Emergent intensity invariance in a physiologically inspired model of the grasshopper auditory system

Jona Hartling, Jan Benda

1 Exploring a grasshopper's sensory world

Our scientific understanding of sensory processing systems results from the distributed accumulation of anatomical, physiological and ethological evidence. This process is undoubtedly without alternative; however, it leaves us with the challenge of integrating the available fragments into a coherent whole in order to address issues such as the interaction between individual system components, the functional limitations of the system overall, or taxonomic comparisons between systems that process the same sensory modality. Any unified framework that captures the essential functional aspects of a given sensory system thus has the potential to deepen our current understanding and fasciliate systematic investigations. However, building such a framework is a challenging task. It requires a wealth of existing knowledge of the system and the signals it operates on, a clearly defined scope, and careful reduction, abstraction, and formalization of the underlying structures and mechanisms.

One sensory system about which extensive information has been gathered over the years is the auditory system of grasshoppers (*Acrididae*). Grasshoppers rely on their sense of hearing primarily for intraspecific communication, which includes mate attraction (D. v. Helversen 1972) and evaluation (Stange and Ronacher 2012), sender localization (D. v. Helversen and Rheinlaender 1988), courtship display (Elsner 1968), rival deterrence (Greenfield and Minckley 1993), and loss-of-signal predator alarm (SOURCE). In accordance with this rich behavioral repertoire, grasshoppers have evolved a variety of sound production mechanisms to generate acoustic communication signals for different contexts and ranges using their wings, hindlegs, or mandibles (Otte 1970). Among the most conspicuous acoustic signals of grasshoppers are their species-specific calling songs, which broadcast the presence of the singing individual — mostly the males of the species — to potential mates within range. These songs are usually more characteristic of a species than morphological traits (Tishechkin and Vedenina 2016; Tarasova et al. 2021), which can

vary greatly within species (Rowell 1972; Köhler et al. 2017). The reliance on songs to mediate reproduction represents a strong evolutionary driving force, that resulted in a massive species diversification (Vedenina and Mugue 2011; Sevastianov et al. 2023), with over 6800 recognized grasshopper species in the *Acrididae* family (Cigliano et al. 2024). It is this diversity of species, and the crucial role of acoustic communication in its emergence, that makes the grasshopper auditory system an intriguing candidate for attempting to construct a functional model framework. As a necessary reduction, the model we propose here focuses on the pathway responsible for the recognition of species-specific calling songs, disregarding other essential auditory functions such as directional hearing (D. v. Helversen 1984; Ronacher, D. v. Helversen, and Helversen 1986; D. v. Helversen and Rheinlaender 1988).

To understand the functional challenges faced by the grasshopper auditory system, one has to understand the properties of the songs it is designed to recognize. Grasshopper songs are amplitude-modulated broad-band acoustic signals. Most songs are produced by stridulation, during which the animal pulls the serrated stridulatory file on its hindlegs across a resonating vein on the forewings (O. v. Helversen and Elsner 1977; Stumpner and O. v. Helversen 1994; D. v. Helversen and O. v. Helversen 1997). Every tooth that strikes the vein generates a brief pulse of sound. Multiple pulses make up a syllable; and the alternation of syllables and relatively quiet pauses forms a characteristic, through noisy, waveform pattern. Song recognition depends on certain temporal and structural parameters of this pattern, such as the duration of syllables and pauses (D. v. Helversen 1972), the slope of pulse onsets (D. v. Helversen 1993), and the accentuation of syllable onsets relative to the preceeding pause (Balakrishnan et al. 2001; D. v. Helversen, Balakrishnan, and Helversen 2004). The amplitude modulation of the song is sufficient for recognition (D. v. Helversen and O. v. Helversen 1997). However, the essential recognition cues can vary considerably with external physical factors, which requires the auditory system to be invariant to such variations in order to reliably recognize songs under different conditions. For instance, the temporal structure of grasshopper songs warps with temperature (Skovmand and Boel Pedersen 1983). The auditory system can compensate for this variability by reading out relative temporal relationships rather than absolute time intervals (Creutzig, Wohlgemuth, et al. 2009; Creutzig, Benda, et al. 2010), as those remain relatively constant across different temperatures (D. v. Helversen 1972). Another, perhaps even more fundamental external source of song variability lays in the attenuation of sound intensity with increasing distance to the sender. Sound attenuation depends on both the frequency content of the signal and the vegetation of the habitat (Michelsen 1978). For the receiving auditory system, this has two major implications. First, the amplitude dynamics of the song pattern are steadily degraded over distance, which limits the effective communication range of grasshoppers to 1 - 2 m in their typical grassland habitats (Lang 2000).

Second, the overall intensity level of songs at the receiver's position varies depending on the location of the sender, which should ideally not affect the recognition of the song pattern. This neccessitates that the auditory system achieves a certain degree of intensity invariance — a time scale-selective sensitivity to faster amplitude dynamics and simultaneous insensitivity to slower, more sustained amplitude dynamics. Intensity invariance in different auditory systems is often associated with neuronal adaptation (Benda and Hennig 2008; Barbour 2011; Ozeri-Engelhard et al. 2018; more general: Benda 2021). In the grasshopper auditory system, a number of neuron types along the processing chain exhibit spike-frequency adaptation in response to sustained stimulus intensities (Römer 1976; Gollisch and Herz 2004; Hildebrandt et al. 2009; Clemens, Weschke, et al. 2010; Fisch et al. 2012) and thus likely contribute to the emergence of intensity-invariant song representations. This means that intensity invariance is not the result of a single processing step but rather a gradual process, in which different neuronal populations contribute to varying degrees (Clemens, Weschke, et al. 2010) and by different mechanisms (Hildebrandt et al. 2009). Approximating this process within a functional model framework thus requires a considerable amount of simplification. In this work, we demonstrate that even a small number of basic physiologically inspired signal transformations — specifically, pairs of nonlinear and linear operations — is sufficient to achieve a meaningful degree of intensity invariance.

Invariance to non-informative song variations is crucial for reliable song recognition; however, it is not sufficient to this end. In order to recognize a conspecific song as such, the auditory system needs to extract sufficiently informative features of the song pattern and then integrate the gathered information into a final categorical percept. Previous authors have proposed a functional model framework that describes this process — feature extraction, evidence accumulation, and categorical decision making — in both crickets (Clemens and Hennig 2013; Hennig et al. 2014) and grasshoppers (Clemens and Ronacher 2013; review on both: Ronacher, Hennig, and Clemens 2015). Their framework provides a comprehensible and biologically plausible account of the computational mechanisms required for species-specific song recognition, which has served as the inspiration for the development of the model pathway we propose here. The existing framework relies on pulse trains as input signals, which were designed to capture the essential structural properties of natural song envelopes (Clemens and Ronacher 2013). In the first step, a bank of parallel linear-nonlinear feature detectors is applied to the input signal. Each feature detector consists of a convolutional filter and a subsequent sigmoidal nonlinearity. The outputs of these feature detectors are temporally averaged to obtain a single feature value per detector, which is then assigned a specific weight. The linear combination of weighted feature values results in a single preference value, that serves as predictor for the behavioral response of the animal to the presented input signal. Our model pathway adopts the

general structure of the existing framework but modifies it in several key aspects. The convolutional filters, which have previously been fitted to behavioral data for each individual species (Clemens and Hennig 2013), are replaced by a larger, generic set of unfitted Gabor basis functions in order to cover a wide range of possible song features across different species. Gabor functions approximate the general structure of the filters used in the existing framework as well as the filter functions found in various auditory neurons (Rokem et al. 2006; Clemens, Kutzki, et al. 2011; Clemens, Wohlgemuth, and Ronacher 2012). The fitted sigmoidal nonlinearities in the existing framework consistently exhibited very steep slopes and are therefore replaced by shifted Heaviside step-functions, which results in a binarization of the feature detector outputs. Another, more substantial modification is that the feature detector outputs are temporally averaged in a way that does not condense them into single feature values but retains their time-varying structure. This is in line with the fact that songs are no discrete units but part of a continuous acoustic stream that the auditory system has to process in real time. Moreover, a time-varying feature representation only stabilizes after a certain delay following the onset of a song, which emphasizes the temporal dynamics of evidence accumulation towards a final categorical decision. The most notable difference between our model pathway and the existing framework, however, lays in the addition of a physiologically inspired preprocessing stage, whose starting point corresponds to the initial reception of airborne sound waves. This allows the model to operate on unmodified recordings of natural grasshopper songs instead of condensed pulse train approximations, which widens its scope towards more realistic, ecologically relevant scenarios. For instance, we were able to investigate the contribution of different processing stages to the emergence of intensity-invariant song representations based on actual field recordings of songs at different distances from the sender. In the following, we outline the structure of the proposed model of the grasshopper auditory pathway, from the initial reception of sound waves up to the generation of a high-dimensional, time-varying feature representation that is suitable for species-specific song recognition. We provide a side-by-side account of the known physiological processing steps and their functional approximation by basic mathematical operations. We then elaborate on two key mechanisms that drive the emergence of intensity-invariant song representations within the auditory pathway.

2 Developing a functional model of the grasshopper song recognition pathway

The essence of constructing a functional model of a given system is to gain a sufficient understanding of the system's essential structural components and their presumed functional roles; and to then build a formal framework of manageable complexity around these two aspects.

Anatomically, the organization of the grasshopper song recognition pathway can be outlined as a feed-forward network of three consecutive neuronal populations (Fig. 1a-c): Peripheral auditory receptor neurons, whose axons enter the ventral nerve cord at the level of the metathoracic ganglion; local interneurons that remain exclusively within the thoracic region of the ventral nerve cord; and ascending neurons projecting from the thoracic region towards the supraesophageal ganglion (Rehbein et al. 1974; Rehbein 1976; Eichendorf and Kalmring 1980). The input to the network originates at the tympanal membrane, which acts as acoustic receiver and is coupled to the dendritic endings of the receptor neurons (Gray 1960). The outputs from the network converge in the supraesophageal ganglion, which is presumed to harbor the neuronal substrate for conspecific song recognition and response initiation (Ronacher, D. v. Helversen, and Helversen 1986; Bauer and Helversen 1987; Bhavsar et al. 2017). Functionally, the ascending neurons are the most diverse of the three populations along the pathway. Individual ascending neurons possess highly specific response properties that contrast with the rather homogeneous response properties of the preceding receptor neurons and local interneurons (Clemens, Kutzki, et al. 2011), indicating a transition from a uniform population-wide processing stream into several parallel branches. Based on these anatomical and physiological considerations, the overall structure of the model pathway is divided into two distinct stages (Fig. 1d). The preprocessing stage incorporates the known physiological processing steps at the levels of the tympanal membrane, the receptor neurons, and the local interneurons; and operates on one-dimensional signal representations. The feature extraction stage corresponds to the processing within the ascending neurons and further downstream towards the supraesophageal ganglion; and operates on high-dimensional signal representations. The details of each physiological processing step and its functional approximation within the two stages are outlined in the following sections.

2.1 Population-driven signal preprocessing

Grasshoppers receive airborne sound waves by a tympanal organ at either side of the body. The tympanal membrane acts as a mechanical resonance filter for sound-induced vibrations (Windmill et al. 200 Malkin et al. 2014). Vibrations that fall within specific frequency bands are focused on different membrane areas, while others are attenuated. This processing step can be approximated by an initial bandpass filter

$$x_{\text{filt}}(t) \;=\; x(t) \,*\, h_{\text{BP}}(t), \qquad f_{\text{cut}} \;=\; 5\,\text{kHz},\, 30\,\text{kHz} \tag{1}$$

applied to the acoustic input signal $x(t)$. The auditory receptor neurons transduce the vibrations of the tympanal membrane into sequences of action potentials. Thereby, they encode the amplitude modulation, or envelope, of the signal (Machens, Prinz, et al. 2001), which likely involves a rectifying nonlinearity (Machens, Stemmler, et al. 2001). This can be modelled as

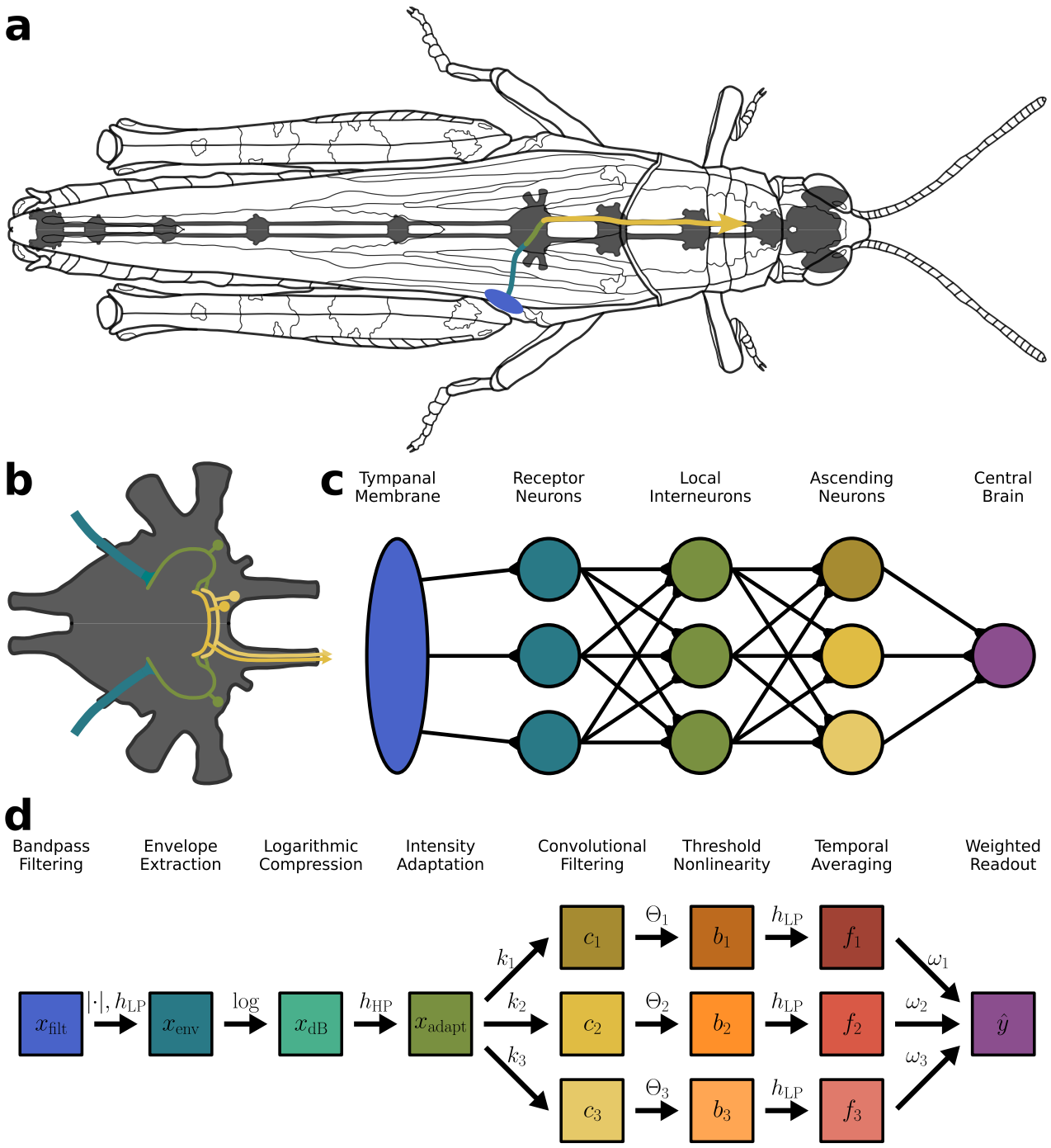

Fig. 1: Schematic organisation of the song recognition pathway in grasshoppers compared to the structure of the functional model pathway. a: Simplified course of the pathway in the grasshopper, from the tympanal membrane over receptor neurons, local interneurons, and ascending neurons further towards the supraesophageal ganglion. **b**: Schematic of synaptic connections between the three neuronal populations within the metathoracic ganglion. **c**: Network representation of neuronal connectivity. **d**: Flow diagram of the different signal representations and transformations along the model pathway. All representations are time-varying. 1st half: Preprocessing stage (one-dimensional). 2nd half: Feature extraction stage (high-dimensional).

full-wave rectification followed by lowpass filtering

$$x_{\mathrm{env}}(t) \;=\; |x_{\mathrm{filt}}(t)| \,*\, h_{\mathrm{LP}}(t), \qquad f_{\mathrm{cut}} \;=\; 500\,\mathrm{Hz} \tag{2}$$

of the tympanal signal $x_{\mathrm{filt}}(t)$. Furthermore, the receptors exhibit a sigmoidal response curve over logarithmically compressed intensity levels (Suga 1960; Gollisch, Schütze, et al. 2002). In the model pathway, logarithmic compression is achieved by conversion to decibel scale

$$x_{\mathrm{dB}}(t) \;=\; 10 \cdot \log_{10} \frac{x_{\mathrm{env}}(t)}{x_{\mathrm{ref}}}, \qquad x_{\mathrm{ref}} \;=\; \max\left[x_{\mathrm{env}}(t)\right] \tag{3}$$

relative to the maximum intensity x_{ref} of the signal envelope $x_{\mathrm{env}}(t)$. Both the receptor neurons (Römer 1976; Gollisch and Herz 2004; Fisch et al. 2012) and, on a larger scale, the subsequent local interneurons (Hildebrandt et al. 2009; Clemens, Weschke, et al. 2010) adapt their firing rates in response to sustained stimulus intensity levels, which allows for the robust encoding of faster amplitude modulations against a slowly changing overall baseline intensity. Functionally, the adaptation mechanism resembles a highpass filter

$$x_{\mathrm{adapt}}(t) \;=\; x_{\mathrm{dB}}(t) \,*\, h_{\mathrm{HP}}(t), \qquad f_{\mathrm{cut}} \;=\; 10\,\mathrm{Hz} \tag{4}$$

over the logarithmically scaled envelope $x_{\mathrm{dB}}(t)$. This processing step concludes the preprocessing stage of the model pathway. The resulting intensity-adapted envelope $x_{\mathrm{adapt}}(t)$ is then passed on from the local interneurons to the ascending neurons, where it serves as the basis for the following feature extraction stage.

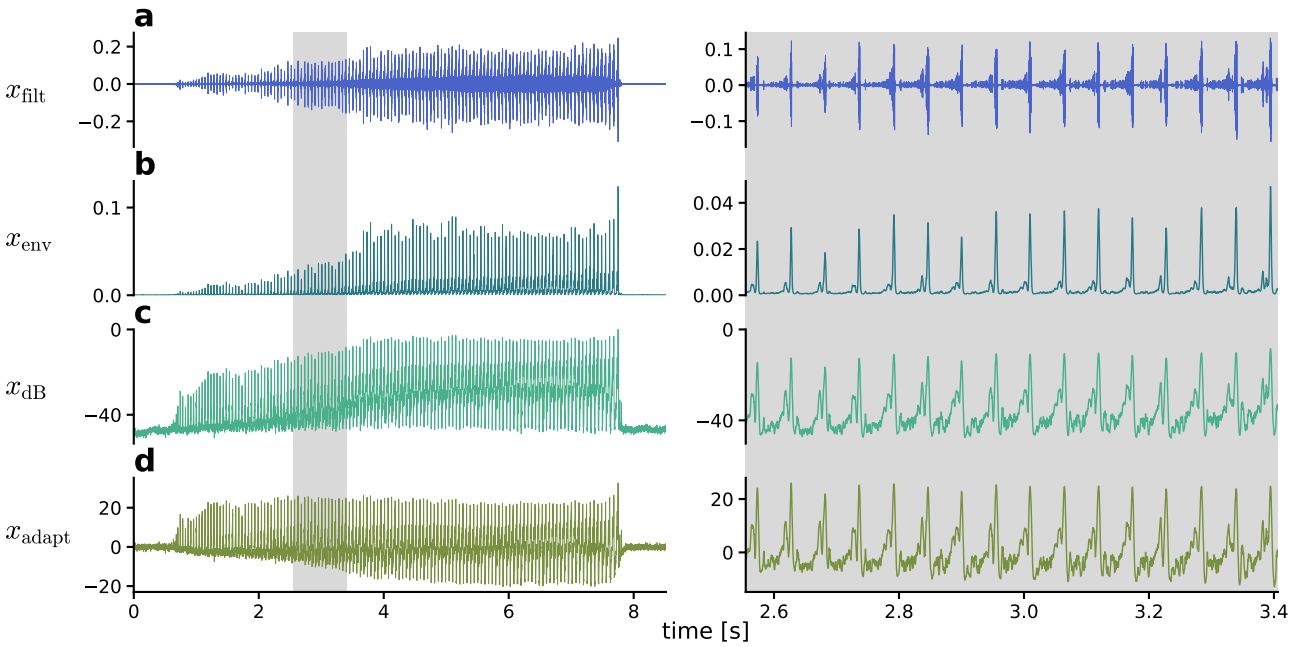

Fig. 2: Representations of a song of *O. rufipes* during the preprocessing stage. **a**: Bandpass-filtered tympanal signal. **b**: Signal envelope. **c**: Logarithmically scaled envelope. **d**: Intensity-adapted envelope.

2.2 Feature extraction by individual neurons

The ascending neurons extract and encode a number of different features of the preprocessed signal. As a population, they hence represent the signal in a higher-dimensional space than the preceding receptor neurons and local interneurons. Each ascending neuron is assumed to scan the signal for a specific template pattern, which can be thought of as a kernel of a particular structure and on a particular time scale. This process, known as template matching, can be modelled as a convolution

$$c_i(t) \;=\; x_{\text{adapt}}(t) \,*\, k_i(t) = \int_{-\infty}^{+\infty} x_{\text{adapt}}(\tau) \,\cdot\, k_i(t \,-\, \tau)\, d\tau \tag{5}$$

of the intensity-adapted envelope $x_{\text{adapt}}(t)$ with a kernel $k_i(t)$ per ascending neuron. We use Gabor kernels as basis functions for creating different template patterns. An arbitrary one-dimensional, real Gabor kernel is generated by multiplication of a Gaussian envelope and a sinusoidal carrier

$$k_i(t,\, \sigma_i,\, \omega_i,\, \phi_i) \;=\; e^{-\frac{t^2}{2\sigma_i{}^2}} \,\cdot\, \sin(\omega_i\, t \,+\, \phi_i), \qquad \omega_i \;=\; 2\pi f_{\text{sin}} \tag{6}$$

with Gaussian standard deviation or kernel width σ_i, carrier frequency ω_i, and carrier phase ϕ_i. Different combinations of σ and ω result in Gabor kernels with different lobe number n,

which is the number of half-periods of the carrier that fit under the Gaussian envelope within reasonable limits of attenuation. The interval under the Gaussian envelope that contains the relevant lobes of the kernel can be defined as Gaussian full-width measured at relative peak height h_{rel}

$$\text{FWRH}(\sigma,\, h_{\text{rel}}) \;=\; 2 \cdot \sqrt{-2 \cdot \ln h_{\text{rel}}} \cdot \sigma, \qquad h_{\text{rel}} \in (0,\, 1] \tag{7}$$

With this, an appropriate carrier frequency ω for obtaining a Gabor kernel with width σ and desired lobe number n can be approximated as

$$\omega(n,\, \sigma,\, h_{\text{rel}}) \;=\; \frac{n \,+\, \beta_0}{4 \cdot \sqrt{-2 \cdot \ln h_{\text{rel}}}}, \qquad n \,\geq\, 2 \;\forall\; n \,\in\, \mathbb{Z} \tag{8}$$

where β_0 is a small positive offset to the near-linear relationship between ω and n to balance the amplitude of the n desired lobes of the kernel — which should be maximized — against the amplitude of the next-outer lobes, which should not exceed the threshold value determined by h_{rel}. For $n = 1$, carrier frequency ω is set to zero, which results in a simple Gaussian kernel. Carrier phase ϕ determines the position of the kernel lobes relative to the kernel center. By setting ϕ to one of only four specific phase values (Tab. 1), we restrict the Gabor kernels to be either even functions (mirror-symmetric, uneven n) or odd functions (point-symmetric, even n) with either positive or negative sign, which refers to the sign of the kernel's central lobe (even kernels) or the left of the two central lobes (odd kernels).

Tab. 1: Values of phase ϕ that are specific for the four major groups of Gabor kernels.

sign	even kernels	odd kernels
$+$	$+\pi\,/\,2$	π
$-$	$-\pi\,/\,2$	0

These four major groups of Gabor kernels allow for the extraction of different types of signal features, such as the presence of peaks (even, $+$), troughs (even, $-$), onsets (odd, $+$), and offsets (odd, $-$) at various time scales. Following the convolutional template matching, each kernel-specific response $c_i(t)$ is passed through a shifted Heaviside step-function $H(c_i \,-\, \Theta_i)$ with threshold value Θ_i to obtain a binary response

$$b_i(t,\, \Theta_i) \;=\; \begin{cases} 1, & c_i(t) \,>\, \Theta_i \\ 0, & c_i(t) \,\leq\, \Theta_i \end{cases} \tag{9}$$

which can be thought of as a categorization into "relevant" and "irrelevant" response values. In the grasshopper, these thresholding nonlinearities might either be part of the processing within the ascending neurons or take place further downstream (SOURCE). Finally, the responses of

the ascending neurons are assumed to be integrated somewhere in the supraesophageal ganglion (Ronacher, D. v. Helversen, and Helversen 1986; Bauer and Helversen 1987; Bhavsar et al. 2017). This processing step can be approximated as temporal averaging of the binary responses $b_i(t)$ by a lowpass filter

$$f_i(t) = b_i(t) * h_{\mathrm{LP}}(t), \qquad f_{\mathrm{cut}} = 1\,\mathrm{Hz} \tag{10}$$

to obtain a final set of slowly changing kernel-specific features $f_i(t)$. In the resulting high-dimensional feature space, different species-specific song patterns are characterized by a distinct combination of feature values, which can be read out by a simple linear classifier.

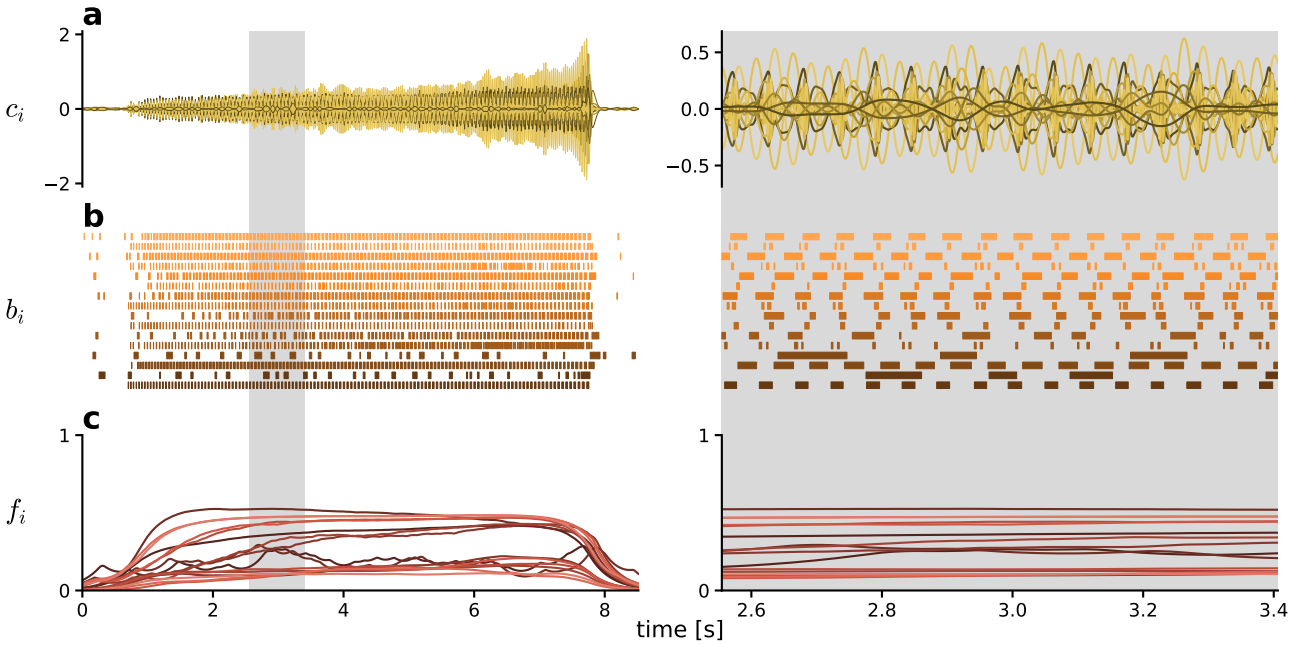

Fig. 3: Representations of a song of *O. rufipes* during the feature extraction stage. Different colors indicate Gabor kernels with different lobe number n and sign, with lighter colors for higher n ($1 \leq n \leq 4$; both $+$ and $-$ per n; two kernel widths σ of 4 ms and 32 ms per sign). **a**: Kernel-specific filter responses. **b**: Binary responses. **c**: Finalized features.

3 Two mechanisms driving the emergence of intensity-invariant song representation

The robustness of song recognition is tied to the degree of intensity invariance of the finalized feature representation. Ideally, the values of each feature should depend only on the relative amplitude dynamics of the song pattern but not on the overall intensity of the song. In the grasshopper, the emergence of intensity-invariant representations along the song recognition pathway likely is a distributed process that involves different neuronal populations, which raises the question of what the essential computational mechanisms are that drive this process. Within

the model pathway, we identified two key mechanisms that render the song representation more invariant to intensity variations. The two mechanisms each comprise a nonlinear signal transformation followed by a linear signal transformation but differ in the specific operations involved, as outlined in the following sections.

3.1 Logarithmic compression & spike-frequency adaptation

The first notable emergence of intensity invariance along the model pathway occurs during the transformation of the signal envelope $x_{\mathrm{env}}(t)$ into the logarithmically scaled envelope $x_{\mathrm{dB}}(t)$ and then into the intensity-adapted envelope $x_{\mathrm{adapt}}(t)$. In order to disentangle the interplay of logarithmic compression and adaptation, $x_{\mathrm{env}}(t)$ can be rewritten as a synthetic mixture

$$x_{\mathrm{env}}(t) \;=\; \alpha \cdot s(t) \,+\, \eta(t), \qquad x_{\mathrm{env}}(t) \;>\; 0 \;\forall\; t \,\in\, \mathbb{R} \tag{11}$$

of a song component $s(t)$ with variable multiplicative scale $\alpha \geq 0$ and a fixed-scale noise component $\eta(t)$. Both $s(t)$ and $\eta(t)$ are assumed to have unit variance. By conversion of $x_{\mathrm{env}}(t)$ to decibel scale (Eq. 3), α turns from a multiplicative scale in linear space into an additive term, or offset, in logarithmic space

$$\begin{aligned} x_{\mathrm{dB}}(t) \;&=\; \log \frac{\alpha \cdot s(t) \,+\, \eta(t)}{x_{\mathrm{ref}}} \\ &=\; \log \frac{\alpha}{x_{\mathrm{ref}}} \,+\, \log \left[s(t) \,+\, \frac{\eta(t)}{\alpha} \right] \end{aligned} \tag{12}$$

which allows for its separation from $s(t)$ but introduces a scaling of $\eta(t)$ by the inverse of α. The subsequent highpass-filtering (Eq. 4) of $x_{\mathrm{dB}}(t)$ can then be approximated as a subtraction of the local offset within a suitable time interval $0 \ll T_{\mathrm{HP}} < \frac{1}{f_{\mathrm{cut}}}$:

$$x_{\mathrm{adapt}}(t) \;\approx\; x_{\mathrm{dB}}(t) \,-\, \log \frac{\alpha}{x_{\mathrm{ref}}} \;=\; \log \left[s(t) \,+\, \frac{\eta(t)}{\alpha} \right] \tag{13}$$

This means that α cannot be entirely eliminated from $x_{\mathrm{adapt}}(t)$, only redistributed between $s(t)$ and $\eta(t)$. In consequence, if α is sufficiently large ($\alpha \gg 1$), $\eta(t)$ is attenuated to the point of being negligible, so that $x_{\mathrm{adapt}}(t)$ represents $s(t)$ in a scale-free manner. If $s(t)$ and $\eta(t)$ are at similar scales ($\alpha \approx 1$), $x_{\mathrm{adapt}}(t)$ largely resembles $x_{\mathrm{dB}}(t)$. However, if α is sufficiently small ($\alpha \ll 1$), $\eta(t)$ masks $s(t)$ even after the intensity adaptation. Therefore, the effective intensity invariance of $x_{\mathrm{adapt}}(t)$ relative to $x_{\mathrm{env}}(t)$ is limited by the initial scaling of $s(t)$ relative to $\eta(t)$; that is, the signal-to-noise ratio (SNR) of $x_{\mathrm{env}}(t)$ with ($\alpha > 0$) and without ($\alpha = 0$)

song component $s(t)$

$$\mathrm{SNR}(\alpha) = \frac{\sigma_x^2}{\sigma_\eta^2} = \frac{\alpha^2 \cdot \sigma_\mathrm{s}^2 + \sigma_\eta^2}{\sigma_\eta^2} = \alpha^2 + 1, \qquad \sigma_\mathrm{s}^2 = \sigma_\eta^2 = 1 \tag{14}$$

which depends quadratically on α if $s(t) \perp \eta(t)$. Overall, the combination of logarithmic compression and adaptation allows for the equalization of different sufficiently large song scales, which is essential for intensity-invariant song representation. However, this mechanism is unable to recover songs that have already sunken below the noise floor, which emphasizes the importance of a sufficiently high SNR at the intial reception of the signal for reliable song recognition.

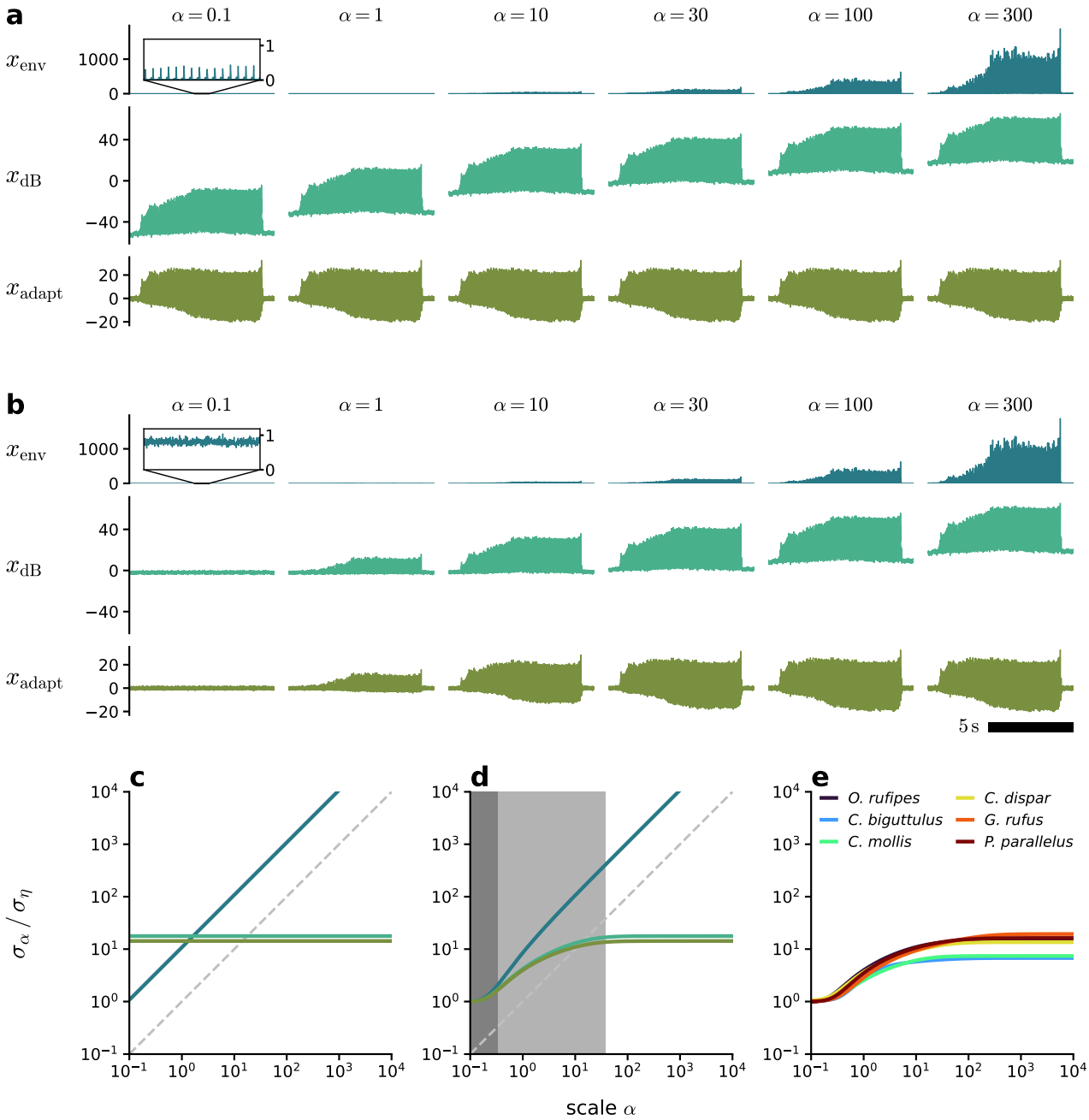

Fig. 4: Intensity invariance by logarithmic compression and adaptation is restricted by the noise floor. Synthetic input $x_{\mathrm{filt}}(t)$ consists of song component $s(t)$ scaled by α with (**c** and **d**) or without (**a** and **b**) additive noise component $\eta(t)$. Input $x_{\mathrm{filt}}(t)$ is transformed into envelope $x_{\mathrm{env}}(t)$, logarithmically compressed envelope $x_{\mathrm{dB}}(t)$, and intensity-adapted envelope $x_{\mathrm{adapt}}(t)$. **Left**: $x_{\mathrm{env}}(t)$, $x_{\mathrm{dB}}(t)$, and $x_{\mathrm{adapt}}(t)$ for different scales α. **Right**: Ratios of the standard deviation of $x_{\mathrm{env}}(t)$, $x_{\mathrm{dB}}(t)$, and $x_{\mathrm{adapt}}(t)$ relative to the respective reference standard deviation σ_η for input $x_{\mathrm{filt}}(t) = \eta(t)$. **a** and **b**: Ideally, if $x_{\mathrm{filt}}(t) = \alpha \cdot s(t)$, then $x_{\mathrm{adapt}}(t)$ is intensity-invariant across all α. **c** and **d**: In practice, if $x_{\mathrm{filt}}(t) = \alpha \cdot s(t) + \eta(t)$, the intensity invariance of $x_{\mathrm{adapt}}(t)$ is limited to sufficiently large α. Shaded area indicates saturation of $x_{\mathrm{adapt}}(t)$ at 95 % curve span.

3.2 Thresholding nonlinearity & temporal averaging

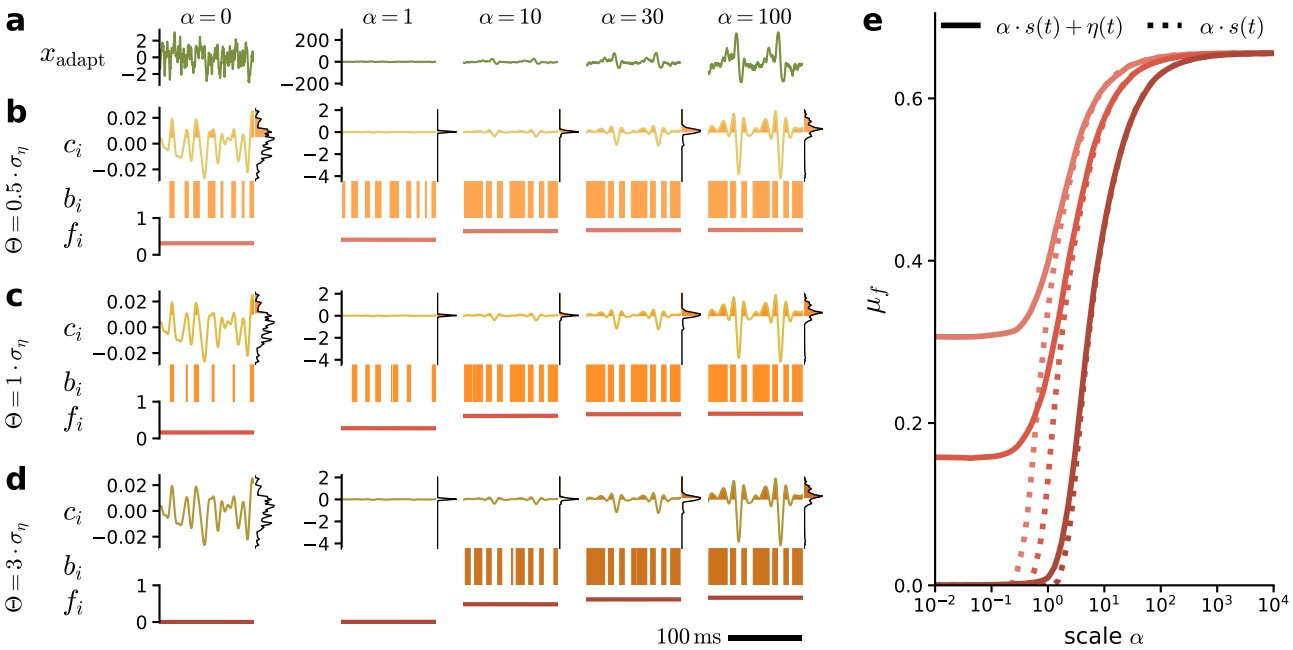

Fig. 5: Intensity invariance by thresholding and temporal averaging depends on both the threshold value and the noise floor. Synthetic input $x_{\text{adapt}}(t)$ consists of song component $s(t)$ scaled by α with additive noise component $\eta(t)$. Input $x_{\text{adapt}}(t)$ is transformed into kernel response $c_i(t)$, binary response $b_i(t)$, and feature $f_i(t)$. Threshold value Θ_i is set to multiples of the reference standard deviation σ_η of $c_i(t)$ for input $x_{\text{adapt}}(t) = \eta(t)$. Darker colors correspond to higher Θ_i. **Left**: $x_{\text{adapt}}(t)$, $c_i(t)$, $b_i(t)$, and $f_i(t)$ for different scales α and threshold values Θ_i. Left-most column is is the pure-noise reference. **Right**: Average value of $f_i(t)$ during the song for the different Θ_i. **a**: Input $x_{\text{adapt}}(t)$. **b-d**: $c_i(t)$, $b_i(t)$, and $f_i(t)$ for the different Θ_i based on the same $x_{\text{adapt}}(t)$ from **a**. **e**: Average value of $f_i(t)$ during the song for the different Θ_i in **b-d**.

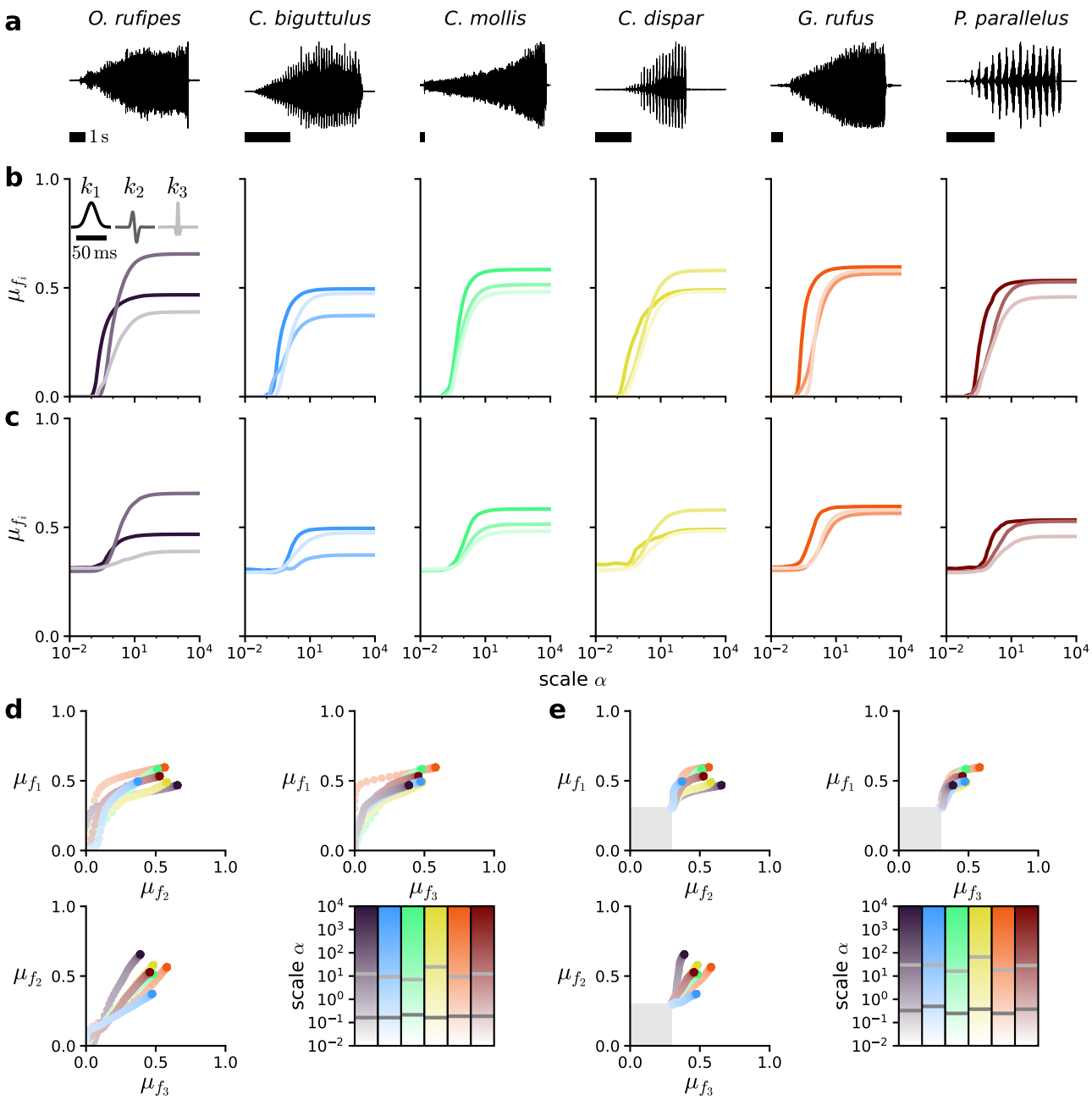

Fig. 6: Feature representation of different species-specific songs saturates at different points in feature space.

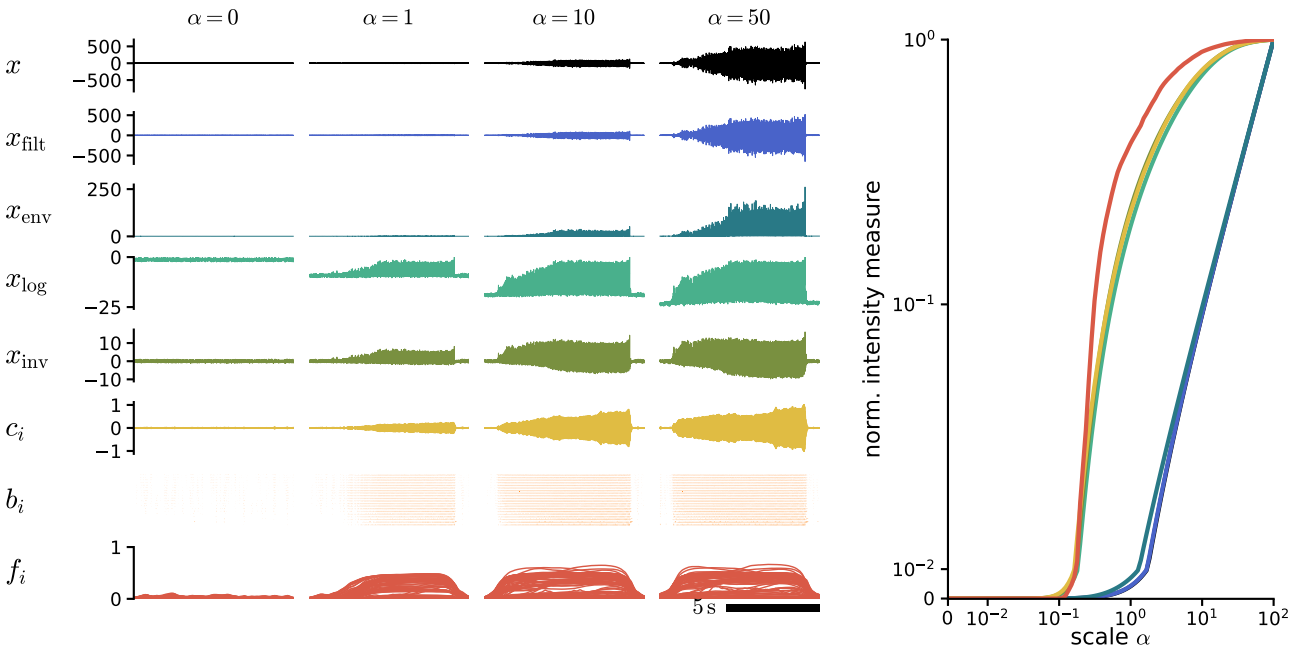

Fig. 7: Step-wise emergence of intensity invariant song representation along the model pathway.

The second key mechanism for the emergence of intensity invariance along the model pathway takes place during the transformation of the kernel responses $c_i(t)$ over the binary responses $b_i(t)$ into the finalized features $f_i(t)$. Kernel response $c_i(t)$ quantifies the degree of similarity between kernel $k_i(t)$ and the preprocessed signal $x_{\text{adapt}}(t)$. The thresholding nonlinearity $H(c_i - \Theta_i)$ categorizes the value of $c_i(t)$ at every time point t into "relevant" ($c_i(t) > \Theta_i$, $b_i(t) = 1$) and "irrelevant" ($c_i(t) \leq \Theta_i$, $b_i(t) = 0$) response values

By passing $c_i(t)$ through the thresholding nonlinearity $H(c_i - \Theta_i)$, its amplitude values are binned into one of two categories (Eq. 9).

: $c_i(t) > \Theta_i$

This mechanism is mediated by the thresholding nonlinearity $H(c_i - \Theta_i)$. By passing $c_i(t)$ through the thresholding nonlinearity (Eq. 9), its probability density $p(c_i, T)$ within some observed time interval T is split around threshold value Θ_i into two complementary parts:

$$\int_{\Theta_i}^{+\infty} p(c_i, T)\, dc_i \;=\; 1 \;-\; \int_{-\infty}^{\Theta_i} p(c_i, T)\, dc_i \;=\; \frac{T_1}{T}, \qquad \int_{-\infty}^{+\infty} p(c_i, T)\, dc_i \;=\; 1 \tag{15}$$

The right-sided part of the split $p(c_i, T)$ corresponds to time T_1 where $c_i(t) > \Theta_i$, while the left-sided part corresponds to time $T_0 = T - T_1$ where $c_i(t) \leq \Theta_i$. The semi-definite integral over the right-sided part of $p(c_i, T)$ represents the ratio of time T_1 to total time T because the indefinite integral of a probability density is normalized to 1. Following the thresholding nonlinearity, the resulting binary responses $b_i(t)$ are lowpass-filtered (Eq. 10) to obtain $f_i(t)$, which can be approximated as temporal averaging over a suitable time interval $T_{\text{LP}} > \frac{1}{f_{\text{cut}}}$

$$f_i(t) \;\approx\; \frac{1}{T_{\text{LP}}} \int_{t}^{t+T_{\text{LP}}} b_i(\tau)\, d\tau \;=\; \frac{T_1}{T_{\text{LP}}}, \qquad b_i(t) \;\in\; \{0,\, 1\} \tag{16}$$

Feature $f_i(t)$

If the lowpass filter (Eq. 10) over $b_i(t)$ is approximated as temporal averaging over a suitable time interval $T_{\text{LP}} > \frac{1}{f_{\text{cut}}}$, then $f_i(t)$ can be linked to a similar temporal ratio of time T_1 during which $b_i(t)$ is 1 within the total averaging interval T_{LP}. Therefore, the value of $f_i(t)$ at every time point t approximately signifies the cumulative probability that $c_i(t)$ exceeds Θ_i during the corresponding averaging interval T_{LP}:

$$f_i(t) \;\approx\; \int_{\Theta_i}^{+\infty} p(c_i, T_{\text{LP}})\, dc_i \;=\; P(c_i \;>\; \Theta_i,\, T_{\text{LP}}) \tag{17}$$

In a sense, $f_i(t)$ resembles a duty cycle of some sort, which quantifies purely temporal relations

in the structure of $c_i(t)$ with no regard for precise amplitude values apart from their relation to Θ_i.

Accordingly, a substantial amount of information about the degree of similarity between signal $x_{\text{adapt}}(t)$ and kernel $k_i(t)$ that is contained in $c_i(t)$ is lost during its transformation into $f_i(t)$. Instead, $f_i(t)$ only retains information about the temporal relation of $c_i(t)$ relative to Θ_i

This loss of amplitude information is the key to the intensity invariance of $f_i(t)$: For a given Θ_i, different scales of $c_i(t)$ can still result in similar T_1 segments depending on the magnitude of the derivative of $c_i(t)$ in temporal proximity to time points at which $c_i(t)$ crosses Θ_i. The steeper the slope of $c_i(t)$ around the threshold crossings, the less T_1 changes with scale variations.

In a sense, $f_i(t)$ resembles a duty cycle of some sort, as it quantifies purely temporal relations in the structure of $c_i(t)$ with no regard for precise amplitude values apart from their relation to Θ_i. This near-complete loss of amplitude information is the key to the intensity invariance of $f_i(t)$: For a given Θ_i, different scales of $c_i(t)$ can still result in similar T_1 segments depending on the magnitude of the derivative of $c_i(t)$ in temporal proximity to time points at which $c_i(t)$ crosses Θ_i. The steeper the slope of $c_i(t)$ around the threshold crossings, the less T_1 changes with scale variations.

4 Discriminating species-specific song patterns in feature space

5 Conclusions & outlook

Song recognition pathway: Grasshopper vs. model:
The model pathway includes a rather large number of Gabor kernels compared to the 15 to 20 ascending neurons in the grasshopper auditory system (Stumpner and Ronacher 1991).

Definition of invariance (general, systemic):
Invariance = Property of a system to maintain a stable output with respect to a set of relevant input parameters (variation to be represented) but irrespective of one or more other parameters (variation to be discarded) → Selective input-output decorrelation

Definition of intensity invariance (context of neurons and songs):
Intensity invariance = Time scale-selective sensitivity to certain faster amplitude dynamics (song waveform, small-scale AM) and simultaneous insensitivity to slower, more sustained amplitude dynamics (transient baseline, large-scale AM, current overall intensity level)

→ Without time scale selectivity, any fully intensity-invariant output will be a flat line

Log-HP: Implication for intensity invariance:
- Logarithmic scaling is essential for equalizing different song intensities
→ Intensity information can be manipulated more easily when in form of a signal offset in log-space than a multiplicative scale in linear space

- Capability to compensate for intensity variations, i.e. selective amplification of output $x_{\text{adapt}}(t)$ relative to input $x_{\text{env}}(t)$, is limited by input SNR (Eq. 14):
→ Ability to equalize between different sufficiently large scales of $s(t)$
→ Inability to recover $s(t)$ when initially masked by noise floor $\eta(t)$

- Logarithmic scaling emphasizes small amplitudes (song onsets, noise floor)
→ Recurring trade-off: Equalizing signal intensity vs preserving initial SNR

Thresh-LP: Implication for intensity invariance:
- Role of song periodicity for feature representation!

- Suggests a relatively simple rule for optimal choice of threshold value Θ_i:
→ Find amplitude c_i that maximizes absolute derivative of $c_i(t)$ over time
→ Optimal with respect to intensity invariance of $f_i(t)$, not necessarily for other criteria such as song-noise separation or diversity between features

- Nonlinear operations can be used to detach representations from graded physical stimulus (to fasciliate categorical behavioral decision-making?):
1) Capture sufficiently precise amplitude information: $x_{\text{env}}(t)$, $x_{\text{adapt}}(t)$
→ Closely following the AM of the acoustic stimulus
2) Quantify relevant stimulus properties on a graded scale: $c_i(t)$
→ More decorrelated representation, compared to prior stages
3) Nonlinearity: Distinguish between "relevant vs irrelevant" values: $b_i(t)$
→ Trading a graded scale for two or more categorical states
4) Represent stimulus properties under relevance constraint: $f_i(t)$
→ Graded again but highly decorrelated from the acoustic stimulus
5) Categorical behavioral decision-making requires further nonlinearities
→ Parameters of a behavioral response may be graded (e.g. approach speed), initiation of one behavior over another is categorical (e.g. approach/stay)

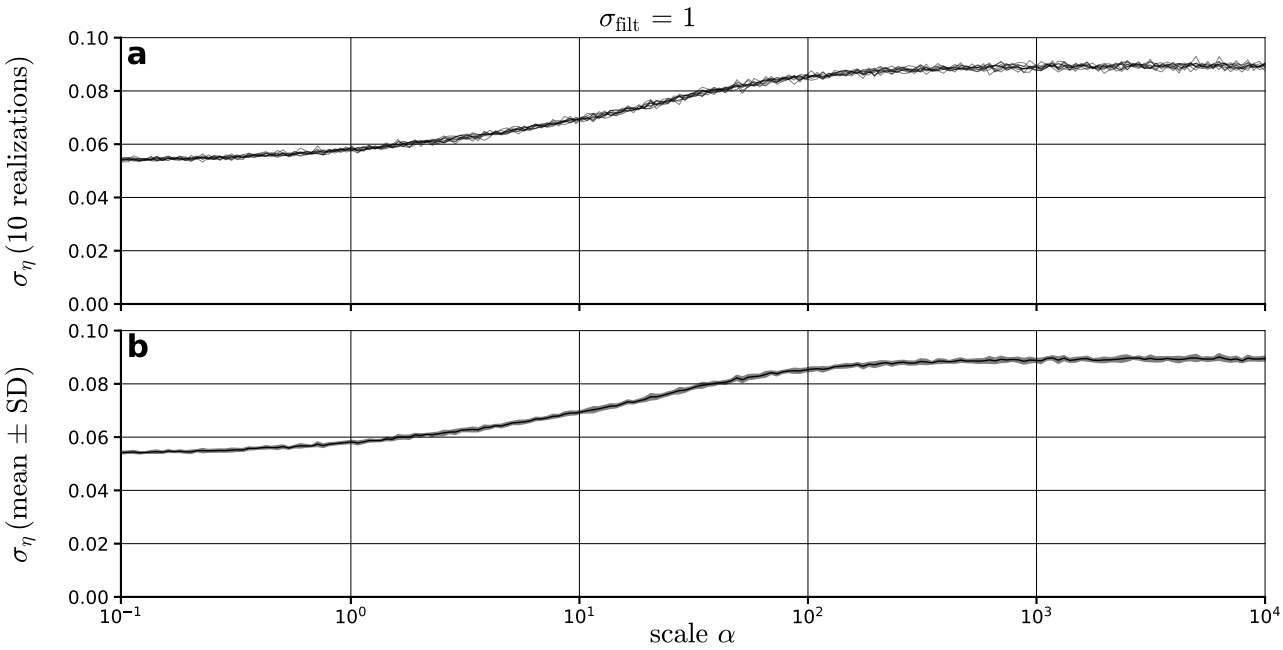
$\sigma_{\mathrm{filt}} = 1$
a
b
σ_η (10 realizations)
σ_η (mean ± SD)
0.00
0.02
0.04
0.06
0.08
0.10
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
10^{4}
scale α

Fig. 8: