# NEURAL AREAS UNDERLYING THE PROCESSING OF VISUAL SPEECH INFORMATION UNDER CONDITIONS OF DEGRADED AUDITORY INFORMATION

*Daniel Callan*, Akiko Callan*, and Eric Vatikiotis-Bateson†*

*Brain Activity Imaging Center, Information Sciences Division, ATR International, Japan
†Communication Dynamics Project, Information Sciences Division, ATR International, Japan

## ABSTRACT

The goal of this study was to localize, using fMRI, the neural processes involved with visual aspects of speech perception under conditions of degraded auditory information. Brain activity underlying aspects of visual speech processing was determined by comparing conditions with visual speech information to appropriate auditory only conditions. Consistent with the idea of a 'mirror neuron system,' results suggest that speech motor areas (Broca's area) of the brain may be involved with the recognition of phonetic gestures inherent in the visual speech signal under conditions of degraded auditory information.

## 1. INTRODUCTION

It is known that observation of visual speech behavior that is concordant with the auditory signal can enhance intelligibility especially under conditions in which the auditory signal is degraded by noise [1]. In a previous case study [2] it was found using EEG that the visual enhancement of speech intelligibility in noise involves high frequency activity in the superior temporal gyrus. The EEG current source density was localized on the surface of the cortex using low-resolution tomography (LORETA) constrained by individual specific volume conductor and source models constructed from anatomical MRI data. This finding is consistent with functional imaging studies of silent speechreading [3,4] and concordant audiovisual speech perception [5]. The EEG case study [2] also revealed the visual enhancement effect to involve a distributed functional network including parietal, occipital, temporal, frontal, and sensorimotor areas representing orofacial structures.

The finding that the visual enhancement effect includes areas of the brain involved with speech production is interesting given recent studies concerning the 'mirror neuron system' [6]. Mirror neurons are thought to contribute to processes underlying both observation and execution of motor action. Interestingly, a mirror neuron system for hand movement imitation has been found to involve Broca's area [6]. It has been proposed that a mirror neuron system (including Broca's area) for gesture recognition underlies human communication and furthermore that it is responsible for the evolution of language in humans [7]. It is maintained that the mirror neuron system underlying speech involves both auditory as well as visual observation of gesture [7].

Broca's area has been shown to be active during speech perception tasks involving such things as phonetic judgement [8], second-language processing [9], processing of rapid presentation of stimuli [10], verbal repetition [10,11], auditory-verbal storage [12], as well as semantic processing [13]. However, under conditions in which the audio speech signal is clear and the processing demands are low there is no or only minimal activity in Broca's area [8,10].

It is maintained that under conditions in which the audio speech signal has been degraded that the mirror neuron system is activated to facilitate speech perception (this is true as long as speech information is still audible). If visual speech gesture information is present under degraded audio conditions it will also activate the mirror neuron system in order to facilitate speech perception.

It is the goal of this study to localize the neural processes involved with visual aspects of speech perception under various conditions of degraded auditory information. The experiment consists of two concordant audiovisual AV conditions (audio with and without noise present), a visual only speech condition (no audio present), two audio only conditions (audio with and without noise present), and a control condition (video of a static face). The audio only conditions were also presented with video of a static face. This was to control for activation of visual areas of the brain resulting from nonlinguistic aspects of the audiovisual stimuli. The audio only conditions served as a referent so that brain activity involved with the contribution of visual aspects of speech perception could be determined. Consistent with the idea of a 'mirror neuron system' it is predicted that under conditions in which there is no auditory speech information, or when it has been degraded by noise, that speech motor areas of the brain including Broca's area will be involved with the recognition of phonetic gestures inherent in the visual speech signal.

# 2. METHODS

## 2.1. Subjects

Four right-handed male native English speakers participated in this study. Subjects were between 30 and 35 years of age. All subjects volunteered to participate in the study and gave informed consent.

## 2.2. Stimuli

The stimuli were similar to those used in [2]. The stimuli consisted of 96 different monosyllabic English words for each of the five conditions (audio only (AO), audio only with noise (AON), audiovisual (AV), audiovisual with noise (AVN), visual only (VO)). The words were spoken by a female native English speaker and were of the form: initial consonant cluster + vowel + final consonant cluster. The words for each condition were controlled for frequency, initial consonant, vowel, and final consonant. Each presentation was one second in duration including facial motion prior to and following the audio speech signal for the word. The words were recorded onto video laser disk for later stimulus presentation. One second of a static face (the same face shown during concordant AV conditions) with slight head movement was used for control and all audio only conditions. It is possible that presentation of audio speech with a face that is not moving in concordance may cause alteration of normal auditory speech processing. Further experiments need to be conducted to examine this issue.

Audio noise used in the experiment consisted of multispeaker babble that was mixed with the speech signal. Multispeaker babble was selected to degrade the auditory signal because its main energy is in the same frequency range as the word stimuli.

## 2.3. Procedure

A block design was used in which 12 words from the same condition were presented (approximately 85-90 dB SPL) at a rate of one every three seconds (inter-stimulus interval equals two seconds). This was followed by presentation of 12 words from another condition and so on. The order of conditions was kept fixed (VO-AO-AVN-control-AON-AV) to increase the fMRI signal sensitivity. Randomization decreases sensitivity by distributing the variance of the task paradigm over multiple frequencies [14]. This pattern was repeated twice for each session. In total there were four sessions. Each session was approximately seven and a half minutes. The task was to identify the word presented. All subjects had practice with the various types of stimuli prior to fMRI scanning. Behavioral results were recorded separately one week after fMRI scanning. For this task subjects were asked to write down the word that they perceived and then push a button to proceed to the next stimuli. The order of presentation for the various conditions was the same as in the fMRI experiment. Subjects were aware that there would be a post-test.

Video from the laser disk for the appropriate conditions was presented synchronously with the corresponding audio sound file. Audio was presented via MR-compatible headphones (Hitachi ceramic transducer headphones). Video was presented by a projector located outside of the MR room to a mirror positioned inside of the head coil just above the subjects' eyes. Stimulus presentation was controlled by specialized computer hardware-software that can drive the laser disk player as well as present audio sound files.

## 2.4. Imaging

Brain imaging was performed using a 1.5 Tesla Marconi Magnex Eclipse scanner. First, high-resolution anatomical T2 weighted images were acquired using a fast spin echo sequence. These scans consisted of 50 contiguous axial slices with a .75x.75x3mm voxel resolution covering the cortex and cerebellum. Second, functional T2* weighted images were acquired using a gradient echo-planar imaging sequence (echo time, 55ms; repetition time, 6000ms; flip angle, 90°). A total of 50 contiguous axial slices were acquired with a 3x3x3mm voxel resolution. The field of view included the cortex and cerebellum.

## 2.5. Data Analysis

Images were preprocessed using programs within SPM99b (Welcome Department of Cognitive Neurology, London UK). Differences in acquisition time between slices were accounted for, movement artifact was removed, and the images were then spatially normalized to a standard space using a template EPI image (Bounding Box, x= -90 to 91 mm, y= -126 to 91 mm, z= -72 to 109 mm; voxel size, 3x3x3 mm). Images were smoothed using a 6-mm FWHM Gaussian kernel.

Regional brain activity for the various conditions was assessed on a voxel-by-voxel basis using SPM. A fixed effect model was employed (the data from all subjects were analyzed together). The data was modeled using a box-car function convolved with the hemodynamic response function. In addition, global normalization and grand mean scaling were carried out. Significance for the AO, AON, AV, AVN, and VO conditions contrasted with the control condition were assessed using a correction for multiple comparisons (corrected T=4.87, $p<0.05$, df=1056). Voxels in Broca's area were assesed using a region of interest correction for multiple comparisons (corrected T=3.58, p <0.05, df=1056). Brain activity involved with the contribution of visual aspects of speech perception was assessed by inclusively masking (p<0.05) the AV, AVN, and VO contrasts with the statistical contrast of the condition at hand compared with the appropriate audio only conditions (AV inclusively masked by AV-AO; AVN inclusively masked by AVN-AON; VO inclusively masked by VO-AO).

## 3. RESULTS

### 3.1. Behavioral Performance

Behavioral performance for each subject was assessed for word identification as well as phoneme cluster identification for all five conditions. The results (Figures 1A-B) indicate an enhancement in identification performance in noise by concordant visual speech information. The results also indicate that although word identification performance is poor for the visual only condition, there are a fair number of phonemes identified correctly.

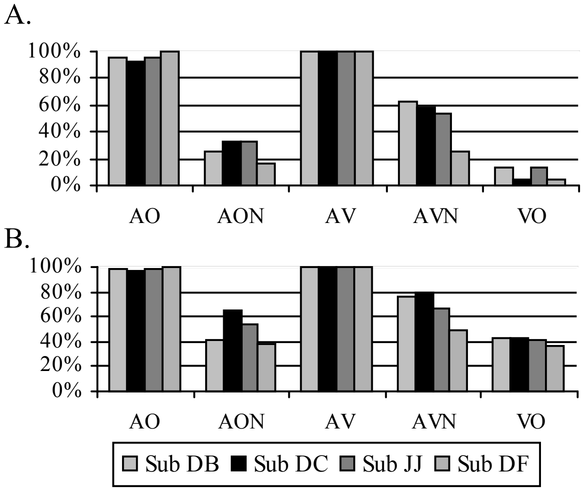


Figure 1: A. Percent correct word identification performance. B. Percent correct phoneme identification performance.

### 3.2. fMRI Analysis

Regional brain activity determined by SPM for the five conditions relative to the control condition are shown in Figure 2. All conditions including the visual only condition show a great deal of activity bilaterally in the superior and middle temporal areas (Brodmann area BA 22, 41, 42, and 21). The conditions containing visual speech stimuli (AV, AVN, VO) show additional activity in the right inferiorposterior temporal lobe (BA 37) and to a lesser degree in BA 19. This is confirmed by inclusive masking (AV inclusively masked by AV-AO; AVN inclusively masked by AVN-AON; VO inclusively masked by VO-AO) (Figure 3). Inclusive masking also revealed a small degree of activity for the VO (inclusively masked by VO-AO) condition in left BA 18 and BA 21 (Figure 3).

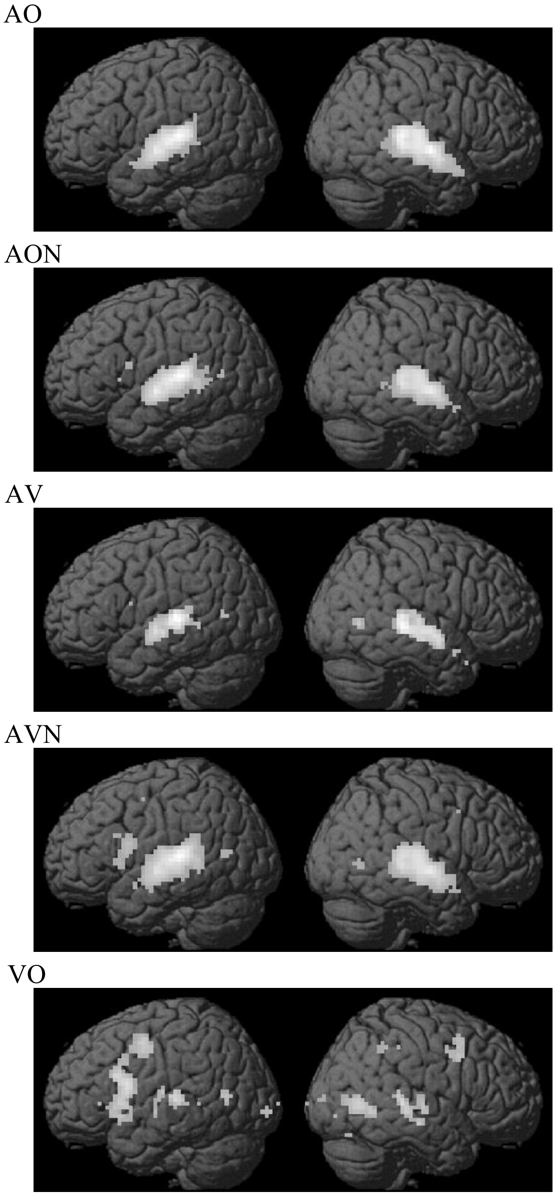
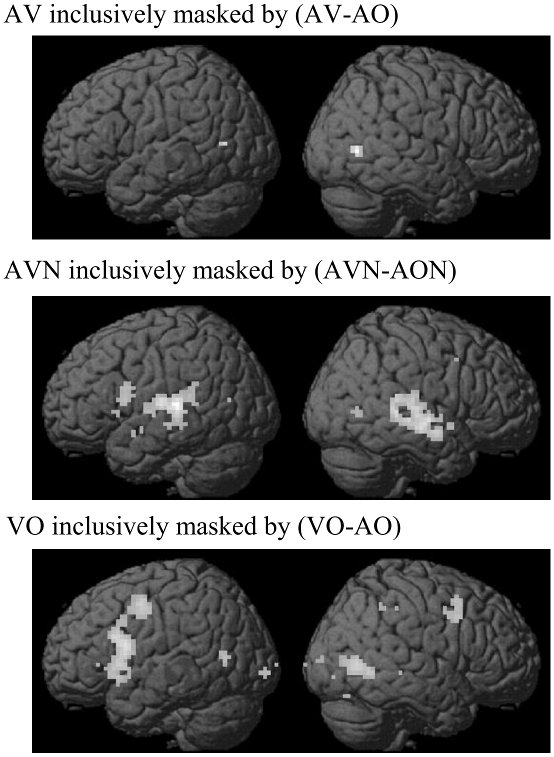


Figure 2: Regional brain activity relative to control condition. Brodmann areas are labeled.

Of primary interest was the finding that Broca's area (BA 44-45) is active for AVN (inclusively masked by AVN-AON as well as inclusively masked by AVN-AON, AVN-AO, and AVN-AV) and VO (inclusively masked by VO-AO as well as inclusively masked by VO-AON, VO-AO, and VO-AV) conditions (Figures 3 and 4). It should also be noted that the AON compared to the control condition contrast (Figure 2) also shows some degree of activity in Broca's area.

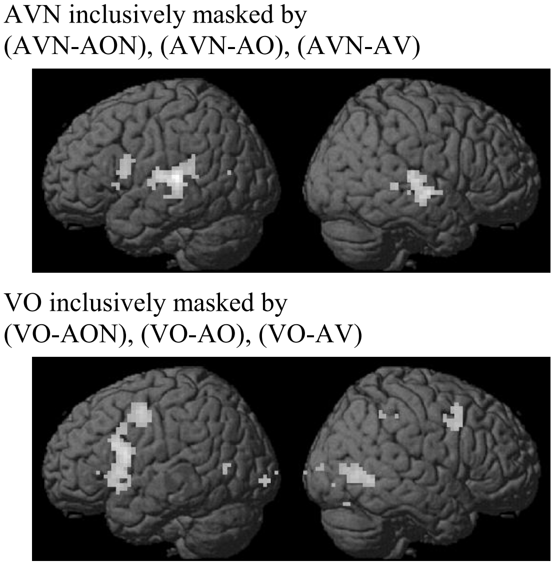


AV inclusively masked by (AV-AO)

AVN inclusively masked by (AVN-AON)

VO inclusively masked by (VO-AO)

Figure 3: Regional brain activity for conditions containing visual speech stimuli inclusively masked by the statistical contrast of the condition at hand compared with the appropriate audio only conditions. Brodmann areas are labeled.



AVN inclusively masked by
(AVN-AON), (AVN-AO), (AVN-AV)

VO inclusively masked by
(VO-AON), (VO-AO), (VO-AV)

Figure 4: Regional brain activity for conditions containing visual speech stimuli inclusively masked by the statistical contrast of the condition at hand compared with AON, AO, and AV conditions. Brodman areas are labeled.

The AVN (inclusively masked by AVN-AON as well as inclusively masked by VO-AON, VO-AO, and VO-AV) condition in addition shows activity bilaterally in BA 21, 22, 41, and 42 (Figure 3 and 4). For the VO (inclusively masked by VO-AO; as well as inclusively masked by VO-AON, VO-AO, and VO-AV) condition additional activity occurs bilaterally in the premotor cortex BA 6, prefrontal cortex BA 9, BA 18, and supplementary motor area SMA (Figures 3 and 4). Activity is also located in left BA47, left anterior insula, left anterior BA 22, right BA 19, left BA 37, and right BA 40 (Figures 3 and 4).

## 4. DISCUSSION

The results of the SPM analysis show a great deal of activity for all conditions in the superior and middle temporal areas when compared to the control static face only condition (Figure 2). Studies have indicated that this area is involved with not only auditory speech processing but also with multimodal audiovisual speech processing [2,5]. It is interesting to note that consistent with other studies investigating silent speechreading [3,4] that the visual only condition showed activity in superior and middle temporal areas, including the primary auditory cortex, (BA 21, 22, 41, 42), even though there was no auditory signal (Figure 2).

In order to find brain activity underlying aspects of visual speech processing, conditions that consist of visual speech information were inclusively masked by the statistical contrast of the condition at hand compared with the appropriate audio only conditions (Figure 3). Activity was found bilaterally in superior and middle temporal areas, including the primary auditory cortex, (BA 21, 22, 41, 42), for the AVN (inclusively masked by AVN-AON as well as inclusively masked by VO-AON, VO-AO, and VO-AV) condition but not for the AV (inclusively masked by AV-AO) condition (Figures 3 and 4). These results are consistent with [2], in which almost identical stimuli were used, demonstrating a multimodal enhancement effect afforded by visual speech information localized in superior and middle temporal areas under conditions in which the audio signal has been degraded by noise.

Consistent with findings in other studies concerning audiovisual speech perception [3,4,5], inclusive masking by the statistical contrast of the condition at hand compared with the appropriate audio only conditions revealed unique brain activity in the right inferiorposterior temporal lobe (BA 37) for all conditions with visual speech information (Figure 3). The region of BA 37 activated includes visual motion processing area V5. One might conclude that activation of this area is a result of lip and jaw movement in conditions containing visual speech information versus conditions in which only a static face is shown. However, activity in BA 37 was present for the VO condition when inclusively masked by the VO-AV contrast (Figure 4). These results are consistent with findings implicating BA 37 as a multimodal processing area [15].

The finding of primary interest was the presence of activity in Broca's area (BA 44-45) in the AVN (inclusively masked by AVN-AON, AVN-AO, and AVN-AV) and the VO (inclusively masked by VO-AON, VO-AO, and VO-AV) conditions (Figure 4). It should also be noted that activity in Broca's area was also present in the AON compared to the control contrast. Consistent with the idea of a 'mirror neuron system' [6,7] the results suggest that speech motor areas of the brain may be involved with the recognition of phonetic gestures inherent in the visual speech signal under conditions in which the audio signal is degraded. The activation of Broca's area for the AVN and VO conditions (Figure 4) cannot be explained by task difficulty alone since behavioral results indicate that the AON condition is more difficult than the AVN condition. In addition to BA 44 the VO condition (Figure 4) also shows unique activity in other speech motor areas (BA 6, 45, 47, anterior insula, and SMA), some of which have been found to be involved with speech perception and multimodal processing [2,3,5,11,16].

One possible explanation for why Broca's area BA 44 was found to be active in this study but not in other studies involving silent speechreading may be because in this study the task required identification of a monosyllabic word (out of hundreds of potential English words) rather than merely identifying a spoken number between one and ten [3,4]. It is likely that the speechreading task used in this study requires greater phonetic processing, which may invoke greater activation of Broca's area, than the task used in other studies.

## 5. ACKNOWLEDGEMENTS

I would like to thank Kevin Munhall for his help in making the laser disk of the stimuli and providing hardware and software to control the laser disk presentation of the stimuli. I would also like to thank Takaaki Kuratate for his help with the stimulus construction. In addition I would like to thank the fMRI technicians Yasuhiro Shimada and Ichiro Fujimoto. Correspondence to dcallan@isd.atr.co.jp.

## 6. REFERENCES

1. Sumby W. and Pollack, I. "Visual contribution to speech intelligibility in noise," *JASA* 26: 212-215, 1954.
2. Callan, D., Callan, A., Kroos, C., and Vatikiotis-Bateson, E. "Multimodal contribution to speech perception revealed by independent component analysis: a single-sweep EEG case study," *Cogn. Brain Res.* 10: 349-353, 2001.
3. Calvert, G., Bullmore, E., Brammer, M., Campbell, R., Williams, S., McGuire, P., Voodruff, P., Iversen, S., and David, A. "Activation of auditory cortex during silent lipreading," *Science* 276: 593-596, 1997.
4. MacSweeney, M., Amaro, E., Calvert, G., Campbell, R., David, A., McGuire, P., Williams, S., Woll, B., and Brammer, M., "Silent speechreading in the absence of scanner noise: an event-related fMRI study," *NeuroReport* 11: 1729-1733, 2000.
5. Calvert, G., Campbell, R., and Brammer, M. "Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex," *Curr. Biol.*10: 649-657, 2000.
6. Iacobini, M., Woods, R., Brass, M., Bekkering, H., Mazziotta, J., and Rizzolatti, G. "Cortical mechanisms of human imitation," *Science* 286: 2526-2528.
7. Rizzolatti, G. and Arbib, M. "Language within our grasp," *TINS* 21 (5): 188-194, 1998.
8. Zatorre, R. and Binder, J. "Functional and structural imaging of the human auditory system," in Toga, A. and Mazziotta, J. (Eds.) *Brain Mapping the Systems*, Academic Press, San Diego, 365-402, 2000.
9. Nakai, T., Matsuo, K., Kato, C., Matsuzawa, M., Okada, T., Glover, G., Moriya, T., and Inui, T. "A functional magnetic resonance imaging study of listening comprehension of languages in human at 3 tesla-comprehension level and activation of the language areas," *Neurosci. Lett.* 263: 33-36, 1999.
10. Price, C., Wise, R., Warburton, E., Moore, C., Howard, D., Patterson, K., Frackowiak, R., and Friston, K. "Hearing and saying: the functional neuro-anatomy of auditory word processing," *Brain* 119: 919-931, 1996.
11. Callan, D., Callan, A., Honda, K., and Masaki, S. "Single-sweep EEG analysis of neural processes underlying perception and production of vowels," *Cogn. Brain Res.* 10: 173-176, 2000.
12. Smith, E. and Jonides, J. "Storage and executive processes in the frontal lobes," *Science* 283: 1657-1661, 1999.
13. Chee, M., O'Craven, K., Bergida, R., Rosen, B., and Savoy, R. "Auditory and visual word processing studied with fMRI," *Human Brain Mapping* 7: 15-28, 1999.
14. Aguirre, G., and D'Esposito, M. "Experimental design for brain fMRI," in Moonen, C. and Bandettini (Eds.) *Functional MRI*, Springer, Berlin, 369-380, 1999.
15. Buchel, C., Price, C., and Friston, K. "A multimodal language region in the ventral visual pathway," *Nature* 394: 274-277, 1998.
16. Fuster, J., Bodner, M., and Kroger, J. "Cross-modal and cross-temporal association in neurons of frontal cortex," *Nature* 405: 347-351, 2000.