NeuroImage 178 (2018) 574–582

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Neural representation of vowel formants in tonotopic auditory cortex

Julia M. Fisher [a,b], Frederic K. Dick [c,d,e], Deborah F. Levy [f], Stephen M. Wilson [f,*]

[a] Department of Linguistics, University of Arizona, Tucson, AZ, USA
[b] Statistics Consulting Laboratory, BIO5 Institute, University of Arizona, Tucson, AZ, USA
[c] Department of Psychological Sciences, Birkbeck College, University of London, UK
[d] Birkbeck-UCL Center for Neuroimaging, London, UK
[e] Department of Experimental Psychology, University College London, UK
[f] Department of Hearing and Speech Sciences, Vanderbilt University Medical Center, Nashville, TN, USA

ARTICLE INFO

Keywords:
Vowels
Formants
Tonotopy
Auditory cortex

ABSTRACT

Speech sounds are encoded by distributed patterns of activity in bilateral superior temporal cortex. However, it is unclear whether speech sounds are topographically represented in cortex, or which acoustic or phonetic dimensions might be spatially mapped. Here, using functional MRI, we investigated the potential spatial representation of vowels, which are largely distinguished from one another by the frequencies of their first and second formants, i.e. peaks in their frequency spectra. This allowed us to generate clear hypotheses about the representation of specific vowels in tonotopic regions of auditory cortex. We scanned participants as they listened to multiple natural tokens of the vowels [ɑ] and [i], which we selected because their first and second formants overlap minimally. Formant-based regions of interest were defined for each vowel based on spectral analysis of the vowel stimuli and independently acquired tonotopic maps for each participant. We found that perception of [ɑ] and [i] yielded differential activation of tonotopic regions corresponding to formants of [ɑ] and [i], such that each vowel was associated with increased signal in tonotopic regions corresponding to its own formants. This pattern was observed in Heschl's gyrus and the superior temporal gyrus, in both hemispheres, and for both the first and second formants. Using linear discriminant analysis of mean signal change in formant-based regions of interest, the identity of untrained vowels was predicted with ~73% accuracy. Our findings show that cortical encoding of vowels is scaffolded on tonotopy, a fundamental organizing principle of auditory cortex that is not language-specific.

Introduction

Cortical encoding of speech sounds has been shown to depend on distributed representations in auditory regions on Heschl's gyrus (HG) and the superior temporal gyrus (STG). Studies using functional MRI (Formisano et al., 2008; Obleser et al., 2010; Kilian-Hütten et al., 2011; Bonte et al., 2014; Arsenault and Buchsbaum, 2015; Evans and Davis, 2015; Zhang et al., 2016) and intracranial electrocorticography (Chang et al., 2010; Pasley et al., 2012; Chan et al., 2014; Mesgarani et al., 2014; Leonard et al., 2016; Moses et al., 2016) have shown that phonemes can be reconstructed and discriminated by machine learning algorithms based on the activity of multiple voxels or electrodes in these regions. Neural data can distinguish between vowels (Formisano et al., 2008; Obleser et al., 2010; Bonte et al., 2014; Mesgarani et al., 2014) and between consonants (Chang et al., 2010; Mesgarani et al., 2014; Arsenault and Buchsbaum, 2015; Evans and Davis, 2015), and there is evidence that phonemic representations in these regions are categorical and reflect the contribution of top-down information (Chang et al., 2010; Kilian-Hütten et al., 2011; Bidelman et al., 2013; Mesgarani et al., 2014; Leonard et al., 2016).

However, little is known regarding the spatial organization of cortical responses that underlie this distributed encoding, even in cases where hypotheses can readily be made based on known principles of auditory cortical organization. The most prominent organizing principle of core auditory regions is tonotopy, whereby there are several continuous gradients between regions in which neurons preferentially respond to lower or higher frequencies (Talavage et al., 2004; Woods et al., 2009; Humphries et al., 2010; Da Costa et al., 2011; Dick et al., 2012; Saenz and

* Corresponding author. Department of Hearing and Speech Sciences, Vanderbilt University Medical Center, 1215 21st Ave S, MCE 8310, Nashville, TN, 37232, USA.
E-mail address: stephen.m.wilson@vanderbilt.edu (S.M. Wilson).

https://doi.org/10.1016/j.neuroimage.2018.05.072
Received 13 February 2018; Received in revised form 29 May 2018; Accepted 30 May 2018
Available online 31 May 2018

Langers, 2013; De Martino et al., 2015). Tonotopic organization also extends to auditory regions beyond the core on the lateral surface of the STG and beyond (Striem-Amit et al., 2011; Moerel et al., 2012, 2013; Dick et al., 2017).

Vowels are pulse-resonance sounds in which the vocal tract acts as a filter, imposing resonances on the glottal pulses, which appear as peaks on the frequency spectrum. These peaks are referred to as formants, and vowels are distinguished from one another largely in terms of the locations of their first and second formants (Peterson and Barney, 1952), which are quite consistent across speakers despite variation in the pitches of their voices, and across pitches within each individual speaker. Because formants are defined in terms of peak frequencies, we hypothesized that vowels may be discriminable based on neural activity in tonotopic regions corresponding to the formants that characterize them.

In animal studies, perception of vowels is associated with increased firing rates of frequency-selective neurons in primary auditory cortex (Versnel and Shamma, 1998; Mesgarani et al., 2008). In humans, natural sounds are encoded by multiple spectrotemporal representations that differ in spatial and temporal resolution (Moerel et al., 2012, 2013; Santoro et al., 2014) such that spectral and temporal modulations relevant for speech processing can be reconstructed from functional MRI data acquired during presentation of natural sounds (Santoro et al., 2017). Therefore it can be predicted that the cortical encoding of vowels, as a special case of natural sounds, would follow the same principles. However, the cortical representation of vowel formants in tonotopic regions has not previously been demonstrated. Magnetoencephalography (MEG) studies have shown differences in source localization between distinct vowels (Obleser et al., 2003, 2004; Scharinger et al., 2011), but findings have been inconsistent across studies (Manca and Grimaldi, 2016), so it is unclear whether any observed differences reflect tonotopic encoding of formants. Neuroimaging studies have almost never reported activation differences between different vowels in univariate subtraction-based analyses (e.g. Formisano et al., 2008; Obleser et al., 2010). As noted above, the imaging and electrocorticography studies that have demonstrated neural discrimination between vowels have done so on the basis of distributed representations (e.g. Formisano et al., 2008; Mesgarani et al., 2014). The patterns of voxels or electrodes contributing to these classifications have been reported to be spatially dispersed (Mesgarani et al., 2014; Zhang et al., 2016).

To determine whether vowel formants are encoded by tonotopic auditory regions, we used functional MRI to map tonotopic auditory cortex in twelve healthy participants, then presented blocks of the vowels [ɑ] (the first vowel in 'father') and [i] (as in 'peak') in the context of an irrelevant speaker identity change detection task. We examined neural responses to the two vowels in regions of interest where voxels' best frequencies corresponded to their specific formants, to determine whether vowel identity could be reconstructed from formant-related activation.

Materials and methods

Participants

Twelve neurologically normal participants were recruited from the University of Arizona community in Tucson, Arizona (age 32.0 ± 5.9 (sd) years, range 26–44 years; 7 male, 5 female; all right-handed; all native speakers of English; education 17.8 ± 1.6 years, range 16–20 years). All participants passed a standard hearing screening (American Speech-Language-Hearing Association, 1997).

All participants gave written informed consent and were compensated for their time. The study was approved by the institutional review board at the University of Arizona.

Structural imaging

MRI data were acquired on a Siemens Skyra 3 T scanner with a 32-channel head coil at the University of Arizona. A whole-brain T1-weighted magnetization-prepared rapid acquisition gradient echo (MPRAGE) image was acquired with the following parameters: 160 sagittal slices; slice thickness = 0.9 mm; field of view = 240 × 240 mm; matrix = 256 × 256; repetition time (TR) = 2.3 s; echo time (TE) = 2.98 ms; flip angle = 9°; GRAPPA acceleration factor = 2; voxel size = 0.94 × 0.94 × 0.94 mm.

Cortical surfaces were reconstructed from the T1-weighted MPRAGE images using Freesurfer version 5.3 (Dale et al., 1999) running on Linux (xubuntu 16.04). Four surface-based anatomical regions of interest (ROIs) were defined using automated cortical parcellation (Fischl et al., 2004). Specifically, HG and the STG were identified in the left and right hemispheres based on the Desikan-Killiany atlas (Desikan et al., 2006).

Tonotopic mapping

Two functional runs were acquired to map tonotopic regions of auditory cortex in each participant. To engage both primary and non-primary auditory areas in meaningful processing (Moerel et al., 2012), the stimuli consisted of bandpass-swept human vocalizations, as previously described by Dick et al. (2012). In brief, vocalization tokens were produced by actors who were instructed to express eight different emotions using the French vowel [ɑ] (Belin et al., 2008). The tokens were spliced together to form sequences of 8 m 32 s. These sequences were then bandpass filtered in eight ascending or descending sweeps of 64 s each. Each sweep involved a logarithmic ascent of the center frequency from 150 Hz to 9600 Hz, or a similar descent. Although the vocalization tokens used the vowel [ɑ], the filtering ensured that there was no trace of the formants of [ɑ] in the tonotopic stimuli. The stimuli were then filtered again to compensate for the acoustic transfer function of the earphones (see below), and were presented at a comfortable level for each participant. To ensure attention to the stimuli, participants were asked to press a button whenever they heard the sound of laughter, which was one of the eight emotional sounds. Additional details are provided in Dick et al. (2012).

Auditory stimuli were presented using insert earphones (S14, Sensimetrics, Malden, MA) padded with foam to attenuate scanner noise and reduce head movement. Visual stimuli (consisting only of a fixation crosshair for the tonotopic runs) were presented on a 24″ MRI-compatible LCD monitor (BOLDscreen, Cambridge Research Systems, Rochester, UK) positioned at the end of the bore, which participants viewed through a mirror mounted to the head coil. Button presses were collected via a fiber optic button box (Current Designs, Philadelphia, PA) placed in the right hand. Stimuli were presented and responses recorded with custom scripts written using the Psychophysics Toolbox version 3.0.10 (Brainard, 1997; Pelli, 1997) in MATLAB R2012b (Mathworks, Natick, MA).

One ascending run and one descending run were acquired. T2*-weighted BOLD echo planar images were collected with the following parameters: 256 volumes; 28 axial slices in interleaved order, parallel to the Sylvian fissure and spanning the temporal lobe; slice thickness = 2 mm with no gap; field of view = 220 × 220 mm; matrix = 110 × 110; TR = 2000 ms; TE = 30 ms; flip angle = 90°; voxel size = 2 × 2 × 2 mm. An additional 10 volumes were acquired and discarded at the beginning of each run, to allow for magnetization to reach steady state and to avoid auditory responses to the onset of scanner noise.

The functional data were preprocessed with tools from AFNI (Cox, 1996). The data were resampled to account for differences in slice acquisition times. Head motion was corrected, with six translation and rotation parameters saved for use as covariates. In the course of head motion correction, all functional runs were aligned with the last volume of the last tonotopy run, which was acquired closest to the structural scan. Then the data were detrended with a Legendre polynomial of degree 2. The functional images were aligned with the structural images using *bbregister* in Freesurfer, and manually checked for accuracy. No