NeuroImage 79 (2013) 351–360

Contents lists available at SciVerse ScienceDirect

NeuroImage

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

Mean-based neural coding of voices

Attila Andics [a,b,c,*], James M. McQueen [a,d], Karl Magnus Petersson [a,b,e]

[a] Max Planck Institute for Psycholinguistics, Nijmegen, P.O. Box 310, 6500 AH, The Netherlands
[b] Donders Institute for Brain, Cognition and Behaviour, Centre for Cognitive Neuroimaging, Radboud University Nijmegen, P.O. Box 9104, 6500 HE, The Netherlands
[c] Comparative Ethological Research Group, Hungarian Academy of Sciences, Eötvös Loránd University, Budapest, Pázmány Péter sétány 1/c, 1117, Hungary
[d] Behavioural Science Institute and Donders Institute for Brain, Cognition and Behaviour, Centre for Cognition, Radboud University Nijmegen, P.O. Box 9104, 6500 HE, The Netherlands
[e] Cognitive Neuroscience Research Group, Institute for Biotechnology & Bioengineering, CBME, University of Algarve, Faro, Campus de Gambelas, 8500-139, Portugal

ARTICLE INFO

Article history:
Accepted 4 May 2013
Available online 9 May 2013

Keywords:
fMRI
Inferior frontal cortex
Prototype-centered representations
Superior temporal sulcus
Voice identity learning

ABSTRACT

The social significance of recognizing the person who talks to us is obvious, but the neural mechanisms that mediate talker identification are unclear. Regions along the bilateral superior temporal sulcus (STS) and the inferior frontal cortex (IFC) of the human brain are selective for voices, and they are sensitive to rapid voice changes. Although it has been proposed that voice recognition is supported by prototype-centered voice representations, the involvement of these category-selective cortical regions in the neural coding of such "mean voices" has not previously been demonstrated. Using fMRI in combination with a voice identity learning paradigm, we show that voice-selective regions are involved in the mean-based coding of voice identities. Voice typicality is encoded on a supra-individual level in the right STS along a stimulus-dependent, identity-independent (i.e., voice-acoustic) dimension, and on an intra-individual level in the right IFC along a stimulus-independent, identity-dependent (i.e., voice identity) dimension. Voice recognition therefore entails at least two anatomically separable stages, each characterized by neural mechanisms that reference the central tendencies of voice categories.

Introduction

Human listeners can recognize individuals from their voices (i.e., auditory percepts of human vocalizations) alone and can rapidly learn new voice identities (i.e., voice-based percepts of person identity). Cortical regions involved in voice recognition have been mapped out, but it is not yet known how those regions represent voice knowledge. Here we test the hypothesis that in category-selective regions voice identities are represented in a prototype-centered voice processing hierarchy. In particular, we ask whether and how cortical activity reflects typicality in newly-learned voice categories. We will refer to this as mean-based neural coding of voices.

Two cortical regions have been reported to be sensitive to conspecifics' vocalizations. These regions are intriguingly similar in the primate and human brain and include regions along the superior temporal sulcus (STS) (in macaques: Petkov et al., 2008; in humans: Belin et al., 2000, 2011; Ethofer et al., 2009b; Grandjean et al., 2005) and the inferior frontal cortex (IFC) (in macaques: Romanski and Goldman-Rakic, 2002; Romanski et al., 2005; in humans: Fecteau et al., 2005; von Kriegstein and Giraud, 2006). Strong anatomical and functional connections have been found between the STS and the ipsilateral IFC in both primates (Hackett et al., 1998; Romanski et al., 1999) and humans (Ethofer et al., 2012). Furthermore, STS and IFC are not only voice-selective but also sensitive to short-term voice stimulus similarity, as demonstrated in rapid fMRI adaptation and carryover effects (STS: Andics et al., 2010, 2013; Belin and Zatorre, 2003; Latinus et al., 2011; Wong et al., 2004; IFC: Andics et al., 2010, 2013; Latinus et al., 2011). Short-term sensitivity here refers to mechanisms typically active within the range of a few seconds (cf., short-term repetition suppression, Epstein et al., 2008). This short-term sensitivity for voice similarity is an important requirement for the ability to tune in to voice stimuli, but it is not sufficient for the representation of long-term voice knowledge. Long-term here refers to processes relying on representations that need to be stored for longer than a few seconds (cf., long-term repetition suppression, Epstein et al., 2008). We adopt this definition in the present study. Neural storage of voice knowledge in the much longer term (e.g. weeks, months) is a topic for future research. Although it seems plausible that category-selective cortical regions are there to represent category knowledge for more than just in the short term, there is little evidence so far that the voice-selective STS and IFC contribute to representing voice knowledge for more than a few seconds.

This study asks whether the STS and IFC perform this function and elaborates on the recent proposal that long-term voice knowledge is represented in the human brain in a prototype-centered way. Mean-based neural coding appears to be a powerful way to represent individual stimuli in a category space (e.g., Panis et al., 2011). A possible mechanism for mean-based coding is neural sharpening (Hoffman and Logothetis, 2009): the coding of central values in relevant object dimensions becomes sparser with more experience. Neural sharpening reflects long-lasting cortical plasticity and so could be

* Corresponding author at: Max Planck Institute for Psycholinguistics, P.O. Box 310, 6500 AH Nijmegen, The Netherlands. Fax: +31 24 3521213.
E-mail address: attila.andics@gmail.com (A. Andics).

1053-8119/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.neuroimage.2013.05.002

used for positioning stimuli in long-term object spaces. For faces, mean-based coding was found behaviorally (Leopold et al., 2001; Rhodes and Jeffery, 2006), in primates (Leopold et al., 2006), and also with human fMRI localizing the mechanism in face-selective fusiform regions (Loffler et al., 2005). It has been argued that mean-based coding can also result from long-term adaptation (Kahn and Aguirre, 2012), a mechanism that is sensitive to stimulus distributions. Recent behavioral (Bruckert et al., 2010; Latinus and Belin, 2011; Latinus et al., 2009; Mullennix et al., 2009; Papcun et al., 1989) and neuroimaging studies (Andics et al., 2010) also suggest that there is mean-based coding for voices. In other words, voice representations appear to be centered around prototypes in long-term memory.

Long-term mean-based coding for voices has nevertheless not yet been demonstrated in voice-selective cortical regions. Andics et al. (2010) found mean-based coding for voices in several regions, but some of these regions (the deep posterior STS and the orbital/insular cortex) are not voice-selective. Other regions (the amygdala and the anterior temporal pole) appear to be involved in the multimodal integration of person identity rather than in pure voice identity processing (Andics et al., 2010; Belin et al., 2011; Latinus et al., 2011). Although recent findings suggested IFC involvement in the representation of long-term stored objects (Latinus et al., 2009), to date there is thus no evidence for long-term mean-based voice encoding in the core category-selective cortical regions, namely the STS and the IFC.

It has been proposed that voice recognition involves not only mean-based voice encoding but also separate processing stages for voice-acoustic and voice identity analysis (Belin et al., 2004, 2011; Bestelmeyer et al., 2012; Charest et al., 2013; Scott and Johnsrude, 2003). This proposal, however, has received little direct support so far in the form of functional-anatomical correspondences between voice-processing stages and voice-selective regions. In the framework of mean-based coding, voice-acoustic analysis corresponds to an identity-independent, supra-individual representation of voice typicality, while voice identity analysis corresponds to an identity-dependent, intra-individual representation of voice typicality. These definitions will be adopted in the present study. Note that typicality is thus defined here with respect to the materials in the experiment, and not judgments of typicality collected, for example, in a rating study.

Recently, Latinus et al. (2011) attempted to dissociate acoustic from identity effects in voice processing, but their design focused on short-term effects of acoustic and identity changes. Short-term acoustic processing was found in both the STS and the IFC and short-term identity processing was found in the IFC only. These short-term effects may be indicators of long-term voice processing mechanisms, but those mechanisms have not yet been tested directly. The present study therefore tested the hypothesis that long-term mean-based voice encoding is present both at voice-acoustic (supra-individual) and at voice identity (intra-individual) levels of processing, and aimed to specify the role of the two core voice-selective cortical regions in these two levels.

We performed an fMRI experiment using a within-subject voice-training paradigm. Listeners were trained on two consecutive weeks to categorize voice stimuli on a voice morph continuum as belonging to either of two talkers characterized by the two continuum endpoints (morph0, morph100). During training the entire continuum was sampled and the acoustic center of the trained stimulus space was identical across weeks (morph50). The feedback during training on week1 and week2 specified different voice identity category boundary locations on each week (morph36 or morph64). After each training session, we could separately manipulate two perceptual properties of the voice stimuli: their perceived acoustic centrality (i.e., degree of prototypicality defined by the acoustic space, independent of identity feedback) and their perceived identity centrality (i.e., degree of prototypicality of a new voice identity, as defined by a voice-training procedure, independent of acoustic properties). Our design also allowed us to separately test for short-term effects (e.g., rapid adaptation indicating stimulus similarity sensitivity in the 0–5 second range) and long-term effects (e.g., long-term adaptation or neural sharpening indicating norm-based coding in the >5 second range) within a single experiment.

We hypothesized that cortical representations of the voice-acoustic space are organized along an acoustically central to acoustically peripheral dimension, and thus should not be modulated by voice identity feedback. Acoustically central stimuli should have sharper neural coding than acoustically peripheral stimuli and hence we predicted that there should be less activity for central than for peripheral stimuli in voice-acoustic regions. We also hypothesized that voice identity representations are organized along a feedback-defined typical to atypical dimension, and that this typicality is fully independent of voice-acoustic properties. We predicted that the activity of voice identity representations generated by identity-typical stimuli should therefore be less than the activity generated by atypical stimuli.

Material and methods

Participants

Eighteen Dutch female listeners (19–24 years) with no reported hearing disorders were paid to complete the experiment. Written informed consent was obtained from all participants. One person was excluded because of a failure to perform the task during training. Two further participants were excluded because of poor learning performance during training (i.e., voice identity categorization performance per morph level did not significantly differ from the 50% chance level in the final training block before scanning, one-sampled, two-tailed $t(14) < 1$, $p > .4$). The analyses presented here were based on the remaining 15 subjects.

Stimulus material

We selected two perceptually similar voices from a voice pool that contained recordings from young male nonsmoking adult native speakers of Dutch with no recognizable regional accents and no speech problems pronouncing Dutch monosyllables (Andics et al., 2007). The voices were unfamiliar to the listeners. Recordings were made in a soundproof booth using a Sennheizer Microphone ME62, a MultiMIX mixer panel, and Sony Sound Forge. All stimuli were digitized at a 16 bit/44.1 kHz sampling rate and were volume balanced using Praat software (Boersma and Weenink, 2007). A single token was selected per voice identity, of the word *mes* (knife). The two tokens were acoustically similar: average pitches were 122 Hz and 113 Hz, and stimulus lengths were 482 ms and 492 ms respectively.

We then created a voice morph continuum using the speech manipulating algorithms of STRAIGHT (Kawahara, 2006). The speech signals were decomposed into three parameters: an interference-free spectrogram, an aperiodicity map and a fundamental frequency (F0) trajectory. These parameters were then logarithmically interpolated segment by segment. Finally, a 100-step stimulus continuum with equidistant intermediate levels was resynthesized. The endpoints (levels morph0 and morph100) were also resynthesized. Average syllable duration was 487 ms (audio samples can be found at http://mpi.nl/people/andics-attila/research).

Training design

Listeners received multiple-phase voice identity training on two consecutive weeks. During the entire course of training, listeners were presented with words from the voice morph continuum and were instructed to make forced-choice decisions on talker identity after every word they heard. To allow initial assignment of talker names (Peter and Thomas) on response buttons to voice identities (voice A and voice B), listeners were presented three naturally