Harvard-MIT Division of Health Sciences and Technology
HST.722J: Brain Mechanisms for Hearing and Speech
Course Instructors: Bertrand Delgutte, M. Christian Brown,
Joe C. Adams, Kenneth E. Hancock,
Frank H. Guenther, Jennifer R. Melcher,
Joseph S. Perkell, David N. Caplan

HST.722 Student Topics Proposal
Erik Larsen

Cortical correlates of audio-visual integration

Introduction

Animals have multiple sensory systems with which they can probe the external environment. Oftentimes, objects or events in the environment produce signals to which more than one sensory system responds, and it is the task of the cerebral cortex to integrate these separate modalities into a unified percept. We take this process for granted, but it is in fact a difficult problem, which becomes obvious when we start to ask ourselves how we would design a system to do the same. The current state of knowledge described in the literature is quite limited and is mostly phenomenological, i.e. describing how auditory and visual inputs can either modulate each other and/or activate association areas of the cortex. There does not seem to be a mechanistic understanding of how this works.

I have collected a number of papers that look at audio-visual integration at different levels, from single-unit studies to cortical activation patterns to behavioral. The methods used range from intracellular recordings to scalp-recorded ERP (event-related potentials) to fMRI / MEG / PET imaging. For your pleasure and convenience, most of these papers are quite short. I have chosen three of them to discuss in class (Discussion papers), three for background, and quite a few for 'further reading' (if you only want to read a few of these, the most interesting for me are marked by ♦).

Behavioral effects of audiovisual integration

Ultimately multi-modal (or multi-sensory) integration should lead to noticeable behavioral effects. For human, the most interesting case is speech perception, which is usually considered an auditory modality but is strongly influenced by visual input. Being able to see the person that is talking to you leads to an increased intelligibility, especially in adverse conditions (high noise, reverberation, competing talkers) as shown by Sumby and Pollack (1954). The visual input gives redundant cues to reinforce the auditory stimulus but also disambiguates some speech sounds which differ in place of articulation but sound similarly (such as /ba/ vs. /da/). Sumby and Pollack (1954) have shown that especially at low acoustic signal-to-noise ratios, the visual signal can dramatically increase word recognition, in some cases from near zero to 70% or 80%. In higher signal-to-noise ratio conditions, the visual signal still contributes, but the absolute effect is smaller because auditory performance alone is already high.

Beside this synergistic effect of auditory and visual input, there are a few other effects that clearly demonstrate the strong interaction of these two modalities. The first is the so-called 'ventriloquist' effect, where a synchronous yet spatially separate audio and visual signal is heard as originating from the visual location. Macaluso et al. (2004) were interested in finding brain areas that mediate this integration of spatially separate yet synchronous information into a single percept, through PET scans of human brains, and identified an area in the right inferior parietal lobule to be activated especially in this

condition. Bushara et al. (2001) used PET scans to identify the right insula being most strongly involved in audiovisual synchrony-asynchrony detection.

Another famous audiovisual 'illusion' is the 'McGurk effect' (McGurk and MacDonald, 1976), where the sound /ba/ is combined with the visual image of a talker articulating /ga/, leading to a robust percept of /da/. The usual explanation for this effect is that the brain tries to find the most likely stimulus given the (in this case conflicting) auditory and visual cues. Note that this effect is extremely robust and not susceptible to extinction, even after hundreds of repetitions. This effect has contributed to the notion that audiovisual integration occurs at a pre-lexical stage, early in the neural processing pathway.

Neural correlates of audiovisual integration for 'simple stimuli'

Given the strong behavioral effects of audiovisual integration described above, it would be very interesting to explore the neural basis for these. However, we will first explore some more basic properties of audiovisual integration.

The canonical view of cortical sensory processing is that each sensory modality has a primary unimodal cortex, several higher-order unimodal association cortices, and that finally the various sensory modalities interact in multimodal association cortices. For the auditory case, the primary auditory cortex is located in the superior temporal gyrus or BA41 (see Fig. 1 for an overview of Brodmann areas). This area is surrounded by a belt and parabelt region, which are the auditory association areas. The middle and inferior temporal gyri (BA 20, 21, 37) are multisensory association areas, mainly auditory and visual. In fMRI studies of human brain activation, these multimodal areas activate uniformly in response to multimodal stimuli. Using high-resolution fMRI, it has recently been shown by Beauchamp et al. (2004) that in fact these multisensory areas (at least in the superior temporal sulcus) contain a patchwork of auditory, visual, and audiovisual areas, and each a few mm in size. It appears that the various unimodal areas send projections to small patches of multisensory cortex, after which the modalities are integrated in the intervening patches.

From human imaging and animal studies it is clear that there are special cortical areas which have multisensory responses. Komura et al. (2005) found that multisensory responses can also be found at lower levels, specifically in the auditory thalamus – the medial geniculate body (MGB). Traditionally, the thalamus is thought of as a relay station between brainstem/spinal cord and cortex, sending signals upward; but it is also known that it receives massive projections from the cortex itself. Komura and coworkers recorded from MGB shell neurons (which receive the cortical projections) in rat during a reward-based auditory spatial discrimination task, which was paired with an irrelevant yet variable light stimulus. Although MBG neurons did not respond to the light stimulus alone, the response was strongly modulated by the visual input, in particular the response was greater when auditory and visual stimuli were matched and smaller when they were conflicting. The matched and conflicting condition also led to a decrease vs. increase in reaction time, respectively, which again demonstrates the utility of integrating multiple modalities (which would usually be in agreement with one another, leading to a faster reaction). Interestingly, MGB shell neurons were also modulated by the amount of reward, although that is beyond the scope of our discussion.

A conceptual difficulty with studies of multisensory integration is that the paradigm does not always permit one to be sure that integration has in fact occurred. For example, in the mentioned study by Macaluso et al. (2004) where spatially separated audio and video was used, it is in principle possible that there was no integration, even though one would expect it based on prior psychophysical data. Also, conditions that are designed to produce multisensory integration often necessarily use somewhat different stimuli than conditions that are not aimed at producing integration. Bushara et al. (2003) devised an ingenious method to circumvent these problems, and were able to use exactly identical stimuli that sometimes produced integration and sometimes did not. By comparing brain activation (BOLD fMRI) from either category they were able to find a correlation between larger activation in specific areas and the (non-)occurrence of integration. Because of these results, they propose that multisensory association areas work in parallel with primary sensory areas, instead of as a higher-level end-station, which is the usual view. This would agree with Komura et al. (2005) who showed that error rates were sensitive to matched or conflicting *simultaneous* audiovisual signals, which also had strong neural correlates (enhanced or depressed rate responses).

A completely different aspect of multisensory effects on sensory processing is attentional cueing, i.e. one modality biases the observer such that the other modality will respond preferentially to the cued object. As an example, consider hearing a familiar voice calling your name from a crowd of people; you will reflexively turn towards the sound location and focus your visual attention of the same location, leading you to find your friend more quickly than had he been silent. Such cueing effects are well documented in psychophysical experiments, and it has been proposed that by directing attention towards an object, the neural signal from that object propagates faster through the brain. This is the proposed neural correlate of judging attended objects to appear earlier than unattended objects, even if they appear simultaneously. McDonald et al. (2005) used a sound cue in a visual task where subjects were required to judge which of two lights switched on earlier. They were able to show by measuring event-related potentials (ERPs) from the scalp that the attended (cued) light yielded a larger ERP than the unattended light, even if they were simultaneous. Somehow this difference in magnitude is translated into a delay in subsequent stages of processing that lead to perception. These findings may again corroborate to some extent the findings of Komura et al. (2005) in that cueing by one modality can influence the response of neurons in the other modality, with subsequent clear-cut behavioral consequences.

Neural correlates of audiovisual integration in speech perception

We have already described how important visual signals are for speech perception, both in the synergistic sense of aiding speech intelligibility (Sumby and Pollack, 1954) as well as in creating illusions such as the McGurk effect (McGurk and MacDonald, 1976). In the previous section we explored how relatively simple audiovisual stimuli that may or may not be fused into single objects can influence neural responses, both for individual neurons as well as for the whole brain. How is this for speech? Does brain activity differ between purely listening to speech as compared to listening *and* seeing speech (beside predictable activation of purely visual sensory areas)? And what about purely seeing speech (no sound)? Does this activate any of the same cortical areas? The latter question

of as academic interest in understanding language processing, but also has a more practical importance in that this is how many deaf people 'listen' to others, i.e. by lipreading, also called speechreading. It would have great value to know what areas are the most important for speech reading, and whether differences in speechreading ability (which are large between people) have a identifiable neural basis.

The classic paper in this context is Calvert et al. (1997). They used fMRI to find brain areas of increased activity when either listening to speech or silent lipreading, and the surprising finding was that silent lipreading activates primary and association auditory cortices. This was interpreted to mean that audiovisual integration occurs at a very early level in the neural pathway, even before association areas are activated (although these may cooperate in parallel, instead of hierarchical, cf. Bushara, 2003). Auditory areas were not activated by closed-mouth, non-speech movements of the lower face. A series of subsequent experiments similar to these were conducted, and with improvements in fMRI technology it became controversial whether primary auditory cortex is indeed activated by silent lipreading. Most investigators failed to find group-averaged activation of primary auditory cortex, although auditory association areas (e.g. BA 42, 22) do reliable activate during silent lipreading. Hall et al. (2005) found such kind of result, with the exception that for some proficient lipreaders, superior temporal gyrus did activate. It seems that currently there is no unambiguous answer to the question whether silent lipreading activates primary auditory cortex; there seems to be evidence pro and contra. Interestingly, Hall et al. (2005) describes other cortical areas that vary in activity as a function of speechreading proficiency. For example, high activity in the left inferior frontal gyrus was associated with poor speechreading ability. The explanation is that the greater task difficulty requires more extensive use of cognitive processes, which are located in the frontal lobe. However, the left inferior frontal gyrus is Broca's area (BA 44, 45), traditionally assumed to support articulatory-based mechanisms of speech production and executive aspects of semantic processing. From these and other recent results, it is becoming increasingly clear that Broca's area is probably also involved in language *comprehension*, supporting the formation of syntactic and semantic structure and syntactic working memory.

Paulesu et al. (2003) also found activation the perisylvian language area and of Broca's area (BA 44) in lipreading, also for non-lexical lipreading (NLLR, formed by playing video backwards). Lexical lipreading (LLR, forward video) differentially activated the more anterior part of Broca's area (BA 45), and also the left inferior temporal cortex. Therefore, Paulesu and coworkers assumed that these two areas might be particularly important for lexical access in lipreading. As in other recent imaging studies of lipreading, they found activation of auditory association areas (but not primary auditory areas) for both LLR and NLLR.

Conversely, auditory input can also activate visual sensory areas, as studied by Giraud and Truy (2002). In both normal and cochlear-implant subjects the fusiform gyrus and early visual cortex (BA 18, 19) was activated by listening to speech (no visual signal). It is assumed that the *expectancy* of visual correlates of speech are responsible for this effect. In cochlear-implant patients, the visual area activation was much greater than for normal-hearing subjects, showing the greater reliance they presumably place on visual cues. This shows that multimodal integration is flexible and can be strengthened when one modality is degraded.

References

Background

- W.H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *J. Acoust. Soc. Am.* **26**(2): 212-215, 1954.
  Nice short paper showing the (large) benefit (in terms of words correctly perceived) of seeing the face (mouth) that is talking to you, with an interesting interaction of the size of the vocabulary.

- H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature* **264**: 746-748, 1976.
  The famous 'McGurk effect' described: an auditory /ba/ paired with a visual /ga/ leads to a perception of /da/.

- G.A. Calvert, E.T. Bullmore, M.J. Brammer, R. Campbell, S.C.R. Williams, P.K. McGuire, P.W.R. Woodruff, S.D. Iversen, A.S. David, "Activation of auditory cortex during silent lipreading," *Science* **276**: 593-596, 1997.
  Classic paper on neuroimaging of speechreading; was the first to show auditory cortex activation from speechreading alone. Later studies have usually found that primary auditory cortex is *not* activated by speechreading alone, throwing some controversy on the issue.

Discussion papers

- K.O. Bushara, T. Hanakawa, I. Immisch, K. Toma, K. Kansaku, and M. Hallet, "Neural correlates of cross-modal binding," *Nature Neurosci.* **6**(2): 190-195, 2003.
  Shows that specific cortical areas are more active (fMRI) when an audiovisual stimulus is *perceived* as one, compared to when *exactly* the same audiovisual stimulus is *perceived* as separate unimodal events.

- G.A. Calvert and R. Campbell, "Reading speech from still and moving faces: The neural substrates of visible speech," *J. Cogn. Neurosci.* **15**(1): 57-70, 2003.
  Visible speech aids speech reception both when moving images are presented as well as still images. In this study, fMRI was used to asses if different cortical areas subserve these two visual categories. They found that similar language-based cortical areas were activated in either case, although stronger activation was seen for the moving images.

- Y. Komura, R. Tamura, T. Uwano, H. Nishijo, and T. Ono, "Auditory thalamus integrates visual inputs into behavioral gains," *Nature Neurosci.* **8**(9): 1203-1209, 2005.
  Single-unit responses in rat MGB (belt region) are modulated by visual stimulus and reward, in an audiovisual discrimination task. Shows that neuron response predicts behavioral response in hit/miss/reject/false alarm conditions; also shows neural and behavioral correlates of task difficulty.

Further reading (♦ most interesting)

- T. Raij, K. Uutela, and R. Hari, "Audiovisual integration of letters in the human brain," *Neuron* **28**: 617-625, 2000.
  MEG imaging of human brain in response to visual, auditory, and audiovisual letters. It was found in the superior temporal sulcus audiovisual responses are suppressed relative to the unimodal responses, in contrast to single-unit responses, which are usually potentiated for multimodal stimuli.

- K.O. Bushara, J. Grafman, and M. Hallet, "Neural correlates of auditory-visual stimulus onset asynchrony detection," *J. Neurosci.* **21**(1): 300-304, 2001.

The authors wanted to identify which brain areas are primarily involved in detection audio-visual asynchrony, and found the right insula to correlate best with the asynchrony detection effort as measured by response latency.

- A.L. Giraud and E. Truy, "The contribution of visual areas to speech comprehension: a PET study in cochlear implant patients and normal-hearing subjects," *Neuropsychologia* **40**: 1562-1569, 2002.
  Cochlear-implant patients (but not normal-hearing subjects) produce strong visual cortex activation when hearing speech.

- E. Paulesu, D. Perani, V. Blasi, G. Silani, N.A. Borghese, U. De Giovanni, S. Sensolo and F. Fazio, "A functional-anatomical model for lipreading," *J. Neurophysiol.* **90**: 2005-2013, 2003.
  A PET-scanning study of lexical and non-lexical (reversed video) lipreading implicates Broca's area in speech perception, as well as some other areas. Associative (non-primary) auditory cortex was activated for both stimulus conditions.

- ♦ E. Macaluso, N. George, R. Dolan, C. Spence, and J. Driver, "Spatial and temporal factors during processing of audiovisual speech: A PET study," *NeuroImage* **21**: 725-732, 2004.
  Using PET scanning the authors identify cortical regions that are sensitive to the synchrony and/or spatial coincidence of visual and auditory speech.

- ♦ M.S. Beauchamp, B.D. Argall, J. Bodurka, J.H. Duyn, and A. Martin, "Unraveling multisensory integration: patchy organization within human STS multisensory cortex," *Nature Neurosci.* **7**(11): 1190-1192, 2004.
  The authors show with high-resolution fMRI that audiovisual association cortex has patches of auditory, visual, and audiovisual cortex, instead of a homogenous audiovisual cortex.

- D.A. Hall, C. Fussell, and A.Q. Summerfeld, "Reading fluent speech from talking faces: Typical brain networks and individual differences," *J. Cogn. Neurosci.* **17**(6): 939-953, 2005.
  Includes a literature overview of neuroimaging studies of speech/lip-reading. Investigates with fMRI which cortical areas are active both in auditory speech and visual speech comprehension, and defines some cortical areas that correlate with individual speechreading proficiency.

- ♦ J.J. McDonald and W.A. Teder-Sälejärvi, F. Di Russo, and S.A. Hillyard, "Neural basis of auditory-induced shifts in visual time-order perception," *Nat. Neurosci.* **8**(9): 1197-1202, 2005.
  Measuring ERP (event-related potentials) from human scalp the authors show that time-order judgments correlate with the magnitude of ERP, not its timing. In particular, they find no evidence for the prior held belief that attended stimuli propagate faster through the CNS than unattended stimuli. Auditory cues were used to bias attention towards one of two visual cues, which could appear simultaneously or with a time delay.

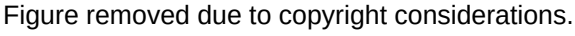

Figure 1. Outline of Brodmann areas with functional attribution. From http://spot.colorado.edu/~dubin/talks/brodmann/brodmann.html.