Neural Correlates of Phonetic Adaptation as Induced by Lexical and Audiovisual Context

Shruti Ullas[1,2], Lars Hausfeld[1,2], Anne Cutler[3], Frank Eisner[4], and Elia Formisano[1,2,5]

Abstract

■ When speech perception is difficult, one way listeners adjust is by reconfiguring phoneme category boundaries, drawing on contextual information. Both lexical knowledge and lipreading cues are used in this way, but it remains unknown whether these two differing forms of perceptual learning are similar at a neural level. This study compared phoneme boundary adjustments driven by lexical or audiovisual cues, using ultra-high-field 7-T fMRI. During imaging, participants heard exposure stimuli and test stimuli. Exposure stimuli for lexical retuning were audio recordings of words, and those for audiovisual recalibration were audio–video recordings of lip movements during utterances of pseudowords. Test stimuli were ambiguous phonetic strings presented without context, and listeners reported what phoneme they heard. Reports reflected phoneme biases in preceding exposure blocks (e.g., more reported /p/ after /p/-biased exposure). Analysis of corresponding brain responses indicated that both forms of cue use were associated with a network of activity across the temporal cortex, plus parietal, insula, and motor areas. Audiovisual recalibration also elicited significant occipital cortex activity despite the lack of visual stimuli. Activity levels in several ROIs also covaried with strength of audiovisual recalibration, with greater activity accompanying larger recalibration shifts. Similar activation patterns appeared for lexical retuning, but here, no significant ROIs were identified. Audiovisual and lexical forms of perceptual learning thus induce largely similar brain response patterns. However, audiovisual recalibration involves additional visual cortex contributions, suggesting that previously acquired visual information (on lip movements) is retrieved and deployed to disambiguate auditory perception. ■

INTRODUCTION

Speech perception is influenced by information other than the acoustic signal itself, such as seeing concurrent lip movements, or the listener's lexical knowledge. These contextual cues not only support speech comprehension but can also create categorically different and novel percepts; consider, for example, the McGurk effect, whereby an auditory syllable (such as /ba/) paired with video of a speaker pronouncing an incongruent syllable (such as /ga/) leads to a perceived new syllable (often /da/; McGurk & MacDonald, 1976). Similarly, when presented with a word containing an unclear syllable (such as a /d/–/t/ blend instead of /d/ in *desk*), listeners are more likely to report hearing a word rather than a nonword (*desk* rather than *tesk*; Ganong, 1980). Audiovisual lipreading cues and lexical knowledge can guide and disrupt perception but can also alter the categorical boundaries of presented phonemes.

Through audiovisual recalibration, listeners presented with video of a speaker pronouncing a syllable, such as /aba/, paired with an ambiguous auditory stimulus (an /aba/–/ada/ mixture) are, after sufficient exposure to the combination, likely to perceive the auditory blend without visual cues as /aba/ (Bertelson, Vroomen, & De Gelder, 2003). Similarly, in lexically guided perceptual retuning, listeners presented with an ambiguous phoneme embedded within words (such as an /s/–/f/ blend in place of /s/ in words such as *horse*) are later likely to identify the /s/–/f/ phoneme blend when it is heard without lexical context as /s/ (Norris, McQueen, & Cutler, 2003).

Both of these approaches allow a glimpse into how speech sound categories can be shifted using contextual cues in addition to the acoustic signal. As audiovisual recalibration can operate through an additional sensory modality (vision), unlike lexical retuning that relies on word recognition within the same sensory channel (audition), the two forms of perceptual learning tend to differ in how they can be induced. In audiovisual processing, the visual cues such as lip movements are available earlier to the listener (Jesse & Massaro, 2010), and thus strong perceptual shifts can be observed after only a few exposure items, but these effects also diminish quickly (Vroomen, van Linden, Keetels, de Gelder, & Bertelson, 2004), whereas lexical cues can lead to longer-lasting, more robust effects, but after long exposures toward one particular phoneme (Eisner & McQueen, 2006). When lexical and audiovisual effects are compared under the same exposure and testing conditions, with short exposures (i.e., eight biasing items) in alternation with short categorization tests on ambiguous items, both adaptation

[1]Maastricht University, [2]Maastricht Brain Imaging Centre, [3]MARCS Institute, Western Sydney University, [4]Radboud University Nijmegen, [5]Maastricht Centre for Systems Biology

Journal of Cognitive Neuroscience 32:11, pp. 2145–2158
https://doi.org/10.1162/jocn_a_01608

effects occur, with audiovisual cues generating larger perceptual shifts than lexical cues (Ullas, Formisano, Eisner, & Cutler, 2020a; van Linden & Vroomen, 2007); the behavioral effects are however not additive (Ullas, Formisano, Eisner, & Cutler, 2020b).

The application of neuroimaging techniques such as fMRI has indicated some of the brain regions involved in category retuning. In general, speech perception employs a network of primarily left-lateralized regions in and around the temporal cortex, particularly within Heschl's gyrus (HG) and planum temporale (PT; Zatorre, Belin, & Penhune, 2002; Binder, 2000; Zatorre, Evans, Meyer, & Gjedde, 1992). Phonetic perception has been linked to activation in HG and PT (Jäncke, Wüstenberg, Scheich, & Heinze, 2002) as well as the superior temporal gyrus (STG) and STS (Formisano, De Martino, Bonte, & Goebel, 2008; Buchsbaum, Hickok, & Humphries, 2001); these areas are also responsible for encoding low-level acoustic–phonetic features and phonemes (Rutten, Santoro, Hervais-Adelman, Formisano, & Golestani, 2019; Leonard & Chang, 2014; Mesgarani, Cheung, Johnson, & Chang, 2014; Chang et al., 2011; Mesgarani, David, Fritz, & Shamma, 2008). STG and STS are also implicated in distinguishing intelligible speech from distorted speech (Davis & Johnsrude, 2003), recognizing consonant–vowel syllables (Liebenthal, Binder, Spitzer, Possing, & Medler, 2005), and identifying phonemic sounds (Liebenthal & Bernstein, 2017). Dual streams of processing may be responsible for acoustic feature processing and gestural motor processing, separated by an anterior–ventral and posterior–dorsal pathway, respectively (Hickok & Poeppel, 2004; Scott & Johnsrude, 2003), although phoneme processing can be bilateral and shared between networks in both the left and right hemispheres (Formisano et al., 2008; Hickok & Poeppel, 2004).

Speech perception extends into frontal and parietal regions as well (Rauschecker & Scott, 2009). Premotor, motor, and parieto-temporal regions are pertinent for representing articulatory gestures and sensorimotor functions (Hickok & Poeppel, 2007), whereas the left inferior frontal gyrus (IFG) is notably linked to speech comprehension and unifying various levels of linguistic information, including phonemes, syllables, and semantics (Hagoort, 2005; Sharp, Scott, Cutler, & Wise, 2005; Poldrack et al., 1999).

When lip movement cues accompany speech, creating audiovisual speech, a similar pattern of activity in the brain can be found across frontal, parietal, and temporal regions (Bernstein & Liebenthal, 2014; Dick, Solodkin, & Small, 2010), with the addition of occipito-temporal contributions (Skipper, Van Wassenhove, Nusbaum, & Small, 2007). Activity in STG and IFG has been observed while listeners experience the McGurk effect (Jones & Callan, 2003), and phoneme boundary shifts resulting from the McGurk effect have been located within STG (Lüttke, Ekman, Van Gerven, & De Lange, 2016). STS may also facilitate perception of noisy audiovisual speech (Beauchamp, 2005), and contextual influences from surrounding sentences on phoneme processing can be exerted by STG and left middle temporal gyrus (MTG; Guediche, Salvata, & Blumstein, 2013). Kilian-Hütten, Vroomen, and Formisano (2011) specifically investigated audiovisual recalibration using fMRI. These authors found that exposure to the audiovisual pairings of ambiguous syllables with videos of lip movements elicited activity in STG, as well as in the inferior parietal lobe (IPL), inferior frontal sulcus, and posterior MTG. Interestingly, activity in response to exposure of adaptor sounds in the same regions predicted activity during test blocks, when ambiguous auditory stimuli were presented in isolation. Furthermore, Kilian-Hütten, Valente, Vroomen, and Formisano (2011) applied multivariate pattern analysis to show that unique patterns of auditory cortex activity reflected the syllable percept (/aba/ and /ada/) for the same acoustic stimulus presented during the test phase.

Similarly, the lexical or Ganong effect has been associated with activity across left and right STG as well as frontal and parietal regions (Myers & Blumstein, 2008). Lexically driven perceptual learning appears to initially depend on frontal and middle temporal regions, followed by later activity in left superior temporal areas when listeners perceive tokens along a continuum of /g/–/k/ whose shift is mediated by exposure to lexical stimuli containing an ambiguous /g/–/k/ (Myers & Mesite, 2014).

Although studies on lexical and audiovisual recalibration have thus indicated involvement of similar brain areas, prior studies did not directly compare the neural underpinnings of the two phenomena. The recalibration or perceptual retuning paradigm allows for the use of the same stimuli during test blocks with either lexical or audiovisual exposure. The ambiguous phoneme blends, to be perceived differently depending on the prior exposure block, can consist of either edited words or videos. The exposure time can also be matched; although lexical retuning studies typically use longer exposure phases to induce a bias, such retuning can take place in shorter time spans and can be observed in shorter test blocks, similar to the typical audiovisual exposure, as well (Ullas et al., 2020a, 2020b; van Linden & Vroomen, 2007).

In this study, lexical and audiovisual recalibration were compared using fMRI, to determine the similarity between the underlying brain regions involved in the two processes using similar testing procedures. As noted above, the existing behavioral studies of audiovisual recalibration and lexical retuning have tended to differ in the amount of exposure time used to induce effects, but they have also differed in the constancy of the bias. Thus, the long exposure phases in lexical retuning have usually served to induce a bias toward only a single phoneme; in contrast, audiovisual recalibration studies have not only used shorter blocks (e.g., eight stimuli) but have also induced a changing phoneme bias throughout the experiment (e.g., Eisner & McQueen, 2006; Vroomen

et al., 2004). This study maintained consistency between the two procedures by using exposure blocks of the same length for both types of stimuli and also allowing the phoneme bias to vary for both. Ambiguous phonemes were presented in identical test blocks, and participants indicated their percept to assess recalibration effects in the same way for each exposure type. This approach of alternating exposure (containing either audiovisual or lexical stimuli, with changing phoneme biases) and test blocks has been shown to be effective in producing both audiovisual recalibration and lexical retuning (see Ullas et al., 2020a, for more details regarding the behavioral outcomes of this approach). By utilizing this procedure, the study aimed to identify the neural commonalities between lexical and audiovisual recalibration under similar experimental constraints, as well as potential unique contributions from multimodal or visual regions for audiovisual recalibration, in contrast to activity within areas of the language network for lexical retuning.

As these two processes likely involve similar cortical areas, we made use of ultra-high-field MRI at 7 T, which provided increased sensitivity in detecting possible differences. Although audiovisual and lexical recalibration have been shown to involve highly similar areas across the temporal cortex as well as parietal, motor, and insular areas, audiovisual recalibration seems, in previous studies, to have been influenced by visual cortex activity as well. For both lexical retuning and audiovisual recalibration, we investigated whether activity within ROIs (in temporal, occipital, inferior-parietal, and insular regions), defined by activity during exposure, could distinguish test blocks with high and low adaptation effects, with higher activation associated with higher behavioral scores.

METHODS

Participants

Twelve participants (nine women, three men) were recruited from Maastricht University to take part in the study (data from one participant were not analyzed because of excessive motion leading to poor-quality MRI data). All participants had normal or corrected-to-normal vision and normal hearing. Participant age range was 21.7–27.3 years (mean age = 24.5 years). Participants gave written informed consent to be scanned and to have their data shared.

Stimuli

The stimulus sets contained a combination of exposure and test stimuli, where exposure stimuli were designed to induce a bias toward a particular phoneme using either lexical or audiovisual (lipreading) information, whereas test stimuli were ambiguous phonemes presented without context, to which listeners could report what phoneme they heard. If recalibration/retuning were successful, responses to test stimuli would be in line with the phoneme bias contained in the prior exposure block (i.e., more perceived /p/ after /p/-biased exposure). Exposure stimuli consisted of audio recordings of words and audio–video recordings of pseudowords, to measure lexical retuning and audiovisual recalibration, respectively. Pseudowords were used to isolate the influence of audiovisual cues without any additional confounds, while also retaining the speech-like structure. All stimuli had the clear portion of the critical phoneme removed (either /op/ or /ot/) and replaced with an ambiguous /op/–/ot/ blend, which was individually chosen from a 10-step /op/–/ot/ continuum.

For lexical stimuli, 16 Dutch words with eight /op/ and eight /ot/ endings were chosen. Most words did not contain any acoustically similar phonemes (i.e., /b/ or /d/) so as to limit retuning effects to the critical phonemes only. Importantly, words were chosen such that only one of the two critical phonemes in the final position could form a word (i.e., *siroop* is a word, but *siroot* is not). There were 4 two-syllable words, 3 three-syllable words, and one monosyllabic word ending in /op/ and /ot/. All stimuli are listed in Table 1.

For audiovisual stimuli, 16 pseudowords were created using WinWordGen (Duyck, Desmet, Verbeke, & Brysbaert,

Table 1. Stimuli Used in the Study, with Corresponding IPA Transcriptions

a. /op/ Words		*b. /ot/ Words*	
Hoop	[hoʊp]	Vloot	[vloʊt]
Siroop	[sɪˈʀoʊp]	Afsloot	[ˈɑfsloʊt]
Aanloop	[ˈaːnloʊp]	Vennoot	[vɛˈnoʊt]
Afkoop	[ˈɑfkoʊp]	Vergroot	[vəʀˈɣʀoʊt]
Wanhoop	[ˈʋɑnhoʊp]	Walnoot	[ˈʋaːlnoʊt]
Geweerloop	[ɣəˈʋeːrˌloʊp]	Hazelnoot	[ˈɦɑzəlnoʊt]
Horoscoop	[ɦɔʀɔˈscoʊp]	Levensgroot	[ˈlɛvənsɣʀoʊt]
Kussensloop	[ˈkʏsənsloʊp]	Middenmoot	[ˈmɪdənmoʊt]

c. /op/ Pseudowords		*d. /ot/ Pseudowords*	
Smoop	[smoʊp]	Vroot	[vʀoʊt]
Aaroop	[ˈaːʀoʊp]	Faloot	[fɑˈloʊt]
Miloop	[ˈmɪloʊp]	Geroot	[ɣəˈʀoʊt]
Onsoop	[ˈɔnsoʊp]	Mevoot	[məˈvoʊt]
Weloop	[ʋəˈloʊp]	Neuloot	[ˈnøːloʊt]
Acenkoop	[ˈɑsəŋkoʊp]	Frieseloot	[ˈfʀisəloʊt]
Lakeroop	[ˈlɑkəʀoʊp]	Leuveroot	[ˈløːvəʀoʊt]
Senkenloop	[ˈsɛŋkənloʊp]	Sanekoot	[ˈsɑnəkoʊt]

IPA = International Phonetic Alphabet.

2004). Pseudowords were matched with words for number of syllables, and lip movements of the speaker indicated /op/ or /ot/ endings, with eight of each.

All stimuli were recorded by a female native Dutch speaker in a sound-attenuated booth. Words and pseudowords were all recorded with both /op/ and /ot/ endings. In addition, *soop* and *soot* (not words in Dutch) were recorded to create an /op/–/ot/ continuum. Video recordings were centered around the speaker's mouth to highlight lip movements during audiovisual exposure.

A continuum of /op/–/ot/ was created, using the *soop* and *soot* recordings, with the speech editing program Praat (Boersma & Heuven, 2001). The final portion of /op/ and /ot/ was each extracted, equated in duration at 44-kHz sampling frequency, and original pitch contours were replaced with the average (at about 230 Hz), similar to previous morphing procedures (van der Zande, Jesse, & Cutler, 2014; Mitterer, Scharenborg, & McQueen, 2013). Consonant bursts and vowel durations of the /op/ and /ot/ tokens were scaled to the same peak amplitude and equated in duration (to 50 msec for the vowel) and then blended together in 10% increments for each step of the continuum. The morphed /op/–/ot/ blends were spliced back onto the /s/ token of *soop*/*soot* for the pretest and test block stimuli. Lexical and audiovisual exposure stimuli were created by splicing these blends at the zero crossing closest to the last 50 msec of the vowel, to reduce potential effects of coarticulatory cues from the preceding vowel. For audiovisual stimuli, the edited pseudowords replaced the audio of the original video recordings, so that the lip movements of the final phoneme /p/ or /t/ were aligned with the ambiguous auditory phoneme. Multiple stimulus sets were created to be able to present listeners with the stimuli containing the phoneme blend perceived to be most ambiguous, on an individual basis.

Behavioral Procedure

During each functional run of the MRI scanning session, participants performed a discrimination task on individually selected, phonetically ambiguous blends. Before the start of the experiment, all participants underwent a pretest to determine the sound along the /op/–/ot/ continuum they perceived to be most ambiguous and to select the most appropriate stimulus set containing this token. The pretest was conducted while participants were already placed in the scanner and using the MRI-compatible earphones, so that participants could become accustomed to the MR environment, sound presentation, and stimuli as closely as possible to the actual scanning session. Participants heard each sound on the continuum a minimum of six times, with sounds at the middle of the continuum presented more often (six times for Steps 1, 2, 9, and 10; eight times for Steps 3 and 8; 12 times for Steps 4–7). For each sound, participants responded with a button press to report whether they heard /op/ or /ot/.

The experimental design was adapted from a similar previous study by van Linden and Vroomen (2007). Stimuli were presented using Presentation software (Version 18.2; NeuroBehavioral Systems). Lexical retuning and audiovisual recalibration were induced in a blocked, counterbalanced design. Each run consisted of eight exposure-test rounds, four rounds of inducing and testing audiovisual recalibration and four rounds of lexical recalibration. In each run, four blocks of audiovisual recalibration were followed by four blocks of lexical recalibration or vice versa. Half of the exposure blocks were biased toward /p/, and the other half were biased toward /t/, so that each run contained two audiovisual /p/-blocks, two audiovisual /t/-blocks, two lexical /p/-blocks, and two lexical /t/-blocks. The phoneme bias of the exposure block alternated every two blocks. Although this procedure can successfully result in both audiovisual and lexical retuning effects, audiovisual cues, compared to lexical, can lead to larger effects (Ullas et al., 2020a; Figure 1).

In an exposure block, eight stimuli were presented with either /p/- or /t/-final bias, indicated by the lip movements of the speaker in the audiovisual version or by the phoneme the word would typically end in for lexical blocks. Four unique items were each presented twice, but the same item was not presented twice in succession. After each exposure block was a test block, containing six stimuli reflecting the most ambiguous token from the /op/–/ot/ continuum and its two neighbors, each presented twice and in random order. Participants were instructed to respond during test blocks for each stimulus with a button press on a button box as soon as the stimuli ended, signaling whether they heard /op/ or /ot/.

MRI Data Acquisition

Participants were scanned in a Siemens 7-T MRI scanner (Siemens Medical Systems) with a head coil (Nova Medical) at the Maastricht Brain Imaging Center. Stimuli were presented binaurally through Sensimetrics MR-compatible earphones (Sensimetrics S14, Sensimetrics Corporation) and played at a comfortable listening volume during silent gaps introduced within image acquisition (see below). Anatomical scans were acquired using a T1-weighted magnetization prepared rapid gradient echo sequence at a 0.6-mm resolution, as well as a proton density image for inhomogeneity correction (echo time = 2.52 msec, repetition time [TR] = 3100 msec, 192 slices). Functional scans were obtained using gradient echo sequence with Multiband 3 and GRAPPA 3 acceleration factor at 1.2-mm isotropic resolution. Eighty-one slices were collected per volume, with a 3000-msec TR (silent gap for sound presentation: 1500 msec, acquisition time = 1500 msec, echo time = 19 msec, field of view = 229 × 229 mm), and 200 volumes per run. Five 10-min runs were completed per participant. Two additional five-volume

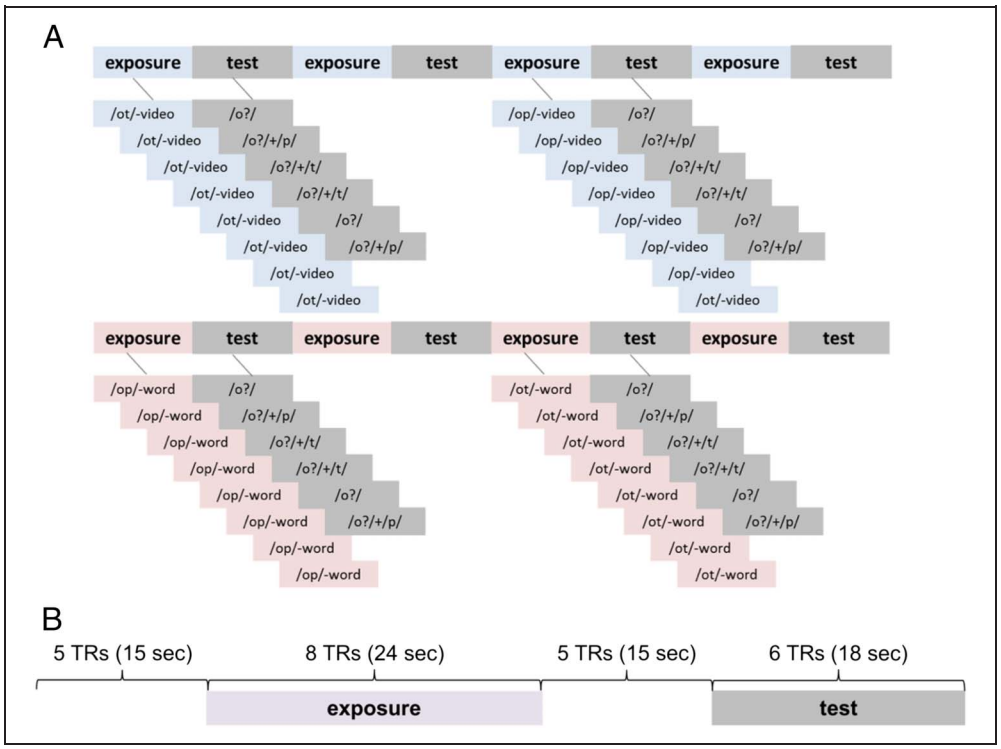

Figure 1. Sample scheme of a run (A). Half of the exposure blocks contained audiovisual stimuli, with half of those containing a bias toward /op/ or /ot/, and the same for the lexical blocks. The same test block followed every exposure block, with the most ambiguous token from the continuum selected from the pretest and its two neighbors, each presented twice. Participants were prompted to indicate by button press after every test item whether they heard /op/ or /ot/. Timings of exposure and test blocks are shown in B; 15-sec gaps, or five TRs, were given between exposure and test blocks. Exposure and test items were presented within the silent gap of each TR.

runs with opposite phase-encoding directions (anterior–posterior and posterior–anterior) were collected for EPI distortion correction.

MRI Data Preprocessing

MRI and fMRI data were preprocessed using BrainVoyager QX v2.8 (BrainInnovation). Anatomical T1 images were scaled using a proton density image to remove distortions. All images were transformed into Talairach space (Talairach & Tournoux, 1988) and interpolated to create 0.5-mm anatomical and 1-mm functional images. Motion correction and slice time correction were performed on all functional runs. To correct for EPI distortions, the data were corrected using the COPE plugin in BrainVoyager (Version 0.5, support.brainvoyager.com/documents/Available_Tools/Available_Plugins/Cope/CopePluginHelp/index.html) and the five-volume "anterior–posterior and posterior–anterior" runs. Additional preprocessing steps included spatial smoothing (8-mm FWHM) as well as temporal high-pass filtering (11 cycles per run) and linear trend removal. Gray-matter and white-matter segmentations were used for surface creation, and functional data were projected onto vertices of the resulting cortical sheet.

MRI Data Analysis

Functional data were analyzed using a random effects general linear model (GLM) including all runs of all participants with separate subject predictors, by convolving the time course of each condition with a hemodynamic response function. Here, predictors reflected six experimental conditions, with audiovisual and lexical exposure, high and low audiovisual test, and high and low lexical test, and additionally, included a baseline predictor for each run. The terms "high" and "low" are meant to distinguish the grouping between greater or fewer bias-consistent responses or responses in accordance with the preceding exposure block. Test blocks were defined as high or low based on behavioral performance, but the median number of bias-consistent responses (in the same direction as the bias of the prior exposure block, i.e., /p/ responses after a /p/-biased block) differed between lexical and audiovisual test blocks. For audiovisual recalibration (median = 4, range = 1), if the participant responded with four or more bias-consistent responses, then this was defined as a high recalibration test block, whereas blocks with fewer than four were defined as low recalibration test blocks. For lexical retuning (median = 3, range = 1), behavioral performance overall indicated a lower median of performance; therefore, three or more bias-consistent responses were categorized as high test blocks, and fewer than three were categorized as low test blocks.

In addition to vertex-wise analyses, we conducted an ROI analysis to examine whether average activity within specific regions could distinguish high versus low recalibration test blocks. ROIs were defined based on individual fixed effects GLMs using the activity during exposure phases. This produced five regions per participant in

auditory, parietal, insula, motor, and (for audiovisual only) visual cortices in both hemispheres. A contrast between high and low recalibration during the respective test blocks (i.e., audiovisual high vs. low recalibration in regions defined by audiovisual exposure) was conducted for each ROI. Paired *t* tests were performed on individual beta estimates reflecting activity during high and low recalibration test blocks within these ROIs.

RESULTS

Behavioral

Pretest responses on the 10-step continuum ranging from /op/ to /ot/ revealed that the sixth step was perceived to be most ambiguous on average (Figure 2).

Responses during test blocks were entered into a generalized linear mixed model with a logistic link using the *lmer* package in R (Version 3.4.1). The factors Phoneme bias during the exposure block, the type of exposure stimuli (lexical or audiovisual, as Condition), and the three test Sounds presented during the test blocks, as well as the phoneme bias of the prior exposure block to account for potential Carryover in effects (where the phoneme bias could be the same as the previous block or different), were entered as fixed effects into the model, and each individual participant was included as a random effect. Interactions were only modeled between the fixed effects variables. All variables were coded to be centered around 0, whereas responses during the test blocks were coded as 0 for /p/ and 1 for /t/. For model selection, the fitting was first performed for a full model including all possible main effects and interactions and followed by fitting of sparser models by iteratively removing slopes of random effects until the model converged and all fixed effects correlations were sufficiently low (less than 0.2). Results are shown in Table 2.

Model results showed a significant intercept, indicating a general tendency to respond with /p/ across all blocks, regardless of other factors. Main effects of Phoneme bias, Sound, and Condition were found to be significant. Phoneme bias was most significant ($p < .0001$), where more /t/ responses were found after /t/-biased exposure blocks than for /p/-biased exposure blocks, indicating successful recalibration with effects in the expected direction. Sound was also found to be significant, where more /t/ responses were observed for the more /t/-sounding test stimuli. Carryover was not found to be significant, so the order of the phoneme bias (which alternated every two blocks and could have led to potential buildup in recalibration effects as a result) did not appear to have any effect on the responses. The main effect of Condition ($p < .001$) indicated that participants showed a stronger response bias toward /t/ across all lexical test blocks than across audiovisual test blocks. Pairwise contrasts were performed for Phoneme bias and Condition, and the difference in amounts of /t/ responses between /t/- and /p/-biased blocks was larger in the audiovisual condition ($p < .0001$) compared to the lexical condition, where the difference was smaller ($p < .05$). Behavioral results are displayed in Figure 3.

fMRI Results

GLM Results

Group GLM results were projected onto a group-averaged brain, created using cortex-based alignment (Goebel, Esposito, & Formisano, 2006). First, contrasts between audiovisual and lexical exposure blocks versus baseline were performed (Figure 4A and C). In addition, contrasts between test blocks after audiovisual or lexical exposure, compared to baseline, were conducted (Figure 4B and D). To identify areas of overlap of conditions, conjunction maps between audiovisual and lexical exposure and between audiovisual and lexical test were also created (Figure 5). All maps were corrected for multiple comparisons by cluster-size threshold ($p_{corr} = .05$), with an initial vertex-wise threshold of $p = .01$.

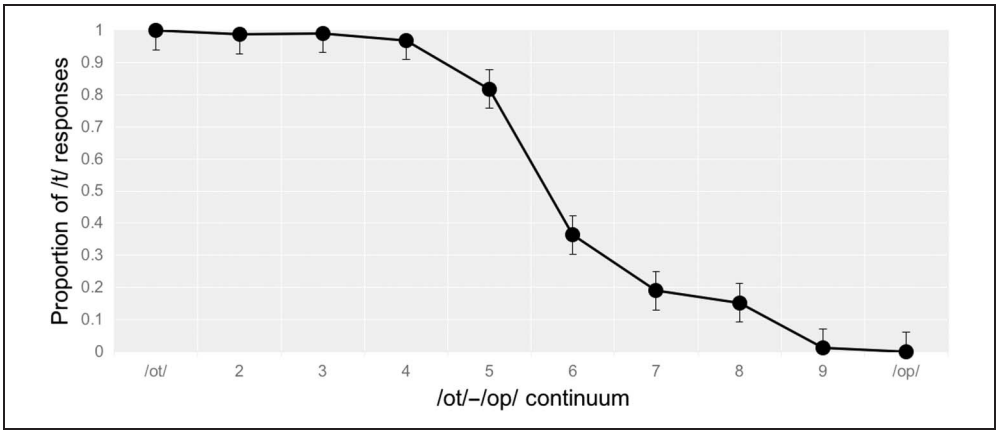

Figure 2. Pretest responses. Responses to each of the 10 steps of the /op/–/ot/ continuum averaged across participants, with error bars indicating standard error.

Table 2. Model Results

	Estimate	*Std. Error*	*z Value*	*Pr(>|z|)*
(Intercept)	−0.24712	0.09087	−2.72	0.00654**
Phoneme bias	0.46576	0.09643	4.83	1.37E-06***
Condition	−0.28243	0.1131	−2.497	0.01252*
Sound	0.55275	0.20655	2.676	0.00745**
Carryover	−0.12575	0.09848	−1.277	0.20162
Phoneme Bias × Condition	0.41196	0.18053	2.282	0.02249*
Phoneme Bias × Sound	−0.08425	0.11304	−0.745	0.45612
Condition × Sound	−0.04496	0.11314	−0.397	0.69107
Phoneme Bias × Condition × Sound	0.10117	0.22583	0.448	0.65416

Model: Response ~ Phoneme Bias × Condition × Sound + Carryover (1 + Phoneme Bias × Condition + Sound + Carryover || Subject).

*$p < 0.05$.

**$p < .01$.

***$p < .001$.

Cluster-size threshold correction was performed with Monte Carlo simulations to estimate the false-positive rates at the cluster level (Goebel et al., 2006).

During audiovisual exposure blocks, significant bilateral engagement was observed in the temporal cortex; in HG, PT, and STG/STS; and in the occipital cortex between V1 and V2 as well as in IFG, insula, IPL, and postcentral gyrus in the left hemisphere and in an occipitotemporal cluster in the right hemisphere (Figure 4A). During lexical exposure blocks, bilateral activation of HG, STG/STS, and insula was found, whereas postcentral gyrus/central sulcus, planum polare (PP), PT, and IPL were also active in the left hemisphere (Figure 4B). Similarly, during test blocks after audiovisual exposure, significant activation was observed bilaterally in HG/Heschl's sulcus, PP, and STG/STS; in insula; and between V1 and V2 as well. IPL and postcentral gyrus/central sulcus were also activated in the left hemisphere (Figure 4B). For test blocks after lexical exposure, significant activation was found across bilateral HG, STG, PT, and insula as well as postcentral

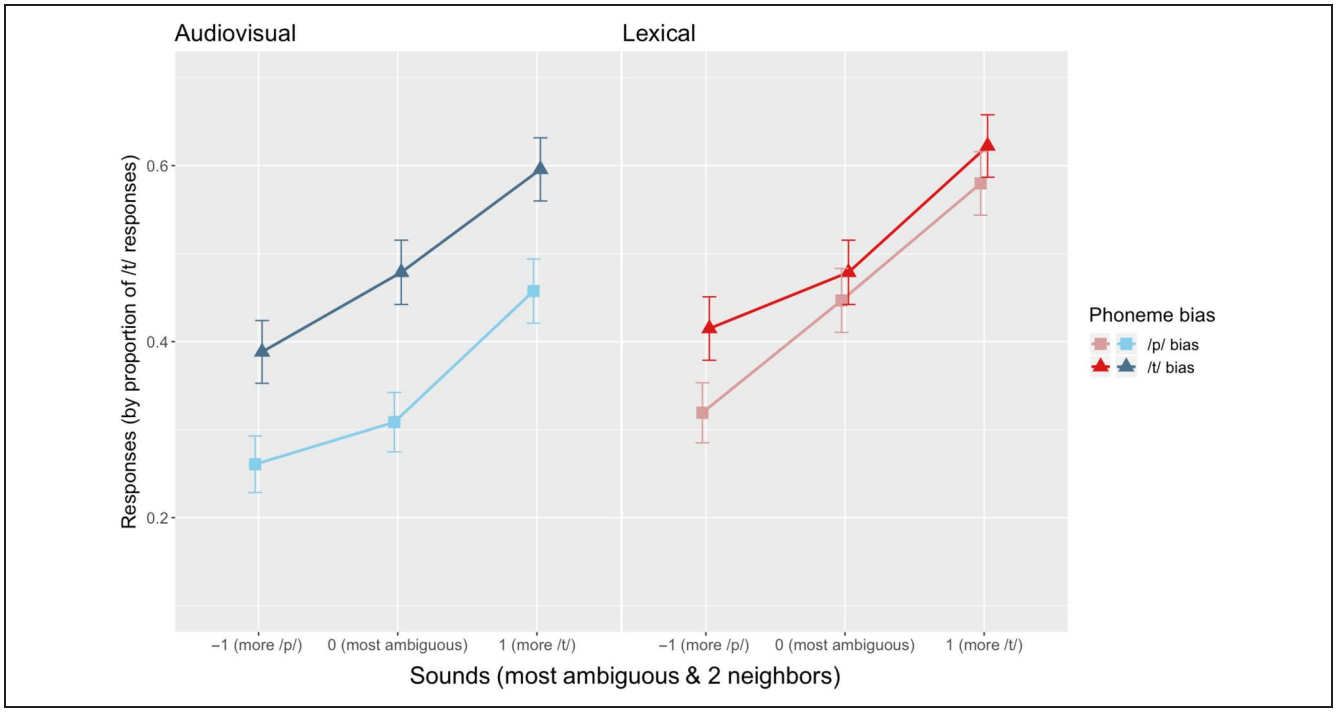

Figure 3. Behavioral results split by type of exposure in the preceding block (lexical and audiovisual), across the three test sounds, with error bars indicating standard error.

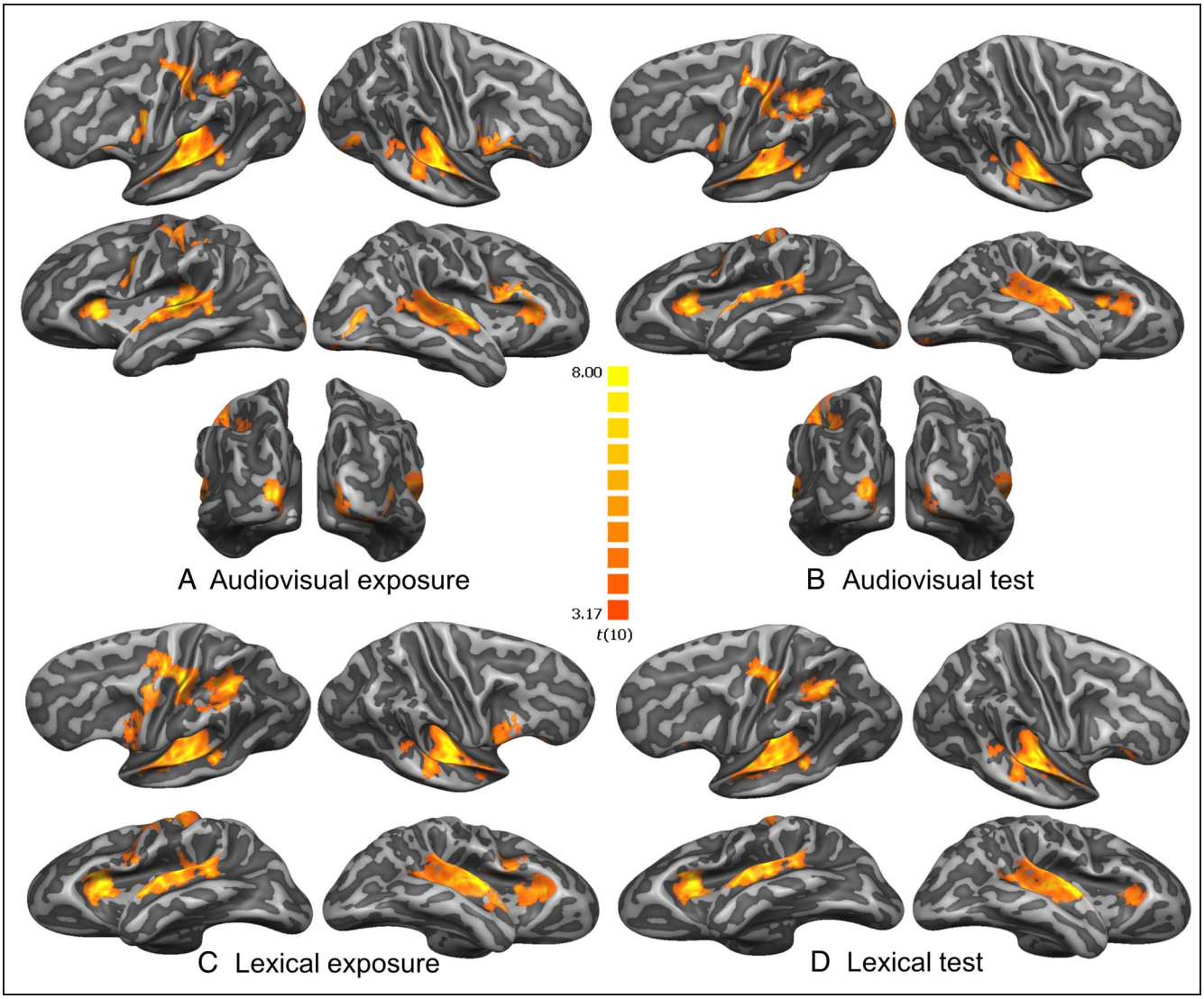

Figure 4. Audiovisual exposure (A), audiovisual test (B), lexical exposure (C), and lexical test (D) blocks versus baseline, with $t(10) > 3.17$, $p < .01$.

gyrus/central sulcus, IPL, and PP in the left hemisphere (Figure 4D). Activations during both exposure types (Figure 5A) and both tests (Figure 5B) were observed consistently in many of the same areas. Table 3 contains a list of all active regions and their respective coordinates (in Talairach space).

ROI Analysis

For the analysis of ROIs (Figure 6A), defined based on activity during exposure blocks, significant differences between high and low recalibration test blocks were found for audiovisual recalibration but not for lexical retuning. As described in the Methods, test blocks were split into high and low based on the median number of bias-consistent responses per condition, which on average resulted in 8.061 audiovisual low blocks ($SD = 2.833$) and 9.129 lexical low blocks ($SD = 2.927$), as well as 11.939 audiovisual high blocks ($SD = 2.561$) and 10.871 lexical high blocks ($SD = 2.771$) per participant. In addition, the positioning of high blocks was calculated to see whether high recalibration blocks may have been in positions where the phoneme bias of the previous exposure block could have had any effect on the recalibration, as the phoneme bias changed every two blocks. For example, if a /p/-biased block was followed by another /p/-biased block, we verified whether the second /p/-block may have potentially led to higher recalibration because of buildup and if all of the high blocks were confounded by this. Of the two possible positions (the first being a change in phoneme bias versus the second being the same phoneme bias as the previous exposure), 67.78% of the first-position blocks were high blocks and 70% of the second-position blocks were high blocks for the audiovisual condition ($p = .344$, paired t test, two-tailed). For the lexical condition, 45.56% of the first-position blocks and 51.11% of the second-position blocks were categorized as high blocks ($p = .179$, paired t test, two-tailed). We concluded that

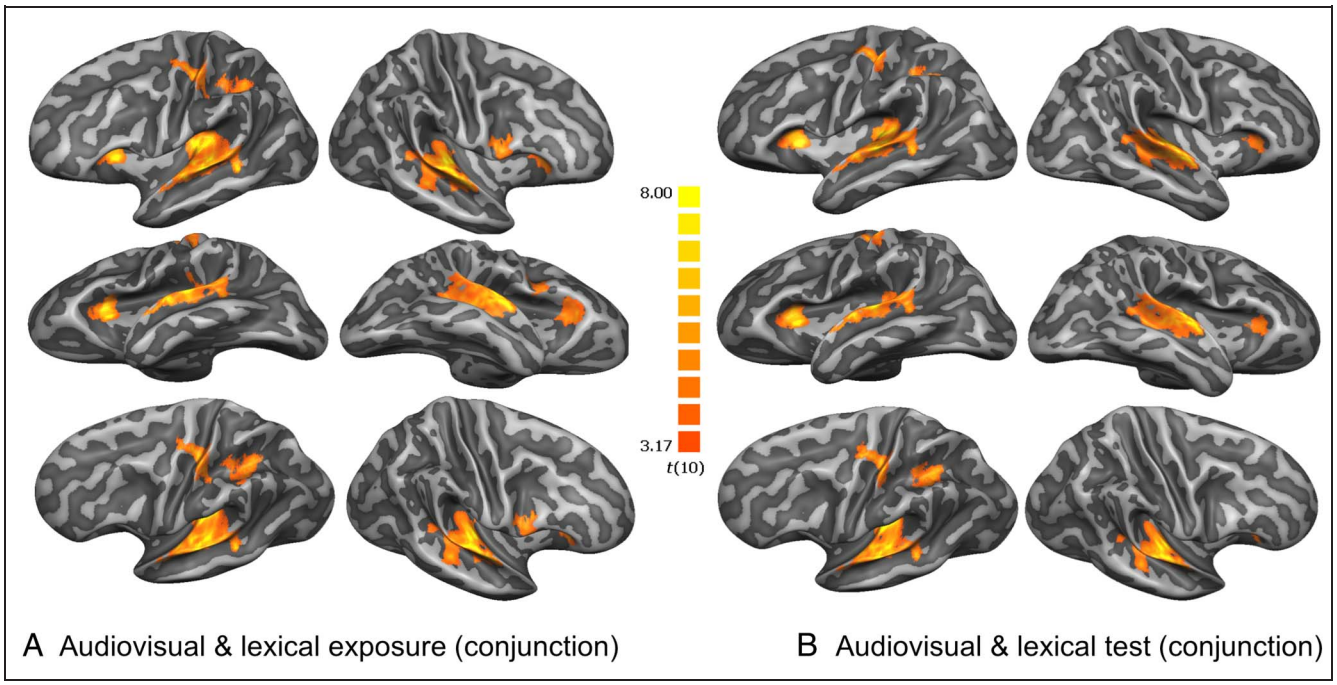

Figure 5. Conjunction maps between audiovisual and lexical exposure (A) between and audiovisual and lexical test (B), with $t(10) > 3.17$, $p < .01$.

there was no significant evidence that high recalibration blocks were confounded by the order of the phoneme biases in the exposures.

In ROIs defined by audiovisual exposure, temporal, insular, and motor (central sulcus) regions as well as STG in the left hemisphere showed a significant difference between high versus low test blocks, whereas insular and parietal clusters showed the same difference in the right hemisphere (Figure 6B). The contrast was also significant for both the left and right occipital ROIs.

Table 3. List of Active Regions during Exposure and Test (as Shown in Figure 4)

	Peak Vertex			
	x	*y*	*z*	*Number of Vertices*
Left hemisphere regions				
Temporal (HG, PT, PP, STG/STS)	−46	−25	6	6340
Frontal (IFG)	−45	3	22	2325
Insula	−27	17	7	1083
Motor (precentral/postcentral gyrus, central sulcus)	−33	−24	44	2258
Occipital (V1/V2)	−12	−90	2	920
Parietal (IPL)	−32	−44	35	2221
Right hemisphere regions				
Temporal (HG, PT, PP, STG/STS)	54	−18	9	5128
Frontal (IFG)	45	5	16	742
Insula	30	24	10	972
Occipital (V1/V2)	10	−85	13	920
Occipito-temporal (BA 19/V3)	39	−67	7	319

All active regions are listed by hemisphere, with average Talairach coordinates of the peak vertex and the average number of contiguous vertices per region, across participants.

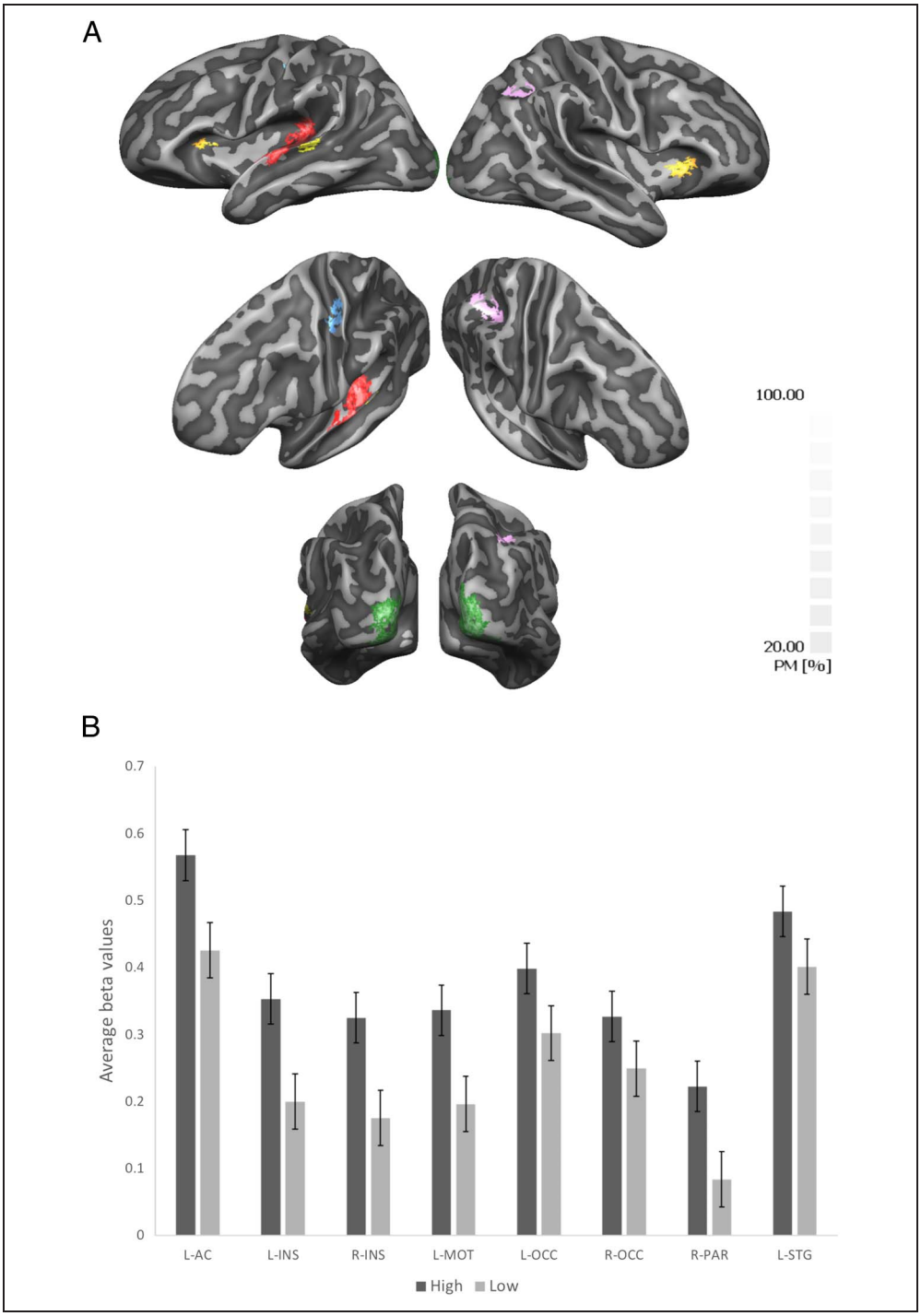

Figure 6. Significant ROIs for high versus low audiovisual recalibration. (A) Probabilistic maps (PM) are shown. Color shadings denote regions with an overlap of at least three participants showing a significant difference ($p < .01$) between high and low audiovisual recalibration. (B) Average beta values by regions, for high and low audiovisual recalibration blocks. Significant differences between high and low blocks were found within temporal/auditory cortex (AC; left), occipital (OCC)/visual cortex (left and right), insula (INS; left and right), motor (MOT; left), parietal (PAR; right) clusters, and STG (left). High recalibration referred to blocks with four or more bias-consistent responses, or responses that were in the same direction as the preceding exposure block (i.e., /p/ responses after /p/-biased exposure), whereas low recalibration included blocks with zero to three bias-consistent responses. High versus low blocks per region were significant at $p < .05$. Error bars indicate standard error. L = left; R = right.

DISCUSSION

Phoneme category recalibration or retuning refers to a process that is an essential part of the celebrated robustness of human speech perception. Listeners can draw on information other than the acoustic signal—lip movements or lexical/semantic knowledge—to adjust boundaries between speech sound categories so that they fit the speech input they are currently hearing, which enables them to adapt to pronunciations they have perhaps never heard. Behavioral evidence (Ullas et al., 2020b) suggests that, despite the apparent similarity, these two adaptation processes may have distinct triggers (coping with noise in the case of audiovisual recalibration and coping with talker novelty in the case of lexical retuning), although both types of adaptation often occur conjointly in real life. In this study, fMRI data were collected as participants underwent both forms of phoneme category adjustments, using lexical and audiovisual cues, respectively, in a counterbalanced, blocked design. The

perceptual boundary between two phonemes, /p/ and /t/, was systematically shifted, using lexical and audiovisual cues, toward either /p/ or /t/. Note that the behavioral results had shown that this procedure resulted in significant effects in both conditions and toward both phonemes, although audiovisual recalibration effects were larger than lexical retuning, in line with previous findings as well (Ullas et al., 2020a; van Linden & Vroomen, 2007).

The analysis of concurrent fMRI measurements showed similarities between audiovisual and lexical exposure blocks, particularly in the temporal cortex across bilateral HG, STG/STS, and PT as well as left IPL and right insula. HG and PT are most likely responsible for acoustic and rudimentary phonetic processing (Obleser & Eisner, 2009; Binder, 2000), whereas nearby STG and STS are likely to represent similar items such as syllables and phonemes (Yi, Leonard, & Chang, 2019; Mesgarani et al., 2008; Jäncke et al., 2002), although they may show overlap in their functions.

Outside the lower-level perceptual areas, insula and IPL activity was also evoked during the audiovisual and lexical exposure blocks. The insula has been proposed to be a part of the articulatory network (Hickok & Poeppel, 2007). Oh, Duerden, and Pang (2014) suggest that the insula also oversees articulation, and other motor-like properties of speech, and is connected to other speech and language regions, including Broca's area. IPL activity may be related to processing audiovisual speech as well as words and pseudowords (Ojanen et al., 2005; Newman & Tweig, 2001). Some areas were uniquely engaged by audiovisual exposure, in the occipital cortex over V1 and V2, whereas lexical exposure was not associated with any unique brain areas. Naturally, the presentation of visual stimuli during the audiovisual blocks elicited activity within the visual/occipital cortex, unlike the lexical blocks where no visual stimuli were presented.

Similar patterns of activation were identified during test blocks after audiovisual and lexical exposure in the temporal cortex, again within HG, STG, and STS. As previously mentioned, these regions are responsible for representing phonemes, syllables, and low-level acoustic information. Activation in these early auditory regions has also been found to undergo top–down modulation by attention to task-relevant acoustic information, such as spectral or temporal features (Rutten et al., 2019). In addition to these functions, Myers and Mesite (2014) reported STG and MTG activity to be strongest for ambiguous items that had been perceptually shifted by exposure to lexical items. Kilian-Hütten, Vroomen, et al. (2011) similarly noted STG as well as IPL, insula, and inferior frontal sulcus to be activated during audiovisual recalibration and that IPL can coordinate higher-order constructive processes in perception. Regions in the parietal lobe may also be involved in detecting phonological changes, distinguishing words from pseudowords, and general linguistic comprehension (Obleser & Eisner, 2009; Newman & Tweig, 2001; Binder et al., 1997). Similarly, the insula can assist in disambiguating degraded speech (Erb, Henry, Eisner, & Obleser, 2013). IPL and insula activation have been reported to underlie text-based recalibration as well (Bonte, Correia, Keetels, Vroomen, & Formisano, 2017). As IPL and insula lie outside the core speech network, they may also be involved in less tangible functions, such as processing abstract linguistic information or multimodal integration (Guediche, Blumstein, Fiez, & Holt, 2014; Dick et al., 2010; Jones & Callan, 2003). The convergence of these regions in this study, as well as the left–right asymmetry we observed in activation strength, consistently align with previous studies of speech perception and retuning/recalibration. In addition, as expected from that prior work, audiovisual cues led to stronger effects than lexical cues.

Although additional activation was also elicited in postcentral gyrus and central sulcus for lexical and audiovisual test blocks, this most likely reflects activity related to the expected button presses. Therefore, it appears unlikely that the activity observed in these regions represents any functions beyond the button presses made during the test blocks; however, motor cortex activity may be reflective of gestural or articulatory movements triggered by speech sounds (Hickok & Poeppel, 2007) and may ease the interpretation of ambiguous speech sounds (Guediche et al., 2014).

Both forms of perceptual learning showed a pattern of reactivation, where many of the same regions active during the exposure blocks were also active during the test blocks, despite the differences in stimuli and task between exposure and test blocks. Namely, this overlap was observed in HG, STG/STS, and left IPL for both audiovisual and lexical test blocks. Both exposure and test blocks evoked activity in the speech network as a result of the presentation of speech (and speech-like) sounds. Most notably, however, the occipital cortex remained active during audiovisual test blocks, although no visual stimuli were presented and a sufficient amount of time was given between exposure and test blocks to allow the BOLD response to return to baseline. The sustained activation in visual cortex suggests that the visual information from the exposure blocks is salient enough to be retained during the subsequent test block, possibly as a form of mental imagery or within an STM loop, as early visual areas are capable of contributing to visual mental imagery (Sparing et al., 2002; Kosslyn, Ganis, & Thompson, 2001). Associative learning may entail involuntary visual learning, or when an association is formed between two stimuli, and can take place within early visual areas such as V1 and V2 (Pearson, 2019). In this study, listeners may thus have formed associations between the ambiguous phonemes and the preceding visual stimuli, with these associations being retrieved and deployed during the test blocks. Kilian-Hütten, Vroomen, et al. (2011) have also noted functional connectivity between occipital regions and left auditory cortex during audiovisual recalibration.

Furthermore, the strong activation of visual cortex during purely auditory test blocks suggests a functional role of visual cortex during audiovisual recalibration and that the auditory cortex does not implement these perceptual shifts on its own.

An ROI analysis revealed a number of regions that were found to be modulated by audiovisual recalibration, including clusters in left temporal, motor, and insular regions, and in right insular and parietal clusters, as well as a larger region spanning V1 and V2. These regions showed a significantly higher hemodynamic activity for test blocks where participants showed larger recalibration effects and a lower activity for less such effect. The relative increase in activity observed during high recalibration blocks points toward more effective identification of the ambiguous sounds, facilitated by top–down contributions from these regions. Activation in a conjunction of both higher- and lower-order regions within and outside the speech network was associated with differences in high and low recalibration performance, which suggests that the process may not be unidirectional, requiring instead a combination of extraction of lower-level acoustic features plus recourse to higher-level semantic and cross-modal representations. The strength of neural activity in these regions seems to be associated with a larger category boundary shift in the same direction as the preceding exposure. Low recalibration blocks appear to be linked with lower levels of activation; however, the relationship between the two is unclear as the underlying cause could reflect a number of factors, such as a lack of attention paid during exposure, the combination of stimuli during exposure not effectively inducing a shift in perception, or fatigue with repeated testing.

The same analysis within the ROIs was not associated with any differences in lexical retuning, corresponding to neither high nor low performance in the test blocks. Participants' generally lower performance during lexical test blocks may have reduced the scope for a significant difference between high- and low-scoring lexical blocks in comparison to the audiovisual test blocks. This might then have translated into the lack of a neural difference as well. In contrast, behavioral audiovisual recalibration effects were larger than lexical, which could have led to higher activation overall compared to lexical test blocks, and thereby increased sensitivity to detecting differences between high and low recalibration within ROIs. Nonetheless, lexical retuning was still elicited under the constraints of the task design (i.e., few exposure items and continuous boundary shifting) and evoked significant patterns of activation across regions known for acoustic–phonetic processing (HG, STG/STS) and higher levels of cognitive engagement (IPL, insula).

Conclusion

This study compared audiovisual recalibration and lexical retuning using high-field fMRI to investigate the underlying similarities and differences in their neural activity. A network of speech-related regions and other higher-order areas emerged as a result of the two forms of perceptual learning, whereas audiovisual recalibration specifically seems to evoke significant visual cortex input during the process, pointing toward a form of involuntary mental imagery, perhaps as a byproduct of associative learning taking place between the visual stimuli and the ambiguous phonemes. In addition, neural activity in several regions spread across the brain was found to be modulated in correspondence with the amount of audiovisual recalibration observed behaviorally. Whereas lexical retuning did not display this pattern across the selected regions, remarkable overlap with audiovisual recalibration was found in temporal, parietal, and insular regions. Evidently, a number of both lower-level regions involved in acoustic–phonetic processing, as well as more complex semantic and cross-modal areas, are involved in these perceptual adjustments. From within and extending beyond the speech network, the strength of the relationship formed between the exposure stimuli and the ambiguous phonemes may therefore be responsible for enabling perceptual shifts. The precise timing and directionality of information processing remain to be investigated; however, our results suggest that not only do recalibration and retuning involve subtly different triggers, but the brain areas responsible for modulating them also involve multiple levels of perceptual organization.

Acknowledgments

This project was supported by Maastricht University, the Netherlands Organization for Scientific Research (NWO) gravitation program Language in Interaction, and a NWO VENI grant (451-17-033 to L. H.).

Reprint requests should be sent to Shruti Ullas, Faculty of Psychology and Neuroscience, Department of Cognitive Neuroscience, Maastricht University, 6200 MD Maastricht, The Netherlands, or via e-mail: shruti.ullas@maastrichtuniversity.nl.

REFERENCES

Beauchamp, M. S. (2005). See me, hear me, touch me: Multisensory integration in lateral occipital–temporal cortex. *Current Opinion in Neurobiology*, *15*, 145–153. **DOI:** https://doi.org/10.1016/j.conb.2005.03.011, **PMID:** 15831395

Bernstein, L. E., & Liebenthal, E. (2014). Neural pathways for visual speech perception. *Frontiers in Neuroscience*, *8*, 386. **DOI:** https://doi.org/10.3389/fnins.2014.00386, **PMID:** 25520611, **PMCID:** PMC4248808

Bertelson, P., Vroomen, J., & De Gelder, B. (2003). Visual recalibration of auditory speech identification: A McGurk aftereffect. *Psychological Science*, *14*, 592–597. **DOI:** https://doi.org/10.1046/j.0956-7976.2003.psci_1470.x, **PMID:** 14629691

Binder, J. R. (2000). Human temporal lobe activation by speech and nonspeech sounds. *Cerebral Cortex*, *10*, 512–528. **DOI:** https://doi.org/10.1093/cercor/10.5.512, **PMID:** 10847601

Binder, J. R., Frost, J. A., Hammeke, T. A., Cox, R. W., Rao, S. M., & Prieto, T. (1997). Human brain language areas identified by functional magnetic resonance imaging. *Journal of*

Neuroscience, *17*, 353–362. **DOI:** https://doi.org/10.1523/JNEUROSCI.17-01-00353.1997, **PMID:** 8987760, **PMCID:** PMC6793702

Boersma, P., & Heuven, V. (2001). Speak and unspeak with PRAAT. *Glot International*, *5*, 341–347.

Bonte, M., Correia, J. M., Keetels, M., Vroomen, J., & Formisano, E. (2017). Reading-induced shifts of perceptual speech representations in auditory cortex. *Scientific Reports*, *7*, 5143. **DOI:** https://doi.org/10.1038/s41598-017-05356-3, **PMID:** 28698606, **PMCID:** PMC5506038

Buchsbaum, B. R., Hickok, G., & Humphries, C. (2001). Role of left posterior superior temporal gyrus in phonological processing for speech perception and production. *Cognitive Science*, *25*, 663–678. **DOI:** https://doi.org/10.1207/s15516709cog2505_2

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, M., & Knight, R. T. (2011). Categorical speech representation in human superior temporal gyrus. *Nature Neuroscience*, *13*, 1428–1432. **DOI:** https://doi.org/10.1038/nn.2641, **PMID:** 20890293, **PMCID:** PMC2967728

Davis, M. H., & Johnsrude, I. S. (2003). Hierarchical processing in spoken language comprehension. *Journal of Neuroscience*, *23*, 3423–3431. **DOI:** https://doi.org/10.1523/JNEUROSCI.23-08-03423.2003, **PMID:** 12716950, **PMCID:** PMC6742313

Dick, A. S., Solodkin, A., & Small, S. L. (2010). Neural development of networks for audiovisual speech comprehension. *Brain and Language*, *114*, 101–114. **DOI:** https://doi.org/10.1016/j.bandl.2009.08.005, **PMID:** 19781755, **PMCID:** PMC2891225

Duyck, W., Desmet, T., Verbeke, L. P. C., & Brysbaert, M. (2004). WordGen: A tool for word selection and nonword generation in Dutch, English, German, and French. *Behavior Research Methods, Instruments, & Computers*, *36*, 488–499. **DOI:** https://doi.org/10.3758/BF03195595, **PMID:** 15641437

Eisner, F., & McQueen, J. M. (2006). Perceptual learning in speech: Stability over time. *Journal of the Acoustical Society of America*, *119*, 1950–1953. **DOI:** https://doi.org/10.1121/1.2178721, **PMID:** 16642808

Erb, J., Henry, M., Eisner, F., & Obleser, J. (2013). The brain dynamics of rapid perceptual adaptation to adverse listening conditions. *Journal of Neuroscience*, *33*, 10688–10697. **DOI:** https://doi.org/10.1523/JNEUROSCI.4596-12.2013, **PMID:** 23804092, **PMCID:** PMC6618499

Formisano, E., De Martino, F., Bonte, M., & Goebel, R. (2008). "Who" is saying "what"? Brain-based decoding of human voice and speech. *Science*, *322*, 970–973. **DOI:** https://doi.org/10.1126/science.1164318, **PMID:** 18988858

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*, 110–125. **DOI:** https://doi.org/10.1037/0096-1523.6.1.110

Goebel, R., Esposito, F., & Formisano, E. (2006). Analysis of functional image analysis contest (FIAC) data with BrainVoyager QX: From single-subject to cortically aligned group general linear model analysis and self-organizing group independent component analysis. *Human Brain Mapping*, *27*, 392–401. **DOI:** https://doi.org/10.1002/hbm.20249, **PMID:** 16596654, **PMCID:** PMC6871277

Guediche, S., Blumstein, S. E., Fiez, J. A., & Holt, L. L. (2014). Speech perception under adverse conditions: Insights from behavioral, computational, and neuroscience research. *Frontiers in Systems Neuroscience*, *7*, 126. **DOI:** https://doi.org/10.3389/fnsys.2013.00126, **PMID:** 24427119, **PMCID:** PMC3879477

Guediche, S., Salvata, C., & Blumstein, S. E. (2013). Temporal cortex reflects effects of sentence context on phonetic processing. *Journal of Cognitive Neuroscience*, *25*, 706–718. **DOI:** https://doi.org/10.1162/jocn_a_00351, **PMID:** 23281778, **PMCID:** PMC3612392

Hagoort, P. (2005). On Broca, brain, and binding: A new framework. *Trends in Cognitive Sciences*, *9*, 416–423. **DOI:** https://doi.org/10.1016/j.tics.2005.07.004, **PMID:** 16054419

Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language. *Cognition*, *92*, 67–99. **DOI:** https://doi.org/10.1016/j.cognition.2003.10.011, **PMID:** 15037127

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*, 393–402. **DOI:** https://doi.org/10.1038/nrn2113, **PMID:** 17431404

Jäncke, L., Wüstenberg, T., Scheich, H., & Heinze, H. J. (2002). Phonetic perception and the temporal cortex. *Neuroimage*, *15*, 733–746. **DOI:** https://doi.org/10.1006/nimg.2001.1027, **PMID:** 11906217

Jesse, A., & Massaro, D. W. (2010). The temporal distribution of information in audiovisual spoken-word identification. *Attention, Perception, & Psychophysics*, *72*, 209–225. **DOI:** https://doi.org/10.3758/APP.72.1.209, **PMID:** 20045890

Jones, J. A., & Callan, D. E. (2003). Brain activity during audiovisual speech perception: An fMRI study of the McGurk effect. *NeuroReport*, *14*, 1129–1133. **DOI:** https://doi.org/10.1097/00001756-200306110-00006, **PMID:** 12821795

Kilian-Hütten, N., Valente, G., Vroomen, J., & Formisano, E. (2011). Auditory cortex encodes the perceptual interpretation of ambiguous sounds. *Journal of Neuroscience*, *31*, 1715–1720. **DOI:** https://doi.org/10.1523/JNEUROSCI.4572-10.2011, **PMID:** 21289180, **PMCID:** PMC6623724

Kilian-Hütten, N., Vroomen, J., & Formisano, E. (2011). Brain activation during audiovisual exposure anticipates future perception of ambiguous speech. *Neuroimage*, *57*, 1601–1607. **DOI:** https://doi.org/10.1016/j.neuroimage.2011.05.043, **PMID:** 21664279

Kosslyn, S. M., Ganis, G., & Thompson, W. L. (2001). Neural foundations of imagery. *Nature Reviews Neuroscience*, *2*, 635–642. **DOI:** https://doi.org/10.1038/35090055, **PMID:** 11533731

Leonard, M. K., & Chang, E. F. (2014). Dynamic speech representations in the human temporal lobe. *Trends in Cognitive Sciences*, *18*, 472–479. **DOI:** https://doi.org/10.1016/j.tics.2014.05.001, **PMID:** 24906217, **PMCID:** PMC4149812

Liebenthal, E., & Bernstein, L. E. (2017). Editorial: Neural mechanisms of perceptual categorization as precursors to speech perception. *Frontiers in Neuroscience*, *11*, 69. **DOI:** https://doi.org/10.3389/fnins.2017.00069, **PMID:** 28261047, **PMCID:** PMC5306389

Liebenthal, E., Binder, J. R., Spitzer, S. M., Possing, E. T., & Medler, D. A. (2005). Neural substrates of phonemic perception. *Cerebral Cortex*, *15*, 1621–1631. **DOI:** https://doi.org/10.1093/cercor/bhi040, **PMID:** 15703256

Lüttke, C. S., Ekman, M., Van Gerven, M. A. J., & De Lange, F. P. (2016). McGurk illusion recalibrates subsequent auditory perception. *Scientific Reports*, *6*, 32891. **DOI:** https://doi.org/10.1038/srep32891, **PMID:** 27611960, **PMCID:** PMC5017187

McGurk, H., & MacDonald, M. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748. **DOI:** https://doi.org/10.1038/264746a0, **PMID:** 1012311

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, *343*, 1006–1010. **DOI:** https://doi.org/10.1126/science.1245994, **PMID:** 24482117, **PMCID:** PMC4350233

Mesgarani, N., David, S. V., Fritz, J. B., & Shamma, S. A. (2008). Phoneme representation and classification in primary auditory cortex. *Journal of the Acoustical Society of America*, *123*, 899–909. **DOI:** https://doi.org/10.1121/1.2816572, **PMID:** 18247893

Mitterer, H., Scharenborg, O., & McQueen, J. M. (2013). Phonological abstraction without phonemes in speech

perception. *Cognition*, *129*, 356–361. **DOI:** https://doi.org/10.1016/j.cognition.2013.07.011, **PMID:** 23973464
Myers, E. B., & Blumstein, S. E. (2008). The neural bases of the lexical effect: An fMRI investigation. *Cerebral Cortex*, *18*, 278–288. **DOI:** https://doi.org/10.1093/cercor/bhm053, **PMID:** 17504782
Myers, E. B., & Mesite, L. M. (2014). Neural systems underlying perceptual adjustment to non-standard speech tokens. *Journal of Memory and Language*, *76*, 80–93. **DOI:** https://doi.org/10.1016/j.jml.2014.06.007, **PMID:** 25092949, **PMCID:** PMC4118215
Newman, S. D., & Tweig, D. (2001). Differences in auditory processing of words and pseudowords: An fMRI study. *Human Brain Mapping*, *14*, 39–47. **DOI:** https://doi.org/10.1002/hbm.1040, **PMID:** 11500989, **PMCID:** PMC6871811
Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*, 204–238. **DOI:** https://doi.org/10.1016/S0010-0285(03)00006-9
Obleser, J., & Eisner, F. (2009). Pre-lexical abstraction of speech in the auditory cortex. *Trends in Cognitive Sciences*, *13*, 14–19. **DOI:** https://doi.org/10.1016/j.tics.2008.09.005, **PMID:** 19070534
Oh, A., Duerden, E. G., & Pang, E. W. (2014). The role of the insula in speech and language processing. *Brain and Language*, *135*, 96–103. **DOI:** https://doi.org/10.1016/j.bandl.2014.06.003, **PMID:** 25016092, **PMCID:** PMC4885738
Ojanen, V., Möttönen, R., Pekkola, J., Jääskeläinen, I. P., Joensuu, R., Autti, T., et al. (2005). Processing of audiovisual speech in Broca's area. *Neuroimage*, *25*, 333–338. **DOI:** https://doi.org/10.1016/j.neuroimage.2004.12.001, **PMID:** 15784412
Pearson, J. (2019). The human imagination: The cognitive neuroscience of visual mental imagery. *Nature Reviews Neuroscience*, *20*, 624–634. **DOI:** https://doi.org/10.1038/s41583-019-0202-9, **PMID:** 31384033
Poldrack, R. A., Wagner, A. D., Prull, M. W., Desmond, J. E., Glover, G. H., & Gabrieli, J. D. (1999). Functional specialization for semantic and phonological processing in the left inferior prefrontal cortex. *Neuroimage*, *10*, 15–35. **DOI:** https://doi.org/10.1006/nimg.1999.0441, **PMID:** 10385578
Rauschecker, J. P., & Scott, S. K. (2009). Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing. *Nature Neuroscience*, *12*, 718–724. **DOI:** https://doi.org/10.1038/nn.2331, **PMID:** 19471271, **PMCID:** PMC2846110
Rutten, S., Santoro, R., Hervais-Adelman, A., Formisano, E., & Golestani, N. (2019). Cortical encoding of speech enhances task-relevant acoustic information. *Nature Human Behaviour*, *3*, 974–987. **DOI:** https://doi.org/10.1038/s41562-019-0739-7, https://doi.org/10.1038/s41562-019-0648-9, **PMID:** 31285622
Scott, S. K., & Johnsrude, I. S. (2003). The neuroanatomical and functional organization of speech perception. *Trends in Neurosciences*, *26*, 100–107. **DOI:** https://doi.org/10.1016/S0166-2236(02)00037-1
Sharp, D. J., Scott, S. K., Cutler, A., & Wise, R. J. S. (2005). Lexical retrieval constrained by sound structure: The role of the left inferior frontal gyrus. *Brain and Language*, *92*, 309–319. **DOI:** https://doi.org/10.1016/j.bandl.2004.07.002, **PMID:** 15721963
Skipper, J. I., Van Wassenhove, V., Nusbaum, H. C., & Small, S. L. (2007). Hearing lips and seeing voices: How cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex*, *17*, 2387–2399. **DOI:** https://doi.org/10.1093/cercor/bhl147, **PMID:** 17218482, **PMCID:** PMC2896890
Sparing, R., Mottaghy, F. M., Ganis, G., Thompson, W. L., Töpper, R., Kosslyn, S. M., et al. (2002). Visual cortex excitability increases during visual mental imagery—A TMS study in healthy human subjects. *Brain Research*, *938*, 92–97. **DOI:** https://doi.org/10.1016/S0006-8993(02)02478-2
Talairach, J., & Tournoux, P. (1988). *Co-planar stereotaxic atlas of the human brain*. New York: Thieme.
Ullas, S., Formisano, E., Eisner, F., & Cutler, A. (2020a). Interleaved lexical and audiovisual information can retune phoneme boundaries. *Attention, Perception, & Psychophysics*, *82*, 2018–2026. **DOI:** https://doi.org/10.3758/s13414-019-01961-8, **PMID:** 31970708
Ullas, S., Formisano, E., Eisner, F., & Cutler, A. (2020b). Audiovisual and lexical cues do not additively enhance perceptual adaptation. *Psychonomic Bulletin & Review*, *27*, 707–715. **DOI:** https://doi.org/10.3758/s13423-020-01728-5, **PMID:** 32319002, **PMCID:** PMC7398951
van der Zande, P., Jesse, A., & Cutler, A. (2014). Cross-speaker generalisation in two phoneme-level perceptual adaptation processes. *Journal of Phonetics*, *43*, 38–46. **DOI:** https://doi.org/10.1016/j.wocn.2014.01.003
van Linden, S., & Vroomen, J. (2007). Recalibration of phonetic categories by lipread speech versus lexical information. *Journal of Experimental Psychology: Human Perception and Performance*, *33*, 1483–1494. **DOI:** https://doi.org/10.1037/0096-1523.33.6.1483, **PMID:** 18085958
Vroomen, J., van Linden, S., Keetels, M., de Gelder, B., & Bertelson, P. (2004). Selective adaptation and recalibration of auditory speech by lipread information: Dissipation. *Speech Communication*, *44*, 55–61. **DOI:** https://doi.org/10.1016/j.specom.2004.03.009
Yi, H. G., Leonard, M. K., & Chang, E. F. (2019). The encoding of speech sounds in the superior temporal gyrus. *Neuron*, *102*, 1096–1110. **DOI:** https://doi.org/10.1016/j.neuron.2019.04.023, **PMID:** 31220442, **PMCID:** PMC6602075
Zatorre, R. J., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: Music and speech. *Trends in Cognitive Sciences*, *6*, 37–46. **DOI:** https://doi.org/10.1016/S1364-6613(00)01816-7
Zatorre, R. J., Evans, A. C., Meyer, E., & Gjedde, A. (1992). Lateralization of phonetic and pitch discrimination in speech processing. *Science*, *256*, 846–849. **DOI:** https://doi.org/10.1126/science.1589767, **PMID:** 1589767