Constructing Invariant Representation of Sound Using Optimal Features and Sound Statistics Adaptation

by

Shi Tong Liu

Bachelor of Science, University of Pittsburgh, 2014

Submitted to the Graduate Faculty of the

Swanson School of Engineering in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2021

University Of Pittsburgh

Swanson School of Engineering

This dissertation was presented

by

Shi Tong Liu

It was defended on

April 13, 2021

and approved by

Neeraj Gandhi, PhD; Professor, Department of Bioengineering

Karl Kandler, PhD, Professor; Department of Neurobiology

Matthew Smith, PhD; Associate Professor, Biomedical Engineering and Neuroscience Institute,

Carnegie Mellon University

Bharath Chandrasekaran, PhD; Professor, Department of Communications

Dissertation Director: Srivatsun Sadagopan, PhD; Assistant Professor, Department of Neurobiology

Constructing Invariant Representation of Sound Using Optimal Features and Sound Statistics Adaptation

Shi Tong Liu, PhD

University of Pittsburgh, 2021

The ability to convey information using sound is critical for the survival of many vocal species, including humans. These communication sounds (vocalizations or calls) are often comprised of complex spectrotemporal features that require accurate detection to prevent mis-categorization. This task is made difficult by two factors: 1) the inherent variability in vocalization production, and 2) competing sounds from the environment. The auditory system must generalize across these variabilities while maintaining sufficient sensitivity to detect subtle differences in fine acoustic structures. While several studies have described vocalization-selective and noise invariant neural responses in the auditory pathway at a phenomenological level, the algorithmic and mechanistic principles behind these observations remain speculative.

In this thesis, we first adopted a theoretical approach to develop biologically plausible computational algorithms to categorize vocalizations while generalizing over sound production and environment variability. From an initial set of randomly chosen vocalization features, we used a greedy search algorithm to select most informative features that maximized vocalization categorization performance and minimized redundancy between features. High classification performance could be achieved using only 10–20 features per vocalization category. The optimal features tended to be of intermediate complexity, offering an optimal compromise between fine and tolerant feature tuning. Predictions of tuning properties of putative feature-selective neurons

matched some observed auditory cortical responses. While this algorithm performed well in quiet listening conditions, it failed in noisy conditions. To address this shortcoming, we implemented biologically plausible algorithms to improve model performance in noisy conditions. We explored two model elements to aid adaption to sound statistics: 1. De-noising of noisy inputs by thresholding based on wide-band energy, and 2. Adjusting feature detection parameters to offset noise-masking effects. These processes were consistent with physiological observations of gain control mechanisms and principles of efficient encoding in the brain. With these additions, our model was able to achieve near-physiological levels of performance. Our results suggest that invariant representation of sound can be achieved based on task-dependent features with adaptation to input sound statistics.

Table of Contents

List of Tables

List of Figures

1.0 Introduction

Many species, including humans, rely on vocalizations to convey information crucial for the survival of the species, such as food availability, predator warnings, mating, etc. As a result, the processing of these conspecific communication sounds is an essential task for the auditory system. Imaging studies in humans found that speech is encoded in a hierarchical fashion, from spectral features such as phonemes in primary auditory areas to more abstract elements such as semantics in the higher areas (Chang et al 2010, Heer et al 2017, Belin et al 2011). A similar preference for conspecific communication sounds has been shown in other vocal species such as macaques (Petkov et al 2008; Perrodin et al 2011) and common marmosets (*Callithrix jacchus*) (Sadagopan et al, 2015). What is less certain however, are the encoding strategies for classifying communication sounds into conceptual categories. There are considerable variations in acoustic parameters such as pitch, speed, etc. between different speakers (for example, Agamaite and Wang 2015). In addition, real-world acoustic environments often contain other extraneous signals that can disrupt the recognition process. This requires the auditory system to encode sound in an invariant manner to generalize across these variations, while paradoxically remain sensitive to the fine differences in the spectrotemporal structure between distinct sound categories. Our goal in this thesis to elucidate how the auditory system can overcome both production and environmental variations to classify sounds, particularly vocalizations, into discrete categories. To accomplish this goal, we will build a computational model of sound categorization to test methods of achieving production and environmental invariant encoding. Testable predictions made using this model may be used to guide future behavioral and electrophysiological studies.

1.1 Encoding and Processing of Communication Sounds in the Auditory Pathway

Understanding how the auditory system encodes and processes vocal sounds such as human speech or animal vocalizations is a long-standing topic in auditory research. Early anatomical studies have identified regions in the temporal and parietal lobe such as superior temporal gyrus (STG) and Wernicke's area that are involved in speech comprehension. Further studies continued to map out the organization in the auditory cortical network by identifying the characteristics and functionality of core, belt and parabelt areas within the temporal cortex (Romanski and Averbeck 2009). A major obstacle in these types of studies, aside from the need for human subjects, is the complex spectrotemporal structure of speech signals. It is more difficult to standardize and parametrically vary speech signals than other simple stimuli such as a sinusoidal tone due to the intricate variations in both frequency and time domain. To simplify the problem, it can be useful to individually examine the spectral and temporal content of speech signals and how each is represented in the brain. The spectral content of speech is largely based around formant frequencies, which are the resonant frequencies of the vocal tract and the energy peaks in the speech signal (Fant 1970). The first and second formant frequencies are often used to distinguish between vowel sounds (Petersen and Barney 1952, Hillenbrand et al 1994). They also play an important, but a lesser role in the identification of consonants (Hillenbrand et al 2001). Tone and pitch information can also be extracted from the spectral domain and be used for speaker identification. Speech and vocalizations are also heavily modulated in amplitude, providing ample temporal cues for recognition. Early psychoacoustical studies have identified that temporal features are the dominant source of information in speech recognition (Van Tasell et al 1987, Rosen 1992). In particular, Shannon et al 1995 showed that even in conditions of greatly reduced spectral information, near-perfect recognition performance can be achieved if temporal features

are preserved. Similar conclusions can be drawn from studies on the recognition performance of cochlear implant users. Hochmair-Desoyer et al 1980 reported that single-channel cochlear implant users can reasonably comprehend unknown sentences based on speech amplitude waveform alone. Given the importance of temporal information in speech recognition, the remainder of the introduction section will place a larger emphasis on regions in the auditory pathway that exhibit selectivity for changes in temporal modulation. Most of the conclusions are drawn from studies using animal subjects, with corroboration from human studies.

1.1.1 Encoding in Auditory Periphery

Neural representation of sound begins at the inner hair cells in the cochlea, where the acoustic signal is transformed from mechanical perturbations into action potentials encoding the relevant information. This information is then carried by the auditory nerve (AN) and act as input to the rest of the auditory pathway. Acoustic information is tonotopically organized, with each AN fiber showing responses to stimuli within a half-octave of its best frequency (BF), meaning it can be viewed as a bank of bandpass filters (Delgutte 1980, Young 2007). At the stage of AN, there is evidence showing the presence of adaption to sound statistics and various nonlinearities such as phase-locking and two-tone suppression (Zhang et al 2001, Zilany et al 2009). Studies on AN response to sinusoidal amplitude modulated (SAM) tones shown that AN phase-lock to the modulation frequency of the tones but are generally poorly tuned (based on spike rate) to variations in those frequencies (Krishna and Semple 2000). A similar phenomenon is observed in the cochlear nucleus, mainly the primary-like and chopper neurons, both of which receive inputs from AN but

show little change in sensitivity to temporal modulations (Young 2007). Overall, there is little evidence suggesting speech processing in the auditory periphery.

1.1.2 Encoding in Central Auditory

In contrast, neurons in the inferior colliculus (IC) and cortical areas such as the primary auditory cortex (A1) exhibit higher selectivity to variations in modulation frequency of SAM tones as measured by their spike rate (Krishna and Semple 2000). This is an indication that the representation of sound transformed from a temporal coding scheme in the auditory periphery regions towards a rate coding scheme in the higher areas. Evidence suggests that this transformation of the encoding scheme is completed at the level of the IC (Krishna and Semple 2000, Langner and Schreiner 1988). Taken together, this indicates that speech & vocalization recognition can start as early as IC. Indeed, there are studies that point to the existence of vocalization responsive neurons in the IC in various vocal animals (Portfors et al 2009; Suta et al 2003; Holmstrom et al 2010; Sadagopan et al 2015). However, the majority of IC neurons respond to multiple categorizes of vocalizations, whereas neurons in cortical areas, such as A1 exhibit higher selectivity for individual vocalization categories (Carruthers et al 2015; Aizenberg and Geffen 2013; Fritz et al 2010; Wang et al. 1995). Neurons in secondary and higher cortical areas show further sharpening of selectivity to call categories (Tian et al 2001; Perrodin et al 2011, Fukushima et al 2014, 2015) Cortical neurons are also more stringent in their preference. For instance, Wang and Kadia 2001 showed that marmoset cortical neurons strongly differentiate between natural and time-reversed vocalizations. This emergence of vocalization selectivity starting in the IC can be thought of as a build-up of complexity in the acoustic features the

vocalization selective neurons are encoding for. For instance, a ubiquitous feature of vocalizations and speech are frequency-modulated FM sweeps (Ryan 1983, Wang et al 1995). While neurons in both the auditory periphery and higher areas show selectivity to FM sweeps, the majority of neurons in IC and higher areas have a directional preference and will fire preferentially or exclusively to FM sweeps either upward or downward direction only (Poon and Chiu 1991, Sadagopan and Wang 2008, Zhang et al 2003). No such directional preferences were observed in AN or the cochlear nucleus.

A gradual build-up of complex spectrotemporal features has also been reported in the human auditory system. A 2014 study in human speech recognition by Mesgarani et al examined electrophysiological recordings of neurons in the superior temporal gyrus (STG) and found neurons encoding for distinct phonetic features (Mesgarani et al 2014). Some of these neurons demonstrated non-linear integration of multiple, lower-level spectrotemporal cues to establish an acoustic-phonetic representation of speech in the STG. Overall, there is a distinct transition in encoding strategy from continuous spectrotemporal signal to discrete conceptual categories as the signal progress through the auditory pathway.

1.2 Effect of Production Variability on The Invariant Representation of Sound

A major challenge in the transition from continuous spectrotemporal variations to discrete sound categories is to account for the sound production variability. Given the bio-mechanical nature of sound production, this variance is nearly inevitable. In human speech, physical differences in vocal cord and tract between individuals can produce significant variations in

formant frequencies of spoken vowels (Wakita 1977, Hillenbrand 1994, Asakawa 2007). These variations can often lead to overlapping formant frequency characteristics between different vowel sounds (Hillenbrand 2001). Without contextual information, categorizing vowels from different speakers using spectral analysis prove to be challenging. Perceptually, however, these variations are well within human speech recognition capabilities (Liberman et al 1967, Hillenbrand 2001). Similar capabilities are observed in other vocal species. For instance, marmosets show variability in phrase frequency between individuals of the same and different gender has also been observed (Agamaite et al 2015).

This invariant representation of sound arises gradually in the auditory pathway, where higher regions in the pathway exhibit more tolerance to acoustic transformations that preserve the identity of the call (Bidelman et al 2013). For instance, a large population of marmoset A1 neurons exhibits invariance to sound levels (Sadagopan and Wang 2008). That is not to say that these variations are "lost" during neural processing, but rather the auditory system is able to differentiate certain acoustic variations as unessential in terms of categorizing the sound. This disparity between the invariant perception of sound and the representation of acoustic variability is the central conflict in the categorization of communication sounds. On one hand, perceptual categorization can be insensitive to even large variations in sound parameters such as sound levels, pitch shifts as a result of sex and age differences, temporal dilation (slow speech), etc. On the other hand, the distinction between similar sound categories is often achieved by tuning into the fine difference within often complex acoustic structures of the signal. This leads to conflicting requirements for simultaneous fine and broad selectivity in sound categorization. How the brain chooses acoustic features that allow for this paradoxical encoding of sound remains largely unclear.

1.3 Effect of Noise on the Invariant Representation of Sound

Aside from the inherent variations during sound production, the acoustic environment has its own set of challenges for the auditory system to overcome. Real-world listening conditions frequently have competing sound sources that can disrupt the recognition process. These sources can arise from multiple talkers, such as in the case of the famous "cocktail party problem", or other elements in the environment such as the rustling of leaves and the humming of fans. A normal functioning auditory system shows robustness to the effects of noise. Studies suggest that noise invariance is an emerging property of the auditory system and that higher areas in the auditory pathway generally show more indifference to the effects of noise (Mesgarani and Chang 2012, Ding and Simon 2013). For the human auditory system, there is evidence showing a decrease in sensitivity to environmental variations in a speech in higher cortical areas (Okada et al 2010, Chang et al 2010, Kell et al 2018, Norman-Haignere & McDermott 2018). Similarly, in other vocal animals, there is evidence supporting this emerging context invariant representation of sound. Rabinowitz et al 2013 examined single unit response to noisy stimuli, both synthetic and natural, in ferret AN, IC, and A1. The results showed that A1 responses changed the least due to noise and AN the most.

To achieve said invariance, the auditory system must first distinguish between signal and noise from an incoming sound signal. In the absence of context or prior knowledge, deviations from the average sound statistics are typically considered to be important signals deserving of attention. However, the auditory system must be able to detect and represent changes in sounds with a wide range of statistical properties to adequately function in a highly variable acoustic environment. For instance, a pin drop in a silent room and sirens in traffic noise are indicative of

important events, but the changes in both the absolute and relative sound levels between the signal and background (contrast) are vastly different. Moreover, the reported dynamic range of auditory neurons [Dean et al 2005; Wen et al 2009] appears ill-equip to adequately encode for such variations in acoustic environment. This "dynamic range problem" highlights a fundamental challenge for the auditory system of efficiently encoding for highly variable parameters [Colburn et al 2003; Wen et al 2009].

Similar to sound production variability, the "dynamic range problem" is not a uniquely auditory phenomenon. The visual system also must represent stimulus with contrast variations that far exceed the limited dynamic range of its cortical neurons due to their response saturation at higher contrast levels [Heeger 1990, 1991]. Heeger 1992 described a solution to this problem using a divisive normalization model observed in cat striated cortex (primary visual cortex, V1). The model states that the response of individual neurons was normalized by the sum of near-by population response. The result is a response that is dynamically adjusted based on stimulus contrast. This effectively changes the dynamic range of the neuron such that it remains sensitivity to relative changes in the stimulus. The consequence of the divisive normalization model is maintaining the same neural response to stimulus in various contrast conditions, allowing further processing to occur without the brain having the attend to variations in the stimulus contrast level, thus achieving contrast invariant visual representation.

The underlying principles of the divisive normalization model prompted the question if analogous mechanisms exist in the auditory system. Rabinowitz et al 2011 observed stimulus contrast-dependent response change in ferret A1 neurons. They quantified this gain change (ratio of output spikes to input current) as adjustments in output nonlinearities that modeled the measured

neuronal responses. These gain changes were well described by the divisive normalization model and served to compensate for the limited dynamic range in cortical neurons.

Adaptation to sound statistics is not restricted to cortical neurons. Dean et al 2005 identified adjustments of neural responses to mean sound level in inferior colliculus (IC). The responses are not completely invariant to sound level, however, suggesting that at least in the level of IC, there is a still need to encode some information about the overall sound level. Similar adaptation to mean sound level was also found at the level of the auditory nerve, albeit at a weaker level compared to IC (Wen et al 2009). These results show that sound statistic adaptation is widespread in the auditory pathway and there is a gradual strengthening of invariant representation towards the higher processing areas.

Besides vision and audition, normalization or normalization-like mechanisms are also involved in population coding of the olfactory system [Olsen et al 2010] and associative memory in the hippocampus [McNaughton & Morris, 1987]. Indeed, normalization appears to be a canonical computation in the brain [Carandini & Heeger, 2012]. In addition, computational models of sensory neurons with normalization (via gain control) accounted for many observed nonlinear behaviors in a "typical" sensory cell [Schwartz & Simoncelli, 2001]. Adaptation to stimulus statistics also complies with the efficient encoding principle of the brain by eliminating the need to maintain copies of neurons for every possible stimulus environment. Taken together, this builds a strong case for normalization, via gain control, playing an important role in building invariant representation in sensory modalities, such as noise invariance in the auditory pathway.

1.4 Computational Modeling Approaches to Understanding Feature Selection for Invariant Representation

Computational principles and information theory play an important role in neuroscience research by providing educated guesses and directions for physiological experiments. Algorithms and methods that prove useful in computational simulations may offer insight into plausible neural mechanisms. In recent times, automatic speech recognition (ASR) algorithms have seen rapid advancements thanks to the tremendous demand for voice-guided technology in our daily lives. As such, it is worth examining if any of the computational principles in ASR may be applicable to a biological system.

Early ASR models were based on signal sequence matching using dynamic time wrapping algorithms (Sakoe and Chiba 1971, Rabiner et al 1978) or hidden Markov models (Levinson 1986, Rabiner 1989). In more recent times, neural networks have risen into prominence in speech recognition thanks to their high performance compared to previous algorithms and advancements in computational power. The principle of neural networks is loosely based on the neural circuit, where each neuron is considered a computational node, and the output of each node is computed as a non-linearly weighted sum of its inputs. Synaptic connections between neurons are simulated as connections between the input/output of nodes. These nodes may also aggregate into layers based on their computational goals. Neural networks excel at classification tasks such as speech recognition and facial identification because they impose a minimum restriction on model inputs, thus better generalize across production and environmental variability. These principles for feature selection can be utilized in studying the sensory system.

Aside from speech recognition algorithms, the computational methods from the visual system also offered insight on addressing production variability. Ullman et al 2002 found that an information-maximization approach to feature selection resulted in autonomously selected facial features that are robust to within-class variations. Facial features were randomly generated with minimal restrains and a set of the most informative features (MIFs) were selected to maximize the mutual information for classification. The MIFs were of intermediate complexity and coincide with results from IC studies, providing a principled explanation for the purpose of these IC features (Ullman et al 2002). In this thesis, we will develop a computational model to test if this approach to feature selection can account for production variability in sound categorization.

A well-known challenge for ASR algorithms, and a potential shortcoming of this model, is their susceptibility to the disruptive effect of noise. While a significant number of noise removal algorithms have been purposed, the success is relatively limited, especially in comparison to the physiological performance level of normal-hearing individuals. Most noise-robust speech recognition algorithms require specified training in noisy conditions to improve their performance, which is inconsistent with the principle of efficient encoding in the brain since this will significantly increase the number of acoustic features the auditory system needs to encode. A more physiologically plausible method is compensation based on prior knowledge about the acoustic environment, which is similar to the effect of contrast gain control in the auditory pathway. The algorithm first estimates the overall noise activity and then maps the noisy inputs to clean speech templates based on learned parameters. This approach requires significantly fewer speech templates while having similar robustness in performance compared to algorithms that are specifically trained in noisy environments, which fits with the principle of efficient encoding in the brain.

1.5 Research Goals

Our primary goal in this thesis is to show physiologically feasible methods of addressing production and environmental variability in sound categorization. To do so, we will first construct a computational model that is capable of accurately classifying different communication sound categories by template matching acoustic inputs with categorical features. The features will be selected such that they maximally differentiate between various sound categories. Based on prior success in the visual system, we hypothesize that this model to be insensitive to production variability and will verify through testing with call samples obtained from various vocal animals (marmosets, macaques, and guinea pigs). We will augment this computational model with a noise compensation algorithm simulating the effects of contrast gain control in the auditory pathway and examine its effectiveness in making the model robust to noise. Finally, based on model results and experimentally obtained physiological and behavioral data, we will explore possible analogous neural mechanisms for achieving production and noise invariant sound categorization.

2.0 Optimal Features for Auditory Categorization

Humans and vocal animals use vocalizations to communicate with members of their species. A necessary function of auditory perception is to generalize across the high variability inherent in vocalization production and classify them into behaviorally distinct categories (‘words’ or ‘call types’). Here, we demonstrate that detecting mid-level features in calls achieves production-invariant classification. Starting from randomly chosen marmoset call features, we use a greedy search algorithm to determine the most informative and least redundant features necessary for call classification. High classification performance is achieved using only 10–20 features per call type. Predictions of tuning properties of putative feature-selective neurons accurately match some observed auditory cortical responses. This feature-based approach also succeeds for call categorization in other species, and for other complex classification tasks such as caller identification. Our results suggest that high-level neural representations of sounds are based on task-dependent features optimized for specific computational goals.

2.1 Introduction

Human speech recognition is a highly robust behavior, showing tolerance to variations in prosody, stress, accents, and pitch. For example, speech features such as formant frequencies exhibit large variations within- and betweenspeakers (Peterson and Barney 1952; Hillenbrand et al 1995), arising from production mechanisms (production variability). To achieve accurate speech recognition, the auditory system must generalize across these variations. This challenge is not uniquely human. Animals produce species-specific vocalizations (calls) with large within- and between-caller variability (Wang 2000) and must classify these calls into distinct categories to produce appropriate behaviors. For example, in common marmosets (Callithrix jacchus), a highly vocal New World primate species, critical behaviors such as finding other marmosets when isolated depend on accurate extraction of call-type and caller information (Epple 1968; Chen et al 2009; Miller et al 2010; Kato et al 2014). Similar to human speech, marmoset call categories overlap in their long-term spectra (Figure 1a), precluding the possibility that calls can be classified based on spectral content alone, and requiring selectivity for fine spectrotemporal features to classify calls. At the same time, marmoset calls also show considerable production variability along a variety of acoustic parameters (Agamaite et al 2015). For example, twitter calls produced by different marmosets vary in such parameters as dominant frequencies, lengths, inter-phrase intervals, and harmonic ratios (Figure 1). Tolerance to large variations in spectrotemporal features within each call type is thus necessary to generalize across this variability. Therefore, there is a simultaneous requirement for fine and broad selectivity for production-invariant call classification. The present study explores how the auditory system resolves these conflicting requirements.

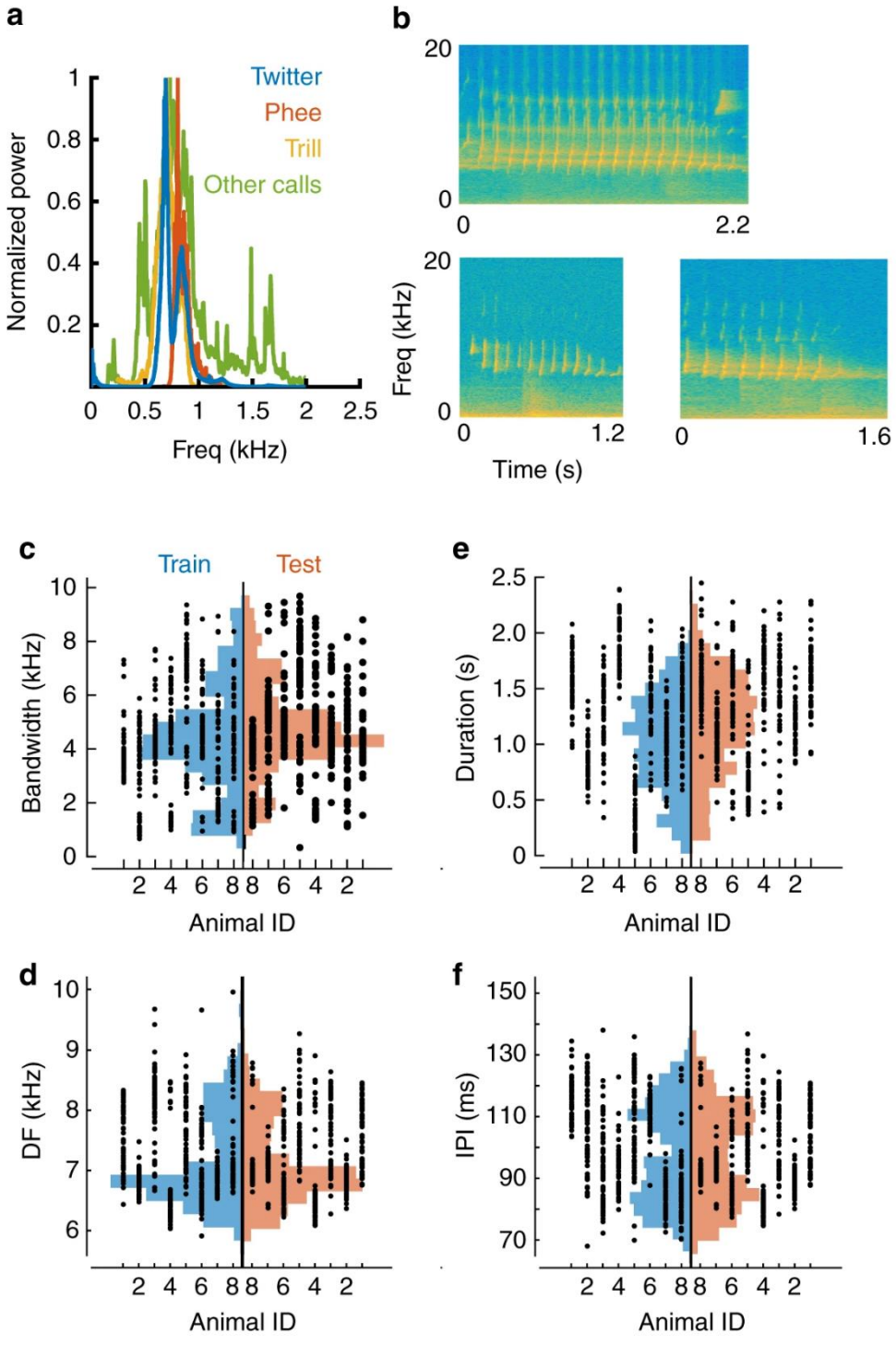

a
Twitter
Phee
Trill
Other calls
Normalized power
Freq (kHz)
b
Freq (kHz)
Time (s)
c
Train
Test
Bandwidth (kHz)
Animal ID
e
Duration (s)
Animal ID
d
DF (kHz)
Animal ID
f
IPI (ms)
Animal ID

Figure 1 Production variability in marmoset calls. **a** The overall spectra of 3 major marmoset call types and other minor call types (grouped as 'Other calls'), showing spectral overlap between call categories. **b** Spectrograms of three twitter calls showing examples of production variability between individuals. **c–f** Production variability of twitter calls quantified along multiple parameters: **c** bandwidth, **d** dominant frequency, **e** duration, and **f** inter-phrase interval. Dots are parameter values of a single call produced by an individual marmoset. Histograms are overall parameter distributions, split into the training (blue) and testing (red) sets. These data show the large production variability captured by the training and test datasets, over which the model must generalize. No systematic bias is evident in calls used for model training and testing

This problem of requiring fine- and tolerant feature tuning, necessitated by high variability amongst members belonging to a category, is not unique to the auditory domain. For example, in visual perception, object categories such as faces also possess a high degree of intrinsic variability (Tsao and Livingston 2008; Jenkins et al 2011; Kramer et al 2018; Ullman et al 2002). To classify faces from other objects, using an exemplar face as a template typically fails because this does not generalize across within-class variability (Ullman et al 2002). Face detection algorithms use combinations of mid-level features, such as regions with specific contrast relationships (Viola and Jones 2004; Sinha 2002), or combinations of face parts (Ullman et al 2002), to accomplish classification. Of these algorithms, the one proposed by Ullman et al. (Ullman et al 2002) is especially interesting because of its potential to generalize to other classification tasks across sensory modalities. In this algorithm, starting from a set of random fragments of faces, the authors used greedy search to extract the most informative fragments that were highly conserved across all faces despite within-class variability. Post hoc analyses revealed that these fragments were mid-

level, i.e., they typically contained combinations of face parts, such as eyes and a nose. The features identified using this algorithm were consistent with some physiological observations, for example at the level of BOLD responses (Lerner et al 2008). While the differences between visual and auditory processing are vast, these results inspired us to ask whether a similar concept – sound categorization using combinations of acoustic features – could be implemented by the auditory system.

The behavioral salience of calls for marmosets (Epple 1968; Chen et al 2009; Miller et al 2010; Kato et al 2014).), and the increasing resources allocated to the processing of calls along the cortical processing hierarchy (Sadagopan et al 2015), suggest that call processing is a computational goal of auditory cortex. Call processing involves detecting the presence of calls in the acoustic input, classifying them into behaviorally relevant categories, extracting information about caller identity, determining the behavioral state of the caller, and developing situational awareness of the environment. Although a number of studies have described call-selective responses at various stages of the auditory pathway, there has been little investigation into how the auditory system goes about solving these problems, both at the algorithmic and mechanistic levels. In this study, we start with the premise that the detection and classification of calls into discrete call types is a critical first step that enables the above computations. Our overall question in this study is to ask how production-invariant call classification can be accomplished in the auditory pathway. Specifically, we test the hypothesis that production-invariant call classification can be accomplished by detecting constituent features that maximally distinguish between call types. Starting from an initial set of randomly selected marmoset call features, we use a greedy search algorithm to determine the most informative and least redundant set of features necessary for call classification. We show that high classification performance can indeed be achieved by detecting

combinations of a small number of mid-level features. We then demonstrate that predictions of tuning properties of putative feature-selective neurons match previous data from marmoset primary auditory cortex. Finally, we show that the same algorithm is equally successful in caller identification with marmoset calls, and in call classification in other species such as guinea pigs (Cavia porcellus) and macaque monkeys (Macaca mulatta). Taken together, our findings suggest that classification of sound categories using mid-level features may be a general auditory computation.

2.2 Results

2.2.1 Intermediate Features are More Informative for Classification

We start with the premise that the first step in call processing is the categorization of calls into discrete call types, generalizing across the production variability that is inherent to calls. Let us consider the example of classifying twitter calls from all other call types. Marmoset twitters can be characterized along several acoustic parameters, such as bandwidth, duration, dominant frequency, and inter-phrase interval (Agamaite et al 2015). In Figure 1c–f, we plot the values of these parameters for individual calls emitted by 8 animals, showing the extent of within- and between-individual variability over which generalization is required for twitter categorization. Similar generalization is required for categorizing the other call types as well (Appendix Figure 1). We first generated 6000 random initial features from the cochleagrams of 500 twitter calls emitted by 8 marmosets (‘training’ set, blue histograms in Figure 1). For the purposes of this study,

a feature is a randomly selected rectangular segment of the cochleagram, corresponding to the spatiotemporal activity pattern of a subset of auditory nerve fibers within a specified time window. For each random feature, we determined an optimal threshold at which its utility for classifying twitters from other calls was maximized. The merit of each feature was taken to be the mutual information value (in bits) at this optimal threshold (Figure 2; Equation 1).

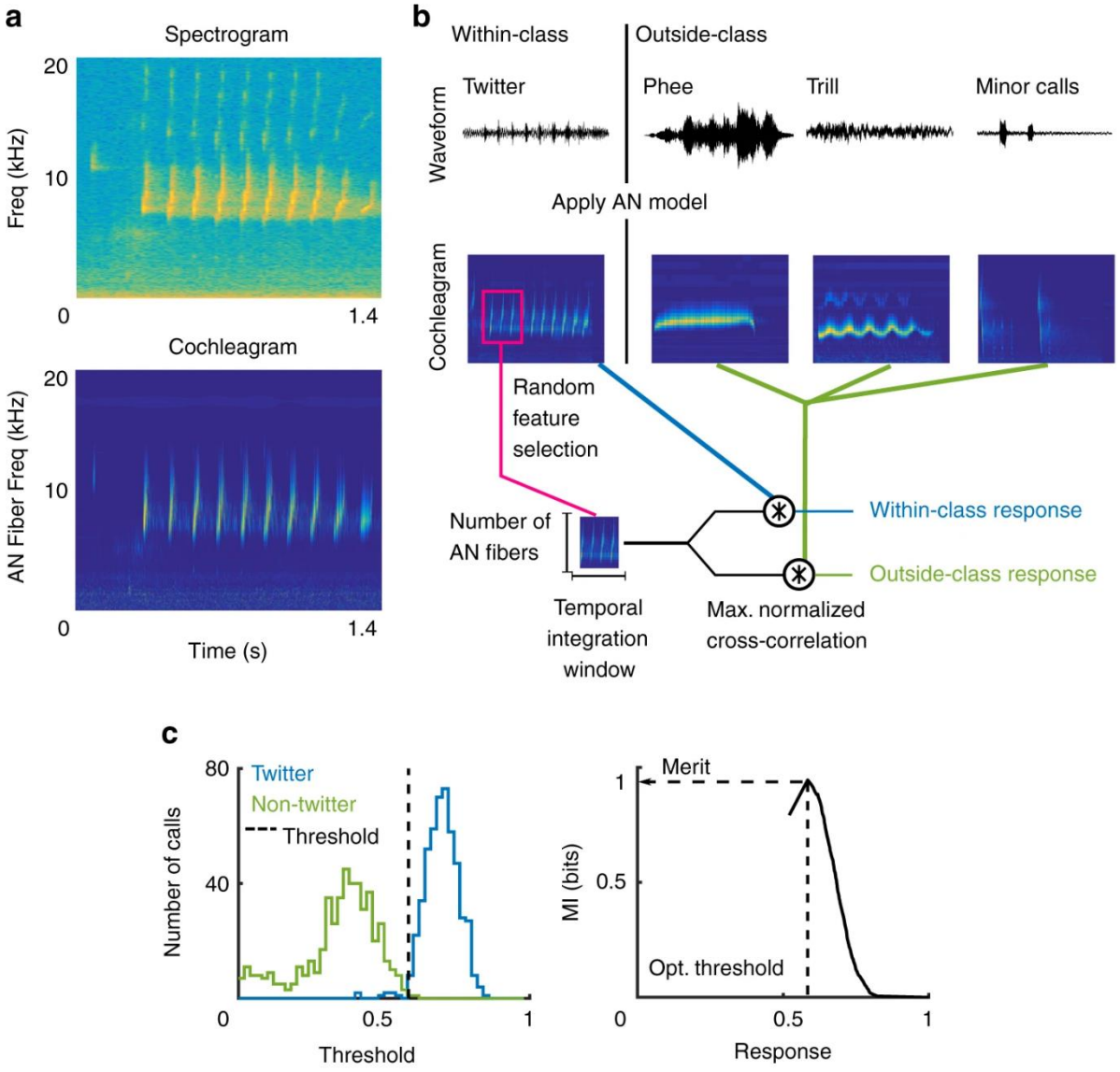


Figure 2 Initial feature generation and evaluation. **a** The spectrogram of a twitter call (top), and its corresponding cochleagram (bottom). Cochleagram color scale denotes firing rates of auditory nerve fibers. **b** Schematic for initial random feature generation for a twitter (within-class) versus other calls (outsideclass) categorization task. Waveforms (top) were converted to cochleagrams (middle). The magenta box outlines a random initial feature picked from the twitter cochleagram shown. The maximum value of the normalized cross-correlation function between each call (within-class—blue, outside-class—green) and each random feature was taken to be the response of a feature to a call. **c** Distributions (top) of a feature's responses to 500 within-class (blue) and 500 outside-class (green) calls. The mutual information (bottom) of a feature computed as a function of a parametrically varied threshold. The dotted line, corresponding to maximal mutual information, is taken to be each feature's optimal threshold

In Appendix Figure 2, we plot the merits of all 6000 initial features as a function of each feature's bandwidth and temporal integration window. Along the margins, we plot the maximum merit of features within each bandwidth- or temporal window bin. These distributions compare the best features from each time bin and show that features of intermediate lengths relative to the total call length show higher merits for call classification. This is an expected consequence of two characteristics of calls: (1) call types overlap in spectral content, so that brief features do not contain sufficient information to separate out categories, and (2) calls have high production variability, so that long features are less likely to be found across all calls belonging to the same category. We observed similar distributions for the classification of other marmoset call types, i.e., for trill vs. other calls, and phee vs. other calls (Appendix Figure 2). We then characterized feature complexity using a kurtosis-based metric (Methods). While features of low merit showed low

complexity values and whole calls showed high complexity values, features of high-merit showed intermediate complexity values. This observation supported the hypothesis that mid-level features of intermediate complexity were most informative for classification (Appendix Figure 2).

2.2.2 Most Informative Features for Classification

Because we generated the initial features at random, many of these have low merit, and many are similar. Therefore, the set of optimal features for classification is expected to be much smaller than this initial set. To determine the set of optimal features that together maximize classification performance, we used a greedy-search algorithm (see Methods). Briefly, we started with the feature of highest merit, and successively added features that maximized pairwise mutual information with respect to the already chosen feature set. We refer to the set of these optimal features as most informative features (MIFs) following the nomenclature of Ullman et al. (Ullman et al 2002, 2004). We determined that call classification could be accomplished using 11 MIFs for twitter vs. all other calls, 20 MIFs for trill vs. all other calls, and 16 for phee vs. all other calls. In Figure 3, magenta boxes outline the top 5 MIFs that are optimal for each of these classification tasks (the first five MIFs in Table 1). The optimal features that we arrive at are mostly intuitive – for example, the top MIFs for classifying twitters detect the frequency contour of individual twitter phrases and the repetitive nature of the twitter call. In some cases, features seemed counter-intuitive— for example, the second MIF for trill classification seems to detect empty regions of the cochleagram. In this theoretical framework, the lack of energy at those frequencies is also informative about the presence of a trill.

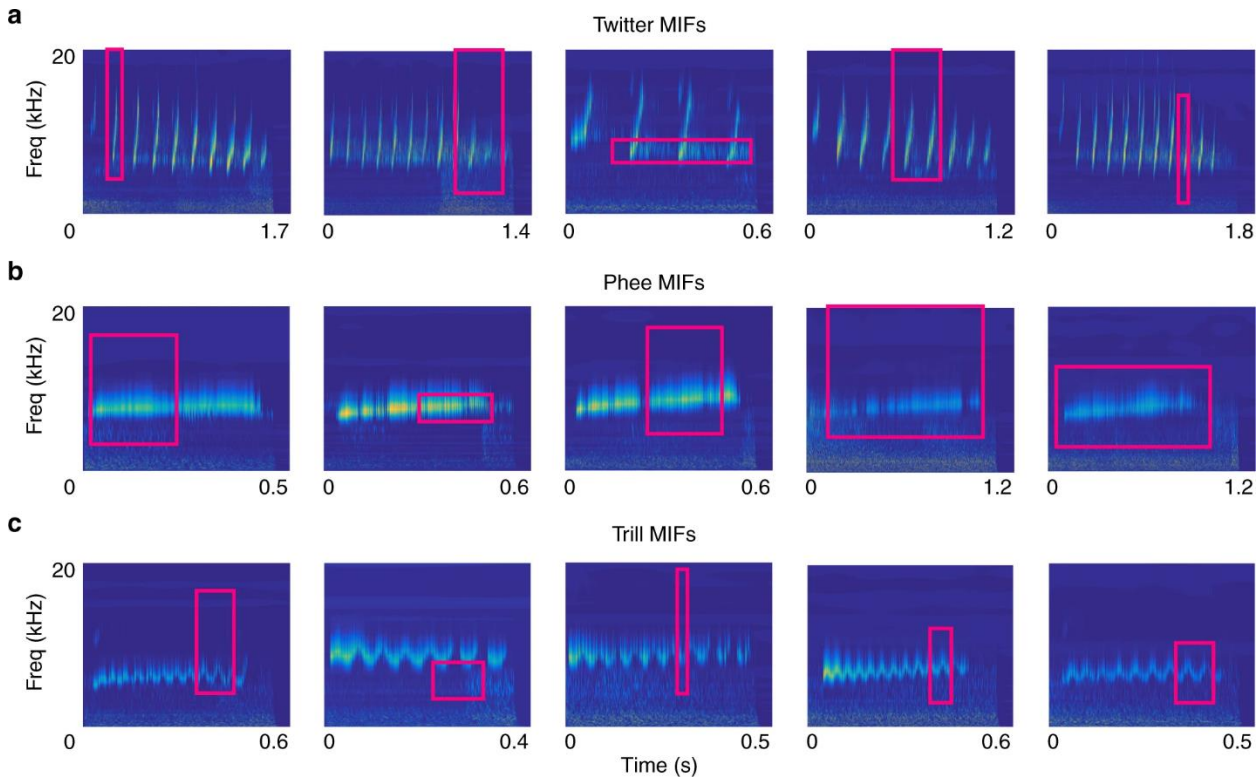


Figure 3 Most informative features for the classification of marmoset calls. Magenta boxes correspond to MIFs for the classification of **a** twitters vs. all other calls, **b** phees vs. all other calls, and **c** trills vs. all other calls, overlaid on the cochleagrams of the parent calls from which the MIFs were obtained

In Tables 1, 2, and 3, we show the pairwise information added by each MIF, the merits, and the weights of the top 10 MIFs for these classification tasks. Note that 1 bit of information corresponds to perfect classification. For twitters, detecting a single feature (the top MIF) was sufficient to gain 0.95 bits of information. Subsequent features probably detected only a few additional twitters without introducing new false alarms. For the other call types, however, the top MIF only provided 0.78 or 0.6 bits of information. Although successive MIFs individually had high merit (second column), they added little information to the top MIF (first column), likely because of redundancy—each MIF could only add a small number of additional hits without

introducing new false alarms. However, detecting these features was crucial for solving the task, as they ultimately elevated the total information to >0.9 bits. The MIFs have positive weights, suggesting that they are informative by virtue of their presence (rather than absence) in the target category. Because we approach very high levels of classification using our pairwise optimization of mutual information, and because joint optimization of mutual information across the entire MIF set is computationally expensive, we used the pairwise-optimized MIF set for all further analyses.

Table 1. Information Content of Twitter MIFs

MIF #	Added Info.	Merit	Weight
1	0.95	0.95	14.58
2	0.01	0.84	12.14
3	0.01	0.44	9.26
4	0.01	0.85	12.49
5	0.01	0.87	12.49
6	<0.01	0.87	12.49
7	<0.01	0.80	11.71
8	<0.01	0.84	12.3
9	<0.01	0.39	8.97
10	<0.01	0.34	8.62

Table 2. Information Content of Phee MIFs

MIF #	Added Info.	Merit	Weight
1	0.78	0.78	10.06
2	0.01	0.67	7.76
3	0.01	0.74	8.65
4	0.01	0.71	8.29
5	0.01	0.75	8.87
6	0.01	0.72	8.39
7	0.01	0.71	8.27
8	<0.01	0.71	8.27
9	<0.01	0.75	8.90
10	<0.01	0.71	8.49

Table 3. Information Content of Trill MIFs

MIF #	Added Info.	Merit	Weight
1	0.60	0.60	7.88
2	0.10	0.12	5.37
3	0.04	0.12	4.40
4	0.04	0.25	7.13
5	0.04	0.53	7.59
6	0.03	0.43	6.18
7	0.03	0.29	7.44
8	0.03	0.27	8.14
9	0.02	0.27	8.26
10	0.02	0.22	7.74

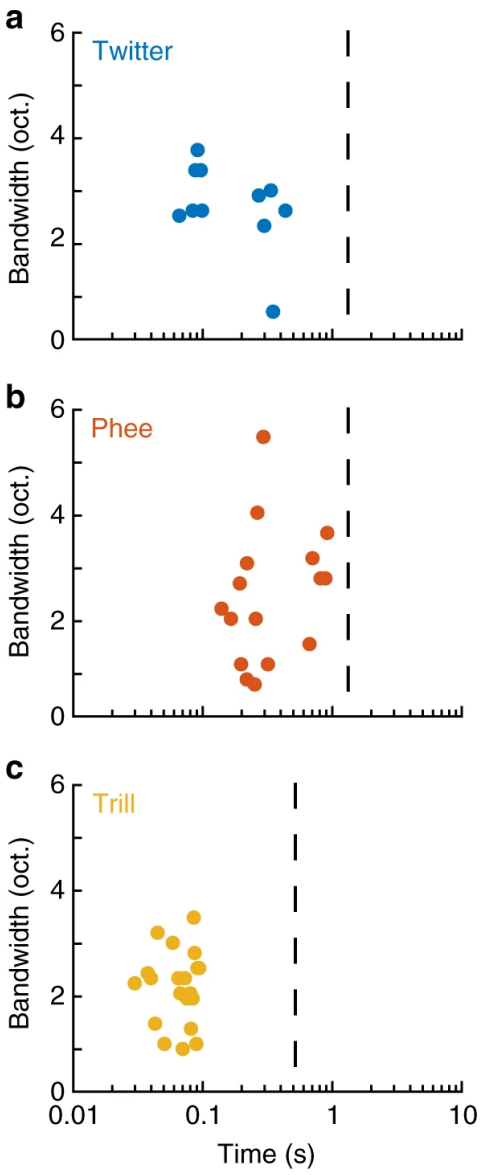


Figure 4 MIFs are of intermediate bandwidths and lengths. Scatter plot of the distribution of all MIFs for **a** twitters, **b** phees, and c trills as a function of their bandwidth and temporal integration period. Dashed line indicates the mean length of each call type. Colors are: blue—twitter, red—phee, yellow—trill

In frequency, MIFs neither encompassed the entire call bandwidth, nor consisted of only few frequency bands. In time, MIFs showed integration windows of the order of hundreds of milliseconds (Figure 4a–c). The mean MIF lengths were 215 ms, 68 ms, and 406 ms for twitters, trills, and phees, respectively. Compared to the average lengths of the calls (twitters: 1.25 s, trills: 0.5 s, phees: 1.27 s), these correspond to 17%, 14%, and 32% of mean call length, respectively. Interestingly, these lengths may correspond to timescales of temporal modulations in calls—for twitters, the sum of mean phrase length and mean inter-phrase interval is ~190 ms; for trills, the mean amplitude modulation period is ~30 ms. Thus, as with the initial feature set, MIFs for call classification were also of intermediate length and complexity.

2.2.3 Accurate Classification of Novel Calls Using MIFs Alone

To validate our model and to test the effectiveness of using only the MIFs for classifying call types, we used a novel set of calls consisting of 500 new within-category and 500 new outside-category calls drawn from the same 8 marmosets. This test call set did not significantly differ from the training set along any of the characterized parameters (red histograms in Figure 1). We conceptualized each MIF as a simulated template-matching neuron whose response to a stimulus was defined as the maximum value of the normalized cross-correlation (NCC) function. This simulated MIF-selective neuron ‘spiked’ whenever its response crossed its optimal threshold, i.e., when an MIF was detected in the stimulus. In Figure 5, we plot the spike rasters of simulated MIFselective neurons for twitter, phee, and trill (top 10 MIFs shown), responding to a train of randomly selected calls from the novel test set. Each spike was weighted by the log-likelihood ratio of the MIF and the weighted sum of responses in 50 ms time bins was taken as the evidence

in support of the presence of a particular call type. Although occasional false positives and misses occurred, over the set of MIFs the evidence in support of the correct call type was almost always the highest. Therefore, production invariant call categorization is a two-step process—first, MIFs are detected in the stimuli, and then each feature is weighted by its log-likelihood ratio to provide evidence for a call type.

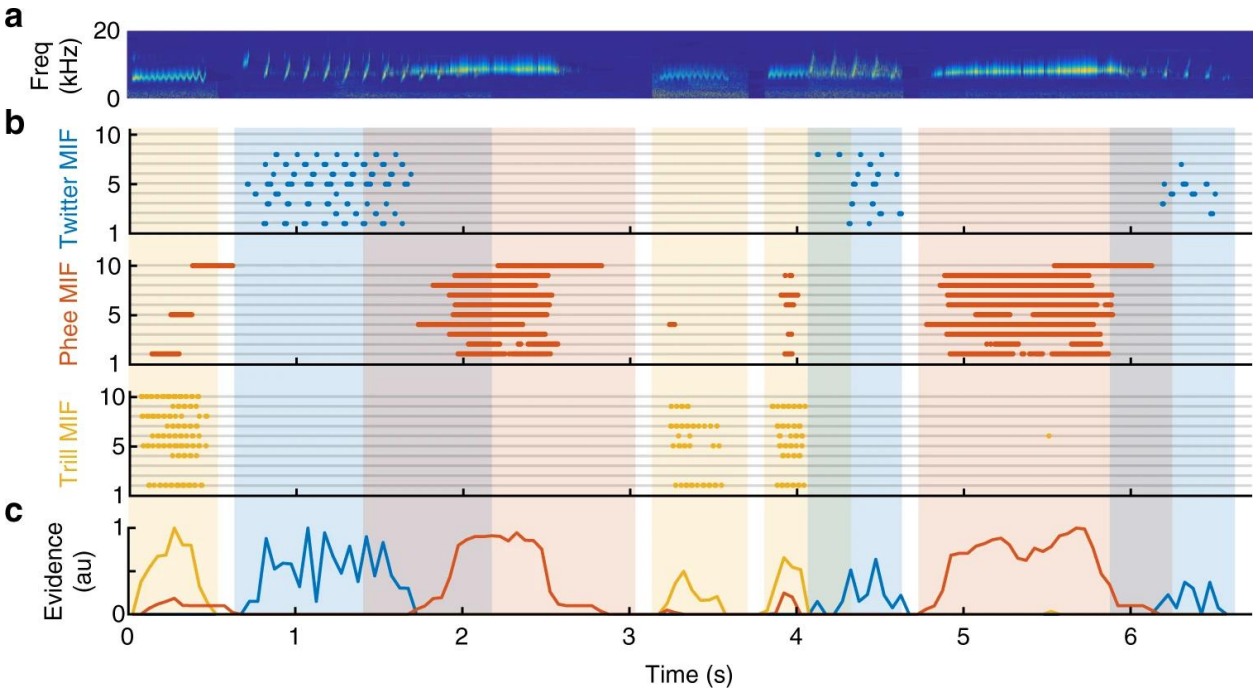


Figure 5 MIF responses to marmoset call sequences. **a** The cochleagram of a sequence of marmoset calls, some of which overlap. **b** Raster plot of the responses of the top 10 MIFs for twitter (top, blue), phee (middle, red), and trill (bottom, yellow). Each dot represents spiking of a putative MIF-selective neuron (i.e. when the response of the MIF exceeds its optimal threshold). **c** The evidence for presence of a particular call type, defined as the normalized sum of the firing rate of all MIF-selective neurons, weighted by their log-likelihood ratio. Over the duration of each call, the call type with the most evidence is considered to be present. Occasional false alarms are usually outweighed by true-positive MIF detections

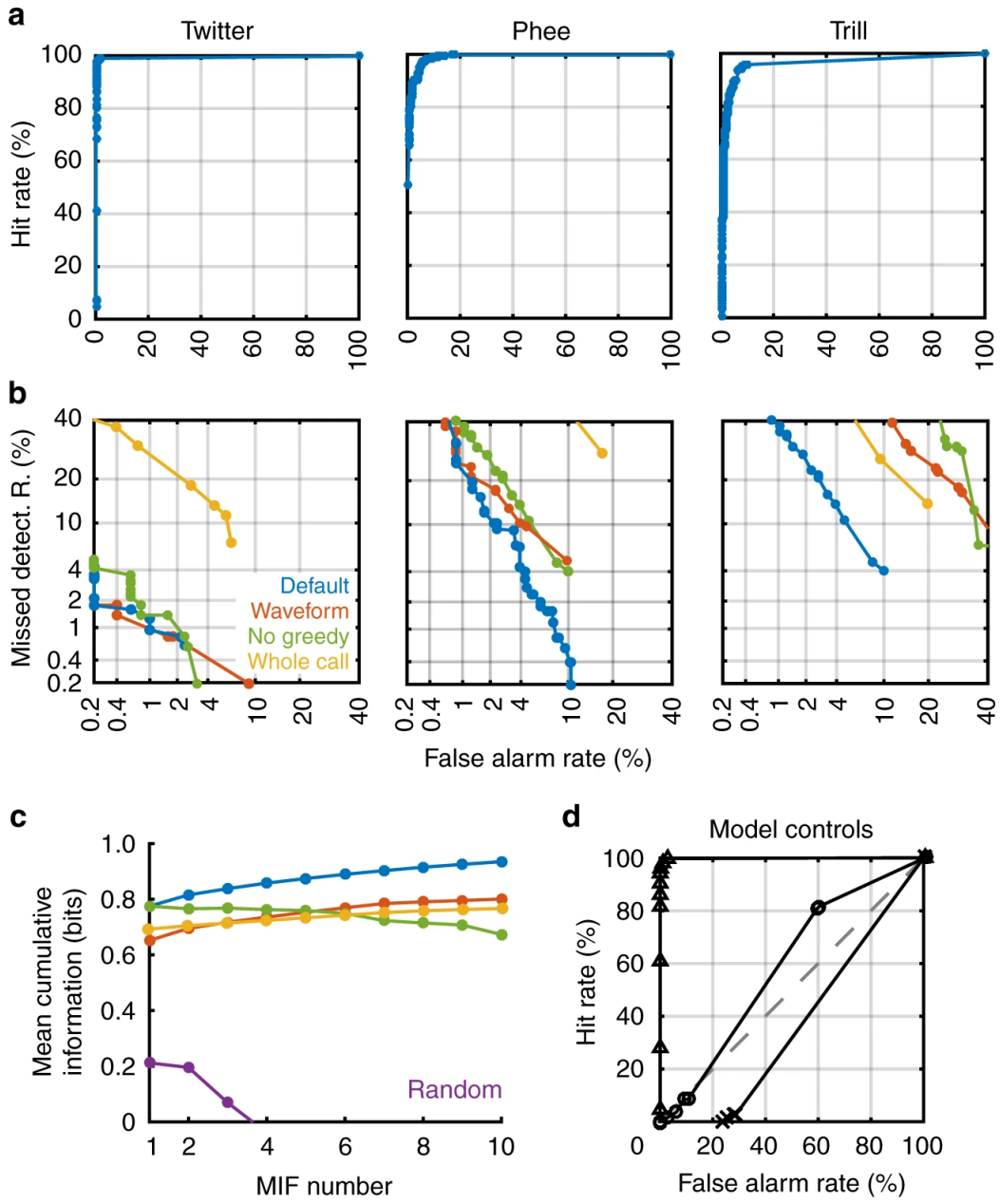


Figure 6 Classification performance and controls. **a** Receiver operating characteristic (ROC) curves for the classification of twitters, phees, and trills using MIFs alone. **b** Detection error tradeoff (DET) curves comparing the default model (blue) to other model variations: (i) MIF-based classification with acoustic waveforms (red), (ii) feature selection without greedy search (green), and (iii) when entire calls are used as features (yellow). **c** Comparison between various model

conditions (same as B) in terms of cumulative information added by each successive feature, averaged across all three call type classification tasks. Random (purple) is the classification of twitters using randomly selected features as MIFs, averaged across 20 trials. **d** ROC curves of three model controls: (i) classification of twitters when the model is trained on the twitters of 4 animals, and tested on twitters from 4 new animals (triangles), (ii) classification of twitters from other twitter calls (circles), and (iii) classification of twitters using trill MIFs (crosses)

We quantified the performance of the entire set of MIFs (n = 11, 16, and 20 for twitter, phee, and trill, respectively) for the classification of novel calls by parametrically varying an overall evidence threshold and computing the hit rate (true positives) and false alarm rate (false positives) at each threshold. From these data, we plotted receiver operating characteristic (ROC) curves (Figure 6a). In these plots, the diagonal corresponds to chance, and perfect performance corresponds to the upper left corner. The MIFs achieved >95% detection rates for all call types with very low false alarm rates.

2.2.4 Control Simulations

First, we ensured that our selection of 6000 initial random features adequately sampled stimulus space. To do so, we iteratively selected sets of MIFs using our greedy search algorithm from initial random sets from which previously picked MIFs were excluded. We found that distinct sets of MIFs that had similar classification performance could be selected in successive iterations (Appendix Figure 3). This suggests that our initial random feature set indeed contained several redundant MIF-like features, confirming the adequacy of our initial sampling.

Second, in order to determine the contributions of various model assumptions and parameters, we repeated this process of random initial feature generation, threshold optimization, and MIF selection in different scenarios. To better visualize these differences, we used detection-error tradeoff curves (Figure 6b), where perfect performance is the lower left corner. In this figure, the performance of the default model, as described above, is plotted in blue. First, when we used the acoustic waveform of calls instead of cochleagrams, classification performance was on average worse (Figure 6b; red), suggesting that phase information in the waveform may be detrimental for classification. Second, we used the features with top merits without greedy-search optimization for classification, and again found that performance compared to the default model was worse (Figure 6b, green). Finally, using entire calls as features, either treating entire individual calls as features ('grandmother cell' model; Figure 6b, yellow), or using the aligned and averaged training call as a single feature (Appendix Figure 4) also resulted in worse performance compared to the intermediate feature-based model.

In Figure 6c, we compare the average cumulative information added by successive features across all three call classification tasks (twitter vs. all other calls, trill vs. all other calls, and phee vs. all other calls) for each control simulation against the performance of the default model. The default model significantly outperformed (at $p < 0.01$, rank-sum test) the no greedy-search model for all classification tasks, after correcting for multiple comparisons (Bonferroni correction). Exact p-values corresponding to default model comparison with the constrained model and the no-greedy-search model were: twitter ($p= 0.000087$ and $p =0.00021$, respectively, rank-sum tests), trill ($p= 0.0058$ and $p= 0.00067$, respectively, rank-sum tests), and phee ($p= 0.00015$ and $p= 0.00021$, respectively, rank-sum tests). While the default model for trill exhibited significantly higher performance compared to the acoustic-waveform model ($p= 0.000091$, rank-sum test), the

default models for twitter and phee did not (p= 0.89 and p= 0.43, respectively, rank-sum tests). These results suggest that our underlying assumptions—using the cochleagram, unconstrained initial feature selection, and MIF optimization using a greedy search—were justified. Twitter MIFs were not qualitatively different when derived from calls emitted by a smaller set of animals (4 animals). Training on a set of 4 animals and testing on the other 4 animals yielded high performance (Figure 6d, triangles), confirming the robustness of using MIFs for categorization of new calls. Twitter MIF performance in classifying twitters from other twitters was near-chance, suggesting that the estimation of mutual information values was unbiased (Figure 6d, circles). Finally, MIFs derived for one task (such as trill vs. other calls) showed chance level performance for other tasks (such as twitter vs. other calls; Figure 6d, crosses), demonstrating the task-dependence of the derived MIFs.

2.2.5 The Precedence of Intermediate Features for Classification

We have previously shown that features of intermediate lengths and complexities possess high individual merits for classification (Appendix Figure 2). We have also shown that the set of MIFs is composed intermediate features (Figure 4a–c). To directly test whether features of intermediate size were indeed the most informative, we re-derived MIFs after constraining the initial set of features to particular time and frequency bins and quantified model performance (Figure 7). When we constrained the features to be only small (<100 ms and <1 oct.) or removed all small features, performance was worse than the default model (Figure 7, top row). Similarly, model performance was worse compared to the default model when we constrained to only large features (>250 ms and >2 oct.) or removed all large features. When we constrained bandwidth and

time independently to be large or small, model performance was worse compared to the default model, with large values being more detrimental (Figure 7, bottom row). As previously discussed, using the largest possible features (whole calls or average call) resulted in poor classification performance as well. These results demonstrate that features of intermediate size indeed provide the best classification performance.

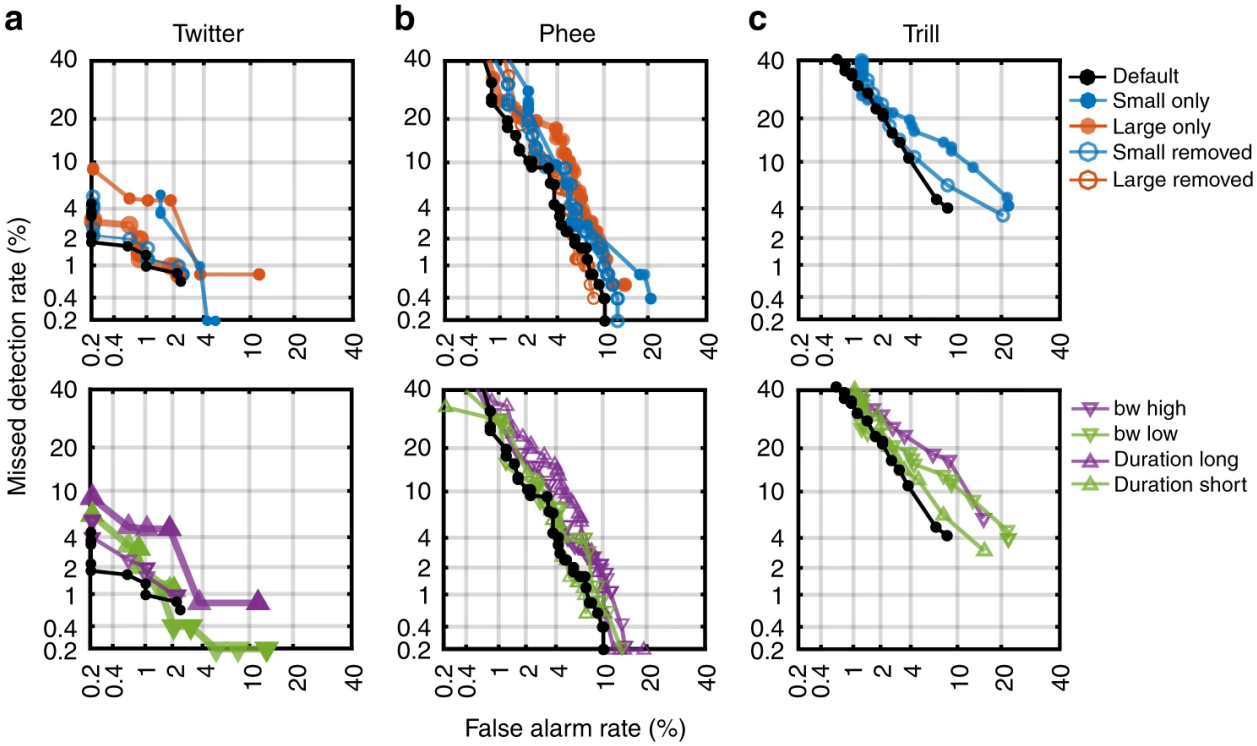


Figure 7 The precedence of intermediate features for classification. DET curves for call classification using features of different sizes, bandwidths, and durations for the classification of **a** twitters, **b** phees, and **c** trills. The default model is in black. Top row shows performance when using small features only (<100 ms and <1 oct; blue discs.) or excluding small features (blue circles), and using large features only (>250 ms and >2 oct.; red discs) or excluding large features (red circles). For trills, some of these conditions fall outside the range of the axes. Bottom row shows performance when feature bandwidths and durations were independently varied. Because

of the short duration of trill calls, we did not test the effect of using only long duration features. Symbols are: purple inverted triangles—high bandwidth features only, green inverted triangle—low bandwidth features only, purple triangle—long-duration features only, green triangle—short-duration features only

2.2.6 MIF Tuning Properties Match Neural Responses from A1 L2/3

MIF tuning properties match neural responses from A1 L2/3. So far, we have demonstrated MIFs derived purely using theoretical principles can achieve high levels of production-invariant call categorization performance. We then asked whether the auditory system uses such an optimal feature-based approach for call classification. To explore this possibility, as a first step, we generated tuning curves of model neurons that were selective for the theoretically derived MIFs, and asked if these tuning curves matched previous experimental observations. In this effort, we were restricted by the appropriateness and availability of previous data. To do so, we first constructed cochleagrams of stimuli, such as trains of frequency-modulated sweeps, amplitude modulated tones, noise bursts, clicks, two-tone combinations, etc. We then used the maximum value of the NCC function as a metric of the model MIF neurons' response to these stimuli, as we did earlier for calls. These responses were conceptualized as membrane potential responses, which elicited spiking only if they crossed each MIF neuron's optimal threshold. We used a power law nonlinearity, applied to the maximum NCC values (see Methods, Equation 2), to determine the firing rate responses of model MIF neurons (Appendix Figure 6). We then compared these model MIF tuning curves to neural data from marmoset primary auditory cortex (A1).

Although, the MIF model did not have prior access to neurophysiological data, we found that model MIF neural tuning recapitulated actual data to a remarkable degree, both at the population and single-unit levels. For example, the population of model MIFs showed high preference for natural calls compared to reversed calls (Figure 8a, bottom), similar to observations by Wang and Kadia (reproduced in Figure 8a, top). The high sparseness of auditory cortical neurons is well-documented (Hromadka et al 2008; Hromadka and Zador 2009; Sadagopan and Wang 2009). The responses of model MIF-selective neurons were also sparse—only few MIF neurons were activated by any given stimulus set, and only after extensively optimizing the parameters of the stimulus set to drive-specific model MIF neurons. For example, in Figure 8b (top), we show a single-unit recording from a marmoset A1 L2/3 neuron that did not respond to most stimulus types (reproduced from Sadagopan and Wang), and only strongly responded to twotone stimuli. Twitter MIFs (Figure 8b, bottom) were similarly not responsive to most stimulus types, and only responded to carefully optimized linear frequency-modulated (lFM) sweeps. None of the model twitter and trill MIF-selective neurons responded to pure tones (Figure 8b, bottom), similar to many A1 L2/3 neurons.

Most strikingly, we could recapitulate some specific and highly nonlinear single-neuron tuning properties as well. Figure 8c (top; reproduced from Sadagopan and Wang) is a single-unit recording from marmoset A1 L2/3 that did not respond to pure tones, but selectively responded to upward lFM sweeps of specific lengths (~80 ms). Responses of at least three of the top 5 twitter

MIF-selective model neurons showed similar tuning for 80 ms long upward lFM sweeps (Figure 8c, bottom). A second peak at ~40 ms was also present in responses of two model twitter MIF-selective neurons, also matching the experimental data. Figure 8d (top; reproduced from Sadagopan and Wang) shows another single-unit recording from marmoset A1 L2/3, where the

neuron did not respond to single lFM sweeps (lightest gray line), but strongly responded to trains of upward lFM sweeps occurring with 50 ms inter-sweep interval. The neuron's response scaled with the number of sweeps present in the train (darker colors correspond to more sweeps). Three of the top 5 twitter MIF-selective neurons also showed remarkably similar tuning (Figure 8d, bottom)—these model neurons did not respond to single sweeps, but responded to trains of at least 2 or more sweeps occurring with a 50 ms inter-sweep interval. Taken together, these data suggest neurons tuned to MIF-like features are present in A1 L2/3. Therefore, we predict that a spectral-content based representation of calls in the ascending auditory pathway becomes largely a feature-based representation in A1 L2/3.

Consistent with the prediction of feature selectivity, we also found neurons in A1 of both marmosets and guinea pigs that respond selectively to conspecific call features. In Figure 9, we present the spike rasters of example single neurons in both marmoset and guinea pig A1 responding to marmoset (Figure 9a) and guinea pig calls (Figure 9b), respectively. We presented multiple exemplars of each call type as stimuli. These example neurons responded at specific time points to a few call stimuli, typically across 1–3 categories. Such responses are consistent with our feature-based model because single features alone do not completely categorize calls, i.e., MIFs do not have 1 bit of information for categorization. Rather, combinations of features weighted by their log-likelihood ratios are necessary to ultimately achieve complete call category information. These data provide promising support for our model, but further experiments are necessary to: (1) determine how informative these neural features are about call category and how they compare with model features, (2) to confirm where such responses arise in the auditory pathway, and (3) to account for possible low-level confounds. Experiments are presently ongoing to address these issues.

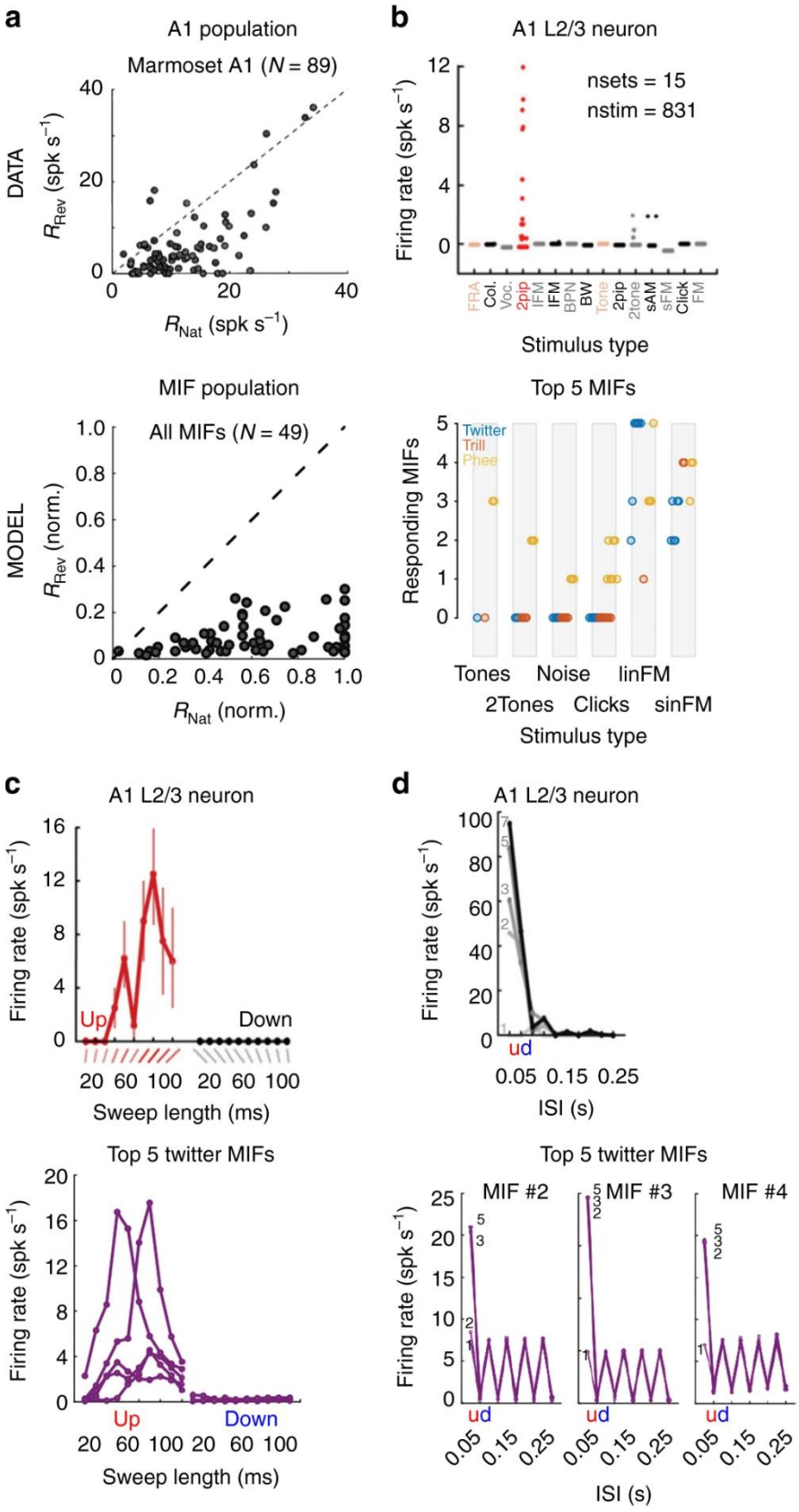

a
A1 population
Marmoset A1 (N = 89)
DATA
R_{Rev} (spk s^{-1})
R_{Nat} (spk s^{-1})
MIF population
All MIFs (N = 49)
MODEL
R_{Rev} (norm.)
R_{Nat} (norm.)
b
A1 L2/3 neuron
nsets = 15
nstim = 831
Firing rate (spk s^{-1})
FRA Col. Voc. 2pip IFM IFM BPN BW Tone 2pip 2tone sAM sFM Click FM
Stimulus type
Top 5 MIFs
Twitter
Trill
Phee
Responding MIFs
Tones Noise linFM
2Tones Clicks sinFM
Stimulus type
c
A1 L2/3 neuron
Firing rate (spk s^{-1})
Up
Down
20 60 100 20 60 100
Sweep length (ms)
Top 5 twitter MIFs
Firing rate (spk s^{-1})
Up
Down
20 60 100 20 60 100
Sweep length (ms)
d
A1 L2/3 neuron
Firing rate (spk s^{-1})
ud
0.05 0.15 0.25
ISI (s)
Top 5 twitter MIFs
MIF #2
MIF #3
MIF #4
Firing rate (spk s^{-1})
ud
0.05 0.15 0.25
ISI (s)

Figure 8 Predictions of putative MIF-neuron tuning properties match cortical data. (**a–d**, top row) Neural data from marmoset A1. (**a–d**, bottom row) Model predictions. (**a**-top) Preference of marmoset A1 responses for natural twitters over time-reversed twitters. (**a**-bottom) Preference of model MIF neurons for natural calls over reversed calls. (**b**-top) Sparse responses of marmoset A1 L2/3 neuron. (**b**-bottom) Sparse responses of MIF neurons. The number of MIF neurons showing responses to the stimulus categories on the x-axis are plotted. Colors correspond to call type (blue—twitter, red—trill, yellow—phee). (**c**-top) Marmoset A1 L2/3 neuron tuned to upward lFM sweeps of a specific length (~80 ms). Error bars correspond to ±1 SD. (**c**-bottom) Twitter MIF neurons show similar tuning. (**d**-top) Marmoset A1 L2/3 neuron that does not respond to single lFM sweeps but shows tuning to trains of upward lFM sweeps with 50 ms inter-sweep interval. Grayscale corresponds to the number of lFM sweeps in the train. (**d**-bottom) Three of the top 5 twitter MIFs showed similar tuning for lFM sweep trains. a-top reproduced from Wang and Kadia (2001), b–d top reproduced from Sadagopan and Wang (2009)

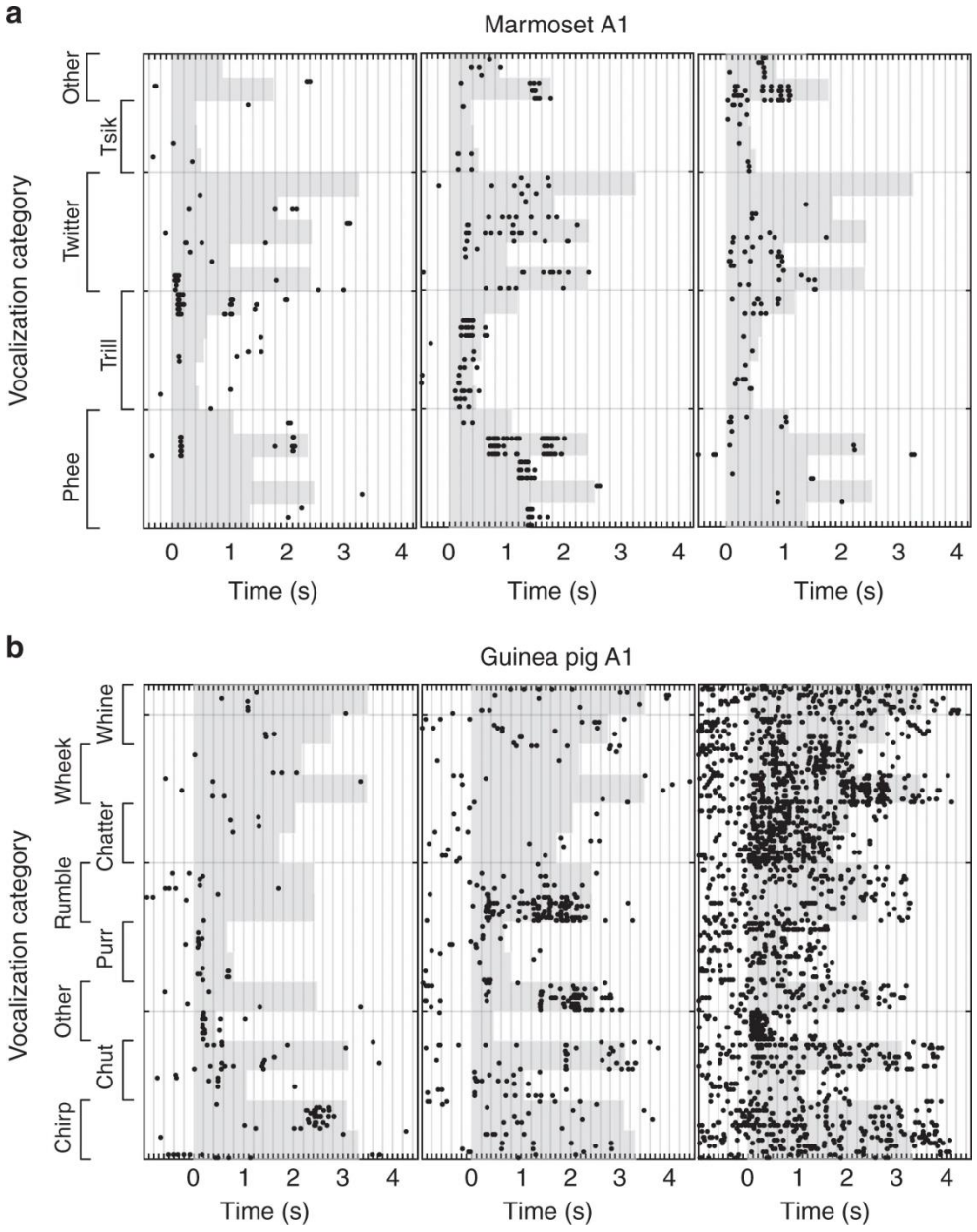


Figure 9 Feature selectivity in cortical neurons. a Spike rasters of three single units from marmoset A1 responding to marmoset call stimuli. Black dots correspond to spikes; gray shading corresponds to stimulus duration (different calls have different lengths). Note that spikes occur at specific times, and in response to 2 or 3 call types, suggesting that the neurons are responding to smaller features within these calls. b Spike rasters of three single units from guinea pig A1 responding to guinea pig call stimuli

2.2.7 Task-dependent MIF Detection as a General Computation

To determine whether MIF-based representations of sounds could also be used for optimally solving other tasks, we performed three proof-of-principle simulations using limited available datasets. First, we tested whether we could accurately determine caller identity using an MIF-based approach. We generated training and test sets of 60 twitters each from eight marmosets, and generated 500 initial random features from the training set. We applied the greedy-search algorithm to determine the MIFs for caller identification in a caller A vs. all other callers task (Figure 10a). We found that similar to call categorization, caller identification could also be achieved using a small number of MIFs (n = 4). If caller identification was performed in a binary fashion (four classifications between two animals each), in half of these tasks, classification could be accomplished using less than 3 MIFs, indicating that the calls of these marmosets probably differed along the frequency axis. This is because if there are clear differences in dominant frequency (for example, Animal 1 vs. 4 in Figure 1d), all features that lie in one animal's frequency range will detect all of that animal's calls and none of the other animal's calls. During the greedy search procedure, these features will be considered redundant and reduced to a single feature. In the other half, more MIFs were required for caller identification, and in general, MIFs were larger than those for call-type classification. This is likely because the differences between twitters produced by these animals are smaller compared to the differences between call types and can only be resolved in a higher dimensional space. Thus, integration over more frequencies and a larger time window may be necessary to resolve caller differences. In Appendix Figure 7, we plot the ROC for caller identification between a pair of marmosets with overlapping dominant frequencies.

The MIF-based approach (n = 20 MIFs) achieved >80% hit rates with <10% false alarm rate for caller identification.

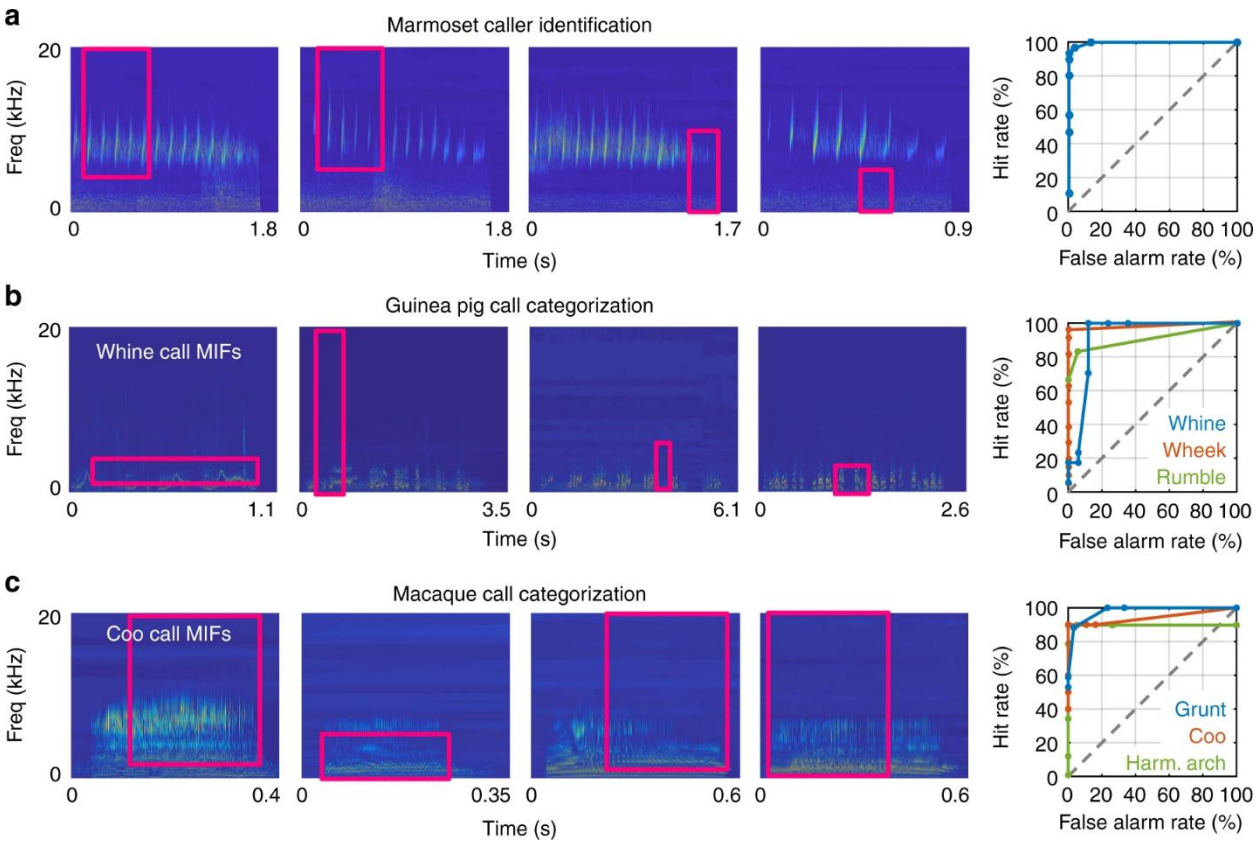


Figure 10 The applicability of MIF-based classification for other auditory tasks. The top four MIFs and ROC curves for: **a** marmoset caller identification (twitter calls), **b** Guinea pig call classification (MIFs for whine calls shown). Colors are: blue—whine, red—wheek, green—rumble. **c** Macaque call classification (MIFs for coo calls shown). Colors are blue—grunt, red—coo, green—harmonic arch

Second, we tested whether MIF-based call classification generalized to other vocal species, using guinea pig and macaque call classification as examples. Guinea pigs are highly vocal rodents that produce seven primary call types (Eisenberg 1974; Berryman 1976; Grimsley et al 2012),

which are highly overlapping in the low frequency end of the spectrum, and show high production variability. We used the MIF-based approach to classify guinea pig call types (whine, wheek, and rumble) from all other guinea pig call types. Similar to marmosets, guinea pig classification could be accomplished using a handful of features (12, 9, and 3 MIFs for whine, wheek, and rumble), and MIF-based classification achieved high performance levels (Figure 10b). Similarly, we implemented the MIF-based algorithm to classify macaque calls (using 5, 4, and 9 MIFs for coos, grunts, and harmonic arches) from a limited macaque call dataset (Hauser 1998) and achieved high classification performance (Figure 10c). These proof-of-principle experiments demonstrate that an MIFbased approach indeed succeeds for different auditory classification tasks and in different species, suggesting that building representations of sounds using task-relevant features in auditory cortex may be a general auditory computation.

2.3 Discussion

In these experiments, we set out to understand the computations performed by the auditory system that enable the categorization of behaviorally critical sounds, such as calls, despite wide variations in the spectrotemporal structure of calls belonging to a category (production variability). We found that the optimal theoretical solution is to detect the presence of informative midlevel features (termed MIFs) in calls. These MIFs generalize over production variability, and conjunctions of MIFs accomplish production-invariant call classification with high accuracy. Critically, the tuning properties of model MIF-selective neurons matched previous recordings from marmoset A1 to a surprising degree. MIF-based classification was also successful for other tasks

(marmoset caller identification), and in other species (guinea pig and macaque call recognition). Our results suggest that the representation of sounds in higher auditory cortical areas is based on the detection of optimal task-relevant features.

An implication of our results is that in higher auditory processing stages, neural representations of sounds serve-specific behavioral purposes. For example, the MIF-based classification approach that we proposed here is targeted to solve well-defined classification problems. At earlier stages of the auditory pathway, however, it may be more important to faithfully represent sounds using basis sets that enable the accurate and complete encoding of novel stimuli. Previous theoretical studies have proposed, for example, that natural sounds can be efficiently encoded using spike patterns, where each spike represents the magnitude and timing of input acoustic features (Smith and Lewicki 2006). However, when optimized to encode the complete waveforms of natural sound ensembles, the kernel functions that elicit each spike show a striking similarity to cochlear filters. The advantage of this approach is that novel stimuli can be completely encoded using these kernel functions. In our approach, the input to our model implements a similar encoding schematic—in the cochleagram, inputs are encoded as spatiotemporal spike patterns, where each spike is the result of cochlear filtering. In this early representation, while information about category identity is present, it is distributed in the activity of many neurons in a high-dimensional space. We propose that in later processing stages, this early representation is transformed into a representation where category identity is more easily separable. By encoding MIF-like features, sound representation in later processing stages is less useful for high-fidelity encoding (although stimulus reconstruction is possible, see Appendix Note and Appendix Figure 5), but is instead goal-oriented. However, this means that each task will require a distinct set of MIFs for optimal performance, and animals likely perform a large number of such

behaviorally relevant tasks. The observed >1000- fold increase between the number of cochlear inputs and auditory cortical neurons may partially result from this necessity to encode a multitude of task-dependent MIFs. Previous theoretical studies have suggested that the generation of redundant and over- complete representations of sounds to solve spatial localization problems might underlie this increase in the number of neurons (Asari et al 2006). Our study proposes another computational reason why such an expanded representation of sounds may be necessary.

Another powerful method to accomplish classification uses hierarchical convolutional neural networks, or deep networks. In these models, layers of filtering, normalization, and pooling operations are cascaded, resulting in individual units exhibiting increasingly complex tuning properties (Rasanene et al 2016; Khalighinejad et al 2017; Kell et al 2018). A final layer reads out class identity. Deep networks can achieve near-human levels of performance on specific tasks, but carry some disadvantages. First, they often require training data of the order of millions of samples. In the visual domain, deep networks appear not to use the same features as humans for object classification (Ullman et al 2016). Finally, an intuitive explanation for how deep network models actually accomplish classification is not yet available. In our approach, we explicitly train our MIF neurons to extract maximally distinguishing features, providing insight into why certain features are represented amongst these neurons. Our model does not require as extensive a training set. We consider our approach complementary to the deep learning approach, in that we aim to provide an explicit and intuitive explanation of why certain features are extracted, as opposed to matching human performance using complex model architectures.

Conceptually, our MIFs may be similar to ‘image signatures’ obtained by recently developed unsupervised methods (Anselmi et al 2016) (see Appendix Discussion). Our approach is complementary to alternative experimental approaches, such the characterization of neural

tuning along an exhaustive list of call parameters (DiMattina and Wang 2006), characterizing call tuning as tuning for regions of the modulation spectrum (Hsu et al 2004; Woolley et al 2005; Stowell and Plumbley 2014), and combinations of these methods in conjunction with machine learning tools (Fukushima et al 2015) (see Appendix Discussion). Our results suggesting auditory cortex as a locus where the neural representation of vocalization sounds generalizes over production variability is consistent with a recent study showing that neurons in the auditory cortex of ferrets show robust responses to vowel identity tolerant to manipulations of various vowel features (Town et al 2018).

Mechanistically, neural selectivity for MIFs may be generated (1) gradually along the ascending auditory pathway, or (2) de novo in cortex. Single-neuron feature selectivity often (but not always, see below) leads to selectivity for one or a few call types, and analyzing call selectivity of neurons at different auditory processing stages could provide insight into where MIF-based representations might be generated in the auditory pathway. In early auditory processing stages, evidence for call selectivity at the single-neuron level is minimal. For example, at the level of the cochlear nucleus, few single neurons in species other than mice show call selectivity (Pollak 2013). At the level of inferior colliculus, a population-level bias in call-selectivity has been reported (Pollak 2013; Portfors et al 2009; Holmstrom et al 2010), but evidence for single-neuron level call-selectivity is equivocal (Suta et al 2003). It is only at the level of auditory cortex where clear single-neuron selectivity for calls or call features has been observed. Therefore, it is quite likely that selectivity for MIF-like features in species with spectrotemporally complex calls is generated at the level of auditory cortex. This is supported by the expansion in the number of cortical neurons mentioned above. Importantly, the cortical emergence of MIF-based representations is also

supported by the fact that MIF-like responses have been observed in the superficial layers of marmoset A1 (Sadagopan and Wang 2009).

We propose the following hierarchical model for auditory processing based on the representation of task-relevant features. In thalamorecipient layers of A1, representation of sound identity is still based on spectral content. This is reflected in the strongly tone-tuned responses of A1 L4 neurons. From these neurons, tuning for MIF-like features may be generated using nonlinear mechanisms, such as combination-sensitivity. For example, the tuning properties of the marmoset A1 responses shown in Figure 8 was determined to be the result of selectivity for precise spectral and temporal combinations of two-tone pips (Sadagopan and Wang 2009). This is also consistent with a recent computational model showing that combinations of spectrotemporal kernels, optimized for representing natural sounds, recreates aspects of experimentally observed spectrotemporal receptive fields from recordings in cat auditory cortex (Mlynarski et al 1981). Further experiments, probing call and feature selectivity in identified layers of A1, are necessary to more precisely address where selectivity for MIF-like features first emerges in the ascending auditory pathway, and at what stage MIFs are combined to result in a categorical read-out. Once categories are detected, further hierarchical processing stages might be necessary to accomplish more sophisticated behavioral goals, such as caller identification, integration of social context with call perception, or decoding the emotional valence of calls.

In conclusion, we propose a hierarchical model for solving a central problem in auditory perception—the goal-oriented categorization of sounds that show high within-category variability, such as speech (Petersen and Barney 1952; Hillenbrand et al 1995) or animal calls (Wang 2000). Our work has broad implications as to where in the auditory pathway categorization begins to emerge, and what features are optimal to learn in categorization tasks. For example, the lack of

distinction of perceptual categories of English /r/ and /l/ by native Japanese spekers might be a consequence of not learning and encoding (MacKain et al 1981; Raizada et al 2010) the optimal features necessary for this /r/-/l/ categorization, as it is not task-relevant for Japanese speech. Our model would predict that /r/-/l/ category learning would cause selective responses to develop for new task-relevant features, and primarily reflected in changes to the A1 L2/3 circuit. Consistent with this hypothesis, a recent study showed that training humans to categorize monkey calls resulted in finer tuning for call features in the auditory cortex (Jiang et al 2018). We therefore suggest that the neural representation of sounds at higher cortical processing stages uses task-dependent features as building blocks, and that new blocks can be added to this representation to enable novel perceptual requirements.

2.4 Methods

2.4.1 Vocalizations

All procedures conformed to the NIH Guide for Care and Use of Laboratory Animals. All marmoset procedures were approved by the Institutional Animal Care and Use Committee (IACUC) of The Johns Hopkins University. All guinea pig procedures were approved by the IACUC of the University of Pittsburgh. We used vocalization recordings from 8 adult marmosets, both male and female, for these experiments. Marmoset calls were recorded from a marmoset colony at The Johns Hopkins University using directional microphones (Agamaite et al 2015). Guinea pig calls were recorded from 3 male and 3 female adult guinea pigs. Two or more guinea pigs with varied social relationships were placed on either side of a transparent divider in a sound attenuated booth. Directional microphones, suspended above the guinea pigs were used to record calls. Calls were recorded using Sound Analysis Pro 2011(Tchernichovski et al 2000), digitized at a sampling rate of 48 KHz, low-pass filtered at 24 KHz, manually segmented using Audacity, and classified into different call types.

2.4.2 Random Feature Generation

All modeling was implemented in MATLAB. We focused on classifying each of three major marmoset call types, twitter, trill, and phee, from all other call types. That is, three main binary classification tasks— twitter vs. all other calls, trill vs. all other calls, and phee vs. all other calls were considered. We set up the categorization tasks as a series of binary classifications based

on the results of an earlier study of visual categorization that demonstrated the advantages of features learnt using multiple binary classifications compared to those learnt using a single multi-way classification. Specifically, in that study, multiple binary classifications resulted in features that were distinctive and highly tolerant to distortions (Akselrod-Ballin et al 2008). For each classification task, we first generated training datasets, which consisted of 500 random within-class calls (e.g., twitters) produced by 8 animals (about 60 calls per animal), and 500 random outside-class calls (e.g., trills, phees, other calls) produced by the same 8 animals. In order to convert sound waveforms of the calls into a physiologically meaningful quantity, we transformed these calls into cochleagrams using a previously published auditory nerve model (Zilany et al 2014) using human auditory nerve parameters with high spontaneous rate. We used human auditory nerve parameters because of the close similarity between marmoset and human audiograms (Osmanski et al 2011). The output of this model was the time-varying activity pattern of the entire population of auditory nerve fibers, and resembles the spectrogram of the call (Figure 2a, b). We then extracted 6000 random features from these 500 within-class cochleagrams. To do so, we randomly chose a center frequency, bandwidth, onset time and length and extracted a snippet of activity from the cochleagram. Each feature thus corresponded to the spatiotemporal pattern of activity of a subset of auditory nerve fibers within a specified time window (magenta box in Figure 2b). We used rectangular feature shapes rather than other shapes to minimize assumptions – for example, an ellipse shaped feature would imply that the weighting of individual auditory nerve fibers changes over time. For twitters, to ensure that smaller features were well-sampled, 2000 of these features were restricted to have a bandwidth less than 1 octave and a duration less than 100 ms. The bandwidth and duration of the remaining 4000 features were not constrained.

2.4.3 Feature Complexity

We characterized feature complexity using the reduced kurtosis of the activity distribution of all auditory nerve fibers contained within a feature. Briefly, if the feature was an empty region of the cochleagram, or a region of uniform activity, the activity of all nerve fibers in all time bins would be about equal. This activity would thus be normally distributed, and show a reduced kurtosis value of zero. At the other extreme, for entire calls, there would be many bins of high activity, and a large number of bins with zero activity, resulting in an activity distribution with very high reduced kurtosis. We hypothesized that midlevel features that represent aspects of calls such as frequency-modulated sweeps or combinations of phrases over time would show intermediate reduced kurtosis values, and be more informative than low-level (tones) or high-level (entire calls) features.

2.4.4 Threshold Optimization

We defined the response of a feature to a call as the maximum value of the normalized cross-correlation (NCC) function between the feature's cochleagram and the call's cochleagram, restricted to the auditory nerve fibers that are represented in the feature. Note that this means features can only be detected in the frequency range that they span, but can be detected anywhere in time within a call. NCC is a commonly used metric to quantify template-match. To compute the NCC, the feature and the cochleagram patch at each lag were normalized by subtracting their respective mean values and dividing by their respective standard deviations before convolving them. This results in a value between −1, signifying that the feature and cochleagram patch at that

lag are completely anticorrelated, and +1, signifying a perfect match between the feature and the cochleagram. Because this is a computation-intensive step, template matching was implemented on an NVIDIA GeForce 980 Ti GPU. For each feature, then, we obtained 500 within-class responses, and 500 outside-class responses (response histograms of an example feature in Figure 2c). To transform these continuous response distributions into a binary detection variable, we used mutual information to quantify the information provided by a feature about the class (within- or outside-class) over a parametrically varied range of thresholds. We computed mutual information following the method of Ullman et al. (Ullman et al 2002), by measuring the frequency of detecting a feature f_i at a given threshold θ_i ($f_i = 1$ if present, 0 if absent) in the within-class ($C = 1$) or outside-class ($C = 0$) cochleagrams as:

$$I(f_i(\theta_i), C) = \sum_{\substack{f_i=\{0,1\} \\ C=\{0,1\}}} p(f_i, C) \log\left(\frac{p(f_i, C)}{p(f_i)p(C)}\right) \quad (1)$$

where $p(C)$ was assumed to be 0.10. We empirically verified that features identified were insensitive to variations of this value. The optimal threshold for each feature was taken to be the threshold value at which the mutual information was maximal, and the merit of each feature was taken to be the maximum mutual information value in bits (Figure 2c). The weight of each feature was taken to be its log-likelihood ratio. At the end of this procedure, each of the initial 6000 features were allocated a merit, a weight, and an optimal threshold at which each individual feature's utility for classifying calls as belonging to within- or outside-class was maximized. Note that merit and weight are distinct quantities that need not be monotonically related. For example, if the lack of energy in a frequency band is indicative of a target category, features that contain energy in this frequency band will be detected often in the other categories, but not in the target

category. The feature will thus have high merit for classification, as it is informative by its absence, but have a negative weight.

2.4.5 Greedy Search

Because we chose initial features at random, many of these features individually provided low information about call category, and many of the best features for classification were similar, or redundant. Therefore, to extract maximal information from a minimal set of features for classification, we used a greedy search algorithm (Ullman et al 2002) to iteratively (1) eliminate redundant features, and (2) pick features that add the most information to the set of selected features. The minimal set of features that together maximize information about call type were termed maximally informative features (MIFs). The first MIF was chosen to be the feature with maximal merit from the set of all 6000 initial random features. Every consecutive MIF was chosen to maximize pairwise added information with respect to the previously chosen MIFs. Note that these consecutive features need not have high merit individually. We iteratively added MIFs until we could no longer increase the hit rate without increasing the false alarm rate. Practically, this meant adding features until total information reached 0.999 bits, or individual features added less than 0.001 bits, whichever was reached earlier. At the end of this procedure, a small set of MIFs, containing the optimal set of features for call classification was obtained.

2.4.6 Analysis and Statistics

To test how well novel calls could be classified using these MIFs alone, we generated from the same 8 animals a test set of 500 within- and outside-class calls that the model had not been exposed to before. We computed the NCC between each test call and MIF, and considered the MIF to be detected in the call if the maximum value of the NCC function exceeded its optimal threshold. If detected, the MIF provided evidence in favor of a test call belonging to a call type, proportional to its log-likelihood ratio. We then summed the evidence provided by all MIFs and generated ROC curves of classification performance by systematically varying an overall evidence threshold. We used the area under the curve (AUC) to compare ROC curves for classification performance by MIFs generated with different constraints (see Results). Statistical significance was evaluated using ranksum tests, with Bonferroni multiple-comparisons corrections, for comparing between these conditions, and for comparing performance to a large number of simulations generated using random MIFs.

2.4.7 Generating Predictions

To generate predictions of the responses of putative MIF-selective neurons to other auditory stimuli, we first generated a large battery of stimuli that have been used in previous recordings from marmoset A1, and computed their cochleagrams as earlier. We then computed the maximum value of the NCC function between the MIF and the stimulus cochleagram. This resulted in response values that could be conceptualized as equivalent to membrane potential (V_m) responses. These were converted to firing rates by applying a power law nonlinearity, of the form:

$$\mathrm{FR} = k \cdot [V_m - \theta]^p \quad (2)$$

where FR is the firing rate response in spk s−1, θ is the MIF's optimal threshold, p is the exponential nonlinearity set to a value of 4, and k is an arbitrary scaling factor.

2.4.8 Call Reconstruction from MIFs

To reconstruct calls, we conceptualized MIFs as MIF-selective neurons, and considered the times at which NCC values exceeded the optimal threshold to be the spike times of these neurons. MIF spike times were computed with a time resolution of 2ms to simulate refractoriness, and alphafunctions were convolved with the spike times to determine the peak time at which each MIF was detected. A copy of the MIF cochleagram was then placed at the peak time, or summed (with log-likelihood weights) if overlapping with a previously placed cochleagram. The accuracy of reconstruction was defined as the NCC between the original stimulus and its reconstructed version at zero lag.

2.4.9 Electrophysiology Methods

Predictions generated from the MIFs were compared to earlier recordings from marmoset A1. All recordings were from the auditory cortex of adult marmosets. Population data comparing natural to reversed twitters were obtained from Wang and Kadia (Wang and Kadia 2001). These experiments were performed in anesthetized marmosets. Single-neuron data regarding feature selectivity were obtaine from Sadagopan and Wang (Sadagopan and Wang 2009). These

recordings were from awake, passively listening marmosets. Single-neuron data regarding feature selectivity in guinea pigs were obtained from adult, head-fixed, passively listening guinea pigs at the University of Pittsburgh. Briefly, a headpost and recording chambers were secured to the skull using dental cement following aseptic procedures. Animals were placed in a double-walled, anechoic, sound attenuated booth. A small craniotomy was performed over auditory cortex. High-impedance tungsten electrodes (3–5MΩ, A-M Systems Inc. or FHC, Inc.) were advanced through the dura into cortex to record neural activity. Stimuli were generated in MATLAB, converted to analog (National Instruments), attenuated, power-amplified (TDT Inc.), and presented from the best location in an azimuthal speaker array (TangBand 4" fullrange driver). Single units were sorted online using a template matching algorithm (Ripple, Inc), and refined offline (MKSort). All analyses were performed using custom MATLAB code.

3.0 Adaptation to Sound Statistics for Noise Invariant Categorization

Accurate processing of behaviorally important sounds such as speech for humans and vocalizations for vocal animals is critical for survival and social interactions. In a previously published model, we had demonstrated that detecting non-redundant spectrotemporal features of intermediate complexity could achieve optimal performance for vocalization categorization. That model was developed and tested in ideal (quiet) listening conditions. In real-world listening conditions, however, this task is often made difficult by the near omnipresence of competing sound sources. Models that are trained and optimized in quiet conditions fail to generalize to such noisy conditions. Physiological observations and results from automatic speech recognition algorithms suggest incorporating adaptation to sound statistics is a possible method for achieving noise invariance. Here, we show that an algorithmic implementation of gain control, a known mechanism for adaptation to sound statistics, improved sound categorization performance in noise. We implemented bottom-up and top-down gain control algorithm, broadly corresponding to subcortical and cortical processes in the auditory pathway. High classification performance could be achieved in noisy environments when top-down gain control, corresponding to contrast gain control found in auditory cortical neurons, was implemented. Our results demonstrate noise invariant categorization of complex sounds can be achieved using biologically plausible mechanisms of adapting to sound statistics.

3.1 Introduction

Recognition of behaviorally important sounds is often performed in the presence of acoustic interference from competing sound sources. The human auditory system, and those of other vocal animals, shows robustness to noise interferences when processing conspecific communication sounds. Our knowledge about the neural mechanisms behind this process, however, remains inadequate. This is exemplified by the performance of speech processors used in auditory prosthetic devices. For instance, cochlear implant users often report great difficulty hearing in noise levels that are trivial for normal hearing individuals, even if they perform comparably to normal hearing individuals in optimal acoustic conditions (Fu et al 2005). Automatic speech recognition algorithms face a similar challenge. High-performing recognition algorithms such as deep neural networks are still susceptible to the effects of noise despite their human-level performance in standard listening conditions. These algorithms, however, can provide a good testing ground for physiologically inspired mechanisms of noise invariant processing. Studies suggest that adaption to sound statistics can be a valid strategy for mitigating the effects of noise. Neurons in the auditory system that exhibit such adaption to stimulus contrast can maintain their response in various acoustic conditions (Dean et al 2005; Watkins and Barbour 2008; Rabinowitz et al 2011). They achieve this adaption by modulating their neural response, i.e. gain control. In conditions of low contrast (noisy environments), neurons increase their gain, thereby expand the dynamic range of their response (Rabinowitz et al 2011, 2013). A large dynamic range increases the neuron's sensitivity to small changes in the stimulus. Since noise elevates overall sound energy and thereby diminishes the level differences between background and signal, an expanded dynamic range can offset this effect. The magnitude of contrast-dependent adaptation strengthens as one

ascends the auditory hierarchy, which coincides with the neural representation of the stimulus being increasing invariant to noise (Rabinowitz et al 2013).

Studies in human and other vocal animals show that noise invariant representation is the result of a transition from a continuous encoding of spectrotemporal properties to a discrete encoding of sound categories (Liberman et al 1967; Chang et al 2010; Ding 2012, 2013). The former method of encoding is typically found in subcortical areas and the auditory periphery, whereas the latter is a staple of cortical sound representation. Given that gain control is found in nearly all stages of auditory processing, it is likely to play an important role in shaping noise invariant representation (Rabinowitz et al 2013). However, it is unclear if the significance of contrast gain control changes in different stimulus encoding schemes. In addition, its neural mechanism of implementation remains to be addressed.

In this study, we aim to investigate these questions by constructing a noise invariant sound categorization model using the principles of contrast gain control. We build on our previous model for sound categorization by implementing contrast gain control algorithms that mimic the observed physiological effects (Liu et al 2019). Briefly, our model uses a template matching algorithm to search for the presence of optimized acoustic features in the stimulus. The optimized features, termed most informative features (MIFs), are selected from an initial set of random features to maximize sound category information and minimize inter-feature redundancy. Classification of the stimulus is based on the number and identity of the MIFs detected by the model. Our information-maximization approach to MIF selection allows the model to achieve high categorization accuracy while accounting for inherent variability between individuals during sound production. We chose this model as the basis for implementing contrast gain control because of the biological relevance of the MIFs. For instance, when used to categorize vocalizations of the

common marmoset (Callithrix jacchus, a highly vocal new world species), the model selected MIFs with predicted tuning properties strikingly similar to those observed in the marmoset primary auditory cortex. Thus, we expect that examining the effects of contrast gain control on MIF detection in noise will yield results with more physiological relevance. Specifically, we implemented a bottom-up and top-down gain control algorithm to our computational model, corresponding to gain control in the subcortical and cortical regions. We interpreted bottom-up gain control as modulation of the input cochleagram (frequency-time representation of the neural activity of auditory nerve fibers in response to acoustic input) based on locally measured sound statistics; and top-down gain control as scaling the MIFs' response to stimulus, simulating contrast-based changes in the gain of putative MIF-detecting neurons.

We show that both forms of gain control increase model performance in noisy conditions. Bottom-up gain control shows a de-noising effect on the input, while top-down gain control increased the range of MIFs' responses, allowing better separation of call type categories. Furthermore, we found that top-down gain control significantly elevates model performance to near physiological levels and may even exceed it in narrowly-defined categorization tasks. Taken together, our results suggest that gain control may indeed be an important neural mechanism for building noise invariant representation.

3.2 Results

3.2.1 Selecting for Robust Features to Increase Performance in Noise

We previously described a feature-based model of sound categorization that accounts for the variability in sound production (See Liu et al 2019 Methods). Briefly, we started with a set of randomly selected acoustic features from a bank of animal calls and evaluated the efficacy of each feature as a classifier of its call type. The classifier functions by thresholding the normalized cross-correlation value between the feature and the input such that values exceeding the threshold signal the presence of a feature. We then selected, using an information-maximization approach, a set of the Most Informative Features (MIFs) for categorizes of that call type. When presented with a novel set of calls obtained from several callers, detecting the set of MIFs is sufficient to achieve very high categorization accuracy with a low false alarm rate in a manner that is robust to production variability. Given that this feature-selection approach can produce a production-invariant representation of call categories, we speculated if these MIFs are also noise invariant. We first tested the categorization performance of MIF sets for the three major marmoset call types (phee, trill, and twitter) using calls masked with Gaussian white noise at various intensity levels. Results are visualized using receiver operating characteristic (ROC) curves in Figure 11A, with lighter shading corresponding to a lower signal-to-noise ratio (SNR). All MIFs were negatively impacted by noise and dropped to chance-level performance (diagonal line) for noise levels greater than -6 dB SNR. Based on these results, we theorized that selecting the MIFs for their performance in the ‘clean’ condition may have resulted in over-optimization and making them susceptible to noise. Given that we have a large sample of acoustic features to draw from, there may exist a set

of MIFs that trade-off accuracy in the ‘clean’ condition for robustness to noise. To test this hypothesis, we selected fifteen additional MIF sets from the pool of 6000 randomly generated acoustic features without replacement (no duplicate features between MIF sets) and tested them in 0 dB SNR noise. The area under the ROC curve (AUC) values in noise were compared against the ‘clean’ condition (Figure 11B). We identified the most robust MIF set for each of the three call types as the one with the smallest difference in AUC between the two conditions, i.e., with the least slope of the SNR-performance relationships (solid black line in Figure 11B). Figure 11C shows the ROC of the most robust MIF sets in the same noise conditions as Figure 11A. Figure 11D shows the amount of increase in AUC (colored area between the curves) using the robust MIF sets compared to the ‘clean’ optimized sets. These results demonstrate that it is possible to select MIF sets that are robust to noise at the cost of a small drop-off in ‘clean’ condition performance. The difference in performance increase between robust and ‘clean’ optimized MIFs for the three call types is likely due to the random nature of the initial acoustic feature generation. All computations from here on will be using the most robust MIF sets in place of the ‘clean’ optimized sets.

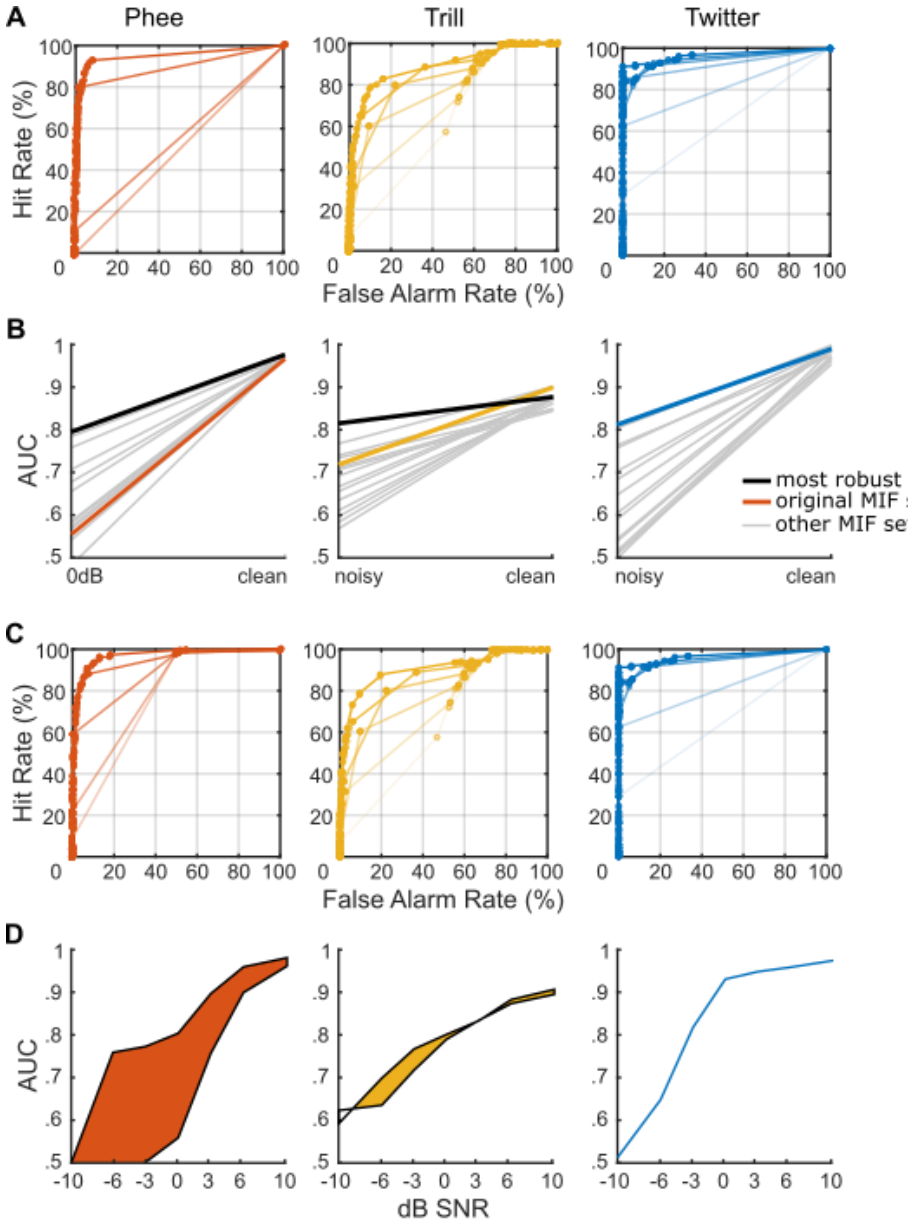


Figure 11 MIF optimization and performance in noise. **A** Receiver operating characteristic (ROC) curves for marmoset phee, trill and twitter calls. The MIFs are tested at 10, 6, 3, 0, -3, -6-, and -10dB SNR. The MIF sets used for these three calls are optimized for categorization in the 'clean' condition. Lighter shading corresponds to lower signal-to-noise ratio (SNR). **B** Performance of alternative MIF sets in both 'clean' and noisy conditions. Colored lines represent the original, 'clean' optimized MIF sets used in **A**. Solid blacklines represent the most robust MIF sets. In the

case of twitter, the original MIF set is also the most robust. **C** ROC curves for the most robust MIF sets tested in the same noise conditions as **A**. **D** Area-under-the-curve (AUC) plots for the original and most robust MIF sets. The colored area corresponds to the values gained by using robust MIF sets for categorization.

3.2.2 Bottom-up Gain Control to 'De-noise' Inputs

To further increase model robustness in noise, we implemented a de-noising algorithm based on the concept of modulating the neural response based on overall and local sound statistics. Because we theorized that putative MIF neurons are likely located in the auditory cortex (see Liu et al 2019 Discussion), this algorithm aimed to simulate the cumulative effect of sound statistics adaptation in the inputs to MIF neurons, i.e., the cumulative effect of gain control in sub-cortical areas. To implement this algorithm, we first measured local sound energy in 2 octave and 200 ms-wide bins of the cochleagram, overlapping by 50% in both frequency and time. We computed the mean activity in each of these blocks, and thresholded the activity in each block by a factor (alpha) that was proportional to the mean activity. All activity below this threshold was set to zero. (see Methods for details). Figure 12A shows an exemplar of a noisy cochleagram (left), the mean value of each frequency-time block (middle), and the de-noised cochleagram after thresholding based on the local mean value (right). To enhance the de-noising process, we optimized the scaling factor alpha to maximize the AUC at each SNR level tested, (Figure 12B). The choice for individually optimizing α is to include the modulatory effect of the overall sound statistics in addition to the local activity mean of μ. Figure 12C shows the effect of optimal thresholding on MIF performance in noise for the three call types. We observed that

thresholding has a small benefit for trill and twitter categorization, but little to no benefit for phee calls. A possible reason for this discrepancy of benefits between call types might be their differences in bandwidth. Twitter calls have the largest bandwidth out of the three call types, meaning that it can span several frequency blocks. In contrast, phee has the smallest bandwidth, meaning the whole call is often contained in one frequency block. When optimizing for thresholds, twitter calls have three α within their bandwidth for adjustments (three degrees of freedom), while phee is often limited to a single α within its bandwidth (one degree of freedom). This discrepancy in degrees of freedom for optimization is likely a major contributing factor to the difference in performance increase between the calls.

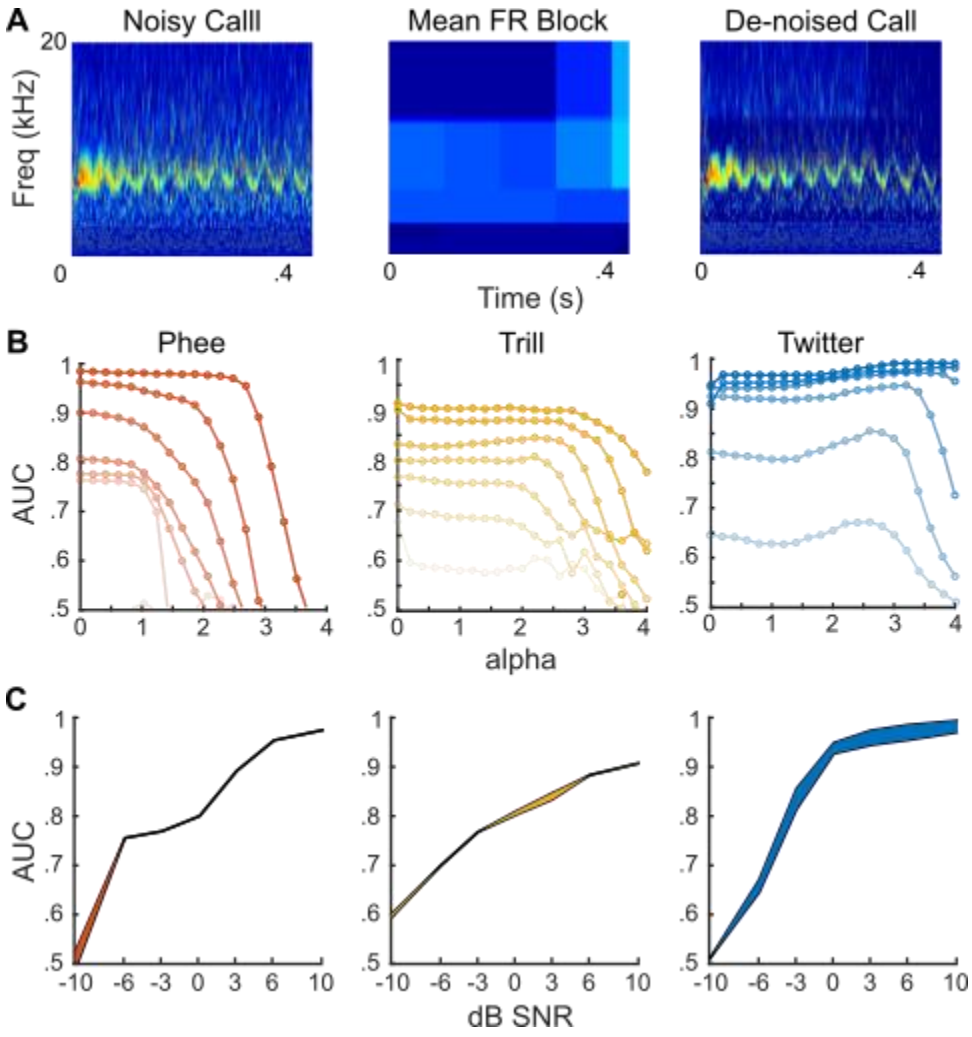


Figure 12 Bottom-up gain control. **A** Example cochleagrams of a noisy trill call (left). The cochleagram is divided into smaller time-frequency blocks (middle), and the mean value of each block is then used as a basis for thresholding the noisy cochleagram. Visual inspection of the resulting cochleagram (right) shows that the trill call is noticeably more visible due to the apparent decrease in background noise level. **B** Optimization curves for the threshold parameter, alpha, for each call type at different noise levels. Lighter shading corresponds to lower SNR. For trill and twitter calls, the optimization curves show local maxima at several SNR levels. While for phee calls, thresholding the noisy cochleagram does not appear to increase AUC (maximum value of optimization curve is 0). **C** AUC plots for categorization with and without cochleagram thresholding. The colored area corresponds to the values gained by thresholding cochleagrams at the optimal alpha value for each SNR level.

3.2.3 Adapting MIFs' Response to Noise

Though selecting for robust features and de-noising the input both improved categorization in noise, these additions were insufficient for the model to reach physiological performance levels in similar conditions. Examining the ROC curves in various noise levels (Figure 11A) showed that for the same false alarm rate, hit rate decreased with increasing noise levels. This suggests that the model's poor performance in noise is the result of its failure to detect features in noise. Figure 13A shows an exemplar of failed feature detection. A feature is considered present in the input if its response (i.e. normalized cross-correlation value between the feature and the input) crosses the MIF's optimal threshold. We observed that the overall shape of the response curve was similar in both 'clean' and noisy conditions, but shifted to sub-threshold levels. The straightforward solution

would be to counteract this shift in response by either proportionally lowering the threshold value (that was previously optimized in clean conditions) or by scaling up the response to achieve supra-threshold detection. Since we theorized that putative MIF detecting neurons are likely located in the auditory cortex, a biologically plausible method of implementing these solutions would be some form of cortical gain control that creates an overall excitatory effect on these MIF neurons. Figure 13B shows a proposed mechanism of cortical gain control that excites pyramidal cells (putative MIF detection neurons) via disinhibition of PV interneurons (reproduced from Willmore et al with permission). We investigated both lowering of threshold (Figure 14) and a multiplicative scaling of the response (Figure 15) approach to noise adaptation.

3.2.4 Adjusting Detection Threshold

MIF threshold values were optimized in the 'clean' condition to maximize hit rate while minimizing false alarms. Noise lowers the MIF response for both within-class and outside-class stimuli. In most cases, the distribution of within-class and outside-class responses remained relatively proportional to each other. Therefore, it might be possible to obtain a similar classification performance from these features by proportionally lowering the threshold. Figure 14A shows this process for an exemplar feature. At each noise level, we varied a parameter (beta) that was multiplied with the threshold previously optimized for clean performance to yield a new threshold. A beta value of 1 corresponded to an unchanged threshold. Figure 14B shows the MIF performance with different thresholds in various noise conditions (lighter shade corresponds to higher noise levels). When we plotted the new threshold value that maximized performance at each SNR as a function of SNR, we observed that it linearly scaled with noise level. Because it is

biologically unrealistic for the system to have knowledge of the stimuli before it is actually recognized, we computed an average beta value across all call types tested, and used this beta value to adjust MIF thresholds at various noise levels. Similar to the subcortical model above, the overall noise level can be computed using neurons with widely-tuned and overlapping receptive fields. Figure 14D shows the increase in performance at various noise levels using the standard threshold ratio. Compared to selecting for robust features and the sub-cortical model (de-noising inputs), the effect of a threshold change was much more profound. For phee and twitter, even at extreme noise levels of -10 dB SNR, the model achieved very high performance levels, even higher than that reported in behavioral experiments (Osmanski et al 2013). Model over-performance is likely a consequence of the limited set of categorization tasks for which our model was trained. Physiological optimization of cortical gain control network requires feedback from higher areas (top-down modulation) regarding the true category of noisy calls. This process is not as efficient or as accurate as the artificial optimization process. Furthermore, the physiological optimization process must account for all behaviorally important sounds, not just the three call types, which is likely to further dilute the benefits of threshold change. Nevertheless, the results suggest that threshold change can play an influential role in noise invariant categorization, both physiologically and computationally.

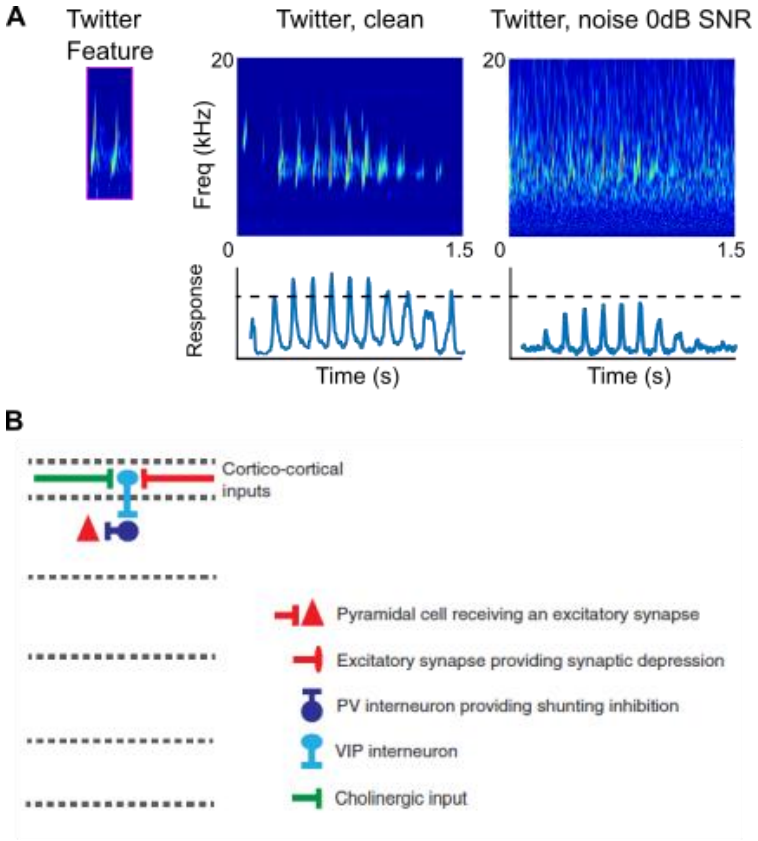


Figure 13 Theoretical computational and neural mechanisms of gain control. **A** Example of a twitter MIF that failed to be detected in noise. Comparing the response curves in for the 'clean' and noisy, we see that there appears to be both a DC shift downward and reduction in the spread. Theoretically, multiplying the MIF response in noise by a constant, which is analogous to the effects of gain control on a putative MIF-detecting neuron, should produce a response curve that is comparable to the clean one. **B** Neural mechanism of contrast gain control as purposed by Wilmore et al 2014 (figure adapted with permission). Cortical-cortical interactions such as attention can recruit vasoactive intestinal polypeptide (VIP)-expressing interneurons, which can excite pyramidal cells via disinhibition of parvalbumin (PV)-expressing interneurons. Putative MIF-detecting neurons are likely to be pyramidal cells located in the auditory cortex, meaning they are the target of cortical contrast gain control via this purposed mechanism.

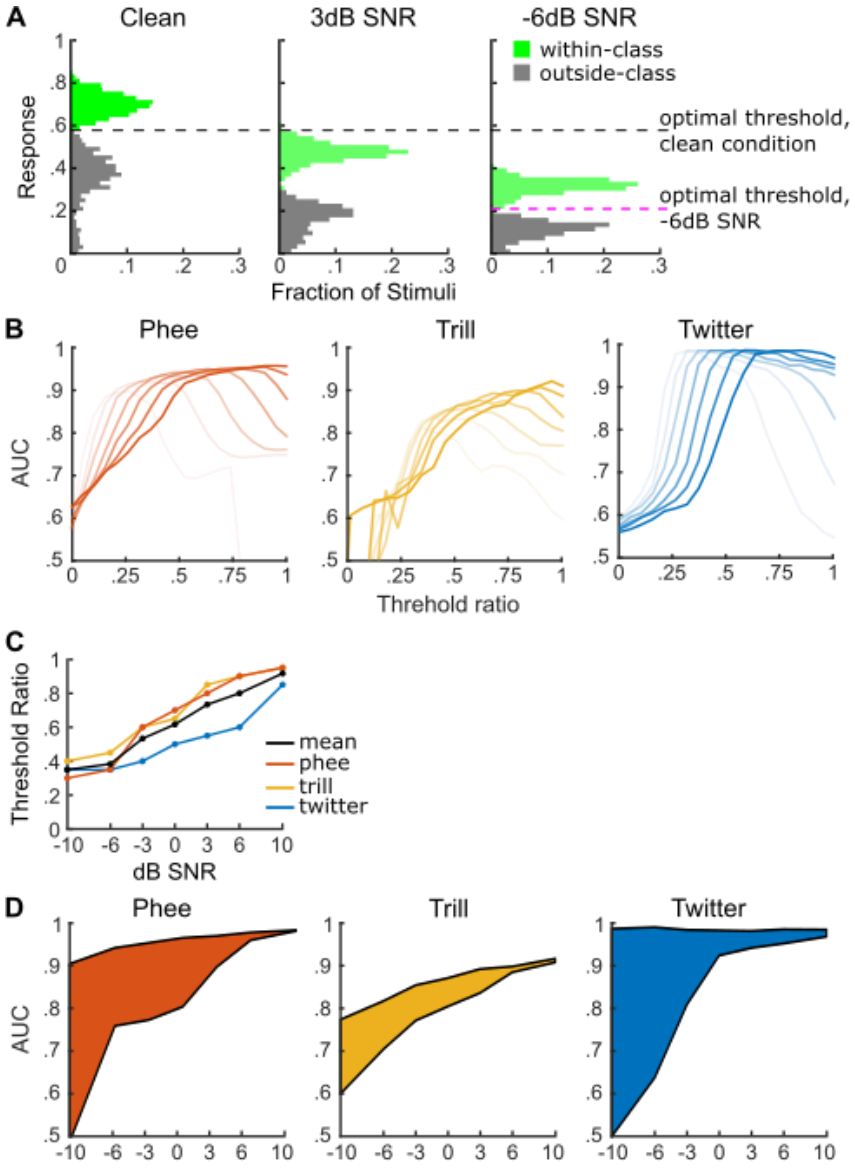


Figure 14. Top-down gain control via threshold change. **A** Schematic of how decreasing the MIF detection threshold can lead to the MIF maintaining its merit as a classifier. MIF detection threshold in the 'clean' condition is optimized to provide maximum separation between the distribution of response for the within-class (same call type as MIF) and outside-class (different call type as MIF) class. As noise level increases, the mean and spread of the response distributions decrease for both class of calls. However, if there is still sufficient separation between these two distributions, the new optimized threshold can allow the MIF to have classification performance

in noisy conditions that is comparable to the clean one. **B** Optimization curves of the ratio between noisy and clean threshold. There is a local maxima for each SNR level and across all three call types. This implies that threshold change is a generally applicable method of improving MIF categorization in noise. **C** Optimal threshold ratio as a function of SNR levels. Curves are color-coded to match the call type. The black line corresponds to the threshold ratio averaged across of all three call types. **D** AUC plots for categorization with and without cochleagram threshold changes. The colored area corresponds to the performance gained. Twitter seemed to benefit the most from threshold changes, which can be attributed to its unique spectrotemporal structure compared to other marmoset call types.

3.2.5 Increase MIFs' Response Gain

A more faithful implementation of cortical contrast gain control is to scale up the MIF response (increased gain) in noisy conditions. This directly stimulates the excitatory effect on the putative MIF detection pyramidal cells. To determine the optimal scaling parameter for each noise condition, we plotted the distribution of MIF responses in noise and the distribution in the 'clean' condition (Figure 15A, exemplar of -3 dB SNR). The diagonal red line indicates the linear mapping of noise response to 'clean' response. The slope of the line is the constant scaling factor β by which the noise response multiple to obtain the corresponding 'clean' response. The result of the linear mapping is shown on the right, with the transformed noise distribution (red) overlaid with the 'clean' distribution (grey). There is a high degree of overlap between the transformed and 'clean' response distributions, which indicates that the gain control MIF is likely to maintain its classification merit in noise. In Figure 15B, we plotted β as a function of SNR for all three call types as well as the average of the three. We observed a negative correlation between β and noise

level (Figure 15B) for all three call types. This is expected because as the noise level increases, so must the response gain in order to offset the effects of noise. Figure 15C shows the increase in AUC (shaded area between the curves) by scaling the MIF responses by β. This implementation of gain control has a significant effect on noise performance and is comparable to those achieved by shifting the threshold.

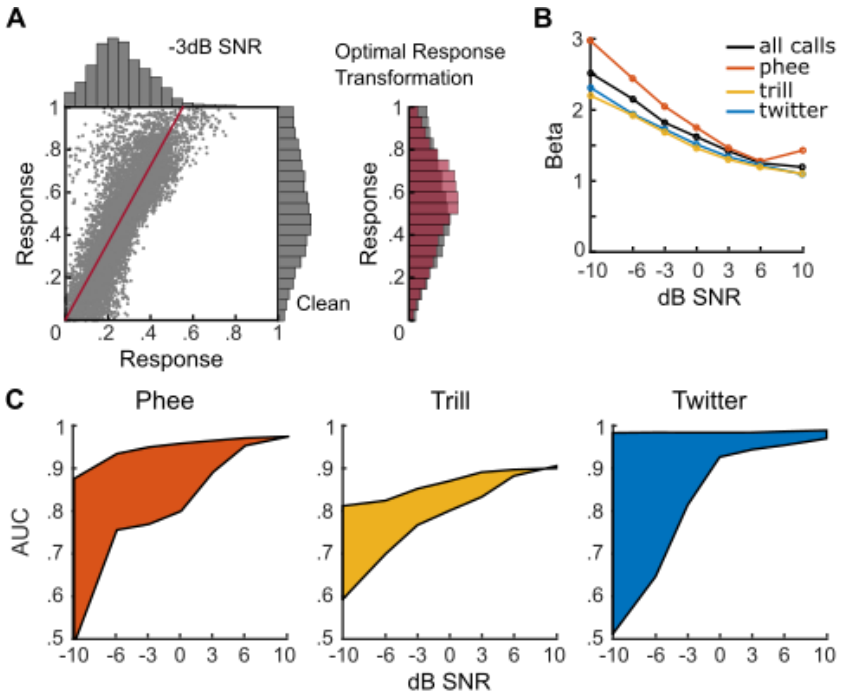


Figure 15. Top-down gain control via MIF response gain. **A** Response distribution for MIFs across all three call types in 3dB noise (along horizontal axis) and clean condition (along vertical axis). The slope of the diagonal red line represents the scaling factor (i.e. response gain) to map the noisy response distribution to the clean one. The transformed noise response (red) is plotted with the clean response (grey), showing good overlap between the two distributions. **B** Response gain (beta) as a function of SNR level. Curves are color-coded to match the call type. The black line corresponds to the threshold ratio averaged across of all three call types. **C** AUC plots for categorization with and without cochleagram threshold changes. The colored area corresponds to

the performance gained. The effects of MIF response gain and threshold change is near identical, which is expected because both methods ultimately make it easier for MIF responses in noise to surpass the detection threshold.

3.3 Discussion

In this paper, we proposed and tested methods to increase the robustness of our sound categorization model in noise. Our overall goal is to test physiological principles of noise invariant representation in a computational setting with the hopes of using the results to provide insight and guidance for future experiments. To that end, we identified adaptation to sound statistic via contrast gain control as a potential strategy for achieving noise-robust categorization (Dean et al 2005; Watkins and Barbour 2008; Rabinowitz et al 2011). Based on the difference in sound encoding schemes in the subcortical and cortical regions, we implemented two separate versions of contrast gain control: bottom-up and top-down, respectively. Bottom-up gain control had the effect of reducing the relative noise level by modulating the input by the local mean activity in a frequency-time block. Top-down gain control increased the model's sensitivity to features obscured by noise, thereby lowering the number of failed detections. Both implementations of gain control increased model performance in noise, especially top-down gain control. The combined effects of these mechanisms are sufficient for the model to achieve physiological performance levels in noise, and even overperforming in well-defined categorization tasks (Osmanski et al 2013). Our results suggest, from a computational perceptive, the efficacy of these neural mechanisms in constructing noise invariant representation of sound in the brain.

Gain control allowed the model to maintain satisfactory performance in a variety of noise conditions with just one set of MIFs per call type. This is consistent with the principle of efficient encoding in the brain (Smith and Lewiski 2006; Holmstrom et al 2010). In contrast, automatic speech recognition algorithms can afford a much liberal use of computational power thanks to technical advancements in recent times. These algorithms often use multi-condition training which selects condition-specific features to deal with variations in the acoustic environment. The equivalent method for our model would be to train a new set of MIFs for each of the noise conditions we tested. While this approach will undoubtedly allow the model to achieve its maximum performance in noise, it significantly increases the number of features and parameters needed. If we consider the vast diversity of noise types and levels in real-world listening environments, the size of the feature bank required is likely magnitudes larger than our current case. Because of this, it is highly unlikely for the auditory system to implement this strategy on a consistent basis for adapting to noisy environments. Noise-specific features may still be encoded, however, if there is prolonged exposure to that condition and/or a significant physiological need for categorization under that condition. Hypothetically, animals over-trained in a reward-driven discrimination task in noise may begin to encode for noises-specific features, provided that their existing mechanisms for noise invariant representation are insufficient for this task. Still, we believe mechanisms like contrast gain control is preferred by the brain for their efficiency because it stretches the scope of the conditions that one feature set can be used for.

We were surprised to discover how well detection threshold change/MIF response gain control worked for categorization in noise. In comparison, the effect of de-noising input appeared lackluster even though both methods are derived from the same physiological observation of sound

statistic adaptation. This discrepancy can be explained by the difference in the target of the modulation. In bottom-up gain control, modifications were made on the input, while the top-down gain control acted directly on the MIFs. From a computational point of view, top-down gain control has an interjection point closer to the output of the categorization model, meaning its impact on the model is less diluted compared to bottom-up gain control. Another way of interpreting this is that bottom-up gain control is dependent on the mean activity level in a local frequency-time block for modulating input, whereas top-down gain control can directly control the threshold/response of the MIFs. The latter has a much more direct effect on whether the MIF will pass its detection threshold. We believe this observation is not just the result of model architecture, but also has potential for broader physiological implications. In the auditory pathway, earlier areas in the pathway typically emphasize continuous encoding of the spectrotemporal information of sound while cortical and higher areas shift to discrete categorical representations (Liberman et al 1967; Chang et al 2010; Ding and Simon 2013). As a result, vocalization/speech feature selective neurons are predominantly located in cortical or higher areas. Therefore, cortical gain control mechanisms can directly modulate the activity of these feature selective neurons and have a greater impact on noise invariance. This is consistent with observations that the effects of gain control and noise invariance strengthen as one ascends the auditory hierarchy (Rabinowitz et al 2013)

Furthermore, we selected parameters in both bottom-up and top-down gain control for maximum model performance in noise. This implies that the objective of gain control in both cortical and subcortical areas is to construct noise invariant representation. While this statement is likely true for cortical gain control, it may not be as valid for subcortical regions. Faithful representation in the early stages of auditory processing may be important for accurate and complete encoding of novel stimuli. In addition, there is a significantly smaller number of

vocalization/speech selective neurons in subcortical areas compared to cortical ones. Therefore, it is possible that gain control in subcortical areas, while helpful for categorization in noise, may be optimized for other purposes. Taken together, this implies that cortical gain control mechanisms are a greater contributor to noise invariant representation than subcortical ones.

We developed this computational model for testing feasibility of neural mechanisms; therefore, we purposely designed the model to operate on broader computational principles with minimal mechanistic assumptions. This is to ensure the model is inclusive to maximum number of possibilities and retain its modular nature for future addition/subtraction based on developing physiological evidence. In other words, our current implementation of gain control is largely agnostic to the specific neural mechanisms. Nevertheless, we will discuss how known mechanisms of gain control fit within our proposed computational framework.

Since the optimization step of our models requires access to overall incoming sound statistics, this suggests that gain control for the purpose of invariant representation of sound are likely the results of inter-neuron/network mechanisms. In auditory cortex, and other sensory systems in general, cortical responses are often shaped by the co-occurrence of synaptic excitation and inhibition [Anderson et al., 2000; Poo & Isaacson, 2009; Wehr & Zador, 2003; Wilent & Contreras, 2004; Zhang et al., 2003]. The ratio between excitation and inhibition is dynamic, but tightly regulated (E/I balance). GABA mediated inhibition increases membrane conductance, and per Ohm's Law, has a divisive effect on membrane potential. If synaptic excitation and inhibition are large in magnitude compared to other factors of membrane conductance, then changes to the E/I ratio can significantly alter membrane potential towards or away from firing threshold [Higley & Contreras, 2006]. Chance et al 2002 also support the gain control capabilities of E/I balance by showing divisive gain modulation of neuronal response by introducing a barrage of excitatory and

inhibitory synaptic conductances. Further studies suggest that overall level of synaptic input appears to be a control signal that modulates the gain of neuronal responses [Chance & Abbott 2005]. These results are consistent with our implementation of gain control as a mechanism for noise-invariant representation. For a putative MIF neuron in auditory cortex (Liu et al 2019), the normalized cross-correlation with sound stimulus corresponds to the driving input of the neuron. The modulatory inputs to the putative MIF neuron arise from other cortical neurons that gauges the overall background activity and exert their affect via intra-cortical connections.

While gain control can be mediated by network mechanisms, within-neuron mechanism such as synaptic depression can also drive gain changes. Fluctuations in rapid firing afferents can mask meaningful changes in slow firing afferents. Neurons can selectivity reduced the gain of rapid and sustained firing afferents via short-term depression [Abbott et al 1997; Rothman et al 2009]. The difference in relative gain between fast and slow afferents serves to emphasize small rate changes, thereby increasing the neuron's sensitivity in low contrast conditions [Abbott et al 1997]. Since synaptic depression is a self-contained mechanism, it can explain neural adaptation to sound statistics at the local level (i.e. contained within the neuron's STRF). However, the time course of gain control observed *in vivo* potentially argue against synaptic depression as the mechanism that drives contrast-invariance. Synaptic depression operates in the milliseconds scale while cortical gain control was observed to be around hundreds of milliseconds [Rabinowitz et al 2012]. Alternatively, the time course can be explained the result of integrating multiple stimuli to accurately gauge the acoustic background.

It is also entirely possible that different gain control mechanisms operate at different stages of the auditory system. In the visual system, both the retina and V1 have been observed to utilize separate gain control mechanisms (Carandini et al 1997; Brown & Masland 2001). These

mechanisms may have different spectral integration range (local or global) and time constants. Therefore, it may be accurate to attribute sound statistic adaptation to the cumulative effect of gain control in general rather than any specific mechanism. Furthermore, gain control in higher processing stages can also be inherited from earlier stages. Thus, it can be difficult to experimentally discern the effects of sub-cortical and cortical gain control mechanisms.

Finally, the mechanisms discussed here are not an exhaustive list of achieving noise invariant representation of sound. Attention has been shown to play a major role in noise discrimination tasks. Studies in ferrets shown neurons changing their spectrotemporal receptive fields (STRFs) in a task-dependent manner to enhance the contrast between background noise and the target signal (Fritz et al. 2007; Atiani et al. 2009; Yin et al. 2014). Spatial cues and contextual information can also be utilized in situations of competing sound sources to parse different streams of sound, such as the case of the cocktail party effect. This classic scenario raises a point about what separates signals from noise. In our model training, we clearly assigned parts of the input as either signal or noise, and they can be well differentiated based on sound statistics. However, in scenarios with multiple talkers such as a cocktail party, the differentiation may not be as well-defined. For one, the noise may be competing vocalizations/speech that shares similar sound statistics as the signal. Since the premise of gain control is that the signal and noise have measurable differences in sound statistics, its effectiveness may be diminished in dealing with overlapping vocalization/speech.

In summary, we showed that adaptation to sound statistics via gain control is an effective approach for improving sound categorization in noise. Computational implementation of gain control has yielded robust performance comparable to physiological levels. Given the presence of

gain control in all stages of the auditory system, it likely plays an important role in building noise invariant representations in the brain.

3.4 Methods

3.4.1 Vocalizations and Noisy Stimulus

All procedures conformed to the NIH Guide for Care and Use of Laboratory Animals. All marmoset procedures were approved by the Institutional Animal Care and Use Committee (IACUC) of The Johns Hopkins University. All guinea pig procedures were approved by the IACUC of the University of Pittsburgh. We used vocalization recordings from 8 adult marmosets, both male and female, for these experiments. Marmoset calls were recorded from a marmoset colony at The Johns Hopkins University using directional microphones (Agamaite et al 2015). Guinea pig calls were recorded from 3 male and 3 female adult guinea pigs. Two or more guinea pigs with varied social relationships were placed on either side of a transparent divider in a sound attenuated booth. Directional microphones, suspended above the guinea pigs were used to record calls. Calls were recorded using Sound Analysis Pro 2011(Tchernichovski et al 2000), digitized at a sampling rate of 48 KHz, low-pass filtered at 24 KHz, manually segmented using Audacity, and classified into different call types. Gaussian white noise was artificially added to the vocalizations using MATLAB software. The parameters of the noise were set to achieve a target signal-to-noise ratio (SNR). We computed the exact SNR value for each of the noisy calls to ensure that none deviated more than ten percent from the target SNR value.

3.4.2 MIF Generation and Selection

See Liu et al 2019 Methods for MIF model details

3.4.3 De-noising the Cochleagram

We divided input cochleagrams into smaller frequency-time blocks that are 200 ms in duration and spans 2 octaves in bandwidth. Each block has a fifty-percent overlap in both time and frequency with the adjacent blocks. This is to ensure adequate coverage of the entire spectrotemporal field of the input. The mean activity, μ of each block is then calculated. Note that since we converted the input from acoustic waveform into cochleagrams via the Auditory Nerve Model (Zilany et al 2014), the values of the cochleagram represents neural firing rate, FR. The mean value of each frequency-time block can therefore be thought of as the average firing rate of a bundle of auditory nerve fibers over a period of time. We then scale μ by a constant α to determine the threshold T for each frequency-time block. Subthreshold values in the block are set to 0). Note that all frequency-time blocks share the same α value, but since μ likely differs from block to block, the threshold will also differ. We search for the best α value for maximum model performance, measured as area-under-the-curve (AUC), using a hill climbing optimization method. Starting with α value very close to 0 (i.e. no change to cochleagram), we incrementally increased α and see if a better AUC value can be reached. To be sure there are no other local maximas, we kept testing larger α values past the optimal value until a steady downward trend is observed or the AUC value falls below 0.5 (near chance rate).

3.4.4 Threshold Optimization

The new MIF threshold value for detection in noise was optimized as a ratio of the threshold value in the 'clean' condition (for derivation of the threshold value in the 'clean' condition, see Methods Liu et al 2019). Since we know the ratio value is bounded between 0 and 1, we used brute-force optimization and systematically tested ratio values within that range to find the one that gave the best AUC. Classification in noise testing was performed with the same set of MIFs, with the only difference being the thresholds were adjusted. Please note that this threshold ratio is applied to all MIFs. We did not individually optimize the threshold in noise for each MIF.

3.4.5 MIF Response Gain

To find optimal gain value, we started with an estimate of the optimal value as the ratio of the mean value of the two distributions. We then scale the MIF distributions in noise by the gain and obtained a transformed distribution. To compare the degree of overlap between the transformed distribution and the 'clean' distribution, we computed the Bhattacharyya distance between the two distributions. We again used the hill climbing optimization methods (testing values that incremental differs from a starting value, look for increase in performance) to find the optimal gain value that gave us the most overlap between the transformed and 'clean' distributions.

4.0 General Discussion

4.1 Summary of Findings

The goal of this thesis was to show physiologically feasible methods of addressing production and environmental variability in sound categorization. The auditory system was seemingly faced with conflicting requirements of fine-tuning for accurate sound categorization and tolerance for variations in sound production and environmental interreferences. We approached this issue from a theoretical point-of-view by using computational principles and modeling to show the efficacy of potential solutions. To preserve physiological relevancy, we attempted to ground these solutions in experimentally observed truth. We aimed to use these results as a basis for future experiments to valuate these proposed methods.

In chapter 2, we showed that features of intermediate complexity can accurately categorize sound while account for production variability. We developed a computational model that can select for these features based on an information-maximization approach. The set of most informative features (termed MIFs) can achieve high accuracy while remaining robust to production variability. Critically, the predicted tuning properties of putative MIF-selective neurons closely matched previous recordings from marmoset A1. The implications of our results are that the brain may be encoding for task-relevant features that served specific behavioral purposes. We also demonstrated the general computational applications of our model for vocalizations of species other than the marmoset and other auditory related tasks such as caller identification.

In chapter 3, we examined the computational efficacy of adapting to input sound statistics as a method of improving model categorization performance in noise. Specifically, we identified gain control as the neural mechanism to achieve this adaptation. We subsequently added gain control inspired algorithms to our existing computational model. The implementation was two-fold: a bottom-up method of gain control, simulating the effects of auditory periphery and subcortical areas; as well as a top-down method of gain control, simulating the effect in cortical and higher areas. Our results showed that both methods of gain control, along with selecting for noise robust features, help improved model performance in noise to near physiological levels. Taken together with results from chapter 2, we have built a computational model that can accurately categorize calls while remaining robust to production and environment variabilities.

4.2 Future Directions

Our future directions with the MIF model are in two broad directions: 1. Physiological experiments to verify the findings and predictions of the model; 2. Algorithmic applications of the model as a call classifier. In these future experiments, we will use guinea pigs (*Cavia porcellus*, GP) as our animal model instead of marmosets. There are several reasons for this switch: 1. GPs are a well-established animal model for studying the auditory system, with rich repertoire of vocalizations and well-defined auditory neural anatomy 2. Replacement of primates with a "lower" species is in accordance with the animal research guidelines for invasive experiments. 3. GP calls share large frequency overlaps with human speech, whereas majority of marmoset calls are in the kHz range.

4.2.1 Testing Model Predictions with Behavioral and Electrophysiological Experiments

Proper selection of informative acoustic features is the basis of our computational model. While we have some initiation about why certain features are important/characteristic to its call category, such as repeating frequency-modulated (FM) sweeps to twitter calls, we have yet to validate whether the auditory system assign the same level of importance to these MIFs. For speech, linguists have identified phonetic features that act as the essential building blocks of spoken language. Studies such as Mesgarani et al 2014 have identified neurons in the human superior temporal gyrus encoding for these phonetic features. The combination of experimental observations and theoretical framework of speech plotted a conceivable path for speech representation in the brain. And this is what we aim to ultimately achieve with our MIF model. To validate the perceptual importance of MIFs, we can set up discrimination tasks for GPs that require them to identify calls with certain MIFs removed (via filtering, noise-masking, or other modifications). The perceptual importance of various MIFs can be analyzed form psychoacoustic curves. Our lab has previously found success with GP discrimination tasks in noise (Montes-Lourido et al 2021) using pupillometry as a metric for perceptual detection. It is a reliable, non-invasive indicator and largely circumvents the need for extensive behavioral training of GPs. We hope to continue these experiments in the future.

Aside from behavioral studies, we can also conduct electrophysiological recordings in GP auditory cortex to find potential evidence for neural encoding of MIF-like features. As we have demonstrated in chapter 2, the predicted tuning properties of putative MIF detection neurons closely reassembles call selective neurons found in marmoset A1. Though these examples only qualify as anecdotal evidence, we can make systematic predictions of tuning properties and draw

meaningful conclusions with a large recording sample of call selective neurons. Our lab has conducted electrophysiological recordings from A1 neurons in awake, passive listening GPs. We characterized their tuning properties and their call selectivity in 'clean' and 'noisy' conditions. We can predict call selectivity of neurons using the MIF model based on the observed tuning properties of the neuron and test these computational predictions against the actual call selectivity of the neurons. Our goal with these experiments is to explain why certain features are encoded and what purpose they serve in the grant scheme of auditory processing. These experiments can also be extrapolated to other auditory regions both up and downstream of A1. The results from these areas can help us to map-out a more complete picture of the features encoded at each processing stage and how these features combine and evolve from one stage to another.

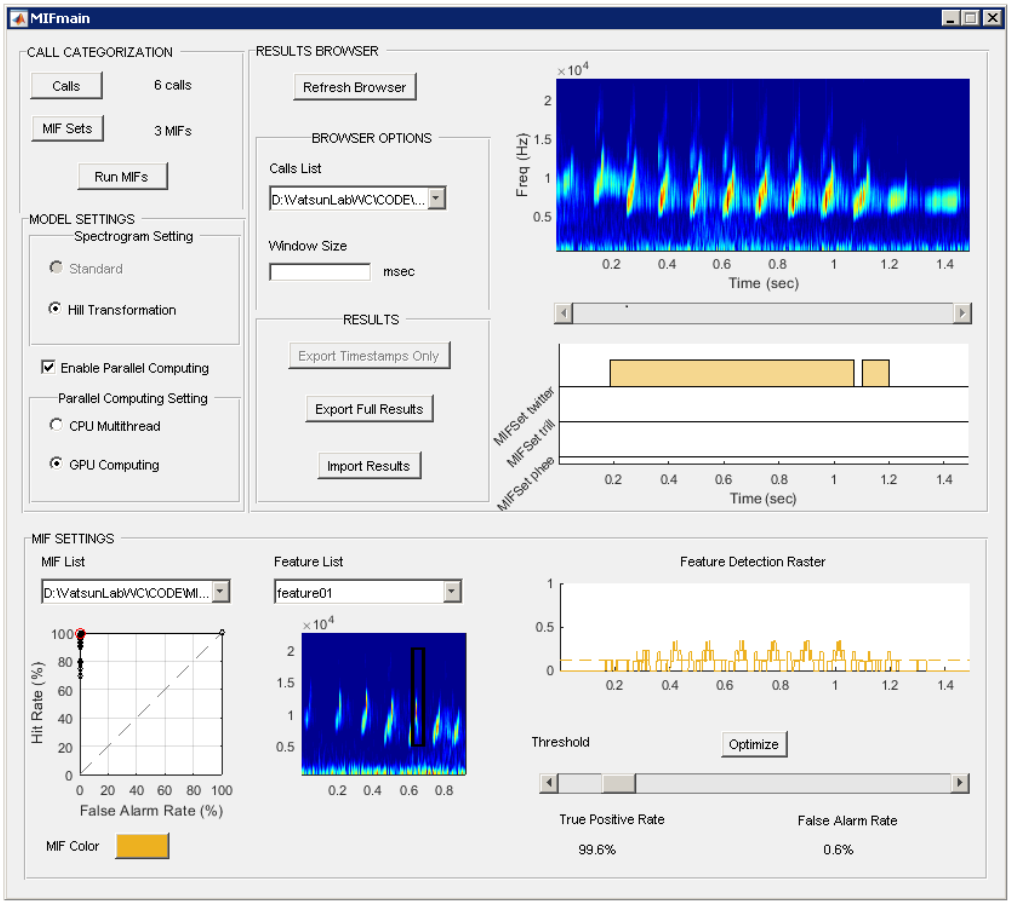
4.2.2 Computational Applications of MIF Model

Building a bank of vocalizations is a critical first step for many auditory experiments. However, it can be a lengthy and tedious process to manually extract and classify calls from recordings. Therefore, we have adapted our model to perform automatic call categorization. The model functions as a graphic-user interface (GUI) in MATLAB (Figure 16)

The model takes sound recording files as input, then attempts to identify calls by searching for the presence of corresponding MIFs. Once identification is complete, the model will output the time stamps associated with each call and allow the user to select method of audio extraction. We also included the ability for users to trained custom MIF sets based on their own vocalization recordings.

This is an ongoing project, with improvement planned based on user feedback. We are currently investigating optimal methods for identifying and extracting overlapping calls. This was a frequently requested issue by people who expressed interest in using our GUI and we are looking to address in the near future. The model in its current state can detect some, but not all overlapping features. We aim to resolve this using smart algorithms to identify potential overlapping calls and adjust necessary model parameters to improve detection. Another feature we can implement is caller identification. We showed in Figure 10A that by changing the classification task, the model can select for MIFs that characterizes speaker identity, albeit at a slightly lower accuracy rate. The challenge with implementing caller identification concurrently with call type classification is that the former is call type invariant (i.e. females typically vocalize at a higher pitch than males regardless of the call made) while the latter should be caller invariant (i.e. both female and male twitter calls should be classified as the same call type). We are exploring methods of integrating the MIF sets for these two tasks together during classification. Our goal is to have the model able to identify both the call type and the caller concurrently.

a


MIFmain
CALL CATEGORIZATION
Calls
6 calls
MIF Sets
3 MIFs
Run MIFs
MODEL SETTINGS
Spectrogram Setting
Standard
Hill Transformation
Enable Parallel Computing
Parallel Computing Setting
CPU Multithread
GPU Computing
RESULTS BROWSER
Refresh Browser
BROWSER OPTIONS
Calls List
Window Size
msec
RESULTS
Export Timestamps Only
Export Full Results
Import Results
Freq (Hz)
Time (sec)
MIFSet twitter
MIFSet trill
MIFSet phee
MIF SETTINGS
MIF List
Feature List
feature01
Hit Rate (%)
False Alarm Rate (%)
MIF Color
Feature Detection Raster
Threshold
Optimize
True Positive Rate
99.6%
False Alarm Rate
0.6%

b

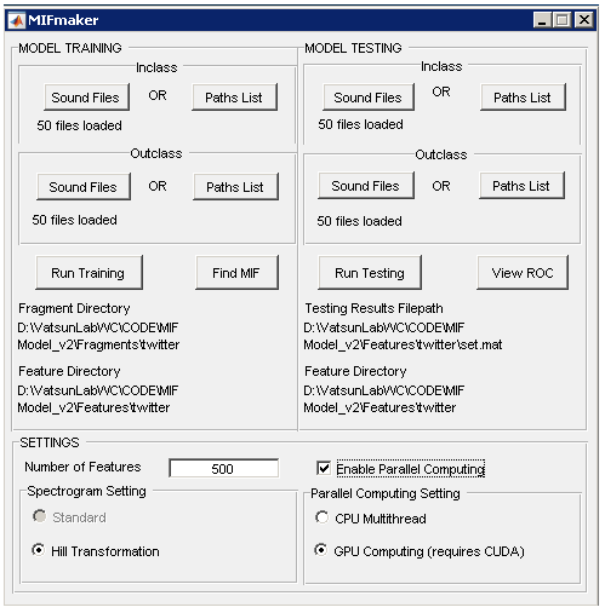

MIFmaker
MODEL TRAINING
Inclass
Sound Files
OR
Paths List
50 files loaded
Outclass
Sound Files
OR
Paths List
50 files loaded
Run Training
Find MIF
Fragment Directory
D:\VatsunLab\VC\CODE\MIF
Model_v2\Fragments\twitter
Feature Directory
D:\VatsunLab\VC\CODE\MIF
Model_v2\Features\twitter
MODEL TESTING
Inclass
Sound Files
OR
Paths List
50 files loaded
Outclass
Sound Files
OR
Paths List
50 files loaded
Run Testing
View ROC
Testing Results Filepath
D:\VatsunLab\VC\CODE\MIF
Model_v2\Features\twitter\set.mat
Feature Directory
D:\VatsunLab\VC\CODE\MIF
Model_v2\Features\twitter
SETTINGS
Number of Features
500
Enable Parallel Computing
Spectrogram Setting
Standard
Hill Transformation
Parallel Computing Setting
CPU Multithread
GPU Computing (requires CUDA)

Figure 16. MIF Model GUI **A**. Example of the GUI categorizing a twitter call. To shorten the computation time, we implemented a spectrogram transform approximation of the auditory nerve model (Rahnman et al 2020). The spectrogram of the input is displayed on the upper right. Below it is the detection rectangle plot showing the time period in which calls are detected (color codes correspond to call types, yellow is twitter). The other two rectangle plots, corresponding to trill and phee, are flat, indicating that these features were not detected. The ROC curve for the selected MIF set is displayed on the lower left. The red circle on the ROC curve corresponds to the estimated hit rate vs false alarm ratio for the current threshold. The threshold can be adjusted using the slider on the lower right based on preference (e.g., trade-off higher false alarm rate for more detection). The feature detection raster shows the MIF response curve overlaid with the current threshold. This is a helpful tool for visualizing where evidence for MIF presence is strongest within the call. **B**. GUI for generating new MIF sets. The user can upload sound files for training and testing the model, as well as selecting parameters associated with MIF generation and testing. Both GUIs have support for parallel computing to expedite the process.

Appendix A

Appendix A.1 Discussions

A.1.1 MIF-based Reconstruction of Call Stimuli

The observation that an MIF-based approach successfully generalizes across production variability implies that most calls belonging to a category will contain one or more of the MIFs. Therefore, we asked how well calls could be reconstructed based on MIFs alone, using twitters as a specific example. To do so, we detected model twitter MIF neuron spiking as described in the main text to the 500 training and 500 test twitters, and convolved these spike times with an alpha function (with a time constant of 20 ms) to detect the peak locations of twitter MIFs within a twitter (Appendix Figure 5A). We then placed copies of MIF cochleagrams at these peak locations, or added copies of MIF cochleagrams to previously placed feature cochleagrams. The final summed cochleagram was taken to be the reconstructed call (Appendix Figure 5B). We evaluated the accuracy of reconstruction as the NCC value at zero lag. The mean reconstruction accuracy was 0.69 (Appendix Figure 5C), suggesting that MIFs were indeed common denominators across twitter calls produced by different animals.

A.1.2 Factors Contributing to the Success of the MIF-based Approach

Three factors were critical in the design and implementation of our approach. First, focusing on a behaviorally critical task (call categorization), and choosing model species with rich

vocal repertoires and behaviors (marmosets and guinea pigs) allowed us to clearly identify a computational goal of cortical processing – call categorization. Previous experiments using both electrophysiological and imaging techniques (Rasuchecker and Tian 2000; Tuab et ak 2001; Romanski and Averbeck 2009; Grimsley et al 2012; Fukushima et al 2014; Petkov et al 2008; Perrodin et al 2011l Sadagopan et al 2015), showing an increase in cortical resources allocated to call processing, validate our choice of call categorization as a critical computational goal in vocal animals. Second, our analyses were based on a large sample of calls recorded from a large number of animals. From this data set, we deliberately oversampled a large number of initial potential features. This ensured that the full extent of production variability was represented in this data set. Third, the greedy search algorithm efficiently identified informative features from a training data set of a few hundred calls. Since clean and labelled training data sets are laborious to generate, the efficiency of greedy search provided a significant methodological advantage.

A.1.3 Limitations of Greedy Search and MIF-based Classification

In this study, we used greedy search and pairwise maximization of information to find optimal features. However, it is possible that the greedy search algorithm does not find an optimal solution because of its inability to overcome local maxima. We do not think this is the case because: 1) the model performs at high accuracy levels, leaving little room for significant improvements, 2) we could arrive at similar sets of MIFs and achieve similar performance levels from different initial feature sets, specifically when highly informative features were excluded (Supp. Figure 3), and 3) we could match or outperform other machine learning based algorithms for marmoset call classification (Turesson et al 2016). Therefore, the implemented greedy search algorithm likely

converges at a true optimal solution. Our MIF-based approach has two limitations. First, the number of auditory tasks that an animal is potentially required to solve is ill-defined. While we mitigate this limitation by choosing ethologically critical tasks such as call categorization, it is likely that we are only probing a small subset of all behaviorally relevant auditory tasks. Consequently, while a subset of neurons in auditory cortex match predictions from our model for call and caller classification, developing a larger bank of natural auditory behavior (for example, predator sounds versus neutral sounds) will allow us to model and predict a larger fraction of cortical responses. Second, our model derives features from the auditory nerve representation of stimuli. It is well-known that this representation is transformed more than once before impinging on cortical neurons. Therefore, the actual representation from which cortical neurons detect features are not accurately modeled here. This limitation arises from the current lack of predictive models for central auditory processing stages. It is possible that the performance of our algorithm will increase if we could accurately model other sub-cortical processing stages.

A.1.4 Alternative Models

Recently, theoretical efforts have been directed at learning invariant representations from small training sets using unsupervised methods (Anselmi et al 2016). In this model, image 'signatures' which serve as a proxy for the probability distribution of an image and its transformations are learnt by leveraging the time correlations of image transformations in the real world to label image identity. Image signatures can be computed by complex cell-like units using Hebbian learning rules. This model predicts that a similar computation might occur in auditory cortex. The MIFs that we have derived for call categorization are similar to the image signatures

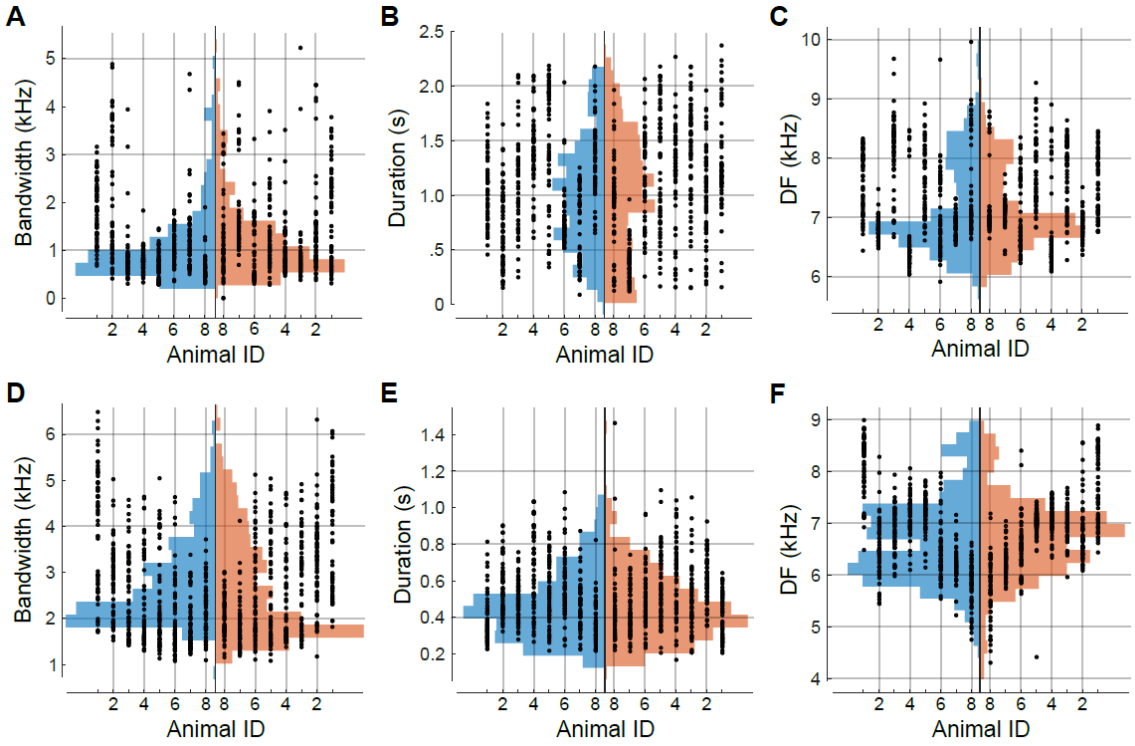
in that they serve as a proxy for the probability distribution of a sound category that has been subjected to production variability. Indeed, vocalizations can be viewed as multivariate probability distributions along multiple call parameters, and MIFs could serve as the 'gist' of a call category around which these variations occur. Similar to image signatures, MIFs seem to be computed by superficial layer auditory cortex neurons. However, differences arise in how MIFs are learnt. Although small sample sizes are adequate, unlike image signatures that are learnt by observing image transformations over time, explicit labeling of the class of input examples is necessary for learning the MIFs of calls. Conceptually, whereas image signatures are learnt by observing within-category transformations, MIFs are learnt by contrasting the distributions of sound categories.

A.1.5 Alternative Experimental Approaches

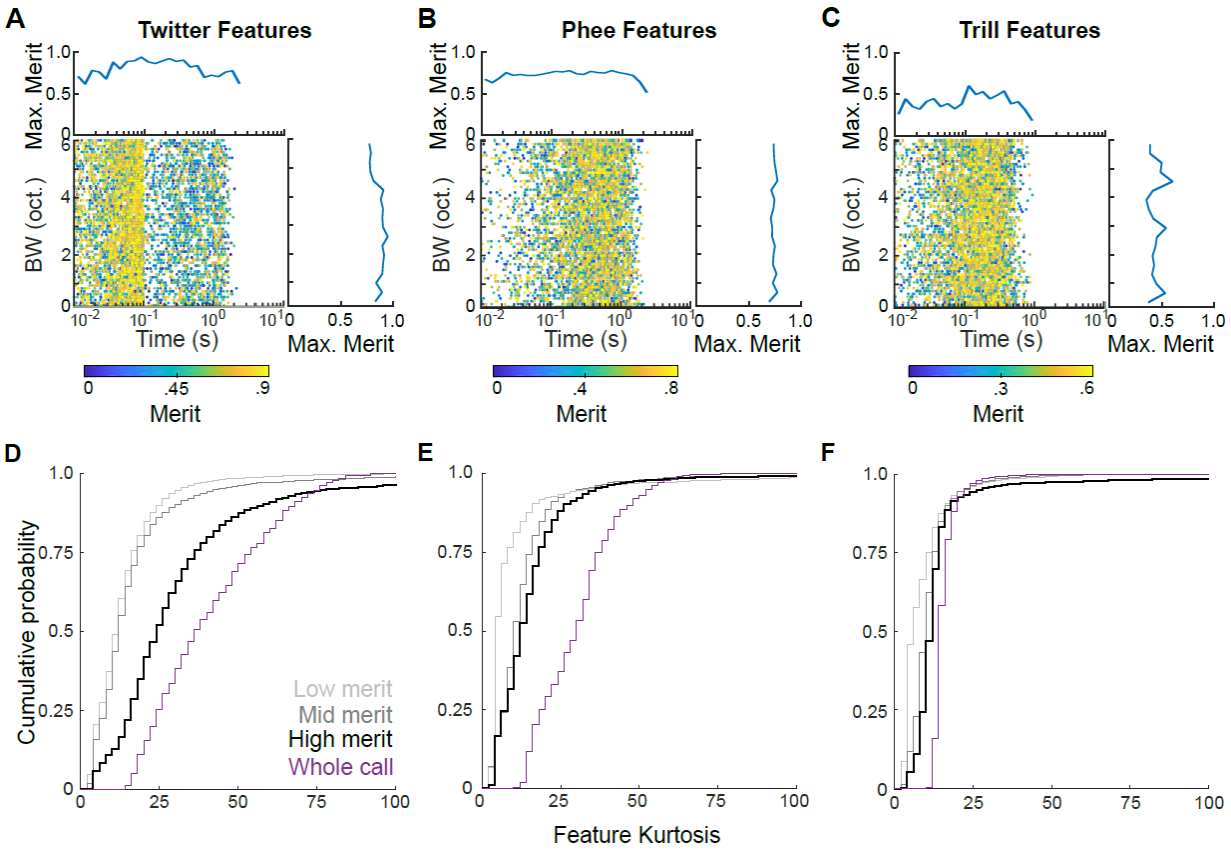
Previous experimental studies have described call selectivity primarily using two methods: 1) characterization of neural tuning along an exhaustive list of call parameters (DiMattina and Wang 2006), and 2) characterizing call tuning as tuning for regions of the modulation spectrum (Hsu et al 2004; Woolley et al 2005; Stowell et al 2014). In the former study, marmoset calls were parametrized along multiple acoustic dimensions. Some of these parameters were common to all call types, such as the length or dominant frequency of a call. The more distinguishing parameters, however, were unique to individual call types, such as the inter-phrase interval for twitters, or sinusoidal frequency modulation rate for trills. Neural tuning to calls was described using tuning to these parameters but did not use the same set of parameters across call types. In our study, different MIFs are used for classification of different call types, but MIFs are parametrized along the same axes – bandwidth and integration window, allowing for a uniform basis for comparisons.

In the latter set of studies, neural tuning for birdsong was described using selectivity for specific frequency and temporal modulations. In this case, tuning could be expressed in a unified stimulus space (of spectral- and temporal modulation rates). Both these methods, however, serve to describe neural tuning, and not to explain why tuning to certain parameters or regions of modulation space are necessary in the first place. Our results suggest that generating selectivity for task relevant features explains why selectivity for stimulus parameters arises in the first place. In a recent study, a combination of the above approaches was used in conjunction with statistical classifier techniques to achieve caller identification for macaque coo calls (Fukushima et al 2015). Caller identification could not be achieved using a single feature alone, where feature referred to a parameter such as fundamental frequency, duration, or location in the modulation spectrum. Rather, a combination of cues was required for high caller identification performance. Our study differs from this study in that our definition of 'feature' is non-parametric, our goal is to generalize over individual identity, and features are contrastive and task-dependent. But similar to this study, a single feature alone was insufficient for call categorization in our study as well.

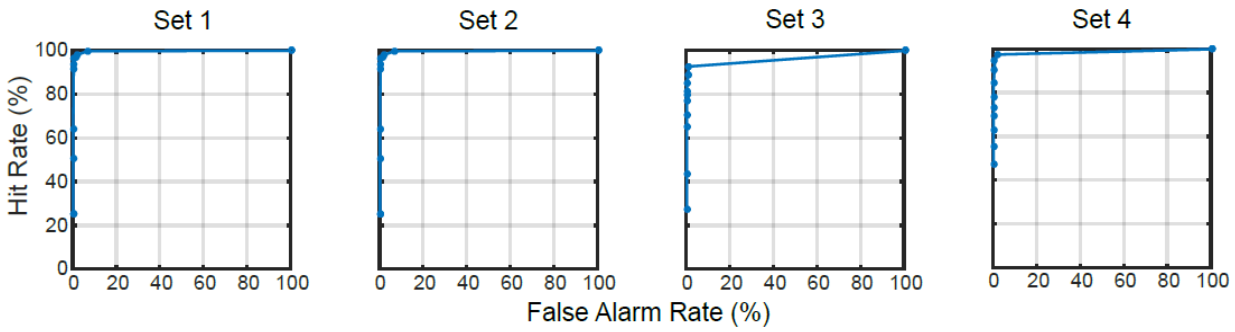
Appendix A.2 Figures



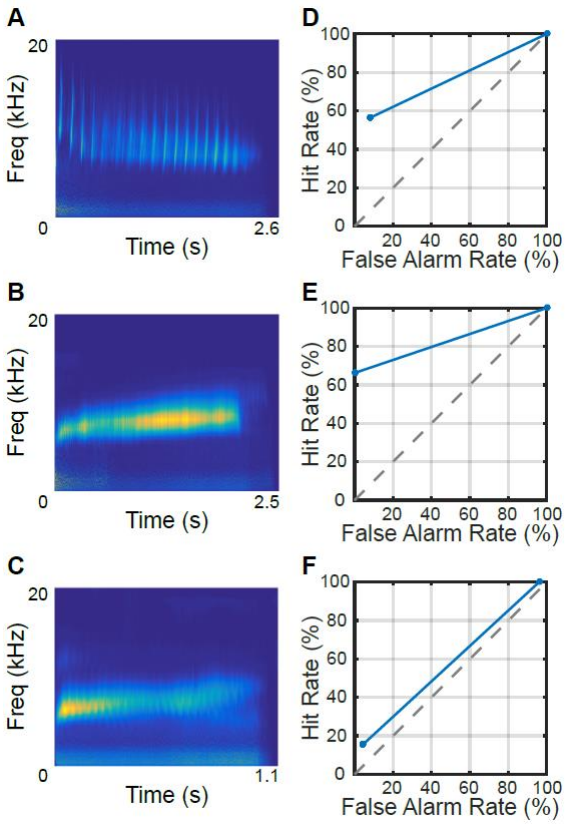
Appendix Figure 1 Production variability of major marmoset call types. **(A-C)** Production variability of phee calls quantified along various parameters: **(A)** bandwidth, **(B)** duration, and **(C)** dominant frequency. Dots depict parameter values for single calls, and histograms indicate the overall distribution of these parameters, split into the training (blue) and testing (red) sets. **(D-F)** Production variability of trill calls quantified as in (**A-C**).



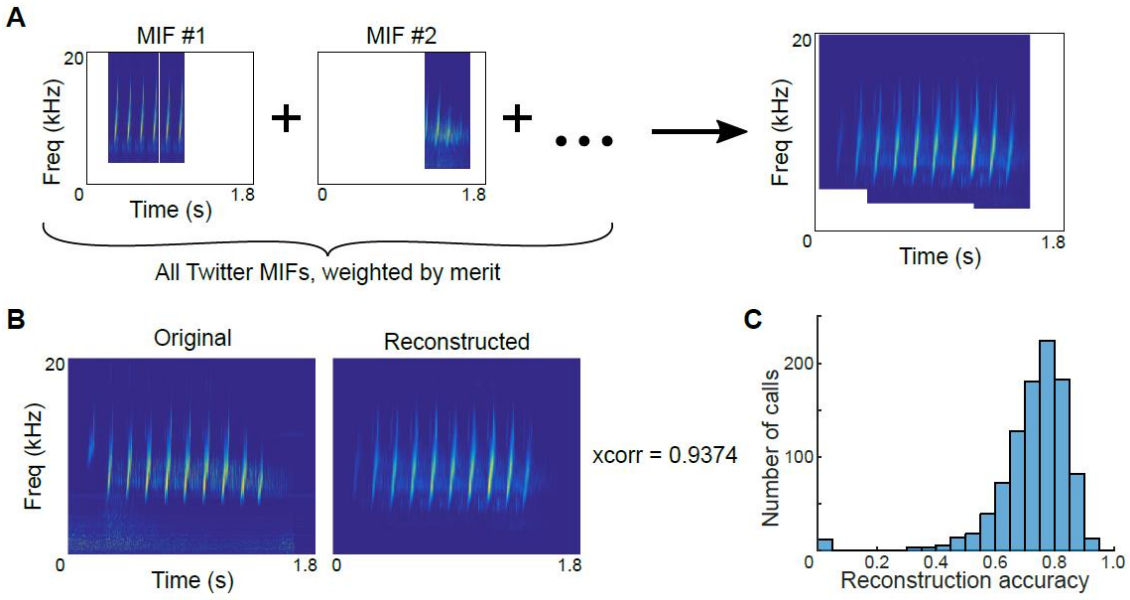
Appendix Figure 2 Information content, complexity, and size of all initial random features. Scatter plot of all 6000 features generated for each call type: twitter **(A)**, phee **(B)**, and trill **(C)**, as a function of their bandwidth and temporal extent. Color scale corresponds to the merit of each feature. Marginal histograms depict the maximum merit in each time- or width-bin. **(D-F)** Features of high merit for classification tend to be of intermediate complexity. Merit vs complexity plot of all randomly generated twitter **(D)**, phee **(E)**, and trill **F)** features. Feature complexity is estimated to be proportional to the reduced kurtosis of the distribution of activity within a feature or call. In these plots, low- or mid-merit features (defined as the bottom 33-%ile (light gray) and 33rd - 66th %-ile (dark gray)) show distributions of low kurtosis values. Whole calls show high kurtosis values (purple). Across call types, high-merit features (top 33%-ile) show intermediate kurtosis values, indicating that high-merit features are of intermediate complexity.



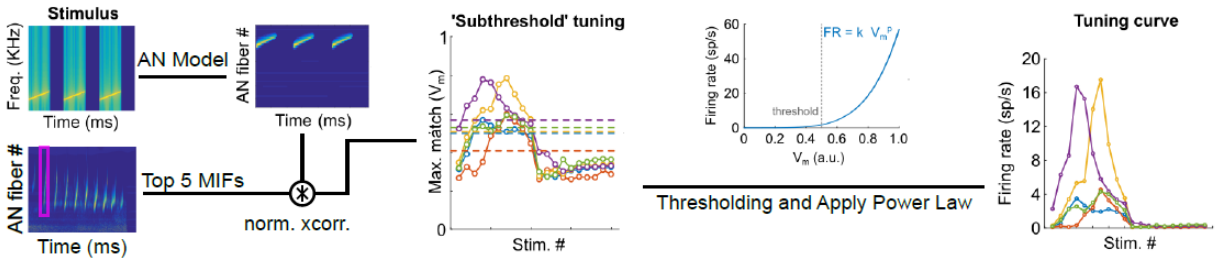
Appendix Figure 3 Similar classification performance obtained using distinct MIF sets. ROC curves for twitter classification using four successive iterations of MIFs, generated by removing all MIFs from the previous set, and selecting MIFs from the remaining features. High performance demonstrates that feature space was adequately sampled, and that the algorithm was not stuck in local maxima.



Appendix Figure 4 Classification using average calls. An average twitter **(A)**, trill **(B)**, and phee **(C)** constructed by aligning and averaging over the calls. **(D-F)** Classification performance using the average call as the single informative feature.



Appendix Figure 5 Reconstruction of twitter calls using only twitter MIFs. **(A)** Cochleagrams of MIFs were placed at the time points at which MIFs were detected within a sample twitter call. All MIF cochleagrams were then summed, weighted by their loglikelihood ratios. **(B)** Cochleagrams of and example original twitter call and its reconstructed version. **(C)** Histogram of the reconstruction accuracy of 1000 twitter calls.



Appendix Figure 6 Simulation of putative MIF-neuron tuning properties. The responses of MIFs to cochleagrams of commonly used auditory stimuli were taken to be the maximum value of the normalized cross-correlation function. A power law nonlinearity was applied to this value to obtain 'tuning curves' of the MIF-neurons to these stimuli.

Bibliography

Abbott, L. (1997). Synaptic Depression and Cortical Gain Control. Science, 275(5297), 221-224.

Agamaite, J. A., Chang, C.-J., Osmanski, M. S. & Wang, X. (2015). A quantitative acoustic analysis of the vocal repertoire of the common marmoset (Callithrix jacchus). J. Acoust. Soc. Am. 138, 2906–2928

Aizenberg, M., & Geffen, M. (2013). Bidirectional effects of aversive learning on perceptual acuity are mediated by the sensory cortex. Nature Neuroscience, 16(8), 994-996.

Akselrod-Ballin, A. & Ullman, S. (2008). Distinctive and compact features. Image Vision. Comput. 26, 1269–1276

Anderson, J., Carandini, M., & Ferster, D. (2000). Orientation Tuning of Input Conductance, Excitation, and Inhibition in Cat Primary Visual Cortex. Journal Of Neurophysiology, 84(2), 909-926. doi: 10.1152/jn.2000.84.2.909

Anselmi F, Leibo JZ, Rosasco L, Mutch J, Tacchetti A, Poggio T. (2016). Unsupervised learning of invariant representations. Theor Comput Sci 633: 112 – 121

Asakawa 2007, Automatic recognition of connected vowels only using speaker-invariant representation of speech dynamics

DiMattina C, Wang X. (2006). Virtual vocalization stimuli for investigating neural representations of species-specific vocalizations. J Neurophysiol 95:1244-1262

Asari, H., Pearlmutter, B. A. & Zador, A. M. (2006) Sparse representations for the cocktail party problem. J. Neurosci. 26, 7477–7490

Atiani S, Elhilali M, David SV, Fritz JB & Shamma SA (2009). Task difficulty and performance induce diverse adaptive patterns in gain and shape of primary auditory cortical receptive fields. Neuron 61, 467–480.

Belin, P., Bestlemeyers, P. E., Latinus, M., & Watson, R. (2011). Understanding voice perception. British Journal of Psychology. 102(4), 711-725

Berryman, J. C. (1976). Guinea-pig vocalizations: their structure, causation and function. Z. Tierpsychol. 41, 80–106

Bidelman, G., Moreno, S., & Alain, C. (2013). Tracing the emergence of categorical speech perception in the human auditory system. Neuroimage, 79, 201-212.

Brown, S., & Masland, R. (2001). Spatial scale and cellular substrate of contrast adaptation by retinal ganglion cells. Nature Neuroscience, 4(1), 44-51.

Carandini, M., & Heeger, D. (1994). Summation and division by neurons in primate visual cortex, Science, 264(5163).

Carandini, M., Heeger, D., & Movshon, J. (1997). Linearity and Normalization in Simple Cells of the Macaque Primary Visual Cortex. The Journal of Neuroscience, 17(21), 8621-8644.

Carandini, M., & Heeger, D. (2011). Normalization as a canonical neural computation. Nature Reviews Neuroscience, 13(1), 51-62.

Carruthers, I., Laplagne, D., Jaegle, A., Briguglio, J., Mwilambwe-Tshilobo, L., Natan, R., & Geffen, M. (2015). Emergence of invariant representation of vocalizations in the auditory cortex. Journal Of Neurophysiology, 114(5), 2726-2740.

Chance, F., Abbott, L., & Reyes, A. (2002). Gain Modulation from Background Synaptic Input. Neuron, 35(4), 773-782.

Chang, E., Rieger, J., Johnson, K., Berger, M., Barbaro, N., & Knight, R. (2010). Categorical speech representation in human superior temporal gyrus. Nature Neuroscience, 13(11), 1428-1432. de Heer, W., Huth, A., Griffiths, T., Gallant, J., & Theunissen, F. (2017). The Hierarchical Cortical Organization of Human Speech Processing. The Journal Of Neuroscience, 37(27), 6539-6557

Chen, H. C., Kaplan, G. & Rogers, L. J. (2009). Contact calls of common marmosets (Callithrix jacchus): influence of age of caller on antiphonal calling and other vocal responses. Am. J. Primatol. 71, 165–170

Colburn, S. H., Carney, L., & Heinz, M. (2003). Quantifying the Information in Auditory-Nerve Responses for Level Discrimination. JARO - Journal Of The Association For Research In Otolaryngology, 4(3), 294-311. doi: 10.1007/s10162-002-1090-6

Cooke, J. E., Kahn, M. C., Mann, E. O., King, A. J., Schnupp, J. W., & Willmore, B. D. (2020). Contrast gain control occurs independently of both parvalbumin-positive interneuron activity and shunting inhibition in auditory cortex. Journal of Neurophysiology, 123(4), 1536-1551.

Dean, I., Harper, N., & McAlpine, D. (2005). Neural population coding of sound level adapts to stimulus statistics. Nature Neuroscience, 8(12), 1684-1689

Delgutte, B. (1980). Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers. The Journal Of The Acoustical Society Of America, 68(3), 843-857

DiMattina, C. & Wang, X. (2006). Virtual vocalization stimuli for investigating neural representations of species-specific vocalizations. J. Neurophysiol. 95, 1244–1262

Ding, N., & Simon, J. (2013). Adaptive Temporal Encoding Leads to a Background-Insensitive Cortical Representation of Speech. Journal Of Neuroscience, 33(13), 5728-5735.

Eisenberg, J. F. (1974). The function and motivational basis of hystricomorph vocalizations. Symp. Zool. Soc. Lond. 34, 211–247

Epple, G. (1968). Comparative studies on vocalization in marmoset monkeys (hapalidae). Folia Primatol. 8, 1–40

Fino E, Packer AM & Yuste R (2013). The logic of inhibitory connectivity in the neocortex. Neuroscientist 19, 228–237.

Friesen, L., Shannon, R., Baskent, D., & Wang, X. (2001). Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants. The Journal Of The Acoustical Society Of America, 110(2), 1150-1163.

Fritz, J., David, S., Radtke-Schuller, S., Yin, P., & Shamma, S. (2010). Adaptive, behaviorally gated, persistent encoding of task-relevant auditory information in ferret frontal cortex. Nature Neuroscience, 13(8), 1011-1019.

Fu, Q., & Nogaki, G. (2005). Noise Susceptibility of Cochlear Implant Users: The Role of Spectral Resolution and Smearing. Journal of The Association For Research In Otolaryngology, 6(1), 19-27.

Fu, Q., Galvin, J., Wang, X., & Nogaki, G. (2005). Moderate auditory training can improve speech performance of adult cochlear implant patients. Acoustics Research Letters Online, 6(3), 106-111.

Turner, C., Gantz, B., Vidal, C., Behrens, A., & Henry, B. (2004). Speech recognition in noise for cochlear implant listeners: Benefits of residual acoustic hearing. The Journal Of The Acoustical Society Of America, 115(4), 1729-1735.

Fukushima M, Saunders R.C., Leopold D.A., Mishkin M., Averbeck B.B. (2014). Differential coding of conspecific vocalizations in the ventral auditory cortical stream. J Neurosci 26:4665-4676

Fukushima, M., Doyle, A. M., Mullarkey, M. P., Mishkin, M. & Averbeck, B. B. (2015). Distributed acoustic cues for caller identity in macaque vocalization. R. Soc. Open Sci. 2, 150432

Grimsley, J. M., Shanbhag, S. J., Palmer, A. R. & Wallace, M. N. (2012). Processing of communication calls in guinea pig auditory cortex. PLoS ONE 7, e51646

Hauser, M. D. (1998). Functional referents and acoustic similarity: field playback experiments with rhesus monkeys. Anim. Behav. 55, 1647–1658

Heeger D. (1991) Nonlinear model of cat in visual cortex. Computational Models of Visual Processing, ed. Landy, M., Movshon, J. A., pp. 119-133. Cambridge, Massachusetts: MIT Press.

Heeger, D. (1992). Normalization of cell responses in cat striate cortex. Visual Neuroscience, 9(2), 181-197.

Hillenbrand, J., Getty, L. A., Clark, M. J. & Wheeler, K. (1995). Acoustic characteristics of American English vowels. J. Acoust. Soc. Am. 97, 3099–3111

Hochmair-Desoyer, I., Hochmair, E., Fischer, R., & Burian, K. (1980). Cochlear prostheses in use: Recent speech comprehension results. Archives Of Oto-Rhino-Laryngology, 229(2), 81-98.

Holmstrom, L. A., Eeuwes, L. B., Roberts, P. D. & Portfors, C. V. (2010). Efficient encoding of vocalizations in the auditory midbrain. J. Neurosci. 30, 802–819

Hromádka, T., Deweese, M. R. & Zador, A. M. (2008). Sparse representation of sounds in the unanesthetized auditory cortex. PLoS Biol. 6, e16

Hromádka, T. & Zador, A. M. (2009). Representations in auditory cortex. Curr. Opin. Neurobiol. 19, 430–433

Hsu, A., Woolley, S. M., Fremouw, T. E. & Theunissen, F. E. Modulation power and phase spectrum of natural sounds enhance neural encoding performed by single auditory neurons. J. Neurosci. 24, 9201–9211 (2004).

Higley, M., Contreras, D. (2006). Balanced Excitation and Inhibition Determine Spike Timing during Frequency Adaptation. Journal Of Neuroscience, 26(2), 448-457.

Jenkins, R., White, D., Van Montfort, X. & Mike Burton, A. Variability in photos of the same face. Cognition 121, 313–323 (2011).

Jiang, X., Chevillet, M. A., Rauschecker, J. P. & Riesenhuber, M. Training humans to categorize monkey calls: auditory feature- and category-selective neural tuning changes. Neuron 98, 405–416 (2018).

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Hagniere, S. V. & McDermott, J. H. A task optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. Neuron 98, 630–644 (2018).

Khalighinejad, B., Cruzatto da Silva, G. & Mesgarani, N. Dynamic encoding of acoustic features in neural responses to continuous speech. J. Neurosci. 37, 2176–2185 (2017).

Killian, N. J., Watkins, P. V., Davidson, L. S., & Barbour, D. L. (2016). The effects of auditory contrast tuning upon speech intelligibility. Frontiers in Psychology, 7. doi:10.3389/fpsyg.2016.01145

Kramer, R. S. S., Manesi, Z., Towler, A., Reynolds, M. G. & Burton, A. M. Familiarity and within-person facial variability: the importance of the internal and external features. Perception 47, 3–15 (2018).

Krishna, B., & Semple, M. (2000). Auditory Temporal Processing: Responses to Sinusoidally Amplitude-Modulated Tones in the Inferior Colliculus. Journal Of Neurophysiology, 84(1), 255-273.

Kato, Y. et al. Vocalizations associated with anxiety and fear in the common marmoset (Callithrix jacchus). Behav. Brain Res. 2275, 43–52 (2014).

Langner, G., & Schreiner, C. (1988). Periodicity coding in the inferior colliculus of the cat. I. Neuronal mechanisms. Journal Of Neurophysiology, 60(6), 1799-1822.

Lerner, Y., Epshtein, B., Ullman, S. & Malach, R. (2008). Class information predicts activation by object fragments in human object areas. J. Cogn. Neurosci. 20, 1189–1206

Levinson, S. (1986). Continuously variable duration hidden Markov models for automatic speech recognition. Computer Speech & Language, 1(1), 29-45.

Liberman, A., Cooper, F., Shankweiler, D., & Studdert-Kennedy, M. (1967). Perception of the Van Tasell, D., Soli, S., Kirby, V., & Widin, G. (1987). Speech waveform envelope cues for consonant recognition. The Journal Of The Acoustical Society Of America, 82(4), 1152-1161

Liu, S.T., Montes-Lourido, P., Wnag, X., & Sadagopan, S. (2019). Optimal features for auditory categorization. Nature Communications, 10(1)

MacKain, K. S., Best, C. T. & Srange, W. (1981). Categorical perception of English /r/ and /l/ by Japanese bilinguals. Appl. Psycholinguist. 2, 369–390

McNaughton, B., & Morris, R. (1987). Hippocampal synaptic enhancement and information storage within a distributed memory system. Trends In Neurosciences, 10(10), 408-415.

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. (2014). Phonetic Feature Encoding in Human Superior Temporal Gyrus. Science, 343(6174), 1006-1010.

Mesgarani, N., & Chang, E. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. Nature, 485(7397), 233-236.

Miller, C. T., Mandel, K. & Wang, X. (2010). The communicative content of the common marmoset phee call during antiphonal calling. Am. J. Primatol. 72, 974–980

Mlynarski, W. & McDermott, J. H. (2017). Learning midlevel auditory codes from natural sound statistics. Neural Comput. 8, 1–39

Montes-Lourido, P., Kar, M., Kumbam, I., & Sadagopan, S. (2021). Pupillometry as a reliable metric of auditory detection and discrimination across diverse stimulus paradigms in animal models. Scientific Reports, 11(1).

Nagel, K., & Doupe, A. (2006). Temporal Processing and Adaptation in the Songbird Auditory Forebrain. Neuron, 51(6), 845-859.

Nie, K., Stickney, G., & Zeng, F. (2005). Encoding Frequency Modulation to Improve Cochlear Implant Performance in Noise. IEEE Transactions on Biomedical Engineering, 52(1), 64-73.

Norman-Haignere, S., & McDermott, J. (2018). Neural responses to natural and model-matched stimuli reveal distinct computations in primary and nonprimary auditory cortex. PLOS Biology, 16(12), e2005127.

Okada, K., Rong, F., Venezia, J., Matchin, W., Hsieh, I., & Saberi, K. et al. (2010). Hierarchical Organization of Human Auditory Cortex: Evidence from Acoustic Invariance in the Response to Intelligible Speech. Cerebral Cortex, 20(10), 2486-2495.

Olsen, S., Bortone, D., Adesnik, H., & Scanziani, M. (2012). Gain control by layer six in cortical circuits of vision. Nature, 483(7387), 47-52.

Osmanski, M. S. & Wang, X. (2011). Measurement of absolute auditory thresholds in the common marmoset (Callithrix jacchus). Hear Res. 277, 127–133

Perrodin C, Kayser C, Logothetis NK, Petkov CI. (2011). Voice cells in the primate temporal lobe. Curr Biol 21:1408-1415

Peterson, G. E. & Barney, H. L. (1952). Control methods used in a study of the vowels. J. Acoust. Soc. Am. 24, 175–184

Petkov CI, Kayser C, Steudel T, Whittingstall K, Augath M, Logothetis NK. (2008). A voice region in the monkey brain. Nat Neurosci 11:367-374

Pollak, G. D. (2013). The dominant role of inhibition in creating response selectivities for communication calls in the brainstem auditory system. Hear Res. 305, 86–101

Portfors, C. V., Roberts, P. D. & Jonson, K. (2009). Over-representation of speciesspecific vocalizations in the awake mouse inferior colliculus. Neuroscience 18, 486–500

Poon, P., & Chiu, T. (2000). Similarities of FM and AM receptive space of single units at the auditory midbrain. Biosystems, 58(1-3), 229-237.

Rabiner, L., Rosenberg, A., & Levinson, S. (1978). Considerations in dynamic time warping algorithms for discrete word recognition. IEEE Transactions On Acoustics, Speech, And Signal Processing, 26(6), 575-582.

Rabiner 1989, A tutorial on hidden Markov models and selected applications in speech recognition

Rahnman, M., Wilmore, B D., King, A. J., & Harper, N. S., (2020) Simple transformations capture auditory input to cortex. Proceedings of the National Academy of Sciences, 117(45), 28442-28451

Rabinowitz, N., Willmore, B., Schnupp, J., & King, A. (2011). Contrast Gain Control in Auditory Cortex. Neuron, 70(6), 1178-1191

Rabinowitz, N., Willmore, B., Schnupp, J., & King, A. (2012). Spectrotemporal Contrast Kernels for Neurons in Primary Auditory Cortex. Journal Of Neuroscience, 32(33), 11271-11284.

Rabinowitz, N., Willmore, B., King, A., & Schnupp, J. (2013) Constructing Noise-Invariant Representations of Sound in the Auditory Pathway. Plos Biology, 11(11), e1001710

Raizada, R. D. S., Tsao, F., Liu, H. & Kuhl, P. K. (2010). Quantifying the adequacy of neural representations for a cross-language phonetic discrimination task: prediction of individual differences. Cereb. Cortex 20, 1–12

Räsänen, O., Nagamine, T. & Mesgarani, N. (2016). Analyzing distributional learning of phonemic categories in unsupervised deep neural networks. Cogscience 2016, 1757–1762

Rauschecker JP, Tian B. (2000). Mechanisms and streams for processing of "what" and "where" in auditory cortex. Proc Natl Acad Sci USA 97:11800-11806

Robinson, B., & McAlpine, D. (2009). Gain control mechanisms in the auditory pathway. Current Opinion in Neurobiology, 19(4), 402-407.

Romanski LM, Averbeck BB. (2009). The primate cortical auditory system and neural representation of conspecific vocalizations. Annu Rev Neurosci. 32:315-346

Rosen, S. (1992). Temporal information in speech: acoustic, auditory and linguistic aspects. Philosophical Transactions of The Royal Society of London. Series B: Biological Sciences, 336(1278), 367-373

Rothman, J., Cathala, L., Steuber, V., & Silver, R. (2009). Synaptic depression enables neuronal gain control. Nature, 457(7232), 1015-1018.

Ryan, M. (1983). Frequency modulated calls and species recognition in a neotropical frog. Journal Of Comparative Physiology A, 150(2), 217-221.

Sadagopan, S., Temiz-Karayol, N. Z. & Voss, H. U. (2015). High-field functional magnetic resonance imaging of vocalization processing in marmosets. Sci. Rep. 5, 10950

Sadagopan, S., & Wang, X. (2008). Level Invariant Representation of Sounds by Populations of Neurons in Primary Auditory Cortex. Journal Of Neuroscience, 28(13), 3415-3426.

Sadagopan, S. & Wang, X. (2009). Nonlinear spectrotemporal interactions underlying selectivity for complex sounds in auditory cortex. J. Neurosci. 29, 11192–11202

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech, And Signal Processing, 26(1), 43-49.

Schwartz, O., & Simoncelli, E. (2001). Natural signal statistics and sensory gain control. Nature Neuroscience, 4(8), 819-825.

Schnupp, J. (2006). Auditory Filters, Features, and Redundant Representations. Neuron, 51(3), 278-280.

Shannon, R., Zeng, F., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech Recognition with Primarily Temporal Cues. Science, 270(5234), 303-304

Sharpee, T., Atencio, C., & Schreiner, C. (2011). Hierarchical representations in the auditory cortex. Current Opinion in Neurobiology, 21(5), 761-767.

Sinha, P. Qualitative representations for recognition. In Proceedings of the Annual Workshop on Biologically Motivated Computer Vision. 249–262 (Springer-Verlag, London, UK, 2002).

Smith, E. C. & Lewicki, M. S. (2006). Efficient auditory coding. Nature 439, 978–982

Stowell, D. & Plumbley, M. D. (2014). Large-scale analysis of frequency modulation in birdsong data bases. Methods Ecol. Evol. 5, 901–912

Suta, D., Kvasnák, E., Popelár, J. & Syka, J. (2003). Representation of species-specific vocalizations in the inferior colliculus of the guinea pig. J. Neurophysiol. 90, 3794–3808

Tchernichovski, O., Nottebohm, F., Ho, C. E., Pesaran, B. & Miltra, P. P. (2000). A procedure for an automated measurement of song similarity. Anim. Behav. 59, 1167–1176

Tian B, Reser, D, Durham A, Kustov A, Rauschecker JP. (2001). Functional Specialization in Rhesus Monkey Auditory Cortex. Science 292:290-293

Town, S. M., Wood, K. C. & Bizley, J. K. (2018). Sound identity is represented robustly in auditory cortex during perceptual constancy. Nat. Commun. 9, 4786

Tsao, D. Y. & Livingstone, M. S. (2008). Mechanisms of face perception. Annu. Rev. Neurosci. 31, 411–437

Turesson HK, Ribeiro S, Pereira DR, Papa JP, de Albuquerque (2016). VHC. Machine learning algorithms for automatic classification of marmoset vocalizations. PLoS One 11: e0163041

Ullman, S., Vidal-Naguet, M. & Sali, E. (2002). Visual features of intermediate complexity and their se in classification. Nat. Neurosci. 5, 682–687

Ullman, S. & Bart, E. (2004). Recognition invariance obtained by extended and invariant features. Neural Netw. 17, 833–848

Ullman, S., Assif, L., Fetaya, E. & Harari, D. (2016). Atoms of recognition in human and computer vision. Proc. Natl Acad. Sci. USA 113, 2744–2749

Viola, P. & Jones, M. (2004). Robust real-time face detection. Int. J. Comput. Vision. 57, 137–154

Wakita, H. (1977). Normalization of vowels by vocal-tract length and its application to vowel identification. IEEE Transactions on Acoustics, Speech, And Signal Processing, 25(2), 183-192.

Wang, X., Merzenich, M., Beitel, R., & Schreiner, C. (1995). Representation of a species-specific vocalization in the primary auditory cortex of the common marmoset: temporal and spectral characteristics. Journal Of Neurophysiology, 74(6), 2685-2706.

Wang, X. (2000). On cortical coding of vocal communication sounds in primates. Proc. Natl Acad Sci USA 97, 11843–11849

Wang, X. & Kadia, S. C. (2001). Differential representation of species-specific primate vocalizations in the auditory cortices of marmoset and cat. J. Neurophysiol. 86, 2616–2620

Watkins, P., & Barbour, D. (2008). Specialized neuronal adaptation for preserving input sensitivity. Nature Neuroscience, 11(11), 1259-1261.

Wehr, M., & Zador, A. (2003). Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex. Nature, 426(6965), 442-446.

Wen, B., Wang, G., Dean, I., & Delgutte, B. (2009). Dynamic Range Adaptation to Sound Level Statistics in the Auditory Nerve. Journal Of Neuroscience, 29(44), 13797-13808.

Wilent, W., Contreras, D. (2004). Synaptic Responses to Whisker Deflections in Rat Barrel Cortex as a Function of Cortical Layer and Stimulus Intensity. Journal Of Neuroscience, 24(16), 3985-3998.

Woolley, S. M., Fremouw, T. E., Hsu, A. & Theunissen, F. E. (2005). Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. Nat. Neurosci. 8, 1371–1379

Yin P, Fritz JB & Shamma SA (2014). Rapid spectrotemporal plasticity in primary auditory cortex during behavior. J Neurosci 34, 4396–4408.

Young, E. (2007). Neural representation of spectral and temporal information in speech. Philosophical Transactions of The Royal Society B: Biological Sciences, 363(1493), 923-945

Zhang, X., Heinz, M., Bruce, I., & Carney, L. (2001). A phenomenological model for the responses of auditory-nerve fibers: I. Nonlinear tuning with compression and suppression. The Journal of The Acoustical Society of America, 109(2), 648-670.

Zhang, L., Tan, A., Schreiner, C., & Merzenich, M. (2003). Topography and synaptic shaping of direction selectivity in primary auditory cortex. Nature, 424(6945), 201-205.

Zilany, M., Bruce, I., Nelson, P., & Carney, L. (2009). A phenomenological model of the synapse between the inner hair cell and auditory nerve: Long-term adaptation with power-law dynamics. The Journal of The Acoustical Society of America, 126(5), 2390-2412.

Zilany, M. S., Bruce, I. C. & Carney, L. H. (2014). Updated parameters and expanded simulation options for a model of the auditory periphery. J. Acoust. Soc. Am. 126, 2390–2412