Neural Processing of Natural Sounds

Authors: Frédéric E. Theunissen and Julie E. Elie.
Department of Psychology and Helen Wills Neuroscience Institute. UC Berkeley.
e-mails: theunissen@berkeley.edu and julie.elie@berkeley.edu

TOC Summary

Natural sounds have unique statistical structure that makes them perceptually salient. The auditory system is finely tuned to this natural structure. Selective neurons found at the higher levels of the auditory system respond selectively to vocalizations used in communication.

On-line Summary

1. Natural sounds include animal vocalizations, environmental sounds such as wind, water and fire noises and non-vocal sounds made by animals and humans for communication. These natural sounds have characteristic statistical properties that make them perceptually salient and that drive auditory neurons in optimal regimes for information transmission.

2. Recent advances in statistics and computer sciences have allowed neuro-physiologists to extract the stimulus-response function of complex auditory neurons from responses to natural sounds. These studies have shown a hierarchical processing that leads to the neural detection of progressively more complex natural sound features and have demonstrated the importance of the acoustical and behavioral contexts for the neural responses.

3. High-level auditory neurons have shown to be exquisitely selective for conspecific calls. This fine selectivity could play an important role for species recognition, for vocal learning in songbirds and, in the case of the bats, for the processing of the sounds used in echolocation. Research that investigates how communication sounds are categorized into behaviorally meaningful groups (e.g. call types in animals, words in human speech) remains in its infancy.

4. Animals and humans also excel at separating communication sounds from each other and from background noise. Neurons that detect communication calls in noise have been found but the neural computations involved in sound source separation and natural auditory scene analysis remain overall poorly understood. Thus, future auditory research will have to focus not only on how natural sounds are processed by the auditory system but also on the computations that allow for this processing to occur in natural listening situations.

5. The complexity of the computations needed in the natural hearing task might require a high-dimensional representation provided by ensemble of neurons and the use of natural sounds might be the best solution for understanding the ensemble neural code.

Preface

We might be forced to listen to a high frequency tone at our audiologist's office or we might enjoy falling asleep with a white-noise machine but the sounds that really matter to us are the voices of our companions or music from our favorite radio station; the auditory system has evolved to process behaviorally relevant natural sounds. Research has shown not only that our brain is optimized for natural hearing tasks but also that using natural sounds to probe the auditory system is the best way to understand the neural computations that allow us to comprehend speech or appreciate music.

Introduction

Until the late 1990s auditory neuroscientists could be divided into two camps. In one camp, following the tradition of the great physicist and sensory physiologist Hermann von Helmholtz[1], classical auditory neurophysiologists used simple synthetic sounds such as pure tones (sound wave generated by a perfect sinusoidal oscillator) to probe the nature of neural responses in the auditory systems. Indeed, just as the Helmholtz resonator separated multi-tone sounds into its frequency components, the principal role of the auditory portion of the inner ear, the cochlea, is to decompose the sound waveform into separate frequency bands[2]. Thus, it is not surprising to learn that auditory neurons responses, at least at the lower levels of the auditory system, have been described and understood in terms of their responses to pure tones of a given frequency (e.g. [3])(Figure 1, upper row). In that classical approach, the frequency tuning curve of auditory neurons takes on a central role and more complex responses are described in terms of specific deviations from the linear summation rule, otherwise known as non-linear responses or contextual effects (reviewed in [4]).

In the second camp, following the tradition of the great ethologist Konrad Lorentz, auditory neuroethologists studied how natural sound stimuli that lead to specific behaviors are represented in the auditory system. One of the key findings from the neuroethologists camp was the discovery of neurons that responded very strongly to natural and behaviorally significant sounds but not necessarily to their simpler components[5-7] (Figure 1, bottom row). In other words, the stimulus-response function describing the neural tuning in these neurons is dominated by the non-linear or contextual effects and not by the frequency tuning curve[8]. Moreover, it appeared that the appropriate "auditory context" to probe the neural system was the natural one. Note that these two "camps" were not antagonistic and we use this term to stress the differences in the two approaches and the contrasting shortcomings of each. There was however relatively little discussion between researchers in each "camp".

Each of these approaches has distinct merit from a methodological viewpoint. The reductionist approach of classical auditory physiologists allows a systematic parameterization of sound stimuli and therefore a clear method for synthesizing stimuli to explore specific mechanistic hypotheses. However, the relevance of results obtained from sounds that an animal rarely hears could always be questioned. Thus, conclusions about the implications of the results for processing behaviorally relevant complex sounds could be criticized as being post-hoc explanations that lack the strength of experimental predictions. In contrast, the behavioral relevance of the neuroethological approach was clearly less problematic and results showing that behaviorally relevant stimuli yield the largest neural responses [7](or the most informative [9,10]) gave support for evolutionary arguments and ultimate explanations; in other words, that the auditory system evolved to optimally process the sounds that matter the most for the survival of the species. However, the lack of a reductionist methodology in the neuroethologists' approach limited the exploration of underlying mechanisms.

One of the recent advances in auditory sciences has been in the merging of these two camps. This merging has been facilitated both by advances in computational approaches used for both sound

and neural data analyses and by advances in experimental techniques. We will review these recent developments. In a first section, we will summarize what we have learned from the statistical analyses of natural sounds. Describing these statistics is important not only to define what is unique about natural sounds but also because this knowledge is needed to analyze the neural responses to these sounds and to determine whether or not the auditory system has evolved to process such sounds in an optimal fashion. In the second section, we will then explain how the use of more recent machine learning techniques has allowed researchers to take into account this statistical structure when estimating neural tuning functions from responses to natural sounds. The third section will focus on the processing of communication calls, the vocalizations emitted by animals in the context of information exchange, that are particularly well represented in the auditory system. In the last section, we will review how progress in experimental methods has also allowed researchers to study hearing processes in more natural listening conditions. Our review will not cover the extensive body of research that specializes in human speech processing and its neural correlate except when general principles are considered and clear parallels can be made.

Statistics of Natural Sounds.

What is a natural sound? The question is particularly relevant given the increased prevalence of anthropomorphic noise in our daily environments that was absent during much of evolution. Natural sounds can be defined as 1) environmental sounds not generated by human made machines, such as the sounds of footsteps, wind, fire and rain, 2) all animal vocalizations, including human speech, and 3) other sounds generated for communication by animals, such as stridulation in crickets[11], buttress drumming by chimpanzees[12] and instrumental music in humans. We will first investigate the properties of isolated sounds and then briefly touch on the statistics of sound mixtures.

Perceptually relevant physical characteristics of isolated natural sounds follow a power law.

Although natural sounds defined in this broad sense have heterogeneous properties they share structure that can be quantified by ensemble statistical analyses (Figure 2). More specifically, it has been observed that the frequency spectra of certain fluctuating physical characteristics of natural sounds follow a 1/f or, more generally, a power law relationship. In other words, some physical characteristic of natural sounds (φ) varies as the power of the frequency (f) such as $\varphi(f) = \alpha f^{-\kappa}$ with α and κ positive constants. It should be clearly noted that this relationship does not hold for the sound spectrum itself but, instead, for slower varying structure such as loudness, measured in the temporal envelope of the sound (Fig. 2, Temporal Envelope Spectrum), or in time varying pitch ("height" of the sound) profiles [13-15]. The power law relationship also holds for the power spectrum of the log of the sound spectrum [15]. This transformation of the sound waveform, called the cepstrum [16], is used to extract spectral structures in the sounds, structures in frequency domain, such as speech formants. Moreover, it has been shown that the frequencies of temporal and spectral modulations in the spectrogram, known as the Modulation Power Spectrum (Fig 2), exhibits specific dependencies beyond those expected from the time-frequency trade-off [17]; natural sounds and vocalizations in particular have higher power at joint low temporal and high spectral modulations frequencies than expected from the product of the marginals: the average power for the same temporal modulation (averaged over all spectral modulation frequencies) multiplied by the average power for the same spectral modulation (averaged over all temporal modulation frequencies) [15]. In other words, many animal vocalizations are dominated by relatively slow sounds with fine harmonic structure.

Physical, behavioural and neural implications of the power law structure.

What are the physical, behavioural and neural implications of this naturally occurring acoustical structure? First, in terms of physical properties, the power law relationship for time varying signals implies that natural sounds have correlations over multiple time scales including very long ones, as reflected by the large energies at low frequencies. In this sense, natural sounds are clearly different from signals that are completely random or uncorrelated, such as white noise signals with flat sound and temporal envelope power spectra, or, at the other end of this spectrum of correlations, signals dominated by a single correlation time, such as those created by a perfect oscillator (e.g. a sound with sinusoidally varying amplitude as in some car alarms). It has also been argued that neither white-noise (a random signal with equal power at all frequencies) nor a pure sine wave can qualify as complex and, thus, information rich or perceptually sophisticated[13,18]. Second, in terms of behaviour, it is interesting to note that the fluctuating physical sound characteristics that show the power law characteristics in natural sounds are those that are directly linked to perceptual attributes. Whereas we are unable to perceive the details of the sound pressure waveform, the time varying amplitude yields a percept of intensity fluctuations, rhythm and timbre; the time varying pitch profile carries the melody in a musical phrase; and the spectral envelope contains critical information for other timbral qualities of sound including speech formants (the high-energy frequency bands in voiced human speech that code the identity of vowels and other phonological information) [19]. Finally, these observed natural statistical structures have implications in terms of neural coding. For example, sound stimuli that have such natural statistics elicit higher information rates measured in auditory neurons relative to matched synthetic sounds which otherwise lack some of the natural statistics[9,10,20]. Interestingly, the spatial and temporal luminance contrast in natural visual scenes also obey power law relations that have also been related to complexity but that are primarily the result of scale invariance [21,22]. This power law relation (1/f) implies that visual scenes have stronger correlations at low spatial and temporal frequencies than at higher frequencies. It has also been shown that early processing in the visual system can reverse this relationship by attenuating low frequencies and boosting high frequencies effectively removing the correlations present in the stimulus images[23]. Such decorrelation is useful to maximize information transmission through a bottleneck such as the optic nerve. Although the physical causes of the power-law relationship observed in natural images and in natural sounds are unrelated, it is highly probable that similar neural efficiency principles apply in both sensory systems. And indeed, a similar decorrelation has been observed in the inferior colliculus where the gain of auditory neurons emphasizes higher temporal and spectral modulation effectively counterbalancing the 1/ f relationship observed in the modulation power spectrum (and not the sound spectrum) of natural sounds[24]. At higher levels of the auditory processing stream, it has been shown that the population of neurons have a maximum gain at intermediate modulation frequencies, in a region that is particularly useful for distinguishing among different natural sounds[25].

The sound spectrum is idiosyncratic for each natural sound class.

As mentioned above, natural sounds exhibit a power-law relationship in the spectrum of particular time-varying features of sounds such as intensity, and this relationship has physical, perceptual and neural implications. This power-law relationship does not exist for the sound spectrum itself since each natural sound class has an idiosyncratic sound spectrum. This does not mean, however, that auditory systems are not sensitive to particular shapes of the sound spectrum of behaviorally relevant sounds. On the contrary, frequency-tuning sensitivity has been shown to be one of the major factors ensuring the sender-receiver match. The neuro-ethological basis of this matched frequency tuning has been particularly well documented in insects and anurans [26]. And, more strikingly perhaps, this frequency tuning adaptation has even been observed in the cochlea of owls [27] and bats [28] where the region of the cochlea that is mechanically tuned to frequencies that are

particularly relevant for the animal is expanded in what has been called an auditory fovea. As we will discuss in more detail in the section of animal vocalizations, the adaptation of the auditory system to the specific structure of conspecific communication calls might be equal or maybe even greater than putative adaptations to more general natural sound statistics.

Natural sounds statistics and the frequency tuning of mammalian auditory nerve fibers.

Looking beyond the matched-tuning found in auditory specialists such as bats and owls, could general natural sound statistics also explain the prototypical frequency tuning observed in the peripheral mammalian auditory system? Primarily as a result of its mechanical properties, the cochlea decomposes sounds into a set of signals centered at increasing frequencies by applying filters of different shapes: narrow band frequency filters for low frequencies and large band frequency filters for high frequencies [2]. Since the frequency power spectrum of specific natural sounds is idiosyncratic, a simple frequency matched-tuning or decorrelation argument cannot be used to provide an adaptive explanation for this relationship. Instead, however, an examination of both the temporal and spectral statistics of different classes of natural sounds can provide an explanation. In particular, environmental sounds and animal vocalizations make two well-defined groups of sounds with different statistical structure: animal vocalizations are dominated by sustained harmonic sounds while environmental sounds are dominated by transient sounds [29]. It has been argued that the shape of the mammalian auditory frequency filters measured at the level of the auditory nerve are optimal at representing the independent components of combinations of animal vocalizations and environmental sounds: the lower frequency narrow-band filters efficiently represent the relatively long but spectrally sharp animal vocalizations and the higher frequency broad-band filters efficiently represent the short but broad environmental sounds. The human speech signal is particular in that it combines sounds from these two classes. One can thus postulate that the physical characteristics of human speech have evolved to be optimally represented at the auditory periphery (while clearly taking into account other constrains) [29].

Statistics of sound mixtures.

Isolated sounds have interesting properties but our brains are more often exposed to complex auditory scenes. Sound mixtures, such as those created by a chorus of insects or a crowd in a loud restaurant, have also their own statistical signature: specifically, the structure that is present in the modulation power spectrum of isolated vocalizations is washed out in sound mixtures whereas the long-time average sound spectrum of isolated sound signals and their mixtures remain similar. Given that the modulation power spectra of background sounds differs from that of foreground sounds, a modulation filter bank - a set of filters in the spectral and temporal amplitude modulations domain - tuned to these differences could be used to separate signal from noise resulting from sound mixtures and such mechanism might be in place in secondary auditory cortical areas [30]. Because the sound spectrum of mixtures and signal are similar, this task would be impossible with a simple frequency filter bank – a set of filters in the sound frequency domain. Sound mixtures also appear to be processed separately from isolated sound signals: whereas the short time detail of isolated sound signals is perceived with high accuracy (allowing for example rather extreme rates of phoneme perception), sound mixtures are perceived and categorized in terms of their long-term statistical properties yielding percepts of sound "texture", defined as the collective result of many similar acoustic events (e.g. rainstorms, insect swarms) [31].

In summary, natural sounds have characteristic statistical properties that can be measured at different levels. All natural sounds have particular slow physical properties such as loudness profiles that obey power law relationships. The sound spectrum does not obey this law but its shape is

nevertheless an important property for the specialized processing of behaviorally relevant sounds. Natural sounds are easily categorized between animal vocalizations and environmental sounds based on differences in terms of relevant time-frequency scales. Sound mixtures lose the fine spectro-temporal modulations seen in isolated natural sounds and are better characterized and perceived in terms of long-term statistical properties. Both the nature of our perception of sounds and neural responses in the auditory system are sensitive to these natural sound statistics.

Stimulus-Response Characterizations.

As explained above, natural sounds are clearly both relevant and efficient stimuli to drive auditory neurons. Moreover, using either theoretical arguments to model processing in the auditory periphery [29,32] or information theoretic measures of empirical data [9,10], the auditory system appears to have evolved for optimal processing of sounds with such statistical properties. These studies, however, shed little light on the actual underlying mechanisms when compared to the explanations provided by the characterization of neuronal stimulus-response functions: *i.e.*. the mathematical formulation that describes how single neurons or neuronal ensembles respond to any given stimulus.

Estimating stimulus-response function using synthetic sounds.

Traditionally, stimulus-response characterizations had been performed with synthetic sounds that would allow the systematic probing of the effect of a single acoustical parameter (e.g. frequency) on neural responses. System identification analysis, the functional description of any arbitrary input-output system, also relied heavily on the use of synthetic sounds and primarily Gaussian white-noise[33]. Noise-like stimuli allow not only an efficient exploration of a large set of possible sounds (e.g. all frequencies within the noise band) but also facilitate the estimation of a neuron's stimulus-response function; with white-noise, the average stimulus before each spike (the spike-triggered average or STA) yields the impulse response of the neuron or, when stimuli are represented in spectrographic form, the neuron's spectro-temporal receptive field (STRF) [34,35]. For neurons that respond linearly to sound features as represented in a spectrogram, the STRF shows the spectro-temporal pattern that would result in the highest firing rates. When the STRF is used as a model, the convolution (a mathematical operation akin to a running-time correlation) between the STRF and the sound spectrogram yields a predicted neural response. The STRF model can be generalized to model the response of any neuron by incorporating non-linear components, as we will describe in more detail below. At lower levels of the auditory processing stream, where neurons are less sensitive to contextual effects, the white-noise approach can yield accurate estimations of the stimulus-response function [36-38]. In these cases, white-noise analyses are used to estimate the stimulus-response function and these functions in turn can explain selectivity for specific vocalizations or the efficient representation of natural sounds in general [29]. At higher levels of the auditory system, however, neural responses can be dominated by contextual effects [8,39,40]; although sound features that drive neurons might be present in white-noise, they might only elicit responses when they are presented in a natural acoustic context, for example, following silence, or following a sequence of other particular sounds or presented jointly with other specific sounds. In other words, neurons become tuned to more complex spectro-temporal patterns that are characteristic of natural sounds but that are poorly sampled in white noise. In those cases, stimulus-responses functions can only be estimated using the appropriate context: behaviorally relevant natural sounds.

Methods for estimating STRFs using natural sounds.

Fortunately, advances in regression techniques and machine learning have allowed the estimation of STRFs using natural sounds (Fig. 3). Great progress has been made on four critical issues. First, natural sound ensembles occupy a limited region of the entire space of possible sounds.

One must therefore be aware that the shape of the estimated STRF will depend on the subset of sounds being sampled and is only valid for sounds sharing the same characteristics as the sampled one. In this case, this issue is simply solved by clearly describing the sampled subset in the space that is relevant for STRFs. For example, if the STRFs are based on spectrograms, the phase and amplitude of the modulation spectrum (i.e. the spectrum of the spectrogram) will describe how the selected natural sounds sample the space [15]. In addition, when one compares STRFs estimated with two distinct sound ensembles (whether they are natural or synthetic), one needs to carefully estimate the STRFs by effectively only using sound features that are found with sufficient frequency in both sound subspaces [41]. Second, in natural sounds, the time-averaged energy of sound features (or equivalently the average intensity and frequency of occurrence) is not uniform through out the subset of sounds being sampled. For this reason, simple estimation techniques such as the STA, which is a straight averaging operation, will yield biased estimates of the STRF. This bias can be removed using the appropriate normalization techniques. These normalization techniques can be thought as a weighted average operation where sound features that are sampled more infrequently are given more weight to compensate for this under-sampling [42,43]. Third, again since natural sounds might only effectively span a small subset of possible sounds, one must carefully match the effective dimensionality of this sampling to the dimensionality of the sound representation. For example an STRF operating on the spectrographic representation of sounds might have 100 slices in time (eg. 100 ms window with 1 ms sampling) and 100 slices in frequency (100 Hz bands between 0 Hz and 10 KHz) for a total of 10,000 time-frequency "pixels", the parameters of the STRF model (the $h(\tau_k,f_j)$ on Fig. 3). Natural sounds might sample these 10,000 dimensions very sparsely yielding very poor estimates for all the 10,000 STRF parameters. This is a well known problem in statistics: if a model has too many parameters (here the number of time-frequency pixels of the STRF) compared to the number of observations (here the number of natural sounds and corresponding neural responses) then the model risks fitting not only the underlying relationship between the stimulus and the neural response but also random fluctuations of the particular data set. To prevent this phenomenon, known as overfitting, regularization techniques must be used [44]. Regularization adds constraints in the form of priors on the model parameters that effectively impose a penalty on model complexity. For example, Principal Component (PC) regression (or subspace regression) and ridge regression implement zero-mean Gaussian priors on the STRF coefficients with a variable variance. By setting a small variance on this prior, STRF parameters will be estimated to be very close to zero during model fitting procedure unless there is robust evidence that they contribute significantly to the prediction of the neural response. PC regression and ridge regression have also analytical solutions (solutions that can be found by solving a mathematical equation) that are computationally very efficient [43,45]. Regularization can also be implemented using other priors on STRF coefficients and iterative algorithms [44,46]. Fourth, the stimulus-response functions of high-level auditory neurons are often dominated by non-linearities that are not captured in the STRF, which, in its simplest form as a model, predicts neural responses from a linear combination of spectro-temporal features. Estimating the nature of the non-linearities is not only important to fully capture the computations performed by the system but is important as they might impact the estimation of the linear STRF with natural stimuli [47,48]. There are many approaches to this problem. Input non-linearities can be incorporated in the chosen representation for the sound stimuli. For example, sound representations can include known non-linearities such as adaptive mechanisms [49,50] or probabilistic expectations [51]. Output non-linearities such as those produced by a spiking threshold can be very efficiently modeled using the generalized linear framework [52] even in combination with input-nonlinearities [53]. Finally, dynamical second order or higher-order non-linearities have been estimated with techniques that yield multi-component STRFs [47,54-57].

The computations in the auditory processing stream revealed by the STRFs.

These methodological advances have allowed auditory neuroscientists to make significant progress in understanding the nature of the auditory computations that are found in the ascending processing stream of both birds [30,41,48,51,58-61] and mammals [38,62-65]. Selectivity for natural sounds is already present at the level of the inferior colliculus (IC) in the sense that IC STRFs show temporal spectral features that are found in behaviorally relevant sounds [38,41,66-68]. Then, novel type of STRFs appear at the level of the primary auditory cortex and one can understand these as models achieving selectivity for more complex and slower acoustical features compared to the simpler STRFs found in the IC and thalamus (principal relay of sensory inputs from the sensory periphery to the cortex) [59,62,63] (see also fig. 4). These changes in STRF go hand in hand with increased selectivity for natural sounds [9,20,25]. Contextual effects also become more important at the higher levels of the auditory system [40-42,48,51,69,70]; these contextual effects manifest themselves as changes in the selectivity for spectro-temporal features due to the presence of particular sounds "outside" a classically estimated STRF [48,71], changes due to expectations about stimulus statistics [51], changes in correlated properties measured in ensemble neurons [41] and changes due to learning and behavioral relevance [69,72]. Again, it is postulated or shown that these auditory contextual effects increase the efficiency of the neural representation for behaviorally relevant natural sounds either at the single [51,69,72] or population level [41,73]. Finally, researchers have begun to understand how complex stimulus-response function found at higher levels of the auditory processing could be used to achieve complex auditory tasks that go beyond "template-matching" between a STRF and an acoustical feature present in natural sounds. For example, the multi-scale time-frequency modulation tuning of the auditory cortex can be used to separate bird song or speech from non-speech signals or noise [30,74,75].

As an alternative to the estimation of linear and non-linear STRFs from responses to natural sounds, researchers have also used synthetic sounds designed to have particular natural statistics. Families of such synthetic natural-like sounds can then be used to isolate the specific natural feature that is particularly important for understanding behavioral or neural responses. This approach has been used, for example, to elucidate the natural sound features critical for phonotaxis in crickets [11], sound texture perception in humans [31,76] and selectivity for conspecific songs in songbirds [9,15].

In summary, analytical and computational advances have allowed auditory researchers to use natural sounds or synthetic natural-like sounds to estimate the stimulus-response functions of high-level auditory neurons. In doing so, they were able not only to extract these functions for neurons that did not respond to white noise or other synthetic stimuli but they were also able to investigate auditory contextual effects and the nature of the computations that generated selective responses for natural sounds.

Animal Vocalizations.

Animal vocalizations as a class of natural sounds have played and continue to play an important role in auditory neurosciences. Historically, the first use of natural sounds in auditory neuroscience came from neuro-ethologists who investigated how conspecific vocalizations or communication signals were selectively processed in the auditory system of auditory specialists. These investigations in model systems led to the discovery of cricket-song selective neurons and their contribution to the females' phonotaxis behavior [77], of call selective neurons in frogs [78] and guinea pigs [79], of song selective neurons in songbirds [7,80-82], of neurons selective to the echolocation signal in bats [5], and of brain regions selective for conspecific calls in primates [83]. Selectivity for conspecific communication calls can be reflected not only in the mean rate of single neurons but also (and sometimes only) in time-varying responses [84,85] or ensemble responses [86,87]. Thus, the auditory system is not only selective to natural sounds in a broad sense but appears

to also exhibit specialized circuitry for the sole purpose of detecting and processing conspecific communication calls. One of the striking results from this line of research has been the relatively high degree of selectivity that has been measured in these vocalization selective neurons [6,88]: systematic manipulations of bird song syllables and bat echolocation calls have shown that this selectivity is achieved by non-linear mechanisms that detect specific temporal or spectral combination of sound features that are uniquely present in specific conspecific vocalizations [5,6,89].

Although such acute selectivity might be useful for auditory tasks that require high fidelity such as the processing of echolocation pulses or for guiding vocal commands in song learning, its utility for processing sounds in terms of their communicative value is more problematic. For example, both primates and songbirds produce alarm calls that need to be categorized correctly in order to guide the appropriate behavior. Such categorization requires invariant responses for all communication calls belonging to the same category as well as recognition of category boundaries [90]. Thus, auditory processing for communication purposes might require not only low-level feature detection processing but also categorization of higher-level structure. Such high-level categorization might involve hierarchical processing steps such as the representation of particular sound features (e.g. formant frequency) that are robust to variation in other physical parameters of the sound (e.g. azimuthal location) and such responses have been found in secondary mammalian and avian auditory areas [91,92]. In terms of higher-level categorization, research in starlings points to a role of the Caudio-Medial Nidopallium (NCM) for classifying behaviorally relevant classes of songs [93] and research in primates suggests that both the Superior Temporal Gyrus (STG) and the ventrolateral Prefontral Cortex (vPFC) could be involved in semantic discrimination [94-98]. Similar cortical areas have been shown in a large body of research to be critical for human speech processing [99]. But it is fair to state that our understanding of the neural mechanisms that generate such high-level categorization of sounds is still at its infancy. Songbirds who have a large repertoire of communication calls that are used in distinct behavioral contexts could also become a powerful animal model to study the neural computations involved in the categorization that is needed in order to extract meaning from variable communication sounds [100,101].

The ontogeny of selective neural responses for vocalizations has also been studied extensively. Although many animal communication calls are innate or have innate characteristics, neural selectivity in the perceptual system for innate calls could arise during development simply as a result of experience and repeated exposure. Moreover vocalizations show learning components both in production, as is the case for song in songbirds [102], and in perception, as it is the case for the interpretation of alarm calls in primates [103], the interpretation of pup calls in mothers versus virgin mice [104], the discrimination of familiar versus unfamiliar contact calls in zebra finches [105], the recognition of individual songs in starlings [106]. Not surprisingly, selective neural responses for natural sounds have been shown to have strong developmental and environmental components; this is true both for lower level selectivity such as that found in primary auditory cortical areas [73,107-109] as well as for the higher-level selectivity found in sensori-motor areas of songbirds [110-112]. Experience during development can also affect perceptual boundaries and their putative neural correlates [113] .

In summary, on one hand, the initial study of the neural representation of conspecific vocalizations in the auditory system has played a crucial role for advancing our understanding of the nature of the non-linear neural responses that are found in the higher auditory areas and for establishing the need to use behaviorally relevant sounds to decipher these computations. On the other hand, research on the nature of invariant representations for vocalization classes and on the link between sound and the perception of meaning is still in its infancy and research in this area could further advance our understanding of the neural mechanisms involved in human speech processing. For example, categorization of sounds for lexical retrieval or for voice recognition requires a combination of filtering (to ignore irrelevant features) and grouping (to allow for variation in the

coding features) that only more complex and non-linear STRFs could achieve. Finally, it is clear that selectivity for natural sound features and vocalizations have both innate and learned components and the relative importance of each factor is an active area of research.

Towards Natural Hearing

Most of the neurophysiology research described above relied on the passive playback of isolated sounds in animals that were either anesthetized or restrained. But natural hearing often involves attention and action on the part of the sender and receiver, such as in bat echolocation [114], the interpretation of alarm calls originating from different individual [115], or communication between mates in bonding behaviors [116]. Moreover, natural hearing also involves the processing of complex auditory scenes. Until recently, the natural sounds that have been analyzed or used in laboratory experiments have been mostly free of natural noise or natural degradations. In the real world, communication signals are most often perceived in unfavorable listening conditions with noise background, distortions due to propagation and echoes [117] and, superposition from other potential acoustical signals [118]. Vocal communication and auditory perception is also affected by the social context, such as in the audience effect [119] or by internal states, such as stress levels. These social and emotional cues can also be mediated by other sensory modalities.

Advances in chronic neural recording techniques have allowed researchers to begin to examine neural processing in these more natural scenarios. Researchers have shown how responses in primary auditory cortex are influenced both by expectations of natural structure in the sound and behavioral relevance, both of which might involve top down modulations [72]. Chronic recordings in awake and vocalizing animals have also been used to obtain neural recordings in auditory areas to the animal's own vocalizations. Such experiments have been performed in bats [120,121], primates [122] and birds [123]. The experiments in bats were crucial to understand how the pulse-echo pair was processed by the auditory system and constitute landmark experiments in that field. In primates and birds, these awake-recordings gave us unique insights on how self-vocalizations are processed for self-monitoring and, in birds, potentially for guiding vocal learning. However, neural recordings in both sender and receivers in the midst of vocal communication bouts such as antiphonal calling in marmosets [124] or duets in social songbirds [116,125] have not yet been performed. Such experiments could be performed in the near future and are needed to advance our understanding of the computations performed in the auditory system for extracting the information content of communication calls.

The auditory processing of communication sounds in noisy backgrounds or in complex auditory scenes is also an active research area [126]. For example, noise invariant neurons, that is neurons which response to a given stimulus is not influenced by the presence of background noise , have been described in the secondary auditory areas of songbirds [30,127] and in primary auditory areas in humans [128]. Noise-invariance has also been shown to emerge in the auditory processing stream as a result of adaptive mechanisms for particular stimulus statistics [129]. Similarly, responses in human auditory cortex use a gain control to emphasize the temporal modulations characteristic of speech. Neurophysiological studies in primates [130,131] and birds [132] have also begun to unravel how multiple auditory streams could be represented in the auditory system.

In summary, auditory neuroscientists have mostly focused their attention to understanding the computations needed to passively recognize and categorize natural sounds and much more research is needed to understand how acoustical signals are processed in active communications and in natural soundscapes. Neurophysiological research in this area is in its infancy but, given the increase in our knowledge achieved from classic playback experiments and the technical advances in chronic recordings, natural hearing research is poised to make giant leaps in the near future.

Conclusion

The use of natural sounds (and in particular conspecific sounds) has had a long tradition in neuroethological research and the findings in these model systems have inspired the more recent development of analytical techniques for both sound analysis and neural data analysis. These developments have allowed auditory neuroscientists to use natural sound stimuli to describe and understand in much greater detail the neural responses of higher-level auditory neurons in both specialists such as crickets, bats and songbirds and in generalists such as guinea pigs, cats, ferrets and non-human primates. Sounds with natural statistics appear to be optimally represented at multiple levels of the auditory system and stimulation with natural sounds facilitated the characterizations of stimulus-response function for neurons that respond poorly to white noise or other simple synthetic sounds. Thus, for the systematic characterization of stimulus-response function, the need to use simple synthetic sounds is not longer required and should even be discouraged. On the other hand, complex synthetic sounds that preserve particular natural statistics and that are designed to systematically investigate the importance of natural statistics provide an additional and powerful insight. Auditory neuroscientists have also been able to begin to relate auditory representations to specific computations needed for recognizing and categorizing behaviorally relevant sounds, such as communication calls.

These past successes will facilitate the design and interpretation of even more naturalistic experiments. In the near future, we see five areas of promising scientific explorations: non-linear computations for invariant representation of communication calls, neurophysiological research in humans, auditory scene analysis, social and multi-modal effects, and investigations of the ensemble neural code. First, neural recordings in animals that actively communicate with other animals will permit both the natural investigation of robust neural representation for call types and a direct assessment of the relationship between sound and meaning. Second, advances in both invasive [133,134] and non-invasive [135] neurophysiological recordings in humans will further allow researchers to make links between animal work and human work. Given the wealth of knowledge in speech and music processing in humans, these links will greatly help with the challenge of understanding the sound to meaning transformations occurring in the auditory system. Third, it is still unclear how the auditory system detects, recognizes and classifies behaviorally relevant signals with degraded signals and multiple sound sources; neural recordings not only with natural sounds but in the natural environment (i.e. in the field) could be performed to study how naturally propagated and corrupted signals are represented. This line of research, however, will also require the statistical characterization of natural auditory scenes, which is a particularly challenging problem. Fourth, we know that communication behavior and auditory perception depends on the social and emotional context and that the physiology of the auditory system can be significantly modulated by hormones [136]. But how the neural code for natural sounds is affected by naturalistic stimuli from other modalities such as vision or self-motion [137] or by modulatory effects from brain systems involved in emotional or stress responses remains in large part unexplored. Finally, although neurophysiologists are now regularly recording the simultaneous activity of many neurons, the role of correlated activity in the ensemble neural code is still unknown [138]. One apparently insurmountable difficulty for studying the ensemble neural code is the explosion in the dimensionality of the problem as a result of combinatorial effects: the number of potential neural activity patterns across neurons becomes so large that investigating the potential role of such patterns becomes impossible. For example, if a single neuron can reliably represent information with 10 different patterns, the code from two such neurons could represent 100 patterns, the code from three neurons 1000 patterns, and so forth. For these combinatorial ensemble patterns to carry unique information about the stimulus (i.e. information beyond the one obtained from the individual responses), the response in one neuron must be correlated with the response in another neuron. A

recent statistical analysis of neural patterns recorded in visual system under natural stimulation showed that ensemble neural responses are indeed correlated but very sparse [139]. In other words, natural scenes appeared to be represented with very few response patterns from all possible combinations that could be possible. These experiments and analyses suggest that using natural stimuli might be the only way out for resolving this dimensionality curse. Although this position might be extreme, given the important role that natural sounds have already played in understanding the auditory system and that questions in natural hearing will require further investigations with natural sounds, auditory neuroscientists might also be well placed to elucidate the nature of ensemble neural code in sensory systems.

References

1. Darrigol O (2003) Number and measure: Hermann von Helmholtz at the crossroads of mathematics, physics, and psychology. Studies in History and Philosophy of Science 34A: 515-573.
2. Robles L, Ruggero MA (2001) Mechanics of the mammalian cochlea. Physiological Reviews 81: 1305-1352.
3. Woolley SM, Casseday JH (2004) Response properties of single neurons in the zebra finch auditory midbrain: response patterns, frequency coding, intensity coding, and spike latencies. J Neurophysiol 91: 136-151. Epub 2003 Oct 2001.
4. Eggermont JJ (2001) Between sound and perception: reviewing the search for a neural code. Hear Res 157: 1-42.
5. Suga N, O'Neill WE, Manabe T (1978) Cortical neurons sensitive to combinations of information-bearing elements of biosonar signals in the moustache bat. Science 200: 778-781.
6. Margoliash D, Fortune ES (1992) Temporal and harmonic combination-sensitive neurons in the zebra finch's HVc. J Neurosci 12: 4309-4326.
7. Margoliash D, Konishi M (1985) Auditory representation of autogenous song in the song system of white-crowned sparrows. PNAS USA 82: 5997-6000.
8. Nelken I, Rotman Y, Bar Yosef O (1999) Responses of auditory-cortex neurons to structural features of natural sounds. Nature 397: 154-157.
9. Hsu A, Woolley SM, Fremouw TE, Theunissen FE (2004) Modulation power and phase spectrum of natural sounds enhance neural encoding performed by single auditory neurons. J Neurosci 24: 9201-9211.
10. Rieke F, Bodnar DA, Bialek W (1995) Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. Proc R Soc Lond B Biol Sci 262: 259-265.
11. Hedwig B (2006) Pulses, patterns and paths: neurobiology of acoustic behaviour in crickets. Journal of Comparative Physiology a-Neuroethology Sensory Neural and Behavioral Physiology 192: 677-689.
12. Arcadi AC, Robert D, Boesch C (1998) Buttress drumming by wild chimpanzees: Temporal patterning, phrase integration into loud calls, and preliminary evidence for individual distinctiveness. Primates 39: 505-518.
13. Voss RF, Clarke J (1975) 1/f noise in music and speech. Nature 258: 317-318.

14. Attias H, Schreiner CE (1997) Temporal low-order statistics of natural sounds. Advances in Neural Information Processing Systems 9: 27-33.
15. Singh NC, Theunissen FE (2003) Modulation spectra of natural sounds and ethological theories of auditory processing. J Acoust Soc Am 114: 3394-3411.
16. Chen JD, Paliwal KK, Nakamura S (2003) Cepstrum derived from differentiated power spectrum for robust speech recognition. Speech Communication 41: 469-484.
17. Cohen L (1995) Time-Frequency Analysis. Englewood Cliffs, New Jersey: Prentice Hall. Chapter 3 Section 4. p50-52
18. Bialek W, Nemenman I, Tishby N (2001) Predictability, complexity, and learning. Neural Computation 13: 2409-2463.
19. Elliott TM, Hamilton LS, Theunissen FE (2013) Acoustic structure of the five perceptual dimensions of timbre in orchestral instrument tones. Journal of the Acoustical Society of America 133: 389-404.
20. Garcia-Lazaro JA, Ahmed B, Schnupp JWH (2011) Emergence of Tuning to Natural Stimulus Statistics along the Central Auditory Pathway. Plos One 6.
21. Srivastava A, Lee AB, Simoncelli EP, Zhu SC (2003) On advances in statistical modeling of natural images. Journal of Mathematical Imaging and Vision 18: 17-33.
22. Ruderman DL (1997) Origins of scaling in natural images. Vision Research 37: 3385-3398.
23. Simoncelli EP, Olshausen BA (2001) Natural image statistics and neural representation. Annu Rev Neurosci 24: 1193-1216.
24. Rodriguez FA, Chen C, Read HL, Escabi MA (2010) Neural modulation tuning characteristics scale to efficiently encode natural sound statistics. J Neurosci 30: 15969-15980.
25. Woolley SM, Fremouw TE, Hsu A, Theunissen FE (2005) Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. Nat Neurosci 8: 1371-1379.
26. Gerhardt HC, Huber F (2002) Acoustic communication in insects and anurans: common problems and diverse solutions. Acoustic communication in insects and anurans: common problems and diverse solutions: University of Chicago Press. pp. i-xi, 1-531.
27. Koppl C, Gleich O, Manley GA (1993) An Auditory Fovea in the Barn Owl Cochlea. Journal of Comparative Physiology a-Sensory Neural and Behavioral Physiology 171: 695-704.
28. Bruns V, Schmieszek E (1980) Cochlear Innervation in the Greater Horseshoe Bat - Demonstration of an Acoustic Fovea. Hearing Research 3: 27-43.
29. Lewicki MS (2002) Efficient coding of natural sounds. Nat Neurosci 5: 356-363.
30. Moore RC, Lee T, Theunissen FE (2013) Noise-invariant Neurons in the Avian Auditory Cortex: Hearing the Song in Noise. Plos Computational Biology 9.
31. McDermott JH, Schemitsch M, Simoncelli EP (2013) Summary statistics in auditory perception. Nature Neuroscience 16: 493-U169.
32. Smith EC, Lewicki MS (2006) Efficient auditory coding. Nature 439: 978-982.
33. Marmarelis PZ, Marmarelis VZ (1978) Analysis of physiological systems: The white noise approach. New York, NY: Plenum.
34. Aertsen AM, Johannesma PI (1981) The spectro-temporal receptive field. A functional characteristic of auditory neurons. Biol Cybern 42: 133-143.

35. Eggermont JJ, Aertsen AM, Johannesma PI (1983) Quantitative characterisation procedure for auditory neurons based on the spectro-temporal receptive field. Hear Res 10: 167-190.
36. Eggermont JJ, Aertsen AM, Johannesma PI (1983) Prediction of the responses of auditory neurons in the midbrain of the grass frog based on the spectro-temporal receptive field. Hear Res 10: 191-202.
37. Schafer M, Rubsamen R, Dorrscheidt GJ, Knipschild M (1992) Setting complex tasks to single units in the avian auditory forebrain. II. Do we really need natural stimuli to describe neuronal response characteristics? Hear Res 57: 231-244.
38. Andoni S, Li N, Pollak GD (2007) Spectrotemporal receptive fields in the inferior colliculus revealing selectivity for spectral motion in conspecific vocalizations. J Neurosci 27: 4882-4893.
39. Ulanovsky N, Las L, Nelken I (2003) Processing of low-probability sounds by cortical neurons. Nat Neurosci 6: 391-398.
40. Asari H, Zador AM (2009) Long-Lasting Context Dependence Constrains Neural Encoding Models in Rodent Auditory Cortex. Journal of Neurophysiology 102: 2638-2656.
41. Woolley SM, Gill PR, Theunissen FE (2006) Stimulus-dependent auditory tuning results in synchronous population coding of vocalizations in the songbird midbrain. J Neurosci 26: 2499-2512.
42. Theunissen FE, Sen K, Doupe AJ (2000) Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. J Neurosci 20: 2315-2331.
43. Theunissen FE, David SV, Singh NC, Hsu A, Vinje W, et al. (2001) Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. Network: Comp Neural Syst 12: 1-28.
44. Sahani M, Linden J (2003) Evidence optimization techniques for estimating stimulus-response functions. In: Becker S, Thrun S, Obermeyer K, editors. Advances in Neural Infomation Processing Systems. Cambridge, MA: MIT Press. pp. 301-308.
45. Hoerl AE, Kennard RW (2000) Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 42: 80-86.
46. David SV, Mesgarani N, Shamma SA (2007) Estimating sparse spectro-temporal receptive fields with natural stimuli. Network-Computation in Neural Systems 18: 191-212.
47. Christianson GB, Sahani M, Linden JF (2008) The consequences of response nonlinearities for interpretation of spectrotemporal receptive fields. Journal of Neuroscience 28: 446-455.
48. Schneider DM, Woolley SMN (2011) Extra-Classical Tuning Predicts Stimulus-Dependent Receptive Fields in Auditory Neurons. Journal of Neuroscience 31: 11867-11878.
49. Gill P, Zhang J, Woolley SM, Fremouw T, Theunissen FE (2006) Sound representation methods for spectro-temporal receptive field estimation. J Comput Neurosci 22: 22.
50. David SV, Shamma SA (2013) Integration over Multiple Timescales in Primary Auditory Cortex. Journal of Neuroscience 33: 19154-19166.
51. Gill P, Woolley SM, Fremouw T, Theunissen FE (2008) What's That Sound? Auditory Area CLM Encodes Stimulus Surprise, Not Intensity or Intensity Changes. J Neurophysiol 99: 2809-2820.

52. Calabrese A, Schumacher JW, Schneider DM, Paninski L, Woolley SMN (2011) A Generalized Linear Model for Estimating Spectrotemporal Receptive Fields from Responses to Natural Sounds. PLoS One 6.
53. McFarland JM, Cui YW, Butts DA (2013) Inferring Nonlinear Neuronal Computation Based on Physiologically Plausible Inputs. Plos Computational Biology 9.
54. Sharpee T, Rust NC, Bialek W (2004) Analyzing neural responses to natural signals: maximally informative dimensions. Neural Comput 16: 223-250.
55. Atencio CA, Sharpee TO, Schreiner CE (2012) Receptive field dimensionality increases from the auditory midbrain to cortex. Journal of Neurophysiology 107: 2594-2603.
56. Depireux DA, Elhilali M (2014) Handbook of Modern Techniques in Auditory Cortex: Nova Science Pub Inc.
57. Ahrens MB, Linden JF, Sahani M (2008) Nonlinearities and contextual influences in auditory cortical responses modeled with multilinear spectrotemporal methods. J Neurosci 28: 1929-1942.
58. Woolley SM, Gill PR, Fremouw T, Theunissen FE (2009) Functional groups in the avian auditory system. J Neurosci 29: 2780-2793.
59. Amin N, Gill P, Theunissen FE (2010) Role of the zebra finch auditory thalamus in generating complex representations for natural sounds. J Neurophysiol 104: 784-798.
60. Nagel KI, Doupe AJ (2008) Organizing principles of spectro-temporal encoding in the avian primary auditory area field L. Neuron 58: 938-955.
61. Kim G, Doupe A (2011) Organized Representation of Spectrotemporal Features in Songbird Auditory Forebrain. Journal of Neuroscience 31: 16977-16990.
62. Miller LM, Escabi MA, Read HL, Schreiner CE (2001) Functional convergence of response properties in the auditory thalamocortical system. Neuron 32: 151-160.
63. Miller LM, Escabi MA, Read HL, Schreiner CE (2002) Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. J Neurophysiol 87: 516-527.
64. Depireux DA, Simon JZ, Klein DJ, Shamma SA (2001) Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. J Neurophysiol 85: 1220-1234.
65. Chi T, Ru P, Shamma SA (2005) Multiresolution spectrotemporal analysis of complex sounds. The Journal of the Acoustical Society of America 118: 887.
66. Escabi MA, Schreiner CE (2002) Nonlinear spectrotemporal sound analysis by neurons in the auditory midbrain. J Neurosci 22: 4114-4131.
67. Escabi MA, Miller LM, Read HL, Schreiner CE (2003) Naturalistic auditory contrast improves spectrotemporal coding in the cat inferior colliculus. J Neurosci 23: 11489-11504.
68. Carlson NL, Ming VL, Deweese MR (2012) Sparse codes for speech predict spectrotemporal receptive fields in the inferior colliculus. PLoS Comput Biol 8: e1002594.
69. Fritz J, Shamma S, Elhilali M, Klein D (2003) Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. Nat Neurosci 6: 1216-1223. Epub 2003 Oct 1228.
70. David SV, Fritz JB, Shamma SA (2012) Task reward structure shapes rapid receptive field plasticity in auditory cortex. Proceedings of the National Academy of Sciences of the United States of America 109: 2144-2149.

71. Rabinowitz NC, Willmore BDB, Schnupp JWH, King AJ (2012) Spectrotemporal Contrast Kernels for Neurons in Primary Auditory Cortex. Journal of Neuroscience 32: 11271-11284.
72. Mesgarani N, David SV, Fritz JB, Shamma SA (2009) Influence of Context and Behavior on Stimulus Reconstruction From Neural Activity in Primary Auditory Cortex. Journal of Neurophysiology 102: 3329-3339.
73. Amin N, Gastpar M, Theunissen FE (2013) Selective and efficient neural coding of communication signals depends on early acoustic and social environment. PLoS One 8: e61417.
74. Mesgarani N, Slaney M, Shamma SA (2006) Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations. Ieee Transactions on Audio Speech and Language Processing 14: 920-930.
75. Mesgarani N, David SV, Fritz JB, Shamma SA (2008) Phoneme representation and classification in primary auditory cortex. Journal of the Acoustical Society of America 123: 899-909.
76. McDermott JH, Simoncelli EP (2011) Sound Texture Perception via Statistics of the Auditory Periphery: Evidence from Sound Synthesis. Neuron 71: 926-940.
77. Libersat F, Murray JA, Hoy RR (1994) Frequency as a Releaser in the Courtship Song of 2 Crickets, Gryllus-Bimaculatus (De Geer) and Teleogryllus-Oceanicus - a Neuroethological Analysis. Journal of Comparative Physiology a-Sensory Neural and Behavioral Physiology 174: 485-494.
78. Feng AS, Hall JC, Gooler DM (1990) Neural Basis of Sound Pattern-Recognition in Anurans. Progress in Neurobiology 34: 313-329.
79. Grimsley JM, Shanbhag SJ, Palmer AR, Wallace MN (2012) Processing of communication calls in Guinea pig auditory cortex. PLoS One 7: e51646.
80. McCasland JS, Konishi M (1981) Interactions between auditory and motor activities in an avian song control nucleus. PNAS USA 78: 7815-7819.
81. Doupe AJ, Konishi M (1991) Song-selective auditory circuits in the vocal control system of the zebra finch. Proc Natl Acad Sci U S A 88: 11339-11343.
82. Grace JA, Amin N, Singh NC, Theunissen FE (2003) Selectivity for conspecific song in the zebra finch auditory forebrain. J Neurophysiol 89: 472-487.
83. Newman J, Wollberg Z (1978) Multiple coding of species-specific vocalizations in the auditory cortex of squirrel monkeys. Brain Res 54: 287-304.
84. Huetz C, Gourevitch B, Edeline JM (2011) Neural codes in the thalamocortical auditory system: from artificial stimuli to communication sounds. Hear Res 271: 147-158.
85. Ter-Mikaelian M, Semple MN, Sanes DH (2013) Effects of spectral and temporal disruption on cortical encoding of gerbil vocalizations. J Neurophysiol 110: 1190-1204.
86. Suta D, Popelar J, Syka J (2008) Coding of communication calls in the subcortical and cortical structures of the auditory system. Physiol Res 57 Suppl 3: S149-159.
87. Wallace MN, Grimsley JM, Anderson LA, Palmer AR (2013) Representation of individual elements of a complex call sequence in primary auditory cortex. Front Syst Neurosci 7: 72.
88. Theunissen FE, Doupe AJ (1998) Temporal and spectral sensitivity of complex auditory neurons in the nucleus HVc of male zebra finches. J Neurosci 18: 3786-3802.

89. Lewicki MS, Konishi M (1995) Mechanisms underlying the sensitivity of songbird forebrain neurons to temporal order. Proc Natl Acad Sci U S A 92: 5582-5586.
90. Marler PR (1982) Avian and primate communication: the problem of natural categories. Neurosci Biobehav Rev 6: 87-94.
91. Walker KMM, Bizley JK, King AJ, Schnupp JWH (2011) Multiplexed and Robust Representations of Sound Features in Auditory Cortex. Journal of Neuroscience 31: 14565-14576.
92. Meliza CD, Margoliash D (2012) Emergence of selectivity and tolerance in the avian auditory cortex. J Neurosci 32: 15158-15168.
93. George I, Cousillas H, Richard JP, Hausberger M (2008) A potential neural substrate for processing functional classes of complex acoustic signals. PLoS One 3: e2203.
94. Petkov CI, Kayser C, Steudel T, Whittingstall K, Augath M, et al. (2008) A voice region in the monkey brain. Nature Neuroscience 11: 367-374.
95. Perrodin C, Kayser C, Logothetis NK, Petkov CI (2011) Voice Cells in the Primate Temporal Lobe. Current Biology 21: 1408-1415.
96. Cohen YE, Russ BE, Davis SJ, Baker AE, Ackelson AL, et al. (2009) A functional role for the ventrolateral prefrontal cortex in non-spatial auditory cognition. Proceedings of the National Academy of Sciences of the United States of America 106: 20045-20050.
97. Gifford GW, 3rd, MacLean KA, Hauser MD, Cohen YE (2005) The neurophysiology of functionally meaningful categories: macaque ventrolateral prefrontal cortex plays a critical role in spontaneous categorization of species-specific vocalizations. J Cogn Neurosci 17: 1471-1482.
98. Cohen YE, Theunissen F, Russ BE, Gill P (2007) Acoustic features of rhesus vocalizations and their representation in the ventrolateral prefrontal cortex. J Neurophysiol 97: 1470-1484. Epub 2006 Nov 1429.
99. Rauschecker JP, Scott SK (2009) Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nature Neuroscience 12: 718-724.
100. Marler P (2004) Bird calls: their potential for behavioral neurobiology. Ann N Y Acad Sci 1016: 31-44.
101. Prather JF, Nowicki S, Anderson RC, Peters S, Mooney R (2009) Neural correlates of categorical perception in learned vocal communication. Nature Neuroscience 12: 221-228.
102. Marler P (1981) Birdsong: the acquisition of a learned motor skill. Trends in Neurosci 4: 88-94.
103. Seyfarth RM, Cheney DL (1986) Vocal Development in Vervet Monkeys. Animal Behaviour 34: 1640-1658.
104. Miranda JA, Liu RC (2009) Dissecting natural sensory plasticity: hormones and experience in a maternal context. Hear Res 252: 21-28.
105. Menardy F, Touiki K, Dutrieux G, Bozon B, Vignal C, et al. (2012) Social experience affects neuronal responses to male calls in adult female zebra finches. Eur J Neurosci 35: 1322-1336.
106. Gentner TQ, Margoliash D (2003) Neuronal populations and single cells representing learned auditory objects. Nature 424: 669-674.
107. Woolley SMN, Hauber ME, Theunissen FE (2010) Developmental Experience Alters Information Coding in Auditory Midbrain and Forebrain Neurons. Developmental Neurobiology 70: 235-252.

108. Hauber ME, Woolley SMN, Cassey P, Theunissen FE (2013) Experience dependence of neural responses to different classes of male songs in the primary auditory forebrain of female songbirds. Behavioural Brain Research 243: 184-190.
109. Sanes DH, Bao SW (2009) Tuning up the developing auditory CNS. Current Opinion in Neurobiology 19: 188-199.
110. Doupe AJ (1997) Song- and order-selective neurons in the songbird anterior forebrain and their emergence during vocal development. J Neurosci 17: 1147-1167.
111. Solis MM, Doupe AJ (1999) Contributions of tutor and bird's own song experience to neural selectivity in the songbird anterior forebrain. J Neurosci 19: 4559-4584.
112. Volman SF (1993) Development of neural selectivity for birdsong during vocal learning. J Neurosci 13: 4737-4747.
113. Kover H, Gill K, Tseng YTL, Bao SW (2013) Perceptual and Neuronal Boundary Learned from Higher-Order Stimulus Probabilities. Journal of Neuroscience 33: 3699-+.
114. Moss CF, Surlykke A (2010) Probing the natural scene by echolocation in bats. Frontiers in Behavioral Neuroscience 4.
115. Cheney DL, Seyfarth RM (1988) Assessment of Meaning and the Detection of Unreliable Signals by Vervet Monkeys. Animal Behaviour 36: 477-486.
116. Elie JE, Mariette MM, Soula HA, Griffith SC, Mathevon N, et al. (2010) Vocal communication at the nest between mates in wild zebra finches: a private vocal duet? Animal Behaviour 80: 597-605.
117. Penna M, Llusia D, Marquez R (2012) Propagation of natural toad calls in a Mediterranean terrestrial environment. Journal of the Acoustical Society of America 132: 4025-4031.
118. Aubin T, Jouventin P (2002) How to vocally identify kin in a crowd: The penguin model. Advances in the Study of Behavior, Vol 31. San Diego: Academic Press Inc. pp. 243-277.
119. Vignal C, Mathevon N, Mottin S (2004) Audience drives male songbird response to partner's voice. Nature 430: 448-451.
120. Kawasaki M, Margoliash D, Suga N (1988) Delay-Tuned Combination-Sensitive Neurons in the Auditory-Cortex of the Vocalizing Mustached Bat. Journal of Neurophysiology 59: 623-635.
121. Suga N, Shimozaw.T (1974) Site of Neural Attenuation of Responses to Self-Vocalized Sounds in Echolocating Bats. Science 183: 1211-1213.
122. Eliades SJ, Wang X (2008) Neural substrates of vocalization feedback monitoring in primate auditory cortex. Nature 453: 1102-1106.
123. Keller GB, Hahnloser RH (2009) Neural processing of auditory feedback during vocal practice in a songbird. Nature 457: 187-190.
124. Miller CT, Wang XQ (2006) Sensory-motor interactions modulate a primate vocal behavior: antiphonal calling in common marmosets. Journal of Comparative Physiology a-Neuroethology Sensory Neural and Behavioral Physiology 192: 27-38.
125. Fortune ES, Rodriguez C, Li D, Ball GF, Coleman MJ (2011) Neural Mechanisms for the Coordination of Duet Singing in Wrens. Science 334: 666-670.
126. Lewicki MS, Olshausen BA, Surlykke A, Moss CF (2014) Scene analysis in the natural environment. Front Psychol 5: 1-21.
127. Schneider DM, Woolley SMN (2013) Sparse and Background-Invariant Coding of Vocalizations in Auditory Scenes. Neuron 79: 141-152.

128. Ding N, Simon JZ (2013) Adaptive Temporal Encoding Leads to a Background-Insensitive Cortical Representation of Speech. Journal of Neuroscience 33: 5728-5735.
129. Rabinowitz NC, Willmore BDB, King AJ, Schnupp JWH (2013) Constructing Noise-Invariant Representations of Sound in the Auditory Pathway. Plos Biology 11: 1710-1710.
130. Middlebrooks JC, Bremen P (2013) Spatial Stream Segregation by Auditory Cortical Neurons. Journal of Neuroscience 33: 10986-11001.
131. Fishman YI, Reser DH, Arezzo JC, Steinschneider M (2001) Neural correlates of auditory stream segregation in primary auditory cortex of the awake monkey. Hearing Research 151: 167-187.
132. Bee MA, Klump GM (2004) Primitive auditory stream segregation: A neurophysiological study in the songbird forebrain. Journal of Neurophysiology 92: 1088-1104.
133. Greenlee JDW, Jackson AW, Chen F, Larson CR, Oya H, et al. (2011) Human Auditory Cortical Activation during Self-Vocalization. Plos One 6.
134. Pasley BN, David SV, Mesgarani N, Flinker A, Shamma SA, et al. (2012) Reconstructing Speech from Human Auditory Cortex. Plos Biology 10.
135. Regev M, Honey CJ, Simony E, Hasson U (2013) Selective and Invariant Neural Responses to Spoken and Written Narratives. Journal of Neuroscience 33: 15978-15988.
136. Al-Mana D, Ceranic B, Djahanbakhch O, Luxon LM (2008) Hormones and the auditory system: a review of physiology and pathophysiology. Neuroscience 153: 881-900.
137. Teramoto W, Sakamoto S, Furune F, Gyoba J, Suzuki Y (2012) Compression of Auditory Space during Forward Self-Motion. Plos One 7.
138. Beckers GJL, Gahr M (2012) Large-Scale Synchronized Activity during Vocal Deviance Detection in the Zebra Finch Auditory Forebrain. Journal of Neuroscience 32: 10594-10608.
139. Schneidman E, Puchalla JL, Segev R, Harris RA, Bialek W, et al. (2011) Synergy from Silence in a Combinatorial Neural Code. Journal of Neuroscience 31: 15732-15741.

Author's Biography

Frédéric Theunissen obtained a B.S. in physics and a Ph.D. in biophysics from UC Berkeley and did his post-doctoral work at UC San Francisco in neurosciences in the laboratory of Dr. Doupe where he began to study the auditory system of songbirds. He is fascinated by the brain, sounds and perception and uses a combination of computational, neurophysiological and behavioral approaches in studies with songbirds, humans and hyenas.

Julie Elie is a French neuroethologist who obtained a B.S. in molecular and cell biology from the Ecole Normale Supérieure and a Ph.D. in biological sciences from the University of Saint-Etienne. She began studying vocal and social behavior of songbirds during her Ph.D work and joined the laboratory of Dr. Theunissen to develop the songbird as a neuroethological model system for studying vocal communication. Her research combines neurophysiological, behavioral and field approaches to understand social and communication processes.

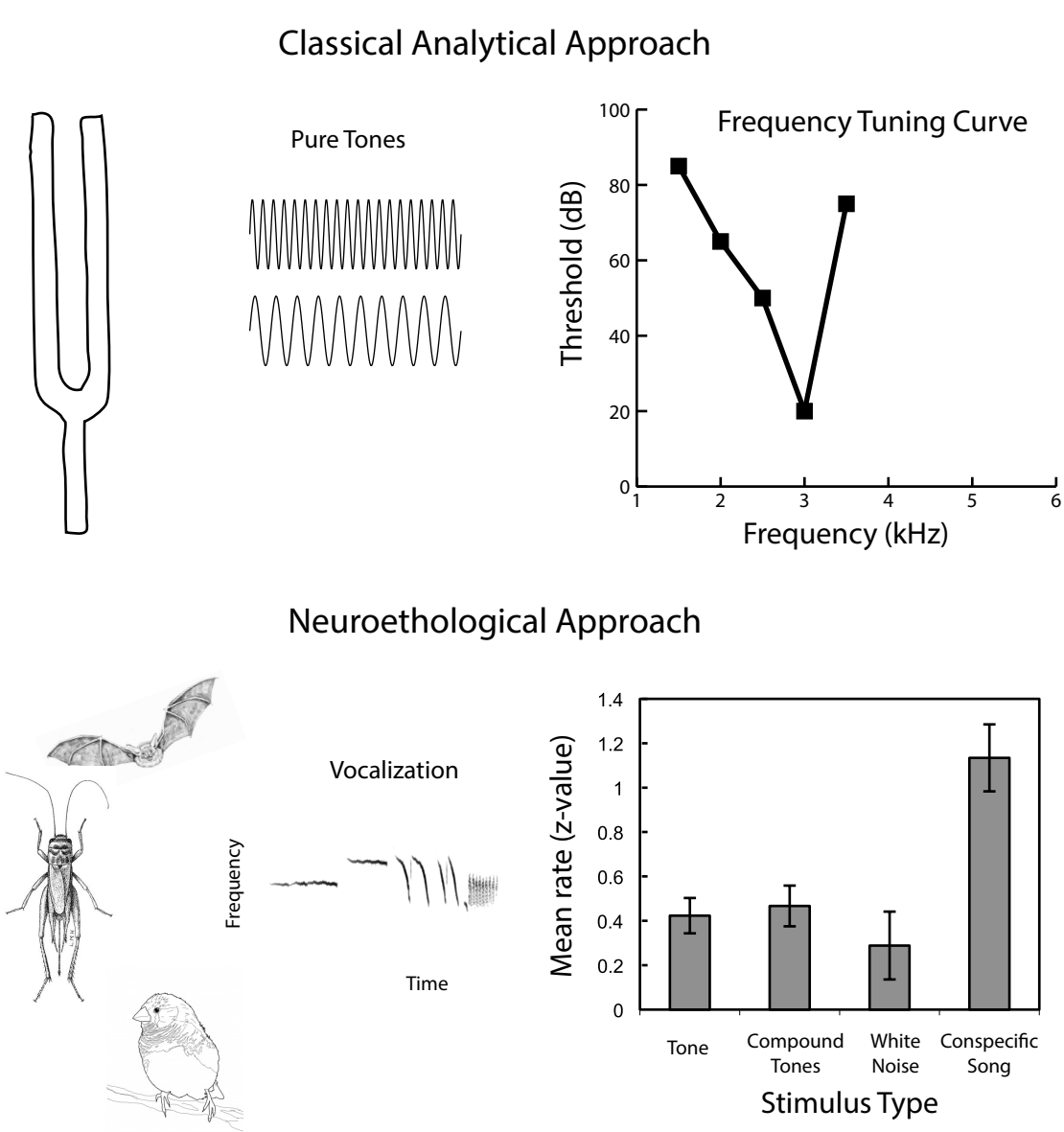

Figure 1. **Historical Approaches to Auditory Neurosciences.** Both the classical analytical approach (top row) and the neuroethological approach (bottom row) are based on analyzing neural responses (right panels) to sounds (middle panels) produced by particular sound sources (left panels). In the classical analytical approach, the sound sources are synthesizers or computers (symbolized by a man-made tuning fork), the sounds are often pure tones (the sine waves shown in the middle) and neural responses are often described as a function of frequency such as in the frequency tuning curve of a neuron (right panel). The frequency tuning curve shows the minimum sound level of pure tones needed to elicit threshold responses. Here we show the tuning curve of a narrow-tuned neuron from the avian inferior colliculus, the MLd (data re-plotted from [3]). This particular neuron is tuned to detect a frequency of 3kHz and the response threshold increases sharply either side of this frequency (hence 'narrow tuned'). The bottom row illustrates the neuroethological approach. Here, the sound sources are often animals vocalizing or generating sounds by other means for communication. These natural sounds are complex signals that are best represented in a time-frequency plot such as the spectrogram of a bird song shown in the middle panel. Responses to these complex sounds are compared to those obtained in response to synthetic sounds, such as pure tones (tones), compound tones (combination of pure tones), white noise or manipulated versions of the species vocalization. The neural data shown in the right bottom graph are from single neurons in the avian primary auditory areas (data re-plotted from [82]). The average spike rates of these neurons, represented here as a z-score (the deviation from the rate obtained in absence of sound stimulus in units of the standard deviation), show that the natural sound, here Conspecific song, is the stimulus type that best excites the neurons.

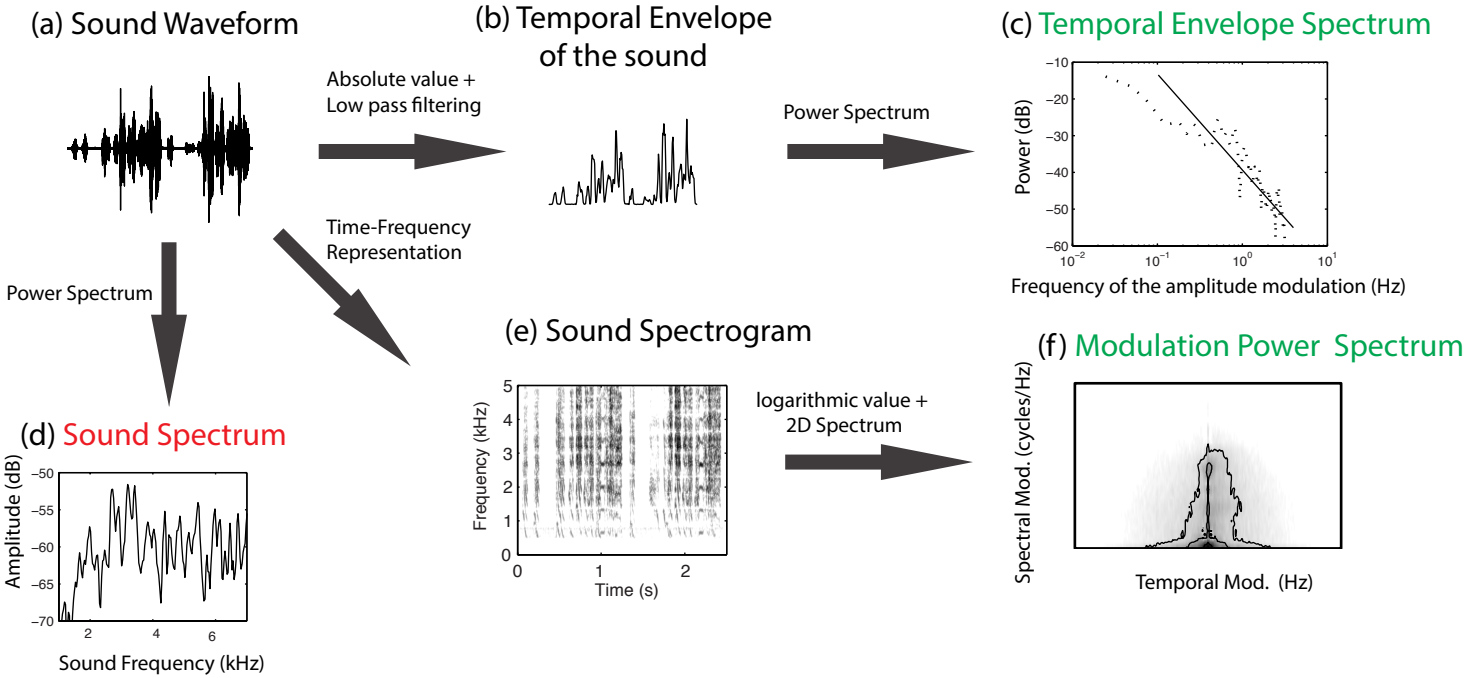

Figure 2. Natural Sound Statistics. Various statistical measurements can be obtained from distinct physical characteristics of sounds. This figure illustrates some of these measurements for zebra finch song and highlights the measures (text in green) that reveal common characteristics of natural sounds and those (text in red) that are specific to each sound class. The sound spectrum (d) is the power of the sound pressure waveform (a) as a function of frequency. This basic spectrum (i.e. obtained without any transformations) shows unique shapes depending on species and call types. Power-frequency curves of natural sounds (sounds spectra) do not obey universal relationships that would be characteristic of all natural sounds. On the other hand (green text), the temporal envelope spectrum (c) obtained by calculating the power spectrum of the temporal envelope (b) obeys a 1/f (f=frequency) relationship (solid line) that is characteristic of all natural sounds [13]: natural sounds are dominated by low frequencies of amplitude modulation. The sound spectrogram (e) is a more intuitive representation obtained by decomposing the sound into time and frequency bins: at each given time point (x-axis) the sound is represented in terms of the amplitude of its frequency components (y-axis). Just as for the basic spectrum, measures on the spectrogram are unique for each natural sound but, the modulation power spectrum (f) obtained from a 2-d spectral analysis of the logarithmic values of the sound spectrogram (e) shows a coarse shape that is characteristic of all animal vocalizations[15].

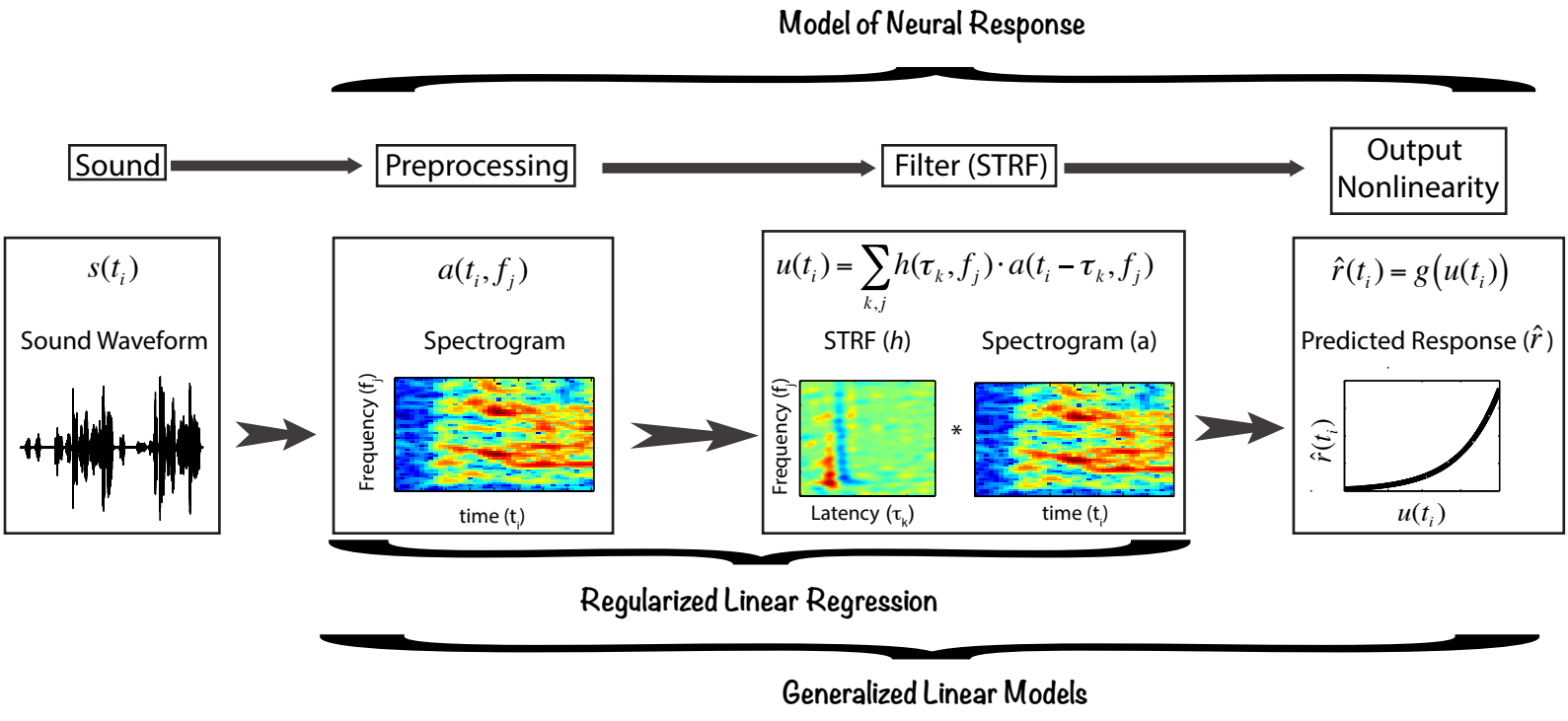

Figure 3. **Stimulus-Response Characterization.** The stimulus-response function of auditory neurons can be estimated using natural sounds and advanced techniques in regression and machine learning. The neural response, r(t) is modeled as a multi-step transformation of the sound stimulus s(t) yielding a predicted neural response . The three steps of this neural model include a pre-processing step, a filtering step and an output non-linearity. The parameters of the three steps are first estimated using a data set of sound stimuli and their corresponding known neural responses, and then the model can be used to predict the neural responses to new sound stimuli ('sound waveform' in the figure). In the pre-processing step, the sound pressure waveform is transformed into a new representation, such as the spectrogram where the amplitude a is expressed as a function of time ti and frequency fj (shown here as an example in the "preprocessing" column; a(ti, fj)), a cochleogram (not shown) which models the filtering and processing occurring at the level of the cochlear nuclei (brainstem nuclei that receive inputs from the cochleae) [49] or higher level processing such as those based on probabilistic expectations (not shown) [51]. The next step involves the estimation of a linear filter (h(τk,, fj) here). Because the new sound representation obtained in the pre-processing step can have many dimensions, regularization regression techniques must be used when estimating the filter to prevent overfitting [43,44,46]. When a time-frequency representation of sound is used, the linear filter h(τk,, fj) obtained is called the spectro-temporal receptive field (STRF, shown here). When the x-axis of the STRF is set up to indicate increasing delay τk from the beginning of a stimulus (shown here), then the STRF represents the neural response obtained to a theoretical impulse stimulus, so called impulse function; when the x-axis is set up to indicate the time preceding a spike, equivalent to a vertical reflection of the previous matrix, then the STRF represents the spectro-temporal features that drive most the neuron (shown in Fig.4). More advanced methods can yield multi-component linear filters (not shown) [55]. In the last step the output of the linear filter u(ti) is transformed into the predicted response using a static non-linearity g(). Generalized linear models can be used to estimate simultaneously the STRF and this non-linear output function for different noise distributions [52].

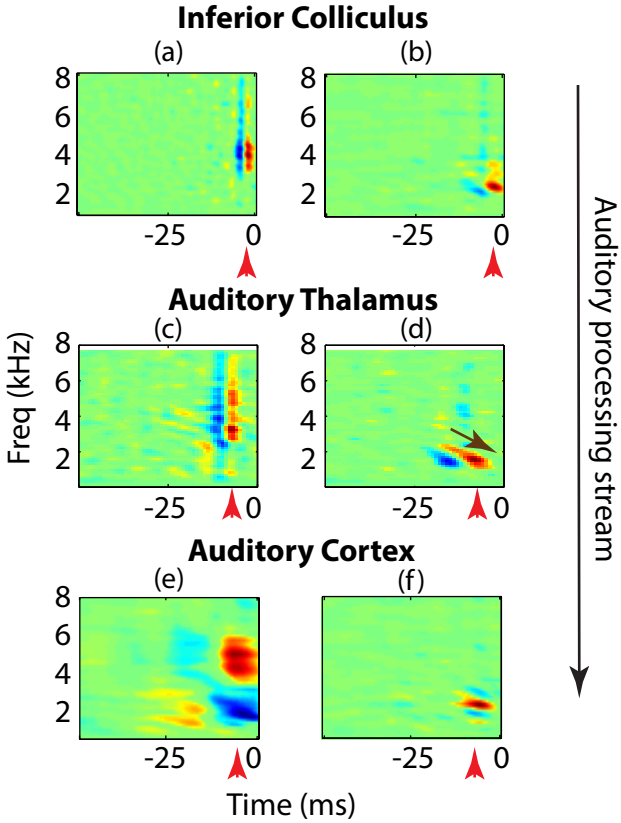

Figure 4. STRFs at Different Levels of the Auditory System. In each region of the auditory system, one finds multiple types of STRFs and in each region some of these types efficiently extract spectro-temporal features of natural sounds. Each row in the figure shows two illustrative STRFs, (shown in pseudocolour in the figure, where red represents the most intense response, and blue the lowest response) found at different levels of the avian auditory system: the inferior colliculus (also known as MLd), the auditory thalamus (also known as nucleus Ovoidalis) and the avian auditory cortex (also known as Field L). Note that for each STRF, the x-axis is the time preceding the response and that therefore the sound features that excite the neuron are read from left to right while the impulse function is read from right to left. As one follows the auditory processing stream, neurons become tuned to slower and more complex features. (a) In the IC some neurons show STRFs with a brief (narrow in time) and large frequency band of inhibition (blue) followed by a brief and large frequency band of excitation (red), such fast broad-band neurons will effectively detect the onset of song syllables and encode the temporal rhythm of song. (b) The narrow-band neuron shown of the right is also selective to the onset of sound but at a particular frequency, around 2.5kHz. The auditory neurons in the thalamus (c-d) exhibit greater latencies than IC neurons: they respond 10-15 milliseconds after the peak energy in the STRF (shown red arrows) while IC neurons responses have latencies around 5-10 ms. Auditory thalamic neurons (c-d) also show greater sensitivity to slower features. The narrow band STRF shown on the right panel (d) is more complex than the one found in IC (b) with frequency tuning that goes down with time (brown arrow). This neuron is sensitive to down-sweeps that are common in zebra finch song syllables. Much slower and more complex STRF appear at the level of the auditory cortex (e-f). The broad-band neuron shown here (e) not only decodes spectral shape at the coarse scale that is useful to represent structures such as formants but is also sensitive to a combination of low frequency sounds (< 3 kHz) followed by high frequency sounds (> 3 kHz). The narrow-band neuron (f) illustrated here shows a sharp excitatory region that is flanked by two inhibitory regions. Such narrow-band neurons are exquisitely tuned to notes of a particular pitch either as pure tones or as harmonic complexes. In the avian auditory system, STRFs that combine excitatory and inhibitory regions at the same time point (as shown in these two examples) appear only at the level of the cortex. Additional avian STRF types and examples can be found in [51,58,59]. Examples in the mammalian auditory system can be found in [24,62-64].