Supporting Information

Giret et al. 10.1073/pnas.1317087111

SI Methods

All experiments were done using methods approved by the Veterinary Office of the Canton of Zurich, Switzerland.

Loop Delay Estimation. Motor and auditory delays were estimated as the time lags of stimulation at which song frequency modulation (FM) or evoked neural responses deviated from baseline levels by more than 3 SDs. To estimate lateral magnocellular nucleus of the anterior nidopallium (LMAN) motor delays, we chronically implanted bipolar 50-μm stainless steel wires separated by about 0.5 mm into LMAN and stimulated using monophasic 0.2-ms current pulses in the range of 100–500 μA. We either used single pulses (Fig. S2*A*) or double pulses separated by 1 ms (Fig. 2*B*).

LMAN Extracellular Recordings in Freely Moving Zebra Finches. We identified LMAN in microdrive implantation surgeries by its antidromic response to electrical stimulation in the robust nucleus of the arcopallium (RA). In each recording session, we let birds first produce undirected songs with lights on, after which we exposed birds to playback of their own songs in the dark.

Spectrograms, Raster Plots, and Spike-Triggered Averages. Sounds were band-pass filtered in the range of 0.3–13 kHz and digitized at 32-kHz sampling rate. They were then Fourier transformed in 16-ms Hamming windows with 12-ms overlap and were displayed as log-power sound spectrograms in Figs. 2, and 3 and Figs. S2 and S4. We produced raster plots in Fig. 3 and Fig. S4 using the real-time display format corresponding to right alignment, i.e., we plotted a spike at time t in between spectrogram columns associated with Fourier windows $[t_0 - 16\ \text{ms},\ t_0]$ and $[t_0 - 12\ \text{ms},\ t_0 + 4\ \text{ms}]$, where t_0 is the largest integer multiple of 4 ms smaller than t. We plotted the spike triggered average (STA) log sound amplitude in Fig. S4 using centered alignment: the STA log amplitude at zero time lag corresponded to the log RMS sound waveform in the window $[t_0 - 8\ \text{ms},\ t_0 + 8\ \text{ms}]$ relative to a spike at time t_0.

Cross-Covariance Analysis. To compute motif cross-covariance (CC) functions, song-related (mean-subtracted) spike trains $\rho_S(t)$ were taken from windows $[0,\ T]$ starting 32 ms before onset of the stereotyped motif, to include hypothetical motor commands that initiate song motifs, and ending coincidentally with the last syllable of the motif $C_{SP}(\tau) = \frac{1}{T}\int_0^T \rho_S(t)\rho_P(t+\tau)dt$. Playback spike trains $\rho_P(t)$ were taken from corresponding windows $[\tau,\ T+\tau]$ of the same motif rendition.

Bout CC functions were computed analogously, with the interval $[0,\ T]$ starting 32 ms before onset of the song bout and ending coincidentally with the last syllable of the bout.

To explore robustness of our covariance analysis, we varied the preonset interval of 32 ms from 4 to 100 ms: our conclusions were not affected by such changes. In one alternative analysis of motif covariance, we zero padded playback spike trains $\rho_P(t)$ outside the boundary $[0,\ T]$ of the stereotyped song motif, in which case $C_{SP}(\tau)$ was normalized by $T - |\tau|$ instead of T to yield an unbiased estimate of covariance. Results were not qualitatively different when using this definition of covariance function.

Peaks in both motif and bout CC functions were purely stimulus driven and therefore could in principle be observed at arbitrary large positive and negative time lags. Conceptually, negative dips in CC functions (instead of local maxima as in Fig. 1) can also be supportive of inverse models, i.e., models in which the feedback is negative and not positive (neurons are inhibited by the sounds they produce and not excited).

LMAN Extracellular Recordings in Anesthetized Zebra Finches

Experiments were performed in two head-fixed anesthetized adult male zebra finches (>90 dph). A few days (typically 1 wk) before the surgery, birds were caught from our colony and placed alone in an isolation chamber. Their song was recorded with a custom Labview (National Instruments) software. For playback, a song bout containing at least three motifs was selected for each bird. Birds were deeply anesthetized with isoflurane (1.5–3%) and received up to four urethane injections in intervals of at least 30 min. Injections consisted of 20–40 μL of 20% urethane in saline injected i.m. or s.c. Birds were placed in a stereotaxic apparatus, and the head angle formed by the flat part of the skull above the beak was set at 65°. Windows were made into the skull near HVC, RA, LMAN, and Area X. A metal plate was fixed on the skull with dental acrylic. A stimulation electrode made of two coated stainless steel wires (diameter: 51 μm) connected to copper wires was implanted into RA [average coordinates from the bifurcation of the midsagittal sinus (lambda): 2.4 mm medio-lateral (ML); −2.8 mm anterior-posterior (AP); −1.7 mm dorso-ventral (DV) with a manipulator angle of −38°]. LMAN and HVC were located based both on stereotaxic coordinates and antidromic activation to RA electrical stimulation (average LMAN coordinates: 1.7 mm ML; 5.0 mm AP; 2.0–2.7 mm DV with a manipulator angle of 55°; average HVC coordinates: 2.0 mm ML; –0.5 mm AP; 0.3–0.6 mm DV with a manipulator angle of −45°). Area X was identified based on steady baseline neuronal activity (average coordinates: 1.9 mm ML; 6.0 AP mm; 2.5–3.5 mm DV with a manipulator angle of 45°). To perform extracellular electrophysiological recordings of single neurons in LMAN, we used glass pipettes filled in with a 3 M saline solution. Glass pipettes were pulled with a P-97 micropipette puller (Sutter Instruments) and had an impedance of 10–14 MΩ. The signal was amplified with an Axoclamp 2B, filtered (300 Hz to 10 kHz) with a LHBF-48X NPI filter, and recorded with a custom Labview software. We used tungsten metal electrodes with impedances of 1.5–4.0 MΩ (Microprobes) to perform extracellular recordings in Area X. The signal was amplified with a custom-made amplifier, filtered (300 Hz to 10 kHz) with a LHBF-48X NPI filter, and recorded with a custom Labview software. A glass pipette filled in with 250 M GABA in saline was implanted into HVC.

While simultaneously recording single units in LMAN and Area X, we performed the following steps. First, RA was stimulated to assess the latency of antidromic spike responses in LMAN; typical stimulation currents evoking responses in LMAN were ∼60 μA, and typical antidromic response latencies were ∼2.0 ms. Second, we played back the bird's own song (BOS) for 5 min with an inter-BOS interval of 2 s. Third, while still recording the same LMAN and Area X units and playing back BOS, we inactivated HVC via pressure injection of GABA (using a Picospritzer). We injected ∼0.1 μL of GABA between each BOS playback during roughly 10 min. Fourth, we stopped the GABA injections but kept recording the same units while playing back BOS for another 10 min.

At the end of the experiment, we made burns at the position of the glass pipette by delivering a 10-μA current for 10 s via a metal electrode (inserted into HVC after removal of the pipette). Similar electrolytic lesions were performed at the last recording sites in LMAN and Area X and at the RA stimulation site. The bird was then euthanized with a 0.05 mL i.m. injection of 20%

pentobarbital (50 mg/mL). Thereafter, the brain was removed and immersed into a 4% solution of paraformaldehyde. The brain was rinsed with a 0.01 M Phosphate Buffer Solution, the hemispheres were separated from each other, and the hemisphere of interest was glued on a metal plate and embedded in 2% agar. Sagittal slices of 80 μm were cut with a Thermo Microm HM650V microtome and mounted on slides. Regions of interest were imaged and inspected for correct placement of electrolytic burns.

Mirroring with Multiple Latencies

We can quantitatively work out a prediction for the mirroring CC function in a simple but conceptually instructive toy model, in the case in which a motor neuron's firing may elicit multiple auditory consequences at a range of latencies. Consider a single motor neuron, as in Fig. S1 *A* and *B*, with time varying firing rate activity $m(t)$. Also, let $a_i(t)$, with $i = 1, \ldots, N$ being the time-varying firing rate activity of a population of N auditory neurons presynaptic to the motor neuron (Fig. S1*B*). Now suppose that each time the motor neuron is active at time t, it may generate multiple song features, which in turn generate subsequent auditory responses of different latencies in each auditory neuron. If we denote the impulse response (to the firing of the motor neuron) of each auditory neuron i by $K_i(u)$ (which is nonzero only for $u > 0$), then during any bout of motor exploration $m(t)$, the response of each auditory neuron, in a simple linear response model, will be

$$a_i(t) = \int_0^\infty du\; K_i(u)\, m(t-u). \qquad \textbf{[S1]}$$

Of course, if there are multiple motor neurons firing, there will be additional contributions to the auditory feedback response. However, when explorations between different motor neurons are uncorrelated, the auditory feedback due to this one motor neuron is the only part of the auditory response that can correlate with the motor neuron's past activity and therefore is the only part of the response that determines the final weights from the presynaptic auditory population to the given motor neuron under a learning rule that has previously been reported to generate inverse models (1, 2).

This learning rule was derived as gradient descent on the following energy function on the vector of synaptic weights $\boldsymbol{w}$

$$E(\boldsymbol{w}) = \left\langle \int_0^\infty ds\; e(s) \left[m(t-s) - \boldsymbol{w} \cdot \boldsymbol{a}(t) \right]^2 \right\rangle_m, \qquad \textbf{[S2]}$$

where $e(s)$ is an eligibility trace, and $\langle \cdot \rangle_m$ denotes an average over the statistics of the motor explorations $m(t)$. This rule can be thought of as attempting to tune the inverse model synaptic weight vector $\boldsymbol{w}$ so that the input to the motor neuron using auditory activity pattern $\boldsymbol{a}(t)$ postdicts any past motor exploration $m(t-s)$, with the importance of postdiction at time lag s weighted by the eligibility trace $e(s)$. We assume the eligibility trace is normalized so that $\int_0^\infty ds\; e(s) = 1$. Gradient descent on this energy function yields a biologically plausible learning rule in which synapses from auditory to motor neurons undergo Hebbian association and hetero-synaptic competition.

Because the energy function is quadratic in the weights, with a unique minimum, it is straightforward to determine the outcome of learning: $\boldsymbol{w} = \boldsymbol{C}^{-1}\boldsymbol{p}$ where

$$\boldsymbol{p}_i = \int_0^\infty ds\; e(s) \left\langle m(t-s)\, a_i(t) \right\rangle_m, \qquad \textbf{[S3]}$$

is the average integrated, eligibility weighted correlation between auditory neuron i and the motor neuron and

$$\boldsymbol{C}_{ij} = \left\langle a_i(t) a_j(t) \right\rangle_m, \qquad \textbf{[S4]}$$

is the equal time correlation between auditory neurons.

Now assume that motor exploration while inverse model synapses $\boldsymbol{w}$ are being learned is temporally uncorrelated: $\left\langle m(t_1) m(t_2) \right\rangle_m = \delta(t_1 - t_2)$. Then using Eq. **S1**, we find

$$\left\langle m(t_1) a_i(t_2) \right\rangle_m = K_i(t_2 - t_1), \qquad \textbf{[S5]}$$

and substituting this into Eq. **S3** yields

$$\boldsymbol{p}_i = \int_0^\infty ds\; e(s) K_i(s). \qquad \textbf{[S6]}$$

Also, $\boldsymbol{C}_{ij}$ in Eq. **S4** reduces to

$$\boldsymbol{C}_{ij} = \int_0^\infty du\; K_i(u) K_j(u). \qquad \textbf{[S7]}$$

Now that we have the learned synaptic weight vector $\boldsymbol{w}$, we can compute the mirroring CC function. Consider for example a simple case where the motor neuron fires at an isolated time during motor production, as in Fig. S1*A*, *Upper*. Call this time $t = 0$. Thus, during motor production, motor activity is $m_M(t) = \delta(t)$. Now consider the motor response during playback of the song. Each auditory neuron i will have a firing rate response $K_i(t)$. This auditory activity will propagate to the motor neuron through the learned inverse model synapses $\boldsymbol{w}$, yielding the motor response to playback, $m_P(t) = \sum_i \boldsymbol{w}_i K_i(t)$. The mirroring CC function is then

$$CC(\tau) \equiv \int dt\; m_M(t)\; m_P(t+\tau) = \sum_i \boldsymbol{w}_i K_i(\tau), \qquad \textbf{[S8]}$$

where again $\boldsymbol{w} = \boldsymbol{C}^{-1}\boldsymbol{p}$, and $\boldsymbol{p}$ and $\boldsymbol{C}$ are given in Eqs. **S6** and **S7**, respectively. This equation yields a general prediction for the mirroring CC function directly in terms of the arbitrary, diverse latency, auditory responses $K_i(t)$ and the eligibility trace $e(t)$.

To understand this prediction intuitively, imagine a situation where each auditory neuron i responds only during a narrow window of time, centered at τ_i, as in Fig. S1*A*, *Lower*. If these windows of time do not overlap significantly, then the auditory correlation matrix $\boldsymbol{C}$ in Eq. **S4** is approximately diagonal. Furthermore, if the auditory responses have roughly equal strength, then $\boldsymbol{C}$ is simply proportional to the identity matrix. Thus, the synaptic weights $\boldsymbol{w}_i$ are simply proportional to the eligibility weighted auditory motor correlations $\boldsymbol{p}_i$. Now suppose the eligibility trace is a monotonically decaying exponential with time constant τ_e: $e(s) = \frac{1}{\tau_e} e^{-s/\tau_e}$. Furthermore, suppose the width of the auditory response $K_i(s)$ is narrow relative to τ_i. Then $\boldsymbol{p}_i$ in Eq. **S6** is approximately proportional to $e^{-\tau_i/\tau_e}$ for all i. Thus, the weight from an auditory neuron to the motor neuron is exponentially suppressed by the latency of the auditory neuron response relative to the motor neuron's firing, shown schematically in Fig. S1*B*. Therefore, in this scenario, the CC mirroring function is

$$CC(\tau) \approx \sum_i e^{-\tau_i/\tau_e} K_i(\tau). \qquad \textbf{[S9]}$$

Therefore, the mirroring CC function is a weighted sum of the auditory responses, where long latency responses are exponentially suppressed by their latency. Thus, if a motor neuron has a complex playback response represented by a diversity of latencies, the mirroring CC function will tend to peak near the earliest

latency, as shown schematically in Fig. S1*C*. Note also that because we are assuming that the propagation delay from auditory neurons to the motor neuron in this particular model is negligible, then the minimum auditory latency τ_1, relative to the motor spike, is the sensorimotor loop delay for the motor neuron, so the mirroring CC function peaks at the neuron's sensorimotor loop delay. This entire argument generalizes straightforwardly to the case where there is a finite propagation delay from auditory neurons to the motor neuron.

This effect of the mirroring CC function peaking at the earliest auditory response latency will also be amplified in the population average CC, across all motor neurons. For example, each motor neuron could have its own earliest playback response latency, yielding a relatively wide distribution of minimum auditory latencies across the motor neuron population, as observed in the main paper. However, as implied by Eq. **S9**, the strength of the peak of each neuron's individual mirroring CC function would be exponentially suppressed by that neuron's own minimum latency. Thus, when averaging over the motor neuron population, motor neurons whose individual CC function peaks earlier will dominate the average relative to those motor neurons whose CC function peaks later (and have longer latency auditory responses). Thus, the population averaged CC function would peak at a temporal lag that would fall within the earlier range of the distribution of motor neuron response latencies to auditory playback (which in this model are equivalent to sensorimotor loop delays), as observed in the main text.

Finally, we note that if a motor neuron has elevated, or suppressed, activity at multiple time points during the song, the mirroring CC function in Eq. **S8** could develop peaks at time points other than the sensorimotor loop delay. Such peaks would then reflect the long range, reproducible, temporal structure of song itself, and could arise for example at $\pm$100 ms, corresponding to the time lag between song syllables. The theory of inverse model learning, however, is completely agnostic about the nature of long-range temporal structures in the final learned song. Thus, the presence or absence of such peaks at other time-lags can neither provide evidence for, nor disprove the theory.

1. Hahnloser R, Ganguli S (2013) Vocal learning with inverse models. *Principles of Neural Coding*, eds Panzeri S, Quiroga P (CRC Taylor and Francis, Boca Raton, FL).

2. Hanuschkin A, Ganguli S, Hahnloser RHR (2013) A Hebbian learning rule gives rise to mirror neurons and links them to control theoretic inverse models. *Front Neural Circuits* 7:106.

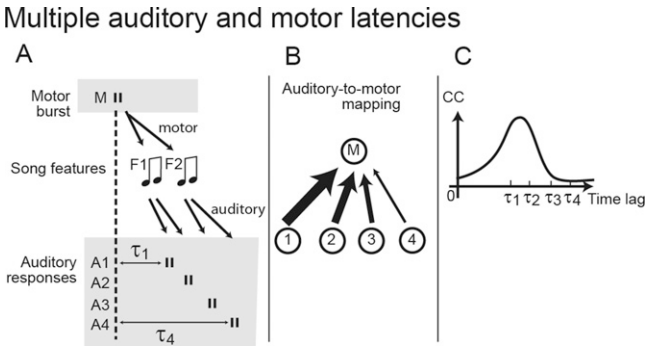

Fig. S1. Auditory-to-motor mapping resulting from multiple auditory and motor latencies. (*A*) In the more realistic situation than in Fig. 1, a motor burst produces multiple song features (F_1, F_2). These features occur with diverse latencies after the burst and lead to auditory responses (in four auditory neurons) with an even wider distribution of time lags (τ_1 to τ_4). (*B*) Eligibility-weighted Hebbian learning strengthens the connection from the first auditory neuron to the motor neuron more than the connections from other auditory neurons with longer response latencies (connection strength is represented by arrow thickness). The reason for this preference of short lags is the monotonically decaying eligibility for synaptic strengthening as a function of the time lag between motor and auditory spikes. (*C*) As a result of the synaptic strengths in *B*, the CC function in the multilatency case exhibits a peak at a time lag near the set of shorter latencies (τ_1 and τ_2).

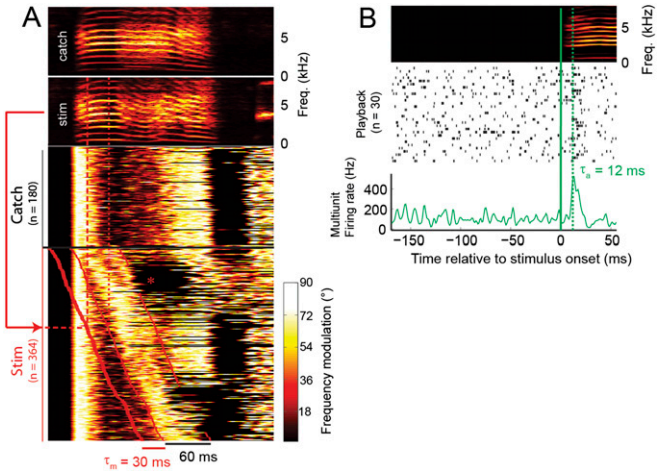

Fig. S2. (*A*) Electrical stimulation in LMAN using 0.2-ms current pulses of 500 μA (single pulse) during song leads to transient distortions of song syllables compared with catch trials (upper vs. lower spectrogram). A stack plot of FM across 364 syllable renditions reveals that for a large range of stimulation times (thick red line) there is a transient FM increase within 30–90 ms (delimited by thin red lines). The red arrow indicates stack position of the stimulated syllable shown on top (labeled Stim). Some stimulation effects are not transient (asterisk). (*B*) Motif-aligned spike raster plot for a LMAN multiunit site (same legend as in Fig. 2*C*).

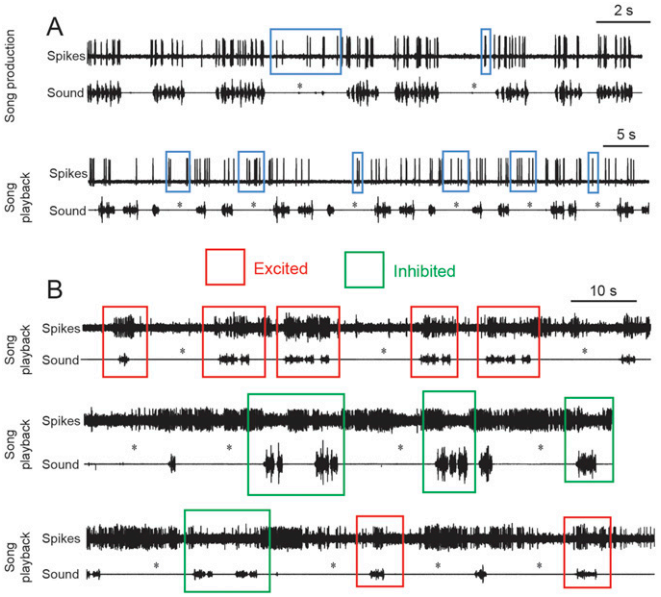

Fig. S3. (*A*) LMAN responses to song production and to song playback. Extracellular voltage trace (spikes) of an LMAN neuron (single unit) recorded during song production (top traces) and during playback of song (bottom traces). Sequentially recorded files have been appended to each other, asterisks mark file boundaries. The blue boxes delimit spontaneous firing in this cell not caused by song or song playback, suggesting a role of LMAN spikes also in offline computations. (*B*) Unusual LMAN single unit that is sometimes excited by song playback (red boxes), and sometimes inhibited (green boxes), with frequent switching in between and no obvious change in behavioral state.

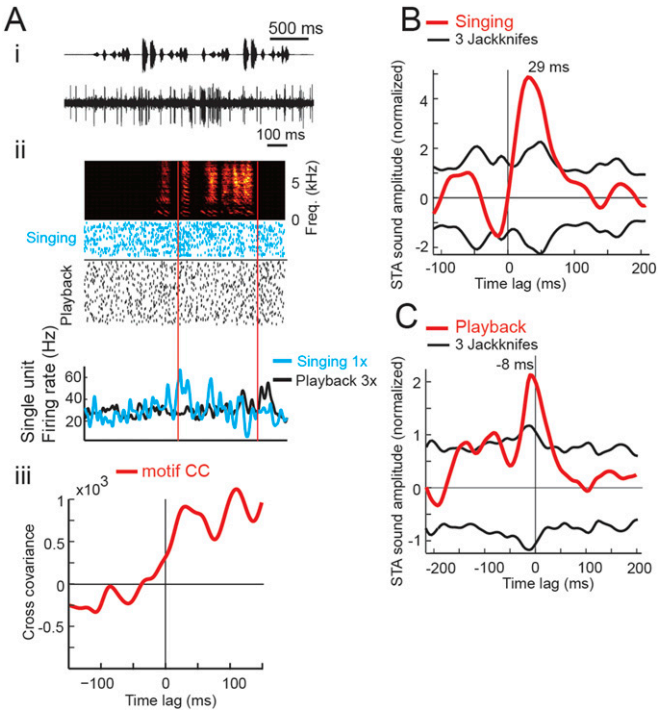

Fig. S4. (*A*) Large positive mirroring offset in an LMAN single unit. (*A, i*) Raw extracellular voltage trace, shown below song oscillogram. (*A, ii*) Log-power song spectrogram (high and low power shown in yellow and black, respectively), and spike raster plots for many different motifs and firing rate curves averaged over all motifs (delimited by red lines). The spike raster plots and firing-rate curves show spiking activity during singing of each motif (blue) and during many playbacks of that motif (black). (*A, iii*) CC function for spike trains from song motifs (thick red line). For this single unit, there is a very long lag between song-evoked firing (blue) at the beginning of the motif and playback-evoked firing (black) at the end of the motif (*A, ii*); nevertheless, there is a positive mirroring offset already in the range of several tens of milliseconds (*A, iii*). (*B* and *C*) Spike-triggered average sound amplitude during song (*B*) and during playback (*C*).

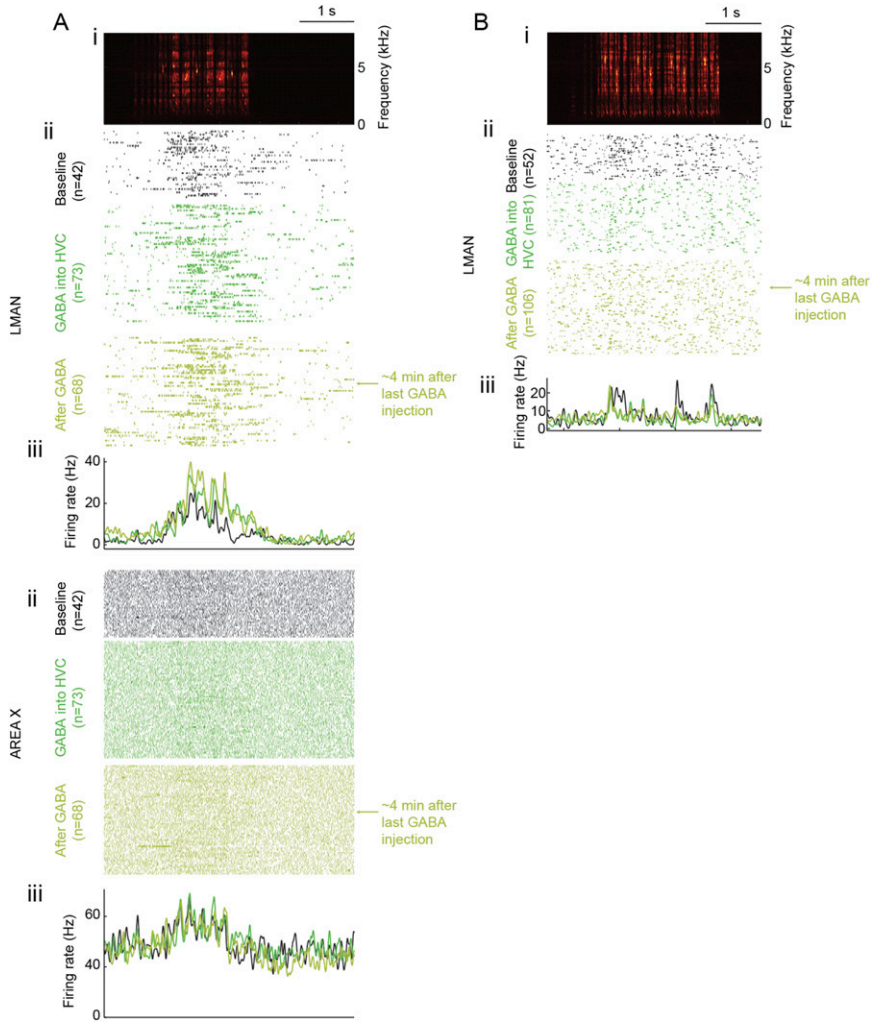

Fig. S5. Paired LMAN and Area X multiunit recording (*A*) and LMAN multiunit recording (*B*) in two anesthetized birds showing auditory responses to the BOS and almost no effect on these responses of HVC inactivation with GABA (250 M in 0.9% saline). (*i*) Log-power spectrogram of the BOS stimulus. (*ii*) The spike raster plot showing spikes recorded during baseline BOS playback (black), during the injection of GABA into HVC while playing back BOS (green), and after the last GABA injection (yellow-green). (*iii*) The firing-rate curves are averages over BOS playback trials (smoothed with a Gaussian kernel of $\sigma = 80$ ms).