Face and Voice Perception: Monkey see, monkey hear

Michael S. Beauchamp
Department of Neurosurgery, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104 USA
Correspondence: michael.beauchamp@pennmedicine.upenn.edu

Primate brains contain specialized areas for perceiving social cues. New research shows that only some of these areas integrate visual faces with auditory voices.

Humans and our primate cousins, Rhesus macaque monkeys, spend a large part of each day interacting with conspecifics. The brains of both species devote a correspondingly large amount of territory to enabling social communication. In this issue, Khandhadia *et al.*[1] report experiments studying macaque areas AF and AM, two nodes of the social brain located in the anterior temporal lobe.

Primate social communication is *multisensory*, meaning that we receive information about the inner state of others from multiple sensory modalities[2]. Two important social cues are visual information from the face of our interaction partner and auditory information from their voice.

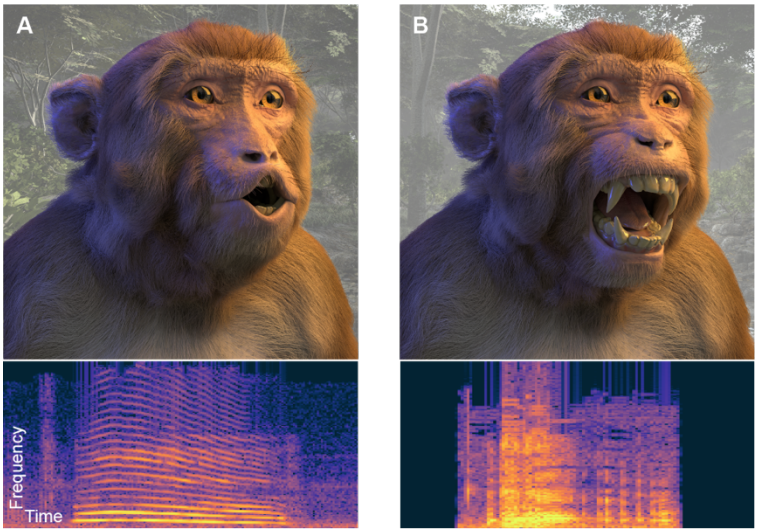

Figure 1. Monkeys communicate with visual and auditory cues.

(A) To express friendly, affiliative intent, monkeys adopt an open-mouth facial posture. This facial gesture is accompanied by a high-pitched cooing sound, spectrogram shown. (B) To express hostility and aggressive intent, monkeys adopt a wide-open grimace, accompanied by a short, harsh bark vocalization. Both images were artificially generated from a parameterized digital 3D model of the macaque face.

In macaques, an affiliative facial expression consists of the mouth opened in an "O" shape accompanied by a high-pitched "coo" call (Figure 1A), while aggression is signaled with a wide, teeth-bared grimace and a guttural "bark" (Figure 1B). For the sender, spreading social cues across modalities makes the signal more resistant to degradation, such as obscuring foliage (visual) or background noise (auditory). For the receiver, combining information across modalities allows for the best possible estimate

of the social signal regardless of environmental conditions, important because social errors (such as confusing hostility for friendliness) can have serious consequences.

Substantial progress has been made in our understanding of the neural substrates of the perception of faces and voices. One milestone was the use of functional magnetic resonance imaging (fMRI) to study brain responses to visually presented faces. In humans, these studies revealed hotspots of face-selective activity in the temporal lobe and elsewhere[3, 4]. In macaques, a similar constellation of hotspots, termed "face patches", was observed with fMRI[5], with the added ability to use invasive electrophysiology to probe the composition of each patch. Single-neuron recording confirmed that individual face-patch neurons were indeed highly face-selective, with most neurons in a patch preferring faces to other visual stimuli[6].

An outstanding question in the field is whether responses in macaque face patches are influenced by auditory voices. Khandhadia *et al.* addressed this question by using fMRI to identify the location of two face patches: AF, in the fundus of the anterior superior temporal sulcus (STS), and AM, on the ventral surface of anterior temporal lobe (Figure 2A).

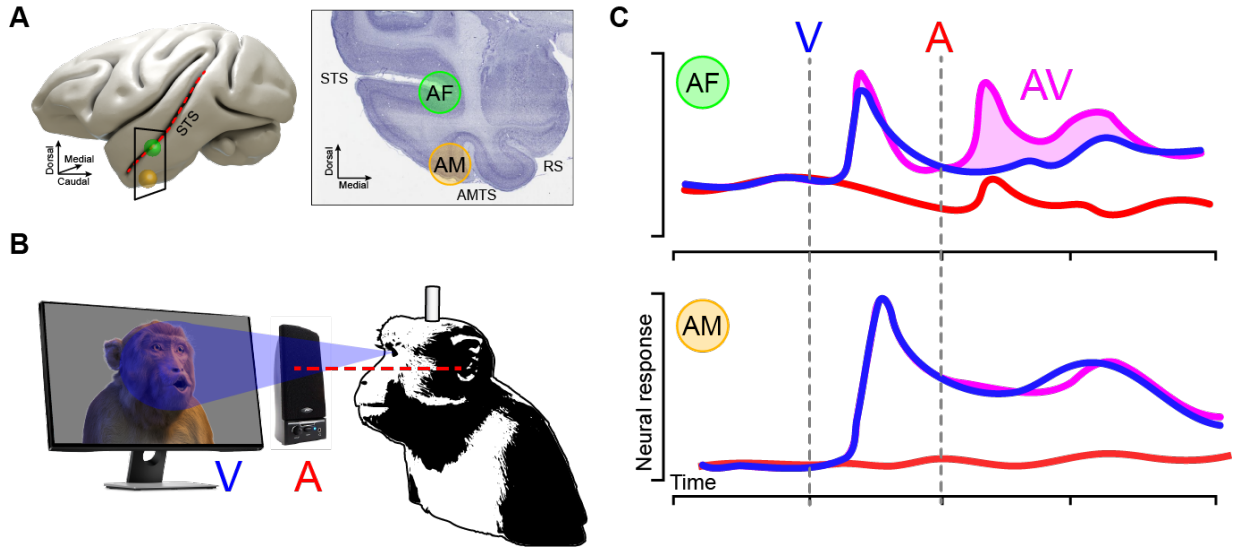

Figure 2. Brain responses to visual faces and auditory voices.

(A) There are multiple areas in the monkey brain specialized for processing social cues. fMRI was used to identify two face patches, AF (anterior fundus, green sphere) and AM (anterior medial, yellow sphere). Dashed red line shows the superior temporal sulcus (STS). Black square shows the location of the anatomical cross-section shown in the right panel (AMTS: anterior medial temporal sulcus; RS: rhinal sulcus). (B) Monkeys viewed face videos (V) and listened to auditory vocalizations (A) while seated in the experimental apparatus. (C) Responses of an example neuron in AF (top plot) and AM (bottom plot) to visual, auditory, and audiovisual recordings. Dashed grey line shows stimulus onset. Shaded area highlights enhanced response during audiovisual stimulation in an AF neuron.

Arrays consisting of 64 microwires were implanted in one face patch in each of four macaques. Compared with traditional microelectrode recording, microwire arrays allow for recording of many more neurons over a time span of days or weeks. The arrays were used to record neuronal responses while monkeys, seated in an experimental chair with head fixed, viewed and listened to monkey faces and voices presented with a computer monitor and speakers (Figure 2B).

Both face patches responded strongly to visual faces and weakly to audiovisual voices (Figure 2C). However, responses to audiovisual faces + voices differed between the patches. In face patch AF, 91 of 119 recorded neurons were multisensory, meaning that they exhibited a significant auditory modulation of their visual response. The most common pattern was auditory enhancement, with the response to audiovisual movies greater than the response to visual movies. In contrast, in face patch AM, neurons responded identically to audiovisual and visual movies, with zero of 55 neurons classified as multisensory.

It might seem surprising that although AF and AM are only 1 cm away from each other in temporal lobe, their responses to multisensory faces are so different. In human temporal lobe, one face-responsive region sits on the lateral surface of the temporal lobe in the posterior STS (pSTS) and another lies on the ventral surface of the temporal lobe in fusiform gyrus (fusiform face area, FFA). The human pSTS is *dynamic*, responding more strongly to moving faces or bodies than to static images[7] with a preference for the visual mouth movements that convey the contents of speech[8] and *multisensory*, responding to auditory, visual and somatosensory stimuli[9] especially spoken voices[10]. In contrast, the FFA shows similar responses to static and dynamic faces[11] with weak or no responses to voices[12]. This corresponds with the observations of Khandhadia *et al.*: audiovisual interactions were observed in face patch AF, in the fundus of the macaque STS, but not AM, on the ventral surface of the temporal lobe, corresponding to the ventral location of human FFA.

In a second set of experiments, Khandhadia *et al.* examined the stimulus properties driving multisensory responses in face patch AF. When the visual face was replaced by a simple disc that expanded and contracted the same way as the mouth, few multisensory interactions were observed. On the other hand, when the voice was replaced with broadband noise with the same amplitude envelope as the original vocalization, multisensory interactions remained prevalent, indicating that the precise frequency contents of the auditory stimulus are not critical for multisensory interactions in AF.

In both experiments, Khandhadia *et al.* found that the presence of the voice increased the response to visually-presented faces for most AF neurons, corresponding to the multisensory enhancement observed in fMRI studies of human pSTS[13, 14]. Other types of multisensory interactions also exist between face and voice. In intracranial studies of the human posterior superior temporal gyrus (pSTG), a visual face can decrease the response to a vocal stimulus, known as multisensory suppression[15, 16]. In another type of interaction, multisensory stimuli can decrease the variability of neuronal responses[17].

The work of Khandhadia *et al.* raises many interesting questions. While monkeys and humans have a dozen or so face-selective regions, Khandhadia *et al.* examined only two of them, spurring curiosity about the multisensory properties of the others. In the studies of Khandhadia *et al.*, stimuli were presented passively, meaning that no information was available about the behavioral relevance of the neural responses. This should be explored in future studies by combining behavioral tasks with causal manipulations such as cooling[18] or microstimulation[19]. Another goal for future studies will be improved neuroethological validity. The monkeys of Khandhadia *et al.* were immobilized in the experimental apparatus and viewed faces on a computer

monitor at a fixed distance, an unnatural arrangement. Advances in wireless neural recording have made it possible to study natural social interactions among large groups of animals, providing new views of social brain computations[20]. Khandhadia *et al.* blaze a trail for future studies of one of the most interesting and important abilities of primates, that of face-to-face communication.

REFERENCES

1. Khandhadia, A.P., Murphy, A.P., Romanski, L.M., Bizley, J.K., and Leopold, D.A. (2021). Audiovisual Integration in Macaque Face Patch Neurons. Current Biology *in press*.
2. Ghazanfar, A.A. (2013). Multisensory vocal communication in primates and the evolution of rhythmic speech. Behav Ecol Sociobiol *67*.
3. Kanwisher, N., McDermott, J., and Chun, M.M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. J Neurosci *17*, 4302-4311.
4. Haxby, J.V., Hoffman, E.A., and Gobbini, M.I. (2000). The distributed human neural system for face perception. Trends Cogn Sci *4*, 223-233.
5. Tsao, D.Y., Freiwald, W.A., Knutsen, T.A., Mandeville, J.B., and Tootell, R.B. (2003). Faces and objects in macaque cerebral cortex. Nat Neurosci *6*, 989-995.
6. Tsao, D.Y., Freiwald, W.A., Tootell, R.B., and Livingstone, M.S. (2006). A cortical region consisting entirely of face-selective cells. Science *311*, 670-674.
7. Beauchamp, M.S., Lee, K.E., Haxby, J.V., and Martin, A. (2003). FMRI responses to video and point-light displays of moving humans and manipulable objects. J Cogn Neurosci *15*, 991-1001.
8. Zhu, L.L., and Beauchamp, M.S. (2017). Mouth and Voice: A Relationship between Visual and Auditory Preference in the Human Superior Temporal Sulcus. J Neurosci *37*, 2697-2708.
9. Beauchamp, M.S., Yasar, N.E., Frye, R.E., and Ro, T. (2008). Touch, sound and vision in human superior temporal sulcus. Neuroimage *41*, 1011-1020.
10. Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P., and Pike, B. (2000). Voice-selective areas in human auditory cortex. Nature *403*, 309-312.
11. Pitcher, D., Dilks, D.D., Saxe, R.R., Triantafyllou, C., and Kanwisher, N. (2011). Differential selectivity for dynamic versus static information in face-selective cortical regions. Neuroimage *56*, 2356-2363.
12. von Kriegstein, K., Kleinschmidt, A., Sterzer, P., and Giraud, A.L. (2005). Interaction of face and voice areas during speaker recognition. J Cognit Neurosci *17*, 367-376.
13. Wright, T.M., Pelphrey, K.A., Allison, T., McKeown, M.J., and McCarthy, G. (2003). Polysensory interactions along lateral temporal regions evoked by audiovisual speech. Cereb Cortex *13*, 1034-1043.
14. Beauchamp, M.S., Lee, K.E., Argall, B.D., and Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. Neuron *41*, 809-823.
15. Besle, J., Fischer, C., Bidet-Caulet, A., Lecaignard, F., Bertrand, O., and Giard, M.H. (2008). Visual activation and audiovisual interactions in the auditory cortex during speech perception: intracranial recordings in humans. J Neurosci *28*, 14301-14310.
16. Karas, P.J., Magnotti, J.F., Metzger, B.A., Zhu, L.L., Smith, K.B., Yoshor, D., and Beauchamp, M.S. (2019). The visual speech head start improves perception and reduces superior temporal cortex responses to auditory speech. Elife *8*, 1-19.
17. Kayser, C., Logothetis, N.K., and Panzeri, S. (2010). Visual enhancement of the information representation in auditory cortex. Curr Biol *20*, 19-24.

18. Plakke, B., Hwang, J., and Romanski, L.M. (2015). Inactivation of Primate Prefrontal Cortex Impairs Auditory and Audiovisual Working Memory. J Neurosci *35*, 9666-9675.
19. Tsunada, J., Liu, A.S., Gold, J.I., and Cohen, Y.E. (2016). Causal contribution of primate auditory cortex to auditory perceptual decision-making. Nat Neurosci *19*, 135-142.
20. Zhang, W., and Yartsev, M.M. (2019). Correlated Neural Activity across the Brains of Socially Interacting Bats. Cell *178*, 413-428 e422.