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

Shruti Ullas¹,², Lars Hausfeld¹,², Anne Cutler³, Frank Eisner⁴, and Elia Formisano¹,²,⁵

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 perceptions; 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 /da/) leads to a perceived new syllable (often /da/; McGurk & MacDonald, 1976). Similarly, when presented with a word containing an unclear syllable (such as /d/-/t/ blend instead of /d/ in desk), listeners are more likely to report hearing a word rather than a nonword (desk rather than task; 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, 2005). 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

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

© 2020 Massachusetts Institute of Technology

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, & Humphries, 2002; Binder, 2000; Zatorre, Evans, Meyer, & Gjede, 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, 2005), 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 be-
tween lexical and audiovisual recalibration under similar
experimental constraints, as well as potential unique con-
tributions from multimodal or visual regions for audio-
visual 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 differ-
ences. 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 stud-
ies, to have been influenced by visual cortex activity as
well. For both lexical retuning and audiovisual recalibra-
tion, 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 re-
cruited from Maastricht University to take part in the
study (data from one participant were not analyzed be-
cause 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 ei-
ther 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 suc-
cessful, 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, respec-
tively. 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 con-
tain 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,
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 coarticularatory 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/.

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
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
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.
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 occipito-temporal 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...
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 4](image-url). Audiovisual exposure (A), audiovisual test (B), lexical exposure (C), and lexical test (D) blocks versus baseline, with $t(10) > 3.17, p < .01$. 

2152 Journal of Cognitive Neuroscience Volume 32, Number 11
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)

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

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.
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 signifi-
cant effects in both conditions and toward both pho-
nemes, although audiovisual recalibration effects were
larger than lexical retuning, in line with previous find-
as 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 prop-
eties 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 pre-
viously mentioned, these regions are responsible for rep-
resenting 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) re-
ported STG and MTG activity to be strongest for ambigu-
ous items that had been perceptually shifted by
(2011) similarly noted STG as well as IPL, insula, and
inferior frontal sulcus to be activated during audiovisual re-
calibration and that IPL can coordinate higher-order
constructive processes in perception. Regions in the pa-
rietial 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). Simi-
larly, 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 out-
side the core speech network, they may also be involved
in less tangible functions, such as processing abstract lin-
guistic 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 post-
central 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 audio-
visual 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 re-
sponse 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 dur-
ing the subsequent test block, possibly as a form of mental
imagery or within an STM loop, as early visual areas are ca-
peable of contributing to visual mental imagery (Sparing
et al., 2002; Kosslyn, Ganis, & Thompson, 2001). Asso-
ciative 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-Hüttener, 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 relative increase in activity observed during high recalibration effects and a lower activity for less such effect. The test blocks where participants showed larger recalibrations, and in right insular and parietal clusters, as well including clusters in left temporal, motor, and insular regions, as well as specific activity in left parietal, frontal, and insular regions. These findings suggest a role for these regions in the process of auditory recalibration.

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**


PMID: 15546770
PMID: 19793570
PMID: 15044149
PMID: 17431404
PMID: 11906217
PMID: 20045890
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
PMID: 24906217
PMCID: PMC4149812
PMID: 28261047
PMCID: PMC5306389
PMID: 15703256
PMID: 27611960
PMCID: PMC5017187
PMID: 1012311
PMID: 24482117
PMCID: PMC450233
PMID: 18247893
Mitterer, H., Scharenborg, O., & McQueen, J. M. (2013). Phonemic abstraction without phonemes in speech
perceptual adjustment to non-standard speech tokens. 


