Published in final edited form as:

Lang Cogn Neurosci. 2016 ; 31(2): 284–302. doi:10.1080/23273798.2015.1101145.

Can you hear me yet? An intracranial investigation of speech and non-speech audiovisual interactions in human cortex

Ariane E. Rhone, Kirill V. Nourski, Hiroyuki Oya, Hiroto Kawasaki, Matthew A. Howard III, and Bob McMurray

Abstract

In everyday conversation, viewing a talker's face can provide information about the timing and content of an upcoming speech signal, resulting in improved intelligibility. Using electrocorticography, we tested whether human auditory cortex in Heschl's gyrus (HG) and on superior temporal gyrus (STG) and motor cortex on precentral gyrus (PreC) were responsive to visual/gestural information prior to the onset of sound and whether early stages of auditory processing were sensitive to the visual content (speech syllable versus non-speech motion). Event-related band power (ERBP) in the high gamma band was content-specific prior to acoustic onset on STG and PreC, and ERBP in the beta band differed in all three areas. Following sound onset, we found with no evidence for content-specificity in HG, evidence for visual specificity in PreC, and specificity for both modalities in STG. These results support models of audio-visual processing in which sensory information is integrated in non-primary cortical areas.

Keywords

Electrocorticography; Cross-modal; Multisensory; Speech

Introduction

Speech perception is challenging. It is well-established that the acoustic signal is highly variable, context-dependent and fleeting. As a result, listeners often complement acoustic information with many sources of non-acoustic information to help resolve phonemes, words, and ultimately a talker's intended message during conversation.

One factor that has received considerable attention is the use of visual information reflecting the talker's speech gestures. The presence of this visual/gestural information improves speech perception under challenging circumstances in both normal hearing and hearing-

Corresponding Author: Ariane E. Rhone (ariane-rhone@uiowa.edu), Department of Neurosurgery, The University of Iowa, 200 Hawkins Drive, Iowa City, IA, 52242, USA, 1-319-335-7049.
Kirill V. Nourski (kirill-nourski@uiowa.edu)
Hiroyuki Oya (hiroyuki-oya@uiowa.edu)
Hiroto Kawasaki (hiroto-kawasaki@uiowa.edu)
Matthew A. Howard III (matthew-howard@uiowa.edu)
Department of Neurosurgery The University of Iowa
Bob McMurray (bob-mcmurray@uiowa.edu)
The University of Iowa Department of Psychological and Brain Sciences Department of Communication Sciences & Disorders
Department of Linguistics The University of Iowa

impaired listeners (Bernstein, Auer, & Takayanagi, 2004; Grant, Walden, & Seitz, 1998; Macleod & Summerfield, 1987; Sumby & Pollack, 1954). Moreover, the long history of work on the McGurk effect (Green & Kuhl, 1988; Macdonald & McGurk, 1978; McGurk & MacDonald, 1976) suggests visual/gestural information is not simply combined with the auditory input in an additive way. Rather, there is a complex interplay of sensory information involved in processing multimodal speech inputs.

There is still considerable theoretical debate around when, where, and how listeners integrate visual/gestural information with phonetic cues in the auditory input (see Besle, Bertrand, & Giard, 2009; Campbell, 2008 for reviews). These debates have implications beyond the matter of audiovisual (AV) integration, as the answers to these questions can inform the larger theoretical debates, such as the issue of whether speech is perceived as primarily an auditory or gestural signal at various stages of processing (Fowler, 1991; Grant et al., 1998; Holt & Lotto, 2008; Ohala, 1996), or the issue of whether speech is a purely bottom up process or engages in top-down data explanation, hypothesis driven or predictive processes (Kleinschmidt & Jaeger, in press; McMurray & Jongman, 2011, in press).

Two questions are crucial for disentangling these debates. First, at what level in the processing chain does visual speech play a role? Gestural models predict that speech is fundamentally perceived in terms of articulatory gestures (Fowler, 1986); such gestures are the distal objects of perception and not tied to any particular modality (and not limited to audiovisual interaction, e.g. Fowler & Dekle, 1991; Gick & Derrick, 2009). Consequently, non-auditory (e.g. visual, tactile) information that can help identify speech gestures should affect processing from the earliest stages¹. In contrast, models like the Fuzzy Logical Model of Perception (Oden & Massaro, 1978) and auditory accounts like that of Holt and Lotto (2010) treat auditory and visual signals as independent channels of information, rather than a single channel whose goal is the representation of a gesture. Auditory and visual inputs are analyzed in their respective, somewhat separate processing streams and integrated at later stages when feature, phoneme or lexical decisions must be made (Massaro & Cohen, 1983). Thus, understanding the level of processing at which visual information impacts speech perception may be helpful for disentangling these theoretical assumptions.

The second critical question is when in the time course of processing does visual information play a role? Recent predictive coding models of perception suggest that rather than passively categorizing the bottom-up signal, observers make active predictions about what they are likely to hear (and see), and that perception is based on the difference between these predictions and the bottom-up signal (Clark, 2013; Friston, 2005; Kumar et al., 2011; Rao & Ballard, 1999; see McMurray & Jongman, 2011 and Kleinschmidt & Jaeger, for applications to speech perception). Visual speech information could play a crucial role in such predictive processes (Arnal & Giraud, 2012; van Wassenhove, 2013) because in many cases, preparatory gestures (e.g., closing the lips before a word initial /b/, raising the tongue before a /d/) are visible before any acoustic signal is produced (Chandrasekaran, Trubanova,

¹These theories should be contrasted with recent studies which have asked whether the *motor system* is involved in speech perception more broadly (Meister et al., 2007; Venezia, Saberi, Chubb, & Hickok, 2012). In contrast, motor theory and gestural accounts argue that *auditory* organization for speech should reflect motor principles.

Stillittano, Caplier, & Ghazanfar, 2009; Schwartz & Savariaux, 2014). Thus, for the listener, the visual speech signal could set up predictions about what is about to be heard. A critical test of such models would be to determine if speech-relevant visual information is available and used prior to sound onset. This question of timing is not independent of the question of levels: If evidence of prediction is observed, it is important to determine at what level of processing the predictive information is integrated. This can help refine predictive coding models by specifying the level at which predictions are made.

These two questions are difficult to address using behavioral techniques alone. Cognitive neuroscience thus may offer a critical route to answering these questions by tagging the effects of visual speech information to known locations in the auditory processing pathway and examining this as a function of time. In this study, we took advantage of electrocorticography (ECoG) to begin to address these questions. We studied awake, behaving humans who were undergoing chronic intracranial monitoring as part of pre-surgical evaluation for treatment of medically intractable epilepsy. These neurosurgical patients were implanted with multi-contact depth electrodes and subdural grid arrays that allowed for simultaneous recordings from primary, non-primary auditory and frontal cortex. Like non-invasive magnetoencephalography (MEG) and scalp-recorded electroencephalography (EEG), ECoG has a very high temporal resolution; moreover, because the electrodes are placed directly on the surface or within the parenchyma of the brain, it also has tremendous spatial precision (albeit limited to areas of the brain covered by the electrode arrays, as determined by clinical considerations). In contrast to MRI, it offers similar spatial precision, but the much higher temporal resolution of electrophysiologic measures, and without the background noise created by the scanner. Thus, this direct invasive recording method was ideally suited for addressing these questions.

In the remainder of this introduction, we briefly review the cognitive neuroscience of AV speech integration. We then present an ECoG study exploring these issues in the context of an experiment in which we measure visual and auditory responses to speech and non-speech stimuli, with emphasis on activation prior to and soon after the onset of acoustic stimulation.

The cognitive neuroscience of AV speech integration

A large literature on multisensory integration in nonhuman animals has demonstrated robust multimodal response profiles even in primary sensory areas long considered “unimodal” (Cappe & Barone, 2005; Driver & Noesselt, 2008; Ghazanfar & Schroeder, 2006; Kayser, Petkov, & Logothetis, 2009). In humans, there has been ongoing debate as to whether or not primary auditory cortex, located in the posteromedial portion of Heschl's gyrus (HG), is driven by (as opposed to modulated by) visual-alone speech. Several functional magnetic resonance imaging (fMRI) studies have provided supporting evidence for activation of primary auditory cortex during silent speech or lipreading (e.g. Calvert et al., 1997; Pekkola et al., 2005), while others (Bernstein et al., 2002; Campbell et al., 2001) have failed to find robust primary auditory cortex activation. Along with increased spatial resolution allowed by modern fMRI (e.g., Pekkola et al., 2005), subject-specific anatomical and functional delineation of primary areas has also provided more direct support for the role of auditory cortex in the processing of (audio)visual speech stimuli (Okada, Venezia, Matchin, Saberi, &

Hickok, 2013). However, fundamental limitations of the method restrict the conclusions that can be drawn about the stage at which these interactions occur. Because the blood-oxygen level dependent (BOLD) contrast in fMRI relies on measurements of metabolic changes secondary to neural activity which evolves over the course of seconds, it is difficult to attribute “activity” in these paradigms to the actual neuronal processing of the multisensory stimulus (Bernstein & Liebenthal, 2014). Moreover, the relatively slow time course of the BOLD response makes it difficult to make inferences about the time course of processing.

Noninvasive electrophysiology (EEG, MEG) offers more detail about the time course of visual interaction with auditory information. Evoked responses occurring approximately 100–200 ms after audio onset have shown modulatory effects of visual speech information. For example, the auditory N_1 - P_2 complex elicited by isolated nonsense syllables (e.g., /pa/, /ta/, /ka/) has reduced amplitude for AV relative to audio-alone stimuli (Besle, Fort, Delpuech, & Giard, 2004; Pilling, 2009; van Wassenhove, Grant, & Poeppel, 2005) and for AV non-speech stimuli with anticipatory visual information (Vroomen & Stekelenburg, 2011, 2009). Because many of the evoked components that show modulation from visual stimuli are thought to have generators in the supratemporal plane (Godey, Schwartz, de Graaf, & Chauvel, 2001; C. Liégeois-Chauvel, Musolino, Badier, Marquis, & Chauvel, 1994; Näätänen & Picton, 1987), on which HG lies, these findings suggest that visual/gestural information can impact early cortical stages of auditory processing. Using MEG, Hertrich et al. (2009) found that dipoles localized to the supratemporal plane were differentially active for particular combinations of auditory-visual stimuli, offering support for multisensory interactions in or near primary auditory cortex. However, the relatively coarse spatial resolution of noninvasive electrophysiology methods limits direct inferences that can be made about the precise localization of these effects.

ECoG can overcome these problems by combining high temporal precision with increased spatial resolution. ECoG has shown a similar pattern of evoked response amplitude reduction to AV relative to audio-alone speech (Besle et al., 2008; Reale et al., 2007) in non-primary auditory areas including those on posterior portion of superior temporal gyrus (STG). These findings support the hypothesis that visual information affects auditory processing relatively early in the cortical processing stream. What remains unclear from these studies is whether visual speech activates auditory cortex at encoding stages or it plays a more modulatory role. That is, does visual information play a driving role at early processing stages because it carries its own information about the speech gesture (e.g., it could directly activate the relevant representations at that level)? Alternatively, visual-gestural organization could be available to modulate auditory-driven representations without directly activating the relevant representations in isolation. Furthermore, it is unclear whether the content of the visual signal is encoded with enough specificity to provide auditory cortical areas with the information necessary to allow predictions to be made about the content of the upcoming signal. That is, can visual speech drive neural correlates of a “percept” in auditory areas, or does it just modulate or interact with the neural activity mediating a primarily auditory-driven percept?

With respect to the issue of timing, the most consistent body of evidence is the aforementioned work showing that the amplitude and latency of the scalp-recorded auditory

N_1 - P_2 evoked potentials and their neuromagnetic counterparts is reduced in the presence of visual information (Arnal, Morillon, Kell, & Giraud, 2009; Besle et al., 2004; van Wassenhove et al., 2005; for a similar argument in the perception of one's own speech during speech motor control, see Houde, Nagarajan, Sekihara, & Merzenich, 2002; Kauramäki et al., 2010). Such results are often interpreted within the predictive coding framework. Visual information before the onset of auditory information serves as a constraint on what the auditory signal may be (Arnal & Giraud, 2012) creating less error between the prediction and the perceptual event and ultimately facilitating processing (van Wassenhove et al., 2005). This interpretation is supported by the fact that the degree of uniqueness in visual mouth shapes has been shown to be related to the reduction in the amplitude of the evoked response (Arnal et al., 2009; van Wassenhove et al., 2005): the more predictable the auditory stimulus is from the visual information, the more reduction is observed. This suggests some specificity to the modulation of the auditory evoked components by visual stimulus.

This approach suffers from two shortcomings. First, the auditory N_1 - P_2 necessarily occurs after the onset of sound. Although relatively early in the sequence of cortical auditory evoked responses, these components are typically hundreds of milliseconds after visual information may become available (Chandrasekaran et al., 2009; but see Schwartz & Savariaux, 2014, for discussion of natural temporal variability in connected speech). Thus, they can only offer indirect support for prediction. Second, by comparing AV to audio-alone conditions, the specificity of the visual information in early auditory areas cannot be directly tested. Here, a comparison between different auditory and/or visual stimuli (e.g., speech versus non-speech) may offer insight into the content of the information in these areas and at these times (Hertrich et al. 2009, 2011; Campbell et al., 2001). Such a comparison is one goal of the present study.

More unambiguous evidence as to the timing would come from evidence of activation of auditory cortex prior to the onset of the sound. Besle et al. (2008) reported evoked responses in the middle temporal gyrus (MTG) prior to acoustic onset, but these responses were not consistent across participants. Moreover, MTG in humans is not a primary auditory area, suggesting that this response – although early in time – may reflect the first stages of getting visual information *to* auditory cortex, which does not address the question of when visual/gestural information is available *in* the auditory cortex. Besle et al. (2008) also report one site on the supratemporal plane which appears to have visual-alone responsiveness, but without precise functional delineation it is unclear whether this response was in a primary or non-primary auditory area, and as with the MTG response, there was not consistent response across subjects. Using ECoG, Schepers et al. (Schepers, Yoshor, & Beauchamp, 2014) found larger high gamma responses in visual cortex to visual relative to audio-visual or audio-alone speech, suggesting that visual information is potentially available for early auditory processing, but the interaction within auditory structures remains unclear. Thus, while ECoG may have the requisite capability, available data does not yet offer the critical evidence that would disentangle competing theories of perception.

Current study

The current study sought to test whether visual/gestural information affects neural activation at early stages of auditory processing, and whether this activity reflects (in part) the content of the visual information (i.e., if it is differentially sensitive to meaningful mouth motion). The use of ECoG allowed us to study neural processing in brain areas that can be anatomically and functionally defined on a subject-by-subject basis with high spatial resolution.

Brain regions—We evaluated neural responses prior to and immediately following the onset of auditory stimuli (acoustic speech and non-speech) in cortical areas implicated in audiovisual speech processing. We defined these areas on a subject-by-subject basis based on the individual anatomical reconstructions (see Methods). ECoG measurement is limited somewhat by clinical concerns which dictate electrode coverage. Nonetheless, across five subjects we were able to accumulate measures in three areas that map closely to the theoretical goals of this investigation. Although audiovisual processing in humans engages a broad swath of cortical and sub-cortical areas (Musacchia, Sams, Nicol, & Kraus, 2006), we focused on three areas specifically engaged in speech processing with an eye toward levels of processing: HG (early auditory processing); STG (phonological and language processing); and precentral gyrus (or PreC, sensorimotor transformations and representations).

The posteromedial two-thirds of HG is the first cortical site that responds to auditory input, and responds both speech and non-speech sounds. The anterolateral portion of HG exhibits markedly different physiological response properties and has been interpreted as non-core (belt and/or parabelt) cortex (Brugge et al., 2008, 2009; Nourski, Steinschneider, McMurray, et al., 2014).

The lateral surface of the STG has been linked to phonological and/or word-form encoding (Boatman, 2004; Gow, Segawa, Ahlfors, & Lin, 2008; Hickok & Poeppel, 2007). It generally responds more robustly to speech than non-speech, and the pattern of activity across STG reflects the phonological content of the speech input with some specificity (Mesgarani, Cheung, Johnson, & Chang, 2014; Nourski, Steinschneider, Oya, et al., 2014; Steinschneider, 2011), though it is also involved in other complex auditory domains (e.g. Liégeois-Chauvel, Peretz, Babai, Laguitton, & Chauvel, 1998; Price, 2000). These findings underscore the fact that STG likely reflects higher level processes that operate on auditory inputs.

Finally, the *precentral gyrus* (PreC) is active in a number of AV speech functional neuroimaging studies (Matchin, Groulx, & Hickok, 2014; Skipper, van Wassenhove, Nusbaum, & Small, 2007). These areas are likely not solely visual processing areas, but have integrative function (Callan, Jones, & Callan, 2014). The timing and specificity of motor cortex involvement in audiovisual – and even auditory-alone – speech (Cogan et al., 2014; Meister, Wilson, Deblieck, Wu, & Iacoboni, 2007; Wilson, Saygin, Sereno, & Iacoboni, 2004) remains an open question (Matchin et al., 2014), and it is not currently clear whether such systems are recruited when speech is presented in non-adverse listening conditions.

Measures—We derived two measures from the local field potential (LFP) generated from electrodes in each of these three areas. First, we computed the power of cortical activity in the high gamma (70–150 Hz) frequency range, which is thought to reflect relatively localized activity of neural ensembles in human cortex (Crone et al., 2006; Steinschneider, Fishman & Arezzo, 2008). High gamma activity in STG has been shown to carry substantial information about the content of the speech signal (Mesgarani et al., 2014; Steinschneider et al., 2011 but see Nourski, Steinschneider et al., 2015). Moreover, activity in the high gamma band correlates with the BOLD fMRI response (Logothetis, Pauls, Augath, Trinath, & Oeltermann, 2001; Mukamel et al., 2005; Niessing et al., 2005), offering us the ability to compare our results with previous AV speech studies.

Second, we computed activity in the beta range (14–30 Hz), which has been hypothesized to represent the influence of long-range functional connections that reflect top-down processing across spatially separated brain regions. Beta activity has been extensively studied in the context of motor planning and imagery (Pfurtscheller & Neuper, 1997; Waldert et al., 2008) and more recently has been suggested to play a more general role in cognition (Engel & Fries, 2010) and in mediating perception in auditory cortex (Arnal & Giraud, 2012; Arnal, Wyart, & Giraud, 2011; Arnal, 2012).

Experiment—To examine the specificity of responses, we presented subjects with both speech and non-speech auditory and visual stimuli, and we presented speech in each modality in isolation. For the auditory stimulus, we used a naturally produced /da/ or a non-speech noise stimulus with a similar amplitude envelope to the /da/. For the visual stimulus, we used either visual /da/ or a non-speech closed-mouth facial gesture (gurning) with the same onset of motion as the /da/2.

Methods

Participants

Participants were five individuals (1 female; mean age: 40.8 years old; age range 32–50) undergoing two-week intracranial ECoG monitoring for the diagnosis and treatment of medically intractable epilepsy. Recordings from each participant were obtained in an electromagnetically shielded hospital room within the Epilepsy Monitoring Unit at the University of Iowa Hospitals and Clinics. Three individuals had electrode placement in the right hemisphere (subjects R212, R232, and R250) and two had electrode placement in the left hemisphere (subjects L206 and L222). All participants were right-handed with left hemisphere language dominance as determined by preimplantation intracarotid amobarbital (Wada) testing. Research protocols were approved by the University of Iowa Institutional Review Board, and participants signed informed consent documents prior to any research recordings. Research activities did not interfere with acquisition of clinical recordings, and participants could withdraw from research activities at any time without consequence for their clinical monitoring or treatment plan. Participants initially remained on their

²Given the constraints of gathering sufficient trials and maintaining vigilance in the patient population, a fully crossed design was not possible. Instead, our design was intended to target specific pairwise comparisons (e.g., visual speech versus non-speech) which captured our research questions in a minimal amount of experimental time and attentional demand for the research participants.

antiepileptic medications but were typically decreased in dosage during the monitoring period at the direction of their treating neurologists until enough seizure activity had been recorded for localization, at which time antiepileptic medications were resumed. No research session occurred within three hours following seizure activity.

Prior to electrode implantation, pure tone audiometry was conducted on each participant, and all had thresholds below 20 dB HL for frequencies between 500 and 4000 Hz. Vision was self-reported as normal or corrected to normal; participants who required glasses wore them during the task. All participants were native speakers of English and none had formal training in lipreading.

Video and audio materials

Stimuli were adapted from Reale et al. (2007). Videos consisted of the lower portion of the face of a native English-speaking female naturally articulating speech syllables or making non-speech closed-mouth facial gestures (gurning). Speech audio was extracted from original video recording. Non-speech (noise) audio was created by extracting the envelope of the naturally produced /da/ token via Hilbert transform and multiplying Gaussian noise by this envelope in MATLAB. In this way, the non-speech audio had the same onset, duration, and amplitude contour of the natural /da/, but lacked spectral structure of speech (Fig. 1). Audio and video signals were combined and saved in audio video interleave format (29.97 frames per second) in VirtualDub (www.virtualdub.org). Table 1 shows audio and video pairings for the six conditions.

Stimulus presentation

Stimulus delivery and randomization was controlled by Presentation software (Version 14.9, www.neurobs.com). Participants were awake and sitting upright in their hospital bed or in a chair in their hospital room at the time of testing. A 19-inch ViewSonic LCD monitor mounted on a mobile cart was positioned approximately 24 inches from the participant's face. Auditory stimuli were delivered through insert earphones (ER4B, Etymotic Research, Elk Grove Village, IL, United States of America) with custom earmolds for all participants except R212, for whom foam ear inserts were used instead.

Written instructions were presented on the monitor and also read aloud to participants prior to the start of the experiment. A brief practice session was included to ensure that participants understood the task. Stimuli were presented at a comfortable volume determined by each participant during the practice session. To maintain vigilance during the experiment, participants were instructed to press a button every time the syllable /tu/ was spoken (either audio-alone, audio-visually, or video-alone); distractor items comprised 20% of all trials. Behavioral responses were recorded on a Microsoft SideWinder gamepad via a button press using the index finger of the ipsilateral hand (relative to primary grid placement) to minimize motor-related activity resulting from the button press. A member of the research team remained in the room during testing to ensure the participants remained focused on the screen.

During intertrial intervals and audio-alone conditions, a light-colored background (matched to average luminance and color of the face stimuli) was presented on the screen. Stimuli

were randomized and presented 40 times each. Data from participant L206 includes only 20 trials per condition due to a technical error. The inter-trial interval was set pseudo-randomly on each trial with a range of 1500 to 2000 ms. During this time a black fixation cross appeared in the center of the screen. Experimental recording sessions lasted approximately 30 minutes.

Electrode placement, localization and recording

Prior to implantation, subjects underwent a whole-brain high-resolution T1-weighted structural MRI (resolution 0.78×0.78 mm, slice thickness 1.0 mm, 2 volumes for averaging) scan. Preimplantation MRIs and postimplantation thin-sliced volumetric computed tomography (CT) scans (resolution 0.51×0.51 mm, slice thickness 1.0 mm) were co-registered using a 3-dimensional rigid fusion algorithm (Analyze version 8.1 software, Mayo Clinic, MN, United States of America). Coordinates for each electrode contact obtained from postimplantation MRI volumes were transferred to preimplantation MRI volumes. Intraoperative photographs were also used to aid in reconstruction.

The recording sites reported here were part of a more extensive set of recording arrays and depth electrodes clinically indicated for identification of seizure foci. Electrode placement was determined solely on the basis of clinical requirements. Consequently, the placement and number of electrodes varied across subjects. Data from the lateral surface of the temporal, frontal, and parietal lobes were recorded from multicontact subdural grid electrodes (AdTech, Racine, WI). The recording arrays consisted of platinum–iridium disc electrodes (2.3 mm diameter) embedded in a silicon membrane. Temporal grids (used in L206, R212, and L222) were arranged in an 8×12 grid with 5 mm interelectrode spacing, yielding a 3.5×5.5 cm array of 96 contacts. Fronto-parietal grids (L206, R212, L222, R232, R250) were arranged in a 4×8 grid with 1 cm interelectrode spacing. Two participants had a depth electrode placed along HG. A clinical depth electrode with 1 cm interelectrode spacing was implanted in L206. Participant R212 was implanted with a hybrid depth electrode containing 14 microwire contacts (2 mm spacing) and four macrocontacts (1 cm spacing); recordings from microwire contacts were not analyzed for the current study. Details of the surgical electrode implantation protocol can be found in Nourski & Howard (2015). A subgaleal contact was used as a reference for all recordings. Electrode sites that demonstrated epileptogenic activity or which were found to have LFP voltage deflections exceeding 3 standard deviations from the within-channel average in more than 20% of trials were excluded from analysis.

ECoG signals were acquired continuously during the experimental session using a TDT RZ2 processor (Tucker-Davis Technologies, Alachua, FL). The ECoG data were amplified, filtered (0.7–800 Hz bandpass, 12 dB/octave rolloff), digitized at a sampling rate of 2034.5 Hz, and stored for subsequent offline analysis. Continuous data were filtered to remove 60 Hz line noise and its harmonics, downsampled to 1000 Hz for computational efficiency and segmented offline into 4 s epochs, from -1 s to 3 s relative to the onset of each trial. Trials with any voltage deflection greater than three standard deviations from the mean were excluded.

Analysis

Time-frequency analysis was implemented using two methods. For visualization, we used multitaper spectral analysis (Thomson, 1982) with three tapers (125 ms windows, 10 ms overlap; time bandwidth: 1.5). For quantification and statistical analysis, event related band power (ERBP) in the beta (14 to 30 Hz) and high gamma band (70 to 150 Hz) was calculated relative to a 600 ms pre-stimulus baseline (900 to 300 ms prior to the onset of the trial to avoid contamination from visual onsets in conditions containing visual stimuli; analytic signal obtained with 100-order zero-delay finite impulse response filter). This was averaged within each of two time windows to compute our dependent variables. The first window was 300 ms long and captured the onset of the visual motion prior to the sound (734 to 1034 ms post stimulus onset). The second window (1168 to 1568 ms post trial onset) captured the onset of auditory stimuli (in conditions in which an audio signal was present).

Site selection—Sites not located on the areas of interest (HG, STG, PreC) were excluded from analysis. Sites on HG were subdivided into core (posteromedial) and non-core (anterolateral) areas based on the morphology and latency of evoked responses and frequency following responses to click train stimuli (Brugge et al., 2008; Nourski et al., 2014). While separate statistical analysis of core and non-core areas of HG was not possible due to the small number of recording sites, we do present an exploratory analysis comparing these two subdivisions.

Analytic strategy—Responses were analyzed using linear mixed effects models using the using LME4 (version 1.1–6; Bates, Maechler, & Bolker, 2013), package in R (version 3.2.0, R Core Team, 2015). We used this approach rather than an individual site-by-site analyses or area-wide ANOVAs for several reasons (see also Nourski et al., 2014). First, site-by-site analyses offer no way to account for the fact that the sites only come from a small number of participants (instead, each site is treated as an independent measure [analogous to an independent subject]). Second, this approach is more consistent with a hypothesis-testing approach with respect to the functional anatomy. Finally, these concerns could have been addressed with an area-wide ANOVA by including subject as a random factor (averaging now across site), but would ignore the fact that data were randomly sampled from multiple sites (within subject). Thus, in order to better control for correlations of repeated measurements obtained from sampling from a large number of recording sites within a small number of subjects, we adopted a mixed effects modeling framework which can simultaneously account for random sampling across recording sites and subject.

The mixed model approach explicitly accounts for this nested structure of the random effects variance by treating site as a random effect that is nested within the random subject effect. A critical difference between mixed models and GLM approaches (like regression and ANOVA) is that mixed models do not estimate effects for individual subjects or sites independently of each other. Rather, they estimate the properties of sites or subjects as a part of a statistical distribution of these effects – that is, they can use all of the data to estimate a subject's or site's contribution. This approach also allows flexibility in including unequal numbers of recording sites for each subject for each area.

The details of each model will be provided in the results section when their results are discussed. However, all of the models shared a basic structure. Each model used sum-coded fixed effects. Additionally, average ERBP in a 300 ms window prior to Window 1 was entered as a covariate (e.g., it did not interact with any of the fixed effects) to account for trial-by-trial variation in overall ERBP. It was significant (and positively correlated) in all of the models and is reported in Table 3 and Table 4 as “ W_0 Value” but will not be discussed further. All models used the Satterthwaite approximation (Satterthwaite, 1946) implemented in the lmerTest (version 2.0-25, Kuznetsova, Brockhoff, & Christensen, 2015) package to compute p -values and degrees of freedom.

Prior to examining the fixed effects, we tested various random effect structures to determine the best fit for these data. Our primary criterion for model selection was to find the model that best fit the data without introducing unnecessary terms. However, given the sparsity of our datasets in some cases, it was possible to overfit the data (e.g., random slopes on subjects with only two subjects). This overfitting often appeared as extremely high correlations between estimates of random slopes ($r = -1$ or $r = +1$), and these models were excluded from consideration. As part of this model selection, we also attempted to find the best fitting (and valid) model across all areas of interest simultaneously, so that models would have equivalent power across brain areas. In one case (modeling beta band activity in HG), the more complex model would not converge; in this instance the simpler random effects structure was used. Specific random effects structures are described with each model.

Results

Visual examination of time-frequency plots showed broad distribution of low frequency power suppression throughout the analysis epoch corresponding to stimulus presentation in all conditions. Example time-frequency plots for the auditory, visual, and audiovisual /da/ are shown in Figure 2 for one participant (L206). High gamma activity was more temporally and spatially localized, occurring in auditory areas (HG, STG) near the onset of the auditory stimulus. Beta increases occurring around the time of auditory onset were common in areas outside of our areas of interest (anterior MTG), and will not be discussed further.

Selected sites

Sites included in subsequent analyses are depicted in Figure 3, and a count of the available recording sites for each area and for each measure is shown in Table 2. Given the moderate number of sites available for analysis on HG, our statistical analysis pooled all sites on HG; specific response profiles for sites in core versus non-core areas of HG will be summarized qualitatively after presentation of quantitative results.

Overview of results

Results are described in three sections. First we examined the pre-audio time window to determine if there was any evidence that 1) visual information elicited responses during this time and 2) whether speech and non-speech stimuli differentially affected responses prior to auditory signal onset. This was conducted across each of the three brain areas (HG, STG, PreC) for both beta and high gamma ERBP. Next we examined the post-audio time window

to investigate the same questions in conditions which contained both auditory and visual stimuli. Finally we conducted more descriptive analysis of the time course of beta and high gamma band power across brain areas and conditions.

Window 1 (Pre-Audio) analysis

For each recording site within our areas of interest, we computed ERBP in the 300 ms immediately following visual motion onset (but prior to auditory onset). We tested whether power in this window was different for visual motion versus no-motion and whether visual speech differed from non-speech (Fig. 4). Because this time window spans the portion of the trial prior to acoustic stimulation, we collapsed across conditions with identical video regardless of the auditory stimulus (which had not yet occurred at this time). This resulted in 40 trials for no-Motion (condition $A_{Sp}V_{\emptyset}$), 80 non-speech-motion trials (conditions $A_{Sp}V_{NS}$ and $A_{NS}V_{NS}$), and 120 speech-motion trials (conditions: $A_{\emptyset}V_{Sp}$, $A_{NS}V_{Sp}$ and $A_{Sp}V_{Sp}$) per subject (half that for L206, see Methods).

These three levels of stimulus were entered into the model as fixed effects using two orthogonal contrast variables. The first coded whether or not there was visual motion (no motion = -0.5 , speech = $+0.25$, non-speech = $+0.25$); the second examined speech versus non-speech visual motion (no motion = 0 ; speech = $+0.5$, non-speech = -0.5). We also added average power in the 100–400 ms window following trial onset as a covariate, as described above. Model selection led to a model with random intercepts of site nested within subject for all of the following models of the pre-audio time window. Complete results of all of the models are shown in Table 3. Figure 4 shows means and standard errors for each condition in each area for both high gamma and beta ERBP.

Heschl's gyrus—The first model examined high gamma activity in HG (Fig. 4, top left panel). This showed no significant main effect of Visual Motion ($p = 0.88$) or Visual Content ($p = 0.88$). For beta, we considered a decrease in power (beta suppression) as an indicator of responsivity. Beta band responses in HG showed effects of both Visual Motion ($B = -0.63$, $p = 0.04$) and Visual Content ($B = 0.37$, $p = 0.04$). Larger beta suppression was found for conditions containing video, and the non-speech video content resulted in the greatest suppression of activity (Fig. 4, bottom left panel).

STG—We next examined high gamma activity in STG (Fig. 4, top middle panel). There was a significant main effect of Visual Content ($B = 0.07$, $p = 0.005$) with more high gamma power for visual speech than non-speech. Although there was no main effect of Visual Motion ($p = 0.165$), this is likely because response to non-speech was not larger than baseline, and the contrast codes used in our models effectively pool both types of visual stimuli together. Thus STG seems to be responding only to visual speech motion in the time window preceding auditory stimulation. The model of beta suppression on STG found a significant effect of Visual Motion ($B = -0.36$, $p < 0.001$) with more suppression for visual motion than no motion (Figure 4, bottom middle panel). However, beta ERBP did not differ between visual /da/ and visual gurning ($p = 0.838$).

Precentral gyrus—Finally, we examined high gamma ERBP in PreC (Fig. 4, top right panel). This analysis found a significant main effect of Visual Motion ($B = 0.19$, $p = 0.001$), with greater power increase for visual motion compared to no-motion. The main effect of Speech was not significant ($p = 0.2$); but the numerical differences between average ERBP in this area suggest an increase for visual speech relative to non-speech motion.

Our analysis of beta suppression on PreC found a significant main effect of Motion ($B = -0.41$, $p = 0.002$) and a significant main effect of Visual Content ($B = -0.19$, $p = 0.02$). Greater beta suppression was observed for conditions containing motion, and beta power in the PreC was most suppressed for speech motion (Figure 4, bottom right panel).

Window 1 summary—Prior to acoustic onset, we found an effect of visual motion in primary auditory cortex (HG), with larger beta suppression when visual motion was present and less suppression for visual /da/ than for visual non-speech. However, this was not observed for gamma band activity. STG showed increased high gamma activity and more beta suppression for visual motion in general and also demonstrated speech specificity, with increased high gamma activity to visual speech than visual non-speech. We found robust evidence for differential responses in PreC with increased high gamma power and beta suppression associated with visual motion in general, with greatest power changes occurring when the visual stimulus was speech.

Our results suggest that even prior to acoustic onset, visual information leads to activation in primary auditory areas; however, there is little speech-specific activation (as reflected by high gamma power increase) in early auditory areas (medial HG). This is consistent with relatively late encoding of visual speech information in the auditory hierarchy, but the pattern of activity observed in beta suppression in HG for visual motion in general suggests visual information may play some role in processing in primary auditory areas.

Window 2 (Post-Audio)

We next examined activity after auditory onset (from 1168 to 1568 ms post-onset; Fig. 5). This corresponds to the portion of the trial in which both an auditory stimulus and visual motion were present. Audio- and video-alone conditions were not included in Window 2 analysis because overall activity levels may be different for multimodal compared with unimodal stimuli. We tested main effects of speech versus non-speech stimuli in auditory and visual modalities. These factors were coded using contrast codes (speech = +0.5, non-speech = -0.5). As with Window 1 analysis, average ERBP in a 300 ms window following trial onset (-634 to -334 prior to Window 1) was also entered as a covariate. Prior to examining fixed effects, we compared several models with different random effect structures to determine the best fit for these data using similar criteria as above. For high gamma power, the best fitting model included random intercepts for subject, and random slopes of the main effects of Audio-Content and Video-Content (but not their interaction) on Site (nested within Subject). For models examining suppression of power in the beta band, the best fitting model for STG and PreC included random intercepts for subject and random slopes of Video-Content on Site (nested within subject), but no slope for Audio-Content. However, this model resulted in overfitting in HG; thus, we used a simpler model (random

intercept on Site nested within Subject) for this band and area only. Figure 5 shows means and standard errors of the mean for each condition and each area. Table 4 contains full model results for this time window.

Heschl's gyrus—We first examined high gamma ERBP in HG. This showed no significant main effects or interactions (see Table 4 and top left panel of Fig. 5). The next model examined beta band activity. As with the pre-audio analysis, we considered lower absolute values (indicating larger suppression relative to baseline) an indication of more responsivity. Beta activity in HG showed a significant main effect of Video-Content ($B = -0.38$; $p = 0.025$) with more beta suppression for visual /da/ conditions than visual gurning. No main effect of Audio-Content and no interaction were found (Fig 5, bottom left panel).

STG—We next examined high gamma ERBP in STG. This model found a significant main effect of Audio-Content ($B = 1.04$; $p < 0.001$), with much greater power for /da/ than non-speech noise. There was also a significant effect of Video-Content ($B = 0.25$; $p < 0.001$), though as Figure 5 (top middle column) shows, it is likely subsumed by the significant interaction ($B = -0.28$, $p < 0.001$). There was greater high gamma power for visual /da/ than visual gurning, but this was reduced when the auditory stimulus was speech.

Models of beta suppression on STG resulted in significant main effects of Audio-Content and Video-Content. Suppression was significantly larger for audio non-speech ($B = 0.35$, $p < 0.001$) than audio speech. In contrast, video speech showed significantly more beta suppression than video non-speech ($B = -0.20$, $p < 0.001$). We also found a significant interaction of Audio-Content and Video-Content ($B = 0.2$, $p = 0.02$). As in the results of the high gamma models, the differences between visual /da/ and visual gurning were largest when the auditory stimulus was non-speech on STG.

Precentral gyrus—The next model examined high gamma ERBP in PreC (Fig. 5, top right panel). This model found a significant main effect of Video-Content ($B = 0.35$, $p = 0.004$), with increased high gamma power for video speech compared to video non-speech. There was no significant effect of Audio-Content ($p = 0.591$), however, the model did show a significant interaction ($B = -0.15$, $p = 0.046$). Pairwise post hoc tests showed no effect of Audio-Content when the visual stimulus was gurning ($p = 0.52$), and an increase for audio non-speech over speech when the video stimulus was /da/ ($B = -11$, $p = 0.03$).

In the PreC, the model found a significant main effect of Video-Content ($B = -0.86$, $p < 0.001$), with significantly more beta suppression observed for /da/ than for gurning (Fig 5, bottom right panel). No main effect of Audio-Content ($p = 0.11$) and no interaction were found ($p = 0.88$).

Window 2 summary—Following audio onset, we find no difference in HG high gamma activity between speech syllables and non-speech analogues. This was observed in both the auditory or visual modality, suggesting a relatively uniform response to all of our conditions. However, beta activity was significantly reduced when the visual stimulus was speech. In non-primary auditory cortex (STG), we find an increase in high gamma power and larger beta suppression for both visual speech and audio speech. In PreC, visual speech also

resulted in increased high gamma power, but there was no influence of audio content. Thus, looking specifically at high gamma, it would appear that HG responds to auditory input non-specifically (similar for speech and non-speech) and is not directly activated by visual information (as reflected in high gamma power), but does show sensitivity to visual content in lower frequency bands. PreC also appears to be processing stimuli in a fairly modality-specific manner, responding selectively to visual speech and not showing substantial differences between auditory input types except a minor increase in high gamma power for auditory non-speech paired with visual /da/. Finally, STG demonstrates sensitivity to the speech versus non-speech content in both modalities.

Time course of cortical responses

While the foregoing analyses document the overall differences among cortical areas, we also conducted a more descriptive analysis of the fine-grained time course of these effects. These were conducted by averaging across sites and subjects within anatomically defined regions, which included posteromedial and anterolateral portions of Heschl's gyrus, STG and PreC.

Effect of audio stimulus—Figure 6A shows the time course of high gamma (top row) and beta (bottom row) power as a function of the auditory stimulus. Visual stimulus was held constant (/da/). As shown statistically, the general morphology and peak high gamma power in HG was similar for both conditions containing an auditory stimulus (solid lines), but no response was observed for visual-alone speech (dashed line). Deviation from baseline in the high gamma band in medial HG appeared to occur only after the auditory stimulus was presented (second shaded window, corresponding to Window 2). In contrast, sites on STG showed differences in high gamma power across all three audio types, with auditory speech having greater high gamma response than auditory non-speech, which was in turn larger than the no-audio condition. Finally, PreC showed similar pattern of response for all types of auditory stimuli in both beta and high gamma bands, highlighting the relative 'indifference' of this area to auditory content.

Effect of visual stimulus—Figure 6B shows the time course of high gamma (top row) and beta (bottom row) power as a function of the visual stimulus, with the auditory stimulus held constant (/da/). Again, medial HG showed almost no difference in response between conditions. However, in lateral HG we observed some beta suppression during Window 1 in the presence of visual motion (of either speech or non-speech type); whereas in the audio-alone condition this was not observed until Window 2. While this must be tempered by the fact that only seven sites (from two participants) contributed to this average, it raises the possibility that beta suppression in lateral HG induced by visual motion could reflect a preparatory signal to auditory cortex (though this does not appear to be specific to speech).

In STG, we find early (Window 1) activity when the visual signal is speech, but we did not see high gamma increases for visual non-speech or audio-alone conditions in this window. Beta responses in STG show earlier and greater suppression for speech versus non-speech visual stimuli. The effect of visual stimulus on responsiveness in sites on the PreC was evident in both high gamma and beta envelope changes in both time windows, and was highly dependent on visual stimulus type. Responses to stimuli without visual information were

considerably delayed, and power changes in both bands are greater for speech than non-speech.

Discussion

This study shows differential effects of auditory and visual speech and non-speech signals on responses recorded from three cortical areas that are known to be important for speech processing in either or both modalities. Critically, because we compared speech to non-speech analogues these differences can be attributed to the content of the visual signal, not just the presence of input. The varied pattern of results across brain regions highlights the complexity of the speech processing system and the distributed nature of multisensory processing across multiple cortical areas. Our central questions were when in time, and what stage in the processing stream visual information plays a role. To address these questions, we first address a key limitation of our study, we then summarize the key findings from each area and finally discuss which type of models our results are compatible with.

Limitations

Although we have shown differences between the neural responses to auditory and visual speech versus non-speech analogues in each modality, we cannot make strong conclusions regarding whether these differences are “speech-specific.” Our non-speech stimuli were designed to differ in key qualities of speech while retaining some properties that were critical to our question of *when* interactions were occurring. In the auditory modality we removed spectral structure while holding timing cues constant. In the visual modality, we used a motion which is not associated with the production of particular speech features while still being biologically plausible human mouth motion. Because our stimulus sets were limited to these few types, it is possible that differences we observe could be driven by spectral richness or low-level visual dynamics. However, the complex interaction of auditory and visual content (Fig. 5, Table 4) hints that physical differences between stimuli are not exclusively driving these effects.

Heschl's gyrus

Data obtained from HG in two subjects showed no systematic difference in high gamma power between our contrasts of interest (i.e., no difference for visual speech versus non-speech) at either Window 1 or Window 2. Any stimulus containing an auditory signal resulted in increased high gamma band power relative to baseline, regardless of whether the content of that signal was speech or non-speech. Similarly, visual-alone speech did not elicit robust high gamma activity in either core (medial) or non-core (lateral) sites on HG. Thus, it does not appear that primary auditory areas (defined on anatomical and functional grounds) are encoding visual/gestural speech information in an auditory-like way.

However, our analysis of the beta band in HG showed extended (Window 1 and 2) suppression for conditions with visual information, with some specificity to the content of the visual information (speech /da/ versus non-speech gurning). Furthermore, our descriptive analyses (see Fig. 6B) suggest that this sensitivity is particularly evident in the anterolateral (non-core) aspect of HG. This area could reflect the transmission of (visual) information to

primary auditory cortex, which could in turn affect sensory processing which occurs after the onset of acoustic stimulation. These results do not offer strong support for models of speech processing in which visual speech information is encoded in core auditory cortex. However, our results hint that visual information may play a modulatory role at this stage. This is consistent with prior literature showing that the visual stimulus can affect auditory responses localized to the supratemporal plane that occur within the first 200 ms following auditory onset (e.g. van Wassenhove et al., 2005; Hertrich et al., 2009; Arnal et al., 2011).

STG

Non-core auditory cortex on STG showed effects of both auditory and visual stimulus content in both time windows. Prior to the onset of acoustic stimulation, STG was responsive to only to meaningful visual speech information, but not to non-speech facial motion (for high gamma band power increase) or a static non-face visual signal (for either high gamma increase or beta suppression). Differences in high gamma power between visual conditions following sound onset were not as large, except in the case of non-speech auditory stimuli where visual speech showed a robust effect. This is consistent with behavioral and fMRI studies showing that visual influence is often most pronounced when the auditory signal is degraded (Callan et al., 2003; McGettigan et al., 2012; Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007; Sumby & Pollack, 1954).

Our findings can be summarized by the idea that the increase in power in STG follows the increase in meaningfulness. When the signal is not meaningful in either modality (auditory noise paired with visual gurning) there is very little high gamma power. If the auditory signal remains non-speech, but meaningful visual speech information is present (auditory noise paired with visual /da/), high gamma power increases considerably. When the auditory stimulus is in itself meaningful (auditory /da/ with either visual gurning or visual /da/), the effect of visual stimulus is reduced considerably, possibly reflecting a ceiling effect. This suggests that non-primary areas, although sensitive to visual information, have the most to gain when the visual signal offers information that is not available in the auditory channel. However, all together, these results suggest that STG serves both to integrate auditory and visual/gestural information, and perhaps to code visual/gestural information before the onset of the auditory stimulus.

Precentral gyrus

Activity recorded from PreC was characterized by strong visual effects. Both before and after the auditory stimulus, there was significantly more activation for visual speech than non-speech and little effect of the auditory signal. This pattern of data suggests that the PreC is sensitive to the visual content. Based on fMRI studies of audiovisual speech processing (e.g. Skipper et al., 2007), it was not surprising to find event-related power changes in motor areas of cortex (specifically, PreC). However, the timing of these responses (see Fig. 6 and Fig. 7) suggests that activity in this area is not secondary to auditory processing, but may reflect an important early stage of multisensory speech processing. In all conditions containing the talker's face, precentral sites show activity as in the pre-audio window studied here. During facial motion but prior to acoustic onset, high gamma activity in PreC increases for both speech and non-speech visual stimuli, with speech motion having highest overall

power (Fig 4). In contrast with STG, precentral sites remain sensitive to visual content after sound onset, but show no specificity for auditory content. Beta activity in this region parallels the high gamma response, with increased suppression for speech relative to non-speech visual stimuli, and no differential effect of auditory stimulus. This activity in the PreC may be reflecting an independent channel of primarily visual analysis.

Implications for theories of AV integration

Taken together, these results do not support a model in which speech is encoded as an inherently multimodal signal at the earliest stages of the auditory cortical processing stream. We did not find either visual-alone high gamma power increases (activation) or visual content specificity in core HG; in summary, HG appears to be – perhaps unsurprisingly – specialized for processing the acoustic aspects of our stimuli.

Conversely, we did not find strong auditory effects in primary motor cortex (PreC). Instead, only non-primary auditory areas on the superior temporal gyrus were sensitive to both factors, with meaningful visual speech content showing distinct advantage (high gamma increase and beta suppression). This is consistent with an integration model in which visual and auditory information are transduced independently and combined at higher levels of processing (reflected here in STG activity, and possibly beta suppression activity in anterolateral HG).

While these clearly favor models in which visual and auditory information are transduced independently and integrated later, our results do not directly support any particular late integration model such as FLMP (Massaro & Cohen 1988). This is because we cannot determine the representations (gestural or non-gestural) that may be relevant at each stage of processing, and it is possible that gestures are computed at a late point, even if each mode is processed independently. In addition, our restricted stimulus set (isolated speech syllable and auditory and visual non-speech analogues) is likely too simple to make strong claims about how levels of processing interact in real-world communication.

The early onset of effects in PreC and the content-specific responses in STG suggest that an important stage of audiovisual integrative processing is occurring in non-primary areas (Besle et al., 2008; Reale et al., 2007). Moreover, the fact that these areas are sensitive to visual speech information prior to the onset of the auditory stimulus appears to be consistent with some predictive role of visual information. However, predictive coding models (e.g., Rao & Ballard, 1999; see McMurray & Jongman, 2011, for an application to speech) typically posit that the content of these predictions is a sensory code (to aid in rapid comparison with the input). However, we find the strongest pre-audio effects in non-primary auditory areas (STG) and motor cortex (PreC), which are not likely to contain a detailed acoustic/sensory code that would be compared to the input in primary auditory cortex. Thus, strong predictive coding models of (audiovisual) speech encoding are not supported by these data. However, content-specific effects occurring in non-primary auditory and motor cortex – even prior to acoustic stimulation – suggest that these areas may play a role in “setting the stage” for rapid multisensory integration of meaningful visual motion with auditory input.

Acknowledgments

We thank Rick Reale for assistance with stimulus creation, and Haiming Chen, Phillip Gander and Rachel Gold for help with data collection. This work was supported by the NIH under grants DC04290, UL1RR024979, and DC008089; and by the Hoover Fund.

Funding acknowledgment: This work was supported by the NIH under grant DC04290, grant UL1RR024979, and grant DC008089; and by the Hoover Fund.

References

1. Arnal LH. Predicting “When” using the motor system’s beta-band oscillations. *Frontiers in Human Neuroscience*. 2012; 6 doi: 10.3389/fnhum.2012.00225.
2. Arnal LH, Giraud AL. Cortical oscillations and sensory predictions. *Trends in Cognitive Sciences*. 2012; 16(7):390–398. doi:10.1016/j.tics.2012.05.003. [PubMed: 22682813]
3. Arnal LH, Morillon B, Kell CA, Giraud AL. Dual neural routing of visual facilitation in speech processing. *The Journal of Neuroscience*. 2009; 29(43):13445–13453. doi: 10.1523/JNEUROSCI.3194-09.2009. [PubMed: 19864557]
4. Arnal LH, Wyart V, Giraud AL. Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. *Nature Neuroscience*. 2011; 14(6):797–801. doi:10.1038/nn.2810. [PubMed: 21552273]
5. Bates, D.; Maechler, M.; Bolker, B. lme4: Linear mixed-effects models using Eigen and Eigen. R Package Version 1.1-7. 2012. Retrieved from <http://lme4.r-forge.r-project.org/>
6. Bernstein LE, Auer ET, Takayanagi S. Auditory speech detection in noise enhanced by lipreading. *Speech Communication*. 2004; 44(1):5–18. doi:10.1016/j.specom.2004.10.011.
7. Bernstein LE, Auer ET Jr, Moore JK, Ponton CW, Don M, Singh M. Visual speech perception without primary auditory cortex activation. *Neuroreport*. 2002; 13(3):311–315. [PubMed: 11930129]
8. Bernstein LE, Liebenthal E. Neural pathways for visual speech perception. *Frontiers in Neuroscience*. 2014; 8 doi: 10.3389/fnins.2014.00386.
9. Besle J, Bertrand O, Giard MH. Electrophysiological (EEG, sEEG, MEG) evidence for multiple audiovisual interactions in the human auditory cortex. *Hearing Research*. 2009; 258(1):143–151. doi:10.1016/j.heares.2009.06.016. [PubMed: 19573583]
10. Besle J, Fischer C, Bidet-Caulet A, Lecaigard F, Bertrand O, Giard MH. Visual activation and audiovisual interactions in the auditory cortex during speech perception: intracranial recordings in humans. *The Journal of Neuroscience*. 2008; 28(52):14301–14310. doi: 10.1523/JNEUROSCI.2875-08.2008. [PubMed: 19109511]
11. Besle J, Fort A, Delpuech C, Giard MH. Bimodal speech: early suppressive visual effects in human auditory cortex. *European Journal of Neuroscience*. 2004; 20(8):2225–2234. doi: 10.1111/j.1460-9568.2004.03670.x. [PubMed: 15450102]
12. Boatman D. Cortical bases of speech perception: evidence from functional lesion studies. *Cognition*. 2004; 92(1):47–65. doi:10.1016/j.cognition.2003.09.010. [PubMed: 15037126]
13. Brugge JF, Nourski KV, Oya H, Reale RA, Kawasaki H, Steinschneider M, Howard MA. Coding of repetitive transients by auditory cortex on Heschl’s gyrus. *Journal of Neurophysiology*. 2009; 102(4):2358–2374. doi: 10.1152/jn.91346.2008. [PubMed: 19675285]
14. Brugge JF, Volkov IO, Oya H, Kawasaki H, Reale RA, Fenoy A, Steinschneider M, Howard MA. Functional localization of auditory cortical fields of human: click-train stimulation. *Hearing Research*. 2008; 238(1):12–24. doi:10.1016/j.heares.2007.11.012. [PubMed: 18207680]
15. Callan DE, Jones JA, Callan A. Multisensory and modality specific processing of visual speech in different regions of the premotor cortex. *Frontiers in Psychology*. 2014; 5:389. doi: 10.3389/fpsyg.2014.00389. [PubMed: 24860526]
16. Callan DE, Jones JA, Munhall K, Callan AM, Kroos C, Vatikiotis-Bateson E. Neural processes underlying perceptual enhancement by visual speech gestures. *Neuroreport*. 2003; 14(17):2213–2218. [PubMed: 14625450]

17. Calvert GA, Bullmore ET, Brammer MJ, Campbell R, Williams SC, McGuire PK, Woodruff PWR, Iversen SD, David AS. Activation of auditory cortex during silent lipreading. *Science*. 1997; 276(5312):593–596. doi: 10.1126/science.276.5312.593. [PubMed: 9110978]
18. Campbell R. The processing of audio-visual speech: empirical and neural bases. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 2008; 363(1493):1001–1010. doi: 10.1098/rstb.2007.2155.
19. Campbell R, MacSweeney M, Surguladze S, Calvert G, McGuire P, Suckling J, Brammer MJ, David AS. Cortical substrates for the perception of face actions: an fMRI study of the specificity of activation for seen speech and for meaningless lower-face acts (gurning). *Cognitive Brain Research*. 2001; 12(2):233–243. doi:10.1016/S0926-6410(01)00054-4. [PubMed: 11587893]
20. Cappe C, Barone P. Heteromodal connections supporting multisensory integration at low levels of cortical processing in the monkey. *European Journal of Neuroscience*. 2005; 22(11):2886–2902. doi: 10.1111/j.1460-9568.2005.04462.x. [PubMed: 16324124]
21. Chandrasekaran C, Trubanova A, Stillitano S, Caplier A, Ghazanfar AA. The natural statistics of audiovisual speech. *PLoS Computational Biology*. 2009; 5(7):e1000436. doi: 10.1371/journal.pcbi.1000436. [PubMed: 19609344]
22. Clark A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*. 2013; 36(03):181–204. doi: 10.1017/S0140525X12000477. [PubMed: 23663408]
23. Cogan GB, Thesen T, Carlson C, Doyle W, Devinsky O, Pesaran B. Sensory-motor transformations for speech occur bilaterally. *Nature*. 2014; 507(7490):94–98. doi:10.1038/nature12935. [PubMed: 24429520]
24. Driver J, Noesselt T. Multisensory interplay reveals crossmodal influences on 'sensory-specific' brain regions, neural responses, and judgments. *Neuron*. 2008; 57(1):11–23. doi:10.1016/j.neuron.2007.12.013. [PubMed: 18184561]
25. Engel AK, Fries P. Beta-band oscillations—signalling the status quo? *Current Opinion in Neurobiology*. 2010; 20(2):156–165. doi:10.1016/j.conb.2010.02.015. [PubMed: 20359884]
26. Fowler CA. Auditory perception is not special: We see the world, we feel the world, we hear the world. *The Journal of the Acoustical Society of America*. 1991; 89(6):2910–2915. doi: 10.1121/1.400729. [PubMed: 1918631]
27. Fowler CA. An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*. 1986; 14(1):3–28.
28. Fowler CA, Dekle DJ. Listening with eye and hand: cross-modal contributions to speech perception. *Journal of Experimental Psychology: Human Perception and Performance*. 1991; 17(3):816. doi: 10.1037/0096-1523.17.3.816. [PubMed: 1834793]
29. Friston K. A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological sciences*. 2005; 360(1456):815–836. doi: 10.1098/rstb.2005.1622.
30. Ghazanfar AA, Schroeder CE. Is neocortex essentially multisensory? *Trends in Cognitive Sciences*. 2006; 10(6):278–285. [PubMed: 16713325]
31. Gick B, Derrick D. Aero-tactile integration in speech perception. *Nature*. 2009; 462(7272):502–504. doi:10.1038/nature08572. [PubMed: 19940925]
32. Godey B, Schwartz D, De Graaf JB, Chauvel P, Liegeois-Chauvel C. Neuromagnetic source localization of auditory evoked fields and intracerebral evoked potentials: a comparison of data in the same patients. *Clinical Neurophysiology*. 2001; 112(10):1850–1859. doi:10.1016/S1388-2457(01)00636-8. [PubMed: 11595143]
33. Gow DW, Segawa JA, Ahlfors SP, Lin FH. Lexical influences on speech perception: a Granger causality analysis of MEG and EEG source estimates. *Neuroimage*. 2008; 43(3):614–623. doi: 10.1016/j.neuroimage.2008.07.027. [PubMed: 18703146]
34. Grant KW, Walden BE, Seitz PF. Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *The Journal of the Acoustical Society of America*. 1998; 103(5):2677–2690. doi: 10.1121/1.422788. [PubMed: 9604361]

35. Green KP, Kuhl PK. The role of visual information in the processing of place and manner features in speech perception. *Perception & Psychophysics*. 1989; 45(1):34–42. doi: 10.3758/BF03208030. [PubMed: 2913568]
36. Hertrich I, Dietrich S, Ackermann H. Cross-modal interactions during perception of audiovisual speech and nonspeech signals: an fMRI study. *Journal of Cognitive Neuroscience*. 2011; 23(1): 221–237. doi:10.1162/jocn.2010.21421. [PubMed: 20044895]
37. Hertrich I, Mathiak K, Lutzenberger W, Ackermann H. Time course of early audiovisual interactions during speech and nonspeech central auditory processing: a magnetoencephalography study. *Journal of Cognitive Neuroscience*. 2009; 21(2):259–274. doi:10.1162/jocn.2008.21019. [PubMed: 18510440]
38. Hickok G, Poeppel D. The cortical organization of speech processing. *Nature Reviews Neuroscience*. 2007; 8(5):393–402. doi:10.1038/nrn2113. [PubMed: 17431404]
39. Holt LL, Lotto AJ. Speech perception within an auditory cognitive science framework. *Current Directions in Psychological Science*. 2008; 17(1):42–46. doi: 10.1111/j.1467-8721.2008.00545.x. [PubMed: 19060961]
40. Houde JF, Nagarajan S, Sekihara K, Merzenich MM. Modulation of the auditory cortex during speech: an MEG study. *Journal of Cognitive Neuroscience*. 2002; 14(8):1125–1138. doi: 10.1162/089892902760807140. [PubMed: 12495520]
41. Kauramäki J, Jääskeläinen IP, Hari R, Möttönen R, Rauschecker JP, Sams M. Lipreading and covert speech production similarly modulate human auditory-cortex responses to pure tones. *The Journal of Neuroscience*. 2010; 30(4):1314–1321. doi: 10.1523/JNEUROSCI.1950-09.2010. [PubMed: 20107058]
42. Kayser C, Petkov CI, Logothetis NK. Multisensory interactions in primate auditory cortex: fMRI and electrophysiology. *Hearing Research*. 2009; 258(1):80–88. doi:10.1016/j.heares.2009.02.011. [PubMed: 19269312]
43. Kleinschmidt DF, Jaeger TF. Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*. 2015; 122(2):148. doi: 10.1037/a0038695. [PubMed: 25844873]
44. Kumar S, Sedley W, Nourski KV, Kawasaki H, Oya H, Patterson RD, Howard MA III, Friston KJ, Griffiths TD. Predictive coding and pitch processing in the auditory cortex. *Journal of Cognitive Neuroscience*. 2011; 23(10):3084–3094. doi:10.1162/jocn_a_00021. [PubMed: 21452943]
45. Kuznetsova, A.; Brockhoff, PB.; Christensen, HB. lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package). R Package Version 2.0-25. 2014. Retrieved from <https://cran.r-project.org/web/packages/lmerTest/index.html>
46. Liégeois-Chauvel C, Musolino A, Badier JM, Marquis P, Chauvel P. Evoked potentials recorded from the auditory cortex in man: evaluation and topography of the middle latency components. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*. 1994; 92(3): 204–214.
47. Liégeois-Chauvel C, Peretz I, Babai M, Laguitton V, Chauvel P. Contribution of different cortical areas in the temporal lobes to music processing. *Brain*. 1998; 121(10):1853–1867. doi: 10.1093/brain/121.10.1853. [PubMed: 9798742]
48. Logothetis NK, Pauls J, Augath M, Trinath T, Oeltermann A. Neurophysiological investigation of the basis of the fMRI signal. *Nature*. 2001; 412(6843):150–157. doi: 10.1038/35084005. [PubMed: 11449264]
49. MacDonald J, McGurk H. Visual influences on speech perception processes. *Perception & Psychophysics*. 1978; 24(3):253–257. doi: 10.3758/BF03206096. [PubMed: 704285]
50. MacLeod A, Summerfield Q. Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology*. 1987; 21(2):131–141. doi: 10.3109/03005368709077786. [PubMed: 3594015]
51. Massaro DW, Cohen MM. Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*. 1983; 9(5): 753. doi: 10.1037/0096-1523.9.5.753. [PubMed: 6227688]

52. Matchin W, Groulx K, Hickok G. Audiovisual speech integration does not rely on the motor system: evidence from articulatory suppression, the McGurk effect, and fMRI. *Journal of Cognitive Neuroscience*. 2014; 26(3):606–620. doi:10.1162/jocn_a_00515. [PubMed: 24236768]
53. McGettigan C, Faulkner A, Altarelli I, Obleser J, Baverstock H, Scott SK. Speech comprehension aided by multiple modalities: Behavioural and neural interactions. *Neuropsychologia*. 2012; 50(5): 762–776. doi:10.1016/j.neuropsychologia.2012.01.010. [PubMed: 22266262]
54. McGurk H, MacDonald J. Hearing lips and seeing voices. *Nature*. 1976; 264:746–748. doi: 10.1038/264746a0. [PubMed: 1012311]
55. McMurray B, Jongman A. What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*. 2011; 118(2):219. doi: 10.1037/a0022325. [PubMed: 21417542]
56. McMurray B, Jongman A. What comes after [f]? Prediction in speech is a product of expectation and signal. *Psychological Science*. in press.
57. Meister IG, Wilson SM, Deblieck C, Wu AD, Iacoboni M. The essential role of premotor cortex in speech perception. *Current Biology*. 2007; 17(19):1692–1696. doi:10.1016/j.cub.2007.08.064. [PubMed: 17900904]
58. Mesgarani N, Cheung C, Johnson K, Chang EF. Phonetic feature encoding in human superior temporal gyrus. *Science*. 2014; 343(6174):1006–1010. doi: 10.1126/science.1245994. [PubMed: 24482117]
59. Mukamel R, Gelbard H, Arieli A, Hasson U, Fried I, Malach R. Coupling between neuronal firing, field potentials, and fMRI in human auditory cortex. *Science*. 2005; 309(5736):951–954. doi: 10.1126/science.1110913. [PubMed: 16081741]
60. Musacchia G, Sams M, Nicol T, Kraus N. Seeing speech affects acoustic information processing in the human brainstem. *Experimental Brain Research*. 2006; 168(1–2):1–10. doi: 10.1007/s00221-005-0071-5. [PubMed: 16217645]
61. Näätänen R, Picton T. The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. *Psychophysiology*. 1987; 24(4):375–425. doi: 10.1111/j.1469-8986.1987.tb00311.x. [PubMed: 3615753]
62. Niessing J, Ebisch B, Schmidt KE, Niessing M, Singer W, Galuske RA. Hemodynamic signals correlate tightly with synchronized gamma oscillations. *Science*. 2005; 309(5736):948–951. doi: 10.1126/science.1110948. [PubMed: 16081740]
63. Nourski KV, Howard MA III. Invasive recordings in the human auditory cortex. *Handbook of Clinical Neurology*. 2015; 129:225–244. doi: 10.1016/B978-0-444-62630-1.00013-5. [PubMed: 25726272]
64. Nourski KV, Steinschneider M, McMurray B, Kovach CK, Oya H, Kawasaki H, Howard MA. Functional organization of human auditory cortex: Investigation of response latencies through direct recordings. *Neuroimage*. 2014; 101:598–609. doi:10.1016/j.neuroimage.2014.07.004. [PubMed: 25019680]
65. Nourski KV, Steinschneider M, Oya H, Kawasaki H, Jones RD, Howard M. a. Spectral organization of the human lateral superior temporal gyrus revealed by intracranial recordings. *Cerebral Cortex*. 2014; 24:340–352. doi: 10.1093/cercor/bhs314. [PubMed: 23048019]
66. Oden GC, Massaro DW. Integration of featural information in speech perception. *Psychological Review*. 1978; 85(3):172. doi: 10.1037/0033-295X.85.3.172. [PubMed: 663005]
67. Ohala JJ. Speech perception is hearing sounds, not tongues. *The Journal of the Acoustical Society of America*. 1996; 99(3):1718–1725. doi: 10.1121/1.414696. [PubMed: 8819861]
68. Okada K, Venezia JH, Matchin W, Saberi K, Hickok G. An fMRI study of audiovisual speech perception reveals multisensory interactions in auditory cortex. *PloS one*. 2013; 8(6):68–959. doi:10.1371/journal.pone.0068959.
69. Pekkola J, Ojanen V, Autti T, Jääskeläinen IP, Möttönen R, Tarkiainen A, Sams M. Primary auditory cortex activation by visual speech: an fMRI study at 3 T. *Neuroreport*. 2005; 16(2):125–128. [PubMed: 15671860]
70. Pfurtscheller G, Neuper C. Motor imagery activates primary sensorimotor area in humans. *Neuroscience Letters*. 1997; 239(2):65–68. doi:10.1016/S0304-3940(97)00889-6. [PubMed: 9469657]

71. Pilling M. Auditory event-related potentials (ERPs) in audiovisual speech perception. *Journal of Speech, Language, and Hearing Research*. 2009; 52(4):1073–1081. doi: 10.1044/1092-4388(2009/07-0276).
72. R Core Team. R: A language and environment for statistical computing. version 3.2.0. R Foundation for Statistical Computing; Vienna, Austria: 2015. <http://www.R-project.org/>
73. Rao RP, Ballard DH. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*. 1999; 2(1):79–87. doi:10.1038/4580. [PubMed: 10195184]
74. Reale RA, Calvert GA, Thesen T, Jenison RL, Kawasaki H, Oya H, Howard MA, Brugge JF. Auditory-visual processing represented in the human superior temporal gyrus. *Neuroscience*. 2007; 145(1):162–184. doi: 10.1016/j.neuroscience.2006.11.036. [PubMed: 17241747]
75. Ross LA, Saint-Amour D, Leavitt VM, Javitt DC, Foxe JJ. Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex*. 2007; 17(5):1147–1153. doi: 10.1093/cercor/bhl024. [PubMed: 16785256]
76. Satterthwaite FE. An approximate distribution of estimates of variance components. *Biometrics Bulletin*. 1946:110–114. doi: 10.2307/3002019. [PubMed: 20287815]
77. Schepers IM, Yoshor D, Beauchamp MS. Electrooculography Reveals Enhanced Visual Cortex Responses to Visual Speech. *Cerebral Cortex*. in press. doi: 10.1093/cercor/bhu127.
78. Schwartz JL, Savariaux C. No, there is no 150 ms lead of visual speech on auditory speech, but a range of audiovisual asynchronies varying from small audio lead to large audio lag. *PLoS Computational Biology*. 2014; 10(7):e1003743. doi: 10.1371/journal.pcbi.1003743. [PubMed: 25079216]
79. Skipper JJ, van Wassenhove V, Nusbaum HC, Small SL. Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex*. 2007; 17(10):2387–2399. doi: 10.1093/cercor/bhl147. [PubMed: 17218482]
80. Steinschneider M. Unlocking the role of the superior temporal gyrus for speech sound categorization. *Journal of Neurophysiology*. 2011; 105(6):2631–2633. doi: 10.1152/jn.00238.2011. [PubMed: 21430279]
81. Steinschneider M, Fishman YI, Arezzo JC. Spectrotemporal analysis of evoked and induced electroencephalographic responses in primary auditory cortex (A1) of the awake monkey. *Cerebral Cortex*. 2008; 18(3):610–625. doi: 10.1093/cercor/bhm094. [PubMed: 17586604]
82. Steinschneider M, Nourski KV, Kawasaki H, Oya H, Brugge JF, Howard MA. Intracranial study of speech-elicited activity on the human posterolateral superior temporal gyrus. *Cerebral Cortex*. 2011; 21(10):2332–2347. doi: 10.1093/cercor/bhr014. [PubMed: 21368087]
83. Nourski KV, Steinschneider M, Rhone AE, Oya H, Kawasaki H, Howard MA, McMurray B. Sound identification in human auditory cortex: Differential contribution of local field potentials and high gamma power as revealed by direct intracranial recordings. *Brain and Language*. 2015; 148:37–50. doi: 10.1016/j.bandl.2015.03.003. [PubMed: 25819402]
84. Sumby WH, Pollack I. Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*. 1954; 26(1):212–215. doi: 10.1121/1.1907309.
85. Thomson DJ. Spectrum estimation and harmonic analysis. *Proceedings of the IEEE*. 1982; 70(9): 1055–1096. doi: 10.1109/PROC.1982.12433.
86. van Wassenhove V. Speech through ears and eyes: interfacing the senses with the supramodal brain. *Frontiers in Psychology*. 2013; 4 doi: 10.3389/fpsyg.2013.00388.
87. van Wassenhove V, Grant KW, Poeppel D. Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America*. 2005; 102(4):1181–1186. doi: 10.1073/pnas.0408949102. [PubMed: 15647358]
88. Venezia JH, Saberi K, Chubb C, Hickok G. Response bias modulates the speech motor system during syllable discrimination. *Frontiers in Psychology*. 2012; 3 doi: 10.3389/fpsyg.2012.00157.
89. Vroomen J, Stekelenburg JJ. Perception of intersensory synchrony in audiovisual speech: Not that special. *Cognition*. 2011; 118(1):75–83. doi:10.1016/j.cognition.2010.10.002. [PubMed: 21035795]

90. Vroomen J, Stekelenburg JJ. Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. *Journal of Cognitive Neuroscience*. 2010; 22(7):1583–1596. doi: 10.1162/jocn.2009.21308. [PubMed: 19583474]
91. Waldert S, Preissl H, Demandt E, Braun C, Birbaumer N, Aertsen A, Mehring C. Hand movement direction decoded from MEG and EEG. *The Journal of Neuroscience*. 2008; 28(4):1000–1008. doi: 10.1523/JNEUROSCI.5171-07.2008. [PubMed: 18216207]
92. Wilson SM, Saygin AP, Sereno MI, Iacoboni M. Listening to speech activates motor areas involved in speech production. *Nature Neuroscience*. 2004; 7(7):701–702. doi:10.1038/nn1263. [PubMed: 15184903]

Figure 1.

Stimulus detail and trial timing. a: Temporal (top) and spectral (bottom) detail of speech syllable /da/ (left) and non-speech noise (right) auditory stimuli. The amplitude envelope of both stimuli is similar, but spectral richness is absent in the non-speech (noise) stimulus. b: Sample frames from speech (top) and non-speech (bottom) video stimuli. Both stimuli begin with neutral mouth-closed still frames; visual motion begins at the same frame for both conditions. c: Combined audio and video trial detail. Video stimuli contained motion prior to the onset of auditory stimulation corresponding with the natural lag between facial motion and vocal production of /da/. d: quantification of lip aperture/spread differences between /da/ and gurning visual stimuli. /da/ contains more up-down spread, while gurning motion is largely side- to side.

Figure 2.

Representative data from one subject (L206). a: Location of implanted electrodes on the lateral surface (left panel) and in the supratemporal plane (right panel). b: ERBP plot for a representative site (location marked by asterisk in panel a) depicting plotting conventions (axes and scales) for panels c-e. Stimulus schematic is shown on top. c-e: Responses from grids on the lateral surface and depth electrode to audio-alone (c), video-alone (d), and audiovisual /da/ (e). HG: Heschl's gyrus.

Figure 3.

Sites included in analysis for each area of interest. MedHG: medial Heschl's gyrus; LatHG: lateral Heschl's gyrus; STG: superior temporal gyrus; PreC: precentral gyrus.

Figure 4.

Pre-Audio responses. Mean high gamma (top) and beta (bottom) ERBP for each area of interest in the time window preceding auditory onset. Due to the low number of recording sites on Heschl's gyrus, one model was fit for all sites. Error bars indicate standard error of the mean. V₀: Audio-alone, V_{NS}: Visual non-speech (gurning), V_{Sp}: Visual speech (/da/).

Figure 5.

Post-Audio responses. Mean high gamma (top) and beta (bottom) event related band power for each area of interest in the time window following auditory onset. Error bars reflect standard error of the mean. A_{NS} = Audio non-speech (noise), A_{Sp} = Audio speech (/da/); see Figures 3 and 4 for additional abbreviations.

Figure 6.

Time course comparison across conditions of interest. a: Effect of auditory stimulus. Envelope of high gamma (top) and beta (bottom) for each region (columns), contrasting auditory stimulus types (visual stimulus /da/ for all plots). b: Effect of visual stimulus. Envelope of high gamma (top) and beta (bottom) contrasting visual stimulus types (auditory stimulus /da/ for all plots). Shaded bars indicate windows used in statistical analysis. See Figure 3 – Figure 5 for abbreviations. Waveforms were smoothed for display only.

Table 1

Stimulus pairs. Six audiovisual stimuli were created by combining auditory speech, non-speech (noise), or no-audio with visual speech, non-speech (gurning), or no-video.

Condition	Audio	Video
$A_{sp}V_{sp}$	/da/	/da/
$A_{sp}V_{ns}$	/da/	gurning
$A_{ns}V_{sp}$	noise	/da/
$A_{ns}V_{ns}$	noise	gurning
$A_{sp}V_{\emptyset}$	/da/	None
$A_{\emptyset}V_{sp}$	none	/da/

Table 2

Number of subjects and recording sites that contributed data to each studied brain area.

	Subjects	Sites
Heschl's gyrus	2	7
Medial	2	3
Lateral	2	4
STG	3	94
Precentral gyrus	5	39

Table 3

Results of linear mixed effects models on data from the Pre-Audio window. W_0 value refers to average ERBP in 300 ms following trial onset which was entered as a covariate (see Method); p -values greater than 0.2 are not shown.

			<i>B</i>	SE	<i>t</i>	df	<i>p</i>
Heschl's gyrus	High gamma	Video-Content	0.02	0.11	0.15	1159	
		Video-Motion	0.03	0.18	0.15	1160	
		W_0 value	0.48	0.03	17.87	1156	<0.001
	Beta	Video-Content	0.37	0.18	2.09	1153	0.037
		Video-Motion	-0.63	0.30	-2.12	1153	0.035
		W_0 value	0.37	0.03	13.01	1158	<0.001
STG	High gamma	Video-Content	0.07	0.02	2.84	19590	0.005
		Video-Motion	-0.06	0.04	-1.39	19590	0.165
		W_0 value	0.39	0.01	57.16	19700	<0.001
	Beta	Video-Content	-0.01	0.04	-0.21	19590	
		Video-Motion	-0.36	0.07	-5.04	19590	<0.001
		W_0 value	0.38	0.01	57.14	19700	<0.001
Precentral gyrus	High gamma	Video-Content	0.12	0.04	1.28	7722	0.2
		Video-Motion	0.19	0.01	3.27	7723	0.001
		W_0 value	0.42	0.01	33.35	5149	<0.001
	Beta	Video-Content	-0.19	0.08	-2.33	7718	0.02
		Video-Motion	-0.41	0.13	-3.04	7719	0.002
		W_0 value	0.51	0.01	47.73	7758	<0.001

Table 4

Results of linear mixed effects models on data from the Post-Audio window. W_0 value refers to average ERBP in 300 ms following trial onset which was entered as a covariate (see Method); p -values greater than 0.2 are not shown. AudCont = Audio-Content; VidCont = Video-Content

		<i>B</i>	<i>SE</i>	<i>t</i>	<i>df</i>	<i>p</i>	
Heschl's gyrus	High gamma	Audio-Content	0.69	0.38	1.81	6	0.122
		Video-Content	0.01	0.17	0.04	7	
		AudCont × VidCont	−0.14	0.23	−0.60	766	
		<i>W</i> ₀ value	0.46	0.03	13.62	767	<0.001
	Beta	Audio-Content	0.01	0.17	0.05	772	
		Video-Content	−0.38	0.17	−2.25	772	0.025
		AudCont × VidCont	0.37	0.34	1.08	772	
		<i>W</i> ₀ value	0.35	0.03	11.47	772	<0.001
STG	High gamma	Audio-Content	1.04	0.11	9.13	93	<0.001
		Video-Content	0.25	0.05	5.15	86	<0.001
		AudCont × VidCont	−0.28	0.05	−5.76	12890	<0.001
		<i>W</i> ₀ value	0.37	0.01	49.47	12950	<0.001
	Beta	Audio-Content	0.35	0.04	8.05	12990	<0.001
		Video-Content	−0.20	0.05	−3.71	84	<0.001
		AudCont × VidCont	0.20	0.09	2.33	12990	0.02
		<i>W</i> ₀ value	0.37	0.01	49.17	13100	<0.001
Precentral gyrus	High gamma	Audio-Content	−0.03	0.06	−0.54	34	
		Video-Content	0.35	0.11	3.10	35	0.004
		AudCont × VidCont	−0.15	0.08	−1.99	5079	0.047
		<i>W</i> ₀ value	0.42	0.01	33.35	5149	<0.001
	Beta	Audio-Content	−0.13	0.08	−1.60	5118	0.111
		Video-Content	−0.86	0.19	−4.53	33	<0.001
		AudCont × VidCont	−0.03	0.17	−0.15	5117	
		<i>W</i> ₀ value	0.44	0.01	36.32	5145	<0.001