Multimodal system for recording individual-level behaviors in songbird groups

L. Rüttimann[1*], J. Rychen[1], T. Tomka[1,2], H. Hörster[1], M. D. Rocha[1], and R.H.R. Hahnloser[1,2*]

[1]Institute of Neuroinformatics, University of Zurich and ETH Zurich, 8057 Zurich, Switzerland.

[2]Neuroscience Center Zurich (ZNZ), University of Zurich and ETH Zurich, 8057 Zurich Switzerland.

*) corresponding authors: rlinus@ini.ethz.ch, rich@ini.ethz.ch

Abstract

In longitudinal observations of animal groups, the goal is to identify individuals and to reliably detect their interactive behaviors, including their vocalizations. However, to reliably extract individual vocalizations from their mixtures and other environmental sounds remains a serious challenge. Promising approaches are multimodal systems that exploit signal redundancy and make use of animal-borne wireless sensors. In this vein, we designed a modular recording system (BirdPark) that yields synchronized data streams. We recorded groups of songbirds with multiple cameras and microphones and recorded their body vibrations with custom low-power frequency-modulated (FM) radio transmitters. We developed a custom software-defined radio receiver with a multi-antenna demodulation technique that increased the signal-to-noise ratio of the received radio signals by 6.5 dB and reduced the signal loss rate due to fading by a factor of 63 to only 0.01% of the recording time compared to single-antenna demodulation. Nevertheless, neither a single vibration sensor nor a single microphone is sufficient by itself to detect the complete vocal output of an individual. Even in the minimal setting of an animal pair, an average of about 3.7% of vocalizations remain undetected within each sensor modality. Our work emphasizes the need for high-quality recording systems and for multimodal analysis of social behavior.

Introduction

Acoustic communication is vital for many social behaviors. However, to study interactions among animals that are kept in groups entails many measurement challenges beyond the already considerable challenges of analyzing longitudinal data of isolated animals[1–3]. One of the key difficulties of group-level behavior research is to perform automatic recognition of individuals and their actions.

Action recognition is the task of detecting behaviors from video or audio, or from signals collected with animal-borne sensors. Video-based action recognition has traditionally been based on posture tracking[4–6] to avoid data-hungry training of classifiers on high-dimensional video data. Recently, posture tracking has greatly improved thanks to deep-learning approaches[7–14]. Action recognition from video requires good visibility of focal animals because visual obstructions tend to hamper recognition accuracy. Given

that freely moving animals may occlude one another, e.g., in birds during nesting, there seems to be a limitation to the usefulness of pure vision-based approaches.

Sound recordings have also been instrumental in action recognition. Sounds can be informative about both vocal and non-vocal behaviors. For example, wing flapping during flying produces a characteristic sound signature. But also preening, walking, and shaking can be recognized from sounds[15]. The task of classifying sounds is known as acoustic scene classification[16]. Similar to vision, microphones record not just the focal animal but also background sounds. Therefore, action recognition and actor identification from sounds alone are challenging tasks, especially when many animals interact with one another. Possible workarounds to these issues are multimodal approaches that combine multiple cameras and microphones, for example, to assign vocalizations from dairy cattle in a barn to individual animals[17]. Other examples are systems that also include motion-tracking devices, for example, to quantify gesture–speech synchrony in humans[18].

Methods that integrate different data modalities have been deemed particularly promising for functional decoding of animal communication[19]. One particular challenge with multimodal systems is the synchronization of the various data streams. Usually, each sensor modality is recorded with a dedicated device that uses its own internal sampling clock, and these clocks tend to drift apart. Furthermore, the recordings must be started at exactly at the same time on all devices, or the individual data streams must be aligned post-recording using markers in the sensor signals or auxiliary synchronization channels, which are labor intensive and error-prone processes[18,20,21].

In general, limitations due to sight occlusions and sound superpositions can be overcome with animal-borne sensors such as accelerometers[22,23], gyroscopes, microphones[24], and global positioning systems (GPS)[23]. In combination with wireless transmitters[24] and loggers[22], these sensors enable the detection of behaviors such as walking, grooming, eating, drinking, and flying, for example, in birds[25], cats[26], and dogs[27], though often with low reliability due to noisy and ambiguous sensor signals[15]. In general, animal-borne transmitter devices are designed to achieve high reliability, low weight, small size, and long battery life, giving rise to a complex trade-off. Among the best transmitters, in terms of battery life, size, and weight, are analog frequency-modulated (FM) radio transmitters. Their low power requirement minimizes the frequency of animal handling and associated handling stress, making them an excellent choice for longitudinal observations of small vertebrates[24,28,29].

Among the challenges associated with FM radio reception is radio signal fading due to relative movements of animal-borne transmitters and stationary receivers. Fading arises when electromagnetic waves arrive over multiple paths and interfere destructively (channel fading)[30], e.g., by reflection of metallic walls. Fading also occurs because every receiver has a direction of zero gain, which may affect the reception of the signal from a moving transmitter. Signal fading can be addressed with antenna diversity, i.e., the use of several antennas. To combine the diverse antenna signals, either the strongest signal is selected, all signals are summed up, or signals are first weighted by their strength and then summed[31]. However, these approaches do not completely prevent fading because the signals can still annihilate. Alternatively, diversity combining is possible with phase compensation, which is the technique of turning signal phases such that phase compensated signals align and sum constructively[32,33]. Phase compensation reduces fading and increases the signal-to-noise ratio of the received signal, and, furthermore, it provides cues for localizing a transmitter[34,35]. We set out to bring the benefits of antenna diversity and phase compensation techniques to ethology research.

To perform individual-level longitudinal observations of social behaviors and to record high-quality multimodal datasets suitable for action recognition, we present a custom recording system for groups of vocalizing animals (BirdPark). We built a naturalistic environment inside a soundproof enclosure that features a set of microphones to record sounds and several video cameras to capture the entire scene. Moreover, all animals wear a miniature low-power transmitter device that transmits body vibrations from a firmly attached accelerometer via an analog, frequency-modulated (FM) radio signal that we receive with several antennas.

Our system is optimized for robust longitudinal recordings of vocal interactions in songbirds. The on-animal accelerometers enable week-long monitoring of vocalizations without a change of battery. The combination of multiple antennas minimizes signal losses and improves the signal-to-noise ratio. All sensor signals are perfectly synchronized, which we achieve by routing dedicated sample triggers to all recording devices (radio receiver, stationary microphone digitizer, and cameras) derived from a central quartz clock using clock dividers. We release our custom recording software and we demonstrate the high data quality and redundant signaling of vocal gestures.

Results

The Recording Arena

We built an arena optimized for audiovisual recordings, minimizing acoustic resonances and visual occlusions (Figure 1A). It provides space for up to 8 songbirds and contains nest boxes, perches, sand baths, food, and water sources. To record the sounds inside the chamber, we installed five microphones. Three video cameras capture the overall scene from three orthogonal viewpoints. In addition, we installed a camera and microphone in each of the two nest boxes. To simplify video analysis, we combined the images from all five cameras into a single video image (Figure 1B). Camera resolutions are sufficient to resolve key points on birds even in mid-flight (Figure 1D).

To record vocalizations, we mounted transmitter devices to birds' backs. Each device comprises an accelerometer that picks up body vibrations from vocalizations and body movements such as hopping and wing flapping[22]. The devices transmit the accelerometer signals as frequency-modulated (FM) radio waves to four antennas inside the recording chamber.

We demodulated the FM radio signals with a custom eight-channel radio receiver that was implemented on a universal software radio peripheral (USRP) with a large field programmable gate array (FPGA). To ensure that the data streams from the video cameras, microphones, and transmitter devices are synchronized, we record videos with industrial cameras and digitize sounds on a National Instruments data acquisition board (NI DAQ), both of which receive sample trigger signals from the USRP. The base clock of the USRP was 200 MHz and was divided by 2^{13} to generate the audio sample rate of 24.414 kHz, and with a further division by 2^{9}, we generated the 47.684 Hz video frame rate.

On the host computer that controlled the acquisition system, we ran two custom applications: BirdVideo, which acquires and writes the video data to a file, and BirdRadio, which acquires and writes the microphone and transmitter device data to a file (see Methods). The generated file pairs are synchronized such that with each video frame there are 512 audio samples. While the recording is gapless, it is split into files of typically 7 minutes duration.

Transmitter device

Our transmitter devices are based on the FM radio transmitter circuit described in Ter Maat et al.[24] (Figure 2A). To distinctly record birds' vocalizations irrespective of external sounds, we replaced the microphone with an accelerometer[22] that picks up body vibrations and acts as a contact microphone. The radio circuit uses a single transistor to modulate the sensor signal $a_{\mathrm{T}}(t)$ onto a radio carrier frequency ω_c (that is set by the resonator circuit properties) with an inductor coil as an emitting antenna. As a result, the accelerometer signal $a_{\mathrm{T}}(t)$ is encoded as (instantaneous) transmitter frequency $\omega_{\mathrm{T}}(t) \simeq \omega_c(t) + ca_{\mathrm{T}}(t)$ with c some constant.

Before mounting a device on a bird, we adjusted the coil to the desired carrier frequency in the range 250–350 MHz by slightly bending the coil wires. Thereafter, we fixed the coil and electronics in dyed epoxy. The purpose of the dyed epoxy is to help identify the birds in the video images, to protect the printed circuit board, and to stabilize the coil's inductance (Figure 2B,C). We mounted the devices on birds using a rubber-string harness adapted from Alarcón-Nieto et al.[36] (Figure 2D).

Radio receiver

To reconstruct the accelerometer signal $a_{\mathrm{T}}(t)$ of a given transmitter device from the received multi-antenna FM signals, we demodulated the latter using a phase-locked loop (PLL), which measures the instantaneous transmitter frequency $\omega_T(t)$. A PLL generates an internal oscillatory signal of variable tracking frequency $\omega(t)$. It adjusts that frequency to maintain a zero relative phase with respect to the received (input) signal. As long as the zero relative phase condition is fulfilled, the PLL's tracking frequency $\omega(t)$ follows the instantaneous transmitter frequency $\omega_T(t)$, i.e., the tracking frequency $\omega(t)$ is the demodulated (transmitter) signal. After high-pass filtering, this signal forms our estimate of the accelerometer signal $a(t)$.

In our diversity combining approach, we construct the PLL's input signal as a combination of all four antenna signals. Namely, we compensate the individual phase offsets of the four antenna signals in such a way that all phases are aligned (we achieve this phase shifting with phase-compensation circuits — one for each antenna). We then form the desired mixture signal by summing the phase-shifted signals; the phase of the summed signal serves as the PLL's error signal that we use to adjust the tracking frequency $\omega(t)$. The variable phase offsets on the four antennas arise from the variable locations and orientations of transmitters relative to receivers.

In summary, our approach to minimizing fading in wireless ethology research is to use a demodulator comprising a PLL and a diversity combining approach based on phase-compensation. Our FM radio demodulation is described in more detail in the following.

We implemented our custom demodulation technique by installing four antennas labeled A, B, C, and D perpendicular to the walls and the ceiling of the chamber (see Methods: Radio reception). We fed the antenna signals into a USRP containing a large FPGA (Figure 3A). The four input stages of the USRP filter out an 80 MHz wide band around the local oscillator frequency ω_{LO}, which we typically set to 300 MHz. These four signals then become digitally available on the FPGA as complex valued signal of 100 MHz sampling rate. We call them the intermediate signals $z^a(t),\ a \in \{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}\}$ (intermediate band in the frequency domain, Methods: Intermediate band, Figure 3B).

On the FPGA, we instantiated eight demodulators indexed by $i \in \{1, \ldots, 8\}$. Each demodulator contained four digital downconverters (DDC) that cut out from the four intermediate signals $z^a(t)$ the four

baseband signals $u_i^a(t)$, $a \in \{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}\}$ around the eight tracking frequencies $\omega_i(t)$, sampled at 781.25 kHz (Methods: Baseband, Figure 3C and D). The digital signals are complex valued, so we interchangeably call them vectors and their phases we refer to as angles.

The PLL is driven by the main vector $u_i^{\mathrm{M}}(t) = u_i^{\mathrm{A}}(t) + u_i^{\mathrm{B}}(t) + u_i^{\mathrm{C}}(t) + u_i^{\mathrm{D}}(t)$ that we formed as the sum of the four baseband vectors. The error signal for the PLL's feedback controller is given by the phase of the main vector $\theta_i(t) = \arg\left(u_i^{\mathrm{M}}(t)\right)$ (Methods: PLL, Figure 3D). When that phase is approximatively zero (PLL is locked), the tracking frequency $\omega_i(t)$ tracks the frequency of transmitter i. We updated the PLL's tracking frequency at a rate of 781.25 kHz.

To avoid destructive interference in the summation of the baseband vectors, we compensated their phases $\alpha_i^a(t) = \measuredangle\left(u_i^{\mathrm{M}}(t), u_i^a(t)\right)$ relative to the main vector. We introduced individual phase offset $\Delta\varphi_i^a(t)$ that were set in feedback loops to drive the phases $\alpha_i^a(t)$ towards zero (Methods: Phase compensation, Figure 3E). Phase offsets were updated at a rate of 1.5 kHz and logged at every video frame.

The PLL and phase compensation form independent control loops. When the PLL is unlocked (e.g., off), the tracking frequency does not match the instantaneous transmitter frequency and the baseband vectors rotate at the difference frequency (Figure 3F left). When the PLL is switched on and locked, the baseband vectors do not rotate, and the main signal displays a phase $\theta_i(t) \simeq 0$ (Figure 3F middle). When the phase alignment is switched on, the baseband signals align, and their sum maximizes the magnitude of the main vector (Figure 3F right).

Because birds' locomotion is slower than their rapid vocalization-induced vibratory signals, the phases $\alpha_i^a(t)$ change more slowly than the tracking frequency $\omega_i(t)$ and therefore we updated phase offsets less often than the PLL's tracking frequency.

Operation

The intermediate band of $z^a(\omega)$ is wide enough to accommodate up to eight FM transmitters. Provided the transmitters' FM carrier frequencies are roughly evenly spaced, the transmitter frequencies do not cross, even during very large frequency excursions. Nevertheless, we limited the tracking frequencies ω_i to the range $[\Omega_i - \Delta\Omega, \Omega_i + \Delta\Omega]$, where Ω_i is the center frequency and $\Delta\Omega = 1$ MHz is the common limiting range of all channels (Figure 3B). The center frequency Ω_i of a channel we manually set at the beginning of an experiment to the associated FM carrier frequency. The limiting range of 1 MHz we found narrow enough for the PLL to rapidly relock after brief signal losses.

During operation, we often observed large excursions of the transmitter frequency. These excursions occurred while birds pecked on the device or while they preened their feathers near the sensor, or when one bird sat on top of another, such as during copulations. The magnitude of these frequency excursions could reach 2.5 MHz (68 dB re 1 kHz), which is much larger than the maximally 3.4 kHz (11 dB re 1 kHz) shifts induced by vocalizations (proximity effect, Figure 4). The large excursions likely were caused by dielectric effects of body movements on the capacitor of the resonator circuit, i.e., the large excursions affected the carrier frequency ω_c that was not constant but modulated by the proximity of body parts. Additionally, the carrier frequency ω_c was subject to temperature and end-of-battery-life drift (see Methods: Transmitter device). While the instability of the carrier frequency ω_c is a disadvantage of the

single-transistor circuit design, advantages are the low device weight (1.5 g) and the small power consumption it enables (the battery lifetime was about 12 days).

Performance of diversity combining technique

We validated the robustness of our diversity combining multi-antenna demodulation method by analyzing the signal-to-noise ratio of the received transmitter signal while recording a zebra finch pair over a full day (2 x 14 h measurement period). We quantified the performance by comparing the radio signal-to-noise ratio (RSNR) of the diversity combining multi-antenna signal with the RSNR of the four single single-antenna signals. The RSNR was defined as the signal power (mean square) divided by the median of the signal variance (the median is taken over the duration of 7-min long data files) of $u_i^a(t)$ (a=A,B,C,D,M). Our recording software BirdRadio logs power and variance trajectories of $u_i^a(t)$ (a=A,B,C,D,M), this allows us to evaluate both single-antenna and multi-antenna RSNR on the same dataset. Furthermore, RSNR is a good measure of the signal quality of the demodulated signal because the noise power of the accelerometer signal $a(t)$ is monotonically decreasing with increasing RSNR (Figure 5A,B).

In theory, the noise power (variance) of $u_i^{\mathrm{M}}(t)$ is four times as large as that of either of the baseband signals $u_i^a(t)$ (a=A..D), assuming independence of radio amplifier noises. The signal power of $u_i^{\mathrm{M}}(t)$ is 16 times that of a single antenna signal $u_i^a(t)$. According to this assumption, we expect a 6 dB RSNR increase for four-antenna demodulation.

By analyzing recorded data of a zebra finches over a full day (2 x 14 h measurement period), we found that the RSNR was about 6.5 dB higher on average than that of the best single antenna (Figure 5C). The total reception time during which the RSNR was critically low (<13 dB, our operational definition of signal fading) was 63 times shorter when using the multi-antenna demodulation than when using the best single-antenna signal (Figure 5D).

To verify that simple whip antennas that we use are not worse than more complex antennas, we recorded a second test dataset of a zebra finch pair (again, 2 x 14 h measurement period) with a single volumetric antenna using single-antenna demodulation. The performance of the volumetric antenna was worse than that of the best simple antenna (Figure 5C,D).

Thus, our diversity combining technique is useful and results in a coherent constructive addition of the four antenna signals and to very robust frequency tracking even when the signal in one antenna vanishes.

Vocal segmentation analysis

Some vocalizations on microphone channels were masked by vocal signals of other birds (as e.g., in Figure 6A) or by mechanical noises (Examples B1/B2 in Figure 6B). Compared to these signals, we observed that vocal signals on the transmitter channels reached up to 7 kHz in optimal cases, but only up to 1 kHz in the worst case (Figure 6A). Some vocalizations were invisible on the acceleration spectrogram (Example A2 in Figure 6B). Others were masked due to simultaneous body movements such as wing flaps or hops (Example A1 in Figure 6B). To quantify these effects relative to vocalizations as detected in the multimodal data stream, we segmented vocal activity by visual inspection of sound and acceleration spectrograms from five microphone and two animal-borne transmitter channels. (Table 1, see Methods for details). We evaluated the number of missed vocalizations on a given channel relative to a consolidated segmentation that was based on visual inspection of all transmitter and microphone

channels combined. We were particularly interested in missed vocalizations that are invisible on a given channel but visible on another channel. Compared to the miss rate of 4.2% (range 2.5%-10.0%, 7 mins data segments from n=2 bird pairs each) associated with a single microphone channel, all microphone channels combined produced a reduction in miss rate by a factor of about 10, which reveals a large benefit of the multi-microphone approach. The percentage of missed vocalizations on a given transmitter channel was similar to that of a single microphone channel: 3.7% (range 1.2%-5.8%, n=4 birds). Note that these numbers are not directly comparable because the miss rate on a microphone channel stems from two birds and the miss rate on a vibration channel from a single bird. A very small percentage of 0.4% (range 0.0%-0.6%, n=2 bird pairs) of vocalizations were missed on all six microphone channels, these vocalizations were all calls masked by loud noises.

The benefit of the transmitter devices is most apparent when vocalizations overlap. While microphone channels picked up mixtures of overlapping vocalizations, transmitter channels did not. We quantified the separation of overlapping vocalizations by visual inspection of acceleration spectrograms from both birds. We found that 3.3% (range 1.9%-10.8%, n=2 bird pairs) of vocalizations showed an overlap with vocalizations in the other bird. Among these overlapping vocalizations, 91.7% (88.9%-100%, n=2 bird pairs) were associated with mixed vocal signals (from both birds) on a given microphone channel and 0.0% were associated with mixed vocal signals on a given transmitter channel. Because mixed vocal signals prevented the straightforward analysis of vocalizations from individual animals, the clean vocal separation provided by accelerometers demonstrates the expected benefit of the animal-borne sensors.

Discussion

We designed and validated a behavioral recording system for up to 8 songbirds, yielding synchronized multimodal data. Our custom diversity combining multi-antenna demodulation technique increases RSNR by 6.5 dB and reduces signal-loss events by a factor of 63 compared with single-antenna demodulation. The wireless devices transmit well-separated vocalizations unless these are masked by large body movements such as wing flaps or because of weak mechanical coupling between the bird and the device, which we were unable to completely eliminate.

While in isolated, non-naturalistic settings, the detection and annotation of vocal segments on a single microphone channel is relatively straight forward, it is much more difficult in social settings and naturalistic environments. We segmented vocal activity by visual inspection of sound and acceleration spectrograms from five microphone and two animal-borne accelerometer channels. With respect to these multimodal observations, we find that no sensor channel by itself achieves a miss rate of less than 1%, with an average of 3.7%, which is a large number given that we took our measurements in ideal settings using the minimal group size of two birds housed in a relatively small, acoustically well isolated environment. Since we might have missed some vocalizations even considering all the channels used, our estimated single-channel miss rates constitute lower bounds, implying that the true miss rate of vocalizations must be higher than what we measured.

By combining transmitter and microphone channels, we showed that vocalization miss rates can be substantially reduced. Currently, there is no reference for miss rates, as to our knowledge, these have not been quantified in previous studies of vocal communication in songbird groups[24,28,29]. Being an important parameter, ideally, the miss rate should be evaluated relative to ground-truth, i.e., the true

segmentation of vocalizations and background noise. Yet, such a ground-truth data set does not (yet) exist. For example, to perfectly measure the vocal output of a bird would require simultaneous measurements of syringeal labia and muscle activity, of sub-syringeal air pressure, and of tracheal air flow[37], which are measurements that have not been performed simultaneously in freely moving group-housed birds. Without such a data set or at least an approximation thereof, quantifications of miss rates will be biased towards zero.

On the question about acceptable miss rates, we anticipate that miss rates scale with the number of animals examined and will depend on the studied behaviors. Also, the maximally admissible miss rate will depend on the specific research question to be examined, which is why future studies will have to revisit this important parameter.

To enable automatic segmentation of longitudinal BirdPark data, new methods will need to be developed. While there are many methods for single channel segmentation available[38–40], only few combine information from multiple sensor channels[40] and none has made use of accelerometer data so far.

Also, it remains open how well vocalizations can be separated from the five stationary microphones alone. In groups of mice and using microphone arrays and a camera, it has been shown that vocalizations can be assigned to individual animals by combining sound source localization and animal position tracking[41]. But because of frequent vocal jamming in songbirds, we doubt these approaches will make the use of animal-borne transmitters obsolete in song-learning studies to be performed in the BirdPark where subtle song changes need to be tracked in developmental time. Our datasets will be useful for the development and evaluation of such methods, as the transmitter and video channels yield well-resolved information at the individual level.

Our system could promote research on the meaning of vocalizations during social behaviors. Of key interest are courtship and reproductive behaviors that have been thoroughly studied[42–45] albeit often without examining the roles of vocal interactions during mating and pair bonding. Similarly, multimodal group-level studies in the BirdPark could help us better understand the learning strategies young birds use while they modify their immature vocalizations to match a sensory target provided by a tutor. Much of our knowledge on song learning stems from research of isolated animals in impoverished environments[46–48], leaving many questions open about the roles of social interactions[49,50] during this process, including by non-singing adult females[51].

Methods

The chamber

All our experiments were performed in a sound-isolation chamber (Industrial Acoustic Company, UK) of dimensions 124 cm (width) x 90 cm (depth) x 130 cm (height). The isolation chamber has a special silent ventilation system for air exchange at a rate of roughly 500 l/min. Two types of light sources made of light emitting diodes (LEDs, Waveform Lighting) are mounted on aluminum plates on the ceiling of the chamber: 1) LEDs with natural frequency spectrum (consuming a total electric power of 80 W); and 2) ultraviolet (UV) LEDs of wavelength 365 nm and consuming a total electric power of 13 W.

The ceiling of the chamber contains three circular holes of 7 cm diameter through which we inserted aluminum tubes of 5 mm wall thickness that serve as holders for three cameras (Basler acA2040-120uc). The tubes also conduct the heat of the LEDs (93 W) and the cameras (13 W) to the outside of the chamber where we placed silent fans to keep the cameras below 55 deg C.

With the cameras, we filmed the arena directly from the top and indirectly from two sides via two tilted mirrors in front of the glass side panels. This setup yields an object distance of about 1 – 1.5 m, which allows for a depth of field that covers the entire arena. Furthermore, the large object distance results in a small perspective distortion.

We installed four microphones on the four side walls of the isolation chamber and one microphone on the ceiling. We mounted two more microphones in the nest boxes and one microphone outside the isolation chamber. Where possible, we covered the sidewalls of the chamber with sound-absorbing foam panels. A door sensor measures whether the door of the chamber is open or closed. The temperature inside the chamber was roughly 26 ° C and the humidity was roughly 24 %.

The arena

Inside the chamber, we placed the bird arena of dimensions 90 cm x 55 cm (floor). To minimize acoustic resonances and optical reflections, the 40 cm high side panels of the arena are tilted by 11 ° and 13 ° toward the inside. Two of the side panels are made of glass for videography and the two opposite panels are made of perforated plastic plates. The floor of the arena is covered with a sound-absorbing carpet. A pyramidal tent made of a fine plastic net covers the upper part of the arena, reaching up to 125 cm above ground. At a height of 35 cm, we attached two nest boxes to the side panels, each equipped with a microphone and a camera. Furthermore, the arena is equipped with perches, a sand-bath, and a food and water tray.

Video acquisition system

To visualize the arena, we used industrial cameras (3 Megapixel, MP) with zoom lenses (opening angles: top view: 45 ° x 26 °, back view: 55 ° x 26 °, side view: 26 ° x 35 °) and exposure times of 3 ms. To visualize nests in the dimly lit nest boxes, we used monochrome infrared cameras (2 MP, Basler daA1600-60um) and fisheye lenses (143 ° x 112 °).

The uncompressed camera outputs (approx. 400 MB/s in total) are relayed via a host computer via a USB3 Vision interface. Each camera receives frame trigger signals generated by the USRP (Figure 1C). A custom program (BirdVideo), written in C++ and using the OpenCV library[52] and FFMPEG[53], undistorts the nest camera images with a fisheye lens model and transforms the five camera images into a single 2976 x 2016 pixel-sized image (Figure 1B). The composite images are then encoded with the h264 codec on a NVIDIA GPU and stored in an MP4 (ISO/IEC 14496-14) file. We used constant-quality compression with a variable compression ratio in the range 150-370, resulting in a data rate of 0.72 MB/s – 1.8 MB/s, depending on the number of birds and their activity in the arena. Compression ratios in this range do not significantly decrease key point tracking performance[54]. The frame rate of the video is about 48 frames per second (frame period 21 ms). The spatial resolution of the main cameras is about 2.2 pixels/mm.

Transmitter device

The accelerometer (Knowles BU21771) senses acceleration with a sensitivity of 1.4 mV/m/s^2 in the 30 Hz – 6 kHz frequency range[55]. Inspired by Ter Maat et al.[24], we performed the frequency modulation of the

accelerometer signal onto a radio carrier using a simple Hartley oscillator transmitter stage with only one transistor and a coil that functions both as the inductor in the LC resonator and as the antenna. The LC resonator determines the carrier frequency ω_c of the radio signal. We set ω_c of our transmitters in the vicinity of 300 MHz, which corresponds to an electromagnetic wavelength of about 1 m and is roughly equal to the dimension of our soundproof chambers, implying that our radio system operates in the near field.

We found that ω_c depends on temperature at an average of -73 kHz/°C (range -67 kHz/°C to -87 kHz/°C, n=3 transmitters). Towards the end of the battery life (over the course of the last 3 days), we observed an increase of ω_c by an average of 500 – 800 kHz. These slow drifts can easily be accounted for by tracking and high-pass filtering the momentary frequency. The measured end-to-end sensitivity of the frequency modulation is 5 kHz/g, with g being the gravitational acceleration constant. The transmitters are powered by a zinc-air battery (type LR41), which lasts at least 12 days. The total weight of the transmitter including the harness and battery is 1.5 g, which is ca. 10% of the body weight of a zebra finch.

We moderately tightened the harness during attachment in a tradeoff between picking up vibratory signals from the singer and preserving the wearer's natural behaviors. It is known that the act of mounting devices can transiently affect zebra finch behavior: right after attachment, the singing rate and amount of locomotion both tend to decrease, they return to baseline within 3 days for sensors weighing 0.75 g[56] and within 2 weeks for sensors weighing 3 g[57].

Radio reception

We mounted four whip antennas of 30 cm length (Albrecht 6157) perpendicular to the metallic sidewalls and ceiling of the chamber (using magnetic feet). We fed the antenna signals to a universal software radio interface (USRP-2940, National Instruments, USA), which comprises four independent antenna amplifiers the gains of which we set to 68 dB, adjustable in the range -20 dB to +90 dB.

Intermediate band

The USRP generates from the amplified radio signal a down-converted signal with an analog bandwidth of 80 MHz around the local oscillator frequency (ω_{LO}). After digitization, this intermediate signal $z^a(t)$ for antenna $a \in \{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}\}$, is a 100 MS/s complex-valued signal (IQ) of 2 x 18 bits precision and is relayed to the FPGA. For details about the analog down conversion and the sampling of complex valued signals, see the NI manual of the USRP and the documentation of our custom radio software.

Digital downconverters (DDCs) and baseband

Every PLL, $i \in \{1, \ldots, 8\}$, makes use of four digital downconverters (DDCs, Figure 3D), one for each antenna a, that cut from the intermediate band a narrow baseband around the tracking frequency ω_i. The 3-stage decimation filter of this down-conversion has a flat frequency response within a ± 100 kHz band. Given the decimation factor of 128, the resulting complex base-band signals $u_i^a(t)$ (the outputs of the DDCs) have a sample rate of 781.25 kHz and a precision of 2 x 25 bits.

Phase locked loop

Each PLL generates its tracking frequency ω_i by direct digital synthesis (DDS) with a 48-bit phase accumulation register and a lookup table. This frequency is dynamically adjusted to keep the phase of the main vector (i.e., the summed baseband signals) close to zero. We calculated the angle $\theta_i =$

$\arg(u_i^{\mathrm{M}})$ of the main vector with the CORDIC algorithm[58] and unwrapped the phase up to ± 128 turns before using it as the error signal for a proportional-integral-derivative (PID) controller. The PID controller adjusts the tracking frequency of the PLL.

The PID controller was implemented on the FPGA in single precision floating point (32-bit) arithmetic. The controller included a limiting range and an anti-windup mechanism[59]. The unwrapping of the error phase was crucial for the PLL to quickly lock-in and to keep the lock even during large and fast frequency deviations of the transmitter. To tune the PID parameters, we measured the closed-loop transfer function of the PLL by adding a white-noise signal to the control signal (tracking frequency), and then adjusted the PID parameters until we observed a low-pass characteristic without overshoot. We achieved a closed loop bandwidth of about 30 kHz.

Phase compensation

For a given PLL i, we compensated the relative phases under which a radio signal arrives at the four antennas a in order to align the four baseband vectors u_i^a. The alignment was achieved by providing a phase offset $\Delta\varphi_i^a$ to each downconverter, where it acts as offset to the phase accumulation register of the direct digital synthesis. To compute the angle of baseband vector relative to the main vector, we rotated the baseband vectors u_i^a by the phase θ_i of the main vector to result in the rotated vector $r_i^a = u_i^a e^{-i\theta_i}$. After averaging the rotated vectors across 512 samples, we computed its angle $\alpha_i^a = \arg(\langle r_i^a \rangle_{512})$. We then compensated that angle by iteratively adding a fraction $\gamma \in [0,1]$ of it to the phase offset: $\Delta\varphi_i^a \coloneqq \Delta\varphi_i^a + \gamma\alpha_i^a$. The parameter γ is the phase compensation gain (Figure 3D), typically set to $\gamma = 0.2$. This iterative update was performed at a rate of about 1.5 kHz (781.25 kHz / 512), which is faster than birds' locomotion (changes in physical position or orientation of the transmitter).

Central control software

On the host computer we run our central control software (BirdRadio) programmed with LabView that acquires the microphone and transmitter signals and writes them to a TDMS file (Figure 1C). Furthermore, BirdRadio sends UDP control signals to BirdVideo, it automatically starts the recording in the morning and stops it in the evening, it controls the light dimming in the sound-isolation chamber in simulation of sunrise and sunset, it triggers an email alarm when the radio signal from a transmitter device is lost, and it automatically adjusts the center frequency of each radio channel every morning to adjust for carrier frequency drift.

Data management

The BirdPark is designed for continuous recordings over multiple months, producing data at a rate of 60 GB/day for 2 birds and 130 GB/day for 8 birds. We implemented the FAIR (findable, accessible, interoperable, and reusable) principles of scientific data management[60] as follows:

Our recording software splits the data into gapless files of 20'000 video frames (ca. 7 mins duration). At the end of a recording day, all files are processed by a script that converts the TDMS files into HDF5 files and augments them with rich metadata. The HDF5 files are self-descriptive in that they contain metadata as attributes and additionally, every dataset in the file contains its description as an attribute. We use the lossless compression feature of HDF5 to obtain a compression ratio of typically 2.5 for the audio and accelerometer data. The script also adds two AAC-compressed microphone channels into the video files. Although this step introduces redundancy, the availability of sound in the video files is very useful during manual annotation of the videos. Furthermore, the script also exports the metadata as a JSON file and

copies the processed data onto a NAS server. At the end of an experiment, the metadata is uploaded onto an openBIS[61] server and is linked with the datafiles on the NAS.

Manual segmentation of vocalizations

We manually segmented vocalizations in recordings of mixed-sex zebra finch pairs (7 mins recordings of n=2 bird pairs). We high-pass filtered (FIR filter with a stopband frequency of 100 Hz and passband frequency of 200 Hz) the raw demodulated transmitter signals and microphone signals and produced spectrograms (window size=384 samples, hop size=96 samples) that we manually segmented using Raven Pro 1.6[62]. We performed three types of segmentations in the following order:

1. Transmitter-based vocal segments: We separately segmented all vocalizations on either transmitter channel, precisely annotating their onsets and offsets. In our released data sets, these segments are referred to as transmitter-based vocal segments. When we were uncertain whether a sound segment was a vocalization or not, we also looked at spectrograms of microphone channels and listened to sound playbacks. If we remained uncertain, the segment was tagged with the label 'Unsure', such segments were treated as (non-vocal) noises and were excluded from further analysis.
2. Microphone-based vocal segments: We simultaneously visualized all microphone spectrograms (from Mic1 to Mic6) using the multi-channel view in Raven Pro (we ignored Mic7, which was located in the second nest that was not accessible to the birds). On those, we annotated each vocal segment on the first microphone channel on which it was visible (e.g., a syllable that is visible on all microphone channels is only annotated on Mic1). Overlapping vocalizations were annotated as a single vocal segment. When we were uncertain whether a sound segment was a vocalization or not, we also looked at spectrograms of accelerometer channels and listened to sound playbacks. If we remained uncertain, the segment was tagged with the label 'Unsure', such segments were treated as (non-vocal) noises and were excluded from further analysis.
3. Consolidated vocal segments: All consistent (perfectly overlapping) transmitter- and microphone-based vocal segments, we labelled as consolidated vocal segments. We then inspected all inconsistent (not perfectly overlapping) segments by visualizing all channel spectrograms. We fixed inconsistencies that were caused by human annotation errors (e.g. lack of attention) by fixing the erroneous or missing transmitter- and microphone-based segments. From the inconsistent (partially overlapping) segments that were not caused by human error, we generated one or several consolidated segments by trusting the modality that more clearly revealed the presence of a vocalization (hence our reporting of 'misses' in Table 1).
   In our released comma-separated value (CSV) files, we give each consolidated vocal segment a Bird Tag (e.g., either 'b15p5_m' or 'b14p4_f') that identifies the bird that produced the vocalization, a Transmitter Tag that identifies the transmitter channel on which the vocalization was identified (either 'b15p5_m' or 'b14p4_f' or 'None'), and a FirstMic Tag that identifies the first microphone channel on which the segment was visible ('Mic1' to 'Mic6', or 'None'). We resolved inconsistencies and chose these tags as follows:
   - If a microphone (-based vocal) segment was paired (partially overlapping) with exactly one transmitter segment, a consolidated segment was generated with the onset time set to the minimum onset time and the offset time set to the maximum offset time of the segment pair. The Bird and Transmitter Tags were set to the transmitter channel name, and the FirstMic Tag was set to the microphone channel name.

- If a transmitter segment was unpaired, a consolidated segment was created with the same onset and offset times. The Bird and TrCh Tags were set to the transmitter channel name, and the FirstMic Tag was set to 'None'.
- If a microphone segment was unpaired, a transmitter channel name was inferred based on the vocal repertoire and noise levels on both transmitter channels. We visually verified that the vocal segment was not the result of multiple overlapping vocalizations (which was never the case). Then we created a consolidated segment with the same on- and offset, set the FirstMic Tag to the microphone Id, the Bird Tag to the guessed transmitter channel name, and the Transmitter Tag to 'None'.
- If a microphone segment was paired with more than one transmitter segment, a consolidated segment was created for each of the transmitter segments. The onsets and offsets were manually set based on visual inspection of all spectrograms. Bird and Transmitter Tags were set to the transmitter channel name and the FirstMic Tag was set to the first microphone channel on which the respective segment was visible.
- We never encountered the case where a transmitter segment was paired with multiple microphone segments.

The statistics in Table 1 were calculated as follows:

vocalizations = (number of consolidated segments)

voc. missed on tr. channel = (number of consolidated segments with Transmitter=None)

voc. missed on all mic. channels = (number of consolidated segments with FirstMic=None)

voc. missed on Mic1 = (number of consolidated segments with FirstMic≠Mic1 and FirstMic≠None)

voc. with overlaps = (number of consolidated segments which overlap with at least one other syllable by at least one spectrogram bin)

The row 'of which are associated with mixed vocal signal on FirstMic' quantifies the number of vocal segments with overlaps that were associated with overlapping (mixed) vocal signals from both birds on the microphone channel specified by the FirstMic Tag of the respective syllable. On all vocal segments with overlaps (as defined above), we visually assessed whether the overlapping vocalization produced by the other bird was visible in the FirstMic spectrogram (implying the microphone signal is corrupted).

The row 'of which are associated with mixed vocal signal on other transmitter ch.' quantifies the number of overlapping segments that were associated with signal leakage of the vocalization into the transmitter channel of the other bird. For all vocal segments with overlaps, we visually assessed whether the vocal signal leaked into the transmitter spectrogram of the other bird (implying the transmitter signal is corrupted).

The assignment to the female/male column was determined by the 'Bird' tag.

Frequency jumps

We detected frequency jumps as deviations of the transmitter tracking frequency ω_i larger than 50 kHz from the running median computed within 0.2 s sliding windows. Frequency jumps were concatenated

when they were separated by less than 50 ms. The modulation amplitude was defined as the maximum absolute deviation.

Animals and Experiments

Mixed-sex zebra finch (Taeniopygia guttata) pairs bred and raised in our colony (University of Zurich and ETH Zurich) were kept in the BirdPark on a 14/10 h light/dark daily cycle for multiple days, with food and water provided ad libitum. All experimental procedures were approved by the Cantonal Veterinary Office of the Canton of Zurich, Switzerland (license numbers ZH054/2020). All methods were carried out in accordance with relevant guidelines and regulations (Swiss Animal Welfare Act and Ordinance, TSchG, TSchV, TVV).

References

1. Lipkind, D. *et al.* Songbirds work around computational complexity by learning song vocabulary independently of sequence. *Nat Commun* **8**, 1–11 (2017).

2. Tchernichovski, O., Mitra, P. P., Lints, T. & Nottebohm, F. Dynamics of the vocal imitation process: How a zebra finch learns its song. *Science (1979)* **291**, 2564–2569 (2001).

3. Kollmorgen, S., Hahnloser, R. H. R. & Mante, V. Nearest neighbours reveal fast and slow components of motor learning. *Nature* **577**, 526–530 (2020).

4. Segalin, C. *et al.* The mouse action recognition system (MARS) software pipeline for automated analysis of social behaviors in mice. *Elife* **10**, (2021).

5. Fujimori, S., Ishikawa, T. & Watanabe, H. Animal behavior classification using DeepLabCut. in *2020 IEEE 9th Global Conference on Consumer Electronics, GCCE 2020* 254–257 (IEEE, 2020). doi:10.1109/GCCE50665.2020.9291715.

6. Perkes, A., Pfrommer, B., Daniilidis, K., White, D. J. & Schmidt, M. Variation in female songbird state determines signal strength needed to evoke copulation. *bioRxiv* 2021.05.19.444794 (2021) doi:10.1101/2021.05.19.444794.

7. Nath, T. *et al.* Using DeepLabCut for 3D markerless pose estimation across species and behaviors. *Nat Protoc* **14**, 2152–2176 (2019).

8. Pereira, T. D. *et al.* SLEAP: Multi-animal pose tracking. *BioRxiv* (2020) doi:10.1101/2020.08.31.276246.

9. Mathis, A., Schneider, S., Lauer, J. & Mathis, M. W. A Primer on Motion Capture with Deep Learning: Principles, Pitfalls, and Perspectives. *Neuron* vol. 108 44–65 Preprint at https://doi.org/10.1016/j.neuron.2020.09.017 (2020).

10. Walter, T. & Couzin, I. D. Trex, a fast multi-animal tracking system with markerless identi cation, and 2d estimation of posture and visual elds. *Elife* **10**, 1–73 (2021).

11. Badger, M. *et al.* 3D Bird Reconstruction: A Dataset, Model, and Shape Recovery from a Single View. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (eds. Vedaldi, A., Bischof, H., Brox, T. & Frahm, J.-M.) vol. 12363 LNCS 1–17 (Springer International Publishing, 2020).

12. Harley, A. W., Fang, Z. & Fragkiadaki, K. Particle Video Revisited: Tracking Through Occlusions Using Point Trajectories. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **13682 LNCS**, 59–75 (2022).

13. Doersch, C. *et al.* TAPIR: Tracking Any Point with per-frame Initialization and temporal Refinement. (2023).

14. Naik, H. *et al.* 3D-POP - An Automated Annotation Approach to Facilitate Markerless 2D-3D Tracking of Freely Moving Birds With Marker-Based Motion Capture. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 21274–21284 (2023).

15. Stowell, D., Benetos, E. & Gill, L. F. On-Bird Sound Recordings: Automatic Acoustic Recognition of Activities and Contexts. *IEEE/ACM Trans Audio Speech Lang Process* **25**, 1193–1206 (2017).

16. Barchiesi, D., Giannoulis, D. D., Stowell, D. & Plumbley, M. D. Acoustic Scene Classification: Classifying environments from the sounds they produce. *IEEE Signal Process Mag* **32**, 16–34 (2015).

17. Meen, G. H. *et al.* Sound analysis in dairy cattle vocalisation as a potential welfare monitor. *Comput Electron Agric* **118**, 111–115 (2015).

18. Pouw, W., Trujillo, J. P. & Dixon, J. A. The quantification of gesture–speech synchrony: A tutorial and validation of multimodal data acquisition using device-based and video-based motion tracking. *Behav Res Methods* **52**, 723–740 (2020).

19. Rutz, C. *et al.* Using machine learning to decode animal communication. *Science* **381**, 152–155 (2023).

20. Zimmerman, P. H., Bolhuis, J. E., Willemsen, A., Meyer, E. S. & Noldus, L. P. J. J. The observer XT: A tool for the integration and synchronization of multimodal signals. *Behav Res Methods* **41**, 731–735 (2009).

21. Dolmans, T. C., Poel, M., Klooster, J. & Veldkamp, B. P. *Data Synchronisation and Processing in Multimodal Research*. *Measuring Behavior 2020-21 Volume* vol. 1 (2021).

22. Anisimov, V. N. *et al.* Reconstruction of vocal interactions in a group of small songbirds. *Nat Methods* **11**, 1135–1137 (2014).

23. Eisenring, E. *et al.* Quantifying song behavior in a free-living, light-weight, mobile bird using accelerometers. *Ecol Evol* **12**, e8446 (2022).

24. Maat, A. Ter, Trost, L., Sagunsky, H., Seltmann, S. & Gahr, M. Zebra finch mates use their forebrain song system in unlearned call communication. *PLoS One* **9**, e109334 (2014).

25. Laich, A. G., Wilson, R. P., Quintana, F. & Shepard, E. L. C. Identification of imperial cormorant Phalacrocorax atriceps behaviour using accelerometers. *Endanger Species Res* **10**, 29–37 (2010).

26. Watanabe, S., Izawa, M., Kato, A., Ropert-Coudert, Y. & Naito, Y. A new technique for monitoring the detailed behaviour of terrestrial animals: A case study with the domestic cat. *Appl Anim Behav Sci* **94**, 117–131 (2005).

27. Gerencsér, L., Vásárhelyi, G., Nagy, M., Vicsek, T. & Miklósi, A. Identification of Behaviour in Freely Moving Dogs (Canis familiaris) Using Inertial Sensors. *PLoS One* **8**, e77814 (2013).

28. Gill, L. F. *et al.* A minimum-impact, flexible tool to study vocal communication of small animals with precise individual-level resolution. *Methods Ecol Evol* **7**, 1349–1358 (2016).

29. Gill, L. F., Goymann, W., Maat, A. Ter & Gahr, M. Patterns of call communication between group-housed zebra finches change during the breeding cycle. *Elife* **4**, (2015).

30. Tse, D. & Pramod, V. *Fundamentals of Wireless Communication*. (Cambridge University Press, 2005).

31. Shatara, R. S. Combined Switched and Phase Aligned Multi-Antenna Diversity System for Signal-Error-Reduction in Mobile Receiving Systems. (Universität der Bundeswehr München, 2003).

32. Voitsun, O., Senega, S. & Lindenmeier, S. Multi-Antenna Diversity Set for Transmission and Reception in Car-to-Car and Car-to-X Communication. in *GeMIC 2020 - Proceedings of the 2020 German Microwave Conference* 80–83 (2020).

33. Senega, S., Nassar, A. & Lindenmeier, S. Automotive antenna diversity system for satellite radio with high phase accuracy in low SNR-scenarios. *Clay Miner* **10**, 578–586 (2018).

34. Haniz, A. *et al.* A Novel Phase-Difference Fingerprinting Technique for Localization of Unknown Emitters. *IEEE Trans Veh Technol* **66**, 8445–8457 (2017).

35. Berdanier, C. A. & Wu, Z. A novel RF emitter localization method through phase information. *IEEE National Radar Conference - Proceedings* 1–5 (2013) doi:10.1109/RADAR.2013.6586137.

36. Alarcón-Nieto, G. *et al.* An automated barcode tracking system for behavioural studies in birds. *Methods Ecol Evol* **9**, 1536–1547 (2018).

37. Goller, F. & Suthers, R. A. Role of syringeal muscles in gating airflow and sound production in singing brown thrashers. *J Neurophysiol* **75**, 867–876 (1996).

38. Cohen, Y. *et al.* Automated annotation of birdsong with a neural network that segments spectrograms. *Elife* **11**, (2022).

39. Lorenz, C., Hao, X., Tomka, T., Rüttimann, L. & Hahnloser, R. H. R. Interactive extraction of diverse vocal units from a planar embedding without the need for prior sound segmentation. *Frontiers in Bioinformatics* **2**, (2022).

40. Steinfath, E., Palacios-Muñoz, A., Rottschäfer, J. R., Yuezak, D. & Clemens, J. Fast and accurate annotation of acoustic signals with deep neural networks. *Elife* **10**, (2021).

41. Oliveira-Stahl, G. *et al.* High-precision spatial analysis of mouse courtship vocalization behavior reveals sex and strain differences. *Scientific Reports 2023 13:1* **13**, 1–18 (2023).

42. Morris, D. The Reproductive Behaviour of the Zebra Finch (Poephila Guttata), With Special Reference To Pseudofemale Behaviour and Displacement Activities. *Behaviour* **6**, 271–322 (2008).

43. Morris, D. The comparative ethology of Grasfinches (Erythrurae) and Mannikins (Amadinae). *Proceedings of the Zoological Society of London* **131**, 389–439 (1958).

44. Caryl, P. G. Sexual behaviour in the zebra finch Taeniopygia guttata: response to familiar and novel partners. *Anim Behav* **24**, 93–107 (1976).

45. Ullrich, R., Norton, P. & Scharff, C. Waltzing Taeniopygia: Integration of courtship song and dance in the domesticated Australian zebra finch. *Anim Behav* **112**, 285–300 (2016).

46. Kollmorgen, S., Hahnloser, R. H. R. & Mante, V. Nearest neighbours reveal fast and slow components of motor learning. *Nature 2020 577:7791* **577**, 526–530 (2020).

47. Lipkind, D. *et al.* Songbirds work around computational complexity by learning song vocabulary independently of sequence. *Nature Communications 2017 8:1* **8**, 1–11 (2017).

48. Tchernichovski, O., Mitra, P. P., Lints, T. & Nottebohm, F. Dynamics of the vocal imitation process: How a zebra finch learns its song. *Science (1979)* **291**, 2564–2569 (2001).

49. Chen, Y., Matheson, L. E. & Sakata, J. T. Mechanisms underlying the social enhancement of vocal learning in songbirds. *Proc Natl Acad Sci U S A* **113**, 6641–6646 (2016).

50. Carouso-Peck, S. & Goldstein, M. H. Female Social Feedback Reveals Non-imitative Mechanisms of Vocal Learning in Zebra Finches. *Current Biology* **29**, 631-636.e3 (2019).

51. Takahashi, D. Y., Liao, D. A. & Ghazanfar, A. A. Vocal Learning via Social Reinforcement by Infant Marmoset Monkeys. *Current Biology* **27**, 1844-1852.e6 (2017).

52. Gary Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools* (2000).

53. Tomar, S. Converting Video Formats with FFmpeg | Linux Journal. *Linux Journal* **2006**, 10 (2006).

54. Mathis, A. & Warren, R. On the inference speed and video-compression robustness of DeepLabCut. *bioRxiv* 457242 (2018) doi:10.1101/457242.

55. Knowles Electronics. TB-26: The use of BU series accelerometers. Preprint at (2017).

56. Gill, L. F. *et al.* A minimum-impact, flexible tool to study vocal communication of small animals with precise individual-level resolution. *Methods Ecol Evol* **7**, 1349–1358 (2016).

57. Anisimov, V. N. *et al.* Reconstruction of vocal interactions in a group of small songbirds. *Nat Methods* **11**, 1135–1137 (2014).

58. Volder, J. E. The CORDIC Trigonometric Computing Technique. *IRE Transactions on Electronic Computers* **EC-8**, 330–334 (1959).

59. Astrom, K. J. & Rundqwist, L. Integrator windup and how to avoid it. in 1693–1698 (Publ by IEEE, 1989). doi:10.23919/acc.1989.4790464.

60. Wilkinson, M. D. *et al.* Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 1–9 (2016).

61. Bauch, A. *et al.* OpenBIS: A flexible framework for managing and analyzing complex data in biology research. *BMC Bioinformatics* **12**, 1–19 (2011).

62. K. Lisa Yang Center for Conservation Bioacoustics at the Cornell Lab of Ornithology. Raven Pro: Interactive Sound Analysis Software (Version 1.5). *Ithaca, NY: The Cornell Lab of Ornithology.* http://www.birds.cornell.edu/raven Preprint at https://ravensoundsoftware.com (2014).

63. Rüttimann, L. *et al.* Behavioral recordings of mixed-sex zebra finch pairs with vocal segment annotations. (2022) doi:10.5281/ZENODO.7105196.

Acknowledgments

We thank Aymeric Nager for his help with the design and construction of the BirdPark and Moritz Wohlhauser for creating the illustration in Figure 2D. Financial support was provided by Swiss national Science Foundation 31003A_182638 (The roles of vocal communication in pair formation and cultural learning in songbirds) and the Swiss National Science Foundation NCCR Evolving Language (Agreement #51NF40_180888).

Author contributions

Conceptualization, J.R., L.R., M.D.R. and R.H.R.H.; System and software development, L.R. and J.R.; animal experiments, L.R. and H.H.; data annotation and data analysis, T.T. and L.R.; writing—original draft, J.R., L.R. and R.H.R.H.; writing—reviewing and editing, J.R., L.R., T.T., M.D.R. and R.H.R.H.

Data and code availability

The raw data files (MP4 videos and HDF5 files that contain the transmitter and microphone data) and the vocal segment annotation files are available for download on Zenodo[63]. This data may be useful for benchmarking machine learning systems on the task of vocal segmentation. The source code of the recording software (BirdRadio and BirdVideo) will be made available for download on publication of this manuscript.

Competing interests

The author(s) declare no competing interests.

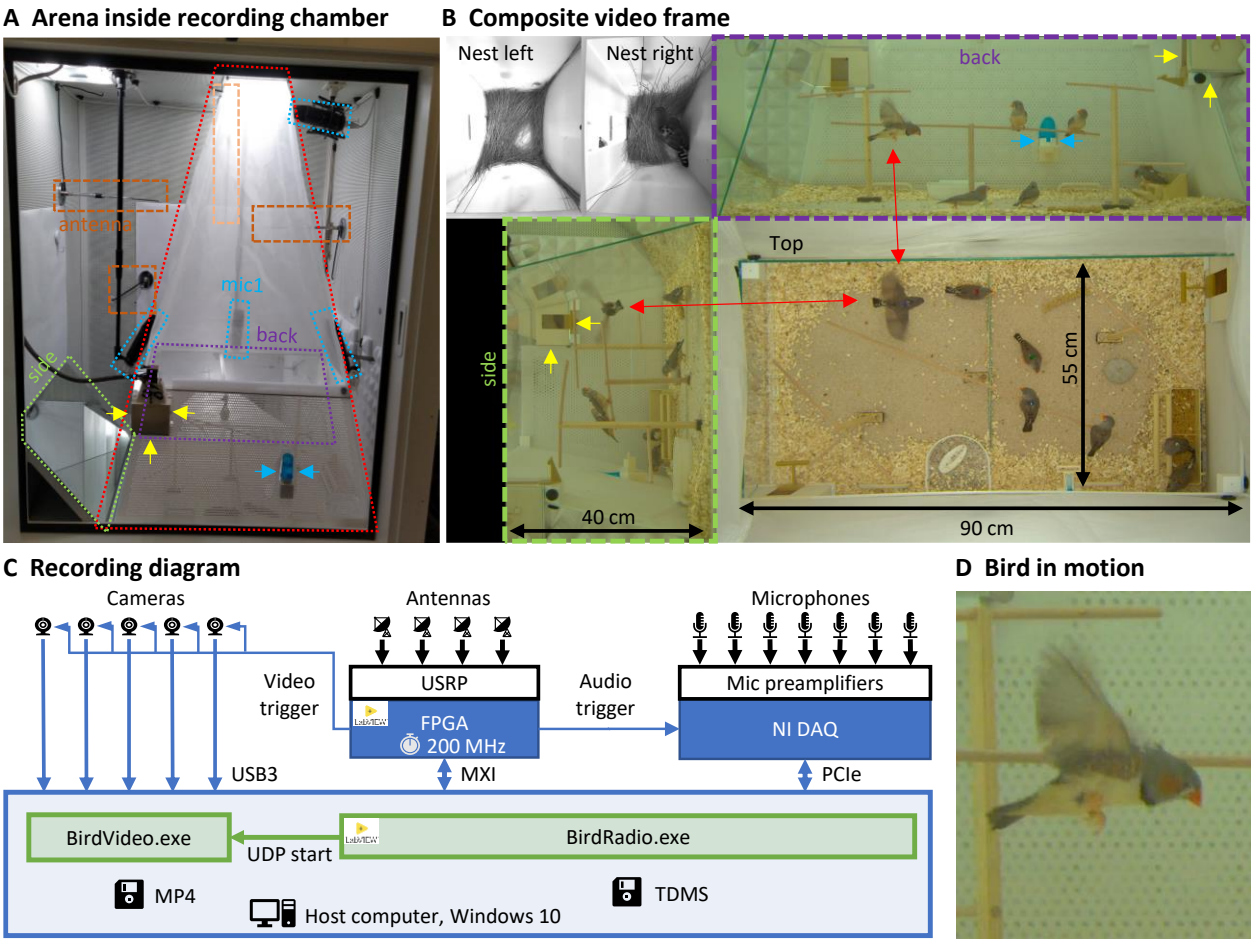


*Figure 1: Recording arena and recording system schematic. (**A**) Inside a soundproof chamber, we built a recording arena (red dotted line) for up to 8 birds. We record the animals' behaviors with three cameras mounted through the ceiling. These provide a direct top view and indirect side and back views via two mirrors (delimited by green and magenta dotted lines). To record the sounds in the chamber, we installed five microphones (blue dotted lines) among all four sides of the cage (one attached to the front door is not visible) and the ceiling, and two small microphones in the nest boxes. The radio signals from the transmitter devices are received with four radio antennas (orange dotted lines) mounted on three side walls and the ceiling. One nest box is indicated with yellow arrows and a water bottle with blue arrows. (**B**) A composite still image of all camera views shows two monochrome nest box views (top left) and three views of the arena (top, side, back) with 8 birds among which one is flying (red arrows). Yellow and blue arrows as in A. (**C**) Schematic of the recording system for gapless and synchronized recording of sound (microphones), acceleration (transmitter devices), and video channels (cameras). The radio receiver is implemented on a universal software radio peripheral (USRP) with a large field programmable gate array (FPGA) that runs at the main clock frequency of 200 MHz. Clock dividers on the FPGA provide the sample trigger for audio recordings and the frame trigger for the cameras. The data streams are collected on a host computer that runs two custom programs, one (BirdRadio) for streaming audio and sensor signals to disk and one (BirdVideo) for encoding video data. (**D**) Zoom-in on an airborne bird, illustrating the spatial and temporal resolution of the camera.*

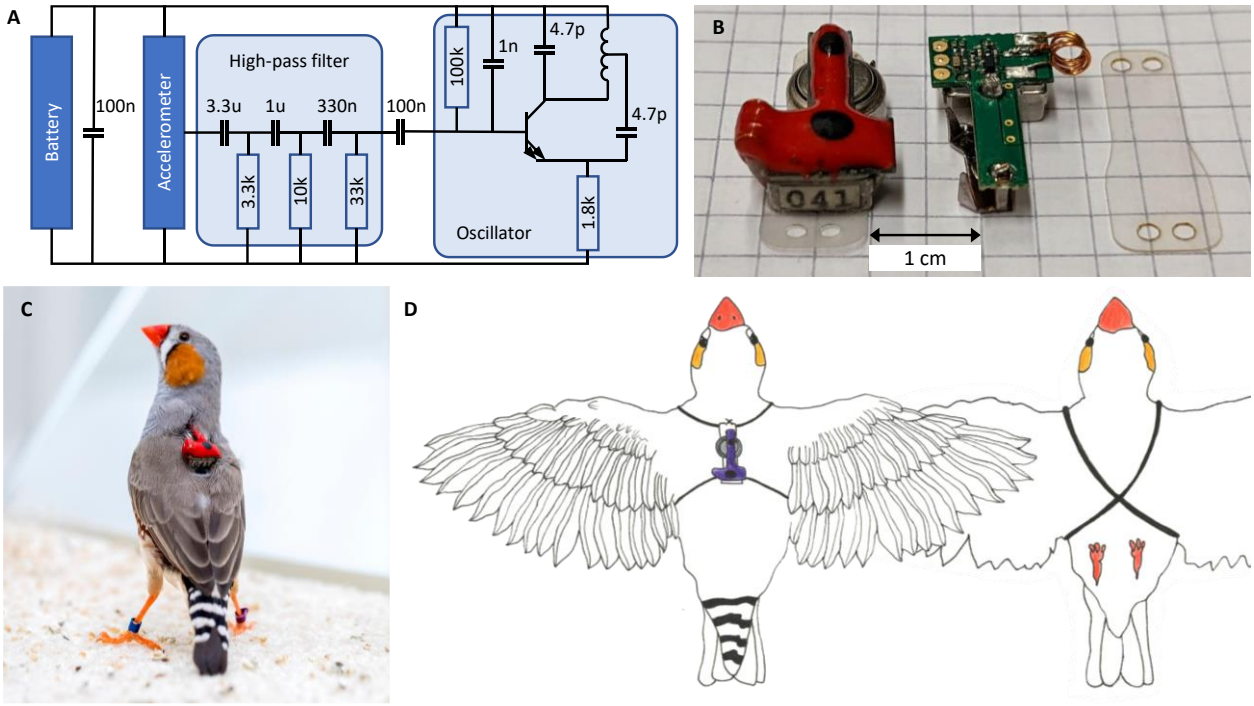


*Figure 2: Transmitter device. (**A**) Schematic of the electronic circuit (adapted from Ter Maat et al.[24]). The analog FM radio transmits the vibration transducer (accelerometer) signal via a high-pass filter (cutoff frequency: 15Hz) followed by a radiating oscillator. (**B**) A fully assembled transmitter (left), and another one without epoxy and battery (middle), and a piece of mounting foil (right). (**C**) Picture of the device mounted on a bird. The transmitters are color-coded to help identify the birds in the video data. (**D**) Schematic of a bird wearing the device. The harness (adapted from Alarcón-Nieto et al.[36]) is made of rubber string (black) that goes around the wings and the chest.*

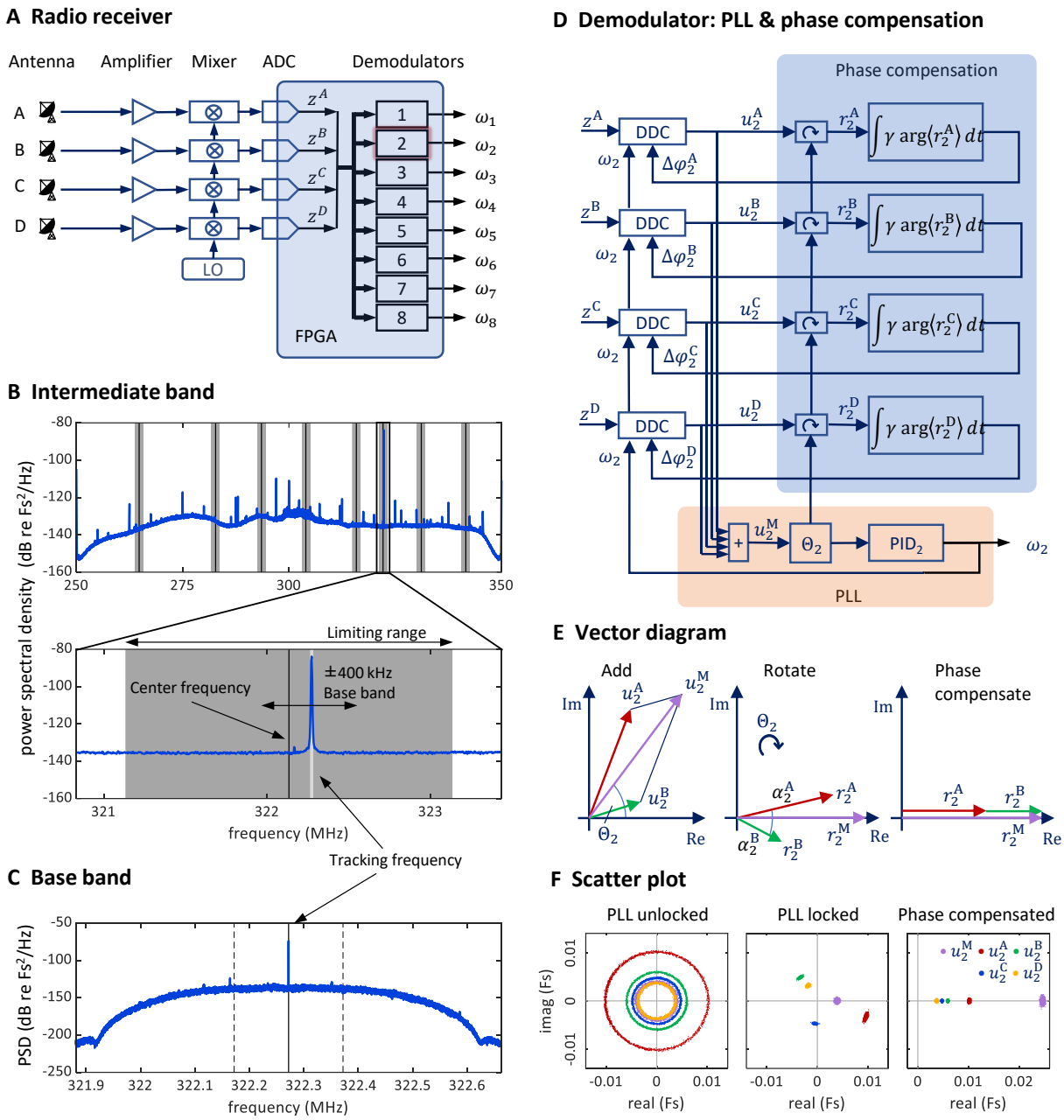


*Figure 3: Radio receiver with PLL demodulators and phase compensation. **(A)** Schematic of the software-defined radio receiver. Analog antenna signals are first down converted by mixing the amplified signals with a local oscillator of frequency ω_{LO}. The resulting intermediate signals have digital representations $z^A(t)$, $z^B(t)$, $z^C(t)$, and $z^D(t)$. From these, eight FM demodulators extract the transmitter tracking frequencies $\omega_i(t)$. **(B)** The power spectral density of the 80 MHz wide intermediate band $z^A(\omega)$ centered on the oscillator frequency $\omega_{LO} = 300$ MHz. The limiting ranges (gray bars) of the 8 channels (transmitters) are typically set to ± 1 MHz of their manually set center frequencies (black vertical lines). A zoom-in (lower graph) to the limiting range of one channel reveals the tracking frequency $\omega_i(t)$ (light grey vertical line) that tracks the large peak at the transmitter frequency. **(C)** The baseband power spectral density $u_i^A(\omega)$ is a down-converted, ±100 kHz wide (dashed vertical lines) flat band around the tracking frequency (solid vertical line) close to the transmitter frequency (the large peak). **(D)** For each channel (transmitter), a demodulator using a PLL (orange shading) and four phase compensation circuits (blue shading) compute the tracking frequency ω_i (shown for $i = 2$). The baseband signals $u_2^a(t)$ are derived from intermediate signals with digital downconverters (DDCs) that operate on a common tracking frequency ω_2. **(E)** Vector diagram in the complex plane illustrating the effects of alignment by the PLL (middle) and of phase compensation (right) on main and baseband vectors (shown for $i = 2$, and only antenna A & B). The phase compensation circuits drive the phases $\alpha_i^a = arg(\langle r_i^a \rangle)$ of rotated baseband signals $r_i^a =$*

*$u_i^a e^{-i\theta_i}$ to zero. **(F)** The four baseband vectors u_2^a are shown without phase compensation (left), after aligning the main vector with the PLL (middle), and after additional phase compensation (right). The result is that all vectors are aligned, and the master vector is of maximal amplitude.*

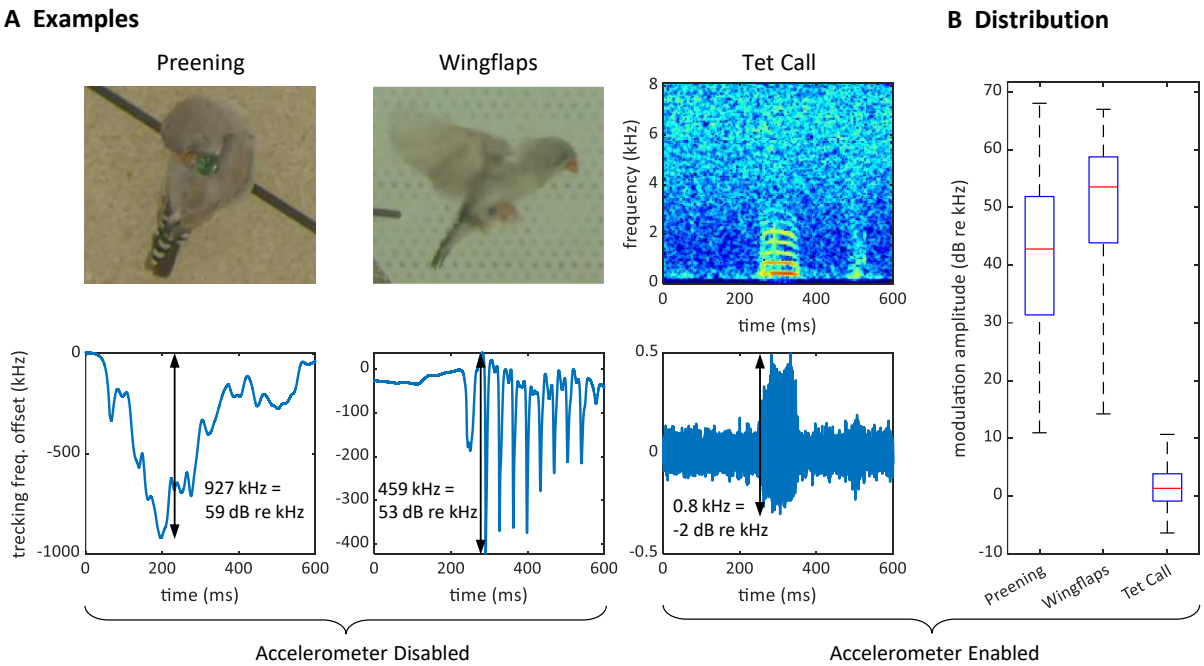
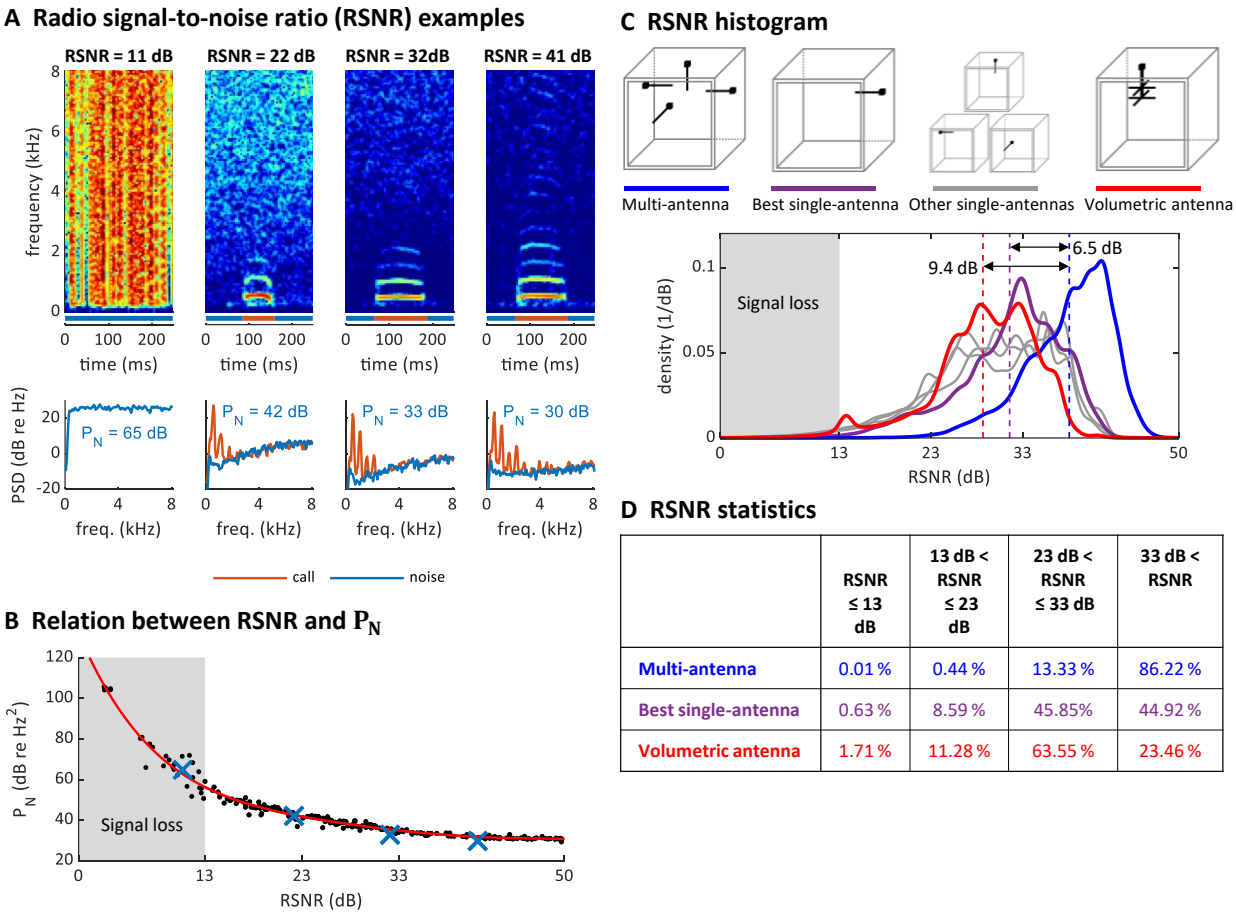


*Figure 4: Proximity affects the transmitter frequency much more than vocalizations. (**A**) On a transmitter device with a disabled (short circuited) accelerometer, the transmitter tracking frequency is strongly modulated by the proximity of the head during preening (left) and by wing movements during flight (middle). In contrast, the modulation of the (high-pass filtered) tracking frequency is much weaker for vocalizations (right) on a transmitter with enabled accelerometer. The modulation amplitudes (black arrows) due to proximity are about 1000 times larger than the vocalization-induced modulation amplitude. (**B**) Distribution of modulation amplitudes for preening (N=62), wing flaps (N=54), and tet calls (N=105). The preening and wing flap events were manually annotated from a set of randomly selected concatenated frequency jumps defined as data segments associated with absolute tracking frequency jumps larger than 50 kHz (see methods).*



	RSNR ≤ 13 dB	13 dB < RSNR ≤ 23 dB	23 dB < RSNR ≤ 33 dB	33 dB < RSNR
Multi-antenna	0.01 %	0.44 %	13.33 %	86.22 %
Best single-antenna	0.63 %	8.59 %	45.85%	44.92 %
Volumetric antenna	1.71 %	11.28 %	63.55 %	23.46 %

*Figure 5: Multi-antenna demodulation achieves the largest radio signal-to-noise ratio (RSNR). (**A**) Examples of acceleration spectrograms that originate from a single calling zebra finch, with RSNRs of 11, 22, 32, and 41 dB (top, from left to right). The bottom plots show the spectra of the accelerometer signal during calls (red lines) and noise (blue lines), computed as time averages from the spectrograms above (time segments indicated by red and blue horizontal bars). The noise power P_N (integral of blue curve, units: dB re 1 Hz^2), is decreasing with increasing RSNR. When the RSNR is below 13 dB, the noise power spectral density is above the signal power of most vocalizations (these become invisible). (**B**) Scatter plot of P_N versus RSNR measurements (dots), plotted for noise segments of diverse RSNR levels. The four examples from **A** are highlighted as blue crosses. A polynomial of degree 5 was fitted to the data (red line). (**C**) Histogram of RSNR over time for the multi-antenna signal (blue line), the single-antenna signal for the antenna with the greatest mean (logarithmic) RSNR (best single-antenna, purple line), all other single-antenna signals (grey lines), and the volumetric antenna signal (red line)**.** The mean (over the measurement period of 2 x 14h) of the multi-antenna RSNR (dashed blue line) is 6.5 dB larger than the mean RSNR of the best single antenna (dashed purple line) and 9.4 dB larger than the mean RSNR of a volumetric antenna (dashed red line). (**D**) Multi-antenna demodulation is significantly better than single-antenna demodulation for the best single antenna and for the volumetric antenna, as demonstrated by the significantly longer reception periods above a given signal-to-noise ratio. The time below the critical RSNR of 13 dB is reduced by a factor of 63 compared to single-antenna demodulation for the best antenna.*

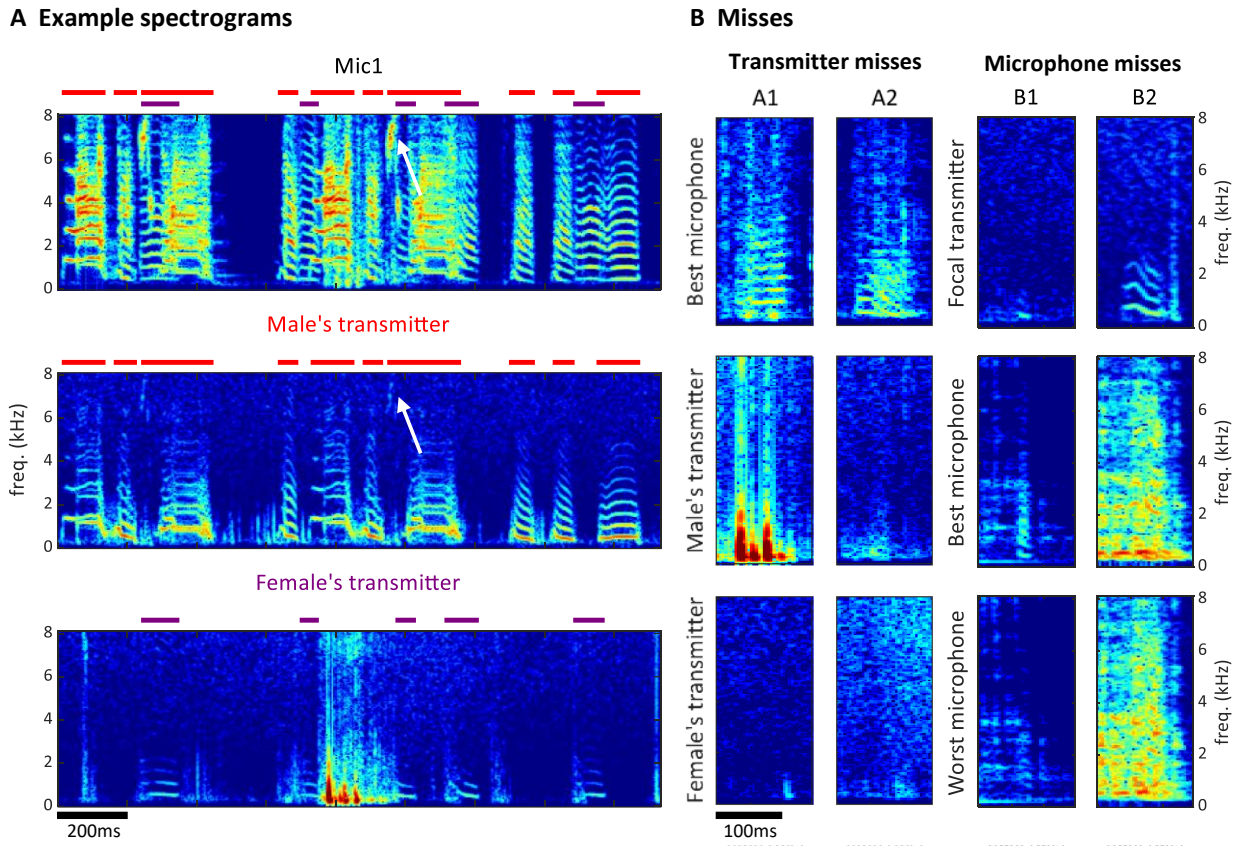


*Figure 6: Vocal signals and their segmentation. (**A**) Example spectrograms of vocalizations produced by a mixed-sex zebra finch pair. The songs and calls overlap on the sound spectrogram of a microphone channel (Mic1, top) but appear well separated on acceleration spectrograms of the male's transmitter channel (middle) and the female's transmitter channel (bottom). Distinct transmitter-based vocal segments are indicated by red (male) and purple (female) horizontal bars on top of the spectrograms. High-frequency vibrations appear attenuated, but even a high-pitched 7 kHz song note (white arrow) by the male is still visible. (**B**) Example vocalizations that are not visible in the transmitter channels (A1: syllable masked by wing flaps, A2: faint signal) or not visible on either some channels (B1: faint syllable masked by noise) or all microphone channels (B2: syllable masked by loud noise).*

	Experiment 1 (7 mins)		**Experiment 2 (7 mins)**		**Total**
	male	**female**	**male**	**female**	
# vocalizations	447	91	487	83	**1108**
# voc. missed on tr. channel	26 (5.8%)	5 (5.5%)	6 (1.2%)	4 (4.8%)	**41 (3.7%)**
# voc. missed on all mic. channels	1 (0.2%)	0 (0.0%)	3 (0.6%)	0 (0.0%)	**4 (0.4%)**
# voc. missed on Mic1 channel	11 (2.5%)	9 (10.0%)	23 (4.7%)	3 (3.6%)	**46 (4.2%)**
# voc. with overlaps	9 (2.0%)	9 (10.0%)	9 (1.9%)	9 (10.8%)	**36 (3.3%)**
of which are associated with mixed vocal signal on given mic. ch.	8 (88.9%)	9 (100%)	8 (88.9%)	8 (88.9%)	**33 (91.7%)**
of which are associated with mixed vocal signal on other transmitter ch.	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)

Table 1: Statistics of missed vocalizations on a transmitter channel (second row, cf. Example A1 and A2 in Figure 6B), on all microphone channels (third row, cf. Example B2 in Figure 6B), and on an individual microphone channel (Mic1, fourth row, cf. Example B1 in Figure 6B). The fifth row ('# voc. with overlaps') quantifies the number of vocalizations that overlap with at least one vocalization of the other bird (as in the example in Figure 6A). For those overlapping vocalizations, we checked by visual inspection whether they were associated with a mixed vocal signal (from both birds) on a given microphone channel (sixth row) and whether the vocalization leaked into the transmitter signal of the other bird (seventh row). See methods for details.