Time as a supervisor: temporal regularity and auditory object learning

Ronald W. DiTullio,[1,2] Eugenio Piasini,[2,3] Pratik Chaudhari,[4] Vijay Balasubramanian*,[1,2,5] and Yale E. Cohen*[6]

[1]*David Rittenhouse Laboratory, Department of Physics and Astronomy, University of Pennsylvania, USA*
[2]*Computational Neuroscience Initiative, University of Pennsylvania, USA*
[3]*Scuola Internazionale Superiore di Studi Avanzati (SISSA), Trieste, Italy*
[4]*Department of Electrical and Systems Engineering, University of Pennsylvania, USA*
[5]*Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA**
[6]*Departments of Otorhinolaryngology, Neuroscience, and Bioengineering, University of Pennsylvania, USA**

Sensory systems appear to learn to transform incoming sensory information into perceptual representations that can inform and guide behavior, or "objects", with minimal explicit supervision. Here, we propose that the auditory system can achieve this goal by using time as a supervisor, i.e., by learning features of a stimulus that are temporally regular. We will show that this procedure generates a feature space sufficient to support fundamental computations of auditory perception. In detail, we consider the problem of discriminating between instances of a prototypical class of natural auditory objects, i.e., rhesus macaque vocalizations. We test discrimination in two ethologically relevant tasks: discrimination in a cluttered acoustic background, and generalization to discriminate between novel exemplars. We show that an algorithm that learns these temporally regular features affords better or equivalent discrimination and generalization than a conventional feature-selection algorithm, i.e., principal component analysis. Our findings suggest that the slow temporal features of auditory stimuli may be sufficient for parsing auditory scenes, and that the auditory brain should utilize these slowly changing temporal features.

Hearing animals parse their auditory world into distinct perceptual representations that inform and guide behavior, i.e., auditory objects. To do so, the auditory system learns to group stimuli with similar spectrotemporal properties into one perceptual object, while simultaneously segregating stimuli with different properties into a different object or objects [1–5]. This learning task is complicated because natural objects can have the same identity despite large variations in physical properties, e.g., their location [6, 7]; they are often encountered in complex, cluttered environments, e.g., noisy rooms filled with speakers [8, 9]; and it is often necessary to generalize to novel objects, e.g., the speech of new acquaintances whose voices have different timbres, accents, and pitches[3, 10]. Despite these challenges, the auditory system appears to learn this stimulus-to-object transformation with minimal explicit supervision.

Intuitively, objects are entities composed of correlated components that change together, so that the whole retains its coherence and identity, even in the presence of other variations [11]. If natural objects are defined in this way, it should be possible to learn and discriminate between them based on temporal cues including: 1) coincidence between the components, 2) continuity in the components, and 3) continuity in the correlations between components. That is, the temporal regularities of an object should define its identity.

This notion, that an object is defined by its temporal properties, is consistent with our current interpretation of the Gestalt principles underlying auditory scene analysis [1, 12]. One of these principles suggests how temporal sequences of stimuli are assigned to common sources or objects. For example, although we hear the sound that occurs when someone steps on the ground, we bind the temporal series of such sounds into "footsteps". Likewise, we can segregate footsteps of someone running in sneakers from someone strolling in wooden clogs, based not only differences in their frequency structure but also on temporal properties like the relative timing between each step. Such Gestalt principles [1, 12] provide important intuitions for what defines an auditory object, and how the auditory system may discriminate between objects. However, they do not provide a unified computational principle that explains how such temporal information could actually be learned and utilized for discrimination by the auditory system. Here, we seek to provide such a computational understanding.

We use rhesus macaque vocalizations as prototypical auditory objects [13] sharing acoustic structure with human speech, bird song, and other sounds produced by animals [14, 15]. We replicate previous results showing these vocalizations are defined by having temporally continuous components. We then expand on these results by showing this structure extends to the continuity of correlations between components and that this structure is significantly less present or virtually absent for other classes of stimuli.

Next, we use Slow Feature Analysis (SFA), a temporal learning algorithm originally used to study vision [16], to infer auditory features from four ethologically relevant classes of macaque vocalizations as well as control stimuli. For our purposes, SFA serves as a convenient method for extracting linear and nonlinear features [17] summarizing the most temporally continuous components and correlations between components of auditory stimuli.

Finally, we train liner support vector machine (SVM) classifiers to separate vocalizations in low-dimensional subspaces of the learned feature space. Our SFA-based classifiers outperform Principal Component Analysis (PCA)-based classifiers which assume that the most variable features, as opposed to the most temporally continuous ones, are informative about object identity. We confirm consistency with our hypothesis: our classifiers perform at chance on white noise, and at intermediate levels on applause, an auditory texture [7, 18] generated by the superposition of claps. These results also hold after addition of clutter to the training and test sets, showing that SFA-based classifiers can solve the "cocktail party problem". Finally, our

* These authors contributed equally to this work

classifiers generalize to novel examplars of all four vocalization classes.

Our results suggest that time can supervise the learning of natural auditory objects. That is, auditory features capturing the most temporally continuous components and correlations of auditory stimuli may suffice for parsing auditory scenes; and implying that the brain should be tuned to these slowly changing temporal features.

I. RESULTS

A. Temporal regularity in natural auditory stimuli

We propose that auditory objects are fundamentally defined by temporal regularity, specifically through the continuity of components of the acoustic signal and of correlations between these components. (Fig. 1A). Thus, we first tested whether these continuities and correlations can be seen directly in auditory stimuli. To do so, we had to select between two broad categories of natural auditory stimuli: acoustic events, such as vocalizations, and textures, such as rain [3, 7, 10, 18]. We focused on acoustic events, exemplified by rhesus macaque vocalizations, because of their ethological relevance [13], prototypical structure [14, 15], and links to previous work on neural correlates of auditory perception [10, 14, 19]. By contrast, textures appear to be defined by longer-time scale/time averaged statistics [7, 18] and thus it is possible the brain processes them differently than acoustic events.

We first characterized the data by computing spectrograms (Fig. 1B) of four ethlogically relevant vocalization classes (Coos, Harmonic Arches, Shrill Barks, and Grunts) and three auditory textures [7] (clutter composed of superpositions of vocalizations, applause and white noise); see Methods - Auditory Stimuli for details. The vocalizations visibly show structure in the frequency domain, i.e., coincidences in which frequencies are present, that appear to vary continuously over time. Although these vocalization appear at first glance to have similar structures, previous work has shown significant quantitative differences between these classes [13, 20]. By contrast, the applause and white noise spectrograms appear unstructured and temporally discontinuous to the eye, with clutter showing some limited structure.

To quantify these observations we computed the modulation power spectra by taking the Fourier transform of the spectrogram [15] (Fig. 1C, Top). Consistent with previous studies of bird calls, human speech, and macaque calls [14, 15], the vocalizations in our dataset are characterized by high power at low spectral and temporal modulation frequencies (Fig. 1C Top), indicating the presence of continuous structure in the acoustic frequencies, and slow variation of this structure. This concentration of power at low modulation frequencies declines somewhat in vocalization "clutter" (a random superposition of 20 vocalizations, analogous to the din in social environments [8, 21]) and still further in applause and white noise.

To test whether vocalizations additionally show continuity of correlations, we conducted a novel analysis: quantifying the spectrotemporal modulation of the matrix of instantaneous pairwise correlations between all frequencies (Fig. 1C, Bottom). In detail, we organized the upper triangular part of the correlation matrix as a time-varying vector, and and computed its modulation power spectrum. We found that for vocalizations and vocalization clutter, the modulation spectra of pairwise correlations is concentrated at low temporal modulation frequencies. In contrast, there was no such concentration for applause and white noise.

These results are consistent with our proposal that auditory objects are defined by the temporal regularity, specifically through the continuity of components of auditory stimuli and of correlations between these component. In the next sections, we test our prediction that an unsupervised algorithm, SFA, can learn features related to these regularities and that these features can support fundamental computations of perception: discrimination and generalization.

B. Learning temporal regularities

Next, we asked whether the identities of auditory objects can be learned just from temporal cues of continuity and the continuity of correlations. To this end, we first passed stimuli through a simple cochlear model ($N = 42$ gammatone spectral filters followed by a temporal filter, normalized to have zero mean and unit variance, see Methods - Simple Cochlear Model). We assembled these filter responses into an auditory response vector $\vec{x}(t) = \{x_0(t), x_1(t), x_2(t) \cdots x_N(t)\}$ where $x_0(t) = 1$ and $x_i(t)$ for $i > 0$ equals the ith cochlear filter response. We then applied Slow Feature Analysis (SFA) [16], a method for extracting temporally continuous features from data.

In detail, consider M features $f_i(\vec{x}(t))$ $(i = 1 \cdots M)$ computed instantaneously from the auditory signal. The set of features $F = \{f_1, f_2, \cdots f_M\}$ is constrained to belong to a function class appropriate to the problem. If $M < N$ we are performing dimensionality reduction of the stimulus, and if $M > N$ we are performing an expansive projection, such as the maps to a higher dimension that facilitate classification in the Support Vector Machine.

Slow Feature Analysis selects the most temporally continuous set of features by minimizing the expected value of the square of the time derivative $\Delta_i = \langle (df_i/dt)^2 \rangle_t$, termed the "slowness", over the specified function class. Here, the expectation value $\langle \cdot \rangle_t$ indicates a time-averaged quantity. To eliminate the trivial constant solution $f_i = c$, the features are constrained to have zero mean and unit variance: $\langle f_i \rangle_t = 0$ and $\langle f_i^2 \rangle_t = 1$. Likewise, to ensure that the each f_i encodes a different feature, we require the features to be uncorrelated, i.e., $\langle f_i f_j \rangle_t = 0$ for $i \neq j$, and in ascending order of slowness ($\Delta_i \leq \Delta_j$ if $i < j$). It is important here that the f_i are instantaneous functions of time, and no temporal filtering is involved. Because of this, the f_i can only be "slow" if the input contains slowly varying information that can be instantaneously extracted, such as object identity.

In this language, our instantaneous auditory stimulus is an N-dimensional vector, and superposition of stimuli is described by addition of their acoustic vectors. With this in mind, we can think of an auditory feature as a direction $\vec{w}$ in the acoustic vector

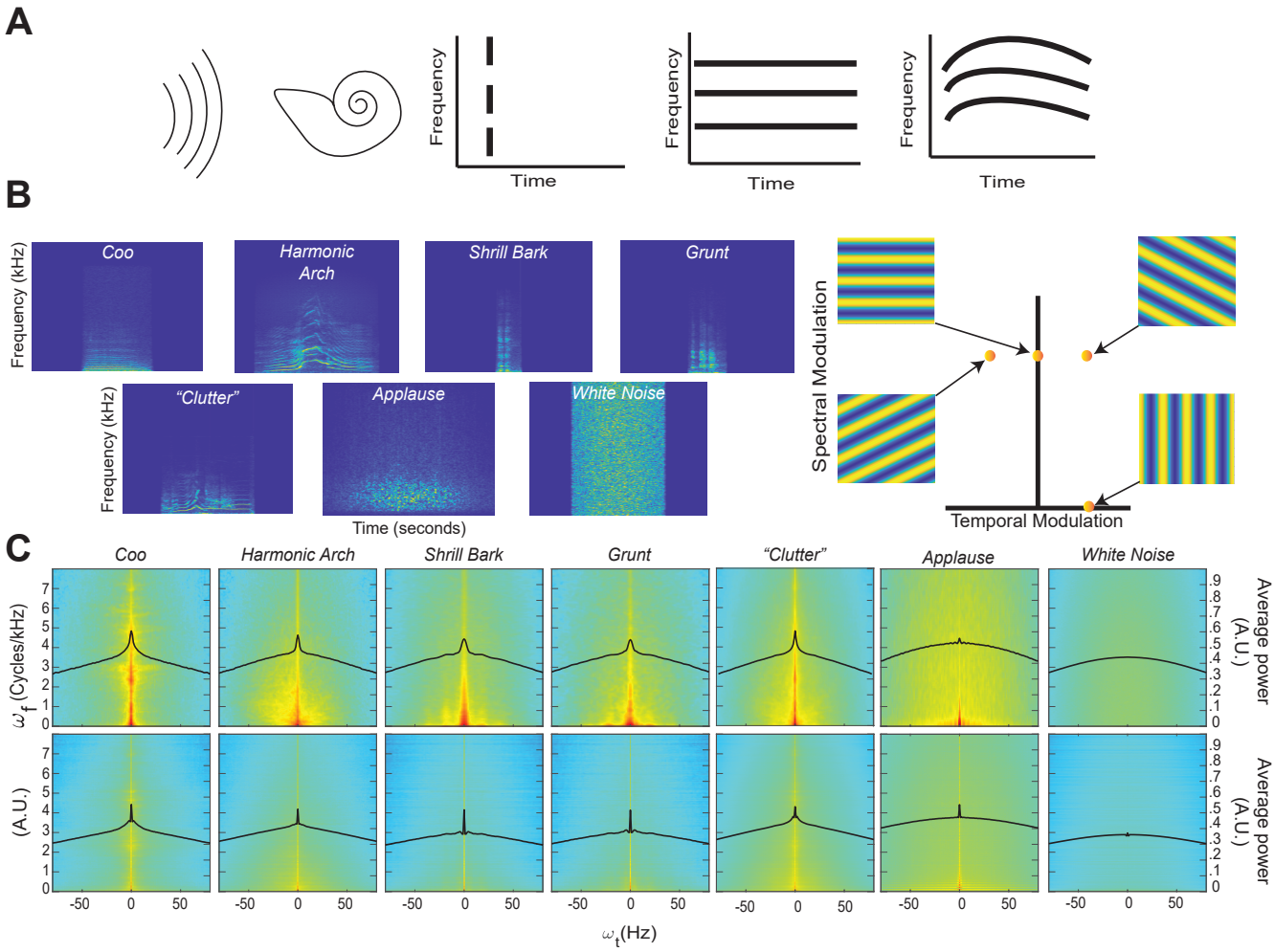

Figure 1: Coincidence, continuity – spectrograms and modulation spectra of auditory stimuli. A. *Far left.* An auditory stimulus (schematized as a set of parallel curved lines) as heard by the auditory system (schematized by the cartoon cochlea). *Left.* Coincident onset of three frequency bands that compromise an auditory stimulus. *Near Right.* Coincident onset and stable persistence of three frequencies that compromise an auditory stimulus. *Far Right.* Continuously varying correlations between the frequency components of an auditory stimulus. **B.** *Left.* An example spectrogram from four different classes of rhesus vocalization, vocalization clutter (see Methods - Auditory Stimuli for more details), applause, and white noise. Each spectrogram displays 1 second of time with the stimulus centered in time for clarity. Color represents the amount of energy at each time-frequency point with warmer colors showing points with the most energy and cooler colors showing the least energy.. *Right.* A spectrogram can be decomposed into a weighted sum of different spectral and temporal modulation rates. Each dot indicates the location in which the indicated spectrogram (arrow) would lie in spectrotemporal modulation space. The spectrogram lying on the abscissa (ordinate) indicate special cases in which there is only temporal (spectral) modulation. Color represents again represent the amount of energy at teach time-frequency point. **C.** *Top Row.* Average spectrotemporal modulation spectra for coos, harmonic arches, grunts, barks, vocalization clutter, applause, and white noise. *Bottom Row.* Same plot as above but now for the frequency-correlation matrix. Regions in red have the most energy, whereas those in light blue have the least energy. In each spectrum, the black line represents the average spectral power as a function of temporal modulation. Modulation spectra is calculated according to previous procedures [14, 15]. For each vocalization class and applause, the average spectra is calculated over all exemplars of that class within our library. For clutter and white noise, 100 examples were generated and the average spectra is calculated over all 100 examples. (see Methods - Auditory Stimuli for more details)

space, and define objects as coincident superpositions of such features. Then we can decompose the stimulus into features via the projections $L_i(t) = \vec{w}_i \cdot \vec{x}(t) = \sum_j w_{ij} x_j(t)$. Slow Feature Analysis conducted with this set of linear features $\vec{L} = \{L_1(t), L_2(t), \cdots\}$ chooses weight vectors $\vec{w}_i$ to give the most temporally continuous projections of stimuli. The zero mean condition on SFA features says that $\langle L_i \rangle_t = 0$ which implies that $w_{io} = -\sum_{j=1}^{N} w_{ij} \langle x_j \rangle_t$, effectively implementing a mean subtraction. This leads to a 42 dimensional linear feature space, the same dimensionality as the input auditory signal. According to our hypothesis, we should be able to train a classifier to detect the presence of auditory object classes based on the coincident presence of such temporally continuous linear SFA (lSFA) features (schematic in Fig. 2).

We also hypothesised that objects are partly defined by

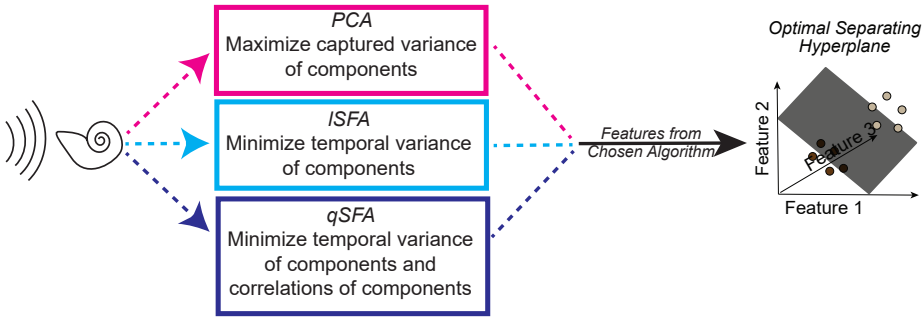

Figure 2: Schematic of one instance of the analysis procedure A set of auditory stimuli are first passed through a simplified cochlear model consisting of a stage of spectral filtering followed by a temporal filtering stage (see Methods - Simple Cochlear Model for more details). The output of this cochlear model is then analyzed by one of three algorithms: PCA, linear SFA, or quadratic SFA. The "top" five features for the selected algorithm are then used to generate a feature space. See text insert in figure for summary of what each algorithm is optimized to find, and thus what "top" means for each algorithm. Finally, test stimuli were projected into the selected algorithm's feature space and a linear support vector machine (SVM) classifier was trained and tested using a 30-fold cross validation procedure. The goal of the classifier is to find a hyperplane that correctly separates the projected data points from one stimulus from the projected data points of another stimulus/other projected stimuli. Except for the analysis on generalization, test stimuli and training stimuli were the same. This procedure was repeated for each set of stimuli so there is no learning transferred between sets. See Methods - Discriminating between vocalizations for more details.

continuity of change in the correlations between components. So we want a function class on which SFA can act to extract quantities related to such slowly changing correlations. To this end, consider that the pairwise correlation between the linear features are $C_{ij} = \langle L_i L_j \rangle_t = \sum_{k,l} w_{ik} w_{jl} \langle x_k x_l \rangle_t$. All of these correlations are linear combinations of the pair correlations in the auditory stimulus $D_{kl} = \langle x_i(t) x_j(t) \rangle_t$ suggesting that we should apply SFA to the nonlinear function space of quadratic polynomials in the acoustic signal. In other words, we use quadratic features of the form: $Q_i(t) = \vec{x}^T V_i \vec{x}$ where $\vec{x}^T$ is the transpose of $\vec{x}$ and V_i is a symmetric weight matrix associated to feature Q_i. Counting the independent components of the weight matrix leads to an 903 dimensional feature space. Since $x_0(t) = 1$ in our conventions, the Q_i are linear combinations of both linear and quadratic monomials in the auditory filter responses, and also include a constant that can be used to implement the mean subtraction required for SFA. Quadratic SFA (qSFA) then selects the weight matrices V_i to extract the most slowly changing quadratic features of the stimulus.

Many standard feature selection algorithms pay no explicit attention to temporal regularity. For example, Principal Components Analysis (PCA) of our data would focus on linear combinations of the cochlear outputs that have the largest variance. Thus, as an algorithmic control we also used PCA to generate a feature space. We then trained linear support vector machine (SVM) classifiers to separate our four vocalization classes and three auditory textures on the basis of the features extracted by linear SFA, quadratic SFA, and PCA (Fig. 2).

C. Auditory discrimination from temporal regularities

Our premise is that temporally continuous features can provide a good low-dimensional representation of object identity. To test this, we constructed a one-shot learning task for discriminating between any two exemplar macaque vocalizations, i.e. a vocalization pair (see details in Methods - Discriminating between vocalizations). Vocalization pairs were sampled uniformly from our four classes. Thus the pair consisted of either the same or different vocalization classes (e.g., coo-coo or coo-grunt), while the individuals producing the pairs could also be the same or different. This task design mirrors our intuition listening to a conversation: two words said by the same speaker are different auditory objects just as the same word said by two different speakers are different auditory objects. We used quadratic SFA on 20 unique pairs to extract the k slowest features for each pair, and then projected the pair into the low dimensional space defined by these features. In this way, over its time course, each stimulus generated a data cloud in the feature space. We trained a linear SVM on 75% of this data to separate the two vocalizations, and tested with the remaining 25% (see Methods for details of cross-validation). This procedure tests whether quadratic SFA with k slowest features is capable of supporting one-shot learning of auditory object discrimination. Classification performance was already significantly above chance (50%) with a single feature (Fig, 3 A. Confidence intervals do not include 50%). With just $7 - 10$ slowest features out of the 903 dimensions of our quadratic SFA feature space, classification accuracy reached about 97% with diminishing returns for the addition of further features (Fig, 3 A). Thus, in subsequent analyses, we restricted ourselves to the first 5 slowest SFA features as performance was consistently above 95%.

Next, we compared the discrimination performance of classifiers trained on feature spaces derived from PCA, linear SFA, and quadratic SFA. To better sample from our library of vocalization pairs, we utilized 100 pairs instead of 20 pairs; otherwise, we followed the same procedure. With 5 features, median classifier accuracy was 85.2%, 94.7%, 98.8 % for PCA, lSFA, and qSFA respectively, all of which were above chance (95% confidence interval (CI) about median for each

algorithm does not include chance level performance 50%: 84.2% - 87.6%, 92.2% - 95.5%, 98.3% - 99.3 %). This high classification accuracy was seen across all vocalization classes and was equivalent whether we conducted within (e.g., two different coo exemplars) or across (e.g., a coo exemplar and a grunt exemplar) vocalization-classes (data not shown). Thus, across all cases discrimination based on SFA, which selects temporally continuous features, outperforms discrimination based on PCA, which selects features with high variance across time. Mann-Whitney U test for all algorithms: $p < 10^{-6}$ H_0: classifier accuracy between each algorithm pair was equal.)

The SFA projection to a slow feature space cannot add information about vocalization identity that is not already present in the data, but it could reduce the information. To assess an upper bound on the discrimination performance, we trained a linear SVM directly on the 42 channel output of the cochlear model. This classifier had high accuracy and low variance (Fig. 3B, second from right; median accuracy: 99.3%; 95 % Confidence Interval: 98.9% - 99.5%). This upper bound was significantly higher than both PCA and lSFA (Mann-Whitney U test for all algorithms: $p < 10^{-18}$, H_0: classifier accuracy between algorithms was equal)(Fig. 3B). However, the classifer based on quadratic SFA had similar accuracy and variance as the upper bound: neither measure showed a statistically significant differences and both algorithms had overlapping CIs (Fig. 3B). This suggests that quadratic SFA has extracted essentially all the information pertinent to vocalization discrimination in its five slowest features.

As a control analysis, we randomly selected 5 channels of the cochlear model and applied the SVM to them. This formulation serves as an alternative null model to chance performance and tests whether the features selected by quadratic SFA could have been found trivially. We found that PCA, linear SFA and quadratic SFA all out-performed this null model (Fig. 3B) (Mann-Whitney U test for all algorithms: $p < 10^{-8}$, H_0: classifier accuracy was less than or equal to null model classifier). Further, linear and quadratic SFA, but not PCA, had lower variance in the classification performance than this null model. (Levene comparison of variance for each algorithm lSFA and qSFA respectively: $p < 0.01$, $p < 10^{-14}$, H_0: classifier variance in accuracy was equal to null model classifier variance). These results confirm that slow features that select for temporal continuity are nontrivial.

Finally, we tested whether the specific correlation structure seen in the vocalizations (Fig. 1) drove the excellent performance of the SFA-based classifier, or whether the performance would be achieved for any set of auditory stimuli. To this end, we repeated the above analyses for with quadratic SFA using exemplars of applause and tokens of white noise (Fig. 3 C and D, respectively). Performance for quadratic SFA applied to applause was still fairly high, but its variability was much higher than the analogous performance for the vocalizations. For white noise tokens, classification performance for quadratic SFA was substantially less than it was for vocalizations. Together, these findings are consistent with the notion that quadratic SFA captures the "slow" correlations within a vocalization (Fig. 1C), but struggles to capture informative structure when such correlations are not inherently present in the auditory stimulus.

Overall, we find that a low-dimensional feature space consisting of temporal continuous components and correlations of auditory stimuli supports excellent performance in one-shot learning of a discrimination task.

D. Discrimination in clutter

Auditory events often occur in the presence of background clutter, such as a din of other events. For example, consider the "cocktail party problem" in which the listener seeks to discriminate vocalizations against a background of other conspecific chatter [8]. Thus we tested whether learning feature from continuity of components and correlations supports auditory discrimination even in such cluttered environments

To this end, we modified our one-shot vocalization-pair discrimination paradigm to include background noise in two ways: (1) in the training data to test whether the algorithms still find useful features in the presence of clutter, and (2) in the test stimuli to test robustness of the learned features to novel clutter. We generated background clutter by summing a random ensemble of vocalizations and systematically varied the signal-to-noise ratio (SNR) between the training or test vocalizations and this background clutter. SNR as the log of the ratio of the average power of the vocalization pair and background clutter (details in Methods - Discriminating between vocalizations embedded in background clutter). In applying SFA and PCA we kept the five slowest or most variable features, respectively, and compared the results, as in the previous section, to: (a) a bound determining by applying an SVM to all 42 cochlear output channels, and (b) a null model that trains an SVM on five random cochlear output channels.

When clutter was added only to the training set, classification was poor at the lowest SNR, but still above chance for all algorithms (Fig. 4 A. confidence intervals do not include chance performance levels, 50%) Indeed, even at relatively low values of SNR (e.g., -3) performance was both above chance and above the null model for all algorithms Mann-Whitney U test for all algorithms: $p < 10^{-12}$, H_0: classifier accuracy was less than or equal to null model classifier). As SNR increased, classification performance increased in all cases but plateaued at a median accuracy of $\sim 85\%$ for PCA (Fig. 4A *top*), and at $\sin 93\%$ for linear SFA (Fig. 4 A *middle*). Quadratic SFA, however, did not plateau and reaching its highest accuracy at the highest two SNRs (Fig. 4 A *bottom*). Linear and quadratic SFA had the similar average accuracy across the tested SNRs (Average with standard error of the mean for each algorithm across all SNRs: PCA $78.0 \pm .454\%$, lSFA $83.3 \pm .528\%$, qSFA $82.2 \pm .519\%$). These results clearly demonstrate that SFA can learn objects embedded in cluttered backgrounds, and surpasses PCA. Interestingly, it seems that temporally continuous linear features, used by linear SFA and included in quadratic SFA, are sufficient for robust performance already at low SNRs. But as the SNR increases, quadratic features which select temporally continuous correlations in auditory stimuli increase in importance, leading to near-perfect performance at high SNR, matching the upper bound determined by the SVM

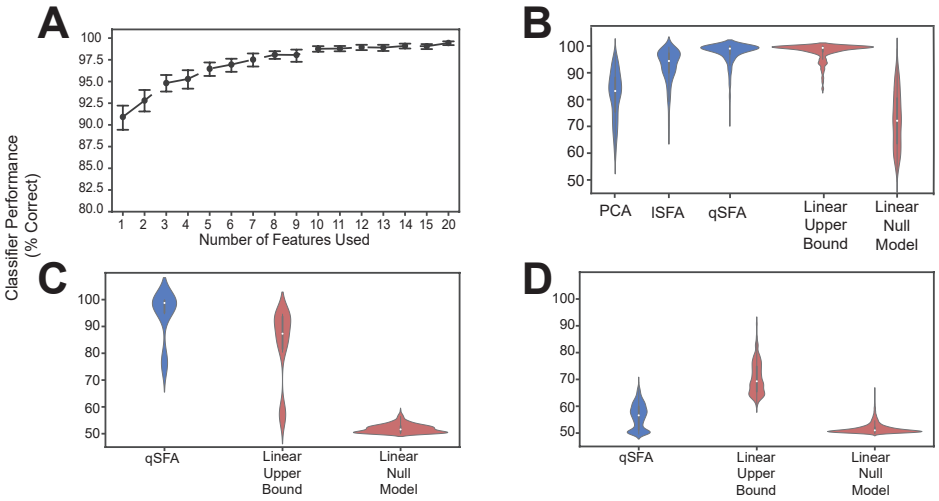

Figure 3: Classifier performance for PCA-, linear SFA-, and quadratic SFA-based classifiers on one-shot discrimination. **A.** Classifier performance of qSFA-based classifier as a function of the number of features for 20 pairs of vocalizations. Black dotes indicate mean performance and error bars indicate the 95 % confidence interval. **B.** Violin plot of PCA-, lSFA-, and qSFA-based classifiers (blue) and linear classifiers applied directly to the data (red) classification performance for 100 unique pairs of vocalizations. The central dot indicates median performance, whereas the whiskers indicate the interquartile range. **C.** Violin plot of qSFA-based classifiers (blue) and linear classifiers applied directly to the data (red) classification performance for 10 unique pairs of different applause tokens. **D.**Same as **C.** but for 100 unique pairs of different tokens of white noise. PCA-, lSFA-, and qSFA-based classifiers as well as the null linear model all use five features. The linear upper bound model uses the full feature space from the output of the cochlear model. See Methods - Discriminating between vocalizations for further details.

applied to the complete cochlear model output.

We obtained similar results when clutter was added only to the test set (Fig. 4 B.). At the lowest SNR, classification was well above chance for each algorithm (Confidence intervals do not include chance performance levels, 50%), and as SNR increased, classification performance increased in all cases (Spearman correlation for each algorithm respectively: .45, .63, .65. for all algorithms $p < 10^{-30}$ H_0: Spearman correlation coefficient is equal to zero). Again, Slow Feature Analysis outperforms PCA, and while linear SFA is more effective at the low SNRs (Average with standard error of the mean of each algorithm across all SNRs: PCA $77.5 \pm .376\%$, lSFA $82.6 \pm .420\%$, qSFA $73.7 \pm .504\%$), at higher SNR quadratic SFA achieves the upper performance bound set by the SVM acting on all 42 outputs of the cochlear model.

Finally, we examined classifier performance in our one-shot learning task when vocalizations were embedded in persistent but changing clutter, i.e., with different training and testing clutter. This approximates the conditions of listening in a crowded restaurant or party (i.e., the cocktail-party problem). Thus, we repeated the above analysis with different clutter tokens of varying SNR in both training and test sets (Fig. 5). At low levels of testing SNR (-7 though -1), performance was above chance for each algorithm and generally equivalent across algorithms. At higher levels of testing SNR (0 through 3), we found that linear SFA had the best average performance across training and test SNRs (Average with standard error of the mean of each algorithm: PCA $77.6 \pm .178\%$, lSFA $82.6 \pm .255\%$, qSFA $78.8 \pm .241\%$).

Interestingly quadratic SFA shows declining performance when the training and test SNRs are mismatched. This is not surprising when the training SNR is large while the test SNR is low, since any learning algorithm will be challenged by this scenario. However, quadratic SFA also performs less well if the training SNR is low while the test SNR is high. This outcome is initially surprising, but may be occurring because at low SNR a one-shot learner without supervision does not actually know what the target object is. Elements of the clutter may in fact appear equally salient, or may coincidentally overlap with the continuous correlations of the target vocalizations. In the latter case, the clutter effectively energetically masks [19, 22] the information in the quadratic features. In natural audition the listener has access to many other cues, e.g. visual cues from the sound source or vocalizing individual, relative motion of background objects, etc., that likely serve to disambiguate the object of interest. Likewise, auditory learners in natural environments will generally have access to multiple examples of the same object in different background clutter, again providing information for identifying the learning target. It will be interesting to extend our one-shot learning paradigm to investigate these directions in the future

E. Generalization to novel exemplars

Animals can generalize to novel exemplars of a particular stimulus class [3, 10]. For example, when we meet a new group of people at a conference, we can recognize and discriminate between the various "hello"s of different individuals despite the novelty of their accents, pitch, and timbre. To approximate this

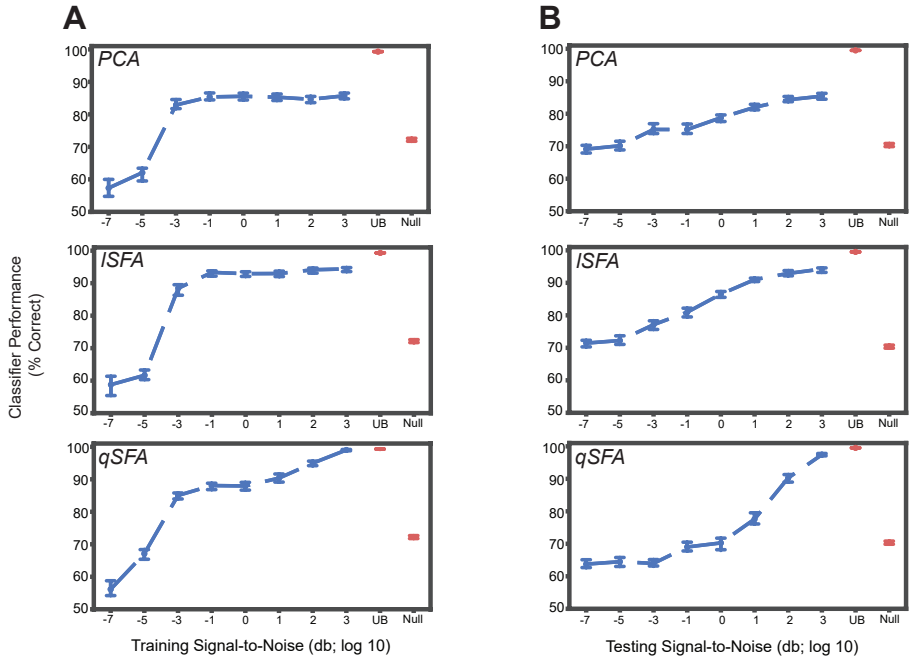

Figure 4: PCA-, linear SFA-, and quadratic SFA- based classifier performance with clutter in the training set and testing set. A. PCA-based, lSFA- and qSFA- based classifier performance as a function of the signal-to-noise ratio (SNR) between the training vocalizations and the background in the *training set*. Blue dots indicate median performance. Error bars indicate the 95% confidence interval. Grey-scale dots indicate mean performance for individual pairs of test vocalizations (e.g., coo vs grunt or shrill bark vs grunt). Red dots and error bars indicate the median performance and 95% confidence intervals for the linear upper bound (UB) and linear null model (Null); for these linear decoders, classifier performance was calculated only at the highest SNR. **B** The same as in **A.** but as a function of the SNR between testing vocalizations and background clutter in the *testing set*. As in previous analyses, PCA-, lSFA-, and qSFA-based classifiers as well as the null linear model use five features. The linear upper bound model uses the full feature space from the output of the cochlear model. All these analyses were performed on 100 unique pairs of vocalizations.

scenario, we used a hold-out training procedure to test whether our one-shot PCA- and SFA-based classifiers can generalize to novel exemplars. Briefly, we had 19 unique exemplars of each vocalization class. We generated 20 unique test pairs of vocalizations for each of the 4 vocalization classes from these 19 exemplars (*i.e.,* 20 pairs of coos, 20 pairs of grunts, *etc.*). For each test pair within a vocalization class, we applied PCA or SFA to the remainin 17 exemplars to generate a feature set. As in previous sections, we kept the five slowest or most variable features for PCA or SFA. Next, we project the test pair time course into the feature space defined by these five features, thereby generating a cloud of data points. Finally, we trained a linear SVM classifer to separate the two vocalizations by using a 75% - 25% train-test split of the data points arising from the projection (see details in Methods - Generalization to novel vocalization exemplars). In other words, we first learn a feature set that applies broadly to a vocalization classes (as opposed to a particular pair in the class), and then use this set to perform one-shot discrimination of a novel pair that did not contribute to generating the features.

Overall, classifiers based on both PCA and SFA generalized significantly above chance across all vocalization classes (Fig. 6; Mann-Whitney U test for each algorithms for each vocal class: $p < 10^{-8}$; H_0: classifier accuracy was less than or equal to chance performance of 50%). Median performance was relatively similar between classifiers for each vocalization class (median and interquartile range for each vocalization and algorithm: *Coo* - PCA: 93.0% , 9.00% ; lSFA: 94.7%, 10.9%; qSFA: 92.1%, 9.85% *Harmonic Arch* - PCA: 68.9% , 11.3 %; lSFA: 71.4%, 15.8%; qSFA: 75.1%, 13.6% *Grunt* - PCA: 68.3%, 9.22%; lSFA: 80.4 %, 17.8 %, qSFA: 73.8%, 18.1% *Shrill Bark* - PCA: 70.0%, 11.1%; lSFA: 76.4%, 16.3%; qSFA: 66.5%, 9.10%. Unlike previous analyses, median performance and variance varied between vocalization classes for all classifiers, with all algorithms performing better on coos than the other three classes (Kruskal-Wallis test on average performance for each vocal class: $p < 10^{-7}$; H_0: classifier accuracy was equal across all vocal classes; post-hoc Mann-Whitney U test for coo vs each other vocal class: $p < 10^{-8}$; H_0: classifier accuracy was equal between coo and other vocal class). This result indicates that, at least for comparisons between PCA- and SFA-based algorithms, within class generalization performance is more sensitive to the statistics of the class than to the classification algorithm. Overall, the above-chance performance for generalization indicates that SFA-based algorithms are able to generalize to novel exemplars, a critical computation in auditory perception.

Performance in this discrimination task is generally lower

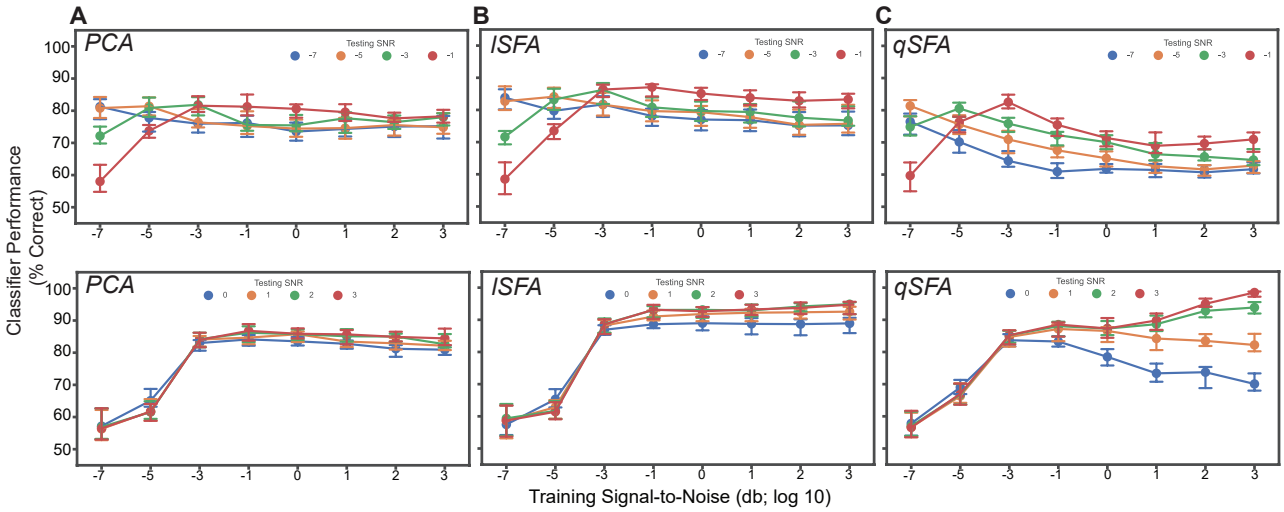

Figure 5: PCA, linear SFA, and quadratic classifier performance with clutter in the training and test set. *Top.* Low testing signal-to-noise ratio (SNR) conditions. *Bottom* High SNR conditions **A.** PCA classifier performance as a function of both training and testing SNR Dots indicate median performance for PCA. Error bars indicate the 95% confidence interval. Colors indicate the level of the SNR in the test set as stated in the legend. **B** and **C:** Same but for lSFA- and qSFA-based classifiers, respectively. As in previous analyses, PCA-, lSFA-, and qSFA-based classifiers all use five features. All these analyses were performed on 100 unique pairs of vocalizations.

than in the tasks studied in the previous sections. The source of this difference may lie in the nature of the slowest features captured by SFA in this case. In the previous one-shot learning examples, we used SFA to extract features specific to a particular pair of vocalizations. In this section, SFA is extracting features that generally characterize a larger ensemble of vocalizations. Extracting these general regularities may result in features that "average out" discriminatory differences between specific pairs of vocalizations, lowering classifier performance. The effect will depend on the particular temporal statistics of each vocalization class. Indeed, previous work has report higher variability between coos as compared to other vocalization classes [20], perhaps explaining the higher discrimination performance for the coo class. It is possible that the discriminatory differences that are absent the five slowest features that we used may be present in the next fastest features of SFA. If this is the case, increasing the number of features used will rapidly boost classifier performance. In effect, we are proposing a temporal feature hierarchy whereby a small number of slow features emphasizing the continuity of components and their correlations of auditory stimuli suffices to broadly support auditory object discrimination, while a set of progressively faster features supports finer separation between objects - a hypothesis that can be tested in future work.

II. DISCUSSION

Our results show that a low-dimensional feature space that selects the temporally continuous components and correlations of auditory stimuli can support the basic computations of auditory perception. We used Slow Feature Analysis to generate such feature spaces, and found that they were able to capture sufficient intra- and inter-class variance of rhesus macaque vocalizations to support discrimination within and between classes. These results were robust to naturalistic clutter within the training and/or testing sets. Furthermore, SFA-based classifiers generalized to novel exemplars within a vocalization class, although more study is required to understand details of which features capture the fine within-class details needed for discrimination, and how many such features we need for high accuracy. Together, these findings suggest that the brain could use temporal continuity and the continuity of correlations of auditory stimuli to learn the identities of auditory objects.

Can such temporal learning be implemented by neural circuits? Several lines of research suggest that this may be possible. First, previous work has shown that neurons can leverage spike-timing dependent plasticity (STDP) with Hebbian learning to learn the slow features favored by SFA in an unsupervised manner [23]. Further work even identified a specific STDP kernel, i.e., a function that identifies how plasticity changes as a function of the relative spike times of the pre- and post-synaptic neurons, that would learn the slow features of auditory stimuli in real time [24]. Second, because the slow features of SFA can be implemented as an equivalent set of low-pass filters, this particular temporal learning algorithm can also be implemented in a neural circuit whose low-pass filter properties are similar to the ones learned by STDP [23].

Alternatively, a temporal learning algorithm of this kind could instead be implemented through the modulation transfer functions (MTFs) of auditory neurons. Previous studies have shown auditory-cortex and forebrain neurons are tuned for particular spectrotemporal modulations, which can be quantified through their modulation transfer function (MTF)

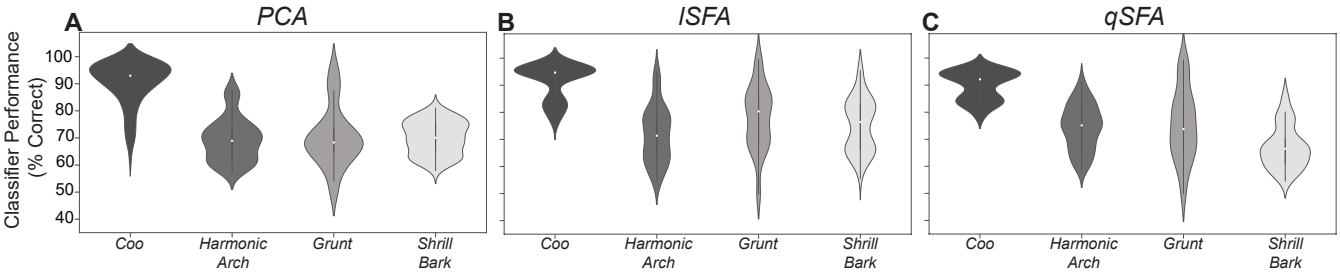

Figure 6: PCA-, linear SFA-, and quadratic SFA-based classifier performance tested on novel pairs of vocalizations A. PCA-based classifier performance as a function vocalization class. Dots indicate median performance for each vocalization class for PCA whereas the whiskers indicate the interquartile range. **B** and **C:** Same but for lSFA and qSFA-based classifiers, respectively. As in previous analyses, PCA-, lSFA-, and qSFA-based classifiers all use five features. These anayses were conducted on 20 novel pairs of vocalizations for each of the four vocalization classes. See Methods-Generalization to novel vocalization exemplars for more details on training and testing sets.

[14, 15, 22, 25–28]. Because the temporally continuous features of auditory stimuli are spectral patterns of increasing temporal modulation, i.e., the slowest feature reflects spectral patterns at the lowest temporal modulation and each faster feature reflects patterns at higher temporal modulations, a neuron's MTF could be reinterpreted as an implementation of an SFA-like computations. In this interpretation, different ensembles of downstream neurons learn to respond to particular combinations of upstream neurons with similar MTFs. Each downstream neuron then captures a particular feature found by SFA, and the ensemble as a whole would create a SFA-like feature space that could be used to support perception.

The literature of MTFs presents further interesting connections to our results. Work relating MTFs to human speech intelligibility has shown that selective filtering of low temporal modulations dramatically impairs speech comprehension [26]. This result aligns exactly with our predictions. Studies in songbirds have shown MTFs that cover the full range of temporal modulations of natural auditory stimuli including conspecific song, speech, and environmental sounds [22]. This range matches what is needed for our suggested implementation of temporal learning in its SFA form through MTFs. However, the majority of neurons in [22] actually show a preference for moderate temporal modulation rather than the low temporal modulations that dominate these natural stimuli (Fig. 1). This seems surprising because the efficient coding principle [2, 29, 30] would appear to suggest that the neural population should concentrate encoding resources on the more prevalent slow temporal modulations. The explanation may lie in our finding that a few slow features, i.e., combinations of MTFs dominated by slow modulations, suffice for good performance on basic discrimination tasks. This implies that relatively few neurons are needed to efficiently encode the spectral patterns seen at the lowest temporal modulations. More neurons could then be deployed to encode moderate temporal modulations which may better support fine discrimination, e.g. between two coos or two grunts. Indeed, the authors of [22] argue that encoding moderate temporal modulations may aid disambiguation of birdsongs as the difference between song temporal modulation spectra is maximal at intermediate modulation rates. Future work should test this hypothesis computationally, e.g., by thoroughly quantifying the effect of adding more features of intermediate temporal frequency in the generalization study of Sec. E for each vocalization class.

In our clutter results, we found that linear SFA generally outperformed quadratic SFA despite the fact that the quadratic features led to the best performance in the absence of clutter. Since linear SFA constructs features from temporally continuous stimulus components, while quadratic SFA additionally includes temporally continuous correlations, we attribute the difference in clutter performance to the selective use of linear features by the former. This suggests that the continuity of correlations between frequencies provides a useful signal about auditory stimuli that is masked or otherwise degraded in high-clutter situations. Thus, if the brain is using coincidence and continuity to learn to discriminate object identities, our results suggest that in the presence of clutter neurons should increase sensitivity to the temporal continuity of stimulus components relative to the continuity of correlations. This idea could be tested experimentally by calculating and comparing the first- and second-, order spectrotemporal receptive fields (STRFs) in cortex while varying the amount of clutter. A first-order STRF computes the reverse correlation between the neural output and the stimulus [25]. A second-order STRF computes the reverse correlation between the output and the same quadratic expansion of the stimulus that appeared in qSFA. STRFs are used to quantify the likelihood that a particular pattern in the stimulus generates a spike. Thus, if our hypothesis is correct, the second-order STRFs of neurons should become *less* predictive and/or the first-order STRFs of neurons should be come *more* predictive of neural response in the presence of clutter.

III. MATERIALS AND METHODS

A. Auditory Stimuli

We obtained rhesus vocalizations from a prerecorded library [31], and limited analysis to four ethnologically defined classes [13, 32] – coos, grunts, shrill barks, and harmonic arches. We chose these classes because they had the most unique exemplars, and thus we could generate the most training data for them.

This library encompasses a substantial fraction of the natural variability inherent in rhesus vocalizations [20], although it is not a complete sampling. Vocalizations were recorded at 50 kHz. To make background "clutter" we superposed 20 randomly selected vocalizations from the above library for the four indicated classes of vocalizations. Each token of clutter was three seconds long and the starting position of each vocalization in the clutter was randomly selected. Starting positions were always selected so that the entire vocalization was present in the clutter. Finally, each token of clutter was amplitude regularized to remove artificial peaks and troughs in the signal that could have biased classifier performance results; see Methods - Discriminating in background clutter. Random tokens of clutter were spot-checked by the research team to ensure quality. We generated white noise bursts with the `wgn` function from MATLAB 2020b's Communications toolbox. 500 ms bursts were generated at a sampling rate of 50 kHz at a power of -20 dBW (set by choosing -20 for the power parameter of `wgn`). Applause exemplars was downloaded from Dr. Joshua McDermott's website: https://mcdermottlab.mit.edu/downloads.html via the link for sound textures. These exemplars have a sampling rate of 20 kHz. All stages of the pipeline that depended on sampling rate were changed accordingly when applause was used.

B. Simple Cochlear Model

We implemented a simple cochlear model to ensure that the inputs to our algorithms approximated the structure of inputs to the macaque auditory system. This model consisted of two stages of filters: 42 gammatone spectral filters followed by a temporal filter that was inspired by previous work [2, 33, 34]. Each of the 42 filter outputs (42 spectral $\times$ 1 temporal filters) was normalized to have zero mean and unit variance.

The gammatone filters were implemented using the `GammatoneFilterbank` function from the python module `pyfilterbank.gammatone` (see Program Language and Code Availability for more details on accessing code). These gammatone filters had center frequencies between 22.9 Hz to 20208 Hz, which covered the range of rhesus hearing. The temporal filter was implemented as a difference of two kernels of the form:

$$g(n) = a\, n^m\, e^{-bn}, \tag{1}$$

in which n is in units of samples and a, b, and m are parameters. The temporal filter was created by taking the difference between g_1 and g_2 with the following parameters: g_1: $a = 1.5$, $b = 0.04$, and $m = 2$; and g_2: $a = 1$, $b = 0.036$, and $m = 2$. This parameter set accounted for some key aspects of cochlear temporal processing [34, 35]. Each filter output was normalized to have zero mean and unit standard deviation. Next, these normalized filter outputs underwent PCA, linear SFA, or quadratic SFA.

C. Discriminating between vocalizations

We tested the ability of the each algorithm to discriminate between pairs of vocalizations in a one-shot learning task. Vocalization pairs could be from the same (e.g., two grunts) or different (e.g., a grunt and a coo) vocalization classes. The order of the vocalization pair was selected randomly and a silent gap (duration: 5 – 50% of the duration of the first vocalization) occurred between the vocalizations to ensure the vocalization did not overlap in time.

We then applied our selected algorithm (*i.e.*, PCA or an SFA variant) to the vocalization pair and the "top" five features of the algorithm were taken. Top features for the SFA algorithms are the top five slowest features (*i.e.*, those with the smallest eigenvalues). The top features for PCA are the top five features that explain the most variance in the data (*i.e.*, those with the largest eigenvalues). For all algorithms, the top five features were found using singular value decomposition SVD, because of the greater numeric stability of SVD for solving generalized eigenvalue problems.

Next, the same pair of vocalizations was projected into the feature spaced defined by these five features. This projection generates a data cloud where each point in the cloud is a moment of one of the vocalization, thus we end up with a number of data points equal to the sum of the number of samples of each vocalization within the pair. We then trained a linear SVM classifier on a randomly selected 75% of the data points and tested performance on the held out 25%. To ensure performance was not an artifact of the particular 75% of the data we randomly selected, we ran a 30 fold cross validation procedure. We designed the cross validation procedure to guarantee that each vocalization in a pair was as evenly represented in the training and testing split as possible accounting for inherent differences in the duration of the vocalizations. To account for any fluctuations in performance not accounted for by the k-fold cross validation, we additionally repeated the above procedure five times and took the average performance across these five repeats. A new instantiation of PCA or an SFA variant was used for each of these five repeats so learning is not transferred across repeats. Note that this is one-shot unsupervised learning because the selected algorithm, i.e., PCA or SFA, only receives one presentation of the vocal pair and that the feature space generated is not modified or otherwise informed by the classifier.

This entire procedure was performed for each unique vocalization pair: 20 pairs in the first analysis of classification performance as a function of the number of features of quadratic SFA and 100 pairs in the second analysis of classification performance as a function of algorithm used.

To establish an upper bound on performance we repeated the above analysis, but applied the linear SVM directly to all 42 outputs of the cochlear model. For the null model, we randomly selected five outputs of the cochlear model to use as inputs to the SVM. The training and testing procedure for the SVM was otherwise exactly the same as it was for PCA, linear SFA, and quadratic SFA.

We utilized the same procedure for the analyses of applause and white noise by just replacing vocalization pairs with the

indicated stimulus. White noise was generated as described in Methods - Auditory stimuli. We generated 100 pairs of unique white noise tokens. Our library of applause exemplars was more limited so we generated 10 pairs of applause exemplars.

D. Discriminating in background clutter

To test discrimination performance in background clutter, we created clutter tokens by by superimposing 20 randomly selected vocalizations that did not include the test pair. Because we minimized the amplitude troughs of this mixture following [21], we reduced the possibility that classifier performance was merely due to test pairs that happen to occur at the time of an amplitude trough of the clutter . A test vocalization pair was generated as described in the previous methods section and then randomly embedded into this background clutter. We then conducted the one-shot learning procedure described above while varying the signal-to-noise-ratio (SNR) between the test vocalization pair and the background clutter. SNR was defined as the log of the ratio between the average powers of the test vocalization pair and was varied from -7 (essentially no signal) to 3 (essentially no background clutter). A new instantiation of PCA or SFA was utilized at each SNR level, so there is no transfer learning. We utilized the same cross-validation and repeats procedure described in the previous methods section.

E. Generalization to novel vocalization exemplars

We tested the ability of each algorithm to generalize to novel vocalizations using a holdout training procedure for each vocalization class. We took 19 exemplars from each vocalization class, and randomly created 20 pairs of vocalizations from these 19 exemplars. We then created unique training sets for each pair comprised of the 17 other vocalization, i.e., the vocalizations not contained in the test pair. The training sets were generated the same way as in previous sections which only utilized pairs, but with more exemplars. In other words, vocalizations were concatenated in time with a random order of exemplars and a random gap duration between each exemplar. Each algorithm was trained on the unique training set for a particular pair. As with the previous analyses only the top five features were kept. The test pair was then projected into this learned feature space. Finally we applied a linear SVM to the projected test pair to determine their linearly separability in this feature space, and then followed the same cross-validation procedure as in previous analyses. As before, the whole process was repeated five times for each pair to average over performance fluctuations not accounted for in the cross validation procedure. We calculated the average performance across these five repeats and analyzed performance as a function of vocalization class and algorithm. This entire procedure was performed for 20 unique pairs of vocalizations for each of the four vocalization classes. As before, a new instantiation of the algorithm was generated for each pair so there was no transfer learning.

F. Programming languages and code availability

Code for generating the stimuli was written in MATLAB 2019A and 2020B. The PCA and SFA algorithms and figure generation were written in Python 3.7. All code will be publicly available via a GitHub repository before final publication. For python code, we made efforts to include links to all requisite modules in the `ReadMe` file and/or include the module in the provided repository. Please contact the corresponding author at rdit@pennmedicine.upenn.edu with any questions regarding the code.

Acknowledgements This work was supported by NIH grant R01EB026945 to VB and YEC from the National Institute on Deafness and Other Communication Disorders, by the Computational Neuroscience Initiative at the University of Pennsylvania. We are grateful to Chethan Parthiban for collaboration in the early stages of this project.

[1] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound* (The MIT Press, 1990).
[2] M. S. Lewicki, Nature neuroscience **5**, 356 (2002).
[3] T. D. Griffiths and J. D. Warren, Nat. Rev. Neurosci. **5**, 887 (2004).
[4] S. A. Shamma, M. Elhilali, and C. Micheyl, Trends in neurosciences **34**, 114 (2011).
[5] L. Krishnan, M. Elhilali, and S. Shamma, PLoS Comput. Biol. **10**, e1003985 (2014).
[6] M. M. Van Wanrooij and A. J. Van Opstal, Journal of Neuroscience **25**, 5413 (2005).
[7] J. H. McDermott and E. P. Simoncelli, Neuron **71**, 926 (2011).
[8] J. H. McDermott, Current Biology **19**, R1024 (2009).
[9] B. Shukla, L. Mendel, and E. Buder, Under Review. The Journal of Speech, Language, and Hearing Research (2022).
[10] J. K. Bizley and Y. E. Cohen, Nat. Rev. Neurosci. **14**, 693 (2013).
[11] J. J. DiCarlo, D. Zoccolan, and N. C. Rust, Neuron **73**, 415 (2012).
[12] C. J. Darwin, Trends Cogn. Sci. **1**, 327 (1997).
[13] A. A. Ghazanfar and L. R. Santos, Nat. Rev. Neurosci. **5**, 603 (2004).
[14] Y. E. Cohen, F. Theunissen, B. E. Russ, and P. Gill, J. Neurophysiol. **97**, 1470 (2007).
[15] N. C. Singh and F. E. Theunissen, J. Acoust. Soc. Am. **114**, 3394 (2003).
[16] L. Wiskott and T. J. Sejnowski, Neural computation **14**, 715 (2002).
[17] H. Sprekeler and L. Wiskott, Available at SSRN 3076122 (2008).
[18] J. H. McDermott, M. Schemitsch, and E. P. Simoncelli, Nature neuroscience **16**, 493 (2013).
[19] B. G. Shinn-Cunningham, Trends in cognitive sciences **12**, 182 (2008).
[20] K. L. Christison-Lagay, S. Bennur, J. Blackwell, J. H. Lee, T. Schroeder, and Y. E. Cohen, Hear. Res. **312**, 128 (2014).
[21] D. M. Schneider and S. M. N. Woolley, Neuron **79**, 141 (2013).

[22] S. Woolley, T. E. Fremouw, A. Hsu, and F. E. Theunissen, Nature neuroscience **8**, 1371 (2005).
[23] H. Sprekeler, C. Michaelis, and L. Wiskott, PLoS Computational Biology **3**, e112 (2007).
[24] G. Bellec, M. Galtier, R. Brette, and P. Yger, Journal of computational neuroscience **40**, 317 (2016).
[25] M. A. Escabı and C. E. Schreiner, Journal of Neuroscience **22**, 4114 (2002).
[26] T. M. Elliott and F. E. Theunissen, PLoS computational biology **5**, e1000302 (2009).
[27] A. Hsu, S. M. Woolley, T. E. Fremouw, and F. E. Theunissen, Journal of Neuroscience **24**, 9201 (2004).
[28] A. Flinker, W. K. Doyle, A. D. Mehta, O. Devinsky, and D. Poeppel, Nature human behaviour **3**, 393 (2019).
[29] H. Barlow, Network: computation in neural systems **12**, 241 (2001).
[30] D. Burr and S. Laughlin, Current Biology **30**, R907 (2020).
[31] M. Hauser, Anim. Behav. **55**, 1647 (1998).
[32] K. L. Christison-Lagay and Y. E. Cohen, Front. Neurosci. **12**, 601 (2018).
[33] X. Zhang, M. G. Heinz, I. C. Bruce, and L. H. Carney, J. Acoust. Soc. Am. **109**, 648 (2001).
[34] S. Tabibi, A. Kegel, W. K. Lai, and N. Dillier, Journal of neuroscience methods **277**, 63 (2017).
[35] R. F. Lyon, A. G. Katsiamis, and E. M. Drakakis, in *Proceedings of 2010 IEEE International Symposium on Circuits and Systems* (IEEE, 2010) pp. 3809–3812.