Invariant recognition of words and naturalistic sounds in non-human primates

Jonathan Melchor, Isaac Morán, Tonatiuh Figueroa and Luis Lemus*

Department of Cognitive Neuroscience, Institute of Cellular Physiology, National Autonomous University of Mexico (UNAM). 04510. Mexico City. Mexico.

Telephone: (+52) 55 5622 5675

Correspondence:

Dr Luis Lemus, lemus@ifc.unam.mx

Keywords

Perceptual constancy, macaques, lexical representation, saliency, behaviour.

JM, IM, TF & LL performed experiments
JM, IM & LL analysed data
LL designed the paradigm
TF programmed the task
JM & LL prepared figures and wrote the paper

Abstract

Our ability to invariably identify words spoken with different accents and timbres requires our perception to tolerate multiple acoustic modulations. However, the mechanisms by which lexical information is perceived to be invariant is unknown. In this study, we explored the ability of two trained rhesus monkeys to recognise many sounds, including multisyllabic words presented in numerous variations. We found that the monkeys' lexical representations are remarkably tolerant of dynamic acoustic changes. Furthermore, we determined that the macaques invariably recognise sounds with formants at close Euclidean distances from a learned category. Based on our results, we propose the possible existence of neuronal circuits responsible for the invariant recognition of lexical representations in macaques.

Introduction

The ability to perceive acoustic categories from sounds that emulate learned sounds, or from acoustic fragments, exists in humans and other animals[1–3]. Although this ability is vital for communication in primates, the perceptual basis of invariant recognition of sounds has not extensively investigated. One possible reason for this is that non-human primates may only demonstrate limited acoustic learning[4–6] , and that therefore their recognition capability may depend on genetically-programmed circuits[7–9]. However, it is known that macaques are capable of learning repertoires of visual categories[10] and even declare the existence of objects with ambiguous or incomplete information[11]. Studies of the inferotemporal and prefrontal cortices of monkeys showed neurons whose categorical responses achieved the grouping of wide variations of images[12–14], which is consistent with perceptual reports[15,16]. Similarly, experiments in the prefrontal cortex and secondary

acoustic areas suggest invariant coding of acoustic categories[17–25]. In this paper, we sought to determine what acoustic parameters stand for the invariant recognition of sounds (IRS) in trained non-human primates[26]. We hypothesised that monkeys invariably recognise sounds of frequency-saliences with dynamics similar to sounds the monkeys had learned[27,28]. To test this, we designed a novel paradigm in which the macaques efficiently reported target (T) sounds presented in sequences that included nontarget (N) sounds. We found that the monkeys recognised sound fragments of frequency modulations in the range of the prominences of the learned sounds as invariants. Our results allowed us to elucidate, for the first time, the acoustic parameters[29–34] that lead to monkeys' IRS. We propose that the ability to group complex stimuli into particular categories is served by chunks of prominent features of the stimuli.

Results

We trained two rhesus monkeys in an acoustic recognition task in order to study the invariant recognition of sounds. During the task, the monkeys obtained a reward for releasing a lever after identifying a T presented after zero, one or two Ns (Fig. 1a-c; see Methods). After two years of training, monkey V recognised seven Ts and twenty-one Ns, and monkey X recognised eleven Ts and ten Ns. Monkeys demonstrated excellent performance with an overall hit rate of 96.8 ± 0.11 (mean ± SEM, one-sample sign test, $p < 0.01$, Supplementary Table 1), and exhibit longer reaction times during false alarms (395.2 ± 128.4 ms), as compared to the reaction times during the hits (281.8 ± 63.8 ms, Kruskal-Wallis test, $p < 0.001$) (Supplementary Fig. 1a). Figs. 1d and 1e present examples of five Ts and five Ns. Monkey V performed better for Ts in the first position, while monkey X

identified Ts faster in the first and third positions (Kruskal-Wallis test, $p < 0.01$) (Supplementary Fig. 1b-c).

The mean frequency of sounds predicted acoustic recognition in monkeys.

To study the monkeys' ability to differentiate Ts from Ns, we inquired the monkeys with sets of morphed sounds created from mixtures of a T and an N in different proportions (see Methods). Fig. 2a illustrates a morphing set in which the N /si/ (i.e. the Spanish word for 'yes') gradually morphed into a T coo monkey call. Fig. 2b shows psychometric functions (PFs) of the probability of recognising a morph as a T. Here, the differential limen (DL) indicates the minimum proportion of T required for recognition of a morph. There were no differences between monkey V's and monkey X's DLs: 11.3 ± 1.2 and 10.93 ± 1.4 (mean ± SEM), respectively (Kruskal-Wallis test, $p = 0.93$; Supplementary Table 2). In order to elucidate the acoustic variables responsible for recognitions, we calculated acoustic functions of morph parameters (e.g. AM, periodicity, entropy and pitch; see methods) to contrast to the PFs. Thus, we derived Pearson acoustic functions (PAFs) from Pearson correlations of each morph and 100% T (Fig. 2c). Therefore, the PAFs express the similarities between the morphs' acoustic modulations and the modulation of T. Nevertheless, as an alternative, we computed acoustic functions of the Euclidean distances (FEDs) between parameters in the morphs and in T (Fig. 2d). Finally, to determine whether recognition of morphs as T depended on Pearson or on proximities to acoustic parameters, we performed Spearman correlations of PAFs and FEDs with the PFs (Figs. 2e and 2f, respectively). The results indicated that FEDs of mean frequencies were strongly correlated with performance. Here, the average rho-values were 0.97 and 0.96 for monkeys V and X,

respectively ($p < 0.05$, Supplementary Table 3), meaning that acoustic saliencies such as the formants drove the monkeys' abilities to recognise sounds.

Invariant recognition arises from variants at Euclidean proximities to learned sounds. To test for the invariant recognition of sounds in macaques, we presented the monkeys with several versions of the learned sounds, e.g. one word uttered by different individuals. We experimented with sets of five versions of each T and N. Fig. 3a presents the T ['pwɛɾ.ta] spectrogram, i.e. the Spanish disyllabic word for door, and five variants (v1-v5). The boxplots in Fig. 3b correspond to the probabilities of recognising the versions as a T. The monkeys recognised 78.0% of the fifty versions above chance (one-sample sign test, $p \leq 0.05$), with no performance differences between the two monkeys: 84.4 and 84.3% hit rate (Mann-Whitney test, $p = 0.148$). To determine whether the recognition of a version was due to the Euclidean proximity between any of its acoustic parameters to a learned sound, we calculated various FEDs from various acoustic parameters. Fig. 3c shows that, using the parameter 'Mean Frequency', the Euclidean distances of ['pwɛɾ.ta] to four of its versions were smaller than the distances of those versions to other learned sounds. The only exception was a version closer to the coo sound. However, the normalised distances showed that the version of ['pwɛɾ.ta] closer to the coo produced the lowest performance (Fig. 3d). Similarly, Fig. 3e shows that the mean frequency of variants of other learned sounds were also closer to the expected category (Spearman correlation, R = 0.92, $p < 0.01$). Moreover, the FEDs of the sounds' mean frequencies explained performance better than PAFs and other acoustic parameters (Fig. 3f).

Sound formants contribute to IRS.

Since the mean frequency is derived from the mean power of the frequencies in a sound, we explored the contribution to IRS of the frequencies with highest power modulations, e.g. the acoustic formants. To do this, we presented the monkeys with sounds of some formants of the learned sounds and their versions. Fig. 4a shows spectrograms of the T [ko.'mi.ða], i.e. the Spanish trisyllabic word for food, and its F1, F2, and F1&F2 formants. Similarly, Fig. 4b shows spectrograms of a version of [ko.'mi.ða], and its formants. The hypothesis was that formants of the learned sounds would suffice to drive the monkeys' recognitions. Moreover, that formants of the versions modulated in the range of the learned sounds would also work for acoustic recognition (Fig. 4c). The monkeys performed for no more than forty presentations of each sound in order to prevent the learning of formants as T or N. Fig. 4d presents the mean performance of the monkeys during the recognition of sounds in Fig. 4a-b. The monkeys significantly identified [ko.'mi.ða], its formants, the versions, and the versions' F1&F2 formants (one-sample sign test, $p < 0.01$). However, the versions' F1 or F2 alone were not sufficient for recognition. Fig. 4e is the same as Fig. 4c but for the category ['xaw.la]. Fig. 4f shows false alarms of ['xaw.la], the versions and formants. Here, F2 of the learned and version sounds, and F1&F2 of the learned sound did not produce a significant number of false alarms. Finally, Figs. 4g-h present the results for other Ts and Ns, and their versions. The monkeys recognised the learned T with a probability of 0.93 ± 0.03, and versions of Ts with a P of 0.73 ± 0.09. Meanwhile, the false alarms of learned N had a P of 0.14 ± 0.06, and for Ns versions P was 0.24 ± 0.16. Overall, 94% of F1&F2 of learned and version sounds were recognised significantly (one-sample sign test, $p < 0.01$) (Supplementary Table 4). These results suggest that the invariant recognition of sounds in

macaques is created from acoustic saliencies modulated in the range of saliencies of learned sounds.

Discussion

We have presented evidence of the invariant recognition of sounds in monkeys. This evidence is mainly supported by the ability of the monkeys to recognise variants to which they had no previous exposure as learned sounds. The learned sounds included words and naturalistic sounds in a broad range of frequencies and temporal modulations. Remarkably, the recognition of the variants was based on saliencies at Euclidean proximities to saliencies of learned sounds. To our knowledge, this is the first demonstration of the ability of monkeys to learn and recognise multiple words and naturalistic sounds stored in long-term memories. Given these findings, we propose that recurrent prominent frequencies create neuronal templates that serve to identify unexperienced sounds of similar prominences.

Macaques learn numerous naturalistic sounds.

A weakness of our study was the lack of semantic relationships to each of the sounds. Perhaps only the conspecific vocalisations have meaning for the monkeys. If this is the case, then the monkeys also had to assign an additional meaning to the vocalisations, i.e. target and nontarget. Moreover, all targets were synonymous with reward, and all Ns with holding the lever. Nevertheless, in our study, the repertoire of frequencies within the Ts and Ns were likely to form diverse lexical representations throughout the superior temporal gyrus. Representations recognised as either releasing, or as holding the lever. Similar associations to behaviour may occur in other communicating animals[1–3]. Further

behavioural studies[26] may address how lexicon serves semantics. Our results differ from previous reports that claim limited acoustic learning in macaques[4–6]. As a consequence, adopting the macaques as a model for acoustic recognition has been limited. In our experience, the training of monkeys was indeed more tenuous and prolonged than training in visual or tactile paradigms, but it was achievable. For instance, the monkeys recognised sounds that included multisyllabic words above a hit rate of 90%. This single result suggested that acoustic circuits cannot be entirely based on genetic programmes[7–9]. Moreover, we verified that the learned sounds remained in long-term memories because the monkeys were capable of solving the task after periods of up to five weeks of rest. However, a realistic possibility was that the monkeys only learned the first or the last chunks of the sounds. Nevertheless, given that the monkeys waited for 0.5 s after the sounds to respond, they likely accumulated all of the available evidence, which is related to previous reports showing that the monkeys needed all available information for discriminating acoustic flutter-frequencies[33]. Nonetheless, our results indicate that the monkeys had longer reaction times during false alarms. Therefore, they probably attended only to a portion of the sounds. However, we have no conclusive evidence of such a possibility. Future experiments using the same syllables at the beginning and the end of the sounds could resolve the matter. For now, we can argue that the observed performance required, at least, fragments posed at correct positions[22,34,35].

Acoustic recognition arises from a rule of proximity.

In order to understand IRS, it is fundamental to discern the range of acoustic variability where a perceptual category remains. Our first hypothesis was that IRS emerged from the similarity in acoustic modulations of a learned sound and its versions. Thus, we first

searched for Pearson correlations between continuous functions of the learned sounds and mixtures of T and N categories. The relationships would suggest the existence of spectrotemporal fingerprints emulated by the morphs. However, we found that subtle differences ruled out the hypothesis. Alternatively, we found that the perceptual constancy of acoustic categories occurred for versions with mean frequencies at Euclidean proximities to the learned categories. This finding coincides with recent reports on vowel identification[17], and is consistent with the notion of formants being crucial for carrying acoustic identities[27–30,36]. One explanation is that the salient formants emerge with semantic information from the persistent fine structure of sounds, such as timbre, which may be responsible for streaming —as in a cocktail party paradigm. In such a scenario, perhaps neuronal responses that adapt to timbre code only for the formants. One possible consequence would be that speakers learn to modulate formants in order to communicate, and not the pitch, nor the timbre, which are more useful for sound localisation, or recognition of conspecifics[37].

Hierarchical processing of sounds

Recordings of neurons in passive untrained macaque have demonstrated that belt area neurons around the core of the auditory cortex (A1) are responsive to band-passed noise[38], FM-sweeps[39], and conspecific vocalisations[40–42]. The belt receives the information contained in vocalisations from simultaneous projections of the neurons which demonstrate sharp frequency tuning in A1. However, belt neurons also respond to reversed monkey calls[43]. This means that these cells do not code for specific sequences of frequencies that provide identity to the acoustic categories. Therefore, if the primary sensory circuits are genetically programmed, they are not responsible for perception[44]. This would suggest and

support the finding of PFC neurons encoding for vocalisations organised in specific frequency sequences[18–20]. Nevertheless, those neurons were observed in non-behaving macaques, so their contribution to acoustic perception remains unclear. Moreover, in order to understand what parameters correlate with acoustic perception[27–30,32,36], experiments using monkeys trained to discriminate the syllables /Bad/ and /Dad/ found categorical responses to linear mixtures of the syllables at the belt[21]. This finding means that belt neurons responded to perceptual categories and not to particular spectrotemporal modulations. Recent fMRI studies in humans and macaques have shown that anterior areas of the superior temporal gyrus respond to monkey vocalisations and rarely to other sounds[22–25,32], suggesting a distributed representation of sounds relevant to behaviour. An important question is whether those representations serve as templates for the recognition of similar sounds. Experiments conducted with behaving ferrets showed that A1 neurons can respond to variations of vowels[17]. However, the neurons were sensitive to input timing, suggesting that the recognition of longer and more complex sounds requires further cortical integration.

Based on our results, we propose that recognition circuits hierarchically integrate patterns of acoustic prominences, including combinations, as in words. Furthermore, recurrent sounds create neuronal templates, sometimes evoked by saliencies of variants. Further experiments may explore semantics using our paradigm. For example, the coding of the meaning of conspecific vocalizations in different brain areas[23,41,43,45]. In conclusion, the behavioural paradigm we present could advance the study of acoustic recognition at the neuronal level, because, in contrast to humans[46], trained monkeys present only a few dozen

acoustic representations, meaning fewer lexical overlaps, which benefits the study of discrete acoustic percepts.

Methods

Ethics statement

All procedures were performed in compliance with the Mexican Official Standard for the Care and Use of Laboratory Animals (NOM-062-ZOO-1999) and approved by the Internal Committee for the Use and Care of Laboratory Animals of the Institute of Cellular Physiology, UNAM (CICUAL; LLS80-16).

Animals and experimental setup

Two adult rhesus macaques (*Macaca mulatta*; one male, 13 kg, ten yrs. old, and one female, 6 kg, ten yrs. old) participated in this study. Typically, each monkey performed ~1000 trials during sessions of three hours (one session per day, six sessions per week). The monkeys received a daily minimum water intake of 20 ml/kg, completed in cage as needed. The monkeys' training lasted approximately two years and concluded after each one recognised more than 20 sounds above an ~85% hit rate. Training and experimental sessions took place in a soundproof booth. The macaque was seated in a primate chair, 60 cm away from a 21" LCD colour monitor (1920 x 1080 resolution, 60 Hz refreshing rate). A Yamaha MSP5 speaker (50 Hz - 40 kHz frequency range) was placed fifteen cm above and behind the monitor to deliver acoustic stimuli at ~65 dB SPL (measured at the monkeys' ear level). Additionally, a Logitech® Z120 speaker was situated directly below

the Yamaha speaker in order to render background white noise at ~55 dB SPL. Finally, a metal spring-lever situated at the monkeys' waist level captured the responses.

Behavioural Task

The acoustic recognition task (ART) consisted of identifying T and N sounds. Fig. 1a presents the elements of the paradigm as follows: First, a grey circle with an aperture of 3° appeared at the centre of the screen, and the monkey pressed and held down the lever. Immediately thereafter, a playback of from 1 to 3 sounds began, and a T was always the last sound (Fig. 1b). After each sound, the monkey kept the lever down for another 0.5 s until the visual cue turned green (G). If the audio was a T, the monkey had 0.8 s to release the lever and receive a drop of liquid. However, releases at other periods constituted a false alarm (FA) that led to the abortion of the trial (Fig. 1c). The task's programming was in LabVIEW 2014 (SP1 64-bits, National Instruments®).

Stimuli

The sounds were recordings from our laboratory or downloads from free internet libraries. They consisted of natural and artificial environmental sounds, e.g. monkey calls, other animal vocalisations and words. All sounds were sampled at 44.1 kHz (cutoff frequencies: 100 Hz to 20 kHz), amplitudes were normalised at -10 dB SPL (RMS), and compressed or elongated to 0.5 s. Fig. 1d presents examples of five T and five N used frequently during the experiments. The morphing sets comprised 11 mixtures of T and an N in proportions ranging from 0% T (i.e. 100% N) to 100% T in 10% increments of T[47,48]. Each morphed sound was repeated randomly ten times but always presented first in a trial. Trials of two or three sounds were completed with T and N. To test for IRS, versions of learned sounds

were presented forty times randomly, but only after the monkeys' training concluded. Finally, we examined the recognition of acoustic salience using F1, F2, and F1&F2 formants of learned sounds and versions. All sounds were processed using Adobe Audition® version 6.0. The morphed sounds were created using the signal processing software STRAIGHT[47].

Analysis

PFs were TanH regressions of the probability of recognising a morph as a T[49]. PAFs and FEDs were functions of Pearson correlations between continuous parameters measured at each morph and the same parameter in 100% T, and the Euclidean distances from each M to 100% T, respectively[17]. Spearman correlations between FAP and FED with PF computed the contribution of acoustic parameters to recognition. Differential limen (DL) was half the difference between the abscissa projected to the PF at 75%, and 25% performance. Reaction times were times of lever releases after the start of G. Logarithmic ratio = log(performance) - log(distance). Behavioural analyses were performed using SigmaPlot® version 12.0 software for Windows (Systat Software, Inc., San Jose, CA, USA), and customised algorithms in MATLAB® 8.5.0.1, R2015a (The Mathworks, Inc). Acoustic metrics were computed using Pratt[50], VoiceSauce[32] (version 1.36), and Sound Analysis Pro[31] .

Acknowledgements

We thank Professors Chandramouli Chandrasekaran, and Victor de Lafuente for valuable comments. We are also grateful for the financial support provided by CONACYT **CB-181301**, **CB-256767**, and *Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica* (*PAPIIT*) **IN207919**. Special thanks to Gerardo Coello and Ana Escalante of

the computing department of the IFC and to Mr Patrick Weill for proofreading the manuscript. Jonathan Melchor Hernández is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM) and was supported by CONACYT 229866. Data in this work are part of his doctoral dissertation.

Disclosure

The authors declare no conflicts of interest

Bibliography

1. Elie, J. E. & Theunissen, F. E. Meaning in the avian auditory cortex: Neural representation of communication calls. *Eur. J. Neurosci.* **41**, 546–567 (2015).
2. Saunders, J. L. & Wehr, M. Mice can learn phonetic categories. *J. Acoust. Soc. Am.* **145**, 1168–1177 (2019).
3. Seyfarth, R., Cheney, D. & Marler, P. Monkey responses to three different alarm calls: evidence of predator classification and semantic communication. *Science (80-.).* **210**, 801–803 (1980).
4. Scott, B. H., Mishkin, M. & Yin, P. Monkeys have a limited form of short-term memory in audition. *Proc. Natl. Acad. Sci.* **109**, 12237–12241 (2012).
5. Wright, A. A. Auditory list memory and interference processes in monkeys. *J. Exp. Psychol. Anim. Behav. Process.* **25**, 284–296 (1999).
6. Fritz, J., Elhilali, M. & Shamma, S. Active listening: Task-dependent plasticity of spectrotemporal receptive fields in primary auditory cortex. *Hear. Res.* **206**, 159–176 (2005).

7. Brockelman, W. Y. & Schilling, D. Inheritance of stereotyped gibbon calls. *Nature* **312**, 634–636 (1984).

8. Owren, M. J., Dieter, J. A., Seyfarth, R. M. & Cheney, D. L. 'Food' Calls Produced by Adult Female Rhesus (Macaca Mulatta) and Japanese (M. Fuscata) Macaques, their Normally-Raised Offspring, and Offspring Cross-Fostered Between Species. *Behaviour* **120**, 218–231 (1992).

9. Zador, A. M. A critique of pure learning and what artificial neural networks can learn from animal brains. *Nat. Commun.* **10**, 3770 (2019).

10. Rajalingham, R., Schmidt, K. & DiCarlo, J. J. Comparison of object recognition behavior in human and monkey. *J. Neurosci.* **35**, 12127–12136 (2015).

11. Roy, J. E., Buschman, T. J. & Miller, E. K. PFC neurons reflect categorical decisions about ambiguous stimuli. *J. Cogn. Neurosci.* **26**, 1283–1291 (2014).

12. Seger, C. A. & Miller, E. K. Category Learning in the Brain. *Annu. Rev. Neurosci.* **33**, 203–219 (2010).

13. DiCarlo, J. J., Zoccolan, D. & Rust, N. C. How does the brain solve visual object recognition? *Neuron* **73**, 415–434 (2012).

14. Bao, P. & Tsao, D. Y. Representation of multiple objects in macaque category-selective areas. *Nat. Commun.* **9**, 1–16 (2018).

15. Wutz, A., Loonis, R., Roy, J. E., Donoghue, J. A. & Miller, E. K. Different Levels of Category Abstraction by Different Dynamics in Different Prefrontal Areas. *Neuron* **97**, 716-726.e8 (2018).

16. Cromer, J. A., Roy, J. E. & Miller, E. K. Representation of Multiple, Independent Categories in the Primate Prefrontal Cortex. *Neuron* **66**, 796–807 (2010).

17. Town, S. M., Wood, K. C. & Bizley, J. K. Sound identity is represented robustly in

auditory cortex during perceptual constancy. *Nat. Commun.* (2018). doi:10.1038/s41467-018-07237-3

18. Romanski, L. M., Averbeck, B. B. & Diltz, M. Neural representation of vocalizations in the primate ventrolateral prefrontal cortex. *J. Neurophysiol.* **93**, 734–747 (2005).

19. Cohen, Y. E., Hauser, M. D. & Russ, B. E. Spontaneous processing of abstract categorical information in the ventrolateral prefrontal cortex. *Biol. Lett.* **2**, 261–265 (2006).

20. Russ, B. E., Ackelson, A. L., Baker, A. E. & Cohen, Y. E. Coding of auditory-stimulus identity in the auditory non-spatial processing stream. *J. Neurophysiol.* **99**, 87–95 (2008).

21. Tsunada, J., Lee, J. H. & Cohen, Y. E. Representation of speech categories in the primate auditory cortex. *J. Neurophysiol.* **105**, 2634–2646 (2011).

22. Leaver, A. M. & Rauschecker, J. P. Cortical representation of natural complex sounds: Effects of acoustic features and auditory object category. *J. Neurosci.* **30**, 7604–7612 (2010).

23. Petkov, C. I. *et al.* A voice region in the monkey brain. *Nat. Neurosci.* **11**, 367–374 (2008).

24. Perrodin, C., Kayser, C., Logothetis, N. K. & Petkov, C. I. Voice cells in the primate temporal lobe. *Curr. Biol.* **21**, 1408–1415 (2011).

25. Perrodin, C., Kayser, C., Abel, T. J., Logothetis, N. K. & Petkov, C. I. Who is That? Brain Networks and Mechanisms for Identifying Individuals. *Trends Cogn. Sci.* **19**, 783–796 (2015).

26. Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A. & Poeppel, D.

Neuroscience Needs Behavior: Correcting a Reductionist Bias. *Neuron* **93**, (2017).

27. Remez, R. E., Rubin, P. E., Pisoni, D. B. & Carrell, T. D. Speech perception without traditional speech cues. *Science* **212**, 947–9 (1981).

28. Furuyama, T., Kobayasi, K. I. & Riquimaroux, H. Acoustic characteristics used by Japanese macaques for individual discrimination. *J. Exp. Biol.* **220**, 3571–3578 (2017).

29. Ghazanfar, A. A. *et al.* Vocal-Tract Resonances as Indexical Cues in Rhesus Monkeys. *Curr. Biol.* **17**, 425–430 (2007).

30. Furuyama, T., Kobayasi, K. I. & Riquimaroux, H. Role of vocal tract characteristics in individual discrimination by Japanese macaques (Macaca fuscata). *Sci. Rep.* **6**, 1–8 (2016).

31. Tchernichovski, O., Nottebohm, F., Ho, C. E., Pesaran, B. & Mitra, P. P. A procedure for an automated measurement of song similarity. *Anim. Behav.* **59**, 1167–1176 (2000).

32. Shue, Y.-L., Keating, P. & Vicenik, C. VOICESAUCE: A program for voice analysis. *J. Acoust. Soc. Am.* **126**, 2221 (2009).

33. Lemus, L., Hernández, A. & Romo, R. Neural codes for perceptual discrimination of acoustic flutter in the primate auditory cortex. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 9471–9476 (2009).

34. Averbeck, B. B. & Romanski, L. M. Probabilistic encoding of vocalizations in macaque ventral lateral prefrontal cortex. *J. Neurosci.* **26**, 11023–11033 (2006).

35. Sainburg, T., Theilman, B., Thielk, M. & Gentner, T. Q. Parallels in the sequential organization of birdsong and human speech. *Nat. Commun.* **10**, 3636 (2019).

36. Fitch, W. T. & Fritz, J. B. Rhesus macaques spontaneously perceive formants in

conspecific vocalizations. *J. Acoust. Soc. Am.* **120**, 2132–2141 (2006).

37. Takahashi, D. Y., Fenley, A. R. & Ghazanfar, A. A. Early development of turn-taking with parents shapes vocal acoustics in infant marmoset monkeys. *Philos. Trans. R. Soc. B Biol. Sci.* **371**, (2016).

38. Rauschecker, J. P. & Tian, B. Processing of band-passed noise in the lateral auditory belt cortex of the rhesus monkey. *J. Neurophysiol.* **91**, 2578–2589 (2004).

39. Tian, B. & Rauschecker, J. P. Processing of Frequency-Modulated Sounds in the Lateral Auditory Belt Cortex of the Rhesus Monkey. *J. Neurophysiol.* **92**, 2993–3013 (2004).

40. Ortiz-Rios, M. *et al.* Widespread and Opponent fMRI Signals Represent Sound Location in Macaque Auditory Cortex. *Neuron* **93**, (2017).

41. Tian, B., Reser, D., Durham, A., Kustov, A. & Rauschecker, J. P. Functional specialization in rhesus monkey auditory cortex. *Science (80-.).* **292**, 290–293 (2001).

42. Rauschecker, J. P. & Tian, B. Mechanisms and streams for processing of 'what' and 'where' in auditory cortex. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 11800–11806 (2000).

43. Recanzone, G. H. Representation of con-specific vocalizations in the core and belt areas of the auditory cortex in the alert macaque monkey. *J. Neurosci.* **28**, 13184–13193 (2008).

44. Lemus, L., Hernández, A., Luna, R., Zainos, A. & Romo, R. Do sensory cortices process more than one sensory modality during perceptual judgments? *Neuron* **67**, 335–348 (2010).

45. Ortiz-Rios, M. *et al.* Functional MRI of the vocalization-processing network in the

macaque brain. *Front. Neurosci.* **9**, (2015).

46. Coupé, C., Oh, Y., Dediu, D. & Pellegrino, F. Different languages , similar encoding efficiency : Comparable information rates across the human communicative niche. (2019).

47. Kawahara, H., Masuda-Katsuse, I. & De Cheveigné, A. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Commun.* **27**, 187–207 (1999).

48. Chakladar, S., Logothetis, N. K. & Petkov, C. I. Morphing rhesus monkey vocalizations. *J. Neurosci. Methods* **170**, 45–55 (2008).

49. Duarte, F. & Lemus, L. The Time Is Up: Compression of Visual Time Interval Estimations of Bimodal Aperiodic Patterns. *Front. Integr. Neurosci.* **11**, (2017).

50. Boersma, P., & Van Heuven, V. Speak and unSpeak with PRAAT. *Glot Int.* **5**, 341–347 (2001).

Figure legends

Fig. 1 Auditory recognition task. a An example of the sequence of events of a trial. First, a visual cue appeared at the center of the screen to indicate that the monkey should press and hold the lever down. After a variable period of 0.5 to 1 s, a playback of 1 to 3 sounds commenced, each followed by a 0.5 s delay and a 0.5 s green cue (G). The monkey obtained a drop of liquid for releasing within 0.7 s of the beginning of the G that followed the T. Releases at other periods aborted the trial. Colour code: orange=T, grey=N, green=release cue. **b** Depictions of sequences of one, two or three sounds. Note that T

always appeared last. **c** The behavioural outcomes after presentations of Ts and Ns. FA, false alarm, CR, correct rejection. **d** Sonograms and spectrograms of five Ts. IPA nomenclature describes Spanish words used in the experiments. **e** same as in (**d**) but for nontarget sounds.

Fig. 2 The mean frequency correlates to target recognitions. a An example of a morphing set in which a N [si] morphed to an T coo, from 0% T to 100% T in increments of 10%. Every morphing set comprised eleven morphs. **b** Monkey V's and monkey X's probabilities of recognising a morph as a T during the morphing set shown in (**a**). Continuous lines correspond to the sigmoidal fit to the average performance during the different morphing sets. **c** Subpanels present Pearson's acoustic functions (PAFs) of various acoustic metrics (see Methods). Same colours as in (**b**). **d** Same as in c but for acoustic functions of Euclidean distances (FEDs). **e** Each dot is a Spearman correlation coefficient (rho) between the psychometric functions and PAFs, for different acoustic metrics. Same colours as in previous panels. Solid bars, monkey V. Unfilled bars, monkey X. **f** Same as in (**e**) but for FEDs.

Fig. 3 Mean frequency proximities between learned sounds and their variants produce perceptual invariance. a Spectrograms of T ['pwer.ta], i.e. the Spanish word for door and five variants. Each variant corresponds to a different speaker (v1-v5). **b** Boxplots of the probability of recognising a variant as a T. Colours at ['pwer.ta] categories correspond to variants at (**a**). **c** Normalised Euclidean distances of variants of ['pwer.ta] to four Ts and five Ns. Colours are the same as in (**a**) and (**b**). Symbols are labelled at the abscissas. **d** Mean monkey performance as a function of Euclidean distances of variants of [‘pwer.ta] to

all Ts and Ns. Same colour code as in (**a-c**). **e** Logarithmic ratio of the recognition of variants at (**b**) and the mean-frequency distance to each T and N, plotted as a function of the probability of recognising a T. Symbols as in (**d**). Upper left, the Pearson correlation coefficient (r) between the logarithmic ratio and behaviour. **f** Similar to (**e**) but for Pearson's r, and for all of the tested acoustic metrics.

Fig. 4. The first and second formants are key for perceptual invariance. a Spectrograms of T [ko.'mi.da], and first, second, and first & second formants. **b** One version of [ko.'mi.da], and its corresponding formants. **c** Comparison of F1 and F2 bandwidth formants of [ko.'mi.da] and the mean of the version's F1 and F2 formants. **d** Monkeys' mean probability of recognising sounds in (a) as T. **e-f** Same as in (**c**) and (**d**) but for N ['xaw.la]. **g-h**, Probability of recognition for F1, F2, F1&F2, learned and variants of Ts and Ns, respectively.

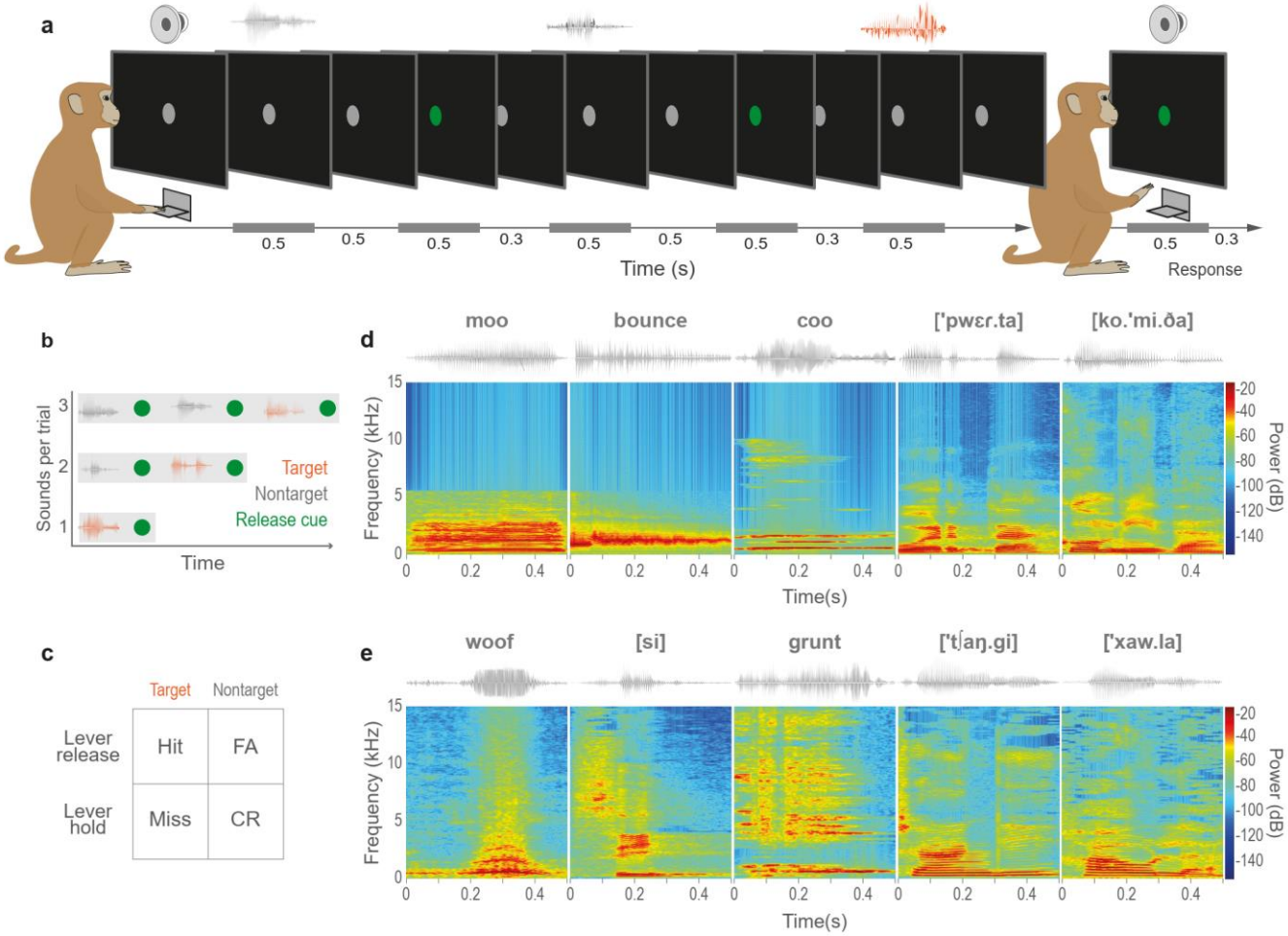

a
0.5
0.5
0.5
0.3
0.5
0.5
0.5
0.3
0.5
0.5
0.3
Time (s)
Response
b
Sounds per trial
3
2
1
Target
Nontarget
Release cue
Time
c
Target
Nontarget
Lever release
Lever hold
Hit
FA
Miss
CR
d
moo
bounce
coo
['pwɛr.ta]
[ko.'mi.ða]
Frequency (kHz)
Power (dB)
Time(s)
e
woof
[si]
grunt
['tʃaŋ.gi]
['xaw.la]
Frequency (kHz)
Power (dB)
Time(s)

Figure 1

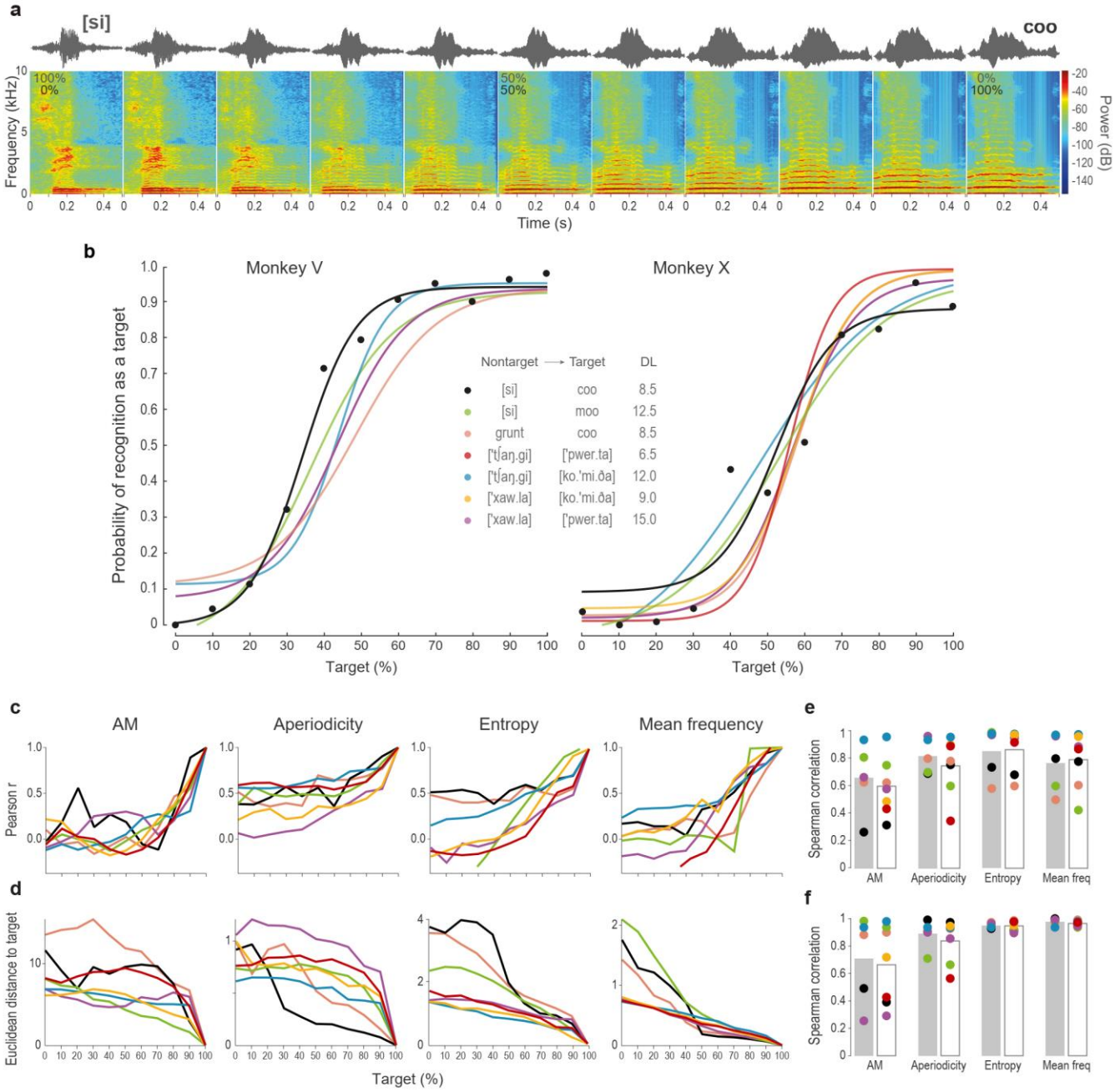

a
[si]
coo
Frequency (kHz)
100%
0%
50%
50%
0%
100%
Power (dB)
Time (s)
b
Monkey V
Monkey X
Probability of recognition as a target
Nontarget → Target DL
[si] coo 8.5
[si] moo 12.5
grunt coo 8.5
['tʃaŋ.gi] ['pwer.ta] 6.5
['tʃaŋ.gi] [ko.'mi.ða] 12.0
['xaw.la] [ko.'mi.ða] 9.0
['xaw.la] ['pwer.ta] 15.0
Target (%)
c
AM
Aperiodicity
Entropy
Mean frequency
Pearson r
d
Euclidean distance to target
Target (%)
e
Spearman correlation
AM
Aperiodicity
Entropy
Mean freq
f
Spearman correlation
AM
Aperiodicity
Entropy
Mean freq

Figure 2

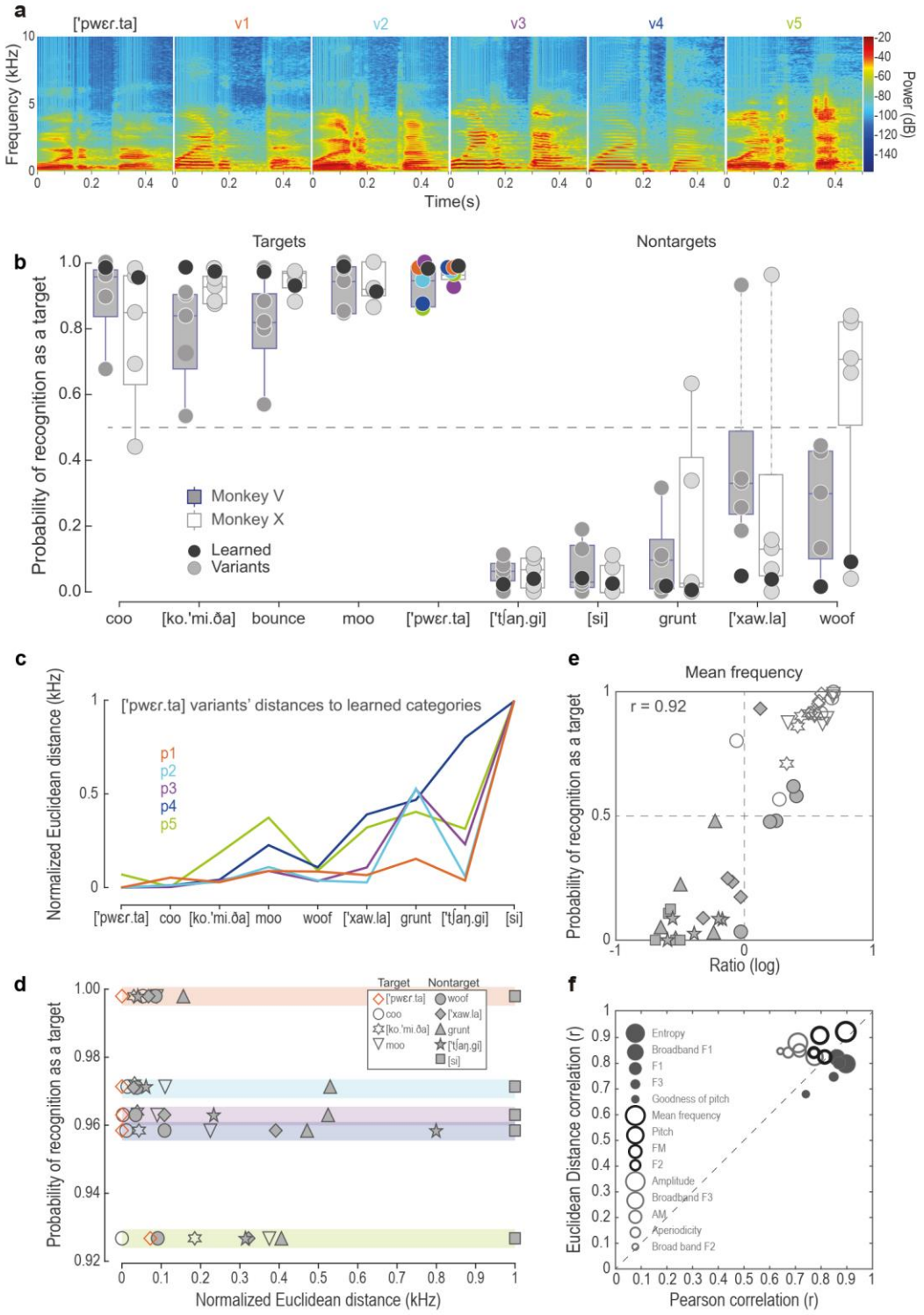

a
['pwɛr.ta]
v1
v2
v3
v4
v5
Frequency (kHz)
Time(s)
Power (dB)
b
Targets
Nontargets
Probability of recognition as a target
Monkey V
Monkey X
Learned
Variants
coo
[ko.'mi.ða]
bounce
moo
['pwɛr.ta]
['tʃaŋ.gi]
[si]
grunt
['xaw.la]
woof
c
['pwɛr.ta] variants' distances to learned categories
Normalized Euclidean distance (kHz)
p1
p2
p3
p4
p5
['pwɛr.ta]
coo
[ko.'mi.ða]
moo
woof
['xaw.la]
grunt
['tʃaŋ.gi]
[si]
d
Target
Nontarget
['pwɛr.ta]
coo
[ko.'mi.ða]
moo
woof
['xaw.la]
grunt
['tʃaŋ.gi]
[si]
Probability of recognition as a target
Normalized Euclidean distance (kHz)
e
Mean frequency
r = 0.92
Probability of recognition as a target
Ratio (log)
f
Entropy
Broadband F1
F1
F3
Goodness of pitch
Mean frequency
Pitch
FM
F2
Amplitude
Broadband F3
AM
Aperiodicity
Broad band F2
Euclidean Distance correlation (r)
Pearson correlation (r)

Figure 3

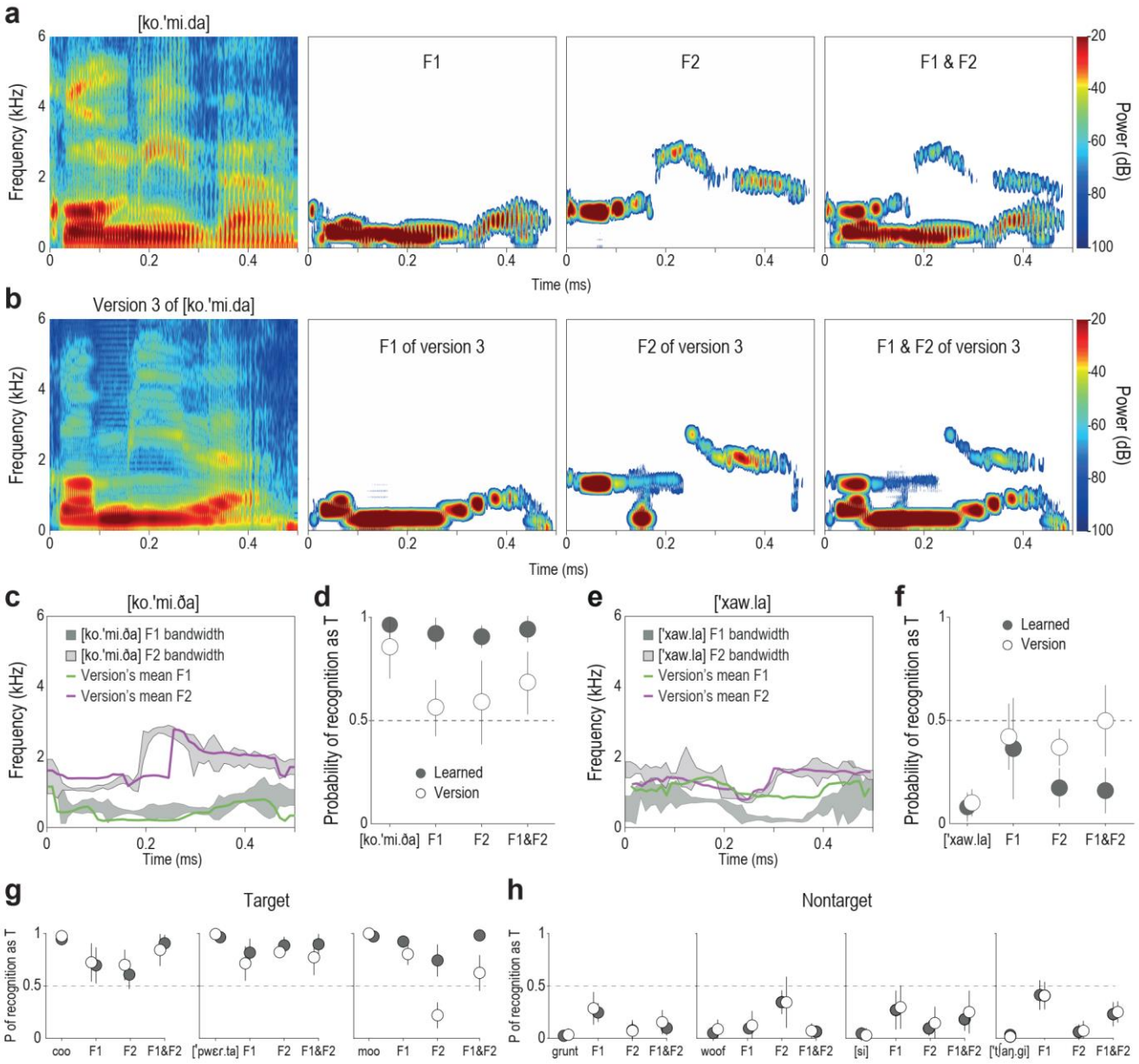
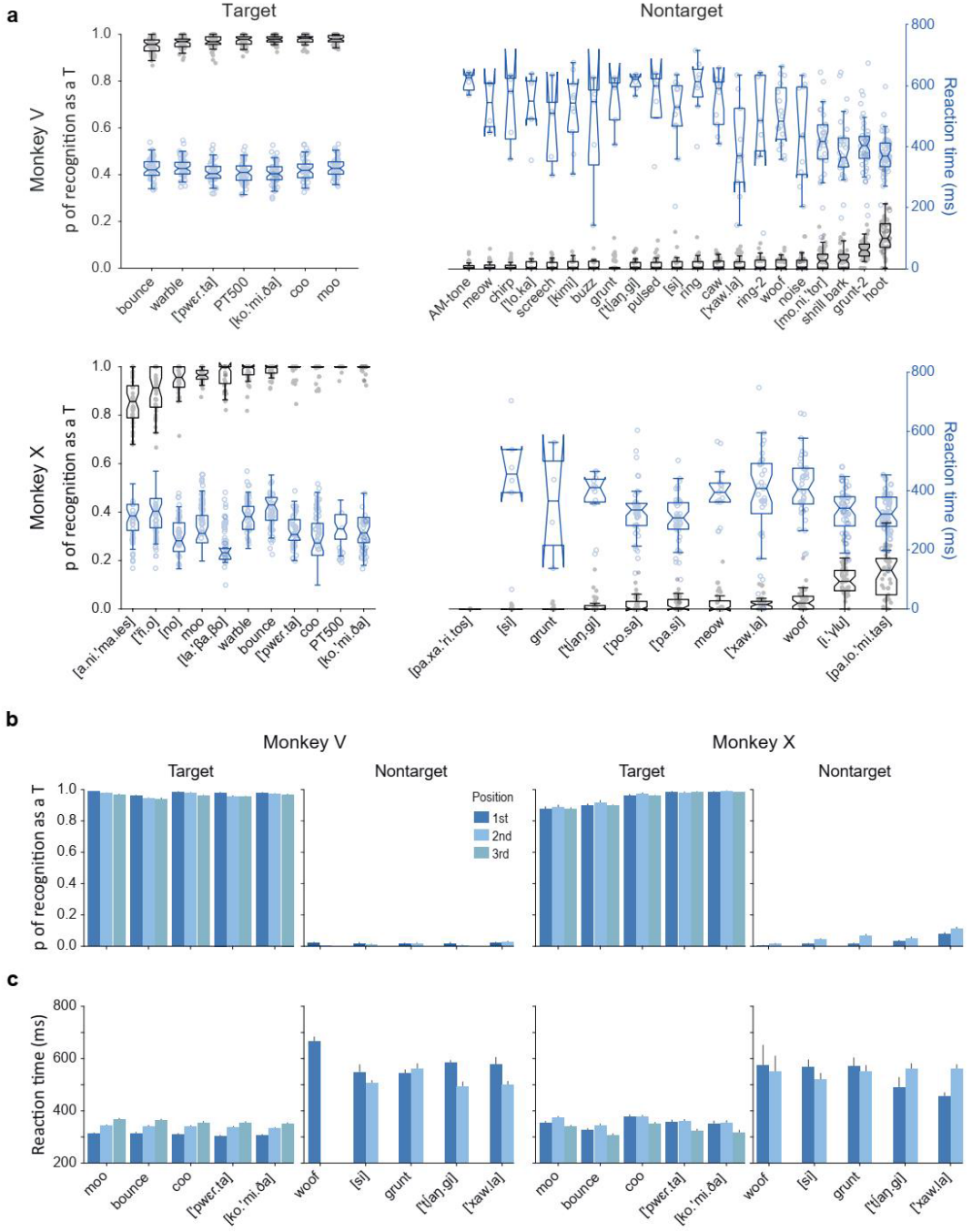

a
[ko.'mi.da]
F1
F2
F1 & F2
Frequency (kHz)
Power (dB)
Time (ms)
b
Version 3 of [ko.'mi.da]
F1 of version 3
F2 of version 3
F1 & F2 of version 3
c
[ko.'mi.ða]
[ko.'mi.ða] F1 bandwidth
[ko.'mi.ða] F2 bandwidth
Version's mean F1
Version's mean F2
d
Probability of recognition as T
Learned
Version
[ko.'mi.ða] F1 F2 F1&F2
e
['xaw.la]
['xaw.la] F1 bandwidth
['xaw.la] F2 bandwidth
f
['xaw.la] F1 F2 F1&F2
g
Target
P of recognition as T
coo F1 F2 F1&F2
['pwɛr.ta] F1 F2 F1&F2
moo F1 F2 F1&F2
h
Nontarget
grunt F1 F2 F1&F2
woof F1 F2 F1&F2
[si] F1 F2 F1&F2
['tʃaŋ.gi] F1 F2 F1&F2

Figure 4



Supplementary Figure 1. Behavioural performance and reaction times (RT). **a** Probability of recognition (left axis) and RT (right axis) for all learned target (T) and nontarget (N) sounds by the monkeys V (top) and X (bottom). All the sounds were 0.5 s in duration. **b** Mean accuracy was high for the three types of trials in both monkeys and sound categories. **c** RT for the five Ts and Ns of monkeys V (left) and X (right).

Supplementary Table 1. Probability of recognition and description of the sound set.

	Name	Description	P of recognition as T	
			Monkey V	Monkey X
Target	coo	Rhesus monkey vocalization	97.76 ± 02.04	98.90 ± 02.82
	warble	Rhesus monkey vocalization	96.33 ± 02.55	97.81 ± 03.99
	[ko.'mi.ða]	Spanish word for food	97.59 ± 01.70	99.37 ± 01.76
	['pwɛr.ta]	Spanish word for door	96.58 ± 02.85	98.80 ± 03.06
	[a.ni.'ma.les]	Spanish word for animals	-	84.97 ± 09.31
	['ři.o]	Spanish word for river	-	89.91 ± 08.84
	[no]	Spanish word for not	-	94.86 ± 05.97
	[la.'βa.βo]	Spanish word for sink	-	96.40 ± 05.16
	moo	Vocal sound of a cow	97.93 ± 01.88	96.34 ± 02.75
	bounce	Bouncing tone	95.01 ± 03.23	98.31 ± 02.54
	PT500	Pure tone of 500 Hz	97.09 ± 02.23	99.24 ± 01.85
Nontarget	grunt	Rhesus monkey vocalization	01.26 ± 02.29	00.19 ± 00.67
	grunt-2	Rhesus monkey vocalization	07.78 ± 04.20	-
	shrill bark	Rhesus monkey vocalization	03.95 ± 04.06	-
	pulsed	Rhesus monkey vocalization	01.31 ± 02.55	-
	['lo.ka]	Spanish word for crazy	00.97 ± 01.67	-
	[kimi]	Spanish pseudoword	01.07 ± 01.88	-
	['tʃaŋ.gi]	Spanish pseudoword	01.22 ± 02.63	01.58 ± 03.42
	[si]	Spanish word for yes	01.37 ± 02.81	00.18 ± 00.54
	['xaw.la]	Spanish word for cage	01.42 ± 02.23	02.50 ± 03.13
	[mo.ni.'tor]	Spanish word for computer monitor	03.91 ± 04.31	-
	[pa.xa.'ri.tos]	Spanish word for birdies	-	00.00 ± 00.00
	['po.sa]	Spanish word for pose	-	02.13 ± 03.51
	['pa.si]	Spanish pseudoword	-	02.33 ± 03.00
	[i.'ɣlu]	Spanish word for igloo	-	12.27 ± 06.06
	[pa.lo.'mi.tas]	Spanish word for popcorn	-	15.31 ± 09.66
	meow	Cat meowing	00.90 ± 01.76	02.39 ± 04.48
	chirp	Bird vocalization	00.92 ± 01.69	-
	screech	Parrot screech	01.01 ± 01.79	-
	caw	Crow squawk	01.57 ± 02.60	-
	woof	Dog bark	01.83 ± 02.34	04.51 ± 07.53
	hoot	Owl hooting	14.29 ± 07.95	-
	AM-tone	Amplitude-Modulated tone (1 kHz)	00.85 ± 01.66	-
	buzz	Mosquito whine	01.10 ± 01.82	-
	ring	Cell phone ring tone	01.41 ± 02.66	-
	ring-2	Ring bell	01.68 ± 02.53	-
	noise	Passband noise (1-4 kHz)	01.87 ± 03.17	-

Data are mean ± SD; (n ≥ 41 sessions).
The IPA nomenclature was obtained with the Spanish automatic phonetic transcriptionist, created for Xavier López Morrás (http://www.aucel.com/pln/transbase.html).

Supplementary Table 2. Differential limen (DL) and sigmoidal fit parameters for the morphing sets.

	Parameters of *tanh* function	Target and nontarget sounds for each morphing						
		[si] to coo	[si] to moo	grunt to coo	['tʃaŋ.gi] to ['pwɛr.ta]	['tʃaŋ.gi] to [ko.'mi.ða]	['xaw.la] to [ko.'mi.ða]	['xaw.la] to ['pwɛr.ta]
Monkey V	c	0.47	0.44	0.53	-	0.51	-	0.53
	α	0.47	0.49	0.42	-	0.43	-	0.42
	θ	33.7	35.9	43.5	-	43.1	-	47.7
	β	0.07	0.05	0.08	-	0.05	-	0.05
	Q	0.65	0.65	0.46	-	0.62	-	0.46
	DL	8.50	12.5	8.50	-	12.0	-	15.0
Monkey X	c	0.47	0.49	0.45	0.50	0.49	0.52	0.51
	α	0.50	0.40	0.55	0.49	0.47	0.47	0.48
	θ	52.6	51.7	47.0	55.3	56.2	57.3	57.3
	β	0.04	0.07	0.03	0.09	0.06	0.07	0.07
	Q	0.49	0.44	0.61	0.80	0.75	0.65	0.70
	DL	15.5	11.0	16.5	6.50	9.50	9.00	8.50

Supplementary Table 3. Spearmans' coefficients (Rho) between psychometric functions (PF) and Pearson Acoustic Function (PAF), and PF and Function of Euclidean distance (FED) for metrics.

Acoustic metric	Rho from PF and PAF			Rho from PF and FED		
	Monkey V (n = 5)	**Monkey X** (n = 7)	**Proportion of Rho with $p < 0.05$**	**Monkey V** (n = 5)	**Monkey X** (n = 7)	**Proportion of Rho with $p < 0.05$**
AM	0.66 ± 0.25	0.58 ± 0.21	58.30	0.70 ± 0.32	0.66 ± 0.28	58.30
Aperiodicity	0.81 ± 0.12	0.74 ± 0.21	83.30	0.88 ± 0.10	0.83 ± 0.16	91.70
Entropy	0.85 ± 0.18	0.86 ± 0.15	83.30	0.94 ± 0.01	0.95 ± 0.03	100.0
Mean Frequency	0.76 ± 0.21	0.78 ± 0.20	66.70	0.97 ± 0.02	0.96 ± 0.02	100.0
Pitch	0.93 ± 0.04	0.91 ± 0.06	100.0	0.94 ± 0.06	0.94 ± 0.04	100.0
Amplitude	0.92 ± 0.06	0.92 ± 0.03	100.0	0.95 ± 0.03	0.93 ± 0.02	100.0
FM	0.86 ± 0.08	0.86 ± 0.06	100.0	0.92 ± 0.05	0.90 ± 0.03	100.0
Goodness of pitch	0.78 ± 0.11	0.81 ± 0.13	91.70	0.84 ± 0.11	0.82 ± 0.11	100.0
First formant	0.86 ± 0.11	0.84 ± 0.16	91.70	0.94 ± 0.03	0.93 ± 0.03	100.0
Second formant	0.61 ± 0.35	0.56 ± 0.29	66.70	0.89 ± 0.03	0.77 ± 0.30	91.70
Third formant	0.76 ± 0.21	0.70 ± 0.19	75.00	0.78 ± 0.23	0.81 ± 0.14	83.30

Data are mean ± SD.
AM, Amplitude modulation; FM, Frequency modulation.

Supplementary Table 4. Probability of recognition of learned and version sounds, and their formants.

Target (T)	Sound		Formant frequency		
			First (F1)	Second (F2)	F1 & F2
	$coo_{learned}$	0.94 ± 0.05	0.69 ± 0.17	0.60 ± 0.13	0.90 ± 0.08
	$coo_{version}$	0.97 ± 0.02	0.72 ± 0.18	0.70 ± 0.14	0.84 ± 0.15
	$moo_{learned}$	0.97 ± 0.04	0.92 ± 0.05	0.74 ± 0.15	0.98 ± 0.03
	$moo_{version}$	1.00 ± 0.00	0.80 ± 0.10	0.22 ± 0.12 (n.s.)	0.62 ± 0.17
	['pwɛr.ta]$_{learned}$	0.96 ± 0.05	0.81 ± 0.13	0.88 ± 0.08	0.89 ± 0.09
	['pwɛr.ta]$_{version}$	0.99 ± 0.01	0.71 ± 0.16	0.82 ± 0.05	0.77 ± 0.16
	[ko.'mi.ða]$_{learned}$	0.96 ± 0.05	0.92 ± 0.07	0.90 ± 0.05	0.94 ± 0.06
	[ko.'mi.ða]$_{version}$	0.85 ± 0.12	0.56 ± 0.13 (n.s.)	0.59 ± 0.20 (n.s.)	0.68 ± 0.15
	$mean_{learned}$	0.96 ± 0.01	0.84 ± 0.10	0.78 ± 0.13	0.93 ± 0.03
	$mean_{version}$	0.95 ± 0.06	0.70 ± 0.10	0.58 ± 0.26	0.73 ± 0.09
	Total	0.95 ± 0.04	0.77 ± 0.12	0.68 ± 0.22	0.83 ± 0.12
Nontarget (N)					
	$grunt_{learned}$	0.02 ± 0.04	0.24 ± 0.08	0.07 ± 0.09	0.09 ± 0.06
	$grunt_{version}$	0.03 ± 0.05	0.28 ± 0.15	0.07 ± 0.08	0.15 ± 0.11
	$woof_{learned}$	0.05 ± 0.05	0.09 ± 0.04	0.34 ± 0.11	0.06 ± 0.04
	$woof_{version}$	0.08 ± 0.09	0.12 ± 0.13	0.34 ± 0.24 (n.s.)	0.07 ± 0.07
	[si]$_{learned}$	0.04 ± 0.06	0.27 ± 0.18	0.09 ± 0.07	0.18 ± 0.11
	[si]$_{version}$	0.02 ± 0.06	0.29 ± 0.20 (n.s.)	0.14 ± 0.15	0.25 ± 0.20
	['tʃaŋ.gi]$_{learned}$	0.03 ± 0.04	0.41 ± 0.14 (n.s.)	0.06 ± 0.07	0.23 ± 0.11
	['tʃaŋ.gi]$_{version}$	0.01 ± 0.03	0.40 ± 0.13 (n.s.)	0.07 ± 0.09	0.25 ± 0.10
	['xaw.la]$_{learned}$	0.07 ± 0.06	0.36 ± 0.24 (n.s.)	0.17 ± 0.09	0.16 ± 0.10
	['xaw.la]$_{version}$	0.10 ± 0.06	0.42 ± 0.15 (n.s.)	0.37 ± 0.08	0.49 ± 0.17 (n.s.)
	$mean_{learned}$	0.04 ± 0.02	0.27 ± 0.12	0.15 ± 0.11	0.14 ± 0.06
	$mean_{version}$	0.05 ± 0.03	0.30 ± 0.12	0.20 ± 0.14	0.24 ± 0.16
	Total	0.05 ± 0.02	0.29 ± 0.11	0.17 ± 0.12	0.19 ± 0.12

T & N	Proportion of sounds recognized significantly above chance level.			
		F1	F2	F1 & F2
Learned	100 %	77 %	100 %	100 %
Version	100 %	55 %	66 %	90 %
Total	100 %	66 %	83 %	94 %

Data are mean ± SD.
n.s., no significantly above chance level (one-sample sign test, $p > 0.05$.).