CHAPTER 3

Neural responses to non-native phonemes varying in producibility: Evidence for the sensorimotor nature of speech perception

3.1 Abstract

Neural responses to unfamiliar non-native phonemes varying in the extent to which they can be articulated were studied with functional magnetic resonance imaging (fMRI). Both superior temporal (auditory) and precentral (motor) areas were activated by passive speech perception, and both distinguished non-native from native phonemes, with greater signal change in response to non-native phonemes. Furthermore, speech-responsive motor regions and superior temporal sites were functionally connected. However, only in auditory areas did activity covary with the producibility of non-native phonemes. These data suggest that auditory areas are crucial for the transformation from acoustic signal to phonetic code, but the motor system also plays an active role, which may involve the internal generation of candidate phonemic categorizations. These 'motor' categorizations would then be compared to the acoustic input in auditory areas. The data suggest that speech perception is neither purely sensory nor motor, but rather a sensorimotor process.

3.2 Introduction

Speech perception involves a transformation from an acoustic signal to a phonetic code, but the nature of the phonetic code—acoustic, articulatory, amodal or some combination—is debated (Liberman and Mattingly, 1985). The motor theory of speech perception proposed that the phonetic code is articulatory in nature, because the striking context-dependency of acoustic cues suggested that only at the level of motor control structures could invariant representations of phonemes be found (Liberman et al., 1967). But much of the evidence presented in support of the motor theory, such as categorical perception, and context-dependency of acoustic cues, is relevant only to other claims of the theory, such as the discreteness of the objects of speech perception, and the cognitive impenetrability of the process. There is much less evidence for the central claim that the phonetic code is articulatory and that the motor system is involved in deriving it. Many researchers have argued against the motor theory, advocating models of speech perception that focus on the auditory system and the acoustic properties of speech (e.g. Kuhl and Miller, 1975 and Stevens, 1981).

However, over the last decade the discovery that motor areas are involved in the representation of observed actions (Rizzolatti and Craighero, 2004) has renewed interest in the motor theory of speech perception. Several recent studies have shown that motor areas are activated by passive speech perception, using transcranial magnetic stimulation (TMS) (Fadiga et al., 2002 and Watkins et al., 2003) and functional neuroimaging (Wilson et al., 2004). In particular, a superior part of ventral premotor cortex (sPMv) has been shown to respond bilaterally to perception of meaningless monosyllables (Wilson et

al., 2004). This region is also involved in speech production, and its location is close (though somewhat anterior and superior) to the location of motor speech areas determined in a meta-analysis of imaging studies (Fox et al., 2001). However, little is known about whether motor areas (and in particular, sPMv) are modulated by particular properties of acoustic inputs, so the extent to which speech perception depends on the motor system remains an open question. It is noteworthy that Broca's area, a premotor area in the posterior inferior frontal gyrus, is not strongly activated by passive listening to meaningless speech (Wilson et al., 2004), and therefore sPMv is the motor area of most interest in the current study. However there is evidence that Broca's area is responsible for modulating motor excitability in speech perception (Watkins and Paus, 2004), and its role in various phonological tasks is well established (Burton et al., 2000 and Bookheimer, 2002).

The objective of this study was to investigate the roles of auditory and motor areas in processing acoustic inputs by using fMRI to examine neural responses to non-native phonemes varying in the extent to which they can be articulated. Each of the world's languages employs a limited set of phonemes from which all the words and morphemes of the language are composed (Kenstowicz, 1994). Infants can discriminate any potential phonetic contrast, but in the first year of life perceptual abilities are honed so that only native contrasts are perceived (Jusczyk, 1997). Likewise infants learn to produce the phonemes of their native language, but not those of other languages. We hypothesized that activity in brain areas involved in transforming the acoustic signal to a phonetic code would differ for native and non-native phonemes, since only for native phonemes can an

accurate internal representation be obtained. Furthermore, if internal representations of phonemes are sensorimotor, then activity in areas involved in deriving a phonetic code might covary with the producibility of novel phonemes, reflecting mismatch between the incoming acoustic input and the predicted acoustic consequences of known phonemes; the degree of mismatch would reflect the extent to which the novel phoneme could be produced. If speech motor areas are modulated by either of these factors (nativeness, producibility), this would bolster the claim that the motor system represents linguistic features of perceived speech.

3.3 Materials and methods

3.3.1 Stimuli

We selected 42 non-English consonants from a variety of languages, and 8 English consonants. The set of non-native phonemes was selected so as to include a range of places of articulation and manners of articulation (Ladefoged and Maddieson, 1996), and to include both phonemes that are relatively easy for English speakers to produce and those that are more difficult. All 50 consonants were produced by an experienced phonetician (Peter Ladefoged) in the environment [ɑCɑ], i.e. each consonant was embedded between two [ɑ] vowels, with stress on the second vowel. For example, if the consonant was [h], this would sound like the English interjection *aha!*. Each phoneme was produced at least three times.

Stimuli were recorded on DAT at 44100 Hz in a soundproof booth, then transferred to a PC. The best token of each stimulus was selected and cropped. The 50 stimuli were then normalized in amplitude by scaling the waveform such that the 97th percentiles of the absolute value of the waveforms were equated. Of the 50 stimuli, 44 were selected for further norming; 6 were discarded due to excessive similarity to others, disfluent production, or excessive similarity to English phonemes.

Two norming studies were performed prior to fMRI scanning, both using monolingual native English speakers. In the first, 15 participants (aged 18–56, mean 27.5, 6 females, 3 left-handed) took part and were paid for their participation. Subjects were asked to listen to the phonemes and attempt to repeat them, then evaluate their performance on a scale from 1 to 4. The experiment was performed on a laptop PC and subjects listened to the stimuli through headphones and made responses into a microphone in a soundproof booth. After several practice trials, the set of 44 phonemes was presented three times in three different random orders. Responses sometimes consisted of producing the closest English phoneme to the non-native phoneme being attempted, for instance, producing [h] instead of the voiceless velar fricative [x]. However more frequently, subjects attended to the phonetic features which distinguished the non-native phonemes from any English phoneme (e.g. the palatal place of articulation which distinguishes [ʎ] from [l]), and attempted to reproduce them with varying levels of success.

The subjects' own ratings for their ability to produce each phoneme were averaged across the three attempts at each phoneme. Furthermore, one of the authors (S.M.W.),

who has phonetic training and linguistic fieldwork experience, rated each trial offline using the same 4-point scale, so that both self-assessed and experimenter-assessed ratings were obtained for each phoneme for each subject. There was a high correlation between these two ratings ($r^2 = 0.71$), so they were averaged together for each phoneme to obtain a single producibility metric. The imaging data were also analyzed using each of these two ratings separately, and very similar results were obtained to those reported below.

In the second norming study, 10 participants (aged 20–30, mean 26.3, 7 females, 1 left-handed) took part. Subjects were asked to listen to the phonemes and rate them on a scale from 1 to 4 as to how "Englishlike" they sounded, or "how much does this sound like it could be a possible sound of English?". The aim of this measure was to quantify two factors which are closely related: firstly, to what extent is each sound novel, i.e. clearly distinct from what is heard in the native language, and secondly, to what extent is each sound perceivable as not being a phoneme of English. As in the first norming study, a laptop PC was used to present the stimuli and collect responses, and the 44 phonemes were presented three times in different random orders, with all ratings averaged across the three repetitions.

Based on these two norming studies, 25 non-native phonemes and 5 native phonemes were selected for the fMRI component of the study. This selection was made with the goal of retaining a range of places and manners of articulation, as well as a continuum of producibility and of Englishlikeness. The 30 phonemes used in the study and the producibility and Englishlikeness measures obtained for them are shown in Table 3.1. Recordings of the phonemes used are available as supplementary materials online. The

Table 3.1 Phonemes used in the study

IPA	Description	Language	Producibility	English-likeness
ǃ	Postalveolar click	Nama	1.57	1.00
ǁ	Alveolar click	Nama	1.58	1.03
ǀ	Dental click	Nama	1.63	1.00
qʼ	Uvular ejective stop	K'ekchi	1.89	1.17
q	Voiceless uvular stop	Aleut	2.29	1.97
qʷʼ	Rounded uvular ejective stop	Montana Salish	2.30	1.03
ɗ̥	Voiceless alveolar implosive stop	Owerri Igbo	2.31	2.30
ʁ	Uvular rhotic	French	2.37	1.93
ɣ	Voiced velar fricative	Greek	2.38	1.83
tʃʼ	Palato-alveolar ejective stop	Quechua	2.38	1.70
ŋ̥	Voiceless velar nasal	Burmese	2.38	1.63
l̥	Voiceless alveolar lateral	Melpa	2.43	1.60
n̤	Breathy voice alveolar nasal	Marathi	2.49	1.63
d̪ʰ	Voiced aspirated dental stop	Hindi	2.61	1.90
ʋ	Voiced labiodental approximant	Isoko	2.74	2.27
r̥	Voiceless alveolar trill	Turkish	2.82	1.33
ᵐʙ	Prenasalized bilabial trill	Kele	2.86	1.23
ɓ̥	Voiceless bilabial implosive stop	Owerri Igbo	2.86	3.43
x	Voiceless velar fricative	German	2.90	2.23
rʲ	Palatalized alveolar trill	Russian	3.09	1.40
r	Alveolar trill	Spanish	3.11	1.47
ŋʷ	Labialized velar nasal	Idoma	3.14	2.03
ʂ	Voiceless retroflex postalveolar fricative	Polish	3.21	3.50
ʎ	Palatal lateral approximant	Italian	3.22	2.10
ʐ̟ʷ	Rounded voiced lamino-postalveolar fricative	Ubykh	3.56	2.37
ɹ	Alveolar approximant	English	3.62	3.93
ʒ	Voiced palatal fricative	English	3.72	2.87
b	Voiced bilabial stop	English	3.77	3.70
z	Voiced alveolar fricative	English	3.77	3.97
w	Voiced labial-velar approximant	English	3.88	3.93

mean duration of the stimuli (including the carrier vowels) was 825 ms (sd = 64 ms), and this did not differ according to nativeness, nor was duration correlated with producibility or Englishlikeness. The non-native phonemes varied widely in place and manner of articulation, and included clicks and trills, as well as stops, fricatives and sonorants with unfamiliar places or manners of articulation, or secondary articulations. Not surprisingly, the correlation between producibility and Englishlikeness was quite high ($r^2 = 0.60$), but these two measures were different enough that they led to different results when used as explanatory variables in the imaging study.

Of the 25 non-native phonemes, only two showed some tendency to be misperceived as English phonemes. These were [ɓ̥], a voiceless bilabial implosive stop, which was often perceived as a [b], and [ʂ], a voiceless retroflex postalveolar fricative, which was often perceived as [ʃ], the voiceless postalveolar fricative in English. In these cases, subjects provided high self-assessed producibility ratings (2.96 and 3.47 respectively) and high Englishlikeness ratings (3.43 and 3.50 respectively). However their actual productions as rated by the experimenter were poorer (2.76 and 2.96 respectively) because they often failed to attend to the features which distinguish these phonemes from perceptually similar English phonemes.

On the other end of the spectrum, to ensure that the three click phonemes were actually perceived as speech sounds, participants in both the norming and the imaging studies were told in advance that some of the sounds would be "clicks from African

languages". The placement of each consonant between two native vowels also contributed to them being perceived as speech.

3.3.2 Scanning procedure

In the fMRI study, 12 monolingual native English speakers (aged 21–37, mean 26.5, 7 females, all right-handed) were scanned. All participants gave informed consent and the study was approved by the UCLA Institutional Review Board.

Functional images were acquired on a 3 T Siemens Allegra scanner at the Ahmanson-Lovelace Brain Mapping Center at UCLA. Phonemes were presented (in intervocalic contexts) during 3 functional runs (TR = 2000 ms; TE = 25 ms; flip angle = 90º; 36 axial slices with interleaved acquisition; 3 × 3 × 4 mm resolution; field of view = 192 × 192 × 144 mm). Each run was 400 seconds in duration (i.e. 200 volumes were acquired), plus 4 seconds to allow for magnetization to reach steady state. Each of the 30 consonants was presented 12 times in total across the 3 runs in a jittered rapid event-related design. The minimum ISI was 2.0 s and the mean ISI was 3.3 s. The minimum ISI between two repetitions of the same phoneme was 20.0 s, and the mean was 86.3 s. Efficient trial placements were determined using custom MATLAB software interfacing with FMRISTAT (Worsley et al., 2002). Stimuli were presented through scanner-compatible headphones at a volume sufficiently loud that the phonemes could be readily perceived over the scanner noise. The volume level was set individually for each subject to a comfortable level during preliminary scans. Participants wore goggles showing a blank screen, so there was no visual stimulation.

Then in a fourth functional run, participants performed a speech production task in order to map mouth motor areas. Scanning parameters were as above, except that this run was only 260 seconds in duration (130 volumes), plus 4 seconds. Subjects were asked to say "ba ba ba..." whenever a central crosshair turned into a circle, and to stop when it returned to a crosshair. The circle appeared 16 times, once every 16 seconds, for 3 seconds each time. Participants were specifically requested to minimize head movement while speaking.

Two anatomical sequences were acquired for registration purposes: high-resolution T2-weighted images coplanar with the functional images (TR = 5000 ms; TE = 33 ms; flip angle = 90°; 36 axial slices; 1.5 × 1.5 × 4 mm resolution; field of view = 192 × 192 × 144 mm); and an MP-RAGE structural volume (TR = 2300 ms; TE = 2.93 ms; flip angle = 8°; 160 sagittal slices; 1.33 × 1.33 × 1.5 mm resolution; field of view = 256 × 256 × 240 mm).

3.3.3 Image analysis

The fMRI data were preprocessed using tools from FSL (Smith et al., 2004). Skull stripping was performed with BET, motion correction was carried out with MCFLIRT, and the program IP was used to smooth the data with a Gaussian kernel (8mm FWHM) and to normalize mean signal intensity across subjects.

Statistical analysis was performed by fitting a general linear model (GLM) with the FMRISTAT toolbox (Worsley et al., 2002). Each of the 30 phonemes was modeled as a separate event type. The design matrix of the linear model was convolved with a

hemodynamic response function (HRF) modeled as a difference of two gamma functions. Temporal drift was removed by adding a cubic spline in the frame times to the design matrix (one covariate per 2 minutes of scan time), and spatial drift was removed by adding a covariate in the whole volume average. Six motion parameters (three each for translation and rotation) were also included as confounds of no interest. Autocorrelation parameters were estimated at each voxel and used to whiten the data and design matrix. The three perception runs within each subject were combined using a fixed effects model.

Voxels where signal change was correlated with producibility were identified by fitting a second GLM at each voxel using the 25 effect size images for each non-native phoneme as the data. An alternative approach in which the 25 phonemes were modeled by one explanatory variable, with a second explanatory variable whose height reflected producibility, produced similar results, which are not reported further. Correlations with Englishlikeness were assessed using the same procedure.

The speech production run was analyzed by coding each speech production instance as a 3-second event, which was then convolved with the HRF. Each pair of volumes acquired during the actual speaking was excluded from the analysis, which is feasible because the delayed hemodynamic response does not peak until several seconds after the subject has stopped speaking. Several studies have shown the utility of this approach for designs that entail task-correlated head movement (e.g. Birn et al., 1999).

Registration was performed with the FSL tool FLIRT. Functional images were aligned to high-resolution coplanar images using an affine transformation with 6 degrees

of freedom. High-resolution coplanar images were aligned to the standard MNI average of 152 brains using an affine transformation with 12 degrees of freedom.

Group analysis was performed with FMRISTAT with a mixed effects (also known as random effects) linear model (Worsley et al., 2002). Standard deviations from individual subject analyses were passed up to the group level. Variance ratio images were not smoothed (i.e. a conventional group analysis was performed). The resulting *t* statistic images were thresholded at $t > 3.106$ ($df = 11$, $p < 0.005$ uncorrected) at the voxel level, with a minimum cluster size then applied so that only clusters significant at $p < 0.05$ (corrected) according to Gaussian random field theory were reported. Statistical parameter maps were displayed as overlays on a high-resolution single subject T1 image ("colin27") using AFNI (Cox, 1996).

A region of interest (ROI) analysis was carried out to examine signal change in (a) motor areas activated by all speech perception versus rest (i.e. sPMv); (b) areas that were activated more by non-native phonemes than native phonemes; and (c) areas where activity was negatively correlated with producibility. The first of these pairs (left/right) of ROIs were defined for each individual subject by thresholding the contrast of listening to all phonemes versus rest, usually at $t > 2.3$, then identifying the relevant activations. For three subjects slightly higher cutoffs were used, to separate the motor clusters from superior temporal clusters, and for one subject a slightly lower cutoff was used since the motor activation in one hemisphere was too weak to reach the 2.3 cutoff. In all cases, there was no difficulty in identifying the relevant clusters. The second and third ROIs were simply based on the areas activated in the group analysis at a cutoff of $t > 3.106$.

Signal change in ROIs was computed by averaging signal change across all voxels in the ROI.

Functional connectivity analyses were conducted by including the timecourses of various ROIs (including left and right speech-responsive motor areas) as additional covariates in the GLM. All ROI timecourses were divided by the whole brain timecourse first to avoid detecting correlations based solely on global signal changes. We also tried an alternative approach in which residuals from ROIs rather than raw timecourses were used as covariates; results obtained using this method were similar to the results reported. The three runs within each subject were combined with fixed effects models, and group analyses were performed with mixed effects models and thresholded as described above.

3.4 Results

3.4.1 Group analyses

For the contrast of all phonemes versus rest, the largest activations were bilateral in the superior temporal gyrus and sulcus (Figure 3.1a, Table 3.2). There were also bilateral activations spanning the border of premotor and primary motor cortex. These motor activations for speech perception overlapped with mouth motor areas activated by speech production, shown with black outlines in the middle panel of Figure 3.1a, replicating previous findings (Wilson et al., 2004). Finally, there was an activation in the right cerebellum.

When responses to native and non-native phonemes were contrasted, there were no areas that were more active for native phonemes. Cortical regions responding more to

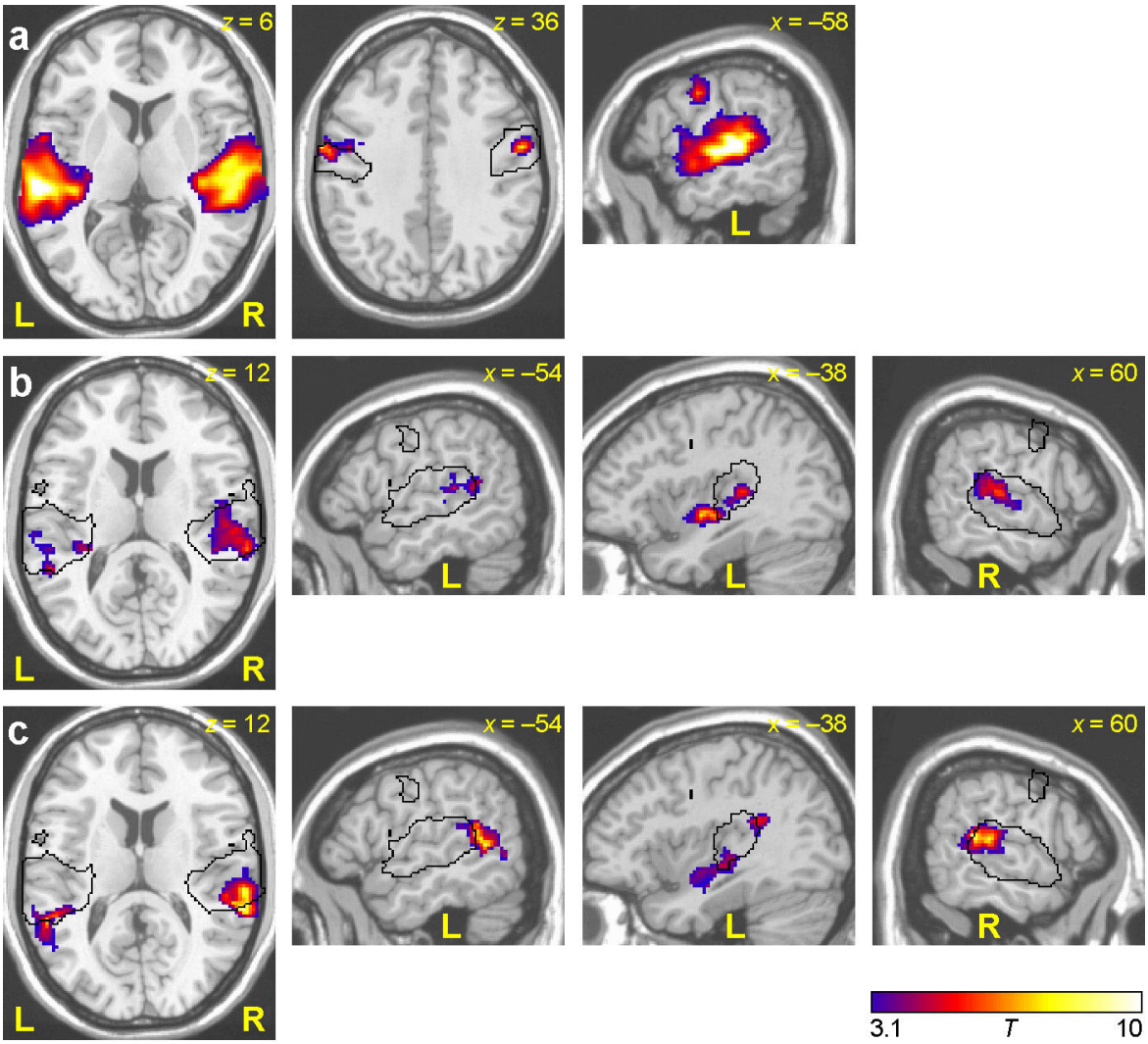

Figure 3.1 Speech-responsive regions and areas sensitive to the factors of nativeness and producibility. (a) Areas activated by listening to all phonemes relative to rest. The black outline on the middle panel shows mouth premotor and primary motor cortex activated by speech production, demonstrating the overlap between motor areas activated by speech perception (i.e. sPMv) and speech production. (b) Areas activated more by non-native phonemes than native phonemes. The black outline here and in panel c shows areas activated by listening to all phonemes relative to rest. (b) Areas where activity was greater the more difficult a phoneme is to produce, i.e. where signal change was negatively correlated with producibility.

Table 3.2 Areas activated in each contrast of interest

Area	MNI coordinates			Extent	Max *t*	Cluster *p*
	x	*y*	*z*	(mm^3)		
All phonemes > rest						
Left superior temporal	–42	–28	12	40904	15.6	< 0.0001
Right superior temporal	48	–12	–2	46712	14.9	< 0.0001
Left pre/primary motor cortex	–62	–4	38	2816	8.1	0.027
Right pre/primary motor cortex	56	–4	38	2952	8.3	0.022
Right cerebellum	18	–68	–26	3640	5.4	0.0088
Non-native phonemes > native phonemes						
Left superior temporal	–38	–8	–6	6552	7.5	0.0005
Right superior temporal	64	–34	10	8888	6.5	0.0001
Negative correlation with producibility						
Left superior temporal	–52	–46	14	6176	8.7	0.0007
Left superior temporal	–42	0	–2	2800	5.9	0.027
Right superior temporal	52	–34	8	7096	14.5	0.0003

non-native phonemes were found bilaterally in the superior temporal lobe (Figure 3.1b, Table 3.2). These regions were largely contained within the areas activated by all phonemes versus rest, but in the left hemisphere extended anteriorly and medially as far as the posterior insula (see third slice).

We next looked for correlations between producibility and signal change for the 25 non-native phonemes. There were no areas showing a positive correlation with producibility. Bilateral superior temporal regions showed a significant negative correlation with producibility (Figure 3.1c, Table 3.2), i.e. the more difficult phonemes

were to produce, the more these areas were active. In the right hemisphere, the area activated was very similar to the area activated for non-native versus native phonemes. In the left hemisphere, an anterior temporal region extending to the posterior insula also mostly overlapped the left temporal area that was more active for non-native than native phonemes (see third slice). However there was one additional left hemisphere area that was negatively correlated with producibility. This area was located in a region posterior to the speech-responsive region (see second and third slices). The peak coordinates of this area correspond very closely to the coordinates of a region called Spt (Sylvian–parietal–temporal) proposed to be involved in mapping between auditory and motor representations (Hickok et al., 2001 and Scott and Wise, 2004).

We considered the possibility that areas responding more for phonemes that are difficult to produce might be responding merely to the novelty of the more unfamiliar phonemes, since it is known that novel auditory stimuli result in greater levels of activation in superior temporal cortex (Opitz et al., 1999). To test this hypothesis, we looked for correlations between Englishlikeness and signal change for the 25 non-native phonemes. No areas were significantly activated; the largest cluster was in the right superior temporal lobe but it was not large enough to pass the cluster size threshold (p = 0.080). Furthermore, when producibility and Englishlikeness were both included in a model as covariates, bilateral superior temporal activations similar to those in Figure 3.1c were found for producibility, but no areas were activated for Englishlikeness (p = 0.97 for the largest cluster).

3.4.2 Region of interest (ROI) analyses

Although the group analyses did not reveal any motor areas differentially activated for native or non-native phonemes, nor any motor areas where activity correlated with producibility, we used a more sensitive ROI approach to examine responses in the motor areas that were activated by speech perception, i.e. sPMv, the same superior part of ventral premotor cortex previously reported to respond to speech sounds (Wilson et al., 2004). We first compared responses to native and non-native phonemes (Figure 3.2a). A repeated measures ANOVA revealed that non-native phonemes activated motor areas more than native phonemes ($F(1, 28) = 4.46$; $p = 0.044$), which is important because it demonstrates that speech-responsive motor regions are sensitive to the distinction between phonemes that are part of the speaker's inventory, and those that are not. The interaction of nativeness by hemisphere (left versus right motor ROI) was not significant ($F(1, 28) = 2.81$; $p = 0.11$). In superior temporal areas, the effect of nativeness was even greater ($F(1, 28) = 15.95$; $p = 0.0004$), and there was also a significant interaction of nativeness by hemisphere ($F(1, 28) = 4.21$; $p = 0.0496$), with the difference between native and non-native phonemes greater in the right hemisphere.

Although motor areas responded more to non-native phonemes, there was no correlation between producibility of non-native phonemes and signal change ($r = –0.20$; $F(1, 23) = 1.02$; $p = 0.32$), nor any interaction with hemisphere ($F(1, 23) = 1.35$; $p = 0.26$) (Figure 3.2b). This contrasts sharply with the case of the superior temporal cortex where robust correlations were found ($r = –0.79$; $F(1, 23) = 38.92$; $p < 0.0001$) (Figure 3.2c). In superior temporal cortex there was also a significant interaction of producibility

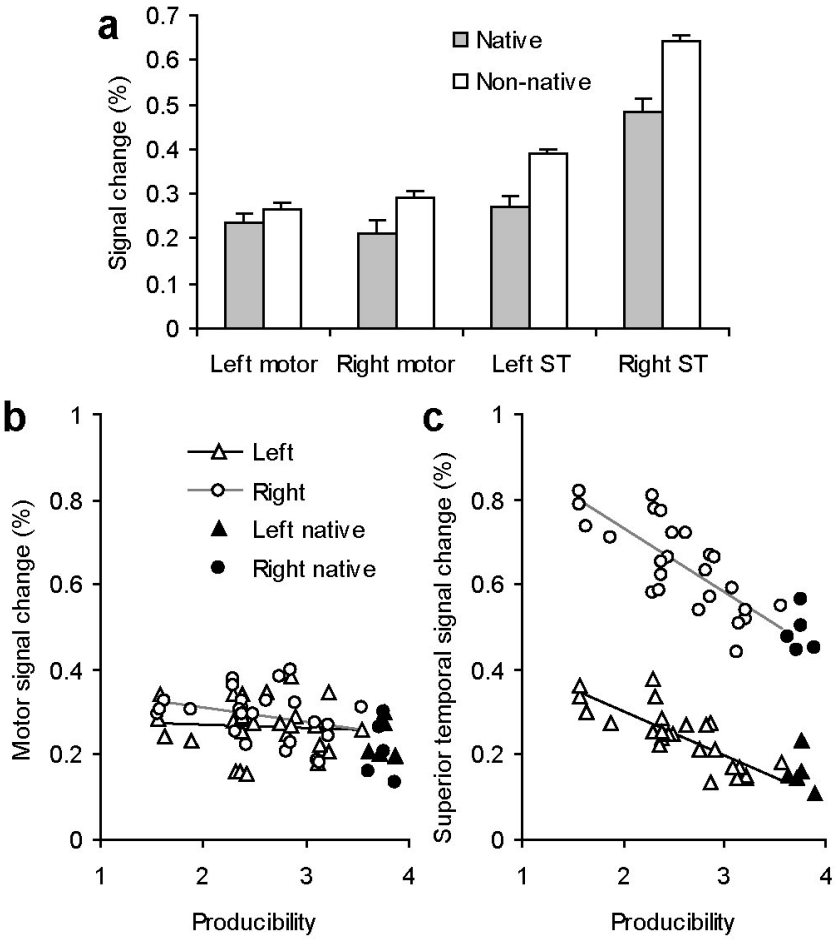

Figure 3.2 Region of interest (ROI) analyses. (a) Signal change for native and non-native phonemes in four regions of interest. Motor ROIs were defined based on individual subjects' maps for all phonemes versus rest; sPMv was identified in each subject. Superior temporal ROIs were defined as the region activated for this contrast in the group analysis (Figure 3.1b). Error bars indicate SEM. (b) Correlational plot of signal change versus producibility in the left and right motor ROIs. Here and in panel c, the five English phonemes are also shown (filled symbols), though they were not used in calculating the correlation. (c) Correlational plot of signal change versus producibility in left and right superior temporal ROIs defined as those areas activated by the negative correlation with producibility in the group analysis (Figure 3.1c).

by hemisphere ($F(1, 23) = 7.05$; $p = 0.014$), such that there was a steeper slope in the right hemisphere. Finally, all ROI analyses were repeated excluding the two phonemes [θ̥] and [ʂ] which were sometimes misperceived as English phonemes, and the same results were obtained from all significance tests.

3.4.3 Functional connectivity analyses

The coactivation of motor and auditory areas in speech perception suggested that these areas might communicate with one another to implement a mechanism of speech perception that is neither motor nor sensory, but rather sensorimotor. We performed a functional connectivity analysis to determine whether there is connectivity between motor areas activated by speech perception (sPMv) and superior temporal regions. Auditory events were included in the models, so correlations do not just reflect common responses to stimuli. For both left (Figure 3.3a) and right (Figure 3.3b) speech-responsive motor regions, correlated regions were found in superior temporal cortex, close to those regions that distinguished native and non-native phonemes (compare Figure 3.1b), or where activity covaried with producibility (compare Figure 3.1c). Likewise, for both left (Figure 3.3c) and right (Figure 3.3d) superior temporal regions defined as voxels where signal change negatively correlated with producibility, we found correlations with speech-responsive motor regions. Our results are consistent with a previous PET study reporting functional connectivity between the planum temporale and the primary motor area for the face (Paus et al., 1996), and with an fMRI study that demonstrated

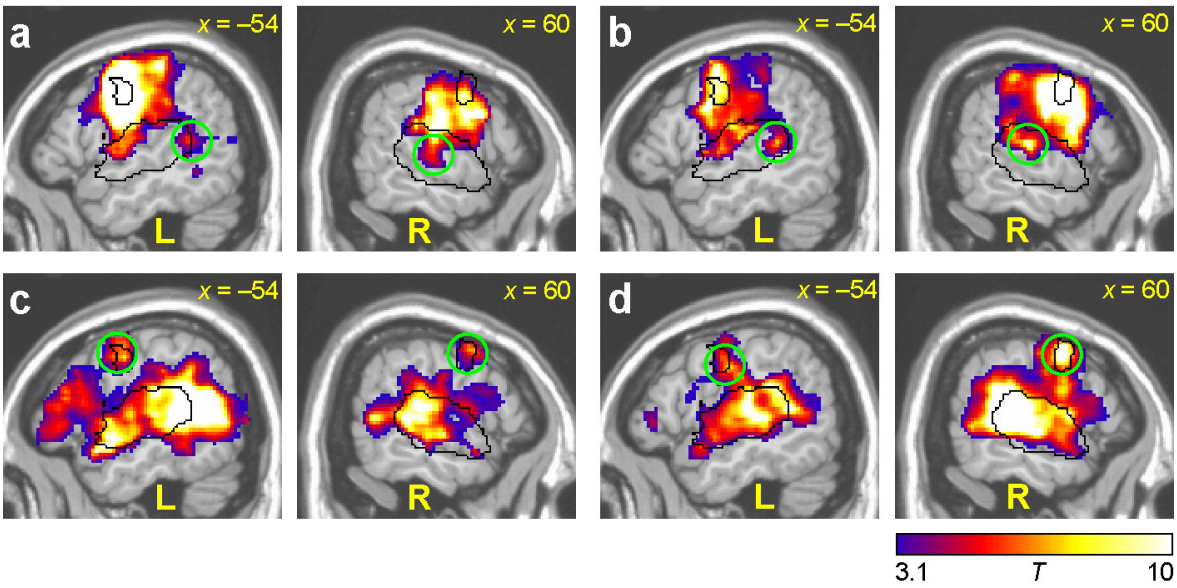

Figure 3.3 Functional connectivity analyses. (a) Areas correlated with the left speech-responsive motor region (sPMv). The same slices are shown here as in Figures 3.1b and 3.1c, and the black outline likewise shows areas activated by all phonemes relative to rest. The green circles show superior temporal areas of interest. MNI coordinates for peak voxels in these areas were (–46, –46, 12) in the left, and (64, –20, 2) in the right hemisphere. (b) Areas correlated with the right speech-responsive motor region (sPMv). MNI coordinates for peak superior temporal voxels were (–46, –36, 8) in the left and (66, –26, 10) in the right hemisphere. (c) Areas correlated with the left posterior superior temporal region where signal correlated with producibility. The green circles show motor areas of interest. MNI coordinates for peak voxels in these areas were (–60, –8, 42) in the left and (62, 2, 46) in the right hemisphere. (d) Areas correlated with the right superior temporal region where signal correlated with producibility. MNI coordinates for peak motor voxels were (–52, –12, 34) in the left and (62, –2, 44) in the right hemisphere.

connectivity between Wernicke's area and a premotor area that is likely mouth-related (Bartels and Zeki, 2005).

3.5 Discussion

These findings suggest that superior temporal auditory areas bilaterally are crucial for the transformation of acoustic speech input to a phonetic code, since only in these areas, and not in motor areas, did signal change correlate with producibility. The central role of bilateral superior temporal cortex in speech perception has been established in numerous imaging and neuropsychological studies (see Hickok and Poeppel, 2000, 2004 and Scott and Wise, 2004 for reviews). Three pieces of evidence, however, point to an important role for speech motor areas, in particular sPMv, in the process: first, motor areas were activated for speech perception relative to rest (Figure 3.1a); second, activity in motor areas differed for native versus non-native phonemes (Figure 3.2a); and third, motor areas were functionally connected to superior temporal cortex (Figure 3.3). The novel finding that motor areas distinguish between native and non-native phonemes is particularly important since it suggests that these regions are sensitive to whether or not phonemes are part of the speaker's inventory, which supports the idea that motor areas play an active role in the speech perception process.

Our results suggest that internal representations of known phonemes are neither purely acoustic nor purely motor, but are sensorimotor in nature. In speech perception, the motor system may be involved in generating internal forward models of native phonemes, whereas the auditory system may be responsible for comparing the acoustic input to the predicted acoustic consequences of phonemes under consideration. We propose that the role of the motor system in speech perception is to generate "top-down" internal models of phonemes under consideration. In other words, when faced with a non-

native phoneme, the motor system attempts to answer the question "how would I say this?" Forward models lead to representations in superior temporal cortex of the predicted acoustic consequences of phonemes under consideration. The superior temporal activity inversely correlated with producibility may be akin to an error signal coding the extent of mismatch between the input and the predicted acoustic consequences of native phonemes under consideration (Haruno et al., 2001). A role for the posterior superior temporal plane in particular in matching auditory input to stored templates has been proposed (Hickok and Poeppel, 2000, 2004, Scott and Wise, 2004; see Warren et al., 2005 for a detailed model). We concur with this view but emphasize a role for the motor system in the online generation of these internal auditory templates (c.f. Callan et al., 2004). According to our account, the motor system can only simulate known phonemes; when hearing a native phoneme, a match is readily obtained, whereas when hearing a non-native phoneme, a match is never obtained, so the motor system is engaged in repeated attempts to model other phonemes, leading to greater motor activity. This would account for the results of the present study: that motor activity only distinguished between native and non-native phonemes, whereas superior temporal activity also coded the extent of mismatch for non-native phonemes.

In superior temporal cortex, much more robust correlations were observed with the producibility metric in comparison to the Englishlikeness metric. This indicates that the greater responses for phonemes that are difficult to produce reflect more than just the unfamiliarity of these phonemes. The Englishlikeness metric also reflects the ability to perceive that a phoneme is distinct from any phoneme in the English inventory, thus this

analysis suggests that the correlations in superior temporal cortex reflect "producibility" more than "perceivability". A number of neurophysiological studies have revealed differences in the neural processing of native and non-native phonemes, using the mismatch negativity (MMN) auditory evoked potential, or its magnetic counterpart (MMNm) (for review, see Näätänen, 2001 and Zhang et al., 2005). The MMN component is elicited by any discriminable auditory change ("deviant") occurring in a train of repetitive ("standard") stimuli (Näätänen, 2001). In a train of native phonemes, deviant native phonemes produced a larger MMN in the left hemisphere than deviant non-native phonemes (Näätänen et al., 1997). Relatedly, linguistically relevant acoustic changes (i.e. crossing a phoneme boundary) produced larger MMNs than changes of equivalent magnitude that did not cross a phoneme boundary (Dehaene-Lambertz et al., 1997). The role of linguistic experience in shaping the MMN has been confirmed in studies in which subjects are trained to discriminate novel phonetic categories; for instance, training of a novel voice onset time contrast led to an increased MMN, larger in the left hemisphere, for the trained stimuli (Tremblay et al., 1997).

Although most studies using MMN paradigms have shown left-hemispheric dominance of the MMN for linguistic stimuli, Shtyrov et al. (1998) reported that under noisy conditions the MMN to deviant phonemes was larger in the right hemisphere. In the present study, phonemes were presented over background scanner noise, and signal change was greater in the right hemisphere for both native and non-native phonemes. Furthermore, in superior temporal areas, the effects of nativeness and producibility were

larger in the right hemisphere. Future studies using sparse scanning could explore the possibility that this right-lateralization is a consequence of the background scanner noise.

Studies based on the MMN have consistently demonstrated greater MMNs for native phonemes, or learned contrasts, whereas in our study we observed increased activity for non-native phonemes. This disparity probably reflects substantial differences in experimental paradigms. Frequently in MMN studies, native phonemes are discriminable as deviants, whereas non-native phonemes cannot be perceptually distinguished from the standards. Under these conditions, it is understandable that there is a greater neural response when the difference is discriminable. On the other hand, in the present study, most non-native phonemes were readily perceivable as non-native, so levels of neural activity instead reflected acoustic processing in some form (e.g. degree of mismatch with known phonemes, as proposed above). Our results are directly consistent with an fMRI study which showed that a non-prototypical example of a vowel sound produced greater activity than a prototypical example in bilateral superior temporal regions (Guenther et al., 2004).

Several imaging studies have investigated the neural consequences of training subjects to discriminate non-native phonetic contrasts. Two studies have shown that after training, numerous areas known to be involved in linguistic processing are recruited, including Broca's area and the anterior insula, premotor cortex, superior temporal regions including Spt, the supramarginal gyrus, and the cerebellum (Callan et al., 2003 and Callan et al., 2004). Along similar lines to our proposal above, Callan et al. (2004) argue that these areas are recruited because they are responsible for instantiating forward and

inverse articulatory-auditory and/or articulatory-orosensory models. In Callan et al. (2004), native English speakers performing the same discrimination task showed less activation in these areas but more activation in anterior superior temporal regions, leading the authors to claim that internal models are more important under adverse conditions (e.g. processing a second language), whereas native speakers make more use of auditory phonetic representations. Another study showed recruitment of the left inferior frontal gyrus and the left caudate nucleus when subjects learned to discriminate between native dental stops and non-native retroflex stops (Golestani and Zatorre, 2004).

Two recent neuroimaging studies have shown greater premotor activity for observation of actions belonging to the observer's motor repertoire compared to those that do not (Buccino et al., 2004 and Calvo-Merino et al., 2005), but not all such studies have obtained this result (Costantini et al., 2005). In contrast, we observed greater motor responses for non-native speech sounds. It is clear that there are major differences between speech perception and the visual perception of actions, so such a discrepancy is not unexpected. Furthermore, the motor area of interest in the present study (sPMv) is not the same region as the premotor regions activated in these action observation studies.

In considering the proposal that the motor system plays an important role in speech perception, it is necessary to consider the fact that patients with Broca's aphasia, who typically have large frontal lesions, have relatively preserved language comprehension (Goodglass, 1993). Although sPMv is distinct from Broca's area, many frontal lesions would extend dorsally to include sPMv. If sPMv is involved in speech perception, then one might expect comprehension deficits to result from these lesions. One possible

explanation is that the motor areas activated by speech perception are bilateral (as are the primary motor areas involved in speech production), and there may be redundancy between the two hemispheres. Most aphasic patients' lesions involve only the left hemisphere. It is possible that in Broca's aphasia, motor areas in the right hemisphere continue to support speech perception, in the same way that speech perception is relatively preserved in patients with unilateral posterior lesions (Hickok and Poeppel, 2004). A second consideration is that many patients with Broca's aphasia actually do show severe phonemic perception deficits under certain conditions (Blumstein et al., 1977, Basso et al., 1977, Miceli et al., 1980 and Caplan et al., 1995). For instance, Basso et al. (1977) found that 20 out of 21 nonfluent patients had deficits (11 severe) on a phoneme identification task involving artificial syllables comprising a voice onset time continuum between *ta* and *da*. However it is not simply the case that the typically good comprehension of patients with Broca's aphasia depends heavily on contextual cues to compensate for phonemic perception deficits, because Miceli et al. (1980) showed that most patients performed well on a single word comprehension task where distractors included phonemic foils. Rather, it appears that deficits are restricted to sublexical speech perception tasks (Hickok and Poeppel, 2004; see also Burton et al., 2000). Precisely which aspects of speech perception are dependent on the integrity of frontal cortical areas remains an important topic for further research, but it is clear that at least some aspects can be severely compromised, which is consistent with a role for the motor system as suggested by the present study and other neuroimaging and TMS studies (Fadiga et al., 2002, Watkins et al., 2003, Wilson et al., 2004 and Skipper et al., 2005).

We observed an activation for speech perception in the right cerebellum, another structure which historically has been thought of as primarily concerned with motor functions. The cerebellar hemisphere contralateral to the language-dominant hemisphere is also known to be involved in a wide range of linguistic functions (Marien et al., 2001 and Jansen et al., 2005), including speech perception (Mathiak et al., 2002). In particular, Mathiak et al. (2002) showed that the right cerebellum was involved in the encoding of durational parameters of perceived speech, consistent with a general role for the cerebellum in time perception (Ivry and Keele, 1989).

In sum, this study confirms the central role of bilateral superior temporal regions in speech perception, since only in these areas did signal change correlate with the producibility of novel phonemes. However there is also evidence for the involvement of motor areas in speech perception, as speech motor areas were activated by passive speech perception, distinguished between native and non-native phonemes, and were functionally connected with superior temporal cortex. Taken together, these findings constitute evidence for the sensorimotor nature of speech perception.