Gradient boosted decision trees reveal nuances of auditory discrimination behaviour

Carla Griffiths[1], Jules Lebert[1], Joseph Sollini[1,2], and Jennifer Bizley[1]

[1]**UCL Ear Institute, 332 Grays Inn Rd, London WC1X 8EE**
[1,2]**Hearing Sciences, University of Nottingham, University Park, Nottingham, NG7 2RD UK**

Corresponding author:
Jennifer Bizley[1]

Email address: j.bizley@ucl.ac.uk

ABSTRACT

Perceptual constancy remains an elusive feature to measure in psychophysical animal studies due to the intractability of correlating trial parameters to trial outcomes and response times. Moreover, the nonlinear nature of our data constrains the potential that traditional statistical models have to elucidate the sensory and non-sensory factors that ultimately shape animal performance. In this study, we present an implementation of gradient-boosted regression and decision trees, a machine learning method, to gain insights into animal behavioural data. We implement this computational modelling method on a task in which 5 ferrets were trained to recognize a target word ("instruments") against a stream of 54 other English non-target word tokens. To test perceptual constance to pitch, we manipulated the mean fundamental frequency (F0) of each trial across the whole series of word tokens (inter-trial F0 roving) or between word tokens (intra-trial F0 roving). We implemented Shapley Additive Explanations and permutation importance for model explainability to determine which factors were significant in our gradient-boosted regression and decision tree model performance. We found that, despite high performance on both intra-trial F0 roved and inter-trial F0 roved task paradigms, F0s that our subjects were introduced to after their initial control F0 period increased the rate of false alarms, and an F0 difference between the target word and the preceding non-target word caused slower reaction times. We also found a subset of non-target distractor words that increased the rate of false alarms and that shared spectrotemporal properties with the target word. Together, our results demonstrate that ferrets, like humans, are capable of showing tolerance to changes in pitch when discriminating complex sounds and that they rely on pitch as a cue to link sounds together over time. Finally, we illustrate the utility of gradient-boosted decision trees to gain insights into behavioural datasets.

Keywords: Shapley Additive Explanations;auditory scene analysis; pitch; ferret, behavioural data analysis in neuroscience;

INTRODUCTION

Perceptual constancy – the ability to recognise objects across identity-preserving variation - is a fundamental feature of sensory perception. Humans can recognise spoken words across considerable spectrotemporal diversity, including temporally stretching or compressing the word, changing pitch, timbral changes, and in background noise (Aulanko et al. 1993) (Haykin and Chen 2005). The ability to identify and discriminate sounds across pitch is likely to be a fundamental property of mammalian audition, as the pitch of a vocal call conveys information about an individual's size, age and emotional state (Hauser 1993) (Charlton, Zhihe, and Snyder 2009). Psychoacoustical studies have demonstrated that, like humans, some animals can generalise across variations in fundamental frequency (F0, the physical stimulus attribute that governs pitch perception). For example, ferrets trained to discriminate artificial vowel sounds with an F0 of 200 Hz maintain their performance at F0s from 150 to 500Hz (Bizley et al. 2013), Town et al., 2015). Both rats (Engineer et al. 2013) and zebra finches trained to discriminate human speech sounds can generalise across different talkers (Ohms et al. 2010).

While speech recognition is robust to variation in voice pitch for non-tonal languages, humans use the pitch of complex sounds to separate simultaneous competing sounds, and to link sounds together over time to form auditory 'streams'. Auditory streaming has been studied in many species, including frogs (Bee and Riemersma 2008), starlings (Bee and Georg M. Klump 2004) (Hulse, MacDougall-Shackleton,

Figure 1. Schematic of the experimental booth. To trigger a trial, ferrets had to nose-poke a centre spout. Catch trials only required the ferret to stay at the centre spout to receive a water reward. In contrast, target trials relied on the ferret to release from the centre spout and lick the lateral spout corresponding to the side of the stimulus presentation (in this example, the left-hand side of the booth) within 2s of the target word being presented.

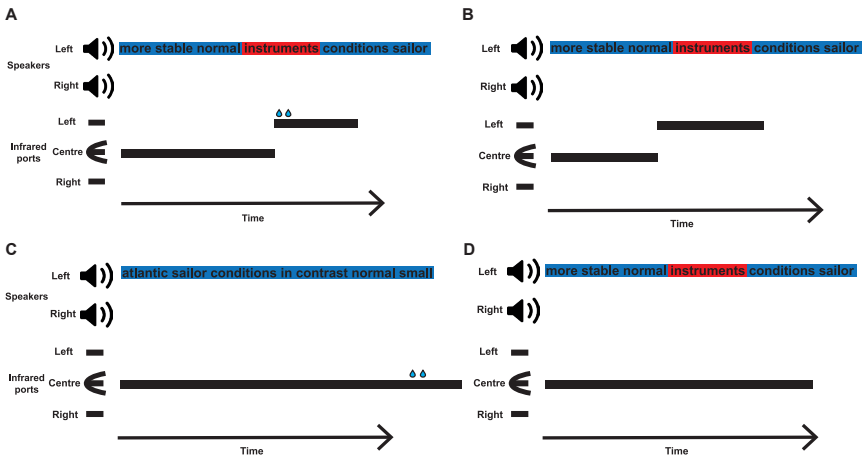

Figure 2. Schematic of trial options in the Go/No-Go task. In catch trials, the target word is not presented. Black rectangles represent the position of the ferret. A, a correct hit response where the side of the audio presentation is the left speaker; B, a false alarm; C, a correct rejection; D, a false alarm during a catch trial

and Wisniewski 1997) and gerbils (Dolležal et al. 2020). Evidence from birds suggests that avians use similar strategies to humans, with differences in intensity and spatial location used to segregate sounds but a greater tolerance to changes in frequency or timing (Dent et al. 2016). Ferrets can also detect the presence of 'mistuning' when a single component of a harmonic complex is shifted in frequency, suggesting that, like humans, harmonicity is a strong grouping in animals (Homma et al. 2016). However, to our knowledge, no one has assessed whether non-human listeners use the pitch of a complex sound in the formation of auditory streams.

Here, we trained ferrets in a task that required they discriminate spectrotemporally complex speech sounds in order to detect a target (the word "instruments") presented within a stream of other words (Sollini and Bizley, in prep.). Ferrets were trained on a male voice and a female voice before variation in F0 was introduced by using STRAIGHT (Kawahara 2006) to pitch-shift words. The whole stream could be shifted upwards or downwards in pitch, to test for perceptual invariance to pitch changes. The F0 of each word within a stream could also be randomly shifted to assess whether pitch contributes to streaming. We predicted that if the common pitch was a streaming cue, this manipulation may decrease performance even if ferrets were tolerant to F0 changes in the target. We used machine learning approaches to analyse the 20,487 trials of behavioural data in order to determine which factors influenced the ferrets' performance and reaction times. We find that while performance was robust to changes in pitch, shifting the F0 of words within a trial significantly slowed reaction times, providing evidence that ferrets, like humans, use pitch to form perceptual streams. We present a novel way to analyse this data by borrowing from a classical machine learning approach, gradient-boosted regression and decision trees and applying these methods to classical behavioural neuroscience data.

MATERIALS AND METHODS

Animals

Subjects were five pigmented ferrets (*Mustela putorius*, female), which started training from 6 months of age and were tested between 18 months and 4 years of age. Animals were maintained in groups of 2 or more ferrets in enriched housing conditions, with regular otoscopic examinations to ensure the cleanliness and health of ears. All animals were trained in behavioural tasks using water as a reward, while cage access to water was regulated. During water regulation, each ferret was tested twice daily from Monday to Friday, with free access to water from Friday afternoon to Sunday afternoon. Each ferret received a minimum of 60 ml/kg of water per day, which was achieved during task performance and supplemented with a wet paste made from water and ground high-protein pellets. Each ferret's weight and water consumption were logged daily throughout the experiment. All experimental procedures were approved by local ethical review committees (Animal Welfare and Ethical Review Board) at University College London and the Royal Veterinary College, University of London, and performed under license from the UK Home Office (Project Licenses PP1253968, 70/7267)

Equipment

We controlled the task and stimulus presentation through an RZ6 controller (Tucker Davis Technology, Florida, USA) using OpenEx and MATLAB (Town et al. 2015) on a windows PC. Both the right and left-hand speakers were calibrated to match the sound levels using a Bruel & Kjaer measuring amplifier (Type 2610). We presented each trial at a mean sound level of 65 dbA; stimuli were scaled to be constant in sound level across trials and talker types.

Stimuli

The sampling rate of both the female and male voice IDs was at 24,414 Hz and was taken from the SCRIBE database (audio recorded at 20,00 Hz). Word tokens were extracted from continuous speech manually.

Task

In a sound discrimination task, we trained five ferrets to recognise the target stimulus (the word 'instruments') against 54 other non-target stimuli (which were also English words). Once trained animals were required to initiate a trial at the center and maintain contact with the central spout until the target was presented, which could occur anytime between 0.5 and 6.5 seconds. The whole trial was comprised of word tokens from the same talker, and presented from either the left or right speaker. Once the target sound was presented, they were required to move to the response spout on the same side as the stimulus presentation. The target word was embedded within a stream containing a variable number of non-target words that occurred before and after the target. A correct response required that the animal released the center spout within 2s of the target word onset and correctly lateralised the sound stream. Catch trials (25%) contained only non-target words and were constructed to be equal in duration to the non-catch trials. On catch trials the animal was required to remain at the center port.

Training

Ferrets were first acclimatized to responding to the word 'instruments' by leaving the center port when it was presented to make a response to the side that the target was presented from. Initially the word 'instruments' was presented in a repeated manner until the ferret responded. Once the ferret was capable of waiting 5 seconds at the center spout for target presentation and accurately lateralising the stimulus, we reduced the target presentation to a single-word token and then gradually introduced non-target words before and after the target. Non-target words were initially presented with a 60 dB attenuation cue that was gradually reduced until animals were performing with the target and non-target sounds at an equivalent sound level. 3/5 animals were trained first on the female and then the male, whereas F2105 and F2002 were trained with both from the beginning of training. All word tokens within a trial were drawn from the same talker, but across trials, the talker identity was randomly drawn. Even once trained, we included a proportion of trials (25-50 %) that included a 10 -20 dB attenuation cue. These trials were excluded from the analysis but helped maintain the animals' motivation to perform the task. 25% of trials were catch trials in which the target word was not presented, and each correct rejection was rewarded with approx. 0.5 ml of water through the nose-poke centre spout (the same spout where ferrets initiated each trial). Baseline training varied in duration from 3 months to 8 months.

Pitch Roving

Animals were considered fully trained once they consistently performed above 70% correct on 0 dBA trials (chance performance is approximately 33% given the 6s trial duration and a 2s response window, i.e. 2s

/ 6s = 1/3). Once trained on the natural ('Control') F0 trials, we introduced F0 (pitch) roving. For each talker, we used STRAIGHT to shift the F0 up or down by 0.4 octaves. This resulted in F0 values of 109 and 144 for the male voice, where the natural F0 was 124 Hz and values of 144 and 251 Hz for the female voice, where the natural F0 was 191 Hz.

In inter-trial roving, the pitch of the entire trial was shifted up or down, whereas in intra-trial roving the F0 value of each word within the trial was randomised. As in training, all word tokens within a trial came from the same talker.

Data Analysis

Any trial has four possible trial outcomes: hit, correct response, miss, and false alarm. We define a hit as releasing from the centre spout within 2s of the target word presentation, a correct rejection as staying at the central spout for the entire duration of a catch trial where no target word is presented, a miss as failing to leave the central spout within 2s of the target word presentation, and a false alarm as releasing from the center spout before target word presentation or the end of a catch trial. We define $p(hit) = n\ hits/(n\ hits + n\ misses)$, and the proportion of false alarms (FA) as $p(FA) = n\ false\ alarms/[n\ hits + n\ misses + n\ correct\ rejections + n\ FA]$. We consider correct responses (CR) as either a hit or a correct reject, where $p(correct) = [n\ hits + n\ correct\ rejections]/[n\ hits + n\ misses + correct\ rejections + n\ FA]$. We also calculated a sensitivity metric (d') (Green and Swets 1966), where $d' = z(p(hits)) - z(p(FA))$, where z represents the normal distribution function. We define reaction time as the central port drop time rather than the peripheral response time relative to the timing of the target word. For analysis of whether word tokens systematically elicited behavioural responses we defined the response time as the time of exit from the central port, relative to trial onset. All data analysis, from behavioural metrics to computational models, was programmed using Python 3.9.

Computational Models

The general approach of the gradient-boosted decision tree model is a form of ensemble learning in which we use an initial weak decision tree of a depth larger than 1 to predict an outcome of a trial based on our behavioural data and then iteratively build upon the error of the first tree (after calculating the loss) by constructing the next tree based on the residuals of the previous tree. Once our loss plateaus or we reach the maximum number of training epochs, we stop training the model and calculate our test accuracy, or how well the model was able to predict our target variable on a held-out test set of data. We chose this method as our data is inherently dense (from long periods of behavioural training and testing) and tabular, which makes gradient-boosted regression and decision trees an excellent candidate for the prediction of categorical and continuous data compared to a nonlinear neural-network-based classifier (Grinsztajn, Oyallon, and Varoquaux 2022).

Linear mixed effect models, and generalized linear models are commonly used alternatives that can allow trial-based analysis of categorical or continuous behavioural data. While powerful, such models can fail to capture non-linear or non-monotonic relationships that might be present in behavioural data. Machine learning approaches offer an alternative model-free approach to uncovering statistical structure in rich behavioural datasets such as those typical of animal behavioural work **??**.

Models were generated using a gradient-boosted regression tree to predict the reaction times relative to the presentation of the target word and absolute reaction times relative to the trial start time using the LightGBM package (Ke et al. 2017). Using LightGBM, we also used gradient-boosted decision trees for the false alarm/correct reject model and the miss/hit model. Implementing gradient-boosted regression and decision trees enables us to capture the nonlinear relationships between behavioural variables and achieve a high degree of test accuracy for model data compared to mixed-effects linear models. To optimize hyperparameters for this model, we implemented a grid search using optuna (Akiba et al. 2019). We define a classification model as the model that predicted whether ferrets would make a correct hit response or miss response after hearing the target word (miss/hit model), and the model predicting whether ferrets would make a false alarm based on the trial characteristics (false alarm/correct reject model). Regression models predicted the reaction time for correct hit trials (reaction time model), and the two models that predicted the absolute release time solely based on the word presentation time of all the tokens in the word bank for both the female and male talker trials (response time models) (each talker type had its own model).

We then determined which features were significant using permutation testing, which shuffles a feature of our data (e.g., the target F0) and then determines the drop in performance the model has as a result of that feature being shuffled. The classification models were tuned using binary log loss with an evaluation metric of binary log loss across 10,000 epochs and implemented early stopping of 100 epochs. The regression models implemented the l2 loss function over 1000 epochs with an early stopping of 100 epochs.

For the classification models, all hyperparameter optimization minimised binary log loss, whereas for the regression (reaction time) model, hyperparameter optimisation minimised the mean-squared error (l2

loss function). For the regression models, the test and train mean-squared error was calculated using 5-fold cross-validation, and for the classification models, the train and test accuracy and balanced accuracy were calculated using 5-fold cross-validation.

To force the model to weight trial types with equal importance, we sub-sampled control F0 trials to match intra and inter-F0 roved trials. To weigh the trial outcomes with equal importance, we sub-sampled hit responses to match the number of miss responses for the miss/hit classification model and sub-sampled non-false alarm responses to match the number of false alarm responses in the false alarm model.

We then used Shapley Additive values to assess parameter influence on the trial outcome. For the classification models, this was the likelihood of a miss/hit or false alarm; for the regression models, this was the reaction time.

To extract the significance of the model features (i.e., feature importances), we used the SHAP package (Lundberg and Lee 2017), an implementation of Shapely Additive Importance features, to elucidate explainability from the typically 'black-box' regression and classification tree models. The SHAP package allowed us to plot partial dependency plots to see how the impact of the model would vary as inter-related features changed (such as talker gender and trial number). We generated permutation importance plots from the sci-kit learn (sklearn) package to extract the model's features that would decimate the model's performance if they were shuffled, thereby establishing which features contributed significantly to model performance. All code used to perform this analysis is available on GitHub (link will be inserted on publication).

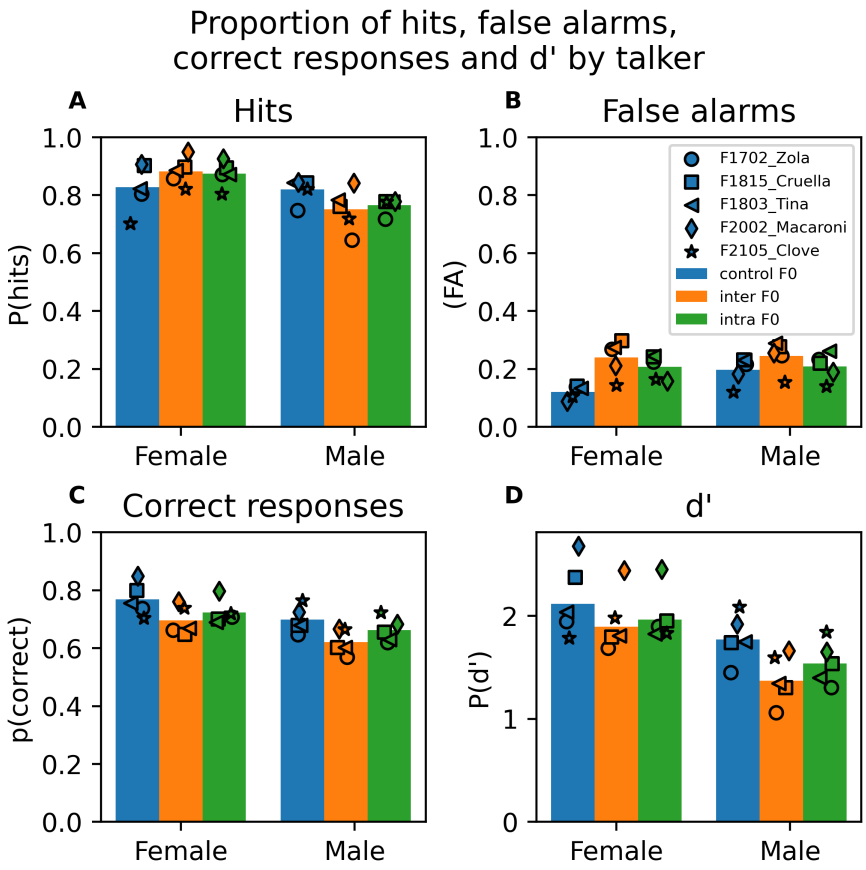

Figure 3. Behavioural metrics across animals distributed by talker type. Bars indicate the across-animal average; symbols show the individual animals. Trials are separated according to the identity of the talker and the pitch roving condition (control = no pitch shifting, inter = F0 shifting of the whole trial, intra = F0 shifting of the tokens within a trial). A, hits B, false alarms. C, the proportion of correct responses across trial types. D, sensitivity index (d').

RESULTS

Ferrets performance during target discrimination robust to pitch shifting

Ferrets were trained to detect the target word "instruments" within a stream of randomly drawn non-target word tokens drawn from a pool of 54 (per talker). On each trial, all the words in a stream came from the same talker and the same position in space. Ferrets were initially trained with a single male and single female voice, and once their performance was stable, trials in which the F0 of the whole trial was shifted ('inter-trial roving') or individual word tokens within the trial were shifted ('intra-trial roving'). After a training period on the natural (control) F0, male and female talkers, the ferrets were switched to interleaved intra-trial and inter-trial F0 roved trials, in which the F0 of each word token could be: 109 Hz, 124 Hz, or 144 Hz for the male talker, and 144 Hz, 191 Hz, and 251 Hz for the female talker. Overwhelmingly, ferrets performed above-chance across the board in both the recognition of the target word and rejection of non-target word stimuli (Figure 3).

We will first provide an overview of the data before using Gradient Boosted decision trees to understand and quantify the factors that shape the animals' performance in this task.

As shown in Figure 3, hit rates were high for both talkers and for both types of pitch-shifted trials, with overall hit rates being higher for the female talker. False alarms were generally low, and equivalent for inter and intra-trial pitch roving, although the trained female talker's natural F0 elicited a lower false alarm rate. Together these effects meant that d' values were highest for both trained talker with a small decrease evident for pitch-roved trials (Figure 3D). Nonetheless, all d' values were well above 1, indicating the animals were well able to perform the task. This is confirmed by simply assessing the outcome of the trials in terms of their probability of being correct which was above 60% for all conditions, where 33% would be considered chance performance (Figure 3C).

When pitch-roved trials were broken down according to their F0 it was apparent that the hit rate increased with F0 such that the highest hit rates were observed for the highest (female talker) F0 trials (Figure 4A). False alarms, in contrast, were apparently lower for the trained F0 values for both the male

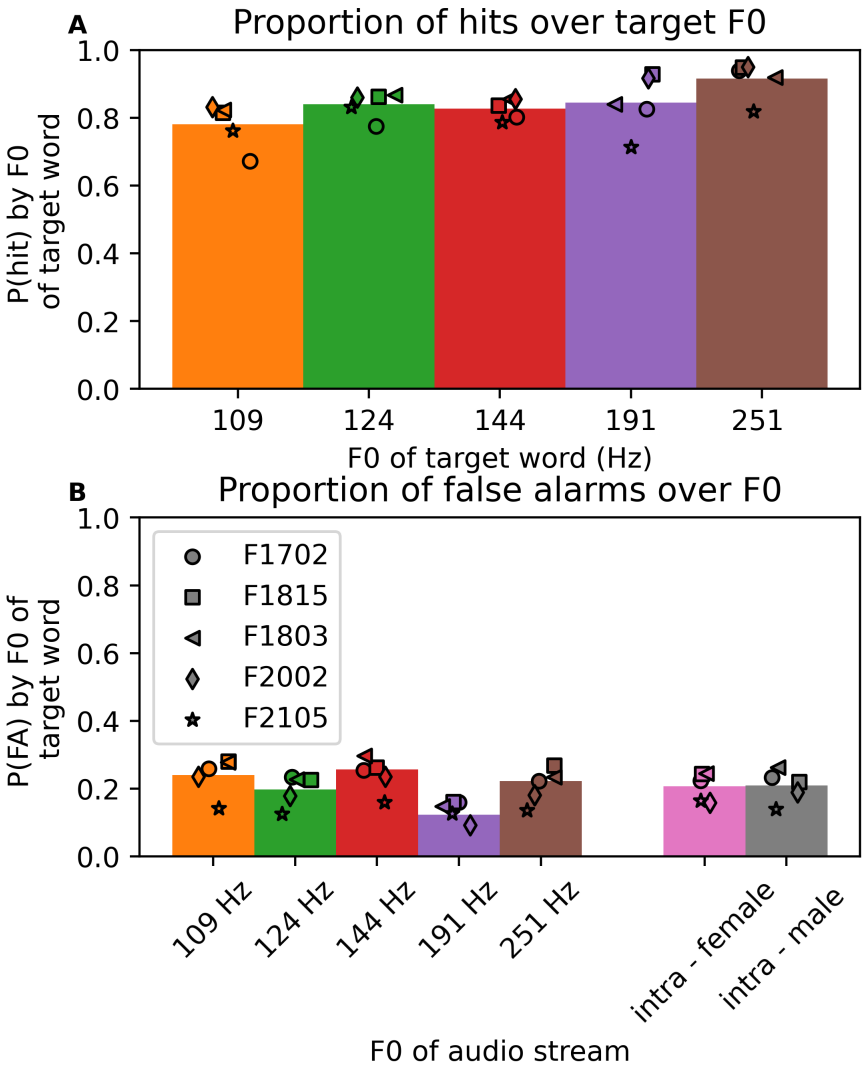

Figure 4. Behavioural metrics of hits and false alarms divided by F0. a, proportion of hits over the F0 of the target word. b, proportion of false alarms over the F0 of the stream for inter-trial roving, and whether the trial had a male or female talker for intra-trial roving.

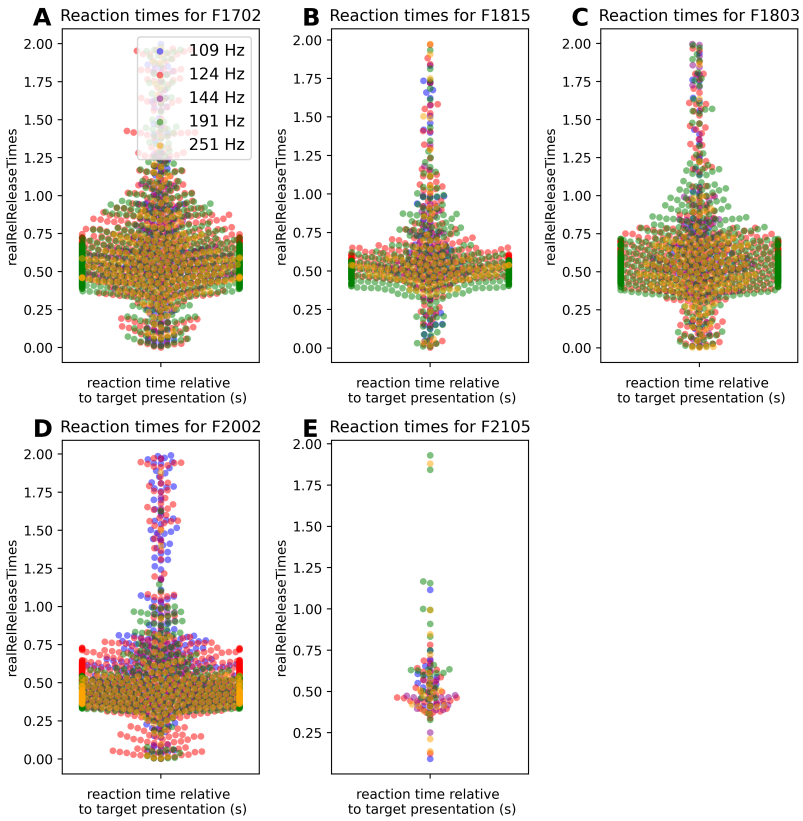

Figure 5. Reaction time distributions by subject for trials in which the target word was presented, correct responses are classified as being within the -2 to +2s range. Each dot is a trial and is coloured by the F0 of the target. A, F1702; B, F1815; C, F1803; D, F2002;, E, F2002; F, F2105

and female talker and intra-trial roved trials eliciting equivalent false alarm rates to inter-trial roved trials (Figure 4B). Reaction times for correct hit responses were centred at 0.5s post-target onset for all subjects, with the targets from the natural female F0 eliciting the fastest reaction times (Figure 5). The trend for lower hit rates at lower F0 and for the female voice to elicit faster reaction times may be a consequence of training, as 3/5 subjects were initially trained on only the female talker. However, while the hearing range of ferrets fully encompasses that of humans, their frequency resolution is poorer and most notably so at the lowest audible frequencies (Sumner et al. 2018), and this effect may negatively contribute to performance for the lowest F0s.

These data show that ferrets can successfully discriminate a target word from non-target words despite variation in F0. While these basic behavioural metrics provide compelling evidence that animals can perform this task, they do not allow us to explore how different aspects of the task influence performance. To consider this in more detail, we next used gradient-boosted decision trees to understand how various factors determine a trial outcome.

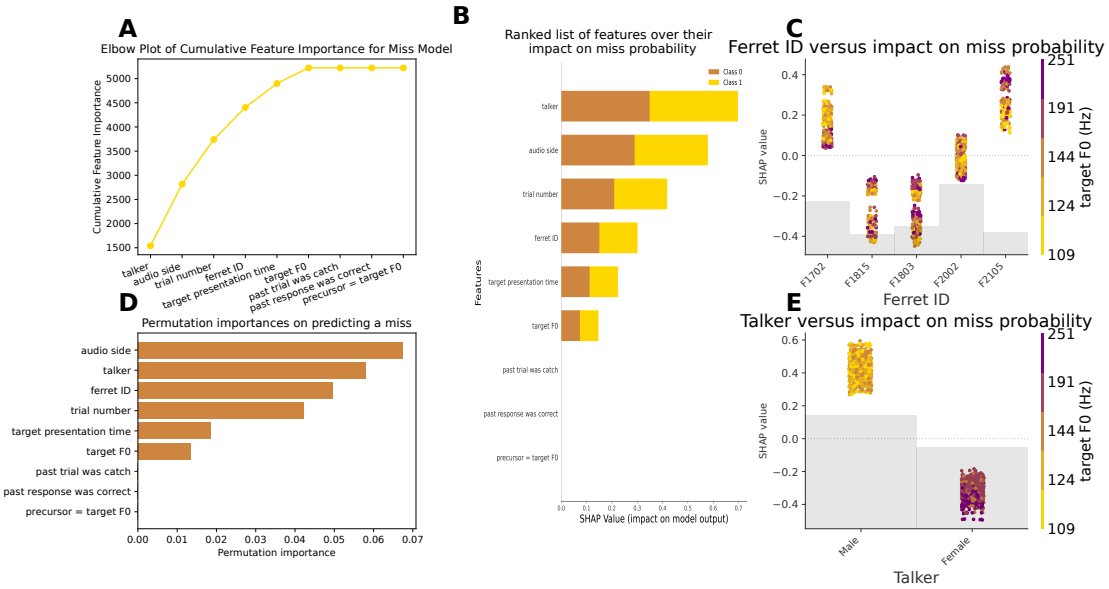

Figure 6. Factors correlated with within-ferret trial parameter preferences drive the miss/hit model; a) the elbow plot of cumulative features over trial features; b) SHAP feature importances of the miss/hit model; c) SHAP partial dependency plot showing the SHAP impact over each precursor F0 color-coded by the time elapsed since the beginning of the trial; d) SHAP partial dependency plot depicting the SHAP impact over each ferret ID color-coded by whether the precursor equalled the target F0.

Target detection is robust across F0

We selected trials in which the animal heard the target sound from those in which the target was not presented, either because they were catch trials or because the animal false alarmed to a word prior to target presentation. Using these trials, we modelled the likelihood of a hit versus a miss using the following: the talker (male/female), the side of the audio presentation, the trial number, the subject ID, the target presentation, the target F0, whether the past trial was a catch trial, whether the past response was a correct response, and whether the F0 of the precursor word token equalled the target F0. We were particularly interested in teasing apart the contribution of the F0 of the target word, the identity of the talker, and whether the model provided evidence for worse performance when the pitch was roved from token-to-token. The performance of the miss/hit model was relatively high despite the sparsity of misses in the behavioural data, with an average balanced accuracy on our training set of 63.22% and an average test balanced accuracy of 61.50%. We considered the following factors to be significant contributors to the model based on whether they contributed a significant increase in the cumulative feature importance plot (A) and if a permutation test that randomised the variable in question showed a significant change in model performance. By these metrics, we excluded the trial history factors (the past trial was correct, and the past trial was a catch trial) and the precursor=target F0 parameter. Out of the factors we incorporated into the model, the talker, the audio side, the ferret ID, the trial number, the target presentation time, and the target F0 were significant features (Figure 6 d).

The top three features that drove the miss/hit model prediction were the talker (with male talker trials increasing the probability of a miss), the side of the audio presentation (which was idiosyncratic across animals, likely reflecting their own individual biases, see supplemental information), and the trial number, with trials occurring later in the session being associated with higher miss rates (Figure **??** a,b,d). Target presentation time, the fourth most significant feature, was not particularly correlated with a higher miss probability unless the target presentation was greater than 5s, where there was a sharp increase in miss probability from 0.2 to 0.8 in 5/5 ferrets. The feature importance metrics, permutation tests, and SHAP feature values were all in concordance with each other, with only minor differences in the order of the top two features. This miss/hit model illustrates that the target word's pitch had a modest influence on the ferrets' performance: only 2/5 animals had stratified miss probabilities based on target pitch (visible as a smooth colour continuum Figure 6c), and these were in the opposite direction to each other. We conclude that ferrets are able to recognise the target word robustly across different F0 values and are no worse at doing so when the F0 of the target word differs from the preceding non-target word.

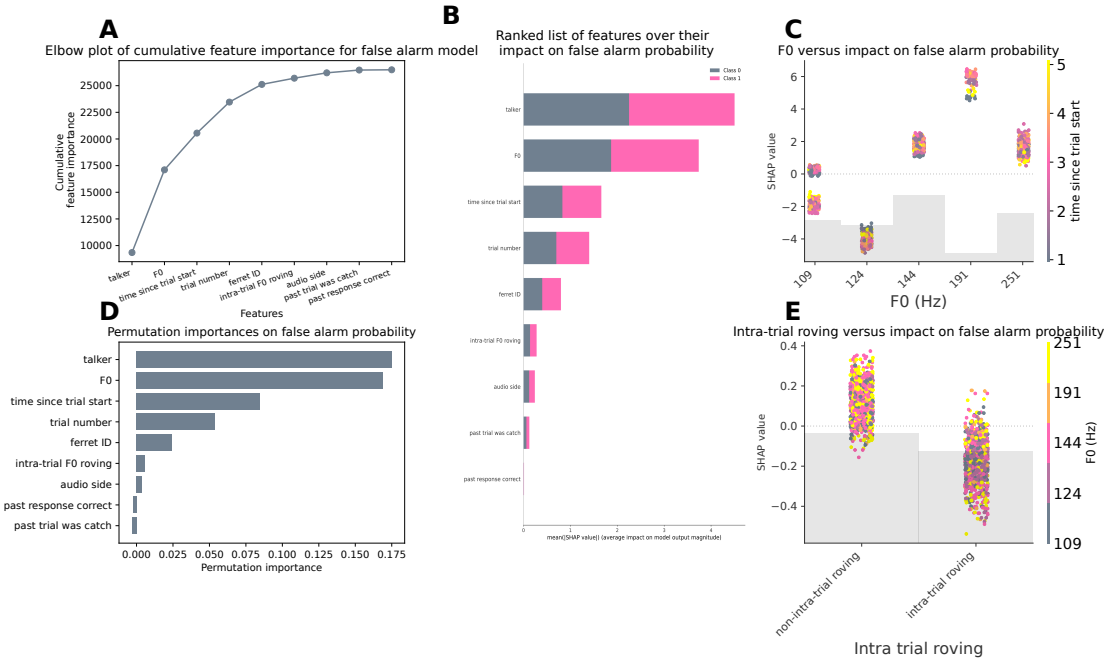

Figure 7. Precursor F0 determines the probability of a false alarm; a) elbow plot depicting the cumulative feature importance of each factor used in the false alarm decision tree model; b) SHAP feature importance plot for the same factors as in a); c, partial dependency plot showing the SHAP value (representing the impact on the probability the trial would be predicted as a false alarm) over ferret ID colour-coded by the precursor F0; d, permutation importance plot (100 shuffles) of the model features of the false alarm model as in a) and b); e, partial dependency plot depicting the SHAP value over the ferret ID colour-coded by whether the trial had intra-trial F0 roving.

False alarms are influenced by F0

Next, we considered the trials in which the target was not presented in order to understand what factors influenced the likelihood of the animal making a false alarm response. We developed a gradient-boosted decision tree (classifier) model to predict whether a subject's trial outcome would be a false alarm based on the following features: the pitch (F0) of the word immediately preceding the target word, known as the precursor word, the trial duration, the time elapsed since the start of the trial, the trial number within the experimental session, the ferret ID, the side of audio presentation, the talker type, whether the past response was correct and a catch trial, and whether there was intra-trial F0 roving. To consider the impact of pitch, we coded the F0 of the word preceding the false alarm and a parameter that coded whether or not the pitch varied within the trial (i.e. whether the trial was intra-trial F0 roved). Out of these features, the following were significant: the talker, the non-target F0, the time elapsed since the trial started, the trial number, the ferret ID, the side of audio presentation (audio side), whether the outcome of the n-1 trial was correct (past response correct), and whether the trial was intra-trial F0 roved (Figure 7 a,b,d). Generally, our false alarm categorical model performed well, with a balanced train accuracy of 81.04% over 5-fold cross-validated and a balanced accuracy on our test data set of 71.24%.

In common with the hit/miss model the talker was the strongest determinant of the likelihood of false alarming, with false alarm rates being higher overall for the male talker (Figure 7A, B). In contrast to the hit/miss model, however, the F0 of the non-target words had a sizeable effect, and whether the trial was intra-trial roving also had a significant, if small, effect, whereby intra-trial roved trials were less likely to elicit false alarms. Trial timing parameters also influenced false alarm rates, with the time elapsed since the start of the trial having a significant influence (longer trial times were associated with a greater likelihood of a false alarm outcome in 5/5 subjects) and trial number was also a significant factor, with early trials in the session were more likely to be associated with a false alarm in 3/5 subjects. The side of the audio presentation was also predictive, but again idiosyncratic across animals.

The above model suggests that while the performance difference for male and female talkers drives both models, the effects of F0 are somewhat dissociable between hits, where F0 has a minor role and false alarms, where F0 has a stronger influence. It is possible that the high levels of performance on hit trials mask the effects of F0 on the task. In the following analysis, we consider whether the animal's reaction time is modulated by acoustic and task factors.

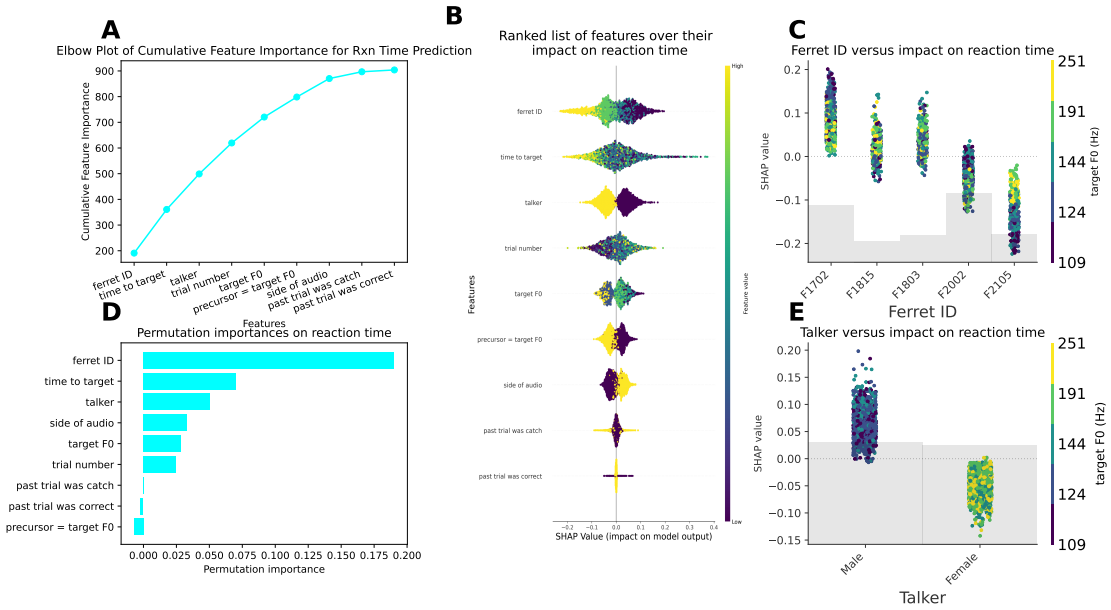

Figure 8. Reaction time models establish a contribution of F0 to target detection. a, feature importances of the hit model; b, SHAP summary plot of ranked feature SHAP values of each factor in the reaction time model; c, partial dependency plot of SHAP impact versus ferret ID color-coded by target F0; d, permutation feature importance of each factor in the model; e, partial dependency plot of SHAP impact color-coded by the target F0.

To explore whether reaction times provided a more sensitive measure of how acoustic and task parameters influenced performance, we modelled the animals' reaction time on trials in which the target was successfully detected (i.e. the target was presented and the animal correctly exited the central port within the 2s response window).

Lower-frequency F0s correlated with slower reaction times

In our reaction time model, we considered the following factors: the ID of the subject, the talker type (male or female), the time to target presentation (within a trial), the trial number (within a session), the side of audio presentation, the target F0, whether the F0 changed from the preceding non-target word to the target word (preceding F0 = target word F0), whether the past trial was a catch trial, and whether the past trial was correct. Our test-set mean squared error (mse) using 5-fold cross-validation was 0.0929s, or alternatively, an error of +/- 0.3s, a relatively high degree of accuracy (train mean-squared error = 0.0854s). From our permutation test, the ferret ID, the talker, the side of audio presentation, the time to target presentation, the target F0, and whether the precursor word equalled the target word were significant factors in this reaction time model (Figure 8 d). Similar to the miss/hit and false alarm/correct reject performance models, both ferret ID and talker ID accounted for the greatest variance; reaction times were longer for the male talker (for whom hit rates were also lower) and, like performance, varied across ferrets (Figure 8b). Other factors that significantly predicted reaction times were the time of the target within a trial (later targets had faster responses), the side of the audio (left responses were slightly faster than right responses in 2/5 ferrets, right faster than left in 2/5 ferrets, 1/5 did not differ). The model was able to dissociate the effects of talker and F0, with the across animal model showing that the natural male F0 had shorter reaction times after the effect of talker (in which male voice trials were overall slower) was taken into account. Breaking down the effect of F0 by ferret (Figure 8c) shows that the effect of target F0 on performance is variable across animals. Reaction times were faster when the preceding non-target word had the same F0 as the target. Factors that did not influence reaction times - as assessed by the permutation test were trial number (although early trials tended to have shorter reaction times) and trial history (the previous trial was a catch trial / previous trial was correct). Therefore, despite equivalent performance in inter and intra-trial roving trials (Figure 3), these data support the idea that ferrets use the identity of the talker, the F0 of the talker, and the consistency of the F0 within a stream interact to shape the animals' reaction times (Figure 8 b and e). In summary, reaction time analysis supports the idea that ferrets use F0 to link sequential sounds into a stream which facilitates target recognition.

Gradient boosted decision tree models demonstrate animals consistently false alarm to a subset of non-target words

Finally, we asked whether subjects consistently false alarmed to particular non-target words. Since the reaction time model clearly demonstrates that multiple factors influence animals' reaction times, and word tokens are presented sequentially in a stream, ascertaining patterns of false alarms is nontrivial (i.e. there will be multiple word tokens in the 2s window that preceded a false alarm and the lack of a fixed response latency to something the animal considers a target limits reverse correlation methods). To tackle this, we exploited the power of the machine learning approach to model the animals' reaction time relative to the onset of the trial based on the timing of word tokens within the trial. We used all trials in which animals left the central port (i.e. hits and false alarms). Reaction time was defined relative to trial onset rather than to target presentation. We separated out trials according to the talker and ran independent models on male talker and female talker trials. Our parameters for the models were the time of each of the 54 non-target words within each trial (although on any given trial, only 1-15 tokens were presented). The prediction accuracy of this model was high across talker types, with a train mse of 0.1329s and a test mse of 0.1331s for the female talker, and a train mse of 0.1301s and a test mse of 0.1312s for the male talker (5-fold cross-validation for both train and test metrics). This high model performance is understandable as the ferrets are well-trained (Figure 3), so we would expect the majority of releases from the centre spout to be highly correlated with the presence of the target word ('instruments', Figure 9b, 10b). Reassuringly, for both male and female talker models, the key driver was the time of the target word presentation. Nonetheless, both feature importance and permutation test metrics illustrate that some words consistently elicit behavioural responses, suggesting that false alarms are not only temporary lapses in attention but that some words are perceived as more similar to the target.

To examine whether all animals false alarmed to the same words we additionally ran the model on each animals' data separately and looked across each animals most predictive words (Figure 9c,10c). These demonstrated results consistent with the all-animal model but also showed some variation across animals. The most false-alarmed to words for the female talker were 'of science', 'in contrast', and 'more stable', whereas for the male talker words 'exposure', 'of science', and 'that generations.' Only 1/3 top three words had a similar duration for both female and male talkers ('of science' and 'that generations', respectively). For the female talker, 'More stable' has a similar frequency spectrum to the target word, whereas 'in contrast' has a similar amplitude waveform to the target word. For the male talker, both 'exposure' and 'of science' cover large swathes of frequency areas, which could easily be confused with the frequency pattern of instruments. However, no unifying feature unites all of the frequently reacted-to-word tokens; thus, our subjects are using complex spectrotemporal properties to decide which word is the target versus solely memorising amplitude waveforms or word timings.

DISCUSSION

We present a novel behavioural paradigm to test whether ferrets can form an invariant representation of a complex sound that is tolerant to changes in F0. We chose English words as our stimuli because words are spectrotemporally rich, have known statistical properties and are likely to have relevance to our animal model as a domesticated companion and working animal. While human speech is undoubtedly more complex than ferret vocalisations, for such a domesticated animal, listening and responding to human vocalizations is not unnatural; ferrets are often trained on tasks in their home environment that are initiated using a human vocalisation (Dancer et al. 2022). Using gradient-boosted regression and classification trees, our data are consistent with ferrets building an F0-tolerant representation of the target word and with their use of F0 to link sounds together over time into a stream.

The factors influencing animals' ability to identify the target accurately were at least partially dissociable from those contributing to false alarms. For target identification (i.e. a hit versus miss), the F0 of the target had only a minimal (if significant, SHAPS score 0.15, Figure 6d) effect, and the F0 of the preceding words did not impact accurate detection. In contrast, F0 was a significant determinant of the likelihood of false alarming (SHAPs score for the FA model; Figure 6). Non-acoustic factors were also shown to shape performance, particularly with respect to false alarms, where timing based factors were strong predictors of false alarms. Trials early in the session more likely to elicit false alarms and false alarm probability increasing through the trial. Both of these effects likely reflect impatience on the part of the ferret, as well as potentially increasing uncertainty for longer-duration trials. For the miss model trial number had a strong impact on performance, with later trials in the session were more likely to be missed; since animals are allowed to complete as many trials as they like, and satiety will increase during the trial it seems likely that misses are mostly shaped by task engagement.

The predictive power of the hit/miss model was lowest overall, at around 63%, likely due to the limited number of misses in our dataset. Our false alarm/correct reject classification model had more trials to work

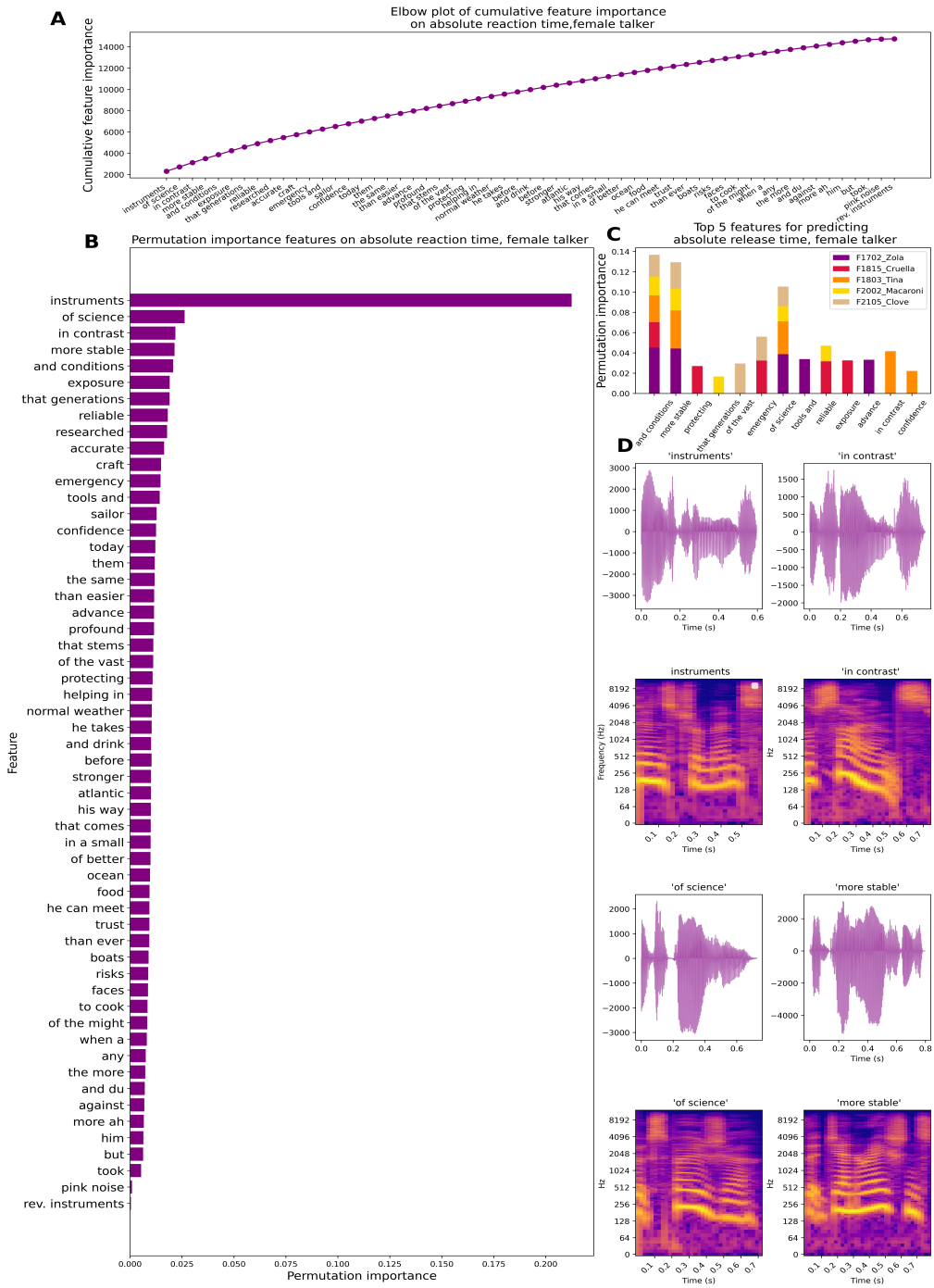

Figure 9. A combination of spectrotemporal properties in word tokens drives false alarms in the female talker. a, elbow plot of cumulative feature importances; b, SHAP summary plot depicting the ranked feature importances of the absolute release time model (as expected, target word presentation drives the release times; c-d, f-g, spectrogram with the wave amplitudes of the top 3 words that contributed to a release from the centre spout along with the target; e, permutation feature importance of the absolute release time model across individual animal models.

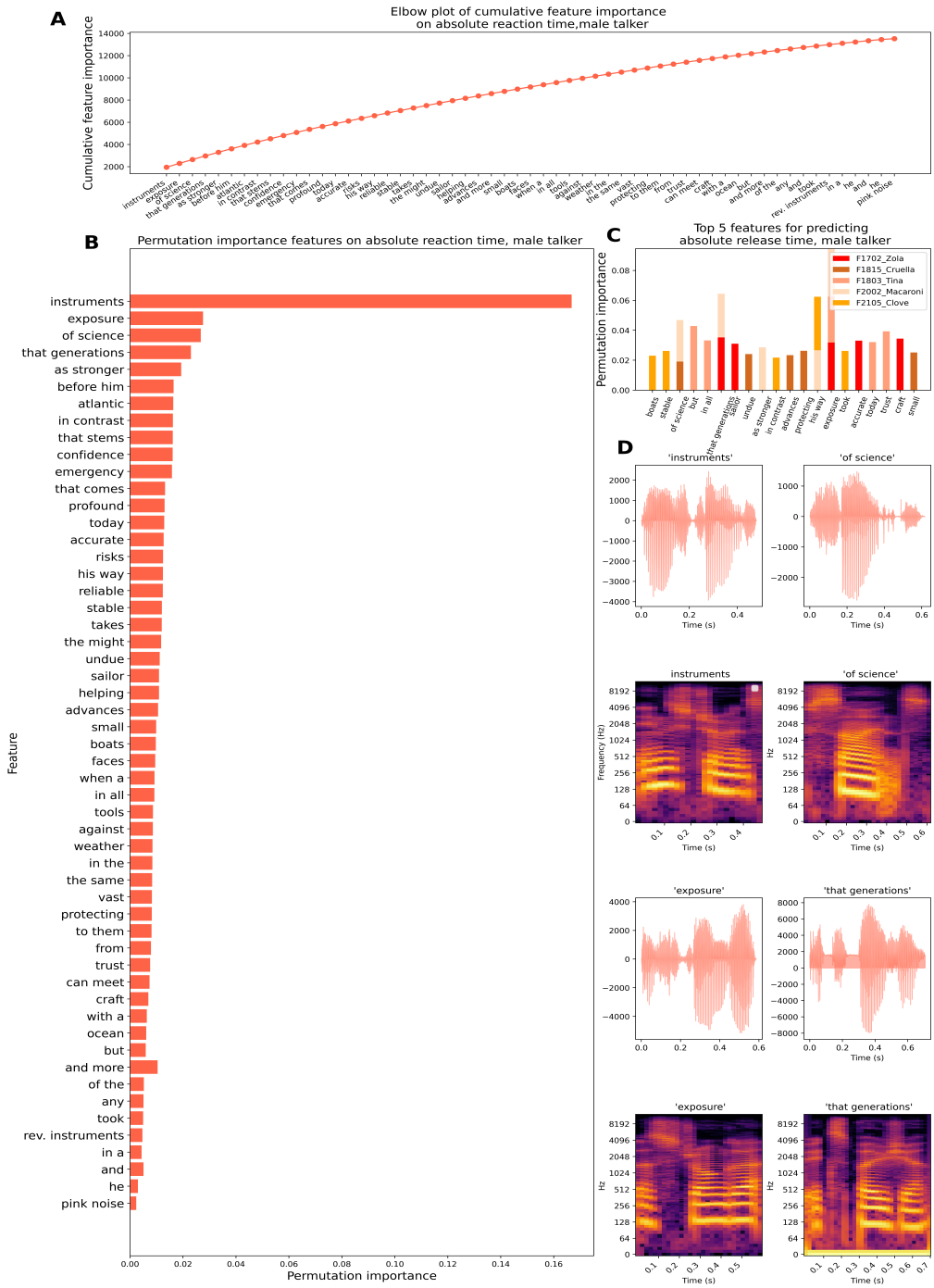

Figure 10. A combination of spectrotemporal properties in word tokens drives false alarms in trials featuring the male talker. a, elbow plot 54of cumulative feature importances; b, SHAP summary plot depicting the ranked feature importances of the absolute release time model (as expected, target word presentation drives the release times; c-d, f-g, spectrogram with the wave amplitudes of the top 3 words that contributed to a release from the centre spout along with the target; e, ranked permutation feature importance of the features of the absolute release time model.

with and greater predictive power (71.24 % balanced accuracy).

While the categorical models discussed above provided important insights into our animals' behaviour, the reaction time analysis yielded a more nuanced perspective of the multiple factors (and their interactions) that shape performance. The reaction time model had high predictive power, with a mean-squared error of 0.0929s%, which in effect, one could think about an accuracy of 1 - 0.3 s/2 s = 85 %. Our ferrets are highly trained and very capable at performing the task, providing bountiful hit responses for our training set and greater certainty that non-sensory factors such as fatigue, after impatience, influence performance. While we could not determine a sizeable effect of pitch-roving within a trial using our categorical performance models, our reaction time model allowed us to see that this manipulation slowed reaction times (Figure 8b). If we interpret changes in reaction time as reflective of uncertainty, then it seems clear that while our subjects were able to compensate for the added pitch uncertainty through training to achieve the same net water reward, token-to-token F0 roving makes auditory object recognition more difficult. The gradient-boosted regression approach allowed us to unpick the complex interdependencies of talker and F0 on our dataset, illustrating that once the talker type was accounted for, reaction times were only slowed for trials in which the male talker's voice was shifted up to 144 Hz (which was also the only common F0 for both talkers).

We extended our use of gradient-boosted regression to try and determine whether there were individual words to which animals reliably false alarmed, following our anecdotal observations that animals sometimes reliably repeated the same errors across correction trials. We therefore asked if we could predict the animals' response time from only the identity (and timing) of the word tokens on an individual trial. If animals never false alarmed then we should be able to predict the timing of the animals' response, based on the timing of the target word, with near-perfect accuracy. While the target word strongly predicted the animals' response time, other words were also predictive of the response timing (Figure 9, 10). The model-free approach offered by gradient-boosted regression is particularly powerful. We did not need to have prior knowledge as to what the animals' reaction time to a non-target word is. Instead, we could use the whole stimuli set of words to predict the reaction time, something that would be difficult to do with a classical regression model.

While each animal showed a slightly different pattern of words, some words were common across ferrets and across the two talkers. Out of the most significant words that the subjects false alarmed to, only one had a temporally similar duration to the target word in both the male and female talker models, suggesting that the ferrets are considering more than just the total duration of each word in making their decision, and are likely instead using a comparing the spectrotemporal pattern of the preceding distractor words to the target word, especially as some of the words carried similar amplitude waveform properties to the respective target word ("in contrast", female talker) and similar frequency spectra ("that generations", male talker). This suggests similar to humans; our animal model is also comparing a variety of spectrotemporal properties of each distractor word to the target word. Indeed, neither pink noise with a matching average frequency spectrum and temporal amplitude dynamics nor "instruments" temporally reversed was correlated with a lick release.

Similar to our perceptual constancy task, Juchter et al. (2022) used a Go/No-Go paradigm to study the perception of single-syllable nonsense words ("logotomes"; both consonant-vowel-consonant [CVCs] and VCVs). Using multidimensional scaling, the authors demonstrated that gerbils and humans performing the same task built a similar underlying 'perceptual map' of speech sounds. By comparing their reaction times with humans performing the same task, Mongolian gerbils perceive vowels and consonants in the same principal component space as humans can learn vowel identities based on frequency differences in the formants and the cadence of the words (Jüchter, Beutelmann, and Georg Martin Klump 2022). This study demonstrates that another low-frequency hearing animal model discriminates word elements similarly to humans. We have built upon this similar work and show that the ferret can perform a complex task where the combination of vowels and consonants is not presented in a consistent A-B-A pattern, much like how domesticated animals are responsive to certain commands in the home setting.

Although ferrets can be trained to a high degree of accuracy in recognising the target word (Figure 3), through using the nonlinear gradient boosted regression tree, we were able to elucidate that they were still using pitch constancy as one of the cues in recognising the target word (Figure 8). We chose the machine learning approach of gradient-boosted regression and decision trees to analyse our behavioural data, as our data is abundant in terms of sample size and tabular in nature. This scenario of structured, dense data is ideal for gradient-boosted decision and regression trees, as this type of method has been often used in recommender systems (Luo et al. 2022) as well as economic predictive modelling for human behaviour in customer loyalty (Machado, Karray, and Sousa 2019). Thus, while it might seem like a niche concept to apply a machine learning technique to animal psychophysics data, the foundational elements of our data (tabular and dense) make gradient-boosted regression and decision trees a good choice of model.

While we do not propose that ferrets naturally form categorical representations of the meaning of speech sounds, our data show they can be trained to learn complex spetrotemporal patterns and that while these are tolerant to the F0 of the voiced elements of these sounds, F0 likely contributes to their ability to link sounds into streams over time. The mechanisms underlying categorical responses to complex sounds and how such representations become tolerant to variation in F0 are not well understood, and by applying machine learning techniques to our behavioural data, we can generate testable hypotheses about the neural representation of such sounds in areas such as auditory cortex.

CONCLUSIONS

This study shows a novel approach to analyzing behavioural animal data: using gradient-boosted regression and decision trees to model the psychoacoustic features underlying our subjects' performance. This approach allowed us to show that while animals were tolerant to changes in F0 of the target word, untrained F0s were generally associated with a slightly higher likelihood of false alarming. When examining reaction times, we found evidence that pitch constancy between the target and preceding non-target word accelerated reaction times relative to when the F0 differed from token to token. Thus, like humans, our behavioural findings suggest that ferrets form pitch-tolerant representations of auditory objects but use consistent pitch as an aid when deciding which groups of sounds constitute a single auditory stream.

AUTHOR CONTRIBUTIONS

CG, JS, and JKB conceived of the original study and the mechanisms of the perceptual constancy task. CG generated the data analysis and code required to produce this paper's gradient-boosted regression and classification models. JL wrote helper functions to extract MATLAB metadata into a standardized Python format. CG and JB wrote the manuscript. All authors have read and agreed to the published version.

DATA AVAILABILITY

Behavioural data can be requested to replicate the results of this paper.

FUNDING

This work was supported in whole, or in part, by a Wellcome Trust/Royal Society Sir Henry Dale Fellowship Grant 098418/Z/12/A to J.K.B.; European Research Council SOUNDSCENE; Biotechnology and Biological Sciences Research Council BB/N001818/1.

ACKNOWLEDGMENTS

We thank the hard work and dedication of the Royal Veterinary College technicians and veterinary staff for supporting our scientific experiments and maintaining our animal colony.

CODE AVAILABILITY

Code used to reproduce the models will be available upon publication.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Akiba, Takuya et al. (July 2019). *Optuna: A Next-generation Hyperparameter Optimization Framework*. arXiv:1907.10902 [cs, stat]. DOI: 10.48550/arXiv.1907.10902. URL: http://arxiv.org/abs/1907.10902 (visited on 06/02/2023).

Aulanko, R et al. (Sept. 1993). "Phonetic invariance in the human auditory cortex". eng. In: *Neuroreport* 4.12, pp. 1356–1358. ISSN: 1473-558X. DOI: 10.1097/00001756-199309150-00018. URL: https://doi.org/10.1097/00001756-199309150-00018 (visited on 03/15/2023).

Bee, Mark A. and Georg M. Klump (Aug. 2004). "Primitive Auditory Stream Segregation: A Neurophysiological Study in the Songbird Forebrain". In: *Journal of Neurophysiology* 92.2. Publisher: American Physiological Society, pp. 1088–1104. ISSN: 0022-3077. DOI: 10.1152/jn.00884.2003. URL: https://journals.physiology.org/doi/full/10.1152/jn.00884.2003 (visited on 05/24/2023).

Bee, Mark A. and Kasen K. Riemersma (Sept. 2008). "Does common spatial origin promote the auditory grouping of temporally separated signal elements in grey treefrogs?" en. In: *Animal Behaviour* 76.3, pp. 831–843. ISSN: 0003-3472. DOI: 10.1016/j.anbehav.2008.01.026. URL: https://www.sciencedirect.com/science/article/pii/S0003347208002194 (visited on 05/24/2023).

Bizley, Jennifer K. et al. (Jan. 2013). "Spectral timbre perception in ferrets: Discrimination of artificial vowels under different listening conditions". In: *The Journal of the Acoustical Society of America* 133.1. Publisher: Acoustical Society of America, pp. 365–376. ISSN: 0001-4966. DOI: 10.1121/1.4768798. URL: https://asa.scitation.org/doi/full/10.1121/1.4768798 (visited on 03/15/2023).

Charlton, Benjamin D., Zhang Zhihe, and Rebecca J. Snyder (2009). "The information content of giant panda, Ailuropoda melanoleuca, bleats: Acoustic cues to sex, age and size". In: *Animal Behaviour* 78. Place: Netherlands Publisher: Elsevier Science, pp. 893–898. ISSN: 1095-8282. DOI: 10.1016/j.anbehav.2009.06.029.

Dancer, Alice M. M. et al. (Nov. 2022). "Pet Owner Perception of Ferret Boredom and Consequences for Housing, Husbandry, and Environmental Enrichment". eng. In: *Animals: an open access journal from MDPI* 12.23, p. 3262. ISSN: 2076-2615. DOI: 10.3390/ani12233262.

Dent, Micheal L. et al. (Feb. 2016). "Cues for auditory stream segregation of birdsong in budgerigars and zebra finches: Effects of location, timing, amplitude, and frequency". In: *The Journal of the Acoustical Society of America* 139.2, pp. 674–683. ISSN: 0001-4966. DOI: 10.1121/1.4941322. URL: https://doi.org/10.1121/1.4941322 (visited on 05/24/2023).

Dolležal, Lena-Vanessa et al. (Mar. 2020). "Release from informational masking by auditory stream segregation: perception and its neural correlate". eng. In: *The European Journal of Neuroscience* 51.5, pp. 1242–1253. ISSN: 1460-9568. DOI: 10.1111/ejn.13794.

Engineer, Crystal T. et al. (2013). "Similarity of cortical activity patterns predicts generalization behavior". eng. In: *PloS One* 8.10, e78607. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0078607.

Green, David M. and John A. Swets (1966). *Signal detection theory and psychophysics*. Signal detection theory and psychophysics. Pages: xi, 455. Oxford, England: John Wiley.

Grinsztajn, Léo, Edouard Oyallon, and Gaël Varoquaux (July 2022). *Why do tree-based models still outperform deep learning on tabular data?* arXiv:2207.08815 [cs, stat]. DOI: 10.48550/arXiv.2207.08815. URL: http://arxiv.org/abs/2207.08815 (visited on 06/01/2023).

Hauser, Marc D. (Sept. 1993). "The Evolution of Nonhuman Primate Vocalizations: Effects of Phylogeny, Body Weight, and Social Context". In: *The American Naturalist* 142.3. Publisher: The University of Chicago Press, pp. 528–542. ISSN: 0003-0147. DOI: 10.1086/285553. URL: https://www.journals.uchicago.edu/doi/abs/10.1086/285553 (visited on 05/24/2023).

Haykin, Simon and Zhe Chen (Sept. 2005). "The Cocktail Party Problem". In: *Neural Computation* 17.9, pp. 1875–1902. ISSN: 0899-7667. DOI: 10.1162/0899766054322964. URL: https://doi.org/10.1162/0899766054322964 (visited on 03/15/2023).

Homma, Natsumi Y. et al. (June 2016). "Mistuning detection performance of ferrets in a go/no-go task". In: *The Journal of the Acoustical Society of America* 139.6. Publisher: Acoustical Society of America, EL246–EL251. ISSN: 0001-4966. DOI: 10.1121/1.4954378. URL: https://asa.scitation.org/doi/abs/10.1121/1.4954378 (visited on 03/15/2023).

Hulse, S. H., S. A. MacDougall-Shackleton, and A. B. Wisniewski (Mar. 1997). "Auditory scene analysis by songbirds: stream segregation of birdsong by European starlings (Sturnus vulgaris)". eng. In: *Journal of Comparative Psychology (Washington, D.C.: 1983)* 111.1, pp. 3–13. ISSN: 0735-7036. DOI: 10.1037/0735-7036.111.1.3.

Jüchter, Carolin, Rainer Beutelmann, and Georg Martin Klump (May 2022). "Speech sound discrimination by Mongolian gerbils". en. In: *Hearing Research* 418, p. 108472. ISSN: 0378-5955. DOI: 10.1016/j.heares.2022.108472. URL: https://www.sciencedirect.com/science/article/pii/S0378595522000430 (visited on 03/15/2023).

Kawahara, Hideki (2006). "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds". en. In: *Acoustical Science and Technology* 27.6, pp. 349–353. ISSN: 1346-3969, 1347-5177. DOI: 10.1250/ast.27.349. URL: http://www.jstage.jst.go.jp/article/ast/27/6/27_6_349/_article (visited on 06/12/2023).

Ke, Guolin et al. (Dec. 2017). "LightGBM: a highly efficient gradient boosting decision tree". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., pp. 3149–3157. ISBN: 978-1-5108-6096-4. (Visited on 03/13/2023).

Lundberg, Scott M and Su-In Lee (2017). "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc. URL: https:

//papers.nips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html (visited on 12/02/2022).

Luo, Jiangwei et al. (Sept. 2022). "LightGBM using Enhanced and De-biased Item Representation for Better Session-based Fashion Recommender Systems". In: *Proceedings of the Recommender Systems Challenge 2022*. RecSysChallenge '22. New York, NY, USA: Association for Computing Machinery, pp. 24–28. ISBN: 978-1-4503-9856-5. DOI: 10.1145/3556702.3556839. URL: https://dl.acm.org/doi/10.1145/3556702.3556839 (visited on 06/01/2023).

Machado, Marcos Roberto, Salma Karray, and Ivaldo Tributino de Sousa (Aug. 2019). "LightGBM: an Effective Decision Tree Gradient Boosting Method to Predict Customer Loyalty in the Finance Industry". In: *2019 14th International Conference on Computer Science & Education (ICCSE)*. ISSN: 2473-9464, pp. 1111–1116. DOI: 10.1109/ICCSE.2019.8845529.

Ohms, Verena R. et al. (Apr. 2010). "Zebra finches exhibit speaker-independent phonetic perception of human speech". eng. In: *Proceedings. Biological Sciences* 277.1684, pp. 1003–1009. ISSN: 1471-2954. DOI: 10.1098/rspb.2009.1788.

Sumner, Christian J. et al. (Oct. 2018). "Mammalian behavior and physiology converge to confirm sharper cochlear tuning in humans". In: *Proceedings of the National Academy of Sciences* 115.44. Publisher: Proceedings of the National Academy of Sciences, pp. 11322–11326. DOI: 10.1073/pnas.1810766115. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1810766115 (visited on 06/02/2023).

Town, Stephen M. et al. (May 2015). "The role of spectral cues in timbre discrimination by ferrets and humans". In: *The Journal of the Acoustical Society of America* 137.5, pp. 2870–2883. ISSN: 0001-4966. DOI: 10.1121/1.4916690. URL: https://doi.org/10.1121/1.4916690 (visited on 06/01/2023).

SUPPLEMENTARY MATERIAL

Trial information

Ferret ID	F1702		F1815		F1803		F2002		F2105	
Talker	M	F	M	F	M	F	M	F	M	F
All trials	2834	2773	1758	1684	3060	3001	6381	3611	2026	2055
Catch trials	717	694	454	439	781	756	1583	918	504	519

Supplementary Table S1. Table of trial type numbers distributed by ferret ID and talker type (M = male talker, F= female talker)

Model Parameters

False alarm categorical model

Parameter	Value
colsample_bytree	0.6964999459813326
subsample	0.9635926861919909
learning_rate	0.4814769779829811
num_leaves	388
max_depth	18
min_child_samples	30
reg_alpha	1.009631439726829
reg_lambda	2.8430971108782197
min_split_gain	0.003052783619672129
bagging_freq	7
feature_fraction	0.9993759866702121
scale_pos_weight	1.1337990379884262
min_child_weight	1.2491722523962963
max_bin	335
min_data_in_leaf	17
min_sum_hessian_in_leaf	0.12530107056740758

Supplementary Table S2. Hyperparameters

Miss/hit model

Parameter	Value
colsample_bytree	0.8144184011423041
subsample	0.9650822598936615
learning_rate	0.24097046270976083
num_leaves	326
max_depth	4
min_child_samples	161
reg_alpha	1.1265361419969506
reg_lambda	2.5610604817723837
min_split_gain	0.0038910469191018263
bagging_freq	3
feature_fraction	0.7289126413310938
scale_pos_weight	1.1010467745609456
min_child_weight	8.080006560459093
max_bin	653
min_data_in_leaf	67
min_sum_hessian_in_leaf	4.6251428904179

Supplementary Table S3. Hyperparameters for the miss/hit gradient-boosted decision tree model

Reaction time model

Parameter	Value
colsample_bytree	0.46168728494506456
alpha	8.758272905706946
n_estimators	82
learning_rate	0.2165288044507529
max_depth	18
bagging_fraction	0.7000000000000001
bagging_freq	0

Supplementary Table S4. Hyperparameters for the reaction time gradient-boosted regression tree model predicting the reaction time from the onset of the target word from the subset of correct hit responses.

Absolute reaction time model - female talker

Parameter	Value
colsample_bytree	0.49787217530410993
alpha	14.65910072586951
n_estimators	100
learning_rate	0.24475049401210153
max_depth	18
bagging_fraction	0.9
bagging_freq	17

Supplementary Table S5. Hyperparameters for the absolute reaction time gradient-boosted regression tree model that predicts the reaction time relative to the female talker type trial start time.

Absolute reaction time model - male talker

Parameter	Value
colsample_bytree	0.5870762820095368
alpha	10.840482953967314
n_estimators	70
learning_rate	0.18038495501541654
max_depth	20
bagging_fraction	0.9
bagging_freq	30

Supplementary Table S6. Hyperparameters for the absolute reaction time gradient-boosted regression tree model that predicts the reaction time relative to the male talker type trial start time.