Dynamic Norm-Based Coding of Vocal Identity by the Primate Voice Patches Neurons

Yoan Esposito (yoan.esposito@univ-amu.fr)
Institut de Neurosciences de la Timone, Aix-Marseille Universite, Marseille, France

Louis-Jean Boe (louisjean.boe@orange.fr)
GIPSA-Lab CNRS, Universite de Grenoble, Grenoble, France

Regis Trapeau (regis.trapeau@univ-amu.fr)
Institut de Neurosciences de la Timone, Aix-Marseille Universite, Marseille, France

Luc Renaud (luc.renaud@univ-amu.fr)
Institut de Neurosciences de la Timone, Aix-Marseille Universite, Marseille, France

Matthieu Gilson (matthieu.gilson@univ-amu.fr)
Institut de Neurosciences de la Timone, Aix-Marseille Universite, Marseille, France

Margherita Giamundo (margherita.giamundo@univ-amu.fr) *
Institut de Neurosciences de la Timone, Aix-Marseille Universite, Marseille, France

Pascal Belin (pascal.belin@univ-amu.fr) *
Institut de Neurosciences de la Timone, Aix-Marseille Universite, Marseille, France

(*) denotes equal senior authorship

Abstract

The neural representation of conspecific vocal identity remains poorly understood in nonhuman primates. Drawing on norm-based coding theories of face identity, we investigated whether a similar mechanism supports voice identity encoding in the macaque brain. We recorded spiking activity from the fMRI-localized voice sensitive region of two awake macaques while presenting synthetic 'coo' vocalizations morphed along a continuum from antivoice to caricature. Dimensionality reduction techniques revealed distinct neuronal subpopulations: for example, one exhibited a very early V-shaped tuning profile, with increasing firing rates as voices deviated from the average, consistent with a norm-based coding strategy; another showed the exact inverse pattern, but with delayed responses. These temporally and functionally distinct subpopulations suggest complementary encoding strategies for representing vocal identity, potentially reflecting both deviation-from-mean and identity-certainty mechanisms. Our findings mirror encoding patterns observed in face-selective regions and provide novel evidence that norm-based coding may be a general principle of high-level social perception across sensory modalities.

Keywords: voice perception; norm-based coding; fMRI-guided electrophysiology; dimensionality reduction.

Introduction

Faces and voices of conspecifics are crucial cues in primate social communication, processed by specialized neural circuits in the temporal cortex (Belin, Zatorre, Lafaille, Ahad, & Pike, 2002). Electrophysiological studies in macaques have identified face and voice patches—clusters of neurons selectively responsive to facial or vocal stimuli (Tsao, Freiwald, Tootell, & Livingstone, 2006; Giamundo et al., 2024).However, the coding strategies used by these neurons remain poorly understood.

In the domain of face perception, evidence supports a norm-based coding (NBC) scheme, where neural responses scale with a stimulus's distance from an average face within a multidimensional "face space" (Leopold, O'Toole, Vetter, & Blanz, 2001; Chang & Tsao, 2017; Koyano, anb D. B. McMahon, Waidman, Russ, & Leopold, 2021). In particular, Koyano et al. showed that face patch neurons exhibit V-shaped tuning, with minimal firing for the average, increasing with distance from it. Whether a similar mechanism underlies voice encoding is unknown. Human fMRI studies suggest that voice sensitive regions show tuning to voice distinctiveness relative to an average (Latinus, McAleer, Bestelmeyr, & Belin, 2013), but direct neuronal evidence is lacking.

Here, we tested for NBC in the fMRI-localized voice-sensitive region of macaques. We recorded single-unit activity in response to parametrically morphed synthetic 'coo' vocalizations that varied in identity distance from an average voice. We predicted a V-shaped neuronal tuning, as observed for faces.

Methods

Auditory stimuli

We generated synthetic macaque "coo" vocalizations using STRAIGHTMORPH software (Belin & Kawahara, 2025). Stimuli were generated by morphing 16 coos to create an average voice, which was then used to construct six identity continua by morphing between the average and six individual coos. Each continuum spanned from anti-voices (–100%) to caricatures (200%), including the original coo (100%) and the average (0%). The full stimulus set included 49 sounds: eight morph levels across six trajectories, plus the shared average coo. The sounds were 467 ms, 22050 Hz mono.

fMRI-guided electrophysiology

We recorded spiking activity from hundreds of single neurons in two adult female rhesus monkeys (Macaca mulatta) using chronic multi-electrode Utah arrays. Voice-sensitive regions were localized using a previous fMRI localizer (Bodin et al., 2021), and arrays were implanted in the voice-selective regions of the rostral superior temporal gyrus. During the recordings, monkeys performed a pure tone detection task while listening to auditory stimuli.

Response time course

Neuronal activity was analyzed using 50 ms sliding windows with a step size of 10 ms, covering a time range from -100 ms to 500 ms relative to stimulus onset, resulting in a total of 60 bins. For each bin, the average firing rate of neurons was calculated and baseline normalized for each identity level. This approach allowed us to examine the temporal dynamics of neuronal responses with fine-grained resolution.

Non-Negative Matrix Factorization

We applied Non-Negative Matrix Factorization (NMF) to explore the decomposition of our neuronal population activity. Specifically, we computed the average firing rate of each neuron for each identity level within the time window from –100 ms to 500 ms relative to stimulus onset, resulting in a matrix of shape [m,t], where m is the number of identity levels and t the number of time bins. This matrix was then normalized by dividing all values by its maximum, yielding a normalized activity matrix with values ranging from 0 to 1. We performed this normalization independently for each neuron, resulting in a final data structure of shape [n,m,t], where n is the number of neurons, m the number of identity levels, and t the number of time bins. This matrix was reshaped to [n,m*t] for NMF computation. To identify the optimal number of factors, we employed a repeated Shuffle Split cross-validation procedure (n=50 splits, test size=50%), assessing the stability and generalization of the NMF decomposition. For each value of k (1 to 10 factors), the NMF model was trained on the training data and subsequently used to transform and reconstruct the held-out test set and the explained variance between the original matrix and the reconstructed one was then computed. The factorization matrix was finally used to reconstruct the average heatmap for each factor.

Results

We recorded single-neuron spiking activity from two female rhesus monkeys implanted with chronic Utah arrays in the fMRI-localized voice sensitive regions. During electrophysiological recordings, monkeys performed a pure tone detection task while hearing synthetic macaque “coo” vocalizations (Fig. 1A).

To characterize the temporal dynamics of identity encoding, we computed time-resolved population activity for each identity level. Neuronal firing rates were normalized (NFR) within a time window from –100 to 500 ms relative to stimulus onset, using 10 ms steps (Figure 1B). Individual neuron heatmaps were averaged to obtain a population-level representation (absolute NFR, Figure 1C). Around 50 ms post-stimulus, a general increase in activity was observed across conditions. By 100 ms, a distance-to-mean pattern emerged, with reduced responses to identity levels near the average. Interestingly, a rebound in activity appeared around 200 ms specifically for the average stimulus.

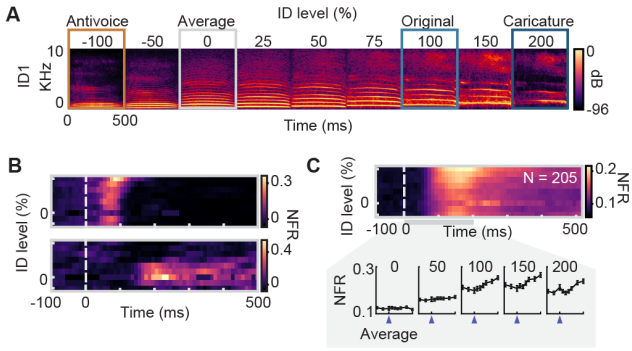

Figure 1: (A) Spectrograms continua created by morphing along the average and original voices (0 : average stimulus; 100 : original coo, -100 : antivoice, opposite acoustical features compared to original coo; 200 : caricature, exagered acoustical features compared to original coo). (B) Two examples neurons (x : time from stimulus onset, y : identity levels, color : NFR). (C) Top panel : population heatmap (axis are the same as in (B), color : absolute NFR, error bars : standard error of the mean). Bottom panel : population tuning curves (x : identity levels, y : absolute NFR)

To assess whether the distinct temporal dynamics observed in our dataset could be identified in a data-driven manner, we applied NMF to the full population activity (Figure 2A). Using shuffle-split cross-validation, we selected a four-factors solution, as adding more components led to overfitting and reduced generalization performance. We then used the component matrix to reconstruct the average activity heatmap for each factor. The results revealed that each factor captured distinct temporal and tuning profiles (Figure 2B). The average normalized activity of the top 20 neurons of each factor revealed consistent and factor-specific temporal and tuning profiles (Figure 2C). To assess sensitivity to distance-to-mean, we conducted a sliding-window Friedman test across four identity levels (–100, 0, 100, and 200) at each time bin. A subpopulation was considered sensitive to distance-from-mean if the test reached significance (FDR-corrected $p < 0.05$) for at least five consecutive bins. Factor 1 corresponded to neurons broadly suppressed following stimulus onset, showing no tuning to distance-from-mean. In contrast, the remaining three subpopulations displayed significant distance-to-mean sensitivity, each with distinct tuning profiles and temporal dynamics. Factor 4 reflected an early V-shaped tuning, with sensitivity emerging rapidly between 70-140 ms. Factor 2 exhibited a late-onset V tuning, peaking between 150-260 ms and 320-400 ms. Finally, factor 3 captured a late rebound of activity specific to the average stimulus, with significant sensitivity observed between 150–260 ms and 280–400 ms.

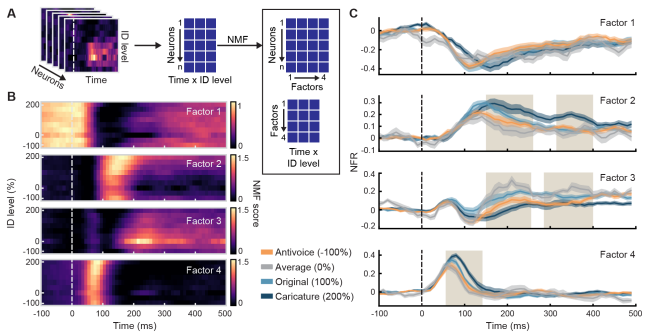

Figure 2: (A) Decomposition analysis methodology (NMF : Non-Negative Matrix Factorization). (B) Factor heatmaps. (C) Plot of the average population activity of the top 20 neurons of each factor (color shaded areas : standard error of the mean, grey shaded areas : significant distance-to-mean sensitivity).

Conclusion

Our findings suggest that macaque voice sensitive region encode vocal identity through a norm-based mechanism. Using data driven approach, we were able to highlight neuronal subpopulations that show distinct tuning and temporal properties in the encoding of the distance to the average stimulus, suggesting complementary encoding strategies for representing vocal identity.

Acknowlegments

This work was funded by Fondation pour la Recherche Medicale AJE201214; Agence Nationale de la Recherche grants ANR-16- CE37-0011-01 (PRIMAVOICE) ANR-16-CONV-0002 (Institute for Language, Communication and the Brain) and ANR-11-LABX-0036 (Brain and Language Research Institute); Excellence Initiative of Aix-Marseille University (AMIDEX) ; and European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement no. 788240).

References

Belin, P., & Kawahara, H. (2025). Straightmorph: A voice morphing tool for research in voice communication sciences. *Open Research Europe*, *4*.

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2002). Voice-selective areas in human 570 auditory cortex. *Nature*, *403*.

Bodin, C., Trapeau, R., Nazarian, B., Sein, J., Degiovanni, X., Baurberg, J., ... Belin, P. (2021). Functionally homologous representation of vocalizations in the auditory cortex of humans and macaques. *Current Biology*, *31*.

Chang, L., & Tsao, D. Y. (2017). The code for facial identity in the primate brain. *Cell*, *169*.

Giamundo, M., Trapeau, R., , Thoret, E., Renaud, L., Nougaret, S., ... Belin, P. (2024). A population of neurons selective for human voice in the monkey brain. *Proceedings of the National Academy of Sciences*, *121*.

Koyano, K. W., anb D. B. McMahon, A. P. J., Waidman, E. N., Russ, B. D., & Leopold, D. A. (2021). Dynamic suppression of average facial structure shapes neural tuning in three macaque face patches. *Current Biology*, *31*.

Latinus, M., McAleer, P., Bestelmeyr, P. E., & Belin, P. (2013). Norm-based coding of voice identity in human auditory cortex. *Current Biology*, *23*.

Leopold, D. A., O'Toole, A. J., Vetter, T., & Blanz, V. (2001). Prototype-referenced shape encoding revealed by high-level aftereffects. *Nat Neurosci*, *4*.

Tsao, D. Y., Freiwald, W. A., Tootell, R. B., & Livingstone, M. S. (2006). A cortical region consisting entirely of face-selective cells. *Science*, *311*.