Electrophysiology of Auditory-Visual Speech Integration

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

[1]Neuroscience and Cognitive Science Program, Cognitive Neuroscience of Language Laboratory, University of Maryland, College Park
[2]Auditory-Visual Speech Laboratory, Walter Reed Army Medical Center, Washington D.C.
vvw@glue.umd.edu

Abstract

Twenty-six native English Speakers identified auditory (A), visual (V), and congruent and incongruent auditory-visual (AV) syllables while undergoing electroencephalography (EEG) in three experiments. In Experiment 1, unimodal (A, V) and bimodal (AV) stimuli were presented in separate blocks. In Experiment 2, the same stimuli were pseudo-randomized in the same blocks, providing a replication of Experiment 1 while testing the effect of participants' expectancy on the AV condition. In Experiment 3, McGurk fusion (audio /pa/ dubbed onto visual /ka/, eliciting the percept /ta/) and combination (audio /ka/ dubbed onto visual /pa/) stimuli were tested under visual attention [1].

EEG recordings show early effects of visual influence on auditory evoked-related potentials (P1/N1/P2 complex). Specifically, a robust amplitude reduction of the N1/P2 complex was observed (Experiments 1 and 2) that could not be solely accounted for by attentional effects (Experiment 3). The N1/P2 reduction was accompanied by a temporal facilitation (approximating ~20ms) of the P1/N1and N1/P2 transitions in AV conditions. Additionally, incongruent syllables showed a different profile from congruent AV /ta/ over a large latency range (~50 to 350ms post-auditory onset), which was influenced by the accuracy of identification of the visual stimuli presented unimodally.

Our results suggest that (i) auditory processing is modulated early on by visual speech inputs, in agreement with an early locus of AV speech interaction, (ii) natural precedence of visual kinematics facilitates auditory speech processing in the time domain, and (iii) the degree of temporal gain is a function of the saliency of visual speech inputs.

1. Introduction

Numerous studies have shown that facial kinematics disambiguate auditory speech in noise [2][3], improves overall intelligibility in normal hearing [4] and hearing-impaired adults [5], and can also alter the auditory percept [1]. A fundamental issue of AV speech pertains to the locus of auditory-visual (AV) speech integration, which has remained a challenge for speech theories. Two major theoretical views have emerged. The *early integration* view postulates a pre-phonetic stage of AV integration, which implies that auditory and visual speech metrics must be compatible pre-phonetically to insure dependency of input channels. As an alternative, the *late integration* (post-phonetic) view proposes an intermediary modality-specific representation [6] where an amodal integrative stage or a metric transformation of visual inputs must be assumed post-phonetically. In both theories, the timing of AV speech integration remains unclear and the neural correlates speculative.

Because the main theoretical issue in AV speech resides in the timing of the integration stage, brain imaging techniques with a good temporal resolution (~1ms), such as electroencephalography (EEG) and magnetoencephalography (MEG), are needed if one aims to link theoretical implications with neurophysiological evidence. A comparison between neural processing and theoretical stance shows that while neural pathways (within and across modalities) act in parallel (and with various time constants), the theoretical approach essentially adopts a serial processing view – but a linearly staged processing model of AV speech underestimates the computational capabilities of the nervous system.

In particular, AV speech processing is not limited to perceptual speech categorization but rather involves various levels of processing with possible interactions at different system levels also specific to multisensory perception (e.g. subcortical and cortical multisensory neural populations). Hence AV speech is a special case of multisensory integration and, as such, should share its principled mechanisms. AV speech is furthermore embedded into a well-studied perceptual framework, the speech system. Consequently, two major types of specific constraints may apply to AV speech, namely physical (inputs-driven and neural architecture) constraints and perceptual (speech) constraints.

A major feature of AV speech resides in the natural timing of events: visual speech often occurs *prior to* the auditory onset. Careful alignment of AV speech inputs is unnecessary to insure AV speech integration, which tolerates as much as ~200ms asynchrony while being elicited more robustly when visual inputs lead the audio [7][8][9]. If timing differences of modality-specific neural pathways can partially account for small lags (auditory information reaches primary auditory cortices in 12-16ms [10][11] and visual inputs the visual cortices in ~50ms [12]), 200ms nevertheless approximates the critical syllabic length common across all languages [13]. In AV speech, recent investigations have suggested a high degree of correlation between lip area and acoustic amplitude envelope [14][15], which corresponds to a low periodicity close to ~4Hz (250ms) [15][16]. Taken together these data support the existence of a ~200ms time constant for syllabicity, which may emerge through specific temporal integrative properties of the cortex [17].

Note that an interesting analogy can be drawn between the 'spatio-temporal coincidence principle' of multisensory integration [18] and AV speech. The spatio-temporal redundancy of AV speech inputs could also provide a first level of perceptual binding independently of the speech nature of the stimuli.

Furthermore, perceptual constraints of speech representation most likely influence AV speech integration for two critical reasons. First, the quality and the quantity of informational content in auditory and visual domains differ greatly. It is now well known that whereas auditory inputs provide sufficient information for full phonetic categorization, visual speech is limited to visemic representation (essentially based on place-of-articulation). Second, the intrinsically different neural coding schemes in the auditory and visual modalities must converge to constitute a unified speech representation at some level. However, independent of the nature of the speech metric, any modality triggering the speech system can likely feed back onto the processing of subsequent speech inputs. Thus, if such perceptual constraints intervene in AV speech integration, the AV profile of modality-specific evoked potentials should differ from that obtained when the same stimuli are presented unimodally. In particular, because visual speech motion precedes auditory speech onsets, effects should be observed in the auditory-evoked potentials.

The goal of our study was to characterize AV speech electrophysiologically. The hypothesis was based on EEG of non-speech AV stimuli [19][20][21] and fMRI of AV speech [22][23], which showed enhanced auditory cortical signals in the presence of visual inputs. An enhancement of the classic auditory evoked potential P1/N1/P2 was thus predicted in AV presentation as compared to A alone. Following the assumption that spatio-temporal coincidence in incongruent speech is likely to be reduced, we predicted a lesser enhancement for McGurk fusion and a different profile from its congruent perceptual counterpart /ta/.

2. Materials and Methods

Twenty-nine normal-hearing native English speakers (13 females, 21.5 years of age) participated in three experiments (four individuals participated in Experiment 1 only, twelve participated in Experiment 1 and Experiment 3, and the remaining ten took part in Experiment 2).

Digital videos of a female talker provided natural congruent AV speech syllables /ka/, /pa/ and /ta/. McGurk tokens were made by dubbing audio /pa/ and /ka/ onto video /ka/ (fusion case, A_pV_k) and /pa/ (combination case, A_kV_p), respectively. Each AV syllable started with a fade-in neutral face. Preparatory facial movements (e.g. aspiration) naturally occurred 350 to 400ms prior to the audio onset. Unimodal conditions consisted of the same stimuli presented either auditorily (A, no video) or visually (V, no audio).

In Experiment 1, A, V and AV /ka/, /pa/, /ta/ and A_pV_k were presented 100 times. Unimodal stimuli (A,V) were presented in the same blocks and bimodal stimuli (AV) in different blocks. Participants reported what they perceived by pressing one of three buttons (3-AFC) according to what they heard (A), saw (V) or heard while watching the video (AV). In all three experiments, choices were /ka/, /pa/ and /ta/.

In experiment 2, the same stimuli (A, V, AV) were pseudo-randomized and presented in the same blocks. The task was identical to Experiment 1 and each stimulus was presented 100 times. Note, however, that in Experiment 2 participants did not know whether audio would be provided when a visual trial was started – as opposed to Experiment 1.

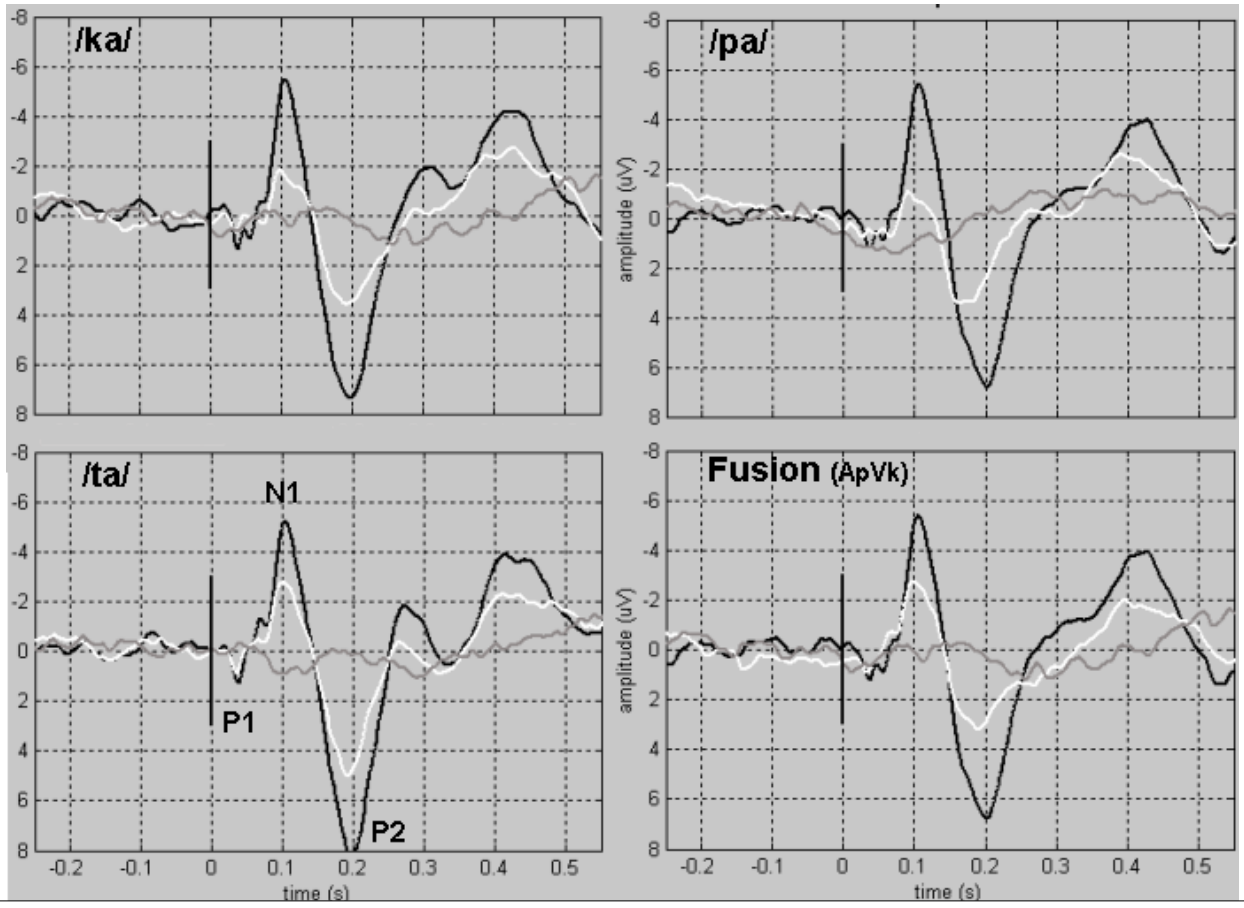

Figure 1: Auditory event-related potentials for unimodal and bimodal syllables (Experiment 1, N=16, BPF: 1-55Hz)

Black traces are audio alone, gray traces are visual alone, and white traces are audio-visual presentations (all recorded from CPz). Classic auditory event-related potentials are P1, N1 and P2 (positivity occurring ~50ms post-auditory onset, negativity at ~ 100ms and second positive deflection at ~200ms, respectively). Note the overall decrease and the latency shift of the auditory complex in AV conditions (latency is most prominent for /pa/).

In Experiment 3, 100 presentations of McGurk fusion (A_pV_k) and combination (A_kV_p) were tested. Participants were asked to report what they saw while ignoring what they heard.

In all three experiments, inter-stimuli intervals were pseudo-randomly chosen among 4 values (750, 1000, 1250 and 1500ms) whether stimuli were A, V or AV. No training was provided prior to the experiments. All participants were right-handed, had no diagnosed hearing problems and had normal or corrected-to-normal vision. The study was carried out with the approval of the University of Maryland Institutional Review Board.

EEG recordings were performed using a Neuroscan system (Neuroscan, Inc.) and 32 Ag/AgCl electrodes mounted on an electro-cap (Electro-Cap International, Inc.). Electrodes were referenced to linked left and right mastoids and AFz served as the ground electrode. Four electrodes recorded ocular artifacts. A continuous AC recording was used (acquisition band-pass filter 1-100Hz) with a sampling rate of 1kHz and a gain setting of 1000. Individual raw data were epoched around the auditory onset for all stimuli, with a pre-audio onset of 2600ms and a post-audio onset of 1895 ms (4096 sample points). Epoched data were submitted to an automatic ocular artifact reduction. Epochs containing amplitudes higher than a 100μV threshold were rejected.

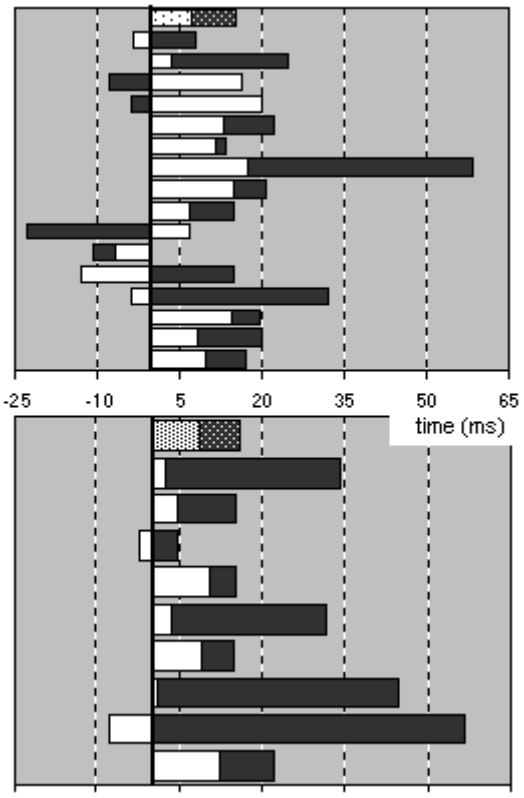

Figure 2: Temporal Facilitation in AV speech (Experiment 1, upper panel and Experiment 2, lower panel)

Transition times between N1/P1 (white) and P2 /N1 (black) were computed on an individual basis for A and AV separately. Values were then averaged across congruent stimuli to form a global comparison between A and AV (A-AV). Each bar corresponds to an individual's values. Dotted bars correspond to the grand average. Positive values reflect a faster transition in AV as compared to A.

Only correctly identified stimuli were considered (for incongruent speech, only /ta/ responses for the fusion case were considered for all experiments and only /pa/ responses for the combination case in Experiment 3). 400ms pre-audio onset for A conditions and 400ms pre-visual onsets for V and AV stimuli were used to baseline correct the entire epoch. Epoched data were then band-pass filtered from 1Hz to 55Hz using a zero-phase shift double-pass Butterworth filter with a 24dB cut-off. An in-house bootstrapping method [24] scripted in Matlab (Mathworks) was then used to resample 300 times individual data for 6 electrodes (CPZ, P7, P8, FCZ, FC3, FC4). Individual bootstrapped mean values (event-related potentials amplitude and latency) were imported in SPSS (11.0.1, SPSS Inc.) for analyses of variance.

All reported amplitude and latency effects were significant in repeated measurements (ANOVAs) performed on unprocessed and bootstrapped data. F and P values reported here are for bootstrapped values for all electrodes. In each analysis of variance, the following factors were used: 2 modalities (A and AV), parameters of 3 event-related potentials (amplitude, or latency of P1, N1, and P2), 6 or 7 stimuli (A and AV /ka/, /pa/ and /ta/, McGurk stimuli could not be included for modality effects) and 6 electrodes (CPz, P7, P8, FCz, FC3, FC4). Reported significance is based upon Greenhouse-Geisser corrections when sphericity could not be assumed and post-hoc Student t-tests for comparisons of means.

3. Results

In Experiments 1 and 2, a significant amplitude reduction of the N1/P2 complex was observed, accompanied by a temporal facilitation of the P1/N1/P2 complex approximating 20ms. Figure 1 shows the grand averaged traces obtained in Experiment 1 at a centro-parietal recording site (CPz) chosen to equally weight the contribution of auditory and visual cortices responses. The amplitude decrease was observed for all AV congruent and incongruent syllables (white traces) when compared to their respective audio alone conditions (black traces). A repeated measures analysis of variance (factors: modality [A, AV], amplitude [P1, N1, P2], electrodes [CPz, P7, P8, FCz, FC3, FC4] and stimuli [ka, pa, ta],) showed a significant interaction of modality with event-related potential (Experiment 1, $F = 53$, $p<0.0001$ and Experiment2, $F = 15.54$, $p<0.002$) These interactions suggest that the amplitude decrease varies with event-related potentials, i.e. with the each processing stage as early as ~50-100ms post auditory-onset.

Figure 2 illustrates the temporal gain between A alone and AV conditions (A minus AV) for N1 latency relative to P1 (white portion) and P2 latency relative to N1 (black portion). N1 and P2 peak latencies were normalized to P1 (N1 latency minus P1 latency) and N1 (P2 latency minus N1) peak latencies, respectively, because we observed inter-individual variability in latencies. The transition times (or relative latencies) thus obtained permit the comparison of the speed of processing time between A and AV conditions for each participant. If AV is faster than A, then the difference between the audio (N1-P1) and the audio-visual (N1-P1) should be positive (i.e. the transition from P1 to N1 in A is longer than in AV). We computed these transitions times for each stimulus and each participant at CPz.

In Experiment 1, (where auditory inputs were paired with visual inputs in the AV block only) effects could be seen as early as the P1/N1 transition (i.e. ~50 to 100ms post-auditory onset). In Experiment 2, (where pseudo-randomized

stimuli presentation lead to a 50% chance of auditory inputs expectation when visual stimuli started), the temporal gain occurred predominantly in the N1/P2 transition (i.e. ~100 to 200ms post-auditory onset). A four-way analysis of variance (factors: modality, latency, electrodes and stimuli) showed a main effect of modality (Experiment 1, $F = 21.69$, $p<0.0001$ and Experiment2, $F = 83.39$, $p<0.0001$). Interactions between modality and stimulus (Experiment 1, $F = 14.06$, $p<0.0001$ and Experiment2, $F = 13.44$, $p<0.001$), and modality and event-related potential (Experiment 1, $F = 14$, $p<0.0001$ and Experiment2, $F = 36.04$, $p<0.0001$) were significant. These interactions show that the temporal gain varies as a function of the auditory processing stage and are a function of the stimuli.

When identification rate was lowest for visual /ka/ (performance rate of ~62%, Experiment 1) a difference between congruent AV /ta/ and illusory /ta/ (audio /pa/ dubbed onto visual /ka/) significantly differed in amplitude at P2 ($p=0.003$) while in Experiment 2, where both /ka/ and /ta/ were confused (identification rate ~70%), differences between the real and the illusory /ta/ were not significant up to ~350ms. Figure 3 provides a comparison between McGurk /ta/ (black trace) and congruent AV /ta/ (white trace). These results suggest that the modulation of the early auditory components is a function of how much information has been extracted in the preceding visual input: the better V/ta/ is categorized, the earlier a difference is observed. Inversely, the more ambiguous visual tokens are, the less the difference in the early stages of auditory processing.

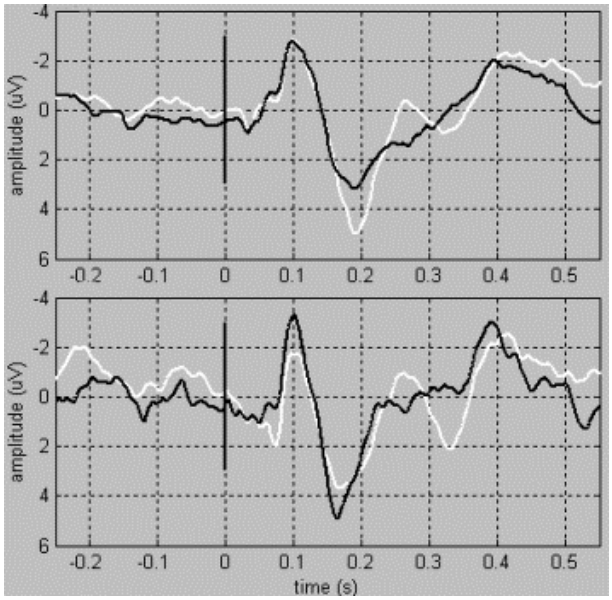

Figure 3: Illusory /ta/ vs. real /ta/ (Experiment 1, upper panel and Experiment 2, lower panel)

Black traces are McGurk /ta/ stimuli and white traces congruent AV /ta/. In Experiment 1, a marked difference between the two traces reaches significance at P2 while in Experiment 2 traces differentiate early on. Note the positivity occurring at ~350ms in congruent AV /ta/ missing in incongruent speech.

Figure 4 shows the EEG traces obtained in fusion (top) and combination (bottom) conditions tested in Experiment 3. While the robust amplitude reduction observed so far may originate from a divided attention effect, the results of Experiment 3 indicate that attentional effects cannot entirely account for such a decrease. While in the fusion case the N1/P2 complex did not significantly differ from that obtained in Experiment 1 (white and black traces respectively, $F=3.339$, $p <0.101$), for the combination case (where visual /pa/ was easily identified) the P1/N1/P2 amplitude is even more reduced than in the congruent AV /pa/ and follows a course similar to the trace obtained for V alone /pa/ from ~50ms prior to auditory onset and peaking at ~50ms post-auditory onset (white and gray traces respectively). These observations also suggest (cf. Figure 3) that the degree of ambiguity in the visual domain drives the magnitude of auditory modulation.

4. Discussion

We show that when using natural AV syllables, the visual modulation of auditory neural processing is observed as temporal facilitation and as amplitude reduction as early as ~50-100ms post-auditory onset. The pattern of activity we found with AV speech differs notably from non-speech AV stimuli [19][20][21] and indicates potential speech specific mechanisms. In particular, we observed an amplitude *reduction* of the auditory N1/P2 complex, whereas most multisensory studies found an enhancement effect. A crucial difference with prior studies resides in their use of non-perceptually-relevant multisensory stimuli for which no categorization or obvious 'meaningful' association could be drawn (e.g. tones paired with morphing or static circles). Consequently, the type of interaction observed could in fact derive from being (or not being) a naturally occurring multisensory signal.

In addition to subcortical multisensory sites of integration (i.e. superior colliculus) a growing body of anatomical and neurophysiological evidence shows that cross-modal interactions in sensory specific cortices are present early on (e.g. [25][26]). It is now widely accepted that early multisensory interactions can affect early processing stages of sensory-specific pathways. Thus, while there is evidence that visual speech alone can access auditory cortices [22], the assumption of speech-specific activation remains vague considering that cross-sensory connectivity is present throughout all sensory modalities i.e. if by design, the neural architecture favors cross-sensory activations, it does not necessarily specify the nature of exchanged information nor its function. The response enhancement of auditory cortices by visual speech found in the fMRI literature (e.g. [23]) is also unclear, as 'supra-additivity' relates to integration in multisensory neural population. The "supra-additive" effect refers to the fact that multisensory neurons (subcortical and cortical) respond to multisensory events with a much higher rate than would be expected by simple summation of their unimodal responses [18]. This effect does not *a priori* apply to cortical auditory neurons and it is therefore unclear how such an enhancement could arise in auditory cortex.

Our results rather converge on the hypothesis of speech-specific visual modulation of auditory processing. A recent finding, where AV vowels showed an amplitude reduction of P1 (or P50) as compared to auditory alone conditions, further supports this claim [27]. Our finding also suggests that AV speech integration occurs early (prior to phonetic integration) as the amplitude modulation and the temporal facilitation of auditory event-related potentials occurs prior to ~100ms and extends to ~200ms. A number of speech studies converge on a ~200ms integration time window [7][8][9][13]. Our results bring further support to the proposal that this time scale corresponds to the time constant for perceptual unit formation in auditory cortex [17][28][29].

We found that visual speech inputs decreased the processing time of the early auditory evoked potentials by approximately ~20ms. The amount of temporal gain shows interindividual variability, and its locus tends to depend upon (1) the expectancy of auditory inputs and (2) the predictability of the auditory stimulus given the accuracy in the visual domain. Because this is the first report of the kind, follow-up studies will be needed to establish more precisely which factors correlate with the temporal gain. It is nevertheless noteworthy that if amplitude increase has often been intuitively associated with neural facilitation (e.g. via increased attentional resources), temporal gain may also play an important role.

Additionally, a trend for an earlier locus of facilitation was found in AV syllables associated with a salient visual input (e.g. /pa/), which suggests that visual information may help predict the auditory input. This is further supported by the effect of participants' expectation level on the locus of the temporal gain observed in Experiment 1 and 2. The natural precedence of visual motion in AV speech (~350-400ms) is likely to initiate the speech processing system and further constrain the extraction of auditory speech inputs. For instance, if a visual /pa/ is easily identified, less auditory information is needed for categorization (essentially voicing information), whereas in the case of an ambiguous /ka/, much auditory information is needed for visual disambiguation. Thus, if the ~200ms auditory time scale is overall modulated by visual inputs, the temporal gain of ~20ms suggests that the fine grain analysis of auditory speech is also subject to visual influence.

In addition, AV integration of incongruent speech differs from congruent speech within 200ms post-auditory onset conditional upon the degree of ambiguity in the visual domain. These results add to prior MEG and EEG findings, in which a mismatch negativity paradigm was used to differentiate congruent and incongruent AV speech [31][32] and lead to significant differences at ~180ms.

While the illusory /ta/ significantly differed from a real /ta/, no obvious trace of specific early integrative mechanism for incongruent speech leading to fusion was found compared to congruent speech. However, in a visual attention paradigm the McGurk combination - where no unique percept is being elicited- showed a highly modulated P1, suggesting an inhibition of the auditory information flow in favor of visual inputs. In the fusion case, a later processing stage differed with a lack of positivity at ~350ms, suggesting that the illusory percept was less representative than a congruent /ta/.

Our results suggest that integration of congruent and incongruent AV speech is not a one-staged process but rather originates through neural interactions at different time scales. In cases in which AV interactions were initiated early on (~50ms), effects were observed over a 200ms time scale. These results point to many aspects of auditory speech perception, and AV speech in particular. Syllabicity, spatio-temporal correlation of AV speech signals and robustness to AV asynchrony converge on a ~200ms time window approximating the suggested time needed for perceptual unit formation. Additionally, a temporal facilitation of ~20ms was found in early stages of auditory processing, a time relevant for phonetic (subsegmental) feature extraction. This result suggests that visual inputs influence auditory processing as early as the (pre) phonetic stage.

While further electrophysiological studies will be needed to precisely define the types of interaction and the nature of the information processed at each time scale, it is interesting to note that the amplitude decrease in our recordings can originate from two major causes: first, an elevation of the auditory thresholds and second, the desynchronization of auditory neural populations. While counter-intuitive, such mechanisms might precisely derive from prior processing in the visual domain. Further studies will focus on this issue.

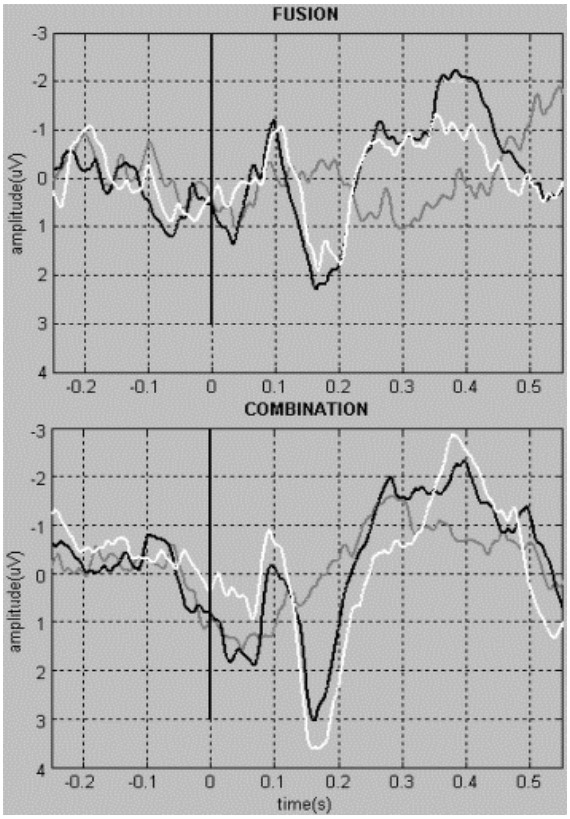

Figure 4 : Effect of visual attention on incongruent AV speech (Experiment 3, N=10)

Black traces are McGurk fusion (upper panel) and combination (lower panel), obtained in Experiment 3. White traces are McGurk fusion (upper panel) and congruent AV /pa/ (lower panel) obtained in Experiment 1. Gray traces are V /ka/ (upper panel) and V /pa/ (lower panel) obtained in Experiment 1. Note the positive deflection in McGurk combination similar to the one observed in visual /pa/ alone.

5. Acknowledgements

This work was funded by NIH grant DC 05660 to DP. We would like to thank Jonathan Z. Simon for his help on bootstrapping. . The opinions or assertions contained herein are the private views of the authors [KG] and are not to be construed as official or as reflecting the views of the Department of the Army or the Department of Defense.

6. References

[1] McGurk, H., and McDonald, J., "Hearing lips and seeing voices", *Nature,* 264, 746-747, 1976.
[2] Sumby, W.H., and Pollack, I., "Visual contribution to speech intelligibility in noise", *J. Acoust. Soc. Am., 26, 212-215,* 1954.

[3] MacLeod, A., and Summerfield, Q., "A procedure for measuring auditory and audio-visual speech-reception thresholds for sentences in noise: rationale, evaluation, and recommendation for use", *British Journal of Audiology, 24, 29-43,* 1990.
[4] Helfer, K.S., "Auditory and audio-visual perception of clear and conversational speech", *Journal of Speech, Language and Hearing Research1, 40, 432-43.* 1997
[5] Grant, K.W., Walden, B.E., Seitz P.F., "Auditory-visual speech recognition by hearing-impaired subjects: consonant recognition, sentence recognition and auditory-visual integration", *J. Acoust. Soc. Am., 103, 2677-2690,* 1998
[6] Massaro, D.W. (1987) Perceiving Talking Faces: From Speech perception to a Behavioral principle. Cambridge, MA: MIT Press, 1987.
[7] Massaro, D.W., Cohen, M.M., Smeele, P.M. (1996). "Perception of asynchronous and conflicting visual and auditory speech", *J. Acous. Soc. Am, 100, 1777-1786,* 1996.
[8] Munhall, K., Gribble, P., Sacco, L., Ward, M., "Temporal constraints on the McGurk effect", *Perception and Psychophysics, 58, 351-362,* 1996.
[9] van Wassenhove, V., Grant, K.W., Poeppel, D., "Temporal window of Integration in bimodal speech" (*submitted)*
[10] Celesia, G.G., "Organization of auditory cortical areas in man", *Brain, 99, 410:414,* 1976.
[11] Yvert, B., Fischer, C., Guénot, M., Krolak-Salmon, P., Isnard, J., Pernier, J., "Simultaneous intracerebral recordings of early auditory thalamic and cortical activity in human", *European Journal of Neuroscience, 16, 1146-1150,* 2002.
[12] Ffytche, D.H., Guy, C.N., Zeki, S., "The parallel visual motion inputs into areas V1 and V5 of human cerebral cortex", *Brain, 118, 1375-1394,* 1995.
[13] Arai, T., and Greenberg, S., "The temporal properties of spoken Japanese are similar to those of English" Proceedings of Eurospeech, Rhodes, Greece, *1011-1014*, 1997.
[14] Summerfield, Q., "Lipreading and audio-visual speech perception", *Phil. Trans. R. Soc. Lond. B, 335, 71-7,* 1992.
[15] Grant, K. and Greenberg, S., (2001), "Speech intelligibility derived from asynchronous processing of auditory-visual information", *Proceedings of the workshop on Auditory-Visual speech processing (AVSP),* 2001.
[16] Rosen, S., "Temporal information in speech: acoustic, auditory and linguistic aspects", *Phil. Trans. R. Soc. Lond. B, 336, 367-373,* 1992.
[17] Poeppel, D., "The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling time'", *Speech Communication.(in press)*
[18] Stein, B.E., and Meredith, A.M., The merging of the senses. *Cambridge, MA: MIT Press,* 1993.
[19] Giard, M.-H., and Peronnet, F., "Auditroy-visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study", *Journal of Cognitive Neuroscience, 11(5), 473-490,* 1999.
[20] Molhom, S., Ritter, W., Murray, M.M., Javitt, D.C., Schroeder, C.E., Foxe, J.J., "Multisensory auditory-visual interactions during early sensory processing in humans: a high-density electrical mapping study", *Cognitive Brain Research, 14, 115-128,* 2002.
[21] Fort, A., Delpuech, C., Pernier, J., Giard, M.-H., "Dynamics of cortico-subcortical cross-modal operations involved in auditory-visual object detection in humans", *Cerebral Cortex, 12, 1031-1039,* 2002.
[22] Calvert, G.A., Campbell, R., Brammer, M.J., "Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex", *Current Biology, 10 (11) 649-656,* 2000.
[23] Calvert, G.A., Bullmore, E.T., Brammer, M.J., Campbell, R., Williams, S.C., McGuire, P.K., Woodruff, P.W., Iversen, S.D., David, A., "Activation of auditory cortex during silent lipreading", *Science, 2776, 593-596,* 1997.
[24] Efron, B., "Bootstrap methods: another look at the jackknife", *Annals of Statistics, 7, 1-26,* 1979.
[25] Watanabe, J., and, Iwai, E., " Neuronal Activity in visual, auditory and polysensory areas in the monkey temporal cortex during visual fixation task", *Brain Research Bulletin, 26, 583-592,* 1991.
[26] Falchier, A., Clavagnier, B., Kennedy, H., "Evidence of multimodal integration in primate striate cortex", *Journal of Neuroscience, 22(13), 5749-5759,* 2002.
[27] Lebib, R.,Papo, D., de Bode, S., Baudonnière P.-M, "Evidence of a visual-to-auditory cross-modal sensory gating phenomenon as reflected by the human P50 event-related brain potential modulation", *Neuroscience Letters, 341, 185-188,* 2003.
[28] Yabe, H., Winkler, I., Czigler, I., Koyama, S., Kakigi R., Sutoh, T., Hiruma, T., Kaneko, S., "Organizing sound sequences in the human brain: the interplay of auditory streaming and temporal integration", *Brain Research, **897**, 222-227*, 2001.
[29] Näätänen, R., Attention and Brain Function. *Hillsdale, NJ: Lawrence Erlbaum Associates.*1992.
[30] Sams, M., Aulanko, R., Hämäläinen, M., Hari, R., Louassmaa, O.V., Lu, S.T., Simola, J., "Seeing speech: visual information from lip movement modifies activity in the auditory cortex." *Neuroscience Letters, 127, 141:145*, 1991.
[31] Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F., Deltenre, P., "Mismatch negativity evoked by the McGurk-MacDonald effect: a phonetic representation within short-term memory." *Clinical Neurophysiology, 113, 495-506,* 2002.