found equally for congruent and incongruent AV pairings, meaning that the effect can occur without audiovisual information producing an integrated percept. This, at least, raises the possibility that the effect may be related to presentation of information in two modalities but not necessarily to integration of AV information. For instance, fixation of a visual stimulus suppresses activity in the auditory brain [15, 16]. Thus, the requirement to observe the screen in audiovisual conditions itself may suppress auditory responses. Furthermore, even unimodal visual speech activates primary auditory cortex regions e.g. [17]. It is possible that such activations alone lead to decreases in auditory responses.

One way to assess the importance of integration mechanisms, compared to the other putative cross-modal processes mentioned above is to assess the importance of temporal synchrony between the modalities. Behavioral evidence shows that synchrony between the modalities is important in AV integration. Small asynchronies of a few tens of milliseconds have little effect, but larger asynchronies greater than 200ms, particularly when the auditory stream leads the visual, severely reduce visual influences on auditory speech perception: McGurk-type influences are diminished [18] as are the audiovisual benefits of speech presented in noise [19].

The brain itself shows differing responses to synchronized and desynchronized AV speech. Macaluso et al. [20], looking at changes in brain haemodynamic responses to AV speech found that synchronous AV speech produces greater activation in the superior temporal cortex than asynchronous AV speech. This finding suggests that the effect on the N1/P2 should also be sensitive to this manipulation: numerous methods have identified the superior temporal region as a multisensory centre [21, 22] and the region is also considered a source of the N1/P2 [14]. Furthermore, Besle et al. [11] and Wassenhove et al. [12] both propose this region as a likely candidate for the AV interactions underlying reductions in N1/P2 amplitude.

## 1.2. Aims

If AV integration processes are responsible for suppression of the N1/P2 then the effect will be dependant on synchrony between the modalities. Specifically, AV speech perceived as being asynchronous should not diminish the auditory response. If, however, the effect is unrelated to integration mechanisms then it should be independent of any synchrony/asynchrony between the modalities. If the effect is only found for synchronous AV speech then it strengthens its claim as a genuine neural correlate of AV speech integration. Experiment 1 recorded ERP responses to speech in unimodal auditory (AO), unimodal visual (VO) and synchronous AV conditions (AV) in a phonetic oddball detection task. N1/P2 amplitudes were found to be lower with AV speech compared to unimodal auditory speech. Experiment 2 recorded ERP responses to speech in synchronous AV, asynchronous AV and AO conditions. It was found that N1/P2 amplitudes were significantly lower with synchronous compared to asynchronous AV speech. Indeed, N1/P2 amplitudes were larger for asynchronous AV speech than for AO speech.

## 2. Experiment 1

### 2.1. Method

#### 2.1.1. Stimuli

Stimuli consisted of high quality audiovisual recordings of a single male talker articulating the speech tokens /pa/ and /ta/ from an initial neutral visual expression. Four good exemplars of each articulated token were edited into video-clips. Each clip starting from the point just before the initial articulation to just after the end of the articulation. A still-frame lasting 1000 ms was presented before the onset of the moving face for each clip. Clips were all 2000 ms in total. Video clips were processed using Discreet Cleaner® (v 1.0) into a sequence of high quality still bitmap images (160 × 210 mm) viewed at an approximate distance of 800 mm. Bitmaps were presented on a 15" LCD monitor at a rate of 25 fps with the corresponding digital audio files presented via Sennhauser headphones. Stimulus presentation was controlled by purpose-written software routines ensuring accurate timing, running on an IBM PC-compatible computer.

#### 2.1.2. Participants

Twelve English speakers age 18 to 30 were used (10 female; 2 male). All had normal hearing and vision.

#### 2.1.3. Procedure

Participants were seated at an approximate distance of about 400 mm from the LCD wearing an EEG cap and headphones. An oddball paradigm was used. In each block participants were presented with 180 standards (/ta/) and 40 deviants (/pa/) and instructed to listen to the spoken sounds while looking at the screen and to press a key each time a deviant was presented. Deviants occurred randomly within the sequence of standards, with the constraint that the first stimulus was never a deviant stimulus and that two deviant stimuli were never presented successively. Stimuli were presented sequentially with a 3500 ms inter-trial interval (+/- 500 ms random jitter). Blocks took approximately 11-12 minutes each. The task was performed in three conditions: AO, VO, AV. In the AV condition the duration between the first moving frame and the recorded speech sound onset was different for each of the four exemplars of /ta/ (values were 210, 222, 260 and 296 ms). Figure 1 shows a schematic diagram of a single AV trial. Two blocks of each condition were given in a random order, with the constraint that no condition was repeated until one block of each condition was already given.

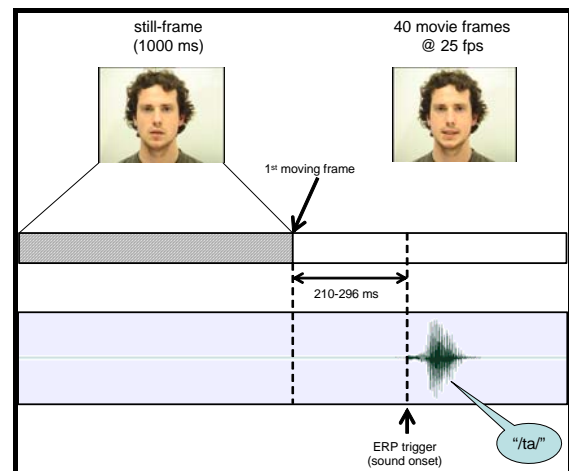


Figure.1: Schematic diagram of a single standard AV trial. A 1000 ms still frame of the face is presented followed by an articulating face. On standard trials (/ta/) the sound-onset occurred between 210-296 ms after presentation of the first moving frame. On VO trials the video stimulus was presented silently. On AO trials only the audio stimulus was presented with the face replaced by a fixation cross. A black screen was presented during the inter-trial interval.

#### 2.1.4. EEG Recoding

Recordings were made at 1000 Hz using a BrainAmp MRplus (BrainProducts GmbH) EEG recording system onto a laptop PC installed with Brain Vision Recorder (V. 1.03). A 34 electrode elastic cap (Falk Minow Services, GmbH) was fitted to the head of the participant. High chloride abrasive paste (Abralyt 20000 HCl, FMS GmbH) was applied between the electrodes and scalp as an electroconductive agent. The ground was positioned at AFz and online referencing was done using Cz. A marker signal designating the auditory speech onset and was recorded with the EEG signal.

## 2.2. Analysis and Results

Recordings were downsampled offline using Brain Vision Analyzer (V 1.05), screened for artifacts, then bandpass filtered using a Butterworth filter (1 - 30 Hz). Data was re-referenced and channel Cz reconstructed using all 34 channels as a new reference. An ocular correction algorithm was applied using FP1 as the EOG channel. Only ERPs to the standards were used in the analysis to avoid contamination from oddball and response related processes. The auditory stimulus onset marked the start of each epoch for averaging purposes. The AO and AV conditions produced a typical N1/P2 wave greatest at central electrode sites in response to the sound onset Grand averaged waveforms for the AV, AO and VO conditions are shown in Figure 2 for three electrodes FC1, FC2 and Cz.

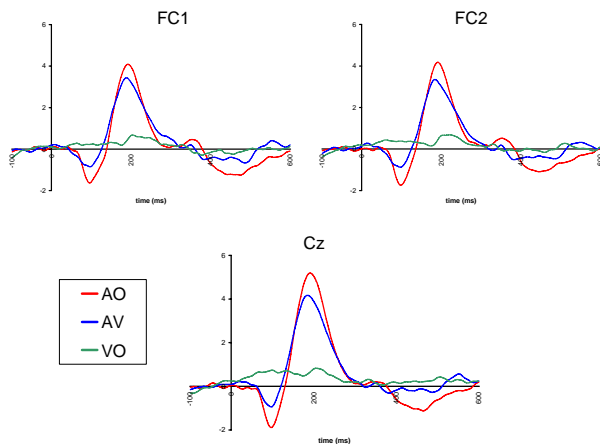


Figure 2: Grand average ERPs as recorded at three sites (FC1, FC2, and Cz) for the AO, AV, and VO conditions. Waveforms have been baseline corrected (-100 ms pre-stimulus baseline) Time (in ms) is given on the abscissa (the ordinate indicates the auditory stimulus onset), and amplitude (in  $\mu V$ ) on the ordinate.

Peak-to-peak measures of the N1/P2 responses were computed for the central electrodes (FC1, FC2, Cz, C3, C4, CP1, CP2). Peaks were detected using a local maxima method. The N1 peak was defined as the largest negative local maxima between 60-140 ms after stimulus onset, and the P2 peak as the largest positive local maxima between 130-300 ms after stimulus onset. The peak-to-peak measure is a quantification of N1/P2 amplitude that is independent of any particular choice of baseline. This measure is displayed in Figure 3 for the AO, AV condition and for the responses in the AO condition summed with responses from the VO condition (AO+VO). The VO condition produced no N1/P2. These measures were subjected to a two-way ANOVA with Condition (AO vs. AV) and electrode (seven levels) as repeated factors. Analysis showed that peak-to-peak amplitude was significantly lower in the AO condition than the AV ( $F=49.49$ ,  $p < .0001$ ). Further analysis demonstrated that auditory response suppression was not a linear effect of summated auditory and visual activity. Comparison between peak-to-peak measures in the AO condition and the AO+VO was non-significant ( $p > .05$ ); the AV condition also showed significantly lower N1/P2 amplitudes compared against AO+VO,  $F=.80.84$ ,  $p < .0001$ .

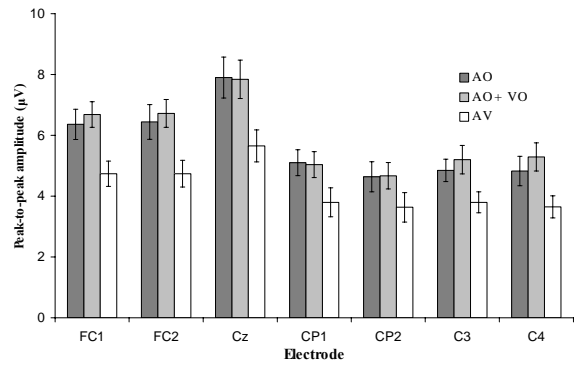


Figure 3: Mean peak-to-peak amplitudes of N1/P2 in the AO, AO+VO and AV conditions in Experiment 1. Error bars show  $\pm 1$  standard error.

## 2.3. Discussion

N1/P2 peak amplitudes were lower overall in the AV condition compared to the unimodal AO condition. This was the case even when the unimodal visual (VO) responses were taken into account. As noted in the introduction, one interpretation of this effect on auditory ERPs is that it reflects the operation of integration mechanisms in multisensory areas, such as the superior temporal cortex, inhibiting N1/P2 generation. An alternative interpretation is that the effect is related to inhibitory effects associated with processing the visual speech information presented on screen. Because AV integration should be dependant on synchrony [18-20], the N1/P2 responses to synchronous and asynchronous AV speech were compared.

## 3. Experiment 2

### 3.1. Method

Twelve subjects aged 18-30, all with normal vision and hearing were used. Experiment 2 had three conditions: AO (the same as the AO condition in experiment 1) and AVsynch (the same as the AV condition in Experiment 1), and AVasynch. In this last condition the auditory signal was moved +200 ms out of phase with the video stream from its recorded position in the AVsynch condition. Participants confirmed that the asynchrony was clearly perceptible in the AVasynch condition.

### 3.2. Results

ERPs were processed and averaged in the same manner described in section 2.2. Waveforms for the AO, AVsynch, and AVasynch conditions are shown in Figure 4 for three electrodes FC1, FC2 and Cz, corrected in the figure by a -100 ms pre-stimulus baseline.

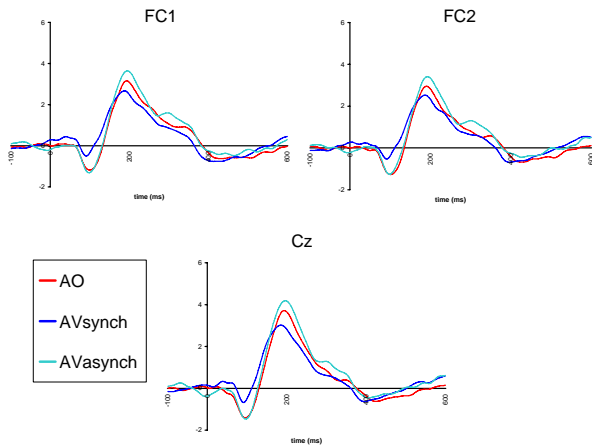


Figure 4: Grand average ERPs on FC1, FC2, and Cz for AO, AVsynch, and AVasynch conditions in Experiment 2. Waveforms are baseline corrected to (-100 ms pre-stimulus baseline Time (in ms) is on the abscissa and amplitude (in  $\mu V$ ) on the ordinate.

Peak-to-peak measures of the N1/P2 in the three conditions were computed as in section 2.2. are given in Figure 5. As in Experiment 1, peak-to-peak suppression was found between the AO and AVsynch conditions ( $F=74.34$ ,  $p < .0001$ ), demonstrating the reliability of the amplitude suppression effect with synchronous audiovisual speech. In contrast, suppression of the auditory response was absent in the AVasynch condition; indeed AVasynch peak-to-peak amplitudes were slightly higher than in the AO condition, though the effect was only marginally significant ( $F=5.41$ ,  $p < .05$ ).

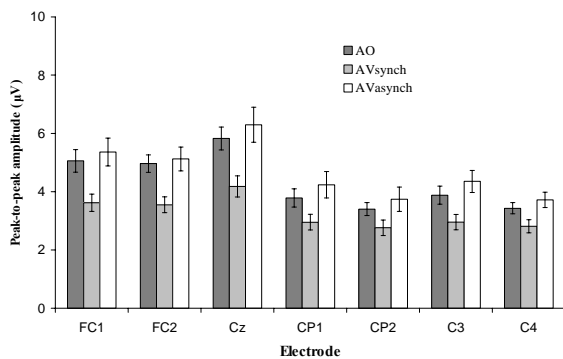


Figure 5: Mean peak-to-peak amplitudes of N1/P2 in AO, AVsynch and AVasynch conditions in Experiment 2. Error bars show  $\pm 1$  standard error.

### 3.3. Discussion

The two audiovisual conditions resulted in significantly different peak amplitudes of the N1/P2 wave. Compared against unimodal auditory speech, responses were suppressed with synchronous audiovisual speech in both Experiment 1 and 2, though no suppression was found with asynchronous AV speech. Thus N1/P2 suppression was contingent on the temporal synchrony between the two modalities. This is evidence against the effect resulting simply from the requirement to observe visual speech information on the screen. Instead results suggest that the effect reflects a genuine process of AV integration in which temporally

coincident (synchronous) visual speech is bound with auditory speech information.

## 4. General Discussion

The results of our asynchronous manipulation can be considered consistent with Wassenhove's 'analysis by synthesis' model of AV speech integration [11]. The model argues that the N1/P2 amplitude decrease reflects a process in which predictive information in the visual signal suppresses redundant auditory information. The visual signal in the AVasynch condition might be diminished in its effects because it occurs temporally closer to the auditory signal. This may mean that the auditory signal is processed before any evaluation process is fully completed. Equally, the evaluation process itself might be sensitive to the cross-modal temporal correspondence of the two signals.

It may be that the temporal association between dynamic changes in the auditory and visual speech signals [23, 24] are critical for this integration process to occur. Cells in the superior temporal cortex might be activated when changes in the two modalities are associated, eliciting auditory cortical response suppression. In temporally displacing the auditory signal (as in the AVasynch condition of Experiment 2), it would decouple the association between the auditory and visual signal diminishing superior temporal activity [20], eliciting no auditory cortical suppression. Perceptibly incongruous, but temporally contiguous AV speech e.g. [12] may still maintain overall associations between the dynamic changes in the two modalities thus leading to suppression in the auditory cortical response.

Unfortunately, ERP signals alone are inadequately specified to provide meaningful information on the relationship between activity in different brain areas. By combining information obtained from ERP and more spatially sensitive imaging techniques (such as fMRI), it might be possible to determine if AV synchrony-specific activity in the superior temporal cortex [20] is associated with decreases in the N1/P2.

This research has identified synchrony of the visual speech signal with the auditory signal as one important factor in producing N1/P2 suppression as an electrophysiological marker of integration. Further work is needed to determine what other properties of the visual speech signal are significant. If, for instance, the dynamic characteristics of visual speech are important, changes in the N1/P2 should be more affected by the video frame rate than the amount of pictorial information in the visual stimulus (as in point light visual speech displays [23, 24]). Furthermore, it is still unknown whether the audiovisual stimuli need to be actively processed for this marker of integration to occur. In the current and previous [10-12] studies the task always emphasized active processing of the stimuli. Besle et al [25] recently showed that visual speech could elicit a MMN response while participants were engaged in performing an unrelated task of detecting changes in a fixation cross stimulus. If modulation of the N1/P2 could be elicited under similar passive conditions it would suggest that early AV integration processes were attention independent.

## 5. Conclusions

Audiovisual presentation of speech leads to suppression of N1/P2 amplitudes. The effect cannot be accounted for by the task requiring the participant to observe the screen. Instead it seems to be a genuine electrophysiological marker of integration. This is strong evidence supporting early integration models of audiovisual speech perception e.g. [7]. Further work is needed to establish what aspects of the visual stimulus and the task conditions are necessary for the effect to occur.

## 6. Acknowledgements

The assistance of Tim Folkard in writing the software, Marco Calabresi in use of the EEG system, and Katrin Krumbholz for help with the analysis is gratefully acknowledged.

## 7. References

- [1] Sumbly, W.H., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* 26, 212–15.
- [2] Schwartz, J.-L., Berthommier, F., Savariaux, C., 2004. Seeing to hear better: evidence for early audio–visual interactions in speech identification. *Cognit.* 93, B69–B78.
- [3] Rosen, S., Faulkner, A., & Wilkinson, L. 1999. “Adaptation by normal listeners to upward spectral shifts of speech: Implications for cochlear implants.” *J. Acoust. Soc. Am.*, 106 (6), 3629–36.
- [4] Reisberg, D., McLean, J., & Goldfield, A. 1987. Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 97–113). Lon.: Erlbaum.
- [5] McGurk, H., & MacDonald, J. 1976. Hearing lips and seeing voices. *Nat.*, 264, 746–48.
- [6] Massaro, D. W., *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*, Cam., MA: Mm Press, 1998.
- [7] Braid, L.D. 1991. Crossmodal integration in the identification of consonant segments. *Q. J. Exp. Psy.*, 43, 647–77
- [8] Green, K.P. 1997. The use of auditory and visual information during phonetic processing: implications for theories of speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 97–113). Lon.: Erlbaum.
- [9] Soto-Faraco, S., Navarra, J., & Alsius, A. 2004. Assessing the automaticity of audiovisual speech integration: Evidence from the Speeded Classification Task. *Cogn.*, 92, B13–B23.
- [10] Besle J., Fort A, Delpuech C, Giard M.H. 2004. Bimodal speech: early suppressive visual effects in the human auditory cortex. *Eur. J. Neurosci* 20:2225–34
- [11] van Wassenhove, Grant, K.W., & Poeppel, D. 2005. Visual speech speeds up the neural processing of auditory speech. *PNAS*, 102, 1181–6
- [12] Klucharev, K., Möttönen, R. & Sams, M. (2003). Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. *Cog. Brain Res.* 18:65–75.
- [13] Hyde M. 1997. The N1 response and its applications. *Audiol Neurootol.* 2(5): 281–307
- [14] Näätänen R, Winkler I. (1999) The concept of auditory stimulus representation in cognitive neuroscience. *Psychol Bull* 6:826–59
- [15] Populin, L.C. & Yin, T.C. 2002. Bimodal Interactions in the superior colliculus of the behaving cat. *J. Neurosci.* 22:2826–34.
- [16] Laurienti P.J, Burdette J.H., Wallace M.T., Yen Y.F., Field A.S., Stein B.E. 2002. Deactivation of sensory-specific cortex by cross-modal stimuli. *J. Cognit. Neurosci.* 14:420–9
- [17] MacSweeney, M., Amaro, E., Calvert, G., Campbell, R., David, A.S., McGuire, P., Williams, S., Woll, B. & Brammer, M. J. Activation of auditory cortex by silent speechreading in the absence of scanner noise: An event-related fMRI study. (2000). *NeuroRep.*, 11 (8), 1729–1733.
- [18] Munhall, K.G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk Effect. *Percept. Psych.*, 58, 351–62.
- [19] McGrath, M., and Summerfield, Q. Intermodal timing relations and audio-visual speech recognition by normal hearing adults, *J. Ac. Soc. Am.* 77: 678–85.
- [20] Macaluso, E., Dolan, R.C., Spence, D. & Driver J. (2004). Spatial and temporal factors during processing of audiovisual speech: a PET study. *Neuroimage.* 725–32.
- [21] Calvert. G. & Campbell. M. 2000. Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Curr. Biol.* 10:649–57
- [22] Callan, D.E. Jones, J.A., Munhall, K. Kroos, C., Callan, A.M., Vatikiotis-Bateson, E. 2004. Multisensory Integration Sites Identified by Perception of Spatial Wavelet Filtered Visual Speech Gesture Information. *J. Cogni. Neurosci.* 16:805–816
- [23] Rosenblum, L.D. & Saldaña, H.M. 1996. An audiovisual test of kinematic primitives for visual speech information. *J Exp. Psy.:HPP*:22:318–31.
- [24] Lachs, L. & Pisoni, D. 2004. Specification of cross-modal source information in isolated kinematic displays of speech. *J. Acoust. Soc. Am.*, 116, 507–18.
- [25] Besle, J., Fort, A., Giard, M.H. 2005. Is the auditory sensory memory sensitive to visual information? *Exp Brain Res.* 166:337–44.