

Hearing Lips and Seeing Voices: How Cortical Areas Supporting Speech Production Mediate Audiovisual Speech Perception

Jeremy I. Skipper^{1,2}, Virginie van Wassenhove³, Howard C. Nusbaum² and Steven L. Small^{1,2}

¹Departments of Neurology and ²Psychology, and the Brain Research Imaging Center, The University of Chicago, Chicago, IL 60637, USA and ³Division of Biology, California Institute of Technology, Pasadena, CA, USA

Observing a speaker's mouth profoundly influences speech perception. For example, listeners perceive an "illusory" "ta" when the video of a face producing /ka/ is dubbed onto an audio /pa/. Here, we show how cortical areas supporting speech production mediate this illusory percept and audiovisual (AV) speech perception more generally. Specifically, cortical activity during AV speech perception occurs in many of the same areas that are active during speech production. We find that different perceptions of the same syllable and the perception of different syllables are associated with different distributions of activity in frontal motor areas involved in speech production. Activity patterns in these frontal motor areas resulting from the illusory "ta" percept are more similar to the activity patterns evoked by AV_{/ta/} than they are to patterns evoked by AV_{/pa/} or AV_{/ka/}. In contrast to the activity in frontal motor areas, stimulus-evoked activity for the illusory "ta" in auditory and somatosensory areas and visual areas initially resembles activity evoked by AV_{/pa/} and AV_{/ka/}, respectively. Ultimately, though, activity in these regions comes to resemble activity evoked by AV_{/ta/}. Together, these results suggest that AV speech elicits in the listener a motor plan for the production of the phoneme that the speaker might have been attempting to produce, and that feedback in the form of efference copy from the motor system ultimately influences the phonetic interpretation.

Keywords: audiovisual speech perception, efference copy, McGurk effect, mirror system, motor system, prediction

The Relationship between Audiovisual Speech Perception and Production

Observable mouth movements profoundly influence speech perception. The McGurk–MacDonald effect is a striking demonstration of this influence: When participants are presented with audiovisual (AV) speech stimuli, they report hearing a phoneme that is neither what they saw nor what they heard but rather a "fusion" of the auditory and visual modalities (McGurk and MacDonald 1976). For example, participants report hearing "ta" when a sound track containing the syllable /pa/ is dubbed onto a video track of a mouth producing /ka/. Another such effect, "visual capture," occurs when listeners *hear* the visually presented syllable (i.e., /ka/ in the prior example).

Other remarkable findings demonstrate the extent to which normal visual cues can affect speech perception. Adding visible facial movements to speech enhances speech recognition comparable with removing up to 20 dB of noise from the auditory signal (Sumbly and Pollack 1954). Multisensory enhancements in intelligibility of degraded auditory speech are anywhere from 2 to 6 times greater than would be expected for the comprehension of words or sentences in the auditory or visual modalities when presented alone (Risberg and Lubker

1978; Grant and Greenberg 2001). Importantly, such effects are not limited to unnatural or degraded stimulus conditions: Visual speech contributes to understanding clear but hard to comprehend speech or speech spoken with an accent (Reisberg et al. 1987).

How do observable mouth movements influence speech perception? Research on mirror neurons in the macaque and a putative mirror system in humans (see Rizzolatti and Craighero 2004 for a review) led us, like others, to propose that observable mouth movements elicit a motor plan in the listener that would be used by the listener to produce the observed movement (Skipper et al. 2005, 2006). Mirror neurons are a small subset of neurons, originally found in the macaque premotor area F5, that fire both during the production of goal-directed actions but also during the observation of similar actions. Similar "mirroring" functionality has been ascribed to the human motor system (Rizzolatti and Craighero 2004).

Indeed, both behavioral and neurophysiological evidence support the notion that the human mirror system and, therefore, the motor system, play a critical role in speech perception when mouth movements are observed. Behaviorally, listeners' perception of the McGurk–MacDonald effect is altered by viewing mouth movements produced by others or by oneself in a mirror (Sams et al. 2005). Similarly, speech production performance is changed or enhanced when producing a syllable and viewing someone saying that syllable compared with when that person is saying a different syllable (Kerzel and Bekkering 2000; Gentilucci and Cattaneo 2005). Neurophysiologically, activation (Campbell et al. 2001; Nishitani and Hari 2002; Olson et al. 2002; Callan, Jones, et al. 2003; Calvert and Campbell 2003; Paulesu et al. 2003; Buccino et al. 2004; Watkins and Paus 2004; Pekkola et al. 2006) and transcranial magnetic stimulation (Sundara et al. 2001; Watkins et al. 2003; Watkins and Paus 2004) of the motor system during the observation of mouth movements have been used to argue for a role of the mirror or motor system in AV speech perception.

Using functional magnetic resonance imaging (fMRI), we have previously shown that AV speech perception activates a network of motor areas including the cerebellum and cortical motor areas involved in planning and executing speech production and areas subserving proprioception related to speech production (Skipper et al. 2005). We also showed that it is primarily the visual aspects of observable mouth movements rather than the auditory content of speech that is responsible for this motor system activity. Auditory speech alone evoked far less activity in the motor system than AV speech, which is typical in speech perception studies that involve no explicit motor responses on the part of the listener (compare Zatorre et al. 1996; Belin et al. 2000, 2002; Burton et al. 2000; Zatorre

and Belin 2001). Furthermore, activity in the motor system was modulated by the visual salience of AV speech: Although an increase in visually distinguishable phonemes in AV stories was associated with a concomitant increase in motor system activity, no such effect occurred during auditory story listening without visual input (Skipper et al. 2005). Together these findings support the notion that it is the visual information about the phonetic content of the stories and not simply the lexical or acoustic properties (or corresponding motor or gestural code for those acoustic properties) per se that induces motor system activity (but see Wilson et al. 2004; Wilson and Iacoboni 2006).

Consistent with other behavioral and neurophysiological research, we concluded that AV speech perception is mediated in part by cortical activity resulting from prior experience producing speech (Stevens and Halle 1967; Liberman and Mattingly 1985). That is, recognition of AV speech involves invoking motor commands that could have generated the observed mouth movements.

Motor Mechanisms of AV Speech Perception

Positing that mirror neurons, the mirror system, or the motor system associated with speech production is involved in AV speech perception is not, however, explanatory. *Recognizing* a mouth movement (in the sense of activity in mirror neurons) and *interpreting* the significance of that movement with respect to speech perception are qualitatively different. Similar neural activity during production and perception of the same behavior certainly demonstrates a correspondence that could reflect matching or recognition of previously produced actions but it does not indicate how the motor system mediates the effect of observable mouth movements on speech perception.

To explain the mechanism by which motor system activity is *understood* by the brain as relevant to speech perception, we have developed a theoretical model of AV speech perception (Skipper et al. 2005; van Wassenhove et al. 2005; Skipper et al. 2006) based on the “hypothesize-and-test” or “analysis-by-synthesis” approach (Stevens and Halle 1967; see Discussion

for further elaboration). Specifically, this model proposes that early multisensory speech representations, derived from sound patterns and observed facial movements, can be thought of as multisensory *hypotheses* (i.e., “unconscious inferences” see Helmholtz 1867; Hatfield 2002)—but not final interpretations—about the phonemes produced by a speaker. These hypotheses, extracted from sensory information, are mapped onto motor commands used in speech production. These motor commands are associated with hypotheses based on past experience producing speech. Activated motor commands then predict the acoustic and somatosensory consequences of executing a speech movement through efference copy (von Holst and Mittelstaedt 1950; sometimes referred to as “corollary discharge,” Sperry 1950 or “forward models” in the motor control literature, Jordan and Rumelhart 1992; Miall 2003; Callan et al. 2004). These internally generated sensory consequences can constrain phonetic interpretation of the incoming sensory information by, for example, lending weight to a particular interpretation of a stretch of utterance corresponding to the sensory representation of the phoneme statistically associated with the production of the hypothesized phoneme.

We have specified the cortical areas that support the mechanism described by this model (Fig. 1; see Skipper et al. 2006 for a review of imaging data associated with these areas). These regions are visual areas, primary auditory cortex (A1), posterior superior temporal (STp) areas, supramarginal gyrus (SMG), somatosensory cortices (SI/SII), ventral premotor (PMv) cortex, and the pars opercularis (POp). The sequence of processing of observable mouth movements begins with a multisensory representation corresponding to a hypothesis in multisensory STp areas (visual area → STp ← A1). This hypothesis is specified in terms of the motor goal of that movement (STp → POp). The motor goal of the movement is mapped to the motor commands that could generate the observed movement in a somatotopically organized manner, in this case the mouth area of PMv cortex (POp → PMv ← M1). These motor commands yield a prediction of both the auditory (PMv → STp) and somatosensory (PMv → SI/SII → SMG → STp) consequences of those commands had they been

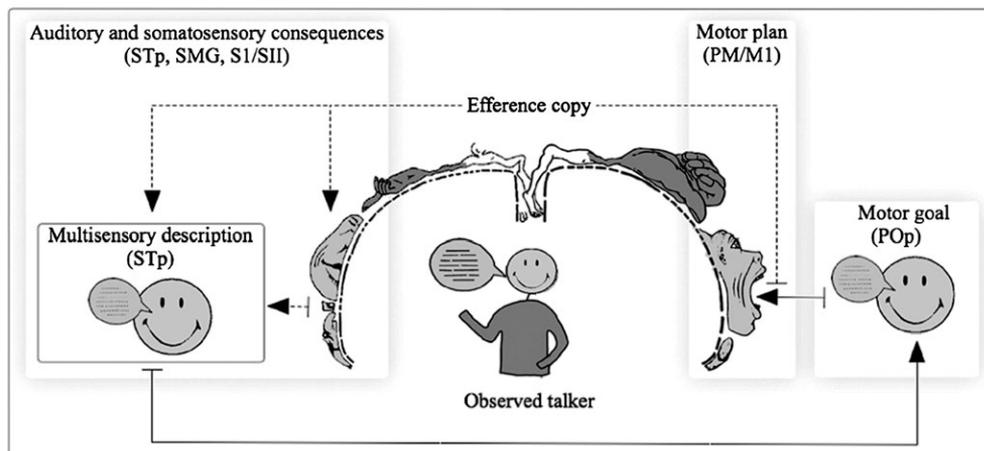


Figure 1. Neurally specified model of AV speech perception as presented in the text. A multisensory description in the form of a hypothesis about the observed talker’s mouth movements and speech sounds (in STp areas) results in the specification (solid lines) of the motor goals of that hypothesis (in the POp the suggested human homologue of macaque area F5 where mirror neurons have been found). These motor goals are mapped to a motor plan that can be used to reach that goal (in PMv and primary motor cortices [M1]). This results in the prediction through efference copy (dashed lines) of the auditory and somatosensory states associated with executing those motor commands. Auditory (in STp areas) and somatosensory (in the SMG and primary and secondary somatosensory cortices [SI/SII]) predictions are compared with the current description of the sensory state of the listener. The result is an improvement in speech perception in AV contexts due to a reduction in ambiguity of the intended message of the observed talker.

produced. These predictions can be used to constrain speech processing by supporting a particular interpretation or hypothesis (STp).

Using event-related fMRI, we tested specific aspects of this model. First, we looked for evidence that observing mouth movements and producing those mouth movements are associated with similar patterns of motor activity. Specifically, neural activity during production of a syllable was expected to be similar to that generated when observing this syllable in an AV condition, or in a silent visual-alone (V) condition. However, neural activity during production of a syllable was expected to be less similar to activity evoked by an audio-only (A) syllable. This would be suggestive of a shared underlying mechanism for production and observation of speech that is based on the presence of observable mouth movements and the ability of the motor system to predict the acoustic and somatosensory consequences of the observed information (van Wassenhove et al. 2005).

Second, if the motor system plays a role in determining perception as proposed by the above model, we expect that different patterns of activity in the motor system evoked by AV stimuli would correspond to different perceptual experiences of those stimuli. For example, the exact same AV stimulus perceived as “ka” or “ta” would be expected to be associated with different patterns of activity in the motor system. Similarly, different AV stimuli perceived as “ka” and “ta” would also be expected to be associated with different patterns of activity in the motor system.

Third, we looked for a particular pattern of neural activity in the motor system that would constitute evidence for the hypothesis-and-test or analysis-by-synthesis model outlined above. To do this we capitalized on the unique property of the McGurk-MacDonald effect, in which sensory aspects of the stimulus do not correspond to participants’ perceptual experience. On our account, motor system activity is proposed to be an early hypothesis about the identity of sensory patterns. If this is the case, patterns of activity during AV speech perception in frontal motor areas would correspond to the participants’ perceptual experience and *not* the physical stimuli transduced by sensory receptors. Therefore, from an early stage of processing, the motor activity for the stimulus that elicits the McGurk-MacDonald effect (i.e., the stimulus that results in an illusory “ta” percept as described above) would more resemble the AV stimulus corresponding to participants’ perception of that stimulus (i.e., /ta/) than the stimuli corresponding to the sensory information that was actually presented (i.e., /pa/ or /ka/; Table 1A).

Fourth, we looked for evidence that the hypothesis about the phonetic identity of a stimulus, reflected in frontal motor system activity, results in a prediction of the sensory consequences of producing those movements and influences sensory cortices through efference copy. Again, using the McGurk-MacDonald effect, we looked for evidence that early activity in sensory areas initially corresponds to a pattern of activity that is consistent with the sensory properties of the stimulus (i.e., /pa/ or /ka/). However, if efference copy is involved in determining perception, subsequent patterns of activity in these sensory regions should come to correspond to a pattern of activity consistent with the motor hypothesis (i.e., /ta/; Table 1B).

The details of the specific analyses that were performed to address these questions are elucidated in greater detail in the Materials and Methods and Results sections below.

Materials and Methods

Participants

Participants were 21 right-handed native speakers of American English with normal hearing and vision and no history of neurological or psychological disturbance. Handedness was determined by the Edinburgh handedness inventory (Oldfield 1971). Participants gave written consent, and the Institutional Review Board of The University of Chicago approved the study.

Task and Stimuli

All tasks and stimuli are described in Table 2. Participants passively listened to and/or watched speech stimuli during 3 separate runs of AV, V, or A stimuli. These runs were presented in a randomized and counterbalanced manner across participants. No explicit motor response was required and no supplementary task (e.g., discrimination, identification, etc.) was performed during this portion of the experiment. AV stimuli were AV_{/pa/}, AV_{/ka/}, and AV_{/ta/} spoken by a female actress filmed from the neck up. The actress made no noticeable facial movements besides those used in articulation. In addition, participants watched and listened to a stimulus designed to elicit the McGurk-MacDonald effect. This stimulus was composed of an audio /pa/ (A_{/pa/}) dubbed onto the video of a face saying /ka/ (V_{/ka/}), henceforth denoted as A_pV_k. Visual-alone stimuli were V_{/pa/}, V_{/ka/}, and V_{/ta/} and were created by removing the audio tracks from the AV stimuli. Audio-alone stimuli were A_{/pa/}, A_{/ka/}, and A_{/ta/} and were created by removing the video tracks from the AV stimuli.

Table 1

Predicted patterns of cortical activity in for the “hypothesis-and-test” or “analysis-by-synthesis” model (described in the Introduction and Fig. 1) in (A) motor areas associated with speech production, and (B) sensory areas

(A)	Time course of activation	Motor areas associated with speech production	
		Acoustic and/or visual	Perceptual/phonetic experience
	Early	NO	YES
	Late	NO	YES
(B)	Time course of activation	Sensory areas	
		Acoustic and/or visual	Perceptual/phonetic experience
	Early	YES	NO
	Late	NO	YES

Note: “YES” indicates that neural activity in an area for an AV stimulus that elicits the McGurk-MacDonald effect resembles activity associated with the pattern of activity elicited by either the “Acoustic and/or visual” properties of that stimulus or the “Perceptual/phonetic experience” of that stimulus. “NO” indicates that neural activity in an area for the stimulus that elicits the McGurk-MacDonald effect does not resemble the pattern of activity associated with either the “Acoustic and/or visual” properties of that stimulus or the “Perceptual/phonetic experience” of that stimulus. “Early” and “Late” refer to the temporal occurrence of activation patterns, as determined by fMRI.

Table 2

Experimental design: scan number, conditions, stimuli, and tasks

Scan #	Condition	Stimuli	Task
Random	AV	A _p V _k , AV _{/pa/} , AV _{/ka/} , and AV _{/ta/}	Watch and listen to speaker
Random	A	A _{/pa/} , A _{/ka/} , and A _{/ta/}	Listen to speaker
Random	V	V _{/pa/} , V _{/ka/} , and V _{/ta/}	Watch speaker
4	AV	A _p V _k , AV _{/pa/} , AV _{/ka/} , and AV _{/ta/}	Watch and listen to speaker
Frequency judgment	aAV	aA _p V _k , aAV _{/pa/} , aAV _{/ka/} , and aAV _{/ta/}	Watch and listen to speaker: 3AFC
6	Speaking	Written “pa,” “ka,” and “ta”	Say “pa,” “ka,” or “ta”

Following the AV, V, and A runs, there was a second AV run corresponding to the fourth functional imaging run. After completion of the second AV run, the experimenter entered the room and placed a 3-button mouse under the participant's right hand. The participant then completed 2 behavioral tasks intended to measure perception of A_pV_k . Prior to placement of the mouse, participants were not aware that they would be making behavioral responses. Participants indicated with a button press which stimulus they heard most frequently: Consistent perception of A_pV_k as "pa," "ka," or "ta" would result in hearing that stimulus twice as often as the others. Following the frequency judgment was an active AV run (aAV) corresponding to the fifth functional imaging run. During the aAV run participants responded "pa," "ka," or "ta" using 3-button alternative forced choice (AFC) to provide a continuous measure of perception of aA_pV_k and recognition of congruent $aAV_{/pa/}$, $aAV_{/ka/}$, and $aAV_{/ta/}$ in the scanning environment (see van Wassenhove et al. 2005 for behavioral results in the AV, A, and V conditions out of the scanning environment).

In a sixth and final run, the participants were asked to produce speech. Participants saw "pa," "ka," or "ta" printed on the screen for 1.5 s. Participants then repeated each syllable at a conversational rate for 12 s until they saw the word "STOP" on the screen. Participants were asked to speak at a normal volume (i.e., to avoid trying to compete with the noise of the scanner), to not open their mouths more than they might in a quiet conversation, and to avoid excessive movement.

Stimuli were always 1.5 s in length. Audio was delivered to the participants at a sound pressure level of 85 decibels (db-SPL) through headphones containing MRI-compatible electromechanical transducers (Resonance 300 Technologies, Inc., Northridge, CA). Participants viewed video stimuli through a mirror attached to the head coil that allowed them to see a screen at the end of the scanning bed. Speech productions were recorded through a microphone attached to the participant's headphones. Participants were monitored with a video camera.

Imaging and Data Analyses

Imaging was performed with blood oxygen level-dependent fMRI at 3 Tesla (GE Medical Systems, Milwaukee, WI) using spiral acquisition (Noll et al. 1995). A volumetric T_1 -weighted inversion recovery spoiled grass sequence was used to provide anatomical images on which landmarks could be found and functional activation maps could be superimposed. The 2 AV runs were each 7 min in length (280 whole-brain images each; a whole-brain image was collected every 1.5 s in this and all conditions) and contained a combined total of 272 stimuli: 68 each of A_pV_k , $AV_{/pa/}$, $AV_{/ka/}$, and $AV_{/ta/}$. The V and A runs were also 7 min in length (280 whole-brain images in each condition) and contained a total of 135 stimuli: 45 each of $/pa/$, $/ka/$, and $/ta/$. The aAV run was 12 min in length (480 whole-brain images) and contained a total of 136 stimuli: 34 each of aA_pV_k , $aAV_{/pa/}$, $aAV_{/ka/}$, and $aAV_{/ta/}$. In each of these runs, stimuli were presented in a randomized event-related manner with a variable interstimulus interval (ISI; mean ISI = ~3 s; minimum ISI = 0 s; maximum ISI = 12 s). The speech production run was 7 min in length (244 whole brain images) and contained 15 randomized blocks of speaking—5, 12-s blocks during which participants spoke "pa," "ka," or "ta."

Images were spatially registered in 3D space by Fourier transformation of each of the time points and corrected for head movement, using the AFNI software package (Cox 1996). Resulting data were analyzed using deconvolution/regression. A deconvolution analysis was used to generate impulse response functions (IRFs) of the fMRI signal on a voxel-wise basis (Glover 1999). There were separate regressors for each time point of A_pV_k , $AV_{/pa/}$, $AV_{/ka/}$, $AV_{/ta/}$, $A_{/pa/}$, $A_{/ka/}$, $A_{/ta/}$, $V_{/pa/}$, $V_{/ka/}$, and $V_{/ta/}$. For the aAV run, regressors were aA_pV_k when it was classified as "ta," aA_pV_k when it was classified as "ka" (see below), $aAV_{/pa/}$, $aAV_{/ka/}$, $aAV_{/ta/}$, and an "other" category in which aA_pV_k was classified as "pa" or $aAV_{/pa/}$, $aAV_{/ka/}$, or $aAV_{/ta/}$ were classified as something other than "pa," "ka," or "ta," respectively. In addition, the "other" category contained a randomly chosen subset of aA_pV_k trials when classified as "ta" or "ka" so that the 2 regressors (i.e., aA_pV_k when classified as "ta" and aA_pV_k when classified as "ka") each contained an equal number of trials. These were then fit using a linear least squares model to each time point of the hemodynamic response. Each stimulus had 12 such regressors, one for each time repetition (TR) between 0 and 11 (i.e., 0–18 s). This analysis resulted in an estimate of the hemodynamic response for each

stimulus relative to a baseline state without a priori assumptions about the IRF. Also incorporated were 3 regressors of interest ("pa," "ka," and "ta") for the speaking task. These regressors were waveforms with similarity to the hemodynamic response, generated by convolving a gamma-variant function with the onset time and duration of the blocks of interest. Finally, the model included one regressor each for the mean signal, linear, and quadratic trends, and 6 motion parameters for each of the 6 runs. Resulting regression coefficients were mean normalized.

Next, cortical surfaces were inflated (Fischl, Sereno, and Dale 1999), registered to a template of average curvature (Fischl, Sereno, Tootell, et al. 1999), and using SUMA (Saad et al. 2004) and AFNI packages (Cox 1996), surface-based analyses of variance (ANOVAs) were performed. ANOVA dependent variables were normalized regression coefficients from the deconvolution/regression analysis. These coefficients were smoothed the equivalent of 4 mm on the surface (Chung et al. 2005). One ANOVA contained 3 factors: syllables from the AV runs (A_pV_k , $AV_{/pa/}$, $AV_{/ka/}$, and $AV_{/ta/}$) and TR following stimulus onset (0–11) as fixed factors and participant as a random factor. Similar ANOVAs were performed for the V and A runs with the exception that syllable contained only 3 levels (e.g., $A_{/pa/}$, $A_{/ka/}$, and $A_{/ta/}$). A 2-factor ANOVA for the speech production run contained the fixed factor syllable ("pa," "ka," and "ta") and participants as a random factor. Overlap analysis (i.e., intersection of active voxels) revealed where activation during AV (both excluding and including A_pV_k in the ANOVA), V, and A perception overlapped speech production of the same syllables. Overlap images were created using images each thresholded at $P < 0.05$ corrected using a false discovery rate (FDR) algorithm (Genovese et al. 2002) and logically conjoined. A paired t -test was conducted on the coefficients from the active run, comparing A_pV_k when classified as "ta" or "ka." Again, images were corrected for multiple comparisons using FDR (corrected to $P < 0.05$).

Additional analyses within each individual were performed within regions of interest (ROIs). ROIs were created with Freesurfer's automated parcellation (Fischl et al. 2004). Regions were then sub-parcellated using landmarks established by Caviness et al. (1996). The precentral gyrus and sulcus were subparcellated into dorsal and ventral sectors, and the superior temporal gyrus and sulcus was divided into 3 sectors (see Skipper et al. 2005 for anatomical definition of these regions). To test specific hypotheses we conducted both correlational and discriminant analysis (Haxby et al. 2001; see Results section for further explication as to why these 2 analyses were conducted and specific applications). To perform these analyses, we split the design matrix into odd and even trials and reran the deconvolution/regression analysis described earlier. Then, within each ROI, regression coefficients for each syllable for odd and even trials from significantly activated voxels across all AV syllables ($P < 0.05$ corrected using FDR) that were also active in the speech production condition ($P < 0.05$ corrected using FDR) were mean corrected by dividing each coefficient by the mean from the regression analysis. For a region to be considered for further analysis, it was required that at least 4 voxels be active within any given region and that more than half of the participants activate it.

First, pairwise correlations were performed on these voxels between A_pV_k and $AV_{/pa/}$, $AV_{/ka/}$, or $AV_{/ta/}$ in each ROI by collapsing over odd and even trials by averaging coefficients. For example, in a given ROI there may have been 60 voxels that were active when participants observed A_pV_k , $AV_{/pa/}$, $AV_{/ka/}$, or $AV_{/ta/}$ and that overlapped the production of "pa," "ka," or "ta." These 60 voxels were then correlated in a pairwise manner. That is, the correlation between each of the 60 voxels for A_pV_k and $AV_{/pa/}$, A_pV_k and $AV_{/ka/}$, A_pV_k and $AV_{/ta/}$ was calculated. Friedman tests were used to test for differences between the correlation between A_pV_k and $AV_{/pa/}$, A_pV_k and $AV_{/ka/}$, and A_pV_k and $AV_{/ta/}$ for each of the voxels in each ROI. A Friedman test is a nonparametric test used to compare observations repeated within the same participants and is an alternative to a repeated measures ANOVA. The Friedman test ranks each correlation. Tukey-type nonparametric multiple comparison post hoc tests (Nemenyi) were used to test for differences between resulting ranks between A_pV_k and $AV_{/pa/}$, A_pV_k and $AV_{/ka/}$, A_pV_k and $AV_{/ta/}$ within each ROI. This analysis resulted in a measure the strength of the correlation of the distribution of activity associated with A_pV_k with the other syllables in an ROI. We performed this analysis both collapsing

over time by averaging the coefficients associated with each point of the IRF in each ROI (e.g., Fig. 3A) and also on each individual time point of each IRF in each ROI (e.g., Fig. 3B–D).

Second, discriminant analysis was conducted on the odd and even trials (Haxby et al. 2001). Correlation coefficients were calculated within syllables (e.g., between even and odd AV_{/ta/}) and across syllables (e.g., between even AV_{/ta/} with odd AV_{/ka/}). Within-syllable correlation coefficients were then compared with each of the across-syllable correlation coefficients. If the within-syllable correlation was larger than that of the across-syllable correlation coefficient, the comparison was counted as a correct identification. If the within-syllable correlation was larger than all across-syllable correlation coefficients, it was identified as correct against all other syllables. A *t*-test was used to test whether the accuracy of identifying a syllable for the whole group exceeded chance for the pairwise comparisons (50%) and chance when corrected against all other syllables (25%).

Results

Behavioral Results

Analysis of participants' responses to the aA_pV_k stimulus during the 3AFC task during run 5 indicated that participants typically labeled A_pV_k as either "ta" or "ka." Therefore, K-means analysis was used to assign participants to 2 different groups based on their responses to aA_pV_k. This resulted in a "ta" (i.e., "fusion") group (*N* = 13) who responded "ta" when presented aA_pV_k, and a "ka" (i.e., "visual capture") group (*N* = 8) who responded "ka" when presented aA_pV_k or who responded "ka" or "ta" with near equal likelihood when presented aA_pV_k. Sixty-two percent of the "ta" group indicated that they heard "ta" most frequently during the frequency judgment and responded "ta" 83% of the time during the 3AFC when presented aA_pV_k. Sixty-three percent of the "ka" group indicated that they heard "ka" most frequently and responded "ka" 61.5% of the time when presented aA_pV_k during the 3AFC. Both groups responded "pa" less than 2% of the time when presented aA_pV_k during the 3AFC. All participants were accurate (>95% correct) in classifying aAV_{/pa/}, aAV_{/ka/}, and aAV_{/ta/}. There were no differences in participants' accuracy in classifying aAV_{/pa/}, aAV_{/ka/}, and aAV_{/ta/}.

Imaging Results

Unless otherwise noted, all analyses were conducted on the "ta" (i.e., "fusion") group of participants. Analyses focused on this group because participants' responses during the behavioral tasks and participants' activation patterns were relatively more homogeneous than those of the "ka" group. That is, the "ka" group was associated with higher variability in both the behavioral responses and activation patterns relative to the "ta" group.

Main Effects of Syllable and Contrasts of the McGurk-MacDonald Syllable with the Other AV Syllables

Above baseline activity for syllables for the AV, A, and V ANOVAs show activation of areas typically associated with both speech perception and speech production (orange and blue in Fig. 2). To investigate if passive viewing of the McGurk-MacDonald syllable elicited a different pattern than passive viewing of the congruent AV syllables, we contrasted A_pV_k with AV_{/pa/}, AV_{/ka/}, or AV_{/ta/} (Tables 3 and 4). When contrasted, A_pV_k was significantly more active than AV_{/pa/} and AV_{/ka/} in more cortical areas than AV_{/ta/}, especially in frontal areas (Table 3). A_pV_k was also significantly less active than AV_{/pa/} and AV_{/ka/} in more cortical areas than AV_{/ta/}, again, especially in frontal regions (Table 4). With respect to these frontal areas, A_pV_k differed from both AV_{/pa/} and AV_{/ka/} in the ventral aspect of the premotor cortex, whereas A_pV_k did not differ from AV_{/ta/} in this area.

Thus, the activation patterns associated with A_pV_k showed a smaller difference in the extent of activity when compared with AV_{/ta/} than when compared with AV_{/pa/} or AV_{/ka/}. Though the relative lack of difference between A_pV_k and AV_{/ta/} is a null result, these findings indicate that the incongruent A_pV_k stimulus produces patterns of cortical activity that are more similar to that of a congruent AV_{/ta/} syllable, especially in frontal areas including PMv cortex (see Olson et al. 2002 for a similar result). This suggests that the motor system treats A_pV_k as if it were the perceived "ta" rather than the observed (i.e., /ka/) or heard (i.e., /pa/) speech. We more explicitly test this below.

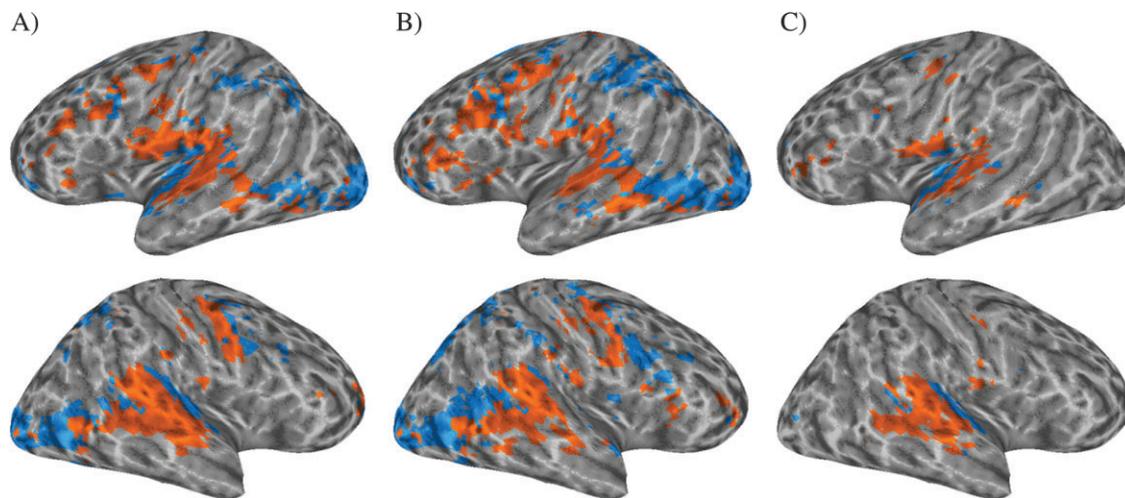


Figure 2. Logical conjunction analyses. Orange indicates regions where activation associated with speaking syllables overlaps with that of activation associated with passively (A) listening to and watching the same congruent AV syllables; (B) watching only the video of these syllables without the accompanying audio track (V); and (C) listening to the syllables without the accompanying video track (A). Overlap images were created using images each thresholded at $P < 0.05$ corrected and logically conjoined. Blue indicates additional regions activated by passive perception alone and not activated by speech production ($P < 0.05$ corrected).

Table 3

Location and center of mass of significantly greater cortical activity for the McGurk-MacDonald syllable (A_pV_k) compared with the other AV syllables (i.e., /pa/, /ka/, or /ta/) for the group

Regions	Hemisphere	MNI center of mass			Cluster size (mm ³)
		x	y	z	
McGurk-Macdonald syllable— $A_pV_k > AV_{/pa/}$					
<i>Temporal and parietal</i>					
Inferior parietal lobule	Left	-44	-54	46	132
STa cortex	Right	55	11	-3	660
SMG	Right	51	-52	32	3950
					Total volume = 4742
<i>Frontal</i>					
Medial frontal gyrus	Left	-2	-19	69	111
Middle frontal gyrus	Left	-2	-20	56	477
Middle frontal gyrus	Left	-35	48	26	224
Middle frontal gyrus	Left	-40	20	47	224
Middle frontal gyrus	Left	-4	-2	64	207
Precentral gyrus	Left	-36	-15	44	148
Superior frontal gyrus	Left	-20	19	50	252
Superior frontal gyrus	Left	-4	23	62	128
Insula	Right	42	-4	-6	1071
Insula	Right	43	-7	13	196
Insula	Right	37	-16	6	195
Precentral gyrus	Right	38	-21	43	120
					Total volume = 3353
McGurk-Macdonald syllable— $A_pV_k > AV_{/ka/}$					
<i>Occipito-temporal</i>					
Lingual gyrus	Right	6	-89	-4	260
<i>Temporal and parietal</i>					
Inferior parietal lobule	Right	48	-33	22	280
SMG	Right	54	-49	26	799
					Total volume = 1079
<i>Frontal</i>					
Medial frontal gyrus	Left	-2	-25	63	639
Medial frontal gyrus	Left	-3	-2	65	355
Middle frontal gyrus	Left	-46	39	-1	608
Middle frontal gyrus	Left	-44	17	44	183
Superior frontal gyrus	Left	-4	8	67	176
Middle frontal gyrus	Right	27	22	49	196
					Total volume = 2157
McGurk-Macdonald syllable— $A_pV_k > AV_{/ta/}$					
<i>Occipito-temporal</i>					
Precuneus	Left	-4	-51	59	196
<i>Temporal and parietal</i>					
Postcentral gyrus	Right	49	-17	30	112
<i>Frontal</i>					
Inferior frontal gyrus	Left	-49	41	10	192
Medial frontal gyrus	Left	-3	-11	61	156
Middle frontal gyrus	Left	-27	21	46	160
Superior frontal gyrus	Left	-23	20	54	336
					Total volume = 844

Overlap of Activity Evoked by the Perceptual Conditions with the Speaking Condition

We next looked for evidence that perception of observable mouth movements results in similar activity as during speech production. This involvement should be detected by comparing activity during AV speech perception with speech production. Stimulus-evoked cortical activation during the passive viewing of congruent AV syllables spatially overlapped with activation found during the speaking of the same syllables, as determined by the intersection of the activity patterns from the main effect of syllable for the AV and speaking ANOVAs (orange in Fig. 2A and Table 5; $P < 0.05$ corrected). Excluding the occipito-temporal areas, Figure 2A and Table 5 show that overlap occurred primarily in anterior superior temporal (STa) cortex, STp, POp, PMv and dorsal premotor (PMd) cortex, SMG, and SI/II. Of these areas, Table 5 indicates that speech perception most extensively overlapped activity elicited by speech production, that is, overlap of greater than 50% of the activated voxels, in STa, STp, and PMv.

Table 4

Location and center of mass of significantly greater cortical activity for the AV syllables (i.e., /pa/, /ka/, or/ta/) compared with the McGurk-MacDonald syllable (A_pV_k) for the group

Regions	Hemisphere	MNI center of mass			Cluster size (mm ³)
		x	y	z	
$AV_{/pa/} > \text{McGurk-Macdonald syllable—}A_pV_k$					
<i>Occipito-temporal</i>					
Fusiform gyrus	Left	-26	-69	-12	132
Inferior occipital gyrus	Left	-28	-94	-8	235
Inferior occipital gyrus	Left	-36	-91	-9	128
Middle occipital gyrus	Left	-47	-69	-10	763
Parahippocampal gyrus	Left	-27	-57	-8	196
Precuneus	Left	-11	-73	20	308
Precuneus	Left	-23	-83	24	224
Precuneus	Left	-21	-63	38	84
Cuneus	Right	11	-83	26	112
Lingual gyrus	Right	5	-69	-5	200
Middle occipital gyrus	Right	43	-68	-8	1244
Middle occipital gyrus	Right	26	-89	-5	84
Precuneus	Right	3	-67	29	1644
					Total volume = 5354
<i>Temporal and parietal</i>					
Angular gyrus	Left	-30	-59	33	699
Inferior parietal lobule	Left	-46	-35	37	476
Middle temporal gyrus	Left	-50	-64	10	735
					Total volume = 1910
<i>Frontal</i>					
Precentral gyrus	Left	-27	-17	66	392
Medial frontal gyrus	Right	4	42	32	148
Medial frontal gyrus	Right	6	52	18	827
Precentral gyrus	Right	45	-10	35	168
					Total volume = 1535
$AV_{/ka/} > \text{McGurk-Macdonald syllable—}A_pV_k$					
<i>Occipito-temporal</i>					
Inferior occipital gyrus	Left	-34	-91	-10	791
Parahippocampal gyrus	Left	-23	-60	-9	356
Precuneus	Left	-19	-71	50	440
Cuneus	Right	2	-73	12	148
Middle occipital gyrus	Right	42	-69	-10	596
Middle occipital gyrus	Right	33	-83	18	208
Parahippocampal gyrus	Right	24	-46	-9	120
					Total volume = 2659
<i>Temporal and parietal</i>					
Angular gyrus	Left	-30	-60	36	476
Inferior parietal lobule	Left	-44	-35	37	464
Middle temporal gyrus	Left	-44	-55	3	148
Inferior parietal lobule	Right	45	-40	44	260
Superior parietal lobule	Right	15	-71	58	344
					Total volume = 1692
<i>Frontal</i>					
Cingulate gyrus	Left	-5	39	-3	480
Lentiform nucleus	Left	-7	4	-5	1008
Middle frontal gyrus	Left	-15	-2	67	280
Cingulate gyrus	Right	2	-58	28	176
Medial frontal gyrus	Right	7	52	15	492
Middle frontal gyrus	Right	23	-16	61	295
Precentral gyrus	Right	33	-3	32	147
					Total volume = 2878
$AV_{/ta/} > \text{McGurk-Macdonald syllable—}A_pV_k$					
<i>Occipito-temporal</i>					
Inferior occipital gyrus	Left	-35	-91	-7	488
Parahippocampal gyrus	Left	-28	-34	-9	424
Precuneus	Left	-29	-64	38	587
Cuneus	Right	4	-69	8	2084
Middle occipital gyrus	Right	43	-70	-8	876
Middle occipital gyrus	Right	36	-81	3	875
Middle occipital gyrus	Right	35	-91	-6	272
Precuneus	Right	26	-75	17	407
Precuneus	Right	2	-70	41	392
Precuneus	Right	27	-81	26	176
					Total volume = 6581
<i>Temporal and parietal</i>					
Inferior parietal lobule	Left	-43	-38	55	576
Inferior temporal gyrus	Left	-42	-67	-3	400
Middle temporal gyrus	Left	-51	-65	10	383
Postcentral gyrus	Left	-45	-22	30	448
STp gyrus	Left	-52	-18	8	471
					Total volume = 2278
<i>Frontal</i>					
Superior frontal gyrus	Left	-16	43	47	244
Insula	Right	40	-9	17	120
Middle frontal gyrus	Right	31	0	37	260
					Total volume = 624

Table 5

Percentage of overlap of cortical activity associated with the AV, A, and V conditions with the speaking condition

Regions	Hemisphere	Percentage of overlap of perceptual conditions with the speaking condition		
		AV speech	A	V
<i>Occipito-temporal</i>				
Anterior occipital sulcus	Left	32	0	30
Inferior occipital gyrus	Left	57	0	55
Inferior occipital sulcus	Left	49	0	56
Middle occipital gyrus and sulcus	Left	35	2	23
Occipital pole	Left	44	0	47
Occipito-temporal gyrus and sulcus	Left	19	0	44
Temporal-occipital sulcus	Left	20	0	37
Inferior occipital gyrus	Right	79	2	14
Middle occipital gyrus and sulcus	Right	37	0	41
Occipital pole	Right	55	3	25
Occipito-temporal gyrus and sulcus	Right	24	0	23
Temporal-occipital sulcus	Right	57	0	17
	Mean	42	1	34
	overlap (%)			
<i>Temporal and parietal regions</i>				
Angular gyrus	Left	17	0	5
STa cortex	Left	53	14	36
Inferior temporal gyrus and sulcus	Left	57	2	22
Intraparietal sulcus	Left	13	0	10
Middle temporal gyrus	Left	21	5	49
Postcentral gyrus and sulcus	Left	19	5	55
STp cortex	Left	55	0	17
SMG	Left	43	20	49
Transverse temporal gyrus and sulcus	Left	90	4	29
Angular gyrus	Right	30	83	19
STa cortex	Right	32	5	18
Inferior temporal gyrus and sulcus	Right	36	1	46
Middle temporal gyrus	Right	36	24	71
Postcentral gyrus and sulcus	Right	11	0	67
STp cortex	Right	79	1	13
Superior parietal lobule	Right	16	27	18
SMG	Right	35	49	8
Transverse temporal gyrus and sulcus	Right	55	63	63
	Mean	39	17	33
	overlap (%)			
<i>Frontal regions</i>				
Cingulate gyrus and sulcus	Left	15	1	11
PMd cortex	Left	16	9	31
Inferior frontal sulcus	Left	33	3	45
Insula	Left	12	8	13
POp	Left	11	7	11
Primary motor cortex	Left	14	7	14
Superior frontal gyrus and sulcus	Left	20	3	13
PMv cortex	Left	22	6	50
PMd cortex	Right	43	0	41
Superior frontal gyrus	Right	15	1	32
PMv cortex	Right	51	3	48
	Mean	23	4	28
	overlap (%)			

Note: Regions are limited to those whose overlap was greater than 10% in the AV condition.

Because A_pV_k is not a naturally spoken syllable and could result in a different pattern of activity compared with the congruent AV syllables, it was excluded from the ANOVA used in the above logical intersection analysis. When the logical intersection analysis was repeated with the A_pV_k stimulus in the ANOVA, neither the activated areas nor the distribution of activity within those areas significantly changed.

Overlap analysis of the activity resulting from V and A stimuli with speech production was used to assess the hypothesis that the recruitment of areas of cortex involved in speech production during AV stimuli is largely due to the participation of the speech production system in the analysis of observable mouth movements. Results indicate that activity associated with the intersection of the V and speech production conditions was found in the same areas identified in the intersection of the AV

and speech production conditions (orange in Fig. 2B and Table 5; $P < 0.05$ corrected). If anything, the V condition yielded a more robust pattern of overlap of activity with the speech production condition. Logical conjunction of activity resulting from the A and speech production conditions, however, showed little overlap except in temporal and parietal areas (orange in Fig. 2C and Table 5; $P < 0.05$ corrected).

Activity in Frontal Regions Associated with Speech Production Corresponds to the Perceived Syllable

The remaining analyses were conducted to understand the computational role of the motor system, operationally defined here as those regions active in both speech perception and production in the frontal lobe, in creating the AV percept. The experimental prediction is that the distribution of motor cortical activity associated with the perception of the “ta” McGurk–MacDonald effect will more closely resemble the distribution of activity for $AV_{/ta/}$ (i.e., the stimulus corresponding to the participants’ “ta” perception) than $AV_{/pa/}$ or $AV_{/ka/}$ for the “ta” group of participants (Table 1). Alternatively, the distribution of activity for the perception of the McGurk–MacDonald stimulus as “ta” in the motor system could resemble the distribution of activity for $AV_{/ka/}$ (i.e., the stimulus corresponding to the visual information about mouth movements in this stimulus) and/or $AV_{/pa/}$ (i.e., the stimulus corresponding to the audio component of the stimulus), suggesting that the motor system more veridically represents the visual or auditory input.

Pairwise correlations were calculated between the distribution of activity associated with A_pV_k and the activity separately associated with each of the $AV_{/pa/}$, $AV_{/ka/}$, or $AV_{/ta/}$ stimuli in the passive task in frontal regions that overlapped speech production (see Table 5 for regions). A 2-way nonparametric Friedman test indicated a significant difference among the pairwise correlations (Friedman ranks test = 14.00, $P = 0.001$). A nonparametric post hoc test of the resulting ranks indicated that—for frontal regions that overlap speech production—activity for A_pV_k was significantly more correlated with the distribution of activity corresponding to $AV_{/ta/}$ than it was with either $AV_{/pa/}$ (Nemenyi = 4.43, $0.005 > P > 0.002$) or $AV_{/ka/}$ (Nemenyi = 4.72, $0.005 > P > 0.002$) (Fig. 3A). Similarly, in frontal regions that overlap speech production, activity resulting from A_pV_k was more correlated with $AV_{/ta/}$ than either $AV_{/pa/}$ or $AV_{/ka/}$ when performing the same analysis over the entire time course of activity for the syllables in these motor regions (P values < 0.05 ; see Fig. 3B for an example).

These analyses were repeated for the “ka” group. Though not significant, a trend was observed in which A_pV_k was more correlated with the distribution of activity corresponding to $AV_{/ka/}$ than either $AV_{/pa/}$ or $AV_{/ta/}$ (Friedman ranks test = 2.25, $P > 0.355$; Friedman rank sum = 13, 19, and 16 for A_pV_k and $AV_{/pa/}$, A_pV_k and $AV_{/ka/}$, and A_pV_k and $AV_{/ta/}$, respectively). This lack of significance was due to higher variability in both the responses and activation patterns for the “ka” group relative to the “ta” group.

Therefore, the distribution of cortical activity evoked by A_pV_k in those frontal regions involved in speech production (for listeners who perceived A_pV_k as “ta”) was more similar in nature to that seen for the veridical $AV_{/ta/}$ than it was to that associated with any other stimulus. This result is consistent with the finding presented above in which the activation patterns associated with A_pV_k showed a smaller difference in the extent

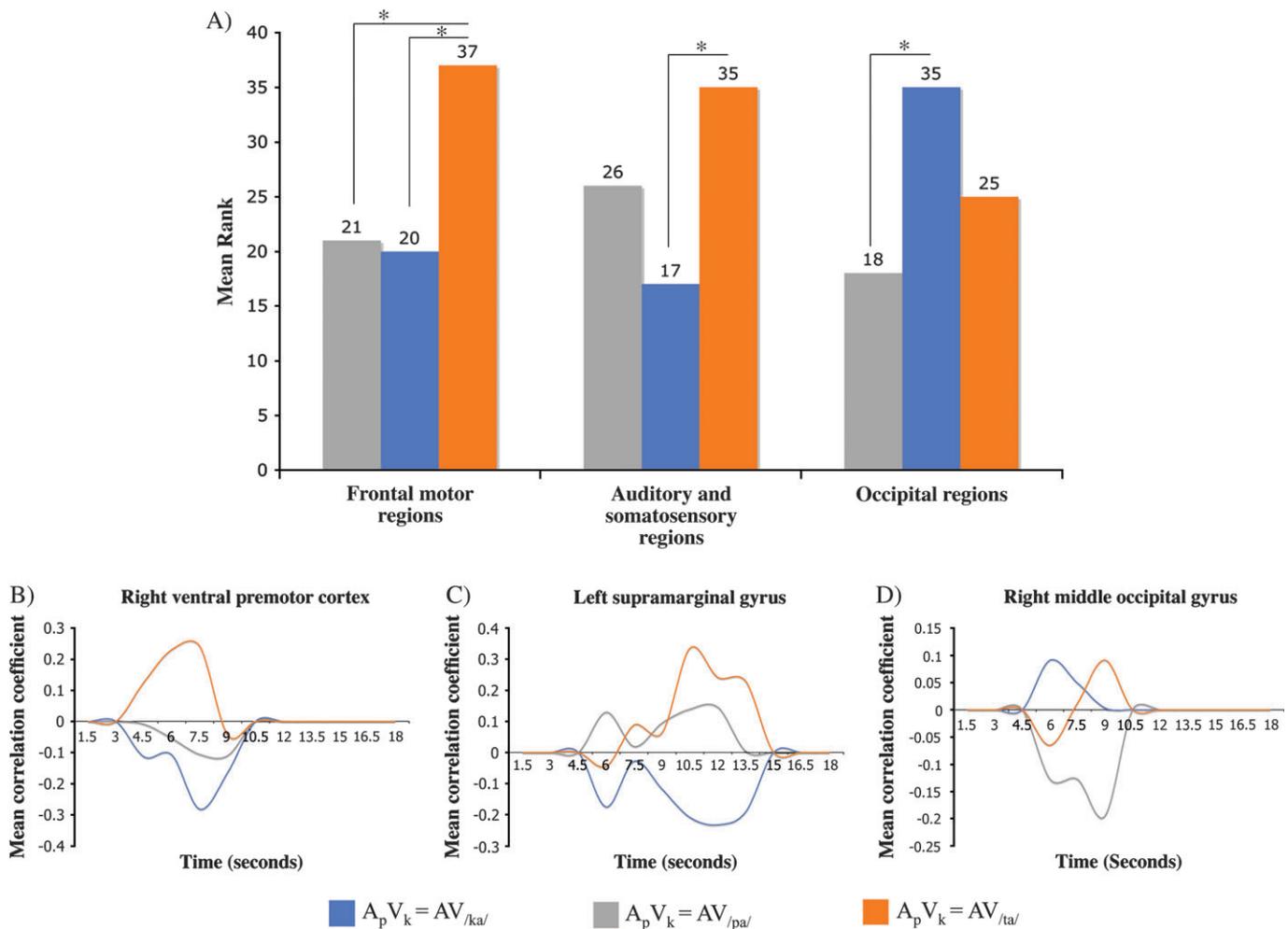


Figure 3. Correlation analyses. Correlation of the distributions of activation associated with passively listening to and watching the incongruent AV syllable made from an audio /pa/ and a visual /ka/ (denoted as $A_p V_k$) and the distributions of activation for $AV_{/pa/}$ (i.e., " $A_p V_k = AV_{/pa/}$ " in gray), $AV_{/ka/}$ (i.e., " $A_p V_k = AV_{/ka/}$ " in blue), or $AV_{/ta/}$ (i.e., " $A_p V_k = AV_{/ta/}$ " in orange) in regions that overlapped speech production. The $A_p V_k$ stimulus elicited the McGurk-MacDonald effect, perceived as "ta" in this group of participants. (A) Correlations analysis when collapsed over the entire time course of activation in all frontal, auditory, and somatosensory sensory, and occipital regions that overlap speech production (Friedman test on pairwise correlations, P values < 0.004; Nemenyi post hoc tests on resulting ranks, $*P$ values < 0.002). This analysis was also conducted at each time point following stimulus onset in the frontal and auditory and somatosensory sensory regions that overlap speech production (see Experimental Procedures). The entire time course of activation is shown in an example (B) motor region, PMv cortex in the right hemisphere; (C) auditory and somatosensory sensory regions that overlap speech production (see Experimental Procedures). The entire time course of activation is shown in an example (D) visual region, the middle occipital gyrus in the right hemisphere (P values < 0.05).

of activity when compared with $AV_{/ta/}$ than when compared with $AV_{/pa/}$ or $AV_{/ka/}$. Conversely, for listeners who perceived $A_p V_k$ as "ka," the trend in motor system activity was more like $AV_{/ka/}$ activity than anything else. These results suggest that activity in frontal motor areas that participate in AV speech perception and production does not simply register visual and/or auditory information but rather represents hypotheses about an *early* integration of AV information.

Furthermore, that $A_p V_k$ was more like a true $AV_{/ta/}$ for participants who perceived "ta" and that $A_p V_k$ was more like a true $AV_{/ka/}$ for participants who perceived "ka" suggests that different hypotheses activate different motor plans resulting in different perceptions. That is, just as producing different syllables requires coordination of different muscles and is therefore mediated by nonidentical neuronal assemblies, the same seems to hold during motor hypotheses testing associated with AV speech perception. To further test this idea, the activation patterns for a subset of trials from the condition in which participants actively classified $A_p V_k$ as "ta" or "ka" were

compared. The classification of $A_p V_k$ as "pa" was excluded from this analysis because this classification occurred on fewer than 2% of the trials. Statistical maps ($P < 0.05$ corrected; Fig. 4) show that when $A_p V_k$ was classified as "ka," significant activation occurred in the middle and inferior frontal gyri and insula. Classifying $A_p V_k$ as "ta" or "ka" yielded cortical activity in spatially adjacent but distinct areas in right inferior and superior parietal lobules, left somatosensory cortices, left PMv, and left M1.

One interpretation of this result is that the observed topography in motor areas could be due to the motor response required of the participants when classifying $A_p V_k$ or to the incongruent or unnatural nature of the $A_p V_k$ stimulus; that is, the observed topography could be an artifact of the task rather than distinct motor hypotheses about AV stimuli. In order to address this concern a discriminant analysis was performed to assess the presence of topographic population codes in these regions for congruent $AV_{/ta/}$, $AV_{/ka/}$, and $AV_{/pa/}$ stimuli, in the condition in which participants made no button responses. Discriminant analysis of activation patterns resulting from these

syllables shows that they are distinguishable from one another in the same motor and somatosensory cortices in which the activation occurred during the active task shown in Figure 4 ($P < 0.05$).

Evidence for Prediction through Efference Copy from Frontal Regions Associated with Speech Production during AV Speech Perception

The proposed model (Fig. 1) predicts that activity in auditory and somatosensory areas might initially (i.e., in early stages of stimuli processing) correspond to the physical properties of the stimulus impinging on the sensory system but subsequently (in later processing and through efference copy) correspond to the hypothesis. If so, activity associated with the “ta” McGurk–MacDonald effect in auditory and sensory areas involved in both speech perception and production should initially resemble the distribution of activity for AV_{/pa/} (the auditory stimulus) and later that of AV_{/ta/} (the fused percept) but not AV_{/ka/} (the visual stimulus). The activity resulting from the perception of McGurk–MacDonald effect as “ta,” however, should be less correlated with AV_{/ka/} because processing of the visual component of the stimulus by these areas is presumably not as robust as processing associated with the auditory component of the stimulus.

To test this prediction, the correlation analysis described in the previous Results section was performed in active temporal

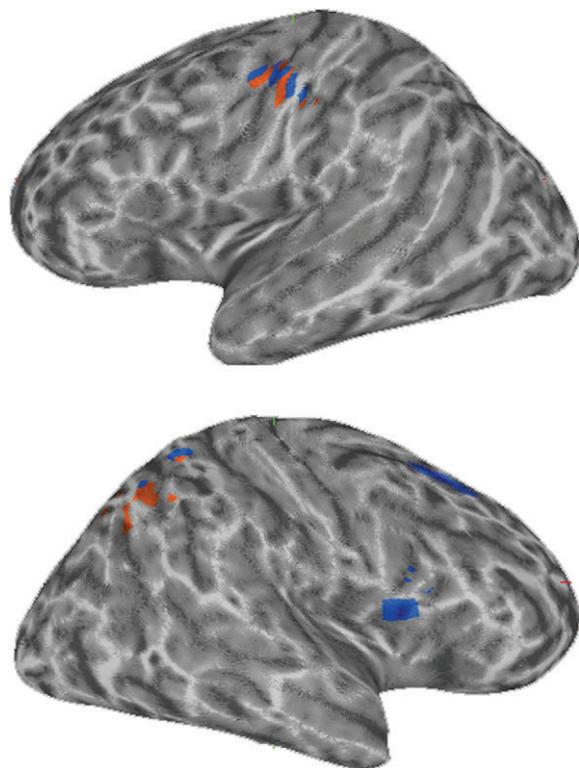


Figure 4. Analysis of the classification condition (i.e., run 5). Contrast ($P < 0.05$ corrected) of activation resulting from hearing a syllable made from an audio /pa/ and a visual /ka/ (denoted as A_pV_k) in one of 2 ways. Blue and orange indicate regions showing differential activation when participants classified A_pV_k as “ka” or “ta,” respectively, in a 3AFC task. Activation when A_pV_k was classified as “ka” is seen in the middle and inferior frontal gyri and insula. Activation when A_pV_k was classified as “ta” or “ka” is in spatially adjacent but distinct areas in the right inferior and superior parietal lobules, left somatosensory cortices, left PMv cortex, and left primary motor cortex.

and parietal areas from the passive AV condition that were also active during speech production for the “ta” group. A 2-way nonparametric Friedman test indicated a significant difference between the pairwise correlations of A_pV_k with the other AV syllables for those participants who perceived A_pV_k as “ta” (Friedman test = 12.46, $P = 0.001$). Post hoc tests indicated that A_pV_k was more highly correlated with AV_{/ta/} than AV_{/ka/} (Nemenyi = 4.99, $0.005 > P > 0.002$) but not significantly different from AV_{/pa/} (Nemenyi = 2.50, $0.2 > P > 0.1$) (Fig. 3A). Similarly, looking over the entire time course of activity in these auditory and somatosensory regions, a 2-way nonparametric repeated measures ANOVA indicated a significant difference between the pairwise correlations of A_pV_k with the other AV syllables at the onset of activity (Friedman test = 17.08, $P < 0.0001$). Post hoc tests indicated that activity evoked by A_pV_k was more highly correlated with activity evoked by AV_{/pa/} than with activity evoked by AV_{/ka/} (Nemenyi = 3.33, $0.05 > P > 0.02$) or AV_{/ta/} for the first 1.5 s of the hemodynamic response (Nemenyi = 5.82, $P < 0.001$). At later time points, however, activity was significantly more correlated with AV_{/ta/} than AV_{/pa/} or AV_{/ka/} (P values < 0.05 ; see Fig. 3C for an example).

Similarly, the distribution of activity for A_pV_k in visual areas was predicted to resemble the distribution of activity for AV_{/ka/} (the visual stimulus) and AV_{/ta/} (the “fused” percept) but not AV_{/pa/} (the auditory stimulus). That is, the visual system receives visual stimulation consistent with AV_{/ka/} but not AV_{/pa/} and shifts to a pattern consistent with the stimulus corresponding to the participant’s perception, AV_{/ta/}. Indeed, a 2-way nonparametric Friedman test indicated a significant difference between the pairwise correlations of A_pV_k with the other AV syllables for those participants who perceived A_pV_k as “ta” (Friedman test = 11.23, $P = 0.004$). Post hoc tests indicated that A_pV_k was more highly correlated with AV_{/ka/} than AV_{/pa/} (Nemenyi = 4.72, $0.005 > P > 0.002$) but not significantly different from AV_{/ta/} (Nemenyi = 2.77, $0.2 > P > 0.1$) (Fig. 3A; see Fig. 3D for an example over the entire time course). This suggests that, like the auditory and somatosensory systems, the visual system shifts from a sensory-based activity pattern (i.e., from /ka/) to one that is more consistent with activity in the motor system.

Above it was shown that A_pV_k was more highly correlated with activity evoked by AV_{/ta/} over the entire time course of activity in motor regions. A_pV_k was, however, more highly correlated with activity evoked by AV_{/pa/} for the first 1.5 s of the hemodynamic response but was thereafter more correlated with AV_{/ta/} in auditory and somatosensory cortices. Here we test whether the strong correlation of A_pV_k with AV_{/ta/} in motor regions precedes this *shift* in the correlation of A_pV_k with AV_{/pa/} to AV_{/ta/} in auditory and somatosensory cortices. Indeed, the correlation of A_pV_k evoked activity with AV_{/ta/} evoked activity in motor regions is significantly stronger than the correlation of A_pV_k with AV_{/ta/} in auditory and somatosensory areas for the first 4.5 s of the hemodynamic response following stimulus presentation (P values < 0.05). Thereafter, however, there is no significant difference between the correlations of A_pV_k with AV_{/ta/} in motor and auditory and somatosensory cortices.

Discussion

The present results show that 1) certain cortical areas active during speech production are also active during both congruent and incongruent AV speech perception, and that this activity primarily occurs when mouth movements are observed by

listeners (Figure 2 and Table 5); 2) the active classification of the McGurk–MacDonald stimulus as 2 different syllables (/ta/ or /ka/) evokes 2 different patterns of activity in frontal cortical areas involved in speech production (Fig. 4); 3) similarly, different congruent AV syllables, when passively viewed and heard, result in different patterns of activity in frontal cortical areas involved in speech production; 4) the pattern of activity evoked by the McGurk–MacDonald syllable in frontal cortical areas involved in speech production most resembles that evoked by the congruent AV syllable that corresponds to participants' perception of the stimulus (Fig. 3); and 5) at the onset of the hemodynamic response, the pattern of activity evoked by the McGurk–MacDonald syllable in sensory areas most resembles the pattern of activity evoked by the congruent AV syllable that corresponds to sensory input. Later, it most resembles the pattern of activity evoked by the congruent AV syllable that corresponds to participants' ultimate perception (see Fig. 3 for an example).

Our results show that aspects of the cortical motor system that are involved in speech production are also involved in perception, suggestive of a shared underlying system for AV speech perception and production (see (1) above) but also indicative of the mechanisms of this system. Motor areas involved in production, when activated differentially, result in different perceptual experiences of AV speech (see (2) and (3) above). This can be explained by a shared mechanism of perception and production in that producing different sounds requires the coordination of different muscle groups or the same groups in different proportions. Furthermore, our results suggest that this motor system activity is involved in determining the ultimate phonetic perception of AV stimuli because the same stimulus perceived in different ways is associated with 2 distinct patterns of activity in the motor areas involved in speech production (see (2) above).

In what follows we relate our interpretation of the results—that the motor system is involved in determining our perceptual experience of AV speech through a shared mechanism of perception and production—to 2 aspects of the model outlined in the Introduction (see also Fig. 1 and Table 1). Recall, our model was formulated to go beyond the claim that a mirror or motor system is associated with both speech production and AV speech perception and to explain the mechanism by which motor system activity determines speech perception. We proposed that motor system activity be thought of as a multisensory *hypothesis* about the phonemes produced by a speaker and that this hypothesis predicts the acoustic and somatosensory consequences of executing a speech movement through efference copy.

We believe that our results show that activity in areas of the motor system associated with speech production during observation of AV speech is a hypothesis about a particular (phonetic) interpretation of the stimulus properties rather than an accurate or veridical representation of stimulus properties. Specifically, by traditional contrasts (Tables 3 and 4), we showed that in terms of patterns of cortical activity, the McGurk stimulus was least different from the congruent AV stimulus /ta/ in PMv cortex. Consistent with this result, we used the McGurk–MacDonald eliciting stimulus to show that the response in the motor system associated with speech production was independent of the physical stimulus impinging on the sensory respecters but, rather, corresponded to the participants' ultimate perception (i.e., the congruent AV stimulus /ta/; see (4) above).

At this point the argument could be made that activation of the motor system occurs after perception in that participants may be imitating or rehearsing the McGurk–MacDonald stimulus after they had perceived it. Similarly, one could argue that activation of the motor system constitutes perception. Contrary to this interpretation, however, we show that whereas the motor system reflects activity consistent with the ultimate perception of a syllable, the auditory and somatosensory systems shift from a sensory-based activity pattern to one that is more consistent with activity in the frontal motor system (see (5) above). Furthermore, activity in frontal motor areas involved in speech production precedes this switch in sensory areas from a sensory-based activity pattern to one more consistent with activity in the motor system (see Results). Thus, rather than interpreting motor system activity as imitation or rehearsal or as constituting perception, we interpret our results as consistent with our model. That is, in our model activity in the motor system associated with speech production constitutes a hypothesis and that this hypothesis is used to predict the acoustic and somatosensory consequences of executing that hypothesis through efference copy.

In the following section we further discuss the significance of these results and their interpretation with respect to theories of speech perception.

Distributed-Interactive and Direct-Mapping Accounts of Speech Perception

The model described in the Introduction for which we have garnered evidence was formulated in order to explain the neural mechanisms by which listeners come to understand speech in face-to-face contexts (Fig. 1, for similar theories see Stevens and Halle 1967; Callan et al. 2004; van Wassenhove et al. 2005). In so doing, this model addresses a specific theoretical problem that has been the focus of speech perception research for decades, namely, the “lack of invariance problem.” Lack of invariance refers to the observation that different acoustic speech signals map onto identical phonetic categories and, conversely, that identical acoustic signals map onto different phonetic categories (e.g., Liberman et al. 1967), meaning that sensory information alone is insufficient to determine the identity of a phoneme (see Nusbaum and Magnuson 1997).

Many theories of speech perception (Fant 1969; Stevens and Blumstein 1978; Klatt 1979; Liberman and Mattingly 1985; McClelland and Elman 1986) and indeed theories of perception more broadly (Hubel and Wiesel 1959, 1977; Barlow 1972; Barlow et al. 1972) argue for a direct mapping of stimulus signals (e.g., acoustic properties) onto perception (e.g., phonetic categories). These theories rest on the notion of tuned neural feature detectors that are presumed to respond selectively to successively more abstract transformations of stimulus input from thalamic systems to primary sensory cortices into secondary and associated sensory cortices until a memory system is contacted representing perception (Whitfield and Evans 1965; Abbs and Sussman 1971; Barlow 1972; Barlow et al. 1972).

With respect to speech perception, a direct-mapping approach cannot computationally manage the nondeterministic relationship between acoustic patterns and phonetic categories (Nusbaum and Magnuson 1997). Just as neural models of bat echolocation (previously thought to exist as an isolated modular system) account for echolocation through complex sensorimotor interactions over distributed neural systems (Suga and

Schlegel 1972; Suga and Shimozawa 1974; Yan and Suga 1996), we suggest that the process of speech perception involves interactions over a distributed set of sensorimotor cortices through mechanisms of both feedforward and feedback processing. In the following section, we discuss how this conclusion specifically relates to 2 influential theories of speech perception, the motor theory of speech perception (Lieberman and Mattingly 1985) and analysis-by-synthesis (Stevens and Halle 1967).

Motor Theories of Speech Perception

The motor theory of speech perception claims to solve the lack of invariance problem by positing that speech perception is directly mediated solely by a motor or gestural code (Lieberman and Mattingly 1985). The theory states that speech perception occurs directly by reference to invariant motor programs for speech production. Thus, *all* speech is directly transduced into a gestural code. Lieberman and Mattingly (1985) further state that there is no auditory processing of speech in service of phoneme perception.

Neuroimaging results from auditory (Wilson et al. 2004; Pulvermuller et al. 2006; Wilson and Iacoboni 2006) and visual (Olson et al. 2002) speech perception alone have been shown to evoke motor cortical activity, and this has been used to argue in favor of the motor theory of speech perception. The present study fleshes out the mechanism underlying these findings in showing that although speech-associated mouth movements (heard or observed) evoke motor cortical activity in speech production brain regions (Fig. 2 and Table 5), speech perception is not simply determined through feedforward or direct-mapping mechanisms. Instead, our results suggest that sensory input interacts with feedback or efference copy from the motor system involved in speech production (see Results and Fig. 3). Although this argues against the specific claims of the motor theory of speech perception (Lieberman and Mattingly 1985), it also argues against purely sensory theories of speech perception (e.g., Fant 1969; Blumstein and Stevens 1981). Furthermore, it raises a question regarding the nature of the interaction between motor and sensory cortices in speech perception.

The “analysis-by-synthesis” model of speech perception developed by Stevens and Halle (1967) addresses the nature of these interactions. In this model, sensory information is used to generate knowledge-based hypotheses about the speech category that the speaker intended. That is, these candidate speech categories are internally “synthesized.” Matching between the synthesized hypothesis and sensory information is used to select the intended speech category. When the discrepancy between the hypothesis and sensory inputs is small, no further processing is necessary. In this case, the mapping between sensory information and phonetic categories is sufficiently deterministic (see Nusbaum and Magnuson 1997). When the discrepancy between the hypothesis and the sensory inputs is large, however, the motor system is recruited. That is, when there is substantial phonetic ambiguity or lack of invariance, the nondeterministic relationship needs to be resolved by additional information provided by the motor system.

There are neurobiological data to support the view that the motor system is actively recruited to resolve phonetic ambiguity associated with the lack of invariance. Callan et al. (2004) have shown that nonnative speakers evoke more activity in motor areas during the perceptual identification of a phonetic

contrast that is ambiguous in their language compared with native speakers of a language in which that contrast is not ambiguous. Native speakers, by comparison, evoked greater activity exclusively in auditory cortices. Callan, Tajima, et al. (2003) have also shown that more ambiguous distinctions for nonnative speakers activate motor cortices to a much larger extent than less ambiguous distinctions.

In our neurobiologically specified version of this analysis-by-synthesis model, the motor system involved in speech production is recruited because it can be used to predict the sensory consequences of the synthesized hypotheses, which can then be matched with incoming sensory information to constrain interpretation. There is considerable behavioral (Knoblich and Flach 2001; Wolpert and Flanagan 2001) and neurophysiological (Zaretsky and Rowell 1979; Bell 1981; Sillar and Skorupski 1986; Sillar and Roberts 1988; Gossard et al. 1991; Wolf and Burrows 1995; el Manira et al. 1996; Blakemore et al. 1998; Wolpert et al. 1998; Bodznick et al. 1999; Iacoboni et al. 2001; Iacoboni 2005) data to support the proposal that prediction and efference copy are general properties of the motor system across vertebrates.

If the motor system is particularly good at prediction then it should be particularly attuned to observable mouth movements because this information can be used to predict auditory sensory information making speech perception easier in everyday environments (see van Wassenhove et al. 2005). This may explain why, in the present experiment, AV and visual speech perception yielded more robust activity in the motor system associated with speech production compared with the auditory condition alone (see Fig. 2).

Summary

In summary, the results of the present experiment point to a model of speech perception in which visual contextual information is used to test hypotheses about the identity of speech sounds. In our hypothesis-and-test or analysis-by-synthesis model, a hypothesis is specified in terms of the motor commands that might elicit the hypothesized movements. These motor commands can then be used to predict the sensory consequences of producing those movements through efference copy. The predicted sensory consequences of the motor hypothesis can be compared with or lend support to a particular interpretation of a stretch of utterance as a sensory representation of that utterance is being built. An interpretation is reached and perception occurs only after integrating the predicted sensory consequences of the motor representation and ongoing sensory processing. It is in this manner that the brain actually comes to see voices and hear (it's own) lips, as suggested by McGurk and MacDonald (1976).

Notes

This work was supported by National Institutes of Health RO1-DC03378. Thanks to the following people who each contributed to this paper in some manner: Michael Arbib, Catherine Barber, Bernadette Brogan, E. Chen, Shahrina Chowdhury, Jean Decety, Fred Dick, Kim Fenn, Tom Hummer, Goulven Josse, Philippa Lauben and Leo Stengel, Matt Longo, Robert Lyons, Xander Meadow, Lisa Patterson, David Poeppel, Anjali Raja, Ana Solodkin, Ryan Walsh, and Nicholas Wymbs. Special thanks to Alison Wiener-Skipper for discussion and much support and Uri Hasson for discussion and advice. *Conflict of Interest:* None declared.

Address correspondence to Jeremy I. Skipper, Department of Neurology, The University of Chicago, MC 2030, 5841 South Maryland Ave., Chicago, IL 60637, USA. Email: skipper@uchicago.edu.

References

- Abbs JH, Sussman HM. 1971. Neurophysiological feature detectors and speech perception: a discussion of theoretical implications. *J Speech Hear Res.* 14:23-36.
- Barlow HB. 1972. Single units and sensation: a neuron doctrine for perceptual psychology? *Perception.* 1:371-394.
- Barlow HB, Narasimhan R, Rosenfeld A. 1972. Visual pattern analysis in machines and animals. *Science.* 177:567-575.
- Belin P, Zatorre RJ, Ahad P. 2002. Human temporal-lobe response to vocal sounds. *Brain Res Cogn Brain Res.* 13:17-26.
- Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B. 2000. Voice-selective areas in human auditory cortex. *Nature.* 403:309-312.
- Bell CC. 1981. An efference copy which is modified by reafferent input. *Science.* 214:450-453.
- Blakemore SJ, Wolpert DM, Frith CD. 1998. Central cancellation of self-produced tickle sensation. *Nat Neurosci.* 1:635-640.
- Blumstein SE, Stevens KN. 1981. Phonetic features and acoustic invariance in speech. *Cognition.* 10:25-32.
- Bodznick D, Montgomery JC, Carey M. 1999. Adaptive mechanisms in the elasmobranch hindbrain. *J Exp Biol.* 202:1357-1364.
- Buccino G, Lui F, Canessa N, Patteri I, Lagravinese G, Benuzzi F, Porro CA, Rizzolatti G. 2004. Neural circuits involved in the recognition of actions performed by nonconspecifics: an fMRI study. *J Cogn Neurosci.* 16:114-126.
- Burton MW, Small SL, Blumstein SE. 2000. The role of segmentation in phonological processing: an fMRI investigation. *J Cogn Neurosci.* 12:679-690.
- Callan DE, Jones JA, Callan AM, Akahane-Yamada R. 2004. Phonetic perceptual identification by native- and second-language speakers differentially activates brain regions involved with acoustic phonetic processing and those involved with articulatory-auditory/orosensory internal models. *Neuroimage.* 22:1182-1194.
- Callan DE, Jones JA, Munhall K, Callan AM, Kroos C, Vatikiotis-Bateson E. 2003. Neural processes underlying perceptual enhancement by visual speech gestures. *Neuroreport.* 14:2213-2218.
- Callan DE, Tajima K, Callan AM, Kubo R, Masaki S, Akahane-Yamada R. 2003. Learning-induced neural plasticity associated with improved identification performance after training of a difficult second-language phonetic contrast. *Neuroimage.* 19:113-124.
- Calvert GA, Campbell R. 2003. Reading speech from still and moving faces: the neural substrates of visible speech. *J Cogn Neurosci.* 15:57-70.
- Campbell R, MacSweeney M, Surguladze S, Calvert G, McGuire P, Suckling J, Brammer MJ, David AS. 2001. 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). *Brain Res Cogn Brain Res.* 12:233-243.
- Caviness VS, Meyer J, Makris N, Kennedy DN. 1996. MRI-based topographic parcellation of human neocortex: an anatomically specified method with estimate of reliability. *J Cogn Neurosci.* 8:566-587.
- Chung MK, Robbins SM, Dalton KM, Davidson RJ, Alexander AL, Evans AC. 2005. Cortical thickness analysis in autism with heat kernel smoothing. *Neuroimage.* 25:1256-1265.
- Cox RW. 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput Biomed Res.* 29:162-173.
- el Manira A, Shupliakov O, Fagerstedt P, Grillner S. 1996. Monosynaptic input from cutaneous sensory afferents to fin motoneurons in lamprey. *J Comp Neurol.* 369:533-542.
- Fant G. 1969. Acoustic theory of speech perception. The Hague, The Netherlands: Mouton.
- Fischl B, Sereno MI, Dale AM. 1999. Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. *Neuroimage.* 9:195-207.
- Fischl B, Sereno MI, Tootell RB, Dale AM. 1999. High-resolution intersubject averaging and a coordinate system for the cortical surface. *Hum Brain Mapp.* 8:272-284.
- Fischl B, van der Kouwe A, Destrieux C, Halgren E, Segonne F, Salat DH, Busa E, Seidman LJ, Goldstein J, Kennedy D, et al. 2004. Automatically parcellating the human cerebral cortex. *Cereb Cortex.* 14:11-22.
- Genovese CR, Lazar NA, Nichols T. 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage.* 15:870-878.
- Gentilucci M, Cattaneo L. 2005. Automatic audiovisual integration in speech perception. *Exp Brain Res.* 167:66-75.
- Glover GH. 1999. Deconvolution of impulse response in event-related BOLD fMRI. *Neuroimage.* 9:416-429.
- Gossard JP, Cabelguen JM, Rossignol S. 1991. An intracellular study of muscle primary afferents during fictive locomotion in the cat. *J Neurophysiol.* 65:914-926.
- Grant KW, Greenberg S. 2001. Speech intelligibility derived from asynchronous processing of auditory-visual information. *Proceedings of the Workshop on Audio-visual Speech Processing*; Scheelsminde, Denmark.
- Hatfield G. 2002. Perception as unconscious inference. In: Heyer D, Mausfeld R, editors. *Perception and the physical world: psychological and philosophical issue in perception.* New York: John Wiley & Sons, Ltd. p. 115-143.
- Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P. 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science.* 293:2425-2430.
- Helmholtz HL. 1867. *Handbuch der physiologischen Optik.* Leipzig: L. Voss.
- Hubel DH, Wiesel TN. 1959. Receptive fields of single neurones in the cat's striate cortex. *J Physiol.* 148:574-591.
- Hubel DH, Wiesel TN. 1977. Ferrier lecture. Functional architecture of macaque monkey visual cortex. *Proc R Soc Lond B Biol Sci.* 198:1-59.
- Iacoboni M. 2005. Understanding others: imitation, language, empathy. In: Hurley S, Chater N, editors. *Perspectives on Imitation: from Cognitive Neuroscience to Social Science, Volume 1: Mechanisms of Imitation and Imitation in Animals.* Cambridge, MA: MIT Press. p. 77-99.
- Iacoboni M, Koski LM, Brass M, Bekkering H, Woods RP, Dubeau MC, Mazziotta JC, Rizzolatti G. 2001. Reafferent copies of imitated actions in the right superior temporal cortex. *Proc Natl Acad Sci USA.* 98:13995-13999.
- Jordan MI, Rumelhart DE. 1992. Forward models: supervised learning with a distal teacher. *Cogn Sci.* 16:307-354.
- Kerzel D, Bekkering H. 2000. Motor activation from visible speech: evidence from stimulus response compatibility. *J Exp Psychol Hum Percept Perform.* 26:634-647.
- Klatt DH. 1979. Speech perception: a model of acoustic-phonetic analysis and lexical access. *J Phon.* 7:279-312.
- Knoblich G, Flach R. 2001. Predicting the effects of actions: interactions of perception and action. *Psychol Sci.* 12:467-472.
- Liberman AM, Cooper FS, Shankweiler DS, Studdert-Kennedy M. 1967. Perception of the speech code. *Psychol Rev.* 74:431-461.
- Liberman AM, Mattingly IG. 1985. The motor theory of speech perception revised. *Cognition.* 21:1-36.
- McClelland JL, Elman JL. 1986. The TRACE model of speech perception. *Cognit Psychol.* 18:1-86.
- McGurk H, MacDonald J. 1976. Hearing lips and seeing voices. *Nature.* 264:746-748.
- Miall RC. 2003. Connecting mirror neurons and forward models. *Neuroreport.* 14:2135-2137.
- Nishitani N, Hari R. 2002. Viewing lip forms: cortical dynamics. *Neuron.* 36:1211-1220.
- Noll DC, Cohen JD, Meyer CH, Schneider W. 1995. Spiral k-space MRI imaging of cortical activation. *J Magn Reson Imaging.* 5:49-56.
- Nusbaum HC, Magnuson J. 1997. Talker normalization: phonetic constancy as a cognitive process. In: Johnson K, Mullennix JW, editors. *Talker variability in speech processing.* San Diego, CA: Academic Press. p. 109-132.
- Oldfield RC. 1971. The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia.* 9:97-113.
- Olson IR, Gatenby JC, Gore JC. 2002. A comparison of bound and unbound audio-visual information processing in the human cerebral cortex. *Brain Res Cogn Brain Res.* 14:129-138.

- Paulesu E, Perani D, Blasi V, Silani G, Borghese NA, De Giovanni U, Sensolo S, Fazio F. 2003. A functional-anatomical model for lipreading. *J Neurophysiol.* 90:2005-2013.
- Pekkola J, Laasonen M, Ojanen V, Autti T, Jaaskelainen IP, Kujala T, Sams M. 2006. Perception of matching and conflicting audiovisual speech in dyslexic and fluent readers: an fMRI study at 3 T. *Neuroimage.* 29:797-807.
- Pulvermuller F, Huss M, Kherif F, Moscoso del Prado Martin F, Hauk O, Shtyrov Y. 2006. Motor cortex maps articulatory features of speech sounds. *Proc Natl Acad Sci USA.* 103:7865-7870.
- Reisberg D, McLean J, Goldfield A. 1987. Easy to hear but hard to understand: a lipreading advantage with intact auditory stimuli. In: Dodd B, Campbell R, editors. *Hearing by eye: the psychology of lipreading.* Hillsdale, NJ: Erlbaum. p. 97-114.
- Risberg A, Lubker J. 1978. Prosody and speechreading. *Speech Transmission Laboratory Quarterly Progress Report Status Report* 4. p. 1-16.
- Rizzolatti G, Craighero L. 2004. The mirror-neuron system. *Annu Rev Neurosci.* 27:169-192.
- Saad ZS, Reynolds RC, Argall B, Japee S, Cox RW. 2004. SUMA: an interface for surface-based intra- and inter-subject analysis with AFNI. *IEEE International Symposium on Biomedical Imaging; Arlington, VA.* p. 1510-1513.
- Sams M, Mottonen R, Sihvonen T. 2005. Seeing and hearing others and oneself talk. *Brain Res Cogn Brain Res.* 23:429-435.
- Sillar KT, Roberts A. 1988. A neuronal mechanism for sensory gating during locomotion in a vertebrate. *Nature.* 331:262-265.
- Sillar KT, Skorupski P. 1986. Central input to primary afferent neurons in crayfish, *Pacifastacus leniusculus*, is correlated with rhythmic motor output of thoracic ganglia. *J Neurophysiol.* 55:678-688.
- Skipper JI, Nusbaum HC, Small SL. 2005. Listening to talking faces: motor cortical activation during speech perception. *Neuroimage.* 25:76-89.
- Skipper JI, Nusbaum HC, Small SL. 2006. Lending a helping hand to hearing: another motor theory of speech perception. In: Arbib MA, editor. *Action to language via the mirror neuron system.* Cambridge, MA: Cambridge University Press. p. 250-285.
- Sperry RW. 1950. Neural basis of the spontaneous optokinetic response produced by visual inversion. *J Comp Physiol Psychol.* 43:482-489.
- Stevens KN, Blumstein SE. 1978. Invariant cues for place of articulation in stop consonants. *J Acoust Soc Am.* 64:1358-1368.
- Stevens KN, Halle M. 1967. Remarks on analysis by synthesis and distinctive features. In: Wathen-Dunn W, editor. *Models for the perception of speech and visual form.* Cambridge, MA: MIT Press. p. 88-102.
- Suga N, Schlegel P. 1972. Neural attenuation of responses to emitted sounds in echolocating rats. *Science.* 177:82-84.
- Suga N, Shimozawa T. 1974. Site of neural attenuation of responses to self-vocalized sounds in echolocating bats. *Science.* 183:1211-1213.
- Sumby WH, Pollack I. 1954. Visual contribution of speech intelligibility in noise. *J Acoust Soc Am.* 26:212-215.
- Sundara M, Namasisvayam AK, Chen R. 2001. Observation-execution matching system for speech: a magnetic stimulation study. *Neuroreport.* 12:1341-1344.
- van Wassenhove V, Grant KW, Poeppel D. 2005. Visual speech speeds up the neural processing of auditory speech. *Proc Natl Acad Sci USA.* 102:1181-1186.
- von Holst E, Mittelstaedt H. 1950. Das Reafferenzprinzip: Wechselwirkungen Zwischen Zentralnervensystem und Peripherie. *Naturwissenschaften.* 37:464-476.
- Watkins K, Paus T. 2004. Modulation of motor excitability during speech perception: the role of Broca's area. *J Cogn Neurosci.* 16:978-987.
- Watkins KE, Strafella AP, Paus T. 2003. Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia.* 41:989-994.
- Whitfield IC, Evans EF. 1965. Responses of auditory cortical neurons to stimuli of changing frequency. *J Neurophysiol.* 28:655-672.
- Wilson SM, Iacoboni M. 2006. Neural responses to non-native phonemes varying in producibility: evidence for the sensorimotor nature of speech perception. *Neuroimage.* 33:316-325.
- Wilson SM, Saygin AP, Sereno MI, Iacoboni M. 2004. Listening to speech activates motor areas involved in speech production. *Nat Neurosci.* 7:701-702.
- Wolf H, Burrows M. 1995. Proprioceptive sensory neurons of a locust leg receive rhythmic presynaptic inhibition during walking. *J Neurosci.* 15:5623-5636.
- Wolpert DM, Flanagan JR. 2001. Motor prediction. *Curr Biol.* 11:R729-R732.
- Wolpert DM, Goodbody SJ, Husain M. 1998. Maintaining internal representations: the role of the human superior parietal lobe. *Nat Neurosci.* 1:529-533.
- Yan J, Suga N. 1996. Corticofugal modulation of time-domain processing of biosonar information in bats. *Science.* 273:1100-1103.
- Zaretsky M, Rowell CH. 1979. Saccadic suppression by corollary discharge in the locust. *Nature.* 280:583-585.
- Zatorre RJ, Belin P. 2001. Spectral and temporal processing in human auditory cortex. *Cereb Cortex.* 11:946-953.
- Zatorre RJ, Meyer E, Gjedde A, Evans AC. 1996. PET studies of phonetic processing of speech: review, replication, and reanalysis. *Cereb Cortex.* 6:21-30.