Norm-Based Coding of Voice Identity in Human Auditory Cortex

Marianne Latinus,1,2,5,* Phil McAleer,1 Patricia E.G. Bestelmeyer,1,3 and Pascal Belin1,4,5,*

1Institute of Neuroscience and Psychology, University of Glasgow, Glasgow G12 8QF, Scotland
2Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN 47405, USA
3School of Psychology, Bangor University, Bangor, Gwynedd LL57 2AS, UK
4Département de Psychologie, Université de Montréal, Montréal, QC H2V 2S9, Canada
5Institut des Neurosciences de La Timone, UMR 7289, CNRS & Université Aix-Marseille, 13005 Marseille, France

Summary

Listeners exploit small interindividual variations around a generic acoustical structure to discriminate and identify individuals from their voice—a key requirement for social interactions. The human brain contains temporal voice areas (TVA) [1] involved in an acoustic-based representation of voice identity [2–6], but the underlying coding mechanisms remain unknown. Indirect evidence suggests that identity representation in these areas could rely on a norm-based coding mechanism [4, 7–11]. Here, we show by using fMRI that voice identity is coded in the TVA as a function of acoustical distance to two internal voice prototypes (one male, one female)—approximated here by averaging a large number of same-gender voices by using morphing [12]. Voices more distant from their prototype are perceived as more distinctive and elicit greater neuronal activity in voice-sensitive cortex than closer voices—a phenomenon not merely explained by neuronal adaptation [13, 14]. Moreover, explicit manipulations of distance-to-mean by morphing voices toward (or away from) their prototype elicit reduced (or enhanced) neuronal activity. These results indicate that voice-sensitive cortex integrates relevant acoustical features into a complex representation referenced to idealized male and female voice prototypes. More generally, they shed light on remarkable similarities in cerebral representations of facial and vocal identity.

Results

Two difficulties arise when approaching the complex problem of voice identity representation: the high dimensionality of the “voice space,” i.e., the large number of acoustical dimensions potentially differentiating speakers, and the dynamic, ever-changing nature of voices. We examined a simpler version of the problem by first using brief syllable stimuli for which the influence of time is minimal and by focusing on a small number of key acoustical measures. Three acoustical dimensions were selected based on their relevance to voice production and perception [11, 15–23] (Figure 1; see also Figure S1 available online): the fundamental frequency of phonation (f0, related to the pitch of voice), formant dispersion (FD, the average frequency difference between formants, related to vocal tract size [17]), and the harmonics-to-noise ratio (HNR, a measure of spectrotemporal regularity); together they defined a three-dimensional acoustical voice space (Figure 2A).

Experiment 1

In Experiment 1, recordings from 32 male and 32 female adult speakers uttering the syllable “had” [24] were subjected to a series of acoustical measures (Table S1; Figure S1; Supplementary Experimental Procedures). Each voice stimulus was represented as a point in the voice space with coordinates corresponding to the stimulus’ average f0, FD, and HNR, Z scored by gender (Figure 2A). We generated male and female prototypical voice stimuli by averaging all 32 same-gender voices by using morphing (Figure 2B). The resulting prototypical voices were characterized by (gender-specific) average f0 and FD values but high HNR values as averaging smoothed out spectrotemporal irregularities (Figure 2A; Table S1); as shown in Figure 2A, they are not located at the barycenter but rather on top of the voice stimulus clouds. Voice composites were also generated for each gender by averaging 2 (n = 16), 4 (n = 8), 8 (n = 4), and 16 (n = 2) different voice stimuli, for a total of 126 stimuli. The Euclidean distance between each stimulus and the gender-specific prototypical voice defined a voice’s “distance-to-mean” (Figures 2A and 2C).

We first asked whether distance-to-mean was related to a behavioral measure of how distinctive each voice sounds to listeners. Normal adult listeners rated each voice for its perceived distinctiveness; this set of voxels (n = 1,096) defined the group-level TVA mask. BOLD signal was then correlated with distance-to-mean (p < 0.001; Spearman’s ρ [CI95%] = 0.73 [0.62 0.81]; R² = 0.53; Figure 2 D; Figure S2 A): correlated with distance-to-mean (p < 0.001; Spearman’s ρ [CI95%] = 0.73 [0.62 0.81]; R² = 0.53; Figure 2 D; Figure S2 A):

Please cite this article in press as: Latinus et al., Norm-Based Coding of Voice Identity in Human Auditory Cortex, Current Biology (2013), http://dx.doi.org/10.1016/j.cub.2013.04.055

*Correspondence: mlatinus@gmail.com (M.L.), pascal.belin@glasgow.ac.uk (P.B.)
Significance was assessed by using a permutation test at each voxel, and corrections for multiple comparison were based on the maximum significant threshold over the whole set of voxels ($p = 0.18$, Supplemental Experimental Procedures) [26, 28].

TVA voxels in both the left (LH) and right (RH) hemispheres showed significant positive correlations at the group level between BOLD signal estimates and distance-to-mean voice stimuli at greater distance from the same-gender prototypical voice eliciting greater BOLD signal (Figure 3B; Table S2). Significant rho values ($p < 0.05$; range $= [0.18, 0.47]$) reached their maximum along right mid-STS ($p_{[95\%]} = 0.47 \pm [0.30, 0.61]$; MNI coordinates $[63 -9 -6]$, Figure 3C). Crucially, distances defined relative to a single, androgynous prototype instead of the two gender-specific prototypes (“distance-to-overall-barycentre”) never explained variance in BOLD signal better than distance-to-mean defined relative to gender-specific prototypes (Supplemental Information). Significant correlations between distance-to-mean and BOLD activity could also be observed at the single participant level, indicating the robustness of the effect (Figure S3). Outside the TVA, at a location close to primary auditory cortex (identified as the maximum of activity in the contrast of all sounds versus silence during the voice localization; MNI coordinates $[51 -6 0]$, correlations between BOLD and distance-to-mean were just below significance ($p_{[95\%]} = 0.17 \pm [-0.01, 0.35]$). In addition, we found that distance-to-mean defined in simpler spaces with fewer dimensions also correlates with TVA activity, indicating that the effect is not restricted to that particular three-dimensional acoustic space (Figure S2B).

These results are consistent with the notion of norm-based coding of vocal identity, but they could also arise from neuronal adaptation effects, i.e., reductions in neuronal activity in response to repeated stimulation [14], that can in some conditions be mistakenly interpreted as prototype effects [13]. Further analyses were run on each voxel of the TVA to disentangle norm-based coding from adaptation effects. Briefly, different regressors were used to model adaptation at different time scales: (1) “distance-to-preceding-stimulus” to model short-term adaptation effects ($\sim 4$ s) occurring between consecutive stimuli during scanning, known to depend on physical similarity; (2) “distance-to-barycentre,” i.e., distance between each voice and the center of the (same-gender) voice cloud, that represent the average position of stimuli presented during an entire block ($\sim 5$ min) to model medium-term adaptation effects; and (3) “distance-to-overall-barycentre” (Figure 3D), i.e., to the average position of all male and female stimuli presented during the experiment ($\sim 20$ min) to model long-term adaptation. As expected, adaptation effects at different time ranges were found to occur in the TVA (Supplemental Information). Crucially, distance-to-mean resulted in significantly stronger correlation with BOLD signal than the adaptation regressors. Conversely, not a single voxel showed significantly larger correlation with either of the adaptation regressors (Supplemental Information). Thus, Experiment 1 provides strong evidence for a relation between TVA activity and acoustical similarity to (gender-specific) voice prototypes that is not simply explained by adaptation effects.

**Experiment 2**

We next sought to replicate these results with different participants and stimuli. In Experiment 2, stimuli consisted of recordings of the word “hello” spoken by 32 male and 32 female adult speakers; i.e., recordings included only natural, unmanipulated stimuli. A new group of healthy adult volunteers was scanned while listening to individual stimuli and performing a pure tone detection task. Male and female prototypical stimuli were generated, following the same procedure as in Experiment 1 (Audio S1), to compute distance-to-mean values for each voice in the three-dimensional acoustical voice space (Table S1). Preprocessing steps and statistical analyses were similar to those of Experiment 1. Again, significant correlations between distance-to-mean and beta estimates of activity were found in several TVA voxels ranging from 0.25 to 0.36 with a maximum in the right hemisphere ($p_{[95\%]} = 0.36 \pm [0.13, 0.55]$, Figure 3D; Table S2). Thus, results of Experiment 1 appear generalizable to other participants and stimulus sets.

**Experiment 3**

Although the correlations observed in Experiments 1 and 2 support the norm-based model, they might still conceivably be caused by other, unknown factors in the stimulation related to both neuronal activity and distance-to-mean. More compelling evidence would be obtained by explicitly manipulating distance-to-mean and testing whether the differences predicted by the prototype-based model are observed. We ran a third experiment in which we used morphing to directly manipulate distance-to-mean while minimizing short-term adaptation effects by equating average distance-to-preceding-stimulus across conditions (Supplemental Experimental Procedures).

For male and female voices independently, the 16 stimuli...
from the Experiment 1 stimulus set with intermediate distance-to-mean values were each morphed with the same-gender prototypical voice to generate a “contracted” (50% closer) and “dilated” (50% away) version of each stimulus (Figures 4A and 4B; Table S1; Audio S2). A new group of normal adult volunteers was scanned while listening to blocks of contracted or dilated stimuli and performing a pure tone detection task. At each TVA voxel, we computed the difference between the BOLD signals for “dilated” and “contracted” blocks. Statistical analyses, similar to those of Experiments 1 and 2, revealed differences in BOLD signal in TVA voxels (Table S2) consistent with the predictions of the norm-based model: voice stimuli induced greater TVA activity when morphed away from the prototype than when morphed toward the prototype by a same amount of acoustical change (significant threshold after correction for multiple comparisons = 0.58; range of significant differences = [0.58 1.00]). Here as well, adaptation effects did not simply explain the observed pattern of results (Figure 4C; Supplemental Information).

Discussion

Like faces, voices can be used to identify a person, yet the neural bases of this ability remain poorly understood. Here we...
provide the first evidence of a norm-based coding mechanism in the auditory modality. Across three experiments using different participants and stimulus sets, we find that voices more distant from (dissimilar to) the same-gender prototypical voice in a three-dimensional f0-FD-HNR acoustical space are perceived as more distinctive, and elicit greater activity in voice in a three-dimensional f0-FD-HNR acoustical space are more distant from (dissimilar to) the same-gender prototypical voice in a three-dimensional f0-FD-HNR acoustical space. The true voice space is likely to include a larger number of more complex representations [30]. These representations appear to be referenced to an internal prototype. They also demonstrate that information on speaker identity is not only distributed over large zones of cortex as suggested by multivariate analyses [5] but can also be encoded locally, because correlations with distance-to-mean were observed at the level of individual TVA voxels. Whether similar mechanisms can be observed for other categories of auditory objects remains to be determined.

The results bring to light interesting similarities in coding mechanisms across sensory modalities [31]. Converging evidence from psychophysical, electrophysiological, and neuroimaging studies indeed indicates that individual faces are represented in part as a function of their distance to a prototypical, average face in a multidimensional “face space” [32–35]. Although the exact nature of the face prototype remains unclear [36], this sparse coding mechanism is thought to offer several advantages including a minimization of energy consumption in response to natural stimulation and an elegant solution to overcome the problem of certain transformations associated with, e.g., viewpoint change or aging. Despite the highly different nature of the sensory input from faces and voices, an analogous mechanism appears to be used to represent a person’s identity across sensory modalities. This does not imply that all cortical processing is similar across vision and audition but illustrates a parsimonious principle of brain organization given the similar nature of the computational problems posed by face and voice identity processing and the fact that information has to be integrated across senses in everyday life [37].

Important information was obtained on the characteristics of the prototypical stimuli, a question that still eludes research on face perception [36]. There are two voice prototypes: one male and one female. A model defining distance-to-mean relative to a single, androgynous prototype explained virtually no variance. The male and female prototypical voices appear well approximated by the morphing-generated average of many same-gender voices, a process resulting in voices with (gender) average f0 and formant frequencies but with high HNR values; i.e., a very common voice but ideally regular and flawless, perceived as highly attractive by listeners [8, 9]. This unanticipated feature of the prototypical voices allows demonstrating norm-based coding and differentiates its effects from those of adaptation occurring over the medium- to long-term (cf. Supplemental Information).

We find that distance-to-mean defined in simpler (two- or one-dimensional) spaces also predicts TVA activity and perceptual ratings of distinctiveness, indicating that the f0-FD-HNR acoustical space used here is not the only valid voice space. The “true” voice space is likely to include a larger number of more complex dimensions, consistent with the intricacy of the voice production apparatus and the many associated

Figure 3. Cerebral Activity in Voice-Sensitive Cortex Correlates with Distance-to-Mean

(A) TVA showing significantly greater fMRI signal in response to vocal versus nonvocal sounds at the group-level used as a mask for further analysis. Color scale indicates T values of the vocal versus nonvocal contrast.

(B) Maps of Spearman correlation between beta estimates of BOLD signal in response to each voice stimulus and its distance-to-mean overlay on the TVA map (black). Color scale indicates significant values (p < 0.05 corrected for multiple comparisons). Note a bilateral distribution with a maximum along the right anterior STS. See Figure S3 for correlation maps in 8 individuals.

(C) Scatterplots and regression lines between estimates of BOLD signal and distance-to-mean at the peak voxel in Experiment 1.

(D) Scatterplots and regression lines between estimates of BOLD signal and distance-to-mean at the peak voxel observed in Experiment 2. Scatterplots are shown for illustration only. See Audio S1 for example of stimuli used in Experiment 2.
acoustical dimensions [15, 17–19] (Figure 1; Figures S2A and S2B). Nevertheless, the f0-FD-HNR space appears an adequate approximation of the true voice space allowing the estimation of a voice’s distance-to-mean from a small number of easily measured acoustical variables.

Note that our results were obtained in the context of a large number of unfamiliar voices. Whether similar mechanisms are involved in coding the identity of familiar voices remains to be established [38], although clinical studies of voice perception [39–41] and face perception studies [42] suggest qualitatively different mechanisms for familiar stimuli. Likewise, the extent to which these results generalize to longer, more complex utterances representative of more natural conditions of conversation remains to be investigated.

These results in turn lead to a range of important new questions. Are the prototypes innate, stored templates? Or are they susceptible to environmental and cultural influences? If yes, to what extent? Could the prototype consist of an average of all voices experienced during one’s life, in which case the frontier between prototype-based coding and long-term adaptation would become blurry? Can similar coding principles be observed in the brain of nonhuman primates, as has been found for faces [35]? While the answers to these questions are yet unknown, our results provide the first evidence of norm-based coding of voice identity in human auditory cortex, a finding with potentially useful applications in voice and speech signal processing [43]. These studies bring to light similarities in encoding strategies between sensory modalities, but also define important characteristics of the internal voice prototypes: they are ideally smooth versions of the male and female prototypical voice.

Supplemental Information

Supplemental Information includes four figures, two tables, Supplemental Data, Supplemental Experimental Procedures, and two audio files and can be found with this article online at http://dx.doi.org/10.1016/j.cub.2013.04.055.

Acknowledgments

This work was supported by BBSRC grants BB/E003958/1 and BB/I022287/1 and ESRC/MRC large grant RES-060-25-0010. The authors are grateful to Marie-Hélène Grosbras, David D. Leopold, Chris I. Petkov, Stefan Schwendenberger, Simon Thorpe, Rufin VanRullen, Gunter Loffler, Scott A. Love, and Robert J. Zatorre for insightful comments and to Hideki Kawahara and Julien Rouger for invaluable help with voice morphing. The authors are thankful for the contributions of anonymous reviewers.

Received: December 4, 2012
Revised: April 4, 2013
Accepted: April 22, 2013
Published: May 23, 2013

References


