Distinct natural syllable-selective neuronal ensembles in the primary auditory cortex of awake marmosets

Huan-huan Zeng[1,3#], Jun-feng Huang[1,2,3#], Zhiming Shen[1,3], Neng Gong[1,3], Yun-qing Wen[1,3], Liping Wang[1,3*] and Mu-ming Poo[1,3*]

[1]Center for Excellence in Brain Science and Intelligence Technology, Institute of Neuroscience, State Key Laboratory of Neuroscience, CAS Key Laboratory of Primate Neurobiology, Chinese Academy of Sciences, Shanghai, China

[2]University of Chinese Academy of Sciences, Beijing, China

[3]Shanghai Center for Brain Science and Brain-Inspired Intelligence Technology, Shanghai, China

[#]These authors contributed equally to this work

[*] Send correspondence to M-m. Poo (mpoo@ion.ac.cn) and L-p. W. (liping.wang@ion.ac.cn)

Abstract

Vocal communication is crucial for animals' survival, but the underlying neural mechanism remains largely unclear. Using calcium imaging of large neuronal populations in the primary auditory cortex (A1) of head-fixed awake marmosets, we found specific ensembles of A1 neurons that responded selectively to distinct monosyllables or disyllables in natural marmoset calls. These selective responses were stable over one-week recording time, and disyllable-selective cells completely lost selective responses after anesthesia. No selective response was found for novel disyllables constructed by reversing the sequence of constituent monosyllables or by extending the interval between them beyond ~1 second. These findings indicate that neuronal selectivity to natural calls exists in A1 and pave the way for studying circuit mechanisms underlying vocal communication in awake non-human primates.

One Sentence Summary: Primary auditory cortex neurons in awake marmosets can encode the sequence and interval of syllables in natural calls.

How neural circuits in the brain process vocal signals in vertebrates is largely unknown. For most primates, calls that mediate interactions among conspecifics are crucial for survival (*1*). Besides calling for foods and alarms, primates use calls to judge the intention and motivational levels of others and modulate their own behaviors appropriately (*2, 3*). As a highly social non-human primate species living in families, marmoset represents a desirable animal model for studying neural substrates underlying complex vocal communication (*4-6*). Previous functional magnetic resonance imaging studies on non-human primates and humans have shown a caudal-to-rostral gradient of vocal sound-selectivity from the primary auditory cortex (A1) to higher auditory areas, with regions in the anterior temporal lobe exhibiting the highest preference for complex vocal sounds or speech (*7-10*). Neurophysiological studies in macaque auditory cortices showed that single neuron responses to calls and other salient sounds are more selective in rostral regions of the superior temporal cortex (the 'rostrotemporal polar area') than in the more caudal A1 area (*11*). Analogous to face cells in the visual system, neurons highly selective to specific calls are thought to reside in higher auditory areas. However, the possibility remains that early stages of the cortical pathway such as A1 could also encode specific calls.

In this study, we examined the coding property of A1 neurons in the common marmoset (*Callithrix jacchus*). We performed two-photon calcium imaging to monitor neuronal activity over a large population of A1 neurons in head-fixed awake animals at the single-cell resolution. This approach allowed us to identify distinct syllable-specific ensembles of layer 2/3 neurons that respond selectively to monosyllables or disyllables found in natural marmoset calls, with stable selectivity over one-week recording period. We also found that these syllable-selective responses are highly susceptible to disruption by anesthesia and that there is a stringent requirement for the sequence and temporal proximity of the two monosyllables constituting the disyllable. These results indicate that auditory processing of natural calls occurs at the earliest stage of the cortical pathway, and underscore the advantage of examining neuronal activity at the single-cell resolution over large neuronal populations in the awake animal.

Calcium imaging of neuronal activity in awake marmoset A1

We first performed imaging of intrinsic optical signals in anesthetized marmosets to identify the A1 area based on its tonotopic organization (Fig. 1A), as described in previous reports (*12*, *13*). We then loaded the synthetic Ca^{2+}-sensitive dye Cal-520AM (*14*) into specific tonotopic A1 areas in head-fixed anesthetized marmosets, and labeled neurons were identified by their soma morphology (see Methods). Two-photon calcium imaging of A1 neuronal activity began 2 hours after dye loading when the marmoset was in the awake state (Fig. 1B). When we presented 3 monosyllables (Phee, Twitter, and Trill) and 2 disyllables (TrillPhee and TrillTwitter) in a random sequence, many individual neurons responded preferentially to one or multiple syllables, as shown by the changes ($\Delta F/F$) in Ca-520AM fluorescence (Fig. 1C). In an alternative approach, we injected tetracycline (Tet)-activated AAV vector expressing genetically encoded calcium indicator GCaMP6f into A1 and performed two-photon imaging 4 weeks after injection and 3 days after Tet feeding (Methods, Fig. 1D). Similar robust differential calcium responses to various syllables were also observed (Fig. 1E), although the number of cells expressing GCaMP6f was in general lower than that of Cal-520AM-loaded cells. Since both calcium imaging approaches yielded similar results, the data were pooled in some analyses.

The GCaMP6 expression approach allows long-term recording of the same population of neurons in the marmoset A1. We have recorded syllable-evoked responses in a marmoset over a 1-week period (on day 1, 4 and 8, Fig. 1F) and found that the preferential responses of various neurons were largely maintained. This relative stability of calcium signals is illustrated by the similarity in normalized $\Delta F/F$ with time for each responsive neuron (Fig. 1G). We did notice, however, a gradual reduction of the absolute magnitude of GCaMP6f signals over the 1-week period, presumably due to the reduced GCaMP6f expression with time in the Tet-on expression system (*15*).

Monosyllable- and disyllable-selective A1 neurons

Among the 5 syllables recorded from our marmoset colony, we chose 4 most common syllables for the standard set for this study: monosyllables Phee (P), Twitter (Tw), Trill (Tr) and disyllable TrillPhee (TrP) (*6*). Each syllable has distinct spectral and temporal dynamics, but all have dominant spectral power at frequencies around 8 to 10 kHz (Fig. 2A). When these syllables

were presented in a random sequence to the awake animal, many cells responded selectively to a specific syllable, as depicted by the examples in Figure 2B. Twitter-selective A1 neurons were previously detected in anesthetized marmosets by electrophysiological recording (*16, 17*). Our calcium imaging in awake animals now uncovered substantial populations of neurons that responded selectively to all 4 standard syllables. We defined a neuron to be syllable-selective when the mean response (ΔF/F) evoked by one syllable was significantly larger than those by the other 3 syllables ($P < 0.05$, ANOVA; see Methods).

Data for all syllable-selective neurons recorded from 3 marmosets are summarized by heat maps, with cells sorted according to the time of syllable-evoked peak ΔF/F signal (Fig. 2C, Fig. 3E and Fig. S2A). Notably, the peak response time for neurons within each syllable-specific ensemble varied widely across the entire syllable duration, and some neurons showed sustained responses after the syllable offset (Fig. 2C). The number of neurons with different peak response times was non-uniform, with higher number of cells with peak responses at one or more distinct times (Fig. 2D). Furthermore, we found substantial variability in the relative sizes of 4 syllable-specific ensembles among the 3 marmosets studied (Fig. 2C, Fig. 3E, Fig. S2A), possibly reflecting different developmental history of individual marmosets (*18-20*).

Among all A1 neurons examined in 3 marmosets using Cal-520AM loading, syllable-selective neurons comprised ~23% (674/2891) of all neurons examined. A small population of responsive neurons (75/326) exhibited similar mean response amplitudes for 2 or 3 syllables (Fig. 2E, with $P > 0.05$, *t*-test; see Methods), and a few neurons showed positive ΔF/F in response to one syllable but negative ΔF/F to another (Fig. S2B). Among single syllable-selective neurons, Twitter neurons were most common, followed by Phee neurons, and TrillPhee and Trill neurons were less common (Fig. 2F). The predominance of Twitter neurons may account for the fact that they were the only type of syllable-selective neurons detected by electrophysiological recording from anesthetized animals (*21*). The syllable selectivity was further quantified using syllable selectivity index (SSI, see Methods), and most syllable-selective neurons exhibited high selectivity (with SSI > 0.33, or 2-fold difference, Fig. 2G).

In the above experiments, the tonotopic property (determined by prior intrinsic optical imaging) of the A1 area chosen for the measurement varied from 2 to 8 kHz. We have also measured responses to pure tones (in the range 0.5-16 kHz) as well as test syllables in two Cal-520AM loaded marmosets (Fig. S3A-D). We found that the 8-kHz area contained more syllable-selective neurons than pure-tone responsive neurons (25% *vs.* 14%), whereas the opposite was true for 2-kHz area (22% *vs.* 38%) (Fig. S3E-H). Furthermore, the percentage of neurons showing both syllable-selective and pure-tone responses was lower in 8-kHz area (6%) than that in 2 kHz area (12%). This is consistent with the dominant spectral power of natural marmoset syllables at 8-kHz. Nevertheless, 2-kHz area still contained a substantial number of syllable-selective neurons.

For 3 marmosets labeled with Cal-520AM, we further examined the spatial distribution of syllable-selective neurons within the imaging area, and found that these cells were largely intermingled (Fig. 2H). However, the nearest-neighbor distances for neurons in the same syllable-specific ensemble were on average smaller than those found for the same number of randomly selected neurons regardless of their syllable selectivity (Fig. 2I, $P < 0.05$, bootstrap), suggesting closer spatial proximity of neurons within each ensemble.

Disyllable-selective neurons in marmoset A1

Marmosets make disyllable calls comprising two temporally linked monosyllables. Consistent with previous reports (*22*), we detected two types of disyllable, TrillPhee (TrP) and TrillTwitter (TrTw), in our marmoset colony (Fig. 3A and 3B). As illustrated in Fig. 3A and 3B, neuronal responses to these two disyllables occurred with a substantial delay, mostly after the onset of the second constituent monosyllable, with a small minority of them also responded weakly to isolated constituent monosyllables (cell 2, Fig. 3A and 3B). Thus, disyllable neurons responded to two temporally linked monosyllables rather than monosyllables themselves. On the other hand, the lack of response of Phee and Twitter neurons (see Fig. S2B and S2C) to the same monosyllable within the disyllable indicates that immediate prior presence of Trill suppressed the response of Phee and Twitter neurons. These findings suggest that higher-order processing via intracortical circuits or top-down feedback may be involved in generating syllable-selective responses.

All imaging data on TrillPhee neurons (Fig. 3C, 2 Cal-520AM marmosets) and TrillTwitter neurons (Fig. 3D, 1 GCaMP6 marmoset) were summarized by the heat map, together with all neurons responsive to the constituent monosyllables (Fig. 3D). These maps clearly demonstrate that TrillPhee and TrillTwitter neurons in general did not respond during the initial presence of Trill and the size of disyllable ensembles was as large as those of their constituent monosyllables.

Previous studies on auditory processing in non-human primates were performed mostly on anesthetized animals (*21*). Single-unit recordings showed that only neurons with transient sound-evoked firing could be found in anesthetized marmosets, but sustained firing was recorded from some neurons in awake animals (*23*). In this study, we have compared the response properties of the same population of syllable-selective neurons before and after anesthesia with a fentanyl cocktail (*12*). As shown by the heat-map for all syllable-selective neurons (Fig. 3E), anesthesia markedly reduced both the amplitude and duration of syllable-evoked responses. TrillPhee neuron became completely non-responsive after anesthesia. Some monosyllable neurons kept their response selectivity but altered their response profiles (for Twitter neuron, Fig. 3F; for Phee and Trill neurons, Fig. S5B). The results for all syllable-selective neurons (n = 62) were summarized by plotting the average syllable-evoked peak responses before and after anesthesia (Fig. 3G). We found that anesthesia resulted in the complete loss of responsiveness in disyllable neurons, and monosyllable neurons were significantly reduced in number and syllable selectivity, as measured by SSIs (Fig. S5C).

Sequence and interval requirement for constituent monosyllables within disyllables

Two critical elements of vocal communication are the temporal sequence and the interval of syllables. We thus further examined whether the disyllable-selective responses of A1 neurons depend on the sequence of and time interval between two constituent monosyllables. We first artificially reconstructed disyllables by reversing the temporal sequence of constituent monosyllables. As shown by two example neurons in Fig. 4A, the selective responses to TrillPhee were completely lost when the sequence of Trill/Phee was reversed to Phee/Trill. One neuron (Fig. 4A, right) also responded with equal amplitude to the isolated Phee. Such loss of disyllable responses after sequence reversal was found for all neurons recorded in a GCaMP6f-expressing

marmoset (Fig. 4B and 4C). Thus, the sequence of constituent monosyllables is critical for disyllable-selective responses.

The requirement of temporal proximity of two constituent monosyllables was further examined by testing the effect of artificially reconstructed disyllables in which the interval between two monosyllables was extended gradually from 10 ms up to 4 sec (Fig. 4D). As shown by example TrillTwitter and TrillPhee neurons (Fig. 4E) and the summary data from 4 disyllable ensembles (Fig. 4F), the peak amplitude of disyllable-selective responses progressively declined as the time interval was extended, and largely disappeared beyond an interval of ~1 sec. For some disyllable neurons, over-extended artificial disyllables could still trigger weak ("residue") responses (Fig. 4E, bottom right panel), although disyllable selectivity was completely lost. These disyllable neurons also responded weakly to isolated monosyllables, with amplitudes similar to those of residual responses. Thus, normal disyllable-selective responses require the temporal proximity of constituent monosyllables to be within ~1 sec.

Novel combination of monosyllables evoked no selective response

We have also constructed novel (artificial) disyllables from two natural monosyllables Twitter and Phee, with the same temporal proximity as those in natural disyllables (Fig. 5A and 4B). These novel disyllables TwitterPhee and PheeTwitter were never observed in our natural marmoset colony. In two marmosets, we found no A1 neuron that showed selective response to either disyllables, and all responsive neurons were selectively responding to either Twitter or Phee (Fig. 5C), with peak amplitudes slightly lower than those evoked by isolated Phee and Twitter, respectively (Fig. 5D-F), suggesting mutual suppressive actions when two monsyllables appeared with close temporal proximity. Thus, disyllable-selective A1 neurons were specific developed for detecting disyllables found in natural calls, rather than any set of temporally linked monosyllables.

Previous studies have shown that selectivity to a specific sensory stimulus could be enhanced and induced by repeated exposure (*24, 25*). We found that after repeatedly exposure of the novel PheeTwitter for 50 or 150 times to an awake marmoset (at 2-s interval), no PheeTwitter-selective response was detected (Fig. 5G). Quantitative analysis of SSIs of neurons that responded to monosyllables Phee and Twitter as well as novel PheeTwitter and TwitterPhee did not change their

SSIs after repetitive exposure to the novel disyllables (Fig. S6). Thus, the circuitry for selective detection of novel disyllable calls could not be shaped simply by the short-term repetitive exposure used in the present study.

Discussion

Previous studies have addressed the mechanisms of cortical coding for conspecific vocal communication in primates (*26*), but whether syllable-selective neurons exist in the tonotopically organized A1 was unclear (*27*). In this study, we found that in both 2- and 8-kHz tonotopic A1 areas of awake marmosets, substantial populations of neurons exhibited selective responses to distinct syllables found in natural marmoset calls. Neurons selectively responding to a monosyllable did not respond to the sequence-reversed syllable (Fig. S7) or to disyllable containing this particular monosyllable, indicating that they were not simply detecting some sound components. Moreover, disyllable responsiveness requires a specific temporal sequence and close proximity of the two constituent monosyllables, consistent with the feature of sequence and interval specificity in vocal communication. That disyllable-selective responses completely disappeared after anesthesia is consistent with the high anesthesia vulnerability of top-down modulation found previously in sensory processing (*28, 29*). Such modulation may be less involved in monosyllable responses that were more persistent after anesthesia. Anesthesia vulnerability of disyllable responses could also be attributed to an overall reduction of neuronal excitation, which prevented the firing of neurons requiring cumulated excitation by sequential monosyllables. Further elucidation of input and output connections of syllable-selective neurons may reveal circuit mechanisms underlying the processing of complex vocal sounds in non-human primates.

Previous studies on auditory processing in A1 have characterized the tonotopic organization and spectra-temporal properties of neuronal responses, involving feed-forward thalamocortical inputs and intracortical processing by local circuits (*30-32*). Several lines of evidence found here point to a more extensive processing of auditory signals than previously recognized. First, substantial fractions of A1 neurons are devoted to detecting complex sound features of natural syllables rather than the frequencies of constituent sounds. Second, responses to artificial

disyllables showed the existence of substantial crosstalk between monosyllables (e.g., Phee and Twitter), since selective responses evoked by each monosyllable was reduced by the immediate prior presence of the other (Fig. 5D-F). Such crosstalk may involve cortical inhibitory circuits within A1. Third, peak activity of individual neurons within each syllable-specific ensemble were found to tile the entire syllable duration in a non-uniform manner (Fig. 2D), suggesting large variation in local recurrent connections or higher-order circuit mechanisms. Finally, disyllable-selective cells are responsive to two monosyllables linked by a proper sequence and an interval less than ~1 sec, and are highly vulnerable to anesthesia. All these findings point to the existence of complex circuitry for temporal sequence and interval processing in A1.

Tonotopic maps in rodent A1 undergo plastic changes following exposure to artificial sounds with specific frequency characteristics (*33*-*35*), indicating plasticity of neural circuits in A1. Syllable-selective responses reported here persisted over the one-week recording period, implicating the stability of underlying neural circuit functions. These circuits are likely to be established during early development for detecting natural sounds relevant to marmosets. Marmoset vocalization undergoes substantial post-natal changes that depend on the social environment, such as the presence of parental vocal feedback (*15*-*17*). In this study, we found that short-term repetitive exposure (up to 20 min) of novel disyllables did not induce the appearance of selectively responsive neurons. However, auditory perceptual learning could result in enhanced cortical response dynamics and mediate improvement of temporal processing in the rat (*36*). Thus, it is possible that circuits for detecting novel syllables and syllable sequences could be established by training adult marmosets in behavioral relevant context or by exposing the marmoset to novel sounds during early development.

In summary, by using optical imaging of large populations of neurons in awake marmosets, we have demonstrated that cortical processing of complex features of vocal sounds occurs in A1. Whether other primary sensory cortices are also capable of processing complex features of natural sensory inputs remains to be explored. With the availability of optical methods for recording neuronal activity at single-cell resolution in non-human primates (*15*, *37*-*39*), together with intracellular recording from awake animals (*40*), further studies of sensory processing in the awake non-human primates is likely to uncover previous unknown cortical mechanisms.

References and notes

1. R. M. Seyfarth, D. L. Cheney, P. Marler, Monkey responses to three different alarm calls: evidence of predator classification and semantic communication. *Science* **210**, 801-803 (1980).
2. L. J. Rogers, L. Stewart, G. Kaplan, Food Calls in Common Marmosets, Callithrix jacchus, and Evidence That One Is Functionally Referential. *Animals (Basel)* **8**, (2018).
3. C. T. Snowdon, Cognitive Components of Vocal Communication: A Case Study. *Animals (Basel)* **8**, (2018).
4. T. Pomberger, C. Risueno-Segovia, J. Loschner, S. R. Hage, Precise Motor Control Enables Rapid Flexibility in Vocal Behavior of Marmoset Monkeys. *Curr Biol* **28**, 788-794 e783 (2018).
5. C. T. Miller *et al.*, Marmosets: A Neuroscientific Model of Human Social Behavior. *Neuron* **90**, 219-233 (2016).
6. S. J. Eliades, C. T. Miller, Marmoset vocal communication: Behavior and neurobiology. *Dev Neurobiol* **77**, 286-299 (2017).
7. S. Sadagopan, N. Z. Temiz-Karayol, H. U. Voss, High-field functional magnetic resonance imaging of vocalization processing in marmosets. *Sci Rep* **5**, 10950 (2015).
8. S. K. Scott, I. S. Johnsrude, The neuroanatomical and functional organization of speech perception. *Trends Neurosci* **26**, 100-107 (2003).
9. C. I. Petkov, N. K. Logothetis, J. Obleser, Where Are the Human Speech and Voice Regions, and Do Other Animals Have Anything Like Them? *Neuroscientist* **15**, 419-429 (2009).
10. P. Belin, C. Bodin, V. Aglieri, A "voice patch" system in the primate brain for processing vocal information? *Hear Res* **366**, 65-74 (2018).
11. Y. Kikuchi, B. Horwitz, M. Mishkin, Hierarchical auditory processing directed rostrally along the monkey's supratemporal plane. *J Neurosci* **30**, 13021-13030 (2010).
12. H. H. Zeng *et al.*, Local homogeneity of tonotopic organization in the primary auditory cortex of marmosets. *Proc Natl Acad Sci U S A* **116**, 3239-3244 (2019).
13. T. Tani *et al.*, Sound Frequency Representation in the Auditory Cortex of the Common Marmoset Visualized Using Optical Intrinsic Signal Imaging. *eNeuro* **5**, (2018).
14. M. Tada, A. Takeuchi, M. Hashizume, K. Kitamura, M. Kano, A highly sensitive fluorescent indicator dye for calcium imaging of neural activity in vitro and in vivo. *Eur J Neurosci* **39**, 1720-1728 (2014).
15. O. Sadakane *et al.*, Long-Term Two-Photon Calcium Imaging of Neuronal Populations with Subcellular Resolution in Adult Non-human Primates. *Cell Rep* **13**, 1989-1999 (2015).
16. X. Wang, M. M. Merzenich, R. Beitel, C. E. Schreiner, Representation of a species-specific vocalization in the primary auditory cortex of the common marmoset: temporal and spectral characteristics. *J Neurophysiol* **74**, 2685-2706 (1995).
17. X. Wang, S. C. Kadia, Differential representation of species-specific primate vocalizations in the auditory cortices of marmoset and cat. *J Neurophysiol* **86**, 2616-2620 (2001).
18. Y. B. Gultekin, S. R. Hage, Limiting parental feedback disrupts vocal development in marmoset monkeys. *Nat Commun* **8**, 14046 (2017).
19. Y. B. Gultekin, S. R. Hage, Limiting parental interaction during vocal development affects acoustic call structure in marmoset monkeys. *Sci Adv* **4**, eaar4012 (2018).

20. D. Y. Takahashi *et al.*, LANGUAGE DEVELOPMENT. The developmental dynamics of marmoset monkey vocal production. *Science* **349**, 734-738 (2015).
21. X. Wang, Cortical Coding of Auditory Features. *Annu Rev Neurosci* **41**, 527-552 (2018).
22. J. A. Agamaite, C. J. Chang, M. S. Osmanski, X. Wang, A quantitative acoustic analysis of the vocal repertoire of the common marmoset (Callithrix jacchus). *J Acoust Soc Am* **138**, 2906-2928 (2015).
23. X. Wang, T. Lu, R. K. Snider, L. Liang, Sustained firing in auditory cortex evoked by preferred stimuli. *Nature* **435**, 341-346 (2005).
24. Y. Li, S. D. Van Hooser, M. Mazurek, L. E. White, D. Fitzpatrick, Experience with moving visual stimuli drives the early development of cortical direction selectivity. *Nature* **456**, 952-956 (2008).
25. F. Engert, H. W. Tao, L. I. Zhang, M. M. Poo, Moving visual stimuli rapidly induce direction sensitivity of developing tectal neurons. *Nature* **419**, 470-475 (2002).
26. L. M. Romanski, B. B. Averbeck, The primate cortical auditory system and neural representation of conspecific vocalizations. *Annu Rev Neurosci* **32**, 315-346 (2009).
27. X. Wang, On cortical coding of vocal communication sounds in primates. *Proc Natl Acad Sci U S A* **97**, 11843-11849 (2000).
28. C. D. Gilbert, M. Sigman, Brain states: top-down influences in sensory processing. *Neuron* **54**, 677-696 (2007).
29. K. V. Nourski *et al.*, Auditory Predictive Coding across Awareness States under Anesthesia: An Intracranial Electrophysiology Study. *J Neurosci* **38**, 8441-8452 (2018).
30. L. I. Zhang, A. Y. Tan, C. E. Schreiner, M. M. Merzenich, Topography and synaptic shaping of direction selectivity in primary auditory cortex. *Nature* **424**, 201-205 (2003).
31. L. Y. Li, Y. T. Li, M. Zhou, H. W. Tao, L. I. Zhang, Intracortical multiplication of thalamocortical signals in mouse auditory cortex. *Nat Neurosci* **16**, 1179-1181 (2013).
32. L. S. Hamilton *et al.*, Optogenetic activation of an inhibitory network enhances feedforward functional connectivity in auditory cortex. *Neuron* **80**, 1066-1076 (2013).
33. H. Nakahara, L. I. Zhang, M. M. Merzenich, Specialization of primary auditory cortex processing by sound exposure in the "critical period". *Proc Natl Acad Sci U S A* **101**, 7170-7174 (2004).
34. Y. K. Han, H. Kover, M. N. Insanally, J. H. Semerdjian, S. Bao, Early experience impairs perceptual discrimination. *Nat Neurosci* **10**, 1191-1197 (2007).
35. L. I. Zhang, S. Bao, M. M. Merzenich, Persistent and specific influences of early acoustic environments on primary auditory cortex. *Nat Neurosci* **4**, 1123-1130 (2001).
36. S. Bao, E. F. Chang, J. Woods, M. M. Merzenich, Temporal plasticity in the primary auditory cortex induced by operant perceptual learning. *Nat Neurosci* **7**, 974-981 (2004).
37. M. Li, F. Liu, H. Jiang, T. S. Lee, S. Tang, Long-Term Two-Photon Imaging in Awake Macaque Monkey. *Neuron* **93**, 1049-1057 e1043 (2017).
38. A. K. Garg, P. Li, M. S. Rashid, E. M. Callaway, Color and orientation are jointly coded and spatially organized in primate primary visual cortex. *Science* **364**, 1275-1279 (2019).
39. R. Ding *et al.*, Targeted Patching and Dendritic Ca(2+) Imaging in Nonhuman Primate Brain in vivo. *Sci Rep* **7**, 2873 (2017).
40. L. Gao, K. Kostlan, Y. Wang, X. Wang, Distinct Subthreshold Mechanisms Underlying Rate-Coding Principles in Primate Auditory Cortex. *Neuron* **91**, 905-919 (2016).

Acknowledgements

We thank Drs. Xiaoqin Wang and Yang Dan for helpful discussion and comments on the manuscript. **Funding:** This work was supported by Strategic Priority Research Program of CAS, Grant No. XDB32000000; Key Research Program of Frontier Sciences [QYZDY-SSW-SMCO01]; International Partnership Program [153D31KYSB20170059]; the Shanghai Key Basic Research Project [16JC1414100;1420201] and Shanghai Municipal Major Project [2018SHZDZX05] and Shanghai Key Basic Research Project [18JC1410100]. **Author contributions:** H.H.Z., L.P.W. and M.M.P. designed the research; H.H.Z. and J.F.H. performed all experiments with Z.M.S., N.G. and Y.Q.W. offered technical help; H.H.Z. analyzed the data; H.H.Z., L.P.W. and M.M.P. wrote the manuscript. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All data necessary to support the paper's conclusions are present in the main text and supplementary materials.

Supplementary Materials

Figures S1-S7

Movies S1-S2

References (41-44)

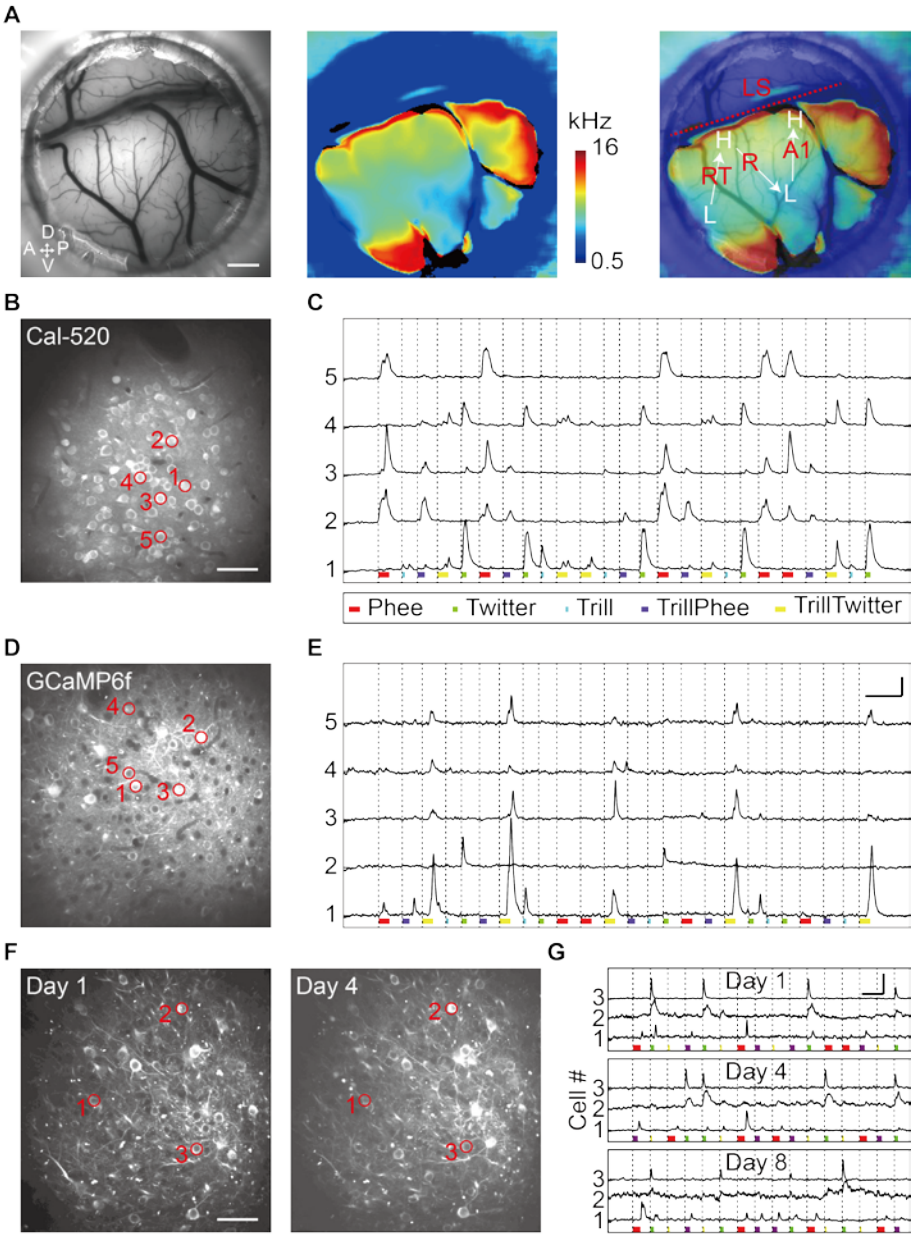

Fig. 1. Two-photon imaging of neuronal activity in awake marmoset A1.

(**A**) Tonotopic map of the auditory cortex obtained by imaging intrinsic optical signals. **Left**: blood vessel map within the imaging window (bar, 1 mm). **Middle**: A tonotopic map revealed by intrinsic optical signals in response to a sequence of pure tone stimuli, color-coded for 21 discrete frequencies in the range of 0.5-16 kHz (same imaging plane as in left). **Right**: Image obtained by merging that in left and middle. LS, lateral sulcus; A1: primary auditory cortex; RT, rostro-temporal field; R, rostral field.

(**B**) Fluorescence image (averaged over 2 min) of a cortical area tonotopically mapped to be 8 kHz-preferring area of A1 of the head-fixed awake marmoset. Calcium-sensitive dye Cal-520AM was loaded into A1 2 hours before imaging. Bar: 50 μm.

(**C**) Relative changes in Cal-520AM fluorescence (ΔF/F) in 5 example cells (marked by circles in B) in response to 5 different call syllables in a random sequence. Stimulus duration marked by the bar below, syllable types coded in colors. **(See corresponding Movie S1)**

(**D** and **E**) Similar to B&C, except that the cortex was injected with tetracycline-dependent AAV expressing GCaMP6f in A1 4 weeks before imaging, and imaging was performed 3 days after tetracycline application. Bars: 5 s and 100% ΔF/F. **(See corresponding Movie S2)**

(**F**) Fluorescence image of an A1 area (averaged over 2 min) of a GCaMP6f-expressing marmoset on day 1 and day 4 of the experiment. Bar: 50 μm.

(**G**) Response profiles of 3 example neurons marked in F, recorded at day 1, 4 and 8. Bars: 5 s and normalized ΔF/F (0 to 1).

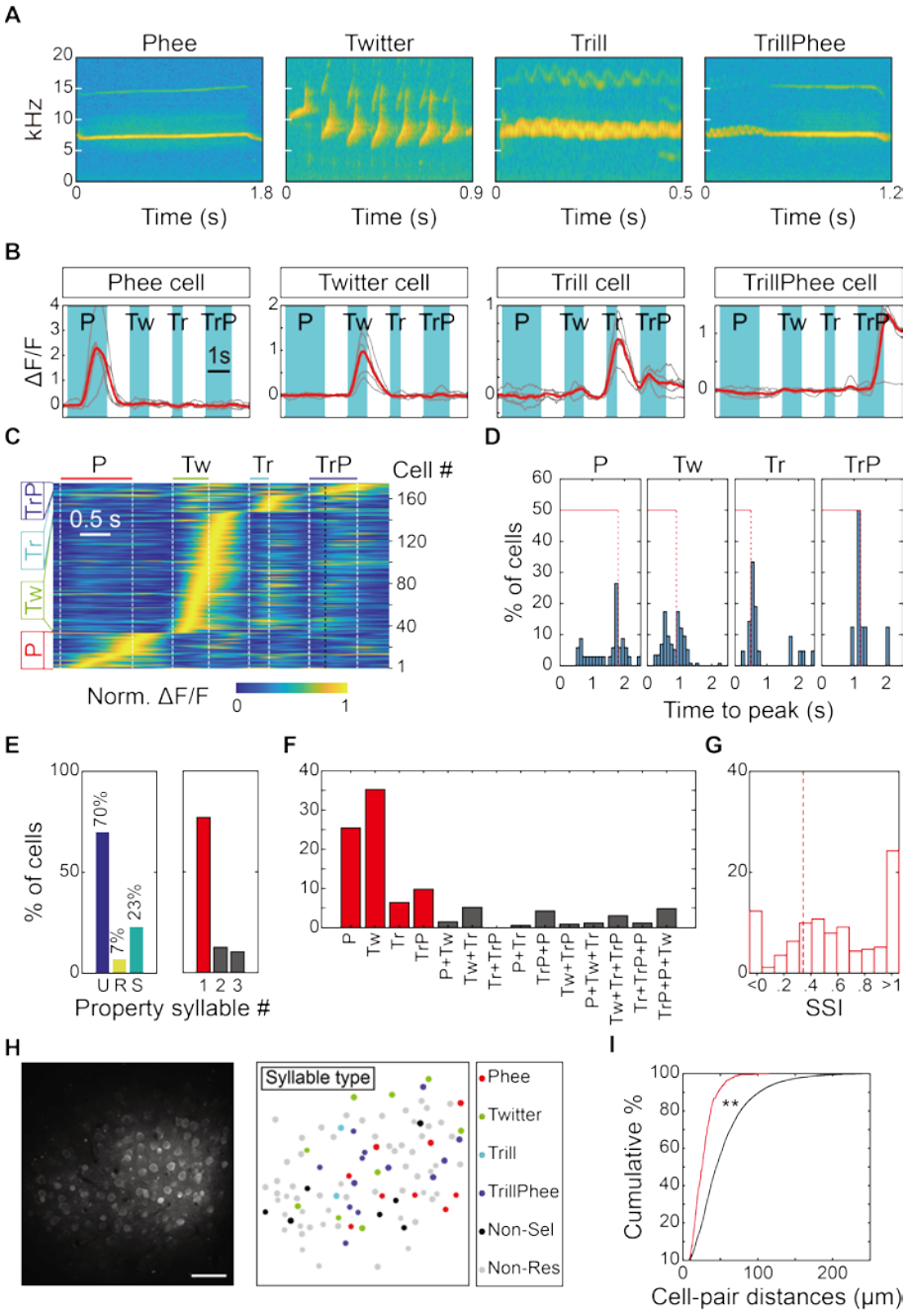

Fig. 2. Syllable-selective cells in awake marmoset A1.

(**A**) Representative spectrograms of 4 standard test stimuli.

(**B**) Fluorescence changes (ΔF/F) recorded in 4 syllable-selective cells in response to test syllables in a random sequence. Gray traces: single trials (n = 5); red traces: average. Cyan shading: syllable duration.

(**C**) Heat map for the activity of all syllable-selective cells in an awake marmoset labeled with Ca-520AM, with ΔF/F normalized for each cell and color-coded by the scale below. The cells are grouped into 4 syllable-selective ensembles, and sorted within each ensemble in an order based on the time of the peak ΔF/F. White dashed lines: syllable onset and offset. Black dashed line: boundary of Trill and Phee components of TrillPhee.

(**D**) Percentages of cells showing different peak-response times within each syllable ensemble shown in C.

(**E**) Statistics of data on syllable-selective cells recorded from 24 imaging fields in 3 marmosets labeled with Cal-520AM. **Left**: Among all cells recorded (n = 2891), the percentages of cells that were unresponsive ("U"), responsive but not syllable-selective ("R") and syllable-selective ("S"). **Right**, the percentages of cells showing syllable selectivity to 1, 2, or 3 syllables among all syllable-selective cells (see Methods).

(**F**) Percentages of cells showing selectivity to single syllable and to different sets of multiple syllables, among all syllable-selective cells.

(**G**) Syllable-selective index (SSI) of all single syllable-selective cells. Red dashed line, SSI = 0.33 (2-fold preference).

(**H**) **Left**, an image of Cal-520AM fluorescence at a recorded region (averaged over 2 min). Bar: 50 μm. **Right**, spatial distribution of all cells in the imaging field, with cell response properties coded in colors.

(**I**) Cumulative percentage plot of nearest-neighbor distances for cells of the same syllable selectivity (red line), and for all cells regardless of syllable selectivity, obtained by bootstrap analysis (black line, see Methods). The difference between two distributions is significant at $P < 0.001$, *Kolmogorov–Smirnov* test).

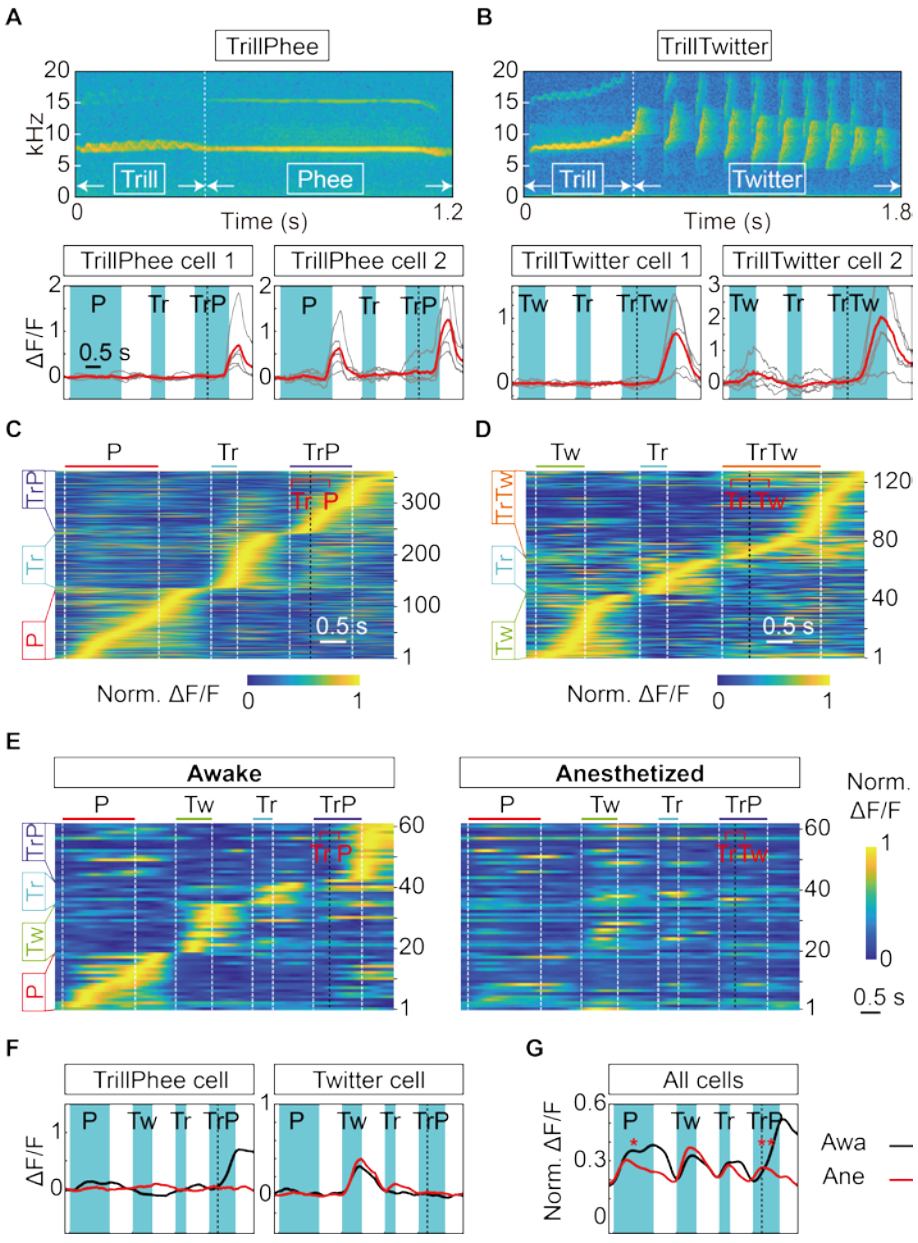

Fig. 3. Properties of disyllable-selective cells and the effect of anesthesia.

(**A** and **B**) Spectrograms of disyllables TrillPhee and TrillTwitter, and selective responses of 2 example cells for each disyllable. Cell 1: responded only to the disyllable; cell 2: also responded to an isolated monosyllable (Phee in A and Twitter in B).

(**C** and **D**) Heat maps of the activity of all cells selectively responding to disyllables (TrillPhee, C; TrillTwitter, D) and constituent monosyllables (Trill, Phee, Twitter), recorded from 3 marmosets labeled with Ca-520AM (n = 2, C) and GCaMP6f (n = 1, D).

(**E**) Heat maps of the activity of all syllable-selective cells recorded from one marmoset, before (left) and after (right) anesthesia, with the same normalization of ΔF/F for each cell. Note the disappearance of disyllable-selective responses after anesthesia.

(**F**) Example cells illustrating syllable-selective responses before (black) and after (red) anesthesia, with each trace depicting averaged signals from 5 trials.

(**G**) Average traces for all syllable-selective cells shown in E before (black) and after (red) anesthesia. The integrated ΔF/F showed significant difference between responses observed before and after anesthesia for disyllable TrillPhee cells and monosyllable Phee cells (P, $P < 0.01$; TrP, $P < 0.001$; Tw, Tr, $P > 0.05$; *t* test).

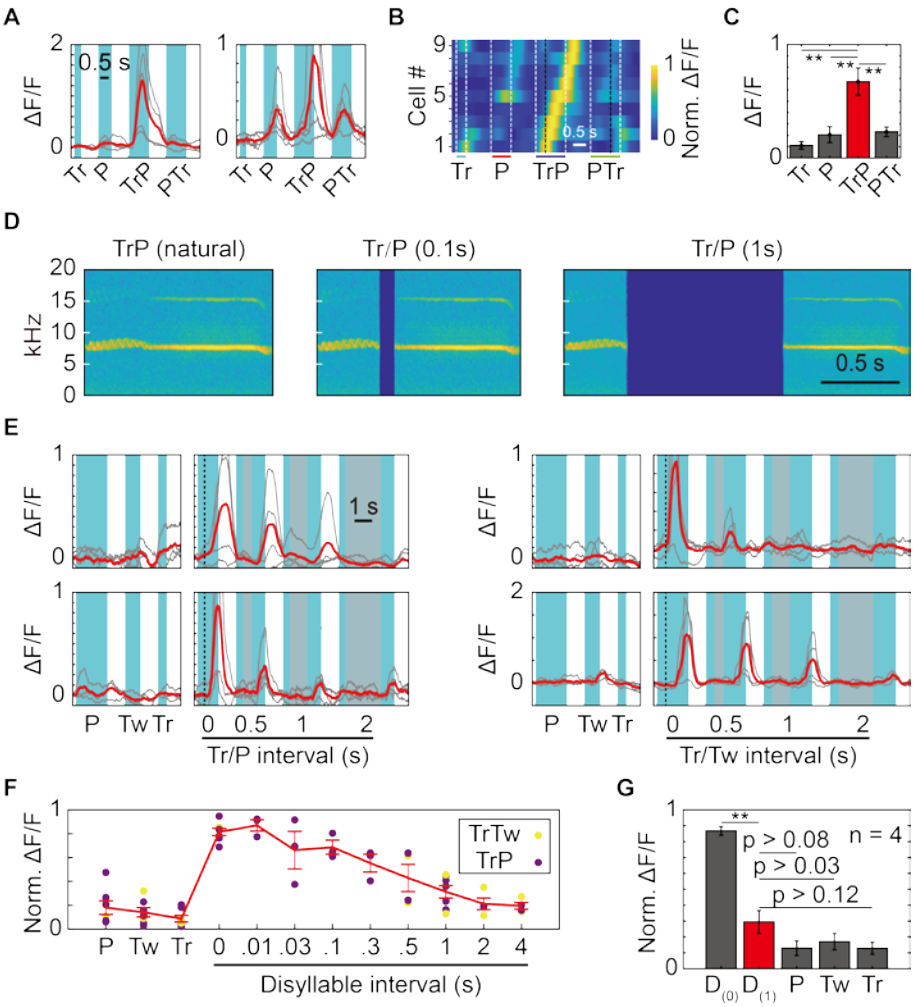

Fig. 4. Dependence of Disyllable Responses on the Sequence and Interval of Constituent Monosyllables.

(**A**) Two examples of TrillPhee-selective neurons showed complete loss of disyllable selectivity when the sequence of Trill/Phee was changed to Phee/Trill.

(**B**) Heat map of TrillPhee-selective neuronal ensemble (n = 9) recorded from one image area of a GCaMP6f-expressing marmoset, showing exclusive responses to TrillPhee but not PheeTrill (ΔF/F normalized for each cell).

(**C**) Summary of average peak values of ΔF/F for all cells shown in B. Data pairs showing significant differences are marked by ** ($P < 0.001$, paired t test).

(**D**) Natural disyllables were reconstructed by artificially extending the interval between two constituent monosyllables, as shown by spectrograms. **Left**, natural disyllable TrillPhee; **Middle** and **Right**: reconstructed disyllables Trill/Phee with 0.1 and 1 sec interval.

(**E**) Two example cells with disyllable-selective responses to natural disyllables (left TrillPhee, right TrillTwitter) and reconstructed disyllables with an interval of 0.5, 1, or 2 s between constituent monosyllables, together with their responses to isolated monosyllables Trill, Phee, and Twitter. Gray lines, individual trials (n = 4); red lines, averages.

(**F**) Summary of all data on responses evoked by reconstructed disyllables with extended intervals from 0.01 to 4 sec (n = 3-7 cells each) and by 3 isolated monosyllables, recorded from one GCaMP6f-expressing marmoset. Red curve: averages at all intervals, with data points depicting normalized peak value of ΔF/F for two disyllables.

(**G**) Averages of normalized peak ΔF/F for data in F, for natural disyllable $D_{(0)}$, extended disyllable at 1-s interval $DS_{(1)}$, and monosyllables. Significant difference (**, $P < 0.01$, paired *t* test).

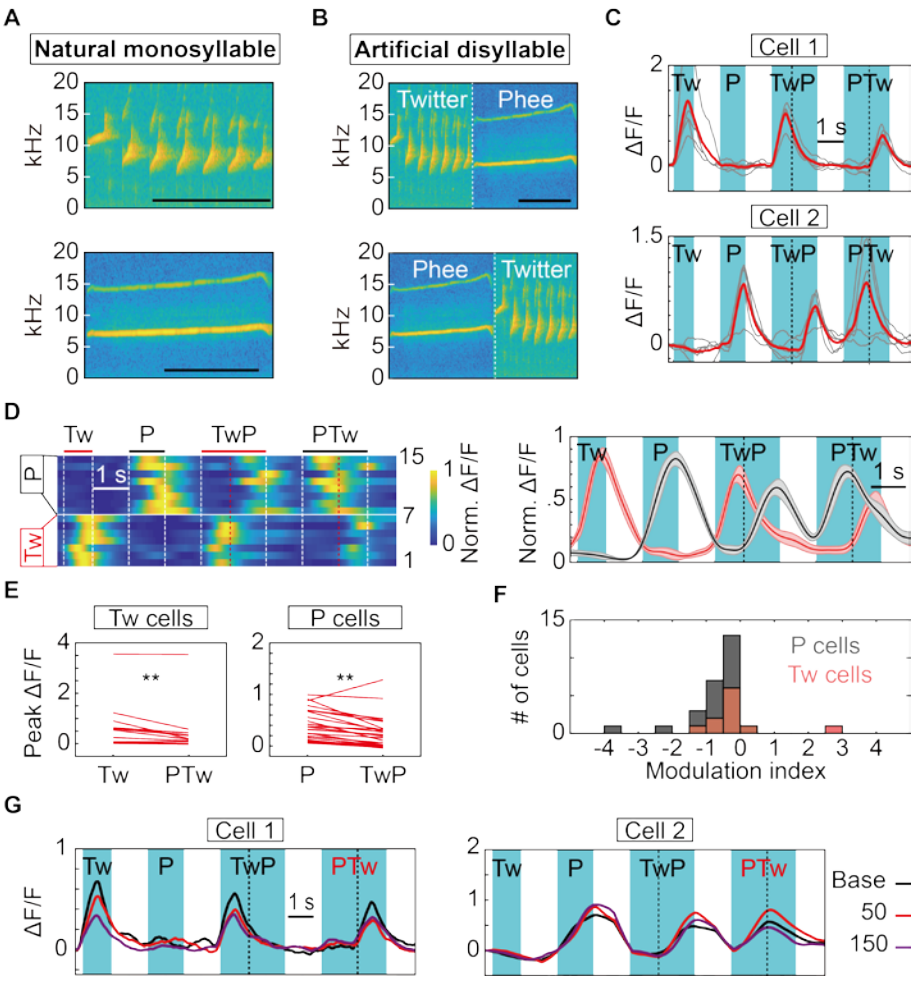

Fig. 5. No Selective Response to Artificially Constructed Novel Disyllables.

(**A** and **B**) Spectrograms of natural Twitter (A, top) and Phee (A, bottom) and novel disyllables (B: Top, TwitterPhee; bottom, PheeTwitter) artificially constructed by linking Twitter and Phee. Bars: 1 s.

(**C**) Single trials (gray lines, n = 5) and mean (red line) fluorescence changes (ΔF/F) evoked by natural monosyllable (Twitter and Phee) and artificial disyllables (TwitterPhee and PheeTwitter) in two example cells. Black dashed line, boundary of Phee and Twitter components.

(**D**) **Left**: Heat map of normalized responses to monosyllables and artificial disyllables for example cells that show selective response to Twitter (n = 7) and Phee (n = 8), and their responses to artificially disyllables. **Right**: Average (± SEM) changes in normalized ΔF/F induced by monosyllables and artificial disyllables for all cells of the Twitter (red) and Phee (black) ensembles, corresponding to the heat map on the left.

(**E**) Comparison of the peak ΔF/F values for individual neurons within the Twitter and Phee ensemble, between responses evoked by isolated monosyllables and the same monosyllable within artificial disyllables. Significant differences were found (**, $P < 0.01$, paired *t* test).

(**F**) The effect of immediate prior presence of another type of monosyllable on the peak ΔF/F values of monosyllable-evoked responses, quantified by the modulation index (see Methods). Note that MIs were predominantly negative for both Twitter and Phee ensembles.

(**G**) The effect of repetitive exposure to novel disyllable PheeTwitter is illustrated by two example neurons. Curves are averaged ΔF/F values prior to (basal, black) and after 50 (red) and 150 (purple) times of repetitive application of PheeTwitter.