Acoustic camera system for measuring ultrasound communication in mice

Jumpei Matsumoto*[1,2], Kouta Kanno*[3], Masahiro Kato[4,5], Hiroshi Nishimaru[1,2], Tsuyoshi Setogawa[1,2], Choijiljav Chinzorig[1], Tomohiro Shibata[6], Hisao Nishijo[1,2]

[1]Department of System Emotional Science, Faculty of Medicine, University of Toyama, Toyama, Japan

[2]Research Center for Idling Brain Science, University of Toyama, Toyama, Japan

[3]Laboratory of Neuroscience, Course of Psychology, Department of Humanities, Faculty of Law, Economics and the Humanities, Kagoshima University, Kagoshima, Japan

[4]Katou Acoustics Consultant Office, Yokohama, Japan

[5]Osawa Memorial Institute of Architectural Environmental Engineering, Kanto Gakuin University, Yokohama, Japan

[6]Department of Human Intelligence Systems, Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology, Kitakyushu, Japan

*Corresponding authors.

Abstract:

Ultrasonic vocalizations in mice have recently been widely investigated as social behavior; however, using existing sound localization systems in home cages, which allow observations of more undisturbed behavior expressions, is challenging. We introduce a novel system, named USVCAM, that uses a phased microphone array and demonstrate novel vocal interactions under a resident-intruder paradigm. The extended applicability and usability of USVCAM may facilitate investigations of social behaviors and underlying physiological mechanisms.

Main:

Ultrasonic vocalizations (USVs) are used for communication by many rodent species (Sales, 2010). Recently, USV communication in mice has received considerable attention as a powerful experimental model to investigate molecular, genetic, and neural mechanisms underlying social behaviors and deficits (Fischer and Hammerschmidt, 2011; Lahvis et al., 2011; Konopka and Roberts, 2016). The advances in sound localization technology in these studies have facilitated the analysis of vocal interactions between multiple subjects (Neunuebel et al., 2015; Sangiamo et al., 2020). However, to date, sound localization systems for mouse USVs (Neunuebel et al., 2015; Heckman et al., 2017; Warren et al., 2018) are built around distributed-microphone arrays (Fig. 1a; distributed-microphone systems), which require a special recording arena that is often equipped with reticulated (acoustic transparent) walls, to carry out behavioral experiments. Therefore, the application of these systems to studies that are ideally performed in home cages, where the animals tend to show more undisturbed expression of social behaviors, is challenging (Kikusui, 2013).

Therefore, we developed USVCAM, which is a novel system for localization and assignment (i.e., identifying which animal vocalizes) of USVs. USVCAM uses a concentrated, phased microphone array (Fig. 1b), which allows recording from only one direction to enable recording in a standard plastic home cage with small modifications. The compact array was fixed near a camera, which allowed a simple acoustic camera design of the sensor assembly. Furthermore, the use of the phased microphone array significantly benefits processing speed to enable near real-time processing because the calculation time window can be as short as a single phase.

The performance and effectiveness of USVCAM were examined by obtaining

recordings from home cages in two strains of mice (C57BL/6 [B6] and ICR), which have distinct vocal communication and social behavior characteristics (Asaba et al., 2014; Golden et al., 2017). To reduce the sound reflection in the home cage for accurate sound localization, we inserted disposable paper towels between a conventional plastic cage and a custom inner cage made of fine stainless mesh (Fig. 1c and Supplementary Fig. 1).

To validate the performance of the localization and assignment of USVs, we recorded from a single male mouse exploring an empty home cage of a female mouse (single mouse experiment), in which we were certain about the source of each USV. For the localization, USVCAM first searches USV segments (continuous peaks in the sound spectrogram) according to a previously proposed USVSEG algorithm (Tachibana et al., 2020; Supplementary Fig. 2). Subsequently, it calculates the averaged spatial spectrum of each USV segment and identifies the peak of the spectrum (Fig. 1d and Supplementary Videos 1 and 2; see Methods for details). The localization errors were quantified as the distance between the snout of the mouse and the spectrum peak (Fig. 1e). The median errors were 1.95 degrees (direction from the microphone array) and 15.8 mm, which were comparable with those reported for previous systems (Neunuebel et al., 2015; Heckman et al., 2017; Warren et al., 2018). The error distributions that were calculated separately for B6 and ICR mice were largely similar (Supplementary Fig. 3), despite differences in the vocal patterns of these strains (Asaba et al., 2014; see also Supplementary Video 1, 2). The localization process for each USV segment took approximately 0.1 s using a consumer personal computer (PC), which is more than 1000 times faster than the computation time (6 min/segment) reported in a previous distributed-microphone system (Warren et al., 2018).

USVCAM assigns a USV segment in two steps (Supplementary Fig. 4). Firstly, as an

initial screening, mice that have peak-snout distances that are longer than the threshold are excluded from the candidate mice emitting the USV segment (Supplementary Fig. 4a). The threshold is set as the 99th percentile of the error histogram obtained from the single mouse experiment (Fig. 1e) so that the frequency of erroneously excluding the correct source mouse is expected to be < 1%. Secondly, the average power at the snout is compared between the candidate mice (Supplementary Fig. 4b). The USV segment is then assigned to the mouse with the highest power if the assignment confidence is higher than 99%. Confidence was calculated as the averaged precision of assignment in similar conditions in simulations where 'virtual' mice were placed in random locations based on data of the single mouse experiments (Supplementary Figs. 5 and 6; see Methods for details).

We confirmed the assignment precision by tests using data of the single mouse experiments and adding one to three virtual mice in random locations within the cage (Supplementary Table 1). We also counted the assigned segments during social interaction experiments between two or three real mice (Supplementary Table 2). Despite the close interaction in the home cage, around half or more segments were assigned in most cases, which is comparable with the performance of distributed-microphone systems (Neunuebel et al., 2015; Warren et al., 2018). The assignment ratios in ICR mice were relatively smaller than those in B6 mice because snout locations were sometimes unavailable when the snouts were hidden under another mouse during interactions between ICR mice; moreover, the ICR mice interacted closely more frequently. Fig. 1f shows an example of an assignment during the interaction between a pair of female ICR mice. Most USV segments were assigned to one of the mice, except for when the snouts of the mice were very close (see also Supplementary Videos 3 and 4). The assigned

segments were finally integrated into syllables based on the gaps between segments (bars under the spectrogram in Fig. 1f and g) to analyze the vocal patterns. Because the USVs from different mice were sometimes crossed (red arrowhead in Fig. 1g), we modified the segmentation algorithm to separate segments at crossing points (Fig. 1g middle and Supplementary Fig. 2) to avoid misassignment and estimate correct vocal patterns even with the overlap (Fig. 1g bottom). We confirmed that the modified algorithm could separate the overlapping syllables better than could the original algorithm by testing in an artificial dataset created by overlaying randomly selected pairs of syllables recorded in the single mouse experiments (Supplementary Fig. 7).

To evaluate the effectiveness of USVCAM for characterizing the social behavior of mice, we analyzed paired interactions under the resident-intruder (R-I) paradigm of each mouse strain (Fig. 2). All mice were tested with both a female (vs-F) and a male (vs-M) as both resident (R) and intruder (I). Fig. 2a shows the mean call rates (number of assigned syllables/min) during each type of session (see Supplementary Table 3 for statistical results). Interestingly, ICR females exhibited USVs even when they interacted with a female as an intruder and when they interacted with a male as a resident. However, it has previously been reported that in other strains the primary sender of vocalizations is the resident and the male during interactions between the same and different sex, respectively, in experiments using devocalization or anesthetization (White et al., 1998; Holy and Guo, 2005; Hammerschmidt et al., 2012). Furthermore, we compared the call rates during different actions of the self and other in the female ICR group, with which vocal interactions were most frequently observed (Fig. 2b and Supplementary Table 4). This result indicates that although the call rate changed depending on the ongoing action, as reported previously (Sangiamo et al., 2020), the call rate was also modulated by the social

context, even during the same types of actions. Finally, we analyzed acoustic features of USVs using dimension reduction by a variational autoencoder (unsupervised learning methods; Goffinet et al., 2021) to determine whether the distribution of acoustic features (i.e., the vocal repertoire) in ICR mice changes depending on social contexts (Fig. 2c and Supplementary Figs. 8 and 9). Results revealed no significant changes in the vocal repertoire depending on social contexts, although the vocal repertoire was dependent on individuals, as previously reported (e.g., Goffinet et al., 2021). To the best of our knowledge, this is the first report conducting a systemic analysis of complex acoustic features of syllables, including those that overlap. This was made possible by effectively combining sound localization technology with the USV segmentation algorithm modified for overlapping USVs (Fig. 1g). Taken together, the results of the USVCAM application revealed a novel characteristic of vocal communication in ICR mice, which demonstrated the effectiveness of USVCAM for characterizing social behavior in various mice strains, such as genetically modified mice used as disease models.

In summary, we validated our novel system using a phased microphone array, named USVCAM, and demonstrated its effectiveness in the application experiment. USVCAM allowed, for the first time, recording and sound localization in home cages, where social behaviors are often observed. The compact acoustic camera design and rapid processing will benefit various experimental conditions, including home cages, and facilitate future studies investigating normal and abnormal vocal communication, social behaviors, and the underlying molecular, genetic, and neural mechanisms.

Methods:

Recording

The sensor assembly (Fig. 1b right) consisted of four ultrasound microphones (TYPE 4158N, ACO, Tokyo, Japan), a video camera (RealSense L515, Intel, CA, USA), and a custom three-dimensional (3D)-printed holder for the sensors. A square phased microphone array (around 8 mm on each side) was composed with the holder. The distance between the camera and the center of the microphone array was approximately 54 × 9 × 6.5 (depth) mm (microphones were placed in front of the camera). The audio data was captured by each microphone, amplified with a four-channel microphone amplifier (BSA-CCPMA4-UT20, Katou Acoustics Consultant Office, Kanagawa, Japan), and sampled at 384 kHz using an analog-digital converter (PCIe-6374, National Instruments, TX, USA). An infrared image (resolution: 640 × 480 pixels; field of view: 70° × 55°) was captured with the camera at 30 Hz. The audio and video data were stored on the same PC (Elitedesk 800 G5 TW, Hewlett-Packard Inc., CA, USA) with timestamps of each data for audio-video synchronization, using a custom recording software written in Python (Python Software Foundation, NC, USA).

Reducing sound reflection in recording environments is crucial for ensuring accurate sound source localization based on the time/phase lags (Fig. 1b). To this end, recordings were performed in a soundproof box (63 × 53 × 84 [height] cm) with 20 mm thickness sound-absorbing melamine foam on the walls and ceiling (Supplementary Fig. 9). We also designed a custom inner cage for home cage recording (Fig. 1d and Supplementary Fig. 11). The inner cage was made from fine stainless mesh (mesh size: 150 mesh/inch). Clean disposable paper towels (Prowipe, Daio Paper Corp, Tokyo, Japan) were inserted into the space (approximately 5 mm) between the inner cage and the plastic

home cage to suppress the influence of sound reflection (see Supplementary Fig. 1 for the effect of the inner cage).

USV segmentation

For the segmentation (detection) of the USVs in the audio data, the USVSEG algorithm (Tachibana et al., 2020) was used with slight modifications. The USVSEG algorithm can robustly segment USVs from background noise by generating a stable spectrogram using the multitaper method and flattening the spectrogram by liftering in the cepstral domain. Although several different algorithms have been proposed for USV segmentation, a recent benchmarking study indicated that the USVSEG is comparable to the state-of-art method (Fonseca et al., 2021). We modified the USVSEG algorithm to separate crossing USVs emitted from different animals (Fig. 1g and Supplementary Fig. 2). First, a binary image of the spectrogram peaks was created (pixel size = 0.5 ms × 750 Hz) using the peaks obtained using the original USVSEG algorithm (Supplementary Fig. 2c). Second, the binary image was dilated twice and eroded once with a 3 × 3 square structuring element to connect the spatially neighboring components using imdilate() and imerode() functions in MATLAB (Mathworks, MA, USA; Supplementary Fig. 2d). Third, the corner points (i.e., the crossing points and edges) were detected using the corner() function in MATLAB (Supplementary Fig. 2e), and the pixels within the rectangles (3 × 15 pixels) centered on the corners were erased to cut segments at the crossing points. Finally, the boundaries of the subsegments were calculated by applying the watershed transform to the image using the watershed() function in MATLAB (Supplementary Fig. 2f), and the spectral peaks were grouped according to the boundaries (Supplementary Fig. 2g). Small segments with ≤ 3.0 ms were excluded from the subsequent analysis. We

validated the algorithm with synthetic data, in which pairs of syllables recorded from a single mouse were overlapped (Supplementary Fig. 7). For test data generation, a pair of syllables recorded in a single mouse experiment were randomly selected, and the peaks of each syllable were detected. The peaks of the two sources were overlayed with a random time shift (± 20 ms). When the peaks from the different sources were close, one stronger peak was selected according to the USVSEG algorithm, and only one peak was selected within a narrow bandwidth. In total, 10,000 (5,000 from B6 and 5,000 from ICR mice data) overlapping syllables were generated, and the segmentation algorithms were applied. To test the effect of the modification of the segmentation algorithms, we compared the results between using the proposed algorithm and the algorithm without corner detection. To quantify the quality of the segmentation, the contamination ratio was defined as $(1/N)\times\sum n_i$, where n_i and N represent the number of contaminated points (i.e., the points from the minor source) in the i-th segment and the total number of points, respectively.

USV localization

Using the conventional (Bartlett, or delay-and-sum) beamformer, the power (P) of sound arriving from a given spatial location ($\boldsymbol{r}$) was calculated as follows (Krim and Viberg, 1996):

$$P(\boldsymbol{r},\omega,t) = |\boldsymbol{w}(\boldsymbol{r},\omega)^H \boldsymbol{x}(\omega,t)|^2$$

$$\boldsymbol{x}(\omega,t) = [X_1(\omega,t),\dots,X_m(\omega,t)]^T$$

$$\boldsymbol{w}(\boldsymbol{r},\omega) = \frac{\boldsymbol{a}(\boldsymbol{r},\omega)}{\sqrt{\boldsymbol{a}(\boldsymbol{r},\omega)^H \boldsymbol{a}(\boldsymbol{r},\omega)}}$$

$$\boldsymbol{a}(\boldsymbol{r},\omega) = \left[e^{-i\omega\tau_1(\boldsymbol{r})},\dots,e^{-i\omega\tau_m(\boldsymbol{r})}\right]^T$$

$$\tau_j(\boldsymbol{r}) = |\boldsymbol{r} - \boldsymbol{r}_j|/c$$

where ω and t are frequency and time indices, respectively. $X_j(\omega,t)$ represents the short-term Fourier transform of the signal captured with the j-th ($j = 1, \ldots, m$) microphone. $\boldsymbol{r}_j$ is the position of the j-th microphone, c is the speed of sound, and τ_j is the expected time delay of the signal arriving at the j-th microphone from the sound source location $\boldsymbol{r}$. T and H denote the transposition and conjugate transposition, respectively. The beamformer shifts the signals captured by the microphones to compensate for the arrival time delays (i.e., the phase lags of the signals) using a steering vector $\boldsymbol{a}$, and sums the shifted signals to calculate the power (P). Thus, the function $P(\boldsymbol{r})$ is expected to be the maximum when the location $\boldsymbol{r}$ overlaps with the actual sound source and is called the spatial spectrum of the sound. Because we are not interested in the absolute power of the signal but rather in the peak locations in the spatial spectrum for sound localization, the following normalized spatial spectrum P_{norm} was used:

$$P_{norm}(\boldsymbol{r}, \omega, t) = \left| \frac{\boldsymbol{w}(\boldsymbol{r}, \omega)^H \boldsymbol{x}(\omega, t)}{|\boldsymbol{x}(\omega, t)|} \right|^2$$

The average spatial spectrum of a given segment P_{seg} (Fig. 1d) was subsequently defined as follows:

$$P_{seg}(\boldsymbol{r}) = \frac{1}{n} \sum_{k=1}^{n} P_{norm}(\boldsymbol{r},\ \omega_k, t_k)$$

where ω_k and t_k represent frequency and time indices, respectively, of the k-th ($k=1, \ldots, n$) peak of a USV segment in the spectrogram. To localize the USV segment, P_{seg} was calculated at each x-y location of the camera image. The depth (z) of the sound source locations were assumed to be constant (i.e., equal to the floor of the cage). To reduce computational load, the image was binned into 5 × 5-pixel bins, and P_{seg} were calculated for each bin. P_{seg} outside the image were also calculated for margins of 100 pixels on each

side to correctly estimate the peaks of the spatial spectrum around the edge of the camera image. The resultant spatial spectrum often showed multiple peaks in the form of a square grid (Supplementary Fig. 12) owing to the periodicity of the sound wave and the square microphone array arrangement; this phenomenon is referred to as spatial aliasing. To avoid spatial aliasing affecting the result, we positioned the microphones as close as possible (Fig. 1b) to maximize the distance between the multiple peaks in the spatial spectrum. Furthermore, we used a USV assignment algorithm that can deal with spatial aliasing (see below). The spatial spectrum was normalized with the z-score normalization to evaluate the saliency of the peaks. Peak locations (local maximums) in the spatial spectrum were searched as bins with equal values across the original spectrum and that filtered using the maximum filter (window size = 5 × 5 bins). A peak with a small height ($z < 1.6$, approximately < 95% in the cumulative distribution function of the standard normal distribution) was excluded from the sound source candidates. A USV segment without any clear peak ($z < 2.3$, approximately < 99% in the cumulative distribution function of the standard normal distribution) in the spatial spectrum was categorized as 'unlocalized' and was excluded from subsequent analysis. The snout-peak distance was defined as the distance from the snout of the mouse and the nearest peak in the spatial spectrum.

USV assignment

Supplementary Fig. 4 shows an overview of the algorithm for USV assignment for each USV segment. Initially, the snout-peak distance was calculated for each mouse, and the mice with snout-peak distances that were within the distance threshold (red dotted line) were selected as candidates of the sound source (Supplementary Fig. 4a). The

268 distance threshold was set as 99th percentiles of the distribution of the snout-peak
269 distances in the single mouse experiments (Fig. 1e and Supplementary Fig. 4a right). If
270 the number of candidates was zero or one, the USV segment was categorized as
271 'unassigned' or assigned to the candidate mouse, respectively. If there was more than one
272 candidate mouse, the following test was conducted among the candidates.

273 To assign the USV segment to one of the candidate mice following the screening
274 procedure above, the average power at the snout of the candidates was first calculated
275 (Supplementary Fig. 4b). The two highest powers among the candidates were compared
276 using a two-tailed Wilcoxon signed-rank test. The resulting p-value (p) was used for the
277 assignment. In addition, we used the distance between the snouts of the best two mice (d)
278 for the assignment because we found using a simulation that, with the same p-value
279 threshold, when the snouts of the two mice become closer, more assignment errors
280 occurred (Supplementary Fig. 5). This may have been caused by an error in sound
281 localization itself (Fig. 1e) or an error in the video-based estimation of snout locations.
282 Using the above two parameters p and d with the number of candidate mice (N_c), the
283 confidence for assigning the segment to the best mouse was calculated as the averaged
284 precision of the assignment for similar conditions in the following simulation using the
285 single mouse experiment data. In the simulation, N_c-1 'virtual' mice were assumed to be
286 at random locations away from the real mouse at a distance d, and the precision (hit
287 count/[hit + error counts]) of the assignment at the p-value threshold p was estimated.
288 The simulations were performed in advance for all possible combinations of p, d, and N_c
289 (Supplementary Fig. 5), and the distribution of the precision was used as a lookup table
290 for the confidence estimation to reduce the computational load. USV segments with a
291 confidence level of > 0.99 were assigned to the best mice, and the others were categorized

292 as ‘unassigned.’

293 The confidence estimation may be inaccurate when the two best mice are located
294 near two different peaks in the spatial spectrum because of spatial aliasing
295 (Supplementary Figs. 12 and 13a). In such cases, although the absolute snout-snout
296 distance (*d*) is large, both mice should be considered good candidates. To ameliorate the
297 problem, we calculated the distance between snouts (*d’*) after converting the snout
298 positions into their relative positions from the nearest peaks (i.e., ‘wrapping’ the positions
299 in a period of the spatial spectrum; Supplementary Fig. 13b) and used *d’* for the
300 confidence estimation.

301

302 *Merging segments into syllables*

303 Rodent USVs consist of syllables, which have tens to hundreds of milliseconds
304 durations with gap intervals (usually of > 30 ms). Previous studies categorizing syllable
305 patterns reported that the patterns differed depending on the behavior, individual, and
306 strain (Holy and Guo, 2005; Matsumoto and Okanoya, 2018; Goffinet et al., 2021). Thus,
307 we merged the short segments (Fig. 1g) into syllables after the assignment. Specifically,
308 segments assigned to one mouse and ‘unassigned’ segments (if they existed) with gap
309 intervals smaller than a given threshold (a minimum gap of 30 ms) were merged into a
310 single syllable. The assignment rate was defined as the ratio of the time-frequency points
311 assigned to the mouse to all the points in the syllable. We only used syllables with an
312 assignment rate of 1.0 for the behavior analysis.

313

314 *Audible broadband vocalization detection*

315 Mice occasionally emit audible broadband vocalizations (BBVs; i.e., ‘squeaks’)

316 during conspecific interactions (Finton et al., 2017). BBVs are loud broadband sounds
317 characterized by a harmonic structure. To prevent misidentifying BBVs as USVs and
318 misassigning USVs that significantly overlap with BBVs, we excluded the ultrasound
319 segments that overlapped with the time intervals of BBVs from the analysis. BBVs were
320 detected automatically using the following simple algorithm. First, the recorded sound
321 was downsampled to 38.4 kHz. Second, the spectrogram (time window = 10 ms;
322 frequency range = 2 to 16 kHz) of the sound was calculated. Third, continuous
323 background noise and transient (impulse-like) broadband noise were reduced by
324 subtracting the median value of each frequency bin and the median value of each time
325 bin, respectively. Fourth, the spectrogram was filtered using a median filter (window size
326 = 0.5 kHz) along the frequency axis. Finally, the maximum power in the spectrogram
327 across the frequency was calculated for each time point, and the time intervals containing
328 BBVs were estimated as the intervals in which the maximum power exceeded a certain
329 threshold value (we used 28 dB in this study). To check the precision of the simple BBV
330 detector, we selected three recording sessions that involved a relatively large number of
331 BBVs (interactions between one pair of male ICR mice and two male-female ICR mouse
332 pairs; see below for details of the recording experiment), and a blinded experimenter
333 compared the results of the automatic detection with the manual annotations
334 (Supplementary Fig. 14). A total of 161 BBVs were detected in the manual annotations.
335 Of these, 160 overlapped with the automatic detection, and one was missed by the detector.
336 The automatic detector had only 22 additional (false-positive) detections. Thus, the results
337 confirmed that the simple detector can effectively exclude time intervals containing BBVs
338 from the analysis. Supplementary Table 5 shows the number of ultrasound segments that
339 overlapped with the time intervals of BBVs in each of the social interaction experiments.

Calibrating microphone positions

For the above USV localization algorithm, microphone positions ($\boldsymbol{r}_j$) needed to be accurately calibrated. For the calibration, we searched the microphone positions that maximized the average power (P_{seg}) at the snout locations for the 20 selected USV syllables in the single mouse experiment, as follows:

$$\underset{\boldsymbol{r}_1,\ldots,\boldsymbol{r}_{20}}{\operatorname{argmax}} \sum_{l=1}^{20} P_{seg}(\boldsymbol{r}_l)$$

where $\boldsymbol{r}_l$ is the snout location at which the l-th selected syllable was emitted. The optimization was performed using the L-BFGS-B algorithm implemented in Scipy (Virtanen et al., 2020). The syllable that emitted different parts in the recording area was selected for appropriate calibration.

Video tracking

USV assignment requires frame-by-frame snout locations of each mouse. USVCAM users can choose any available high-precision video tracking software (such as DeepLabCut [Lauer et al., 2021], Social LEAP Estimates Animal Poses [SLEAP; Pereira et al., 2021], and Mouse Action Recognition System [MARS; Segalin et al., 2021]) to estimate snout locations. In this study, we used AlphaTracker (Chen et al., 2020) for tracking the locations of the snout and the other body parts. The software can relatively robustly track the locations of body parts of interacting mice using deep neural networks. We prepared labeled data to train the networks by manually annotating body parts (snout, tail-base, and left and right ears) and the bounding box of the mice in 1694 and 996 randomly selected frames from the videos of B6 and ICR mice, respectively. Different

363 networks were trained for tracking B6 and ICR mice. The outputs of AlphaTracker were
364 manually curated to correct occasional errors (e.g., switching mouse identities and
365 flipping snout and tail-base locations) using custom software written in Python. The
366 resultant trajectories of the snouts were filtered using a median filter (window size = 0.16
367 s) and used for USV assignment. The trajectories of the bounding box, tail-base, left and
368 right ears, and snout were filtered using a locally estimated scatterplot smoothing
369 (LOESS) filter (time window = 0.5 s) and used for behavioral event classifications (see
370 below). Supplementary Video 5 shows an example of the filtered trajectories used for
371 behavioral event classification.
372
373 *Audible sound generation from recorded USVs*
374 For intuitive data presentation, the audible sound was created according to the
375 results of the USV segmentation and integrated with the videos (Supplementary Videos
376 1–4) using a sound synthesis method proposed by Carruthers et al. (2013). First, the
377 maximum peak of the USV segments at each time point was extracted. Then, the sound
378 *x(t)* was generated as:
379 $$x(t) = a(t)\sin\left[2\pi\int_0^t f(\tau)d\tau\right]$$
380 where *t* is time, and *a(t)* and *f(t)* are the amplitude and frequency of the peak, respectively.
381 Thus, the generated sound contains no background noise. The frequency was linearly
382 mapped from 0 to 192 kHz to 1 to 6 kHz to make the sound audible. The generated sounds
383 were only used for visualization.
384
385 *Animals and experimental schedule*
386 All animal experiments were performed with institutional ethics approval (from

387 the Institutional Animal Use Committee of Kagoshima University #L21007). We used 40
388 adult mice: 10 males and 10 females each of the C57BL/6J (B6) and ICR strains. Mice
389 were purchased from Japan SLC (Shizuoka, Japan) at the age of 8 weeks and housed
390 alone in a cage (182 × 260 × 128 mm, CREA Japan, Tokyo, Japan) equipped with the
391 custom inner cage for 1 week. Food (5L37 Rodent LabDiet EQ, PMI Nutrition
392 International, MO, USA) and water were supplied *ad libitum*, and the animals were kept
393 under a standard 12 h:12 h light-dark cycle. Soft paper chips were used for bedding (Japan
394 SLC). Experiments were mainly conducted in the light phase, except for some of the
395 recordings for system validation (the single mouse and three mice experiments, see
396 below), which were conducted during the dark phase. The environment was maintained
397 at a constant temperature (22–25°C) and humidity (50% ± 5%).

398 After a 1 week habituation period, the social interactions between mice in the
399 home cages were recorded. On the first day, each mouse was allowed to interact with
400 another mouse of the same sex and strain (M-M and F-F contexts), both as resident and
401 intruder (nine mice as residents and one mouse as an intruder were tested first). On the
402 second day, each mouse was allowed to interact with another mouse of a different sex but
403 the same strain (M-F and F-M contexts). In these cases, all individuals were used as both
404 a resident and an intruder, and the order in which these roles were applied to the
405 experiment was counterbalanced. For the recording, the home cage with a resident mouse
406 was placed in the soundproof recording box, an intruder mouse was placed in the home
407 cage, and the behaviors of the mice were recorded for approximately 3 min. After the 2
408 days of paired social interaction recording, one mouse of each strain was recorded in a
409 single mouse condition to obtain data for system validation and determine parameters for
410 the USV assignment. In the single mouse experiment, a male mouse was placed in an

empty home cage of a female mouse, and its vocalizations were recorded for approximately 10 min. We also tested the recording of a three-mouse interaction of each strain for approximately 5 min for system validation (Supplementary Table 2).

Data analysis

The number of syllables assigned completely (assignment rate = 1.0) was counted for each mouse for each recording. The call rate was calculated by dividing the syllable count by the recording duration. In addition, to check the relationship between USVs and specific actions, we defined the following five behavior events based on video tracking results: *approach*, *leave*, and *contact with the tail base*, *trunk*, and *snout* (Supplementary Fig. 15). Then, call rates during different behavior events were calculated separately. The call rates during contact with the snout were not analyzed because of difficulty in USV assignment when the snouts were very close. To quantify and compare the acoustic features (patterns) of syllables, we used a variational autoencoder (VAE), according to the method proposed by Goffinet et al. (2021). In this method, VAE learns to map single-syllable spectrogram images onto 32 latent features in an unsupervised manner. We reconstructed a spectrogram of an assigned syllable using the frequency and amplitude of each point in the segments of the syllable to enable the analysis of the acoustic feature of the syllables even when it temporally overlapped with the syllables emitted from the other mouse. In total, 7960 single-syllable spectrogram images were reconstructed and used for VAE training and analysis. The distribution of the syllables in latent space was visualized using Uniform Manifold Approximation and Projection (UMAP) and the difference in the distributions (vocal repertoires) between a pair of different experimental conditions was quantified using maximum mean discrepancy

(MMD), according to a previous study (Goffinet et al., 2021). Distributions with fewer than 10 syllables were excluded from the MMD analysis. Statistical tests were performed using R (The R Foundation, IN, USA) and MATLAB.

References

1. Asaba A, Okabe S, Nagasawa M, Kato M, Koshida N, Osakada T, Mogi K, Kikusui T. Developmental social environment imprints female preference for male song in mice. PLoS One. 2014 Feb 5;9(2):e87186. doi: 10.1371/journal.pone.0087186.
2. Carruthers IM, Natan RG, Geffen MN. Encoding of ultrasonic vocalizations in the auditory cortex. J Neurophysiol. 2013;109(7):1912-1927. doi:10.1152/jn.00483.2012
3. Chen Z, Zhang R, Zhang YE, Fang HS, Zhou H, R RR, Bal A, Padilla-Coreano N, Keyes L, Tye KM, Lu C. AlphaTracker: A Multi-Animal Tracking and Behavioral Analysis Tool. Preprint at https://www.biorxiv.org/content/10.1101/2020.12.04.405159v1 (2020)
4. Finton CJ, Keesom SM, Hood KE, Hurley LM. What's in a squeak? Female vocal signals predict the sexual behaviour of male house mice during courtship. Animal Behaviour. 2017 Apr;126:163-175.
5. Fischer J, Hammerschmidt K. Ultrasonic vocalizations in mouse models for speech and socio-cognitive disorders: insights into the evolution of vocal communication. Genes Brain Behav. 2011 Feb;10(1):17-27. doi: 10.1111/j.1601-183X.2010.00610.x.
6. Fonseca AH, Santana GM, Bosque Ortiz GM, Bampi S, Dietrich MO. Analysis of ultrasonic vocalizations from mice using computer vision and machine learning. Elife. 2021 Mar 31;10:e59161. doi: 10.7554/eLife.59161.
7. Goffinet J, Brudner S, Mooney R, Pearson J. Low-dimensional learned feature spaces quantify individual and group differences in vocal repertoires. Elife. 2021 May 14;10:e67855. doi: 10.7554/eLife.67855.
8. Golden SA, Aleyasin H, Heins R, Flanigan M, Heshmati M, Takahashi A, Russo SJ,

464 Shaham Y. Persistent conditioned place preference to aggression experience in adult
465 male sexually-experienced CD-1 mice. Genes Brain Behav. 2017 Jan;16(1):44-55.
466 doi: 10.1111/gbb.12310.
467 9. Hammerschmidt K, Radyushkin K, Ehrenreich H, Fischer J. The structure and usage
468 of female and male mouse ultrasonic vocalizations reveal only minor differences.
469 PLoS One. 2012;7(7):e41133. doi: 10.1371/journal.pone.0041133.
470 10. Heckman JJ, Proville R, Heckman GJ, Azarfar A, Celikel T, Englitz B. High-
471 precision spatial localization of mouse vocalizations during social interaction. Sci
472 Rep. 2017 Jun 7;7(1):3017. doi: 10.1038/s41598-017-02954-z.
473 11. Holy TE, Guo Z. Ultrasonic songs of male mice. PLoS Biol. 2005 Dec;3(12):e386.
474 doi: 10.1371/journal.pbio.0030386.
475 12. Kikusui T. Analysis of male aggressive and sexual behavior in mice. Methods Mol
476 Biol. 2013;1068:307-18.
477 13. Konopka G, Roberts TF. Animal Models of Speech and Vocal Communication
478 Deficits Associated With Psychiatric Disorders. Biol Psychiatry. 2016 Jan
479 1;79(1):53-61. doi: 10.1016/j.biopsych.2015.07.001.
480 14. Krim H, Viberg M. Two decades of array signal processing research: the parametric
481 approach. IEEE Signal Processing Magazine. 1996 July;13(4):67-94. doi:
482 10.1109/79.526899.
483 15. Lahvis GP, Alleva E, Scattoni ML. Translating mouse vocalizations: prosody and
484 frequency modulation. Genes Brain Behav. 2011 Feb;10(1):4-16. doi:
485 10.1111/j.1601-183X.2010.00603.x.
486 16. Lauer J, Zhou M, Ye S, Menegas W, Nath T, Rahman MM, Santo VD, Soberanes D,
487 Feng G, Murthy VN, Lauder G, Dulac C, Mathis MW, Mathis A. Multi-animal pose

488 estimation and tracking with DeepLabCut. Preprint at
489 https://www.biorxiv.org/content/10.1101/2021.04.30.442096v1 (2021).
490 17. Matsumoto YK, Okanoya K. Mice modulate ultrasonic calling bouts according to
491 sociosexual context. R Soc Open Sci. 2018 Jun 20;5(6):180378. doi:
492 10.1098/rsos.180378. PMID: 30110406; PMCID: PMC6030292.
493 18. Neunuebel JP, Taylor AL, Arthur BJ, Egnor SE. Female mice ultrasonically interact
494 with males during courtship displays. Elife. 2015 May 28;4:e06203. doi:
495 10.7554/eLife.06203.
496 19. Pereira TD, Tabris N, Li J, Ravindranath S, Papadoyannis ES, Wang ZY, Turner DM,
497 McKenzie-Smith G, Kocher SD, Falkner AL, Shaevitz JW, Murthy M. SLEAP:
498 Multi-animal pose tracking. Preprint at
499 https://www.biorxiv.org/content/10.1101/2020.08.31.276246v1 (2020)
500 20. Sales GD. Ultrasonic calls of wild and wild-type rodents. In: Handbook of
501 Mammalian Vocalization: An Integrative Neuroscience Approach (Brudzynski SM,
502 ed), pp 77–88. Elsevier, 2010.
503 21. Sangiamo DT, Warren MR, Neunuebel JP. Ultrasonic signals associated with
504 different types of social behavior of mice. Nat Neurosci. 2020 Mar;23(3):411-422.
505 doi: 10.1038/s41593-020-0584-z. Epub 2020 Feb 17.
506 22. Segalin C, Williams J, Karigo T, Hui M, Zelikowsky M, Sun JJ, Perona P, Anderson
507 DJ, Kennedy A. The Mouse Action Recognition System (MARS) software pipeline
508 for automated analysis of social behaviors in mice. Elife. 2021 Nov 30;10:e63720.
509 23. Tachibana RO, Kanno K, Okabe S, Kobayasi KI, Okanoya K. USVSEG: A robust
510 method for segmentation of ultrasonic vocalizations in rodents. PLoS One. 2020 Feb
511 10;15(2):e0228907. doi: 10.1371/journal.pone.0228907.

24. Warren MR, Sangiamo DT, Neunuebel JP. High channel count microphone array accurately and precisely localizes ultrasonic signals from freely-moving mice. J Neurosci Methods. 2018 Mar 1;297:44-60. doi: 10.1016/j.jneumeth.2017.12.013. Epub 2018 Jan 5.

25. White NR, Prasad M, Barfield RJ, Nyby JG. 40- and 70-kHz vocalizations of mice (Mus musculus) during copulation. Physiol Behav. 1998 Feb 15;63(4):467-73. doi: 10.1016/s0031-9384(97)00484-8.

26. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat İ, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P; SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020 Mar;17(3):261-272. doi: 10.1038/s41592-019-0686-2.

Data availability:

The datasets generated and/or analyzed for the current study are available from the corresponding author on reasonable request. Part of the raw data is available in a public repository as example data to test the codes (doi:10.6084/m9.figshare.17121275).

Code availability:

The codes of the USVCAM system (for recording, localization, and assignment) are available in a public repository (https://github.com/MatsumotoJ/usvcam). Other codes used for further analysis are available from the corresponding authors on reasonable request.

Acknowledgments:

We would like to thank Ryosuke Tachibana for providing helpful explanations of the USVSEG algorithm, and Takafumi Tamura and the staff at Engineering Machine Shop, University of Toyama, for technical assistance for manufacturing the mesh inner cage and the 3D-printed holder.

This work was supported by Grant-in-Aid for Scientific Research from Japan Society for the Promotion of Science (nos. 16H06534, 21K06438, 19H04912), the Firstbank of Toyama Scholarship Foundation Research Grant, Tokyo Biochemical Research Foundation, Takeda Science Foundation, Grant for Basic Science Research Projects from The Sumitomo Foundation, and RIKEN Collaborative Technical Development in Data-driven Brain Science.

Author contributions:

J.M. developed the USVCAM system. T.Sb. advised on algorithm selection. M.K. advised on the design of the microphone array. K.K. collected experimental data. J.M., K.K., C.C., T.Sg., H.Nm., and H.Nj. analyzed data. J.M., K.K., H.Nm., H.Nj., T.Sb., and T.Sg. wrote the manuscript. All authors contributed to the article and approved the submitted version.

Competing interests statement:

M.K. developed and sells the custom microphone amplifier used in this study.

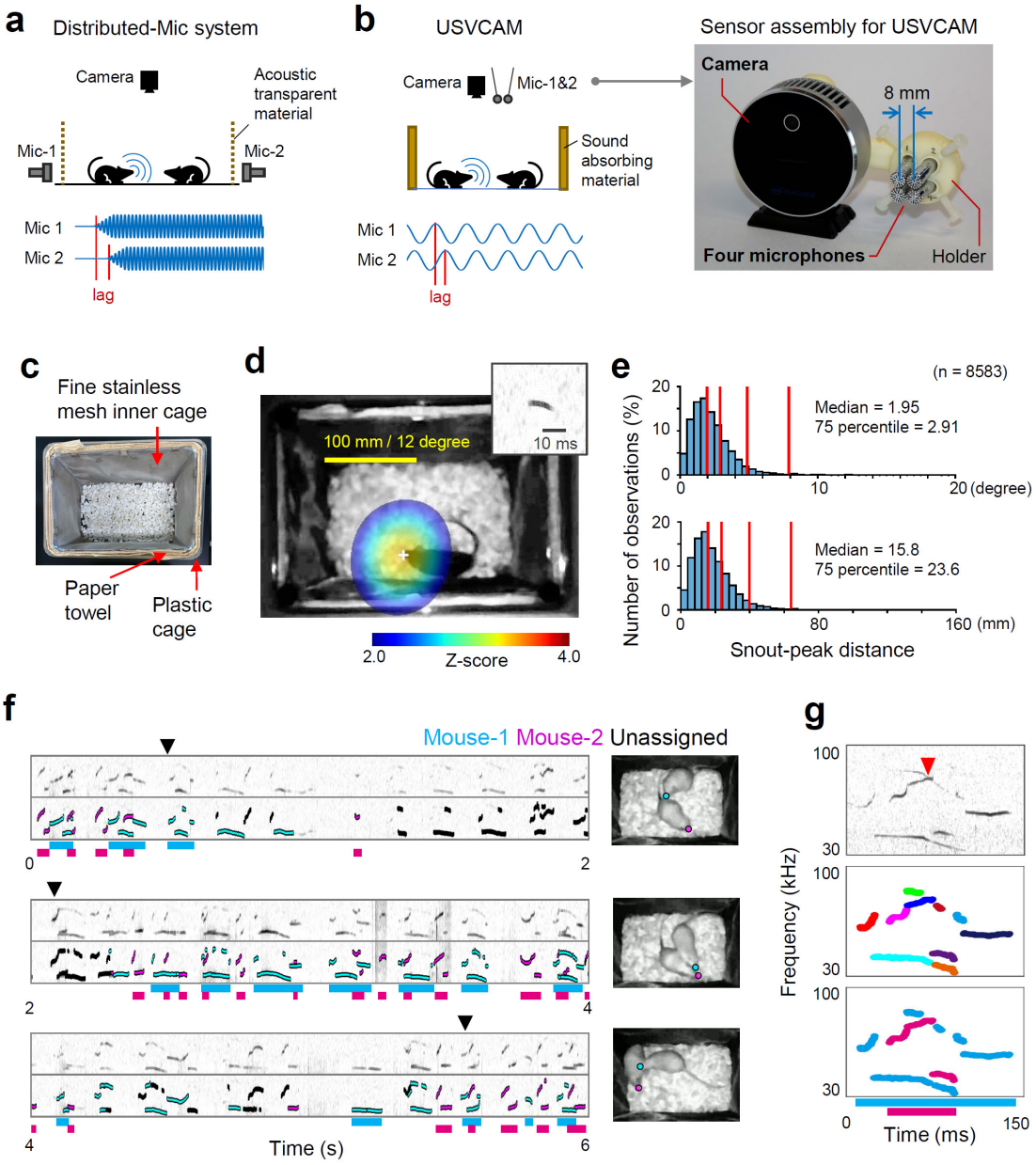
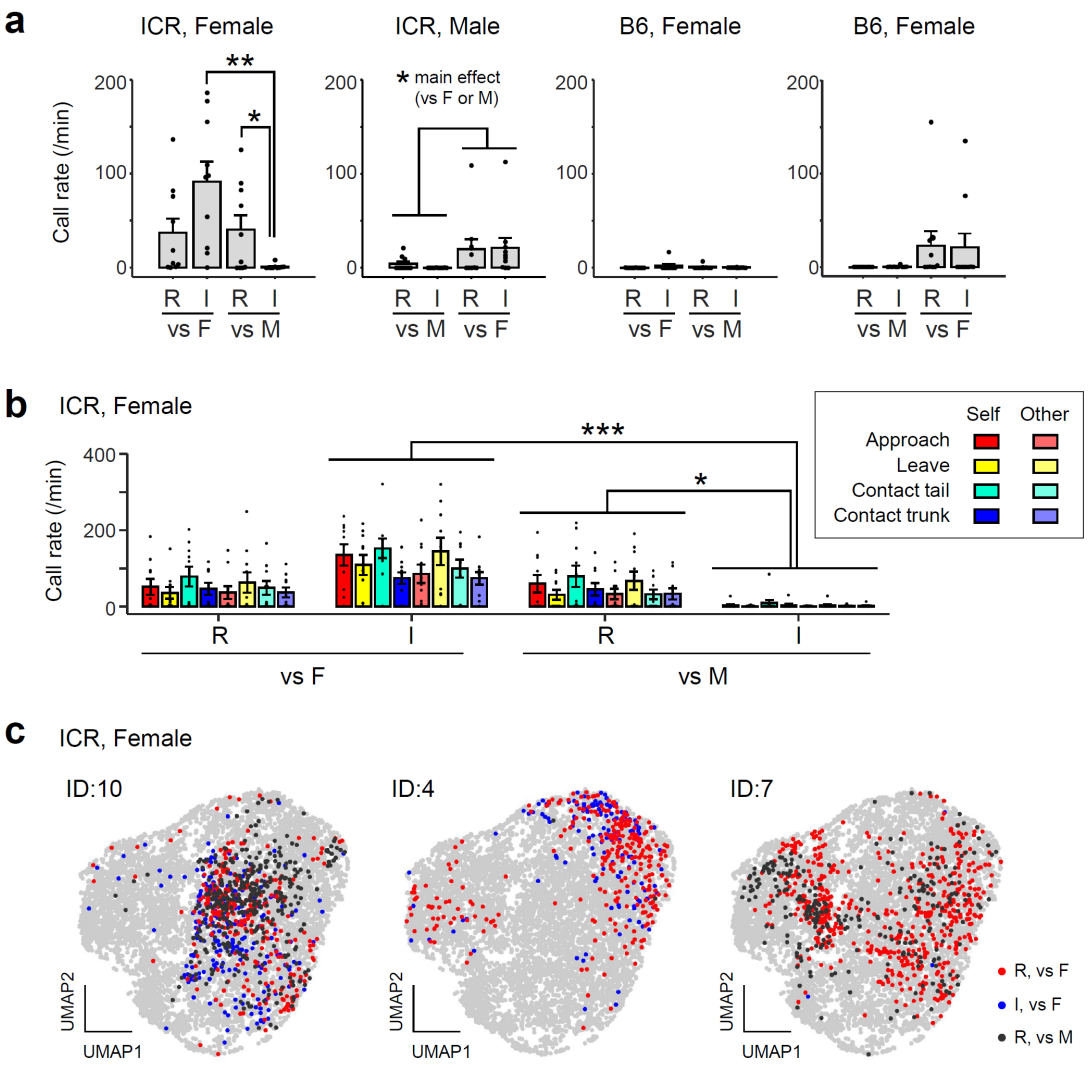


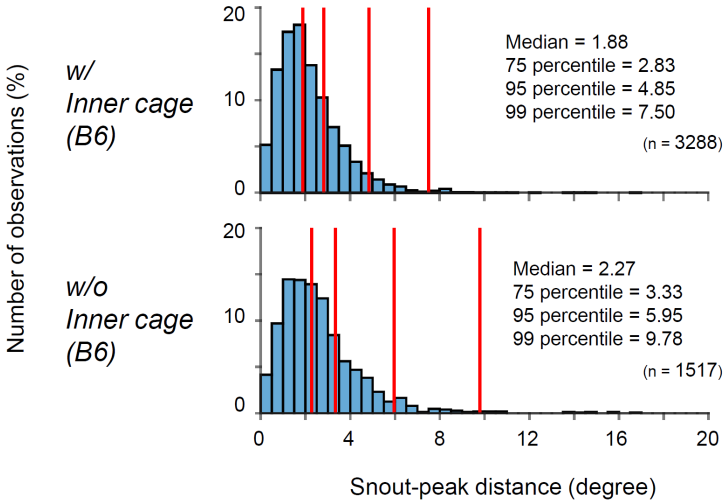
Figure 1. USV localization and assignment using USVCAM.
(a) A schematic showing the setup (top) and the signals received from the microphones (bottom) using the distributed-microphone (Mic) system. Only two microphones are shown for simplicity. Because the resolution of sound localization depends on the time lags of the sound arrival, microphones are located on the sides of the recording chamber, with acoustic transparent walls to maximize the lags. **(b)** A similar schematic (left) and a picture of the sensor assembly (right) of the USVCAM. Because USVCAM utilizes phase lags of sound waves for sound localization, the microphone array can be set in one place. **(c)** A home cage equipped with the custom inner cage. **(d)** An example of sound

localization of a USV segment (inset). The white cross signifies the peak of the spatial spectrum. **(e)** Distributions of the localization errors are shown in degrees (top) and millimeters (bottom). Red vertical lines indicate 50th, 75th, 95th, and 99th percentiles of the distributions, respectively. **(f)** An example of a USV assignment. The original spectrograms and those overlayed with the assignment results are shown. Bars under the spectrograms indicate syllables assigned to each mouse. The video frames at the black arrows are shown on the right. The snout positions are labeled by colored circles. **(g)** An example of the segmentation and assignment of the overlapping USVs emitted from different mice. Top, spectrogram; middle, the segmentation result (different colors indicate different segments); bottom, the assignment result. The frequency (y-axis) ranges of all spectrograms in the figure are 30 to 100 kHz.

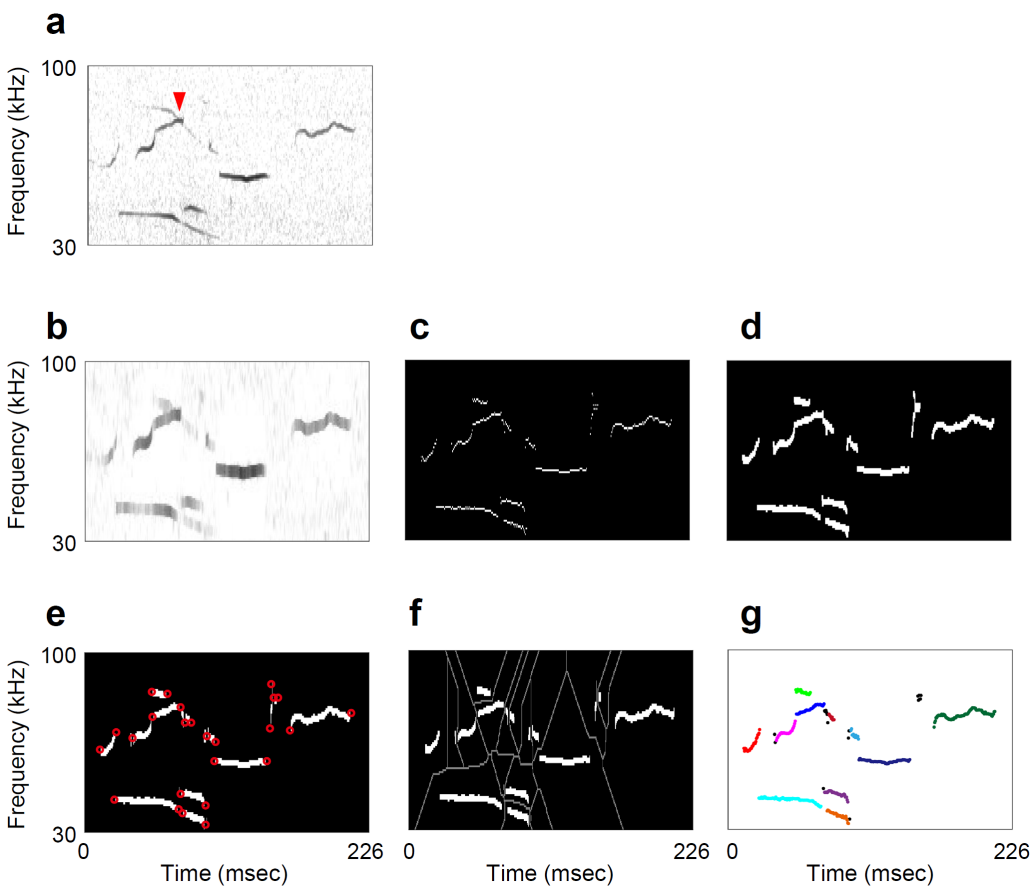


586 **Figure 2. Analysis of mouse pair interactions under the resident-intruder (R-I)**
587 **paradigm using USVCAM.**
588 (**a**) Comparisons of call rates (number of assigned syllables per minute) of subjects in
589 different social contexts. R, the subject was a resident; I, the subject was an intruder; vs
590 F, the partner was a female; vs M, the partner was a male. Each dot represents an
591 individual mouse. Error bars, standard error of the mean (s.e.m.); $^{**}p < 0.01$, $^{*}p < 0.05$,
592 simple main effects analysis. (**b**) Call rates of female ICR mice during different actions
593 by the subjects (self) and partners (other). See Supplementary Figure 15 for the definition
594 of the actions. Each dot represents an individual mouse. Error bars, s.e.m.; $^{***}p < 0.001$,
595 $^{*}p < 0.05$, simple main effects analysis. (**c**) UMAP projection of the acoustic features of
596 syllables extracted using the VAE (Goffinet et al., 2021). Examples of three ICR female
597 mice are shown (ID, identity number of the mouse). Each point represents an assigned

syllable. Red, blue, and black indicate the syllables of the subject recorded in different sessions. Gray points are all the other syllables recorded during the pair interaction experiments.

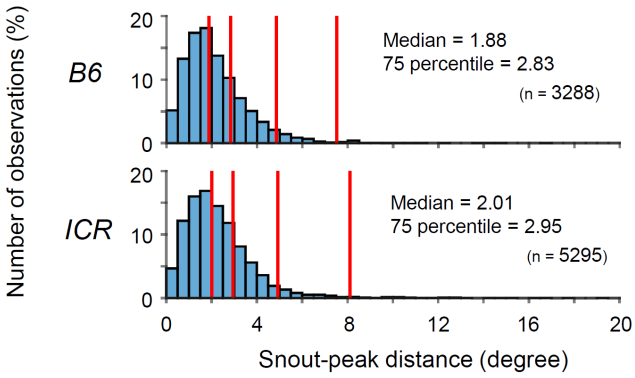


Supplementary Figure 1. Error histograms similar to those in Figure 1e, in the home cage with (top) and without (bottom) the inner cage and paper towels in a B6 mouse. Note that more outliers were observed without the inner cage (bottom), which demonstrates the effect of the inner cage.

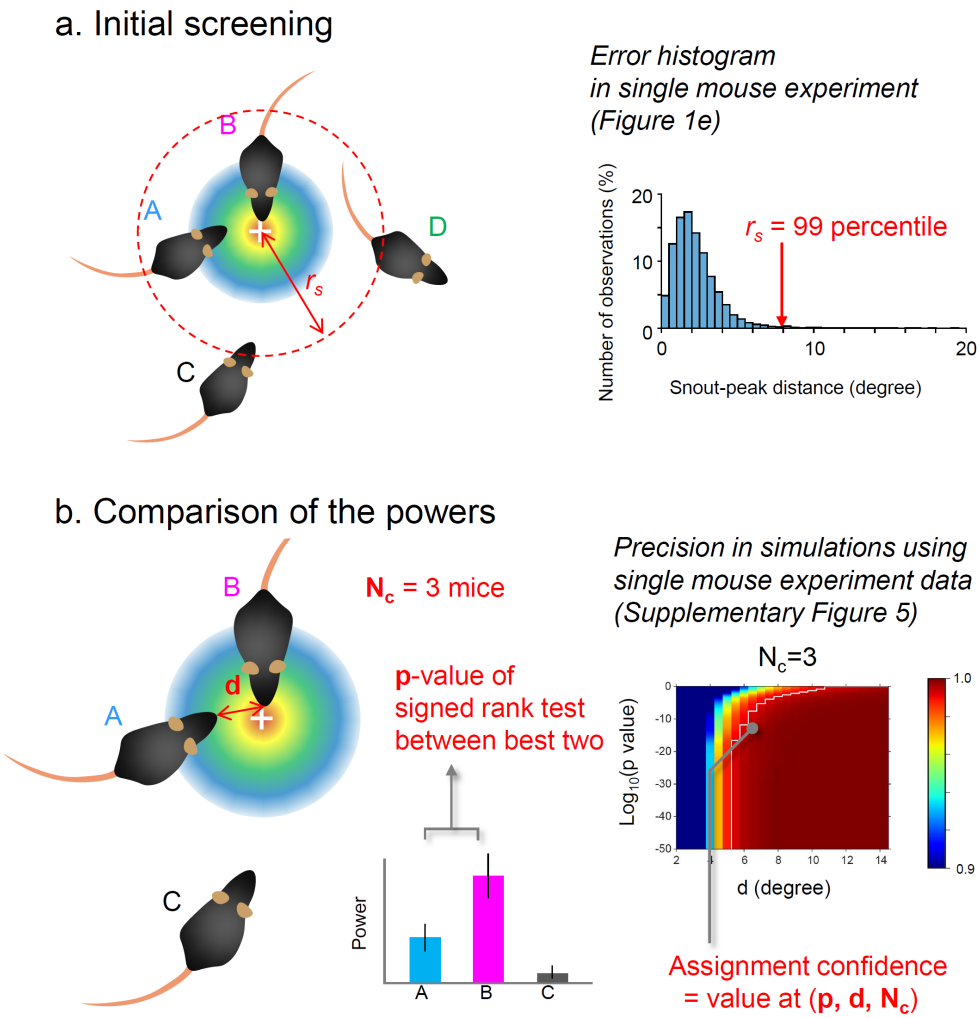


Supplementary Figure 2. USV segmentation.

(a) A spectrogram of example USVs. Note that in this example, calls emitted from two different mice crossed at the red arrow. **(b–g)** The segmentation process. First, a filtered spectrogram was calculated **(b)**, and peaks in the spectrogram were extracted **(c)** using the USVSEG algorithm. Then, morphological dilation and erosion were repeatedly applied to connect the continuous peaks **(d)**. To separate the segments at the crossing points, corners of the images were detected and removed (**e**; red circles indicate the detected corners). Finally, the watershed algorithm was applied to find the boundary of the segments (**f**; gray curves indicate the boundaries of the segments). The resultant segments are shown in **g**. The colors represent different segments. Segments that were too short (≤ 3 ms; colored black) were excluded from subsequent localization processes.

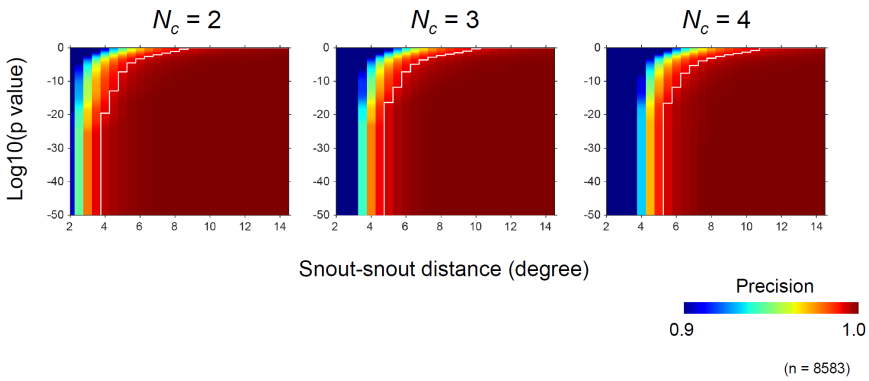


Supplementary Figure 3. Error histograms similar to those in Figure 1e, calculated separately for B6 (top) and ICR (bottom) mice.

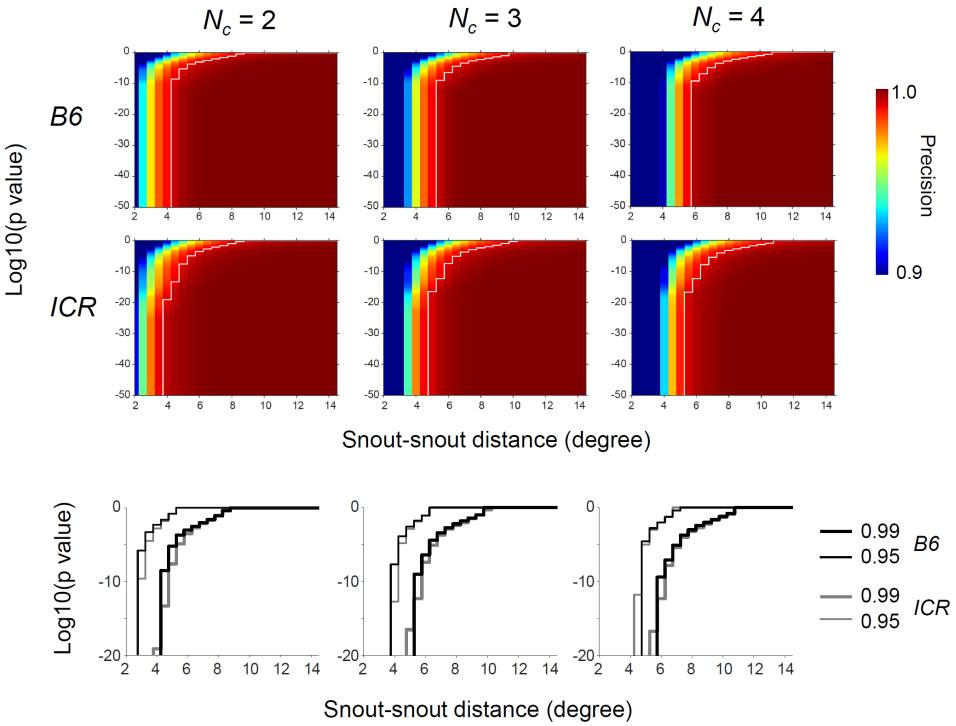


Supplementary Figure 4. USV assignment.

Assignment was performed in two steps: initial screening (**a**) and comparison of the sound powers at the snouts of mice (**b**). (**a**) First, mice with peak-snout distances that were longer than the threshold (r_s) were excluded from the candidates. A–D, IDs of mice. The white cross indicates the peak of the spatial spectrogram. (**b**) The USV segment was then assigned to the mouse with the highest power if the assignment confidence was higher than 99%. The confidence value was estimated using the *p*-value of the comparison between the best two mice (*p*), the distance between the snouts of the best two mice (*d*), and the number of candidate mice (N_c), based on simulations using the data of the single mouse experiments. See Methods for details.

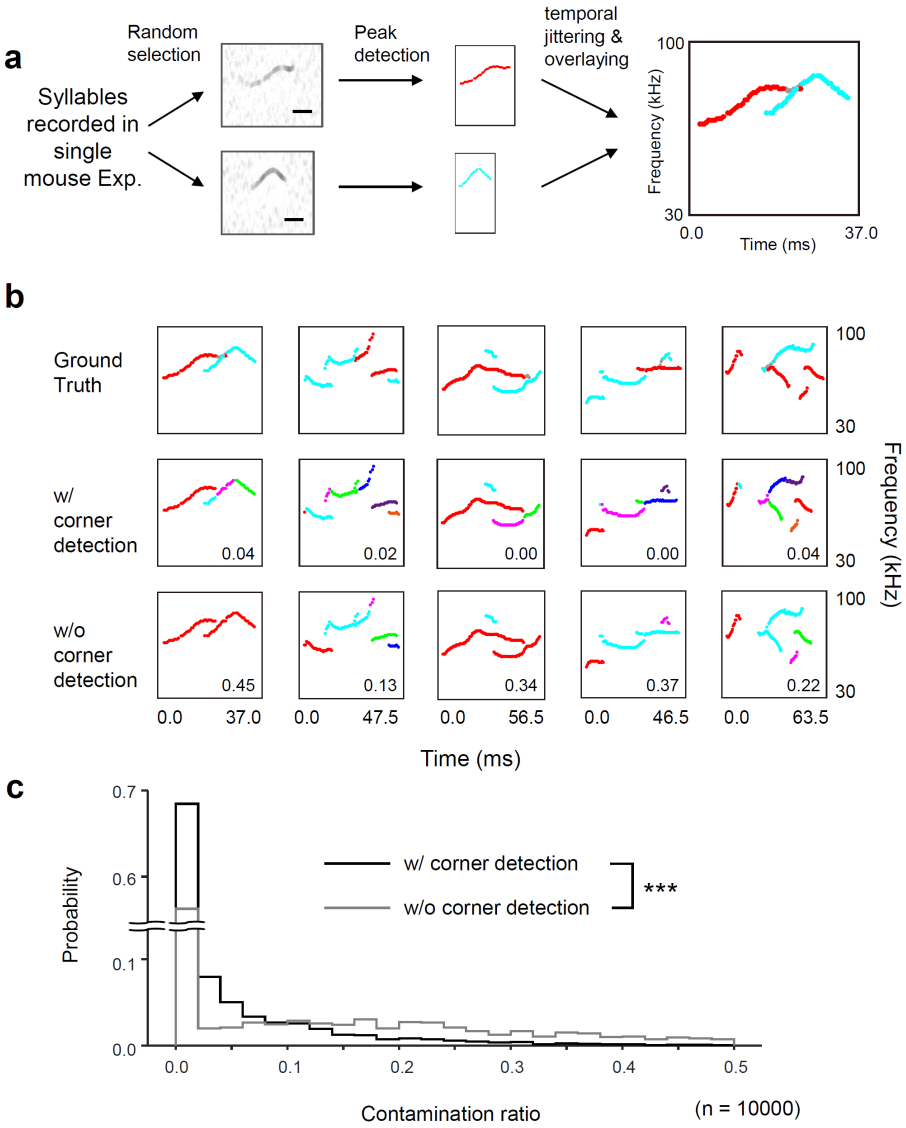


641 **Supplementary Figure 5. Distribution of assignment precision in the simulation**
642 **using data obtained from the single mouse experiments.**
643 p, *p*-value of the sign-rank test comparing the sound powers at the snouts between the
644 best two mice. N_c, the number of candidate mice. White curves indicate the border of 0.99.

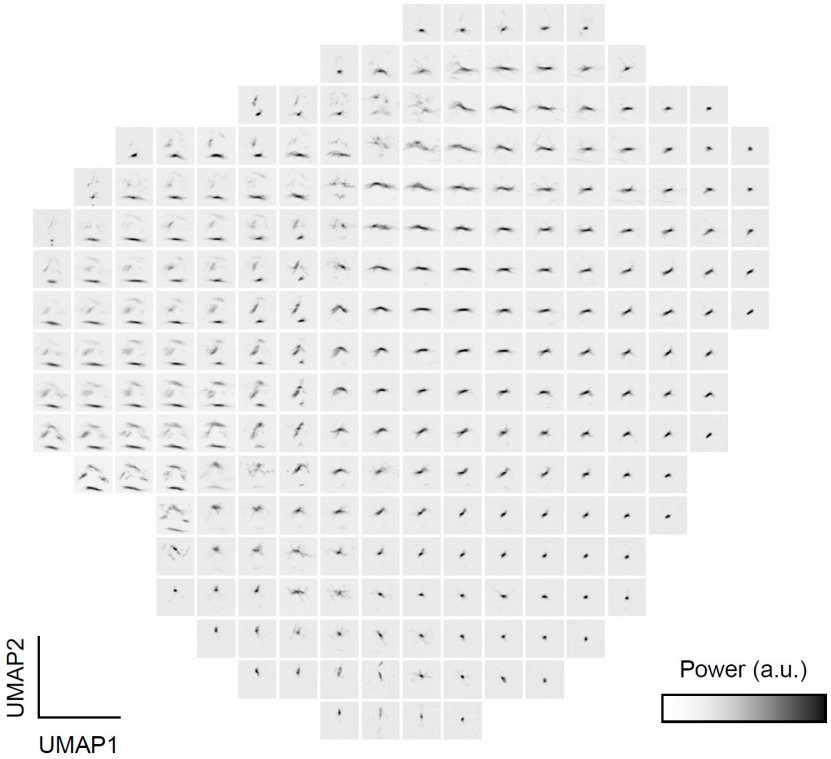


Supplementary Figure 6. Distribution of assignment precision similar to that in Supplementary Figure 5, calculated separately for B6 and ICR mice.

Top: distribution of the precision. Bottom, comparison of the borders of 0.95 and 0.99 between strains. Because the simulation results for B6 and ICR mice were largely similar to those above, we used a combined distribution (Supplementary Fig. 5) for the assignments in this study.

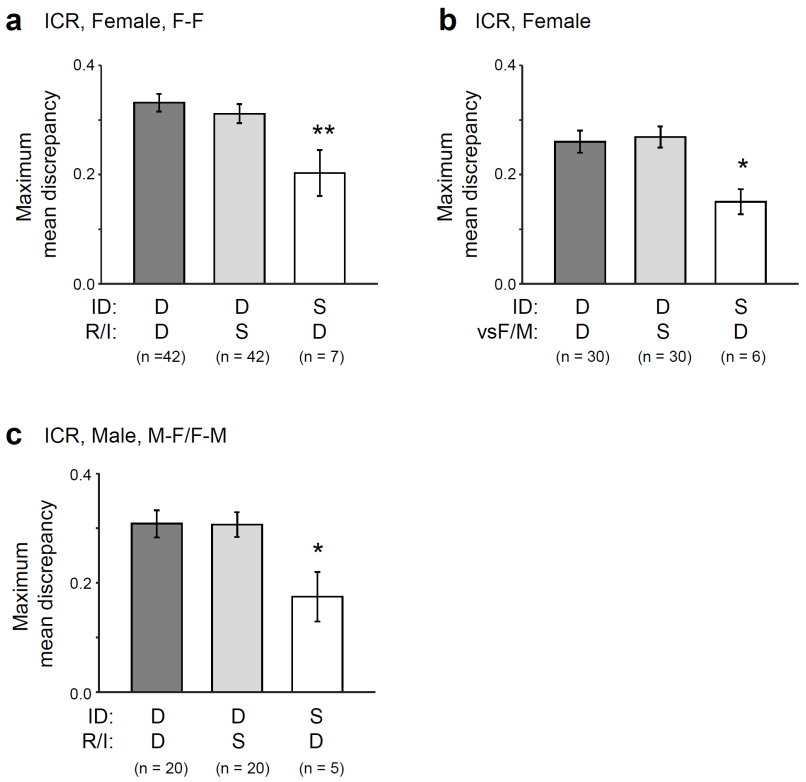


656 **Supplementary Figure 7. Effectiveness of the modified segmentation algorithm.**
657 (**a**) Test data generation. Scale bars in the spectrograms = 20 ms. Gray peaks were
658 excluded to simulate the peak search in the USVSEG algorithm. (**b**) Examples of the
659 ground truth labels (top) and segmentation with (middle) and without (bottom) corner
660 detection (Supplementary Figure 1e) for separation at the crossing points. Different colors
661 of the ground truth indicate different sources. Different colors in the segmentation results
662 indicate different segments. The number in each segmentation result indicates the
663 contamination ratio. (**c**) The distribution of contamination ratios in the segmentation with
664 (black) and without (gray) corner detection. ***$p = 1.9 \times 10^{-117}$, two-tailed Wilcoxon
665 rank-sum test.
666



Supplementary Figure 8. Mean spectrograms of the syllables for each location on the UMAP projection in Fig. 2c.

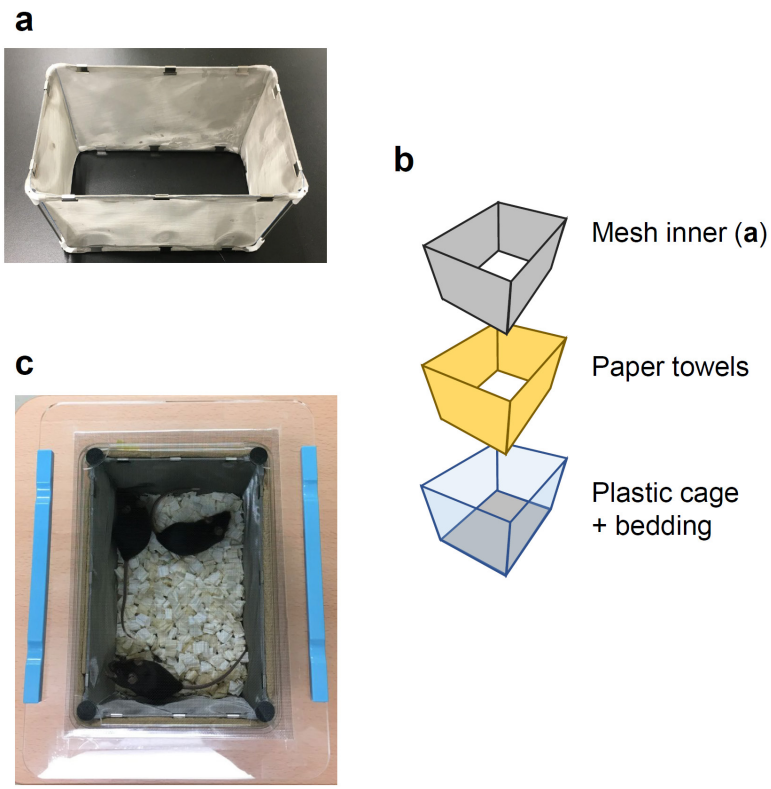
The frequency range of the spectrograms is 30 to 100 kHz, and the time range is 128 ms.



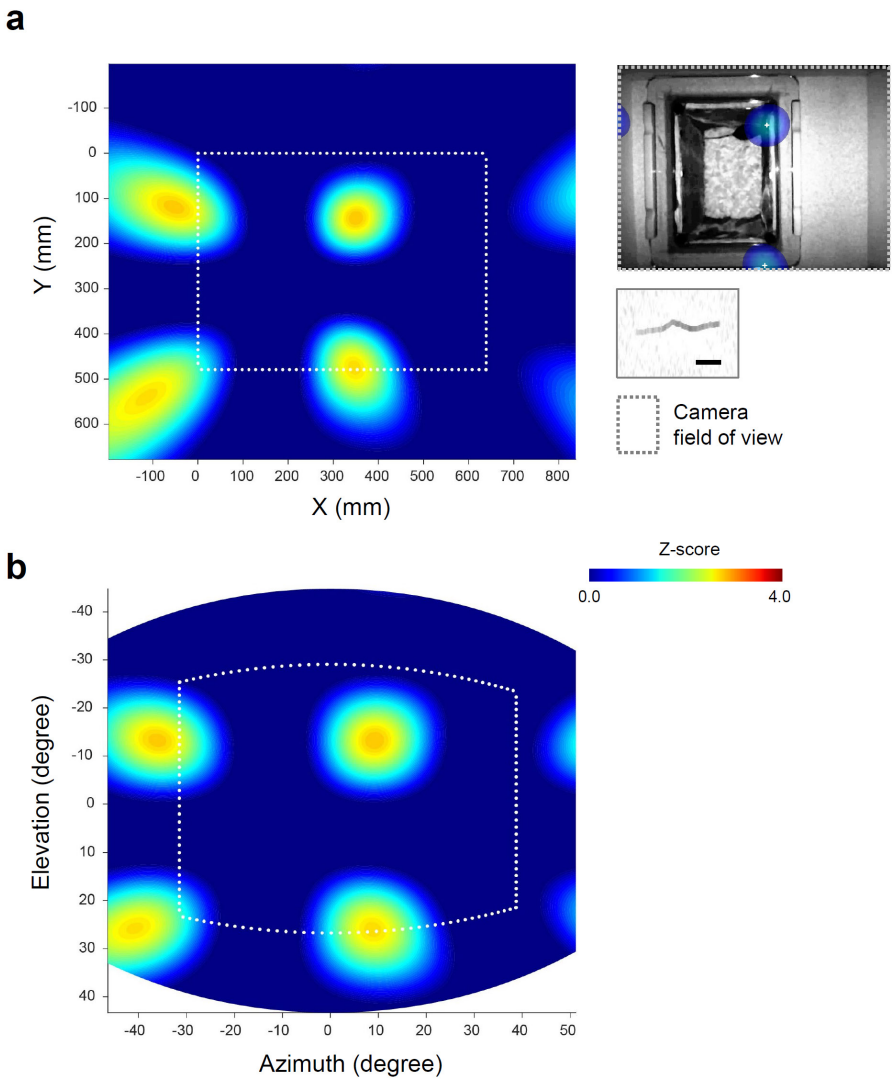
Supplementary Figure 9. Quantitative analysis of the similarity of vocal repertoires within the same social contexts or individuals.

Dissimilarities in vocal repertoire calculated as maximum mean discrepancy (MMD) between pairs of distributions of the features of syllables (Fig. 2c) emitted by individuals. D, different; S, same. Error bars, s.e.m. (**a**) Comparisons of mean MMDs between groups of pairs of different IDs and different R/I labels (dark gray), pairs of different IDs and the same R/I labels (light gray), and pairs of the same IDs and different R/I labels (white) in female ICR mice in F-F sessions. $^{**}p < 0.01$, difference from pairs of different IDs and different R/I labels (dark gray) with Dunnett's test after one-way analysis of variance (ANOVA) ($F[2,88] = 4.154$; $p = 0.0189$). (**b**) Comparison of mean MMDs between groups of pairs of different IDs and different vsF/M labels (dark gray), pairs of different IDs and the same vsF/M labels (light gray), and pairs of the same IDs and different vsF/M labels (white) in female ICR mice in F-F, F-M, and M-F sessions. $^{*}p < 0.05$, difference from pairs of different IDs and different vsF/M labels (dark gray) with Dunnett's test after one-way ANOVA ($F[2,63] = 3.261$; $p = 0.0449$). (**c**) Similar comparison as (**a**) in male ICR mice in F-M and M-F sessions. $^{*}p<0.05$, difference from pairs of different IDs and different R/I labels (dark gray) with Dunnett's test after one-way ANOVA ($F[2,42] = 3.504$; $p = 0.0391$).

Supplementary Figure 10. The soundproof box used in this study.
The floor was made of polyvinyl chloride. The walls and ceiling were covered with sound-absorbing melamine foam to reduce sound reflections.

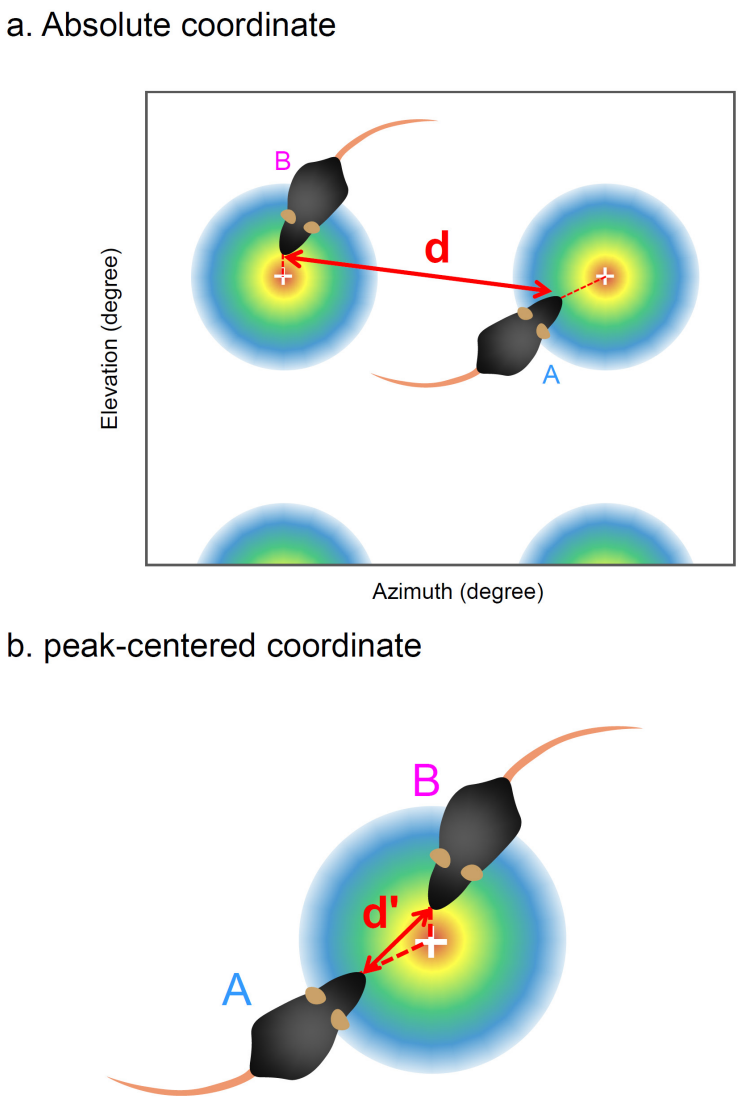


Supplementary Figure 11. Details of the home cage recording.
(**a**) The inner cage was made of fine stainless mesh. (**b**) An illustration of the layers of the cage. (**c**) The cage lid used for recording. A clear mesh screen was glued onto an acrylic frame. Rubber feet were attached to the corners of the lid to pin down the inner cage. The light blue bars are weights.



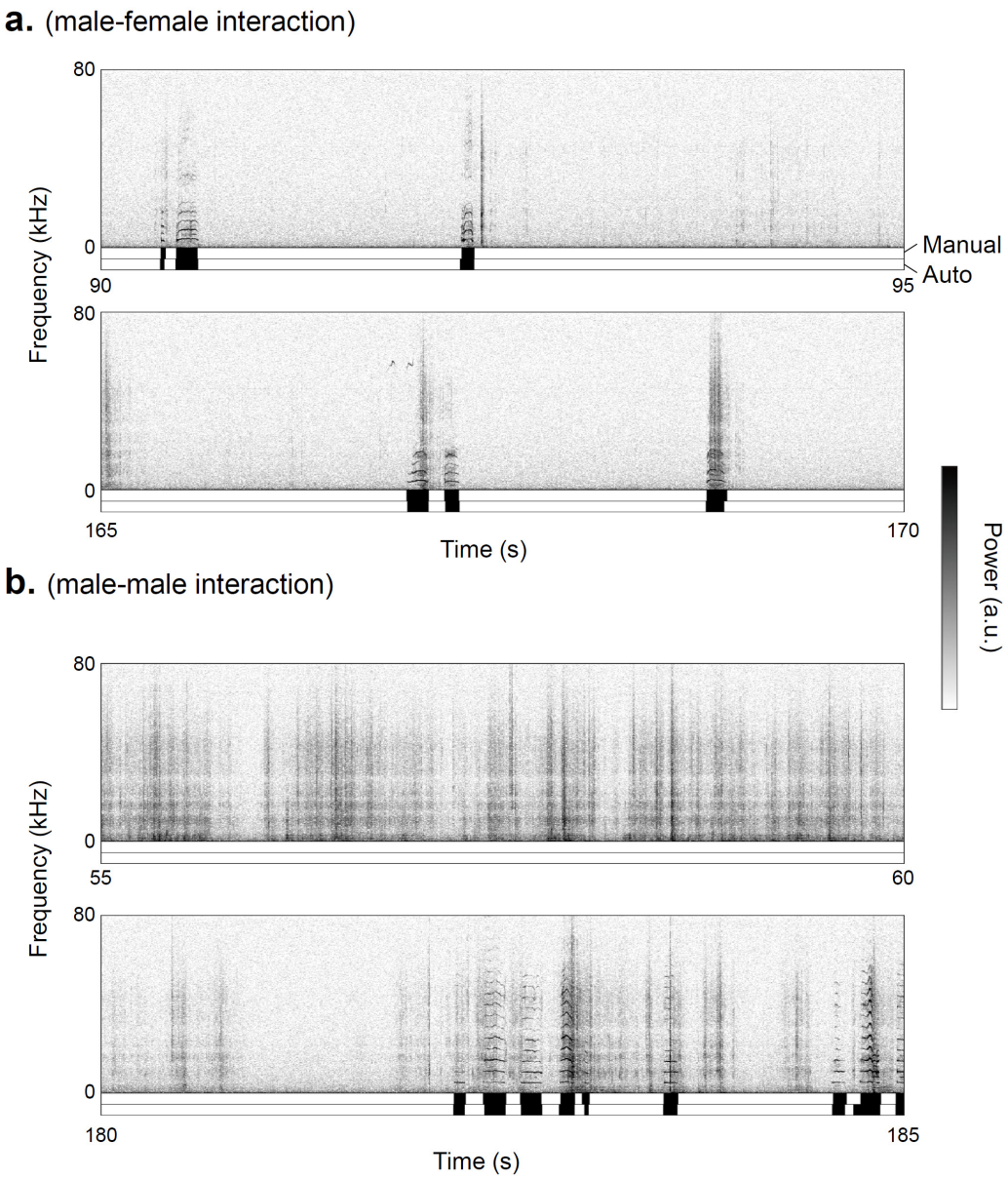
Supplementary Figure 12. An example of the spatial spectrum shown in a wide field of view.

(**a**) Spatial spectrum around the camera field of view. Inset shows the spatial spectrum overlayed on the video frame (top) and sound spectrogram (frequency range = 30 to 100 kHz, scale bar = 20 ms) of a USV segment. (**b**) The same spatial spectrum projected in the direction from the center of the microphone array. The peaks are located in a square grid.

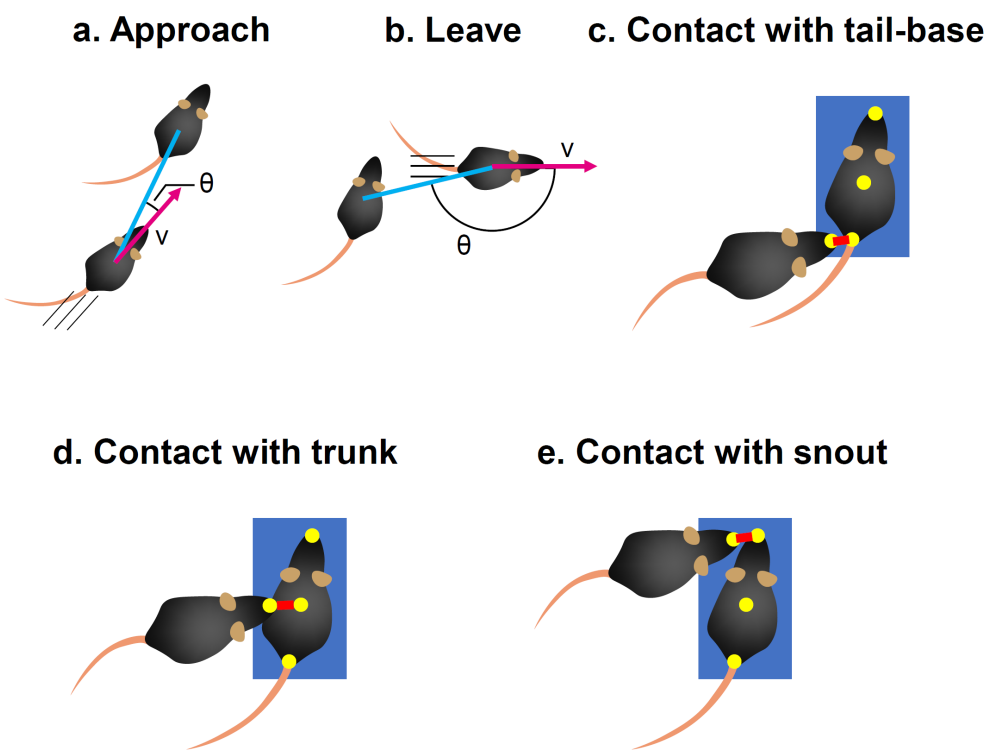


Supplementary Figure 13. Spatial 'wrapping' for fixing snout-snout distance for the assignment.

(**a**) An illustration of a case where mice are near different peaks. (**b**) The mouse locations in **a** are aligned to the peak-centered coordinates. The snout-snout distance in this space (*d'*) was used for estimating assignment confidence. A and B, IDs of mice.



Supplementary Figure 14. Examples of BBV detection using the automatic detector. Examples of 5 s spectrograms are shown in male-female (**a**) and male-male (**b**) ICR mice interactions. Black bars in the first and second rows below the spectrogram indicate manual annotation and automatic detection of BBVs, respectively.



Supplementary Figure 15. Definitions of the behavioral events.
Approach (**a**) was counted when a subject's running speed (v) was > 50 mm/s, and the angle between the partner location and running direction (θ) was < 45 degrees. *Leave* (**b**) was counted when v was > 50 mm and θ was > 135 degrees. *Contact with the tail-base* (**c**), *trunk* (**d**), and *snout* (**e**) was counted when a subject's snout was inside the bounding box of the partner (blue shaded area), and the body part (yellow point) of the partner closest to subject's snout was the tail-base, the center of the bounding box, and the snout, respectively. The trunk position was calculated as the center of the bounding box.

742 **Supplementary Table 1. Assignment performance in the simulations based on the**
743 **single mouse experiment.**
744

			Assigned			Unassigned		
	Sub.	Total	Hit	Error	Precision	Low Conf.	Unlocalized	Pos. N/A
B6	2	3384	2873 (84.9%)	5 (0.1%)	**99.8%**	410 (12.1%)	65 (1.9%)	31 (0.9%)
	3	3384	2537 (75.0%)	15 (0.4%)	**99.4%**	736 (21.7%)	65 (1.9%)	31 (0.9%)
	4	3384	2213 (65.4%)	23 (0.7%)	**99.0%**	1052 (31.1%)	65 (1.9%)	31 (0.9%)
ICR	2	5678	4777 (84.1%)	17 (0.3%)	**99.6%**	501 (8.8%)	293 (5.2%)	90 (1.6%)
	3	5678	4347 (76.6%)	32 (0.6%)	**99.3%**	916 (16.1%)	293 (5.2%)	90 (1.6%)
	4	5678	3956 (69.7%)	31 (0.5%)	**99.2%**	1308 (23.0%)	293 (5.2%)	90 (1.6%)

745
746 Sub., number of subjects in the cage; one is the real mouse and the others are virtual
747 mice randomly positioned in the cage.
748 Total, total number of USV segments detected.
749 Hit, number of segments correctly assigned to the real mouse.
750 Error, number of segments incorrectly assigned to one of the virtual mice.
751 Precision, percentage of Hit ÷ (Hit + Error).
752 Low Conf., number of segments in which assignment confidences were < 0.99.
753 Unlocalized, number of unlocalized segments.
754 Pos. N/A, number of segments emitted when the snout positions could not be estimated
755 because of video-based tracking issues.
756 Audible call, number of segments that overlapped with audible calls.
757 Percentages in parentheses represent ratios relative to the total number of segments.
758
759
760

Supplementary Table 2. Assignment performance in the home cage social interaction experiment.

				Unassigned		
	Sub.	Total	Assigned	Low Conf.	Unlocalized	Pos. N/A
B6	2 (F-F)	102	84 (82.4%)	14 (13.7%)	2 (2.0%)	2 (2.0%)
	2 (M-M)	30	26 (86.7%)	3 (10.0%)	0 (0.0%)	1 (3.3%)
	2 (F-M)	1795	1420 (79.1%)	339 (18.9%)	14 (0.8%)	22 (1.2%)
	2 (M-F)	2518	1823 (72.4%)	483 (19.2%)	29 (1.2%)	183 (7.3%)
	3 (M-FF)	970	578 (59.6%)	263 (27.1%)	52 (5.4%)	77 (7.9%)
ICR	2 (F-F)	16052	9813 (61.1%)	3203 (20.0%)	468 (2.9%)	2568 (16.0%)
	2 (M-M)	743	191 (25.7%)	288 (38.8%)	27 (3.6%)	237 (31.9%)
	2 (F-M)	8030	3881 (48.3%)	1760 (21.9%)	236 (2.9%)	2153 (26.8%)
	2 (M-F)	2897	1497 (51.7%)	759 (26.2%)	109 (3.8%)	532 (18.4%)
	3 (M-FF)	1607	569 (35.4%)	804 (50.0%)	69 (4.3%)	165 (10.3%)

Sub., number of subjects in the cage; the first and second letters in the parentheses indicate resident and intruder(s). M, one male; F, one female; FF, two females.
Total, total number of USV segments detected.
Assigned, number of segments assigned to one of the mice.
Low Conf., number of segments in which assignment confidences were < 0.99.
Unlocalized, number of unlocalized segments.
Pos. N/A, number of segments emitted when the snout positions could not be estimated because of video-based tracking issues.
Percentages in parentheses represent ratios relative to the total number of segments.

775 **Supplementary Table 3. Detailed results of the two-way ANOVA in Fig. 2a.**

776

		F(1,9)	*p*-value
ICR, Female	(vs F/M)	8.1058	0.0192*
	(R/I)	0.2727	0.6141
	(vs F/M) x (R/I)	10.0044	0.0115*
ICR, Male	(vs F/M)	5.1275	0.0498*
	(R/I)	0.0641	0.8059
	(vs F/M) x (R/I)	0.1553	0.7027
B6, Female	(vs F/M)	0.2791	0.6101
	(R/I)	0.5974	0.4594
	(vs F/M) x (R/I)	2.0187	0.1891
B6, Male	(vs F/M)	2.2182	0.1706
	(R/I)	0.0535	0.8223
	(vs F/M) x (R/I)	0.1290	0.7277

777

778 vs F/M, whether the partner was a female or a male; R/I, whether the subject was a
779 resident or an intruder; *$p < 0.05$.

780

781 **Supplementary Table 4. Detailed results of the three-way ANOVA in Fig. 2b.**

782

	df	F	*p*-value
(Action)	(7, 63)	12.815	4.1×10^{-10}***
(vs F/M)	(1, 9)	10.022	0.011*
(R/I)	(1, 9)	0.253	0.63
(Action) x (vs F/M)	(7, 63)	2.064	0.061
(Action) x (R/I)	(7, 63)	1.450	0.20
(vs F/M) x (R/I)	(1, 9)	9.705	0.012*
(Action) x (vs F/M) x (R/I)	(7, 63)	4.895	0.00019***

783

784 Action, the type of actions; vs F/M, whether the partner was a female or a male; R/I,
785 whether the subject was a resident or an intruder; *$p < 0.05$; ***$p < 0.001$.

786

Supplementary Table 5. Number of ultrasound segments that overlapped with BBVs.

	Sub.	# of segments
B6	2 (F-F)	0
	2 (M-M)	1
	2 (F-M)	0
	2 (M-F)	3
	3 (M-FF)	0
ICR	2 (F-F)	99
	2 (M-M)	959
	2 (F-M)	834
	2 (M-F)	1999
	3 (M-FF)	1236

Sub., number of subjects in the cage; the first and second letters in the parentheses indicate resident and intruder(s). M, one male; F, one female; FF, two females.

Supplementary Video 1. An example of USV localizations emitted from a male B6 mouse in a home cage of a female mouse.

Supplementary Video 2. An example of USV localizations emitted from a male ICR mouse in a home cage of a female mouse.

Supplementary Video 3. An example of USV assignment during an interaction between a male (cyan) and a female (magenta) B6 mouse in a home cage.

Supplementary Video 4. An example of USV assignment during an interaction between two female ICR mice in a home cage.

Supplementary Video 5. An example of the video tracking result used for behavioral event classification.