Gradient boosted decision trees reveal nuances of auditory discrimination behavior

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

Pitch is a fundamental feature of a person's voice, and a hallmark of human voice processing is recognizing a word regardless of voice pitch. Pitch continuity substantially improves our ability to hear a person's voice in noise. However, it is not clear whether pitch continuity is uniquely human or whether it can be considered a more general feature of the mammalian auditory system. To assess this, we trained ferrets to report a target word's presence, timing, and lateralization within a stream of consecutively presented non-target words. To assess the animals' ability to generalize across pitch, we manipulated the fundamental frequency (F0) of the speech stimuli across trials, and to assess the contribution of pitch to streaming, we roved the F0 from word token-to-token. We then implemented gradient-boosted regression and decision trees on the trial outcome and reaction time data to understand the behavioral factors behind the ferrets' decision-making. While ferrets were able to accurately perform the task across all pitch-shifted conditions, our models reveal subtle effects of shifting F0 on performance, with within-trial pitch shifting elevating false alarms and extending reaction times. While our models demonstrated that false alarms are primarily driven by non-acoustic factors, through our novel application of well-established machine learning algorithms, we were able to identify a subset of words that animals consistently confused with the target word. Overall, gradient-boosted regression and decision trees allowed us to tease apart acoustic and behavioral factors that shaped sound discrimination performance in this task, demonstrating that ferrets can identify complex sounds across variations in pitch but that they, like humans, utilize the F0 of sounds as a streaming cue.

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

AUTHOR SUMMARY

Hearing is one of the aspects of life that we often take for granted until it is too late. According to the World Health Organization, by 2050, 1 in 10 people will have life-changing hearing loss. However, our understanding of how even normal brain mechanisms support hearing remains limited. This remains challenging because such work relies upon animal models, and most behavioral tasks implemented in animals remain impoverished versions of everyday challenges humans face when listening to speech in the presence of other competing sounds. Here, we used a novel behavioral paradigm designed to include this complexity and present a novel implementation of machine learning algorithms to understand how trained ferrets perform this task. Gradient-boosted regression and decision trees are well-established machine learning methods that do not require users to predetermine interaction effects. Through this machine learning method, we find that ferrets can perform the task across variations in voice pitch and benefit from continuity of pitch within a sound sequence to make their decisions. Our results suggest that this machine learning approach can analyze behavioral data in animal models and that ferrets can process complex aspects of sound stimuli to make choices under changing sound environments similar to humans.

INTRODUCTION

Psychophysics paradigms in non-human animals are often designed to yield tractable datasets for relating brain and behavior. Auditory cortex is thought to be critical for discriminating sounds based on combinations of spectrotemporal features and for organizing sounds into auditory scenes Bizley and Cohen 2013, Griffiths and Warren 2004). However, inactivation studies frequently yield contradictory results that

often fail to demonstrate a role for auditory cortex in listening tasks (Slonina, Poole, and Bizley 2022). Most common laboratory-based paradigms rely on artificial stimuli presented within simple tasks – such as two-alternative forced choice paradigms in which animals must discriminate a single sound token, or go/no-go tasks in which animals detect a change in a repeating sequence of sounds. While such paradigms offer tight experimental control, they potentially fail to engage auditory cortex, providing an explanation for why inactivating auditory cortex apparently has no impact (Slonina, Poole, and Bizley 2022). Yet animals can be trained to perform more complex tasks, generating rich behavioral datasets. Significant advances have been made in the analysis of behavioral data, and in particular in identifying non-sensory factors that shape performance (Ashwood et al. 2022 Roy et al. 2021). Here, we report a novel application of machine learning methods to parse a large dataset of behavioral responses from animals trained in a sound discrimination task. Specifically, we sought to understand whether 1) trained ferrets can generalize learned discrimination across variations in pitch and 2) whether, like humans, animals use the pitch as a streaming cue to link sounds together over time. Ferrets were trained to detect the word "instruments" within a stream of other randomly drawn non-target words (Sollini and Bizley, in prep.). Within a trial, all word tokens were drawn from a single female or male voice, and the whole stream could be shifted upwards or downwards in fundamental frequency (F0, which determines pitch). The F0 of each word within a stream could also be randomly shifted to assess whether pitch contributes to streaming. We used gradient-boosted regression and decision trees to analyze 20,487 trials of behavioral data. This novel application of machine learning algorithms allowed us to demonstrate 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. Moreover, this approach allowed us to identify words that ferrets consistently confused with the target word suggesting that errors were not simply random lapses in attention.

RESULTS

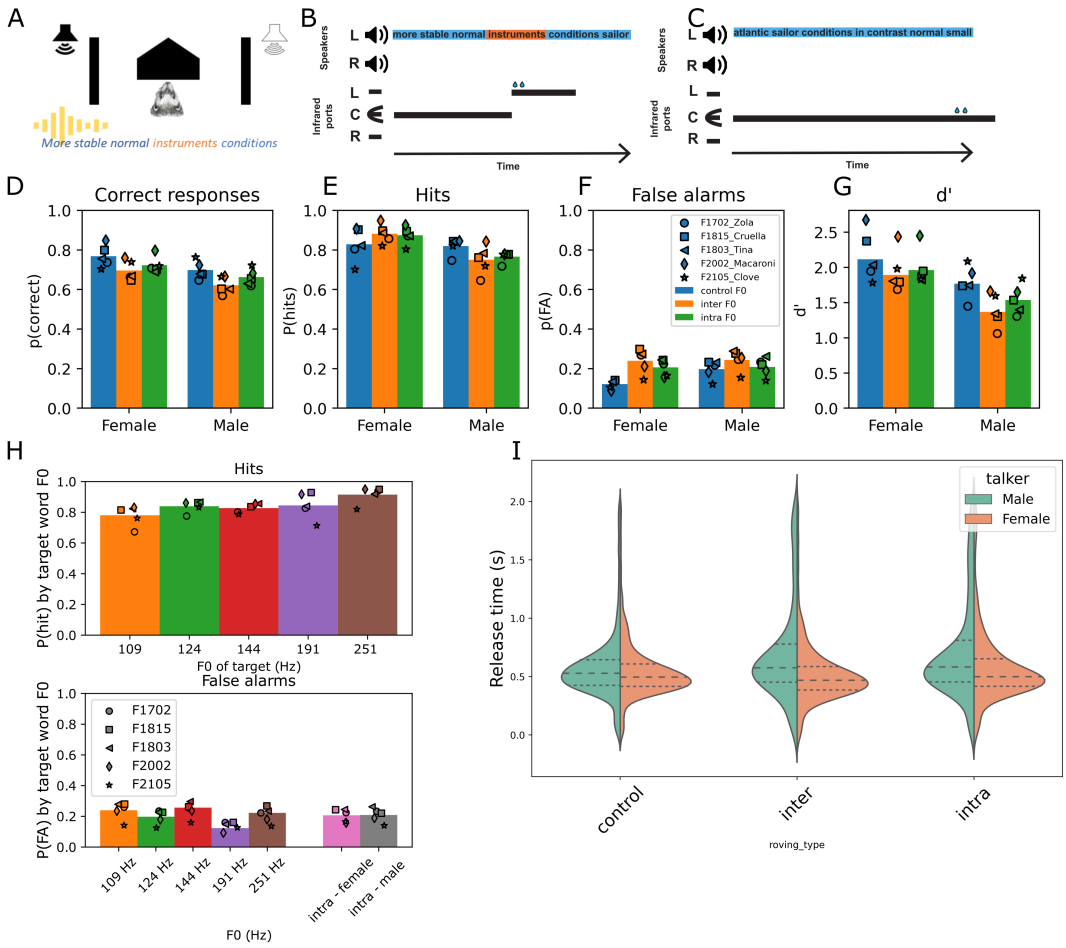


Figure 1. A; Schematic of the experimental booth. To trigger a trial, ferrets had to nose-poke a center port that contained an IR sensor and water port. This triggered the presentation of a stream of words from either the left or right speaker. (B) Ferrets were trained to remain at the center until the presentation of the target word ('instruments'), and received a water reward at a lateral port if they correctly released within 2s of target presentation and responded to the lateral port whose side matched that of the speech stream. (C) Catch trials did not contain the target word, and the ferret was rewarded for remaining at the central port for the duration of the trial. (D) Behavioral 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). D-G Behavioral metrics across animals: D % correct over all trials, E hits, F false alarms G sensitivity (d'). (H) impact of F0 on hit rate (top) and false alarm rate (bottom). False alarm rates are plotted separately for intra-trial pitch roving because the F0 changed from token to token, making it impossible to assign a false alarm to a distractor of a given F0.I; Violin plot of reaction times during correct responses on trials in which the target was correctly identified for all animals, separated by talker type.

Ferrets can discriminate speech sounds and their performance is robust to pitch shifting

Ferrets were trained to detect the target word "instruments" within a stream of randomly drawn non-target word tokens. Subjects initiated a trial by nose-poking in a central port that contained an infrared sensor and water delivery spout and were required to remain at the center port until the presentation of the target word. On each trial, all tokens came from the same talker and position in space, and ferrets were rewarded for responding at the lateral port adjacent to the speaker within 2s of the target word (Fig. 1A,B). On catch trials, in which only non-target words were presented, ferrets were rewarded for remaining at the central port (Fig.1C). Ferrets were trained with a single male and single female voice. Once performance was stable, trials were introduced 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'). 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.

Ferrets' were able to learn and perform the task across control and F0-shifted conditions; performance ranged from 57% -85% correct for all animals and conditions, where 33% would be considered chance performance (Fig.1D). Hit rates were generally high (Fig.1E) and false alarms low (Fig. 1F) for both talkers and for both types of pitch-shifted trials. Overall, performance was higher for the female voice, with a small decrease in d' evident for pitch-roved trials compared to natural F0 ones (Fig.1G). Nonetheless, all d' values were well above 1, indicating the animals were well able to perform the task.

To understand whether ferrets form a pitch-tolerant representation of the target word, we considered the impact of F0 changes, with inter and intra-trial roving eliciting slightly lower percentage correct values and d' overall, with slightly higher hit rates and false alarms for the pitch-shifted female voice and lower hit rates and higher false alarms for the male voice (see Fig.S1 for bias values). For both male and female trials, the performance on inter and intra-trial roved trials was equivalent (Fig.1D-F). When the performance was broken down according to the actual F0 value we observed there was a modest influence of F0 on hit rates, such that the highest hit rates were observed for the female talker's up-shifted F0 trials (Fig.1H). False alarms, in contrast, were lower for the trained F0 values for both the male and female talkers.

Reaction times varied by ferret and according to the talker (Figure S1B). 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 too may limit performance at the lowest F0s.

The basic behavioral metrics above show that ferrets can successfully discriminate a target word from non-target words despite variation in F0. We turn to gradient-boosted decision trees to further consider how acoustic and non-acoustic factors influence individual trial outcomes. We chose this machine learning approach as our data is abundant in sample size and tabular. While its application to animal behavioral work is novel, this scenario of structured, dense data is ideal for gradient-boosted decision and regression trees, as this type of method has often been used in recommender systems (Luo et al. 2022) as well as economic predictive modeling for human behavior in customer loyalty (Machado, Karray, and Sousa 2019). We considered two types of models – decision tree models that performed categorical discriminations, considering whether responses to targets were misses or hits and whether responses to non-target words were false alarms or correct rejections, and decision-tree regression models to predict continuous reaction time data. While we are interested in the impact of F0 shifting on performance, other factors of interest and nuisance variables may shape the animals' overall performance. A machine learning approach is ideal because they do not require users to predetermine interaction effects in their model, and we can consider multiple stimulus features (such as the talker and pitch of the word, as well as the trial history parameters of was the previous trial correct, was the previous trial a catch trial) and non-stimulus features (such as the timing of the trial within the session, the time of the target word within the trial, and the side that the animal was required to respond) that may influence performance but do not inform our experimental hypothesis.

Talker identity drives miss responses

We modelled the likelihood of a miss vs hit using only trials in which the animal heard the target sound (i.e., excluding false alarms and catch trials) based on; the talker (male/female), the side (left/right) of the audio presentation, the trial number (in the session), the subject identity (ID), target presentation time (within the trial), the target F0, whether the previous trial was a catch trial, whether the previous response was correct, and whether the F0 of the non-target word preceding the target matched that of the target (non-target F0=target F0).

The performance of the miss/hit model was reasonable despite the sparsity of miss responses in the behavioral data, with an average balanced accuracy on our training set of 63.22% and an average test balanced accuracy of 61.50%. We eliminated factors that either did not significantly increase the cumulative feature importance plot (Fig.2A) or if a permutation test that randomized the variable in question did not impact model fit (Fig.2D). Thus, trial history factors (the past trial was correct or a catch trial) and the prior non-target F0=target F0 parameter were eliminated. For the remaining features, the feature importance metrics, permutation tests, and SHAP feature values were all in concordance with each other, with only minor differences in the ranking of features. The top three features were the talker (the male talker increased the probability of a miss, Fig.2C), the side of the audio presentation (which was idiosyncratic across animals, likely reflecting their own individual biases, see Fig.S2B), and the trial number (with trials occurring later in the session being associated with higher miss rates). While significant, the target presentation time within the trial (Fig.S2A) did not show a strong relationship across animals, as shown by the lack of stratification in the SHAPs plot examining the target presentation time for each ferret. The

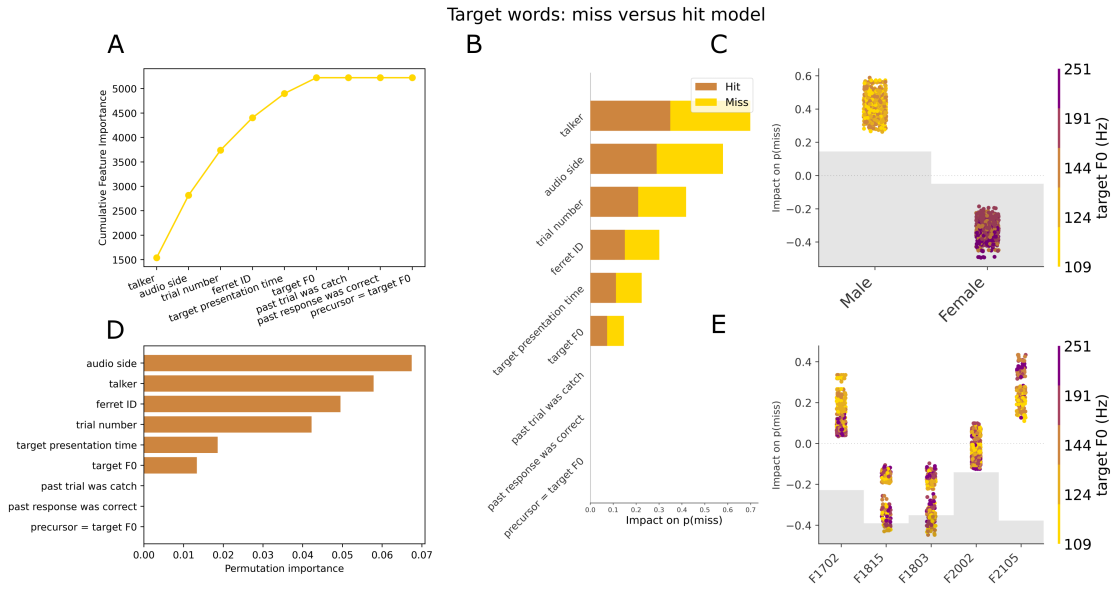


Figure 2. Factors correlated with within-ferret trial parameter preferences drive the miss/hit model; A, the elbow plot of cumulative features over trial features; B, SHAP feature importances of the miss/hit model; C, SHAP partial dependency plot showing the SHAP impact over each talker type color-coded by target F0; D, permutation importance bar plot of the features in the correct hit/miss model; e) SHAP partial dependency plot depicting the SHAP impact over each ferret ID color-coded by target F0.

F0 of the target sound also had a small but significant effect, which varied by ferret (Fig. 2E). Only 2/5 animals had stratified miss probabilities based on target pitch (F1702 who was more likely to miss low F0 trials, and F2105 who was more likely to miss high F0 trials, visible as a smooth color continuum in these two animals). Whether the non-target word that preceded the target word was matched in F0 did not significantly influence the likelihood of missing. We conclude that the identity of the talker was the single biggest stimulus factor that altered the likelihood of missing, with the F0 of the target word having a modest effect in some animals. Changing the F0 from word token to word token did not change the likelihood of correctly detecting the target.

False alarms are influenced by talker identity and F0

Next, we modelled whether a subject would false alarm all trial types, based on the following features: the talker, the pitch (F0) of the trial or for intra-trial roved trials the F0 of the last non-target word in the trial, the side of audio presentation, the trial duration, the time elapsed since the start of the trial, the trial number within the experimental session, the ferret ID, whether the past response was correct, whether the past trial was a catch trial, and whether there was intra-trial F0 roving. The false alarm model had above-chance accuracy (mean test accuracy of 61.54% over 5-fold cross-validation; balanced accuracy 61.46%), and returned the following as the most significant contributors: the time elapsed since the trial started, the trial number, the ferret ID, the non-target F0, the audio side, and whether the trial was intra-trial F0 roved (Fig. 3A, B, D).

In contrast to the miss model, the strongest determinants of whether an animal was likely to false alarm were timing parameters (time in the trial and trial number within the session) and the individual ferrets. Partial dependency plots (Fig S3) showed that two ferrets were more likely to false alarm early in the trial, one late in the trial and two animals showed unstratified responses, implying they were not systematically influenced by this parameter (FigS3A). Trial number, although significant, did also not show clear stratification when considered by animal (FigS3H).

The speech sound F0 and talker both impacted the likelihood of FA, with the partial dependency plot showing that low F0 words spoken by the female talker were most likely to elicit a FA, while the trained F0 for the female talker was least likely to elicit a FA (Fig.3C, Fig S3C). The audio side and intra-trial roving also contributed predictive power: the audio side was again idiosyncratic across animals (FigS3.B). Whether or not word tokens within a trial varied in F0 (i.e. intra-trial roving) contributed a small but significant effect, with 4 / 5 ferrets showing an increased likelihood of false alarming when the F0 of words varied from token to token (Fig.3E).

In summary, the FA model suggests that non-acoustic factors are the key drivers in whether animals false alarm with a small contribution of acoustic factors, particularly the untrained.

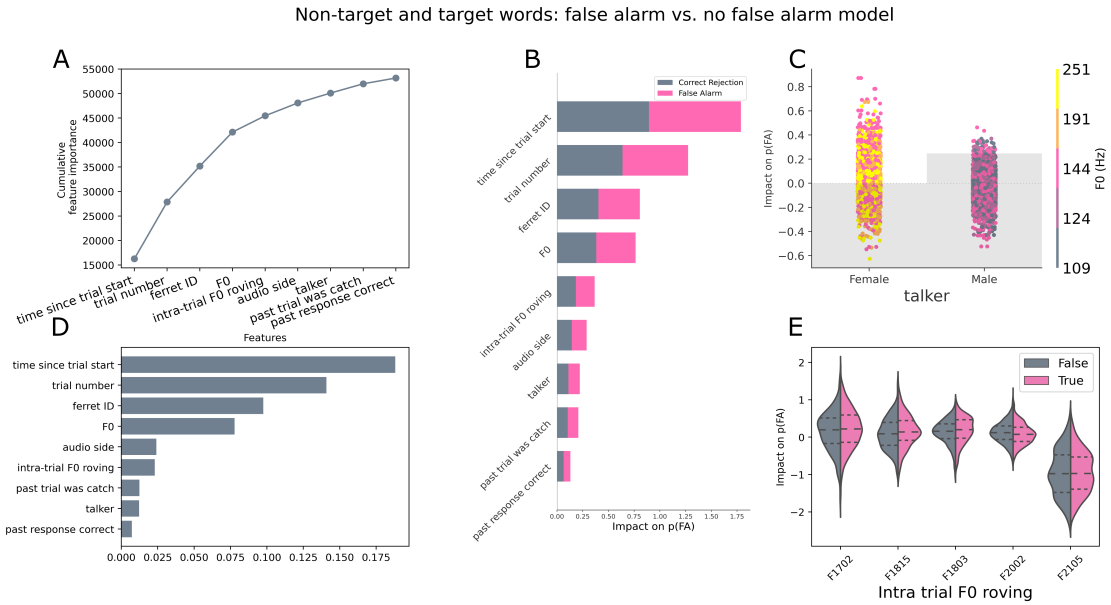


Figure 3. Precursor F0 determines the probability of a false alarm; A, elbow plot depicting the cumulative feature importance of each factor used in the false alarm decision tree model; B, SHAP feature importance plot for the same factors as in A; C, partial dependency plot showing the SHAP value (representing the impact on the probability the trial would be predicted as a false alarm) over ferret ID color-coded by the 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 whether the trial was intra-trial roved color-coded by F0. Gray bars illustrate the relative proportion of trials across categories.

Gradient boosted regression of reaction time data reveals the impact of pitch on target detection and streaming

Given our performance measures were generally quite high with, in particular, a very limited number of miss trials with which to explore whether F0 changes impacted performance, we focused next on reaction time (RT) measures. To explore whether RTs provided a more sensitive measure of how acoustic and task parameters influenced performance, we used gradient-boosted regression (Ke et al. 2017). In our RT model, derived from responses from correct non-catch trials, we considered the following factors: ferret ID, talker (male or female), 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 0.0947s, or alternatively, a median absolute error of +/- 0.1260s, a relatively high degree of accuracy (train mean-squared error = 0.0895s). From our permutation test, the ferret ID, the talker, the side of the audio presentation, the time to target presentation, the target F0, and whether the F0 of the previous word equaled the target word were significant factors in this reaction time model. Similar to the miss/hit and false alarm/correct reject performance models, both ferret ID and talker ID accounted for a large amount of model variance; reaction times were longer for the male talker (in 4/5 ferrets, see Supplemental S4D, female faster in F2105) and varied systematically across ferrets (Fig.4C). Overall, later targets had faster responses, Figure 4B, 3/5 ferrets showed this effect, 1/5 had faster reaction times for earlier targets, and 1 showed no difference, Fig.S4A) Other factors that significantly predicted reaction times were the time of the target within a trial (later targets had faster responses, Fig.4B), the side of the audio (idiosyncratically across animals; 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, FigS4B. The model dissociated the effects of talker and F0, with the effect of F0 being somewhat idiosyncratic across ferrets with three ferrets showing the fastest reaction times for the pitches associated with the talker with the fastest reaction times (F1702, F2002 and F2105) and two showing only very modest F0 effects (Fig 4C). Reaction times were faster when the preceding non-target word had the same F0 as the target in 4/5 animals (Fig. S4C). Factors that did not influence reaction times - as assessed by the permutation test were the trial number and trial history factors (the previous trial was a catch trial / correct). Therefore, despite equivalent performance in inter and intra-trial roving trials by applying gradient boosted regression to the reaction time data we observe that ferrets' reaction times are faster when pitch provides a consistent streaming cue (Fig.4B,E).

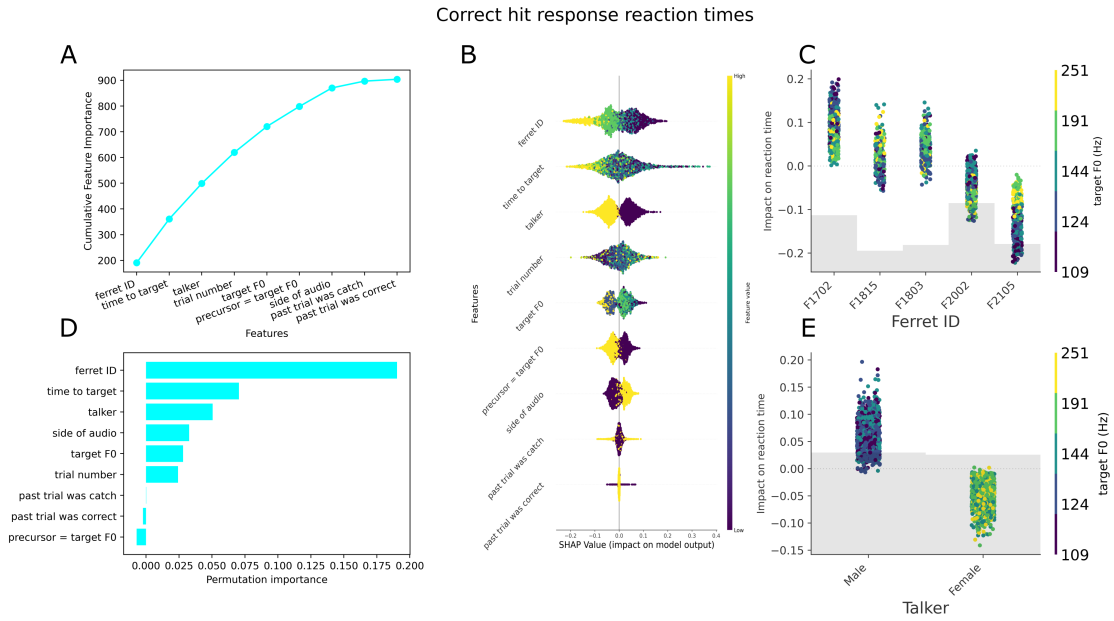


Figure 4. Reaction time models establish a contribution of F0 to target detection. A, feature importances of the hit model; B, SHAP summary plot of ranked feature SHAP values of each factor in the reaction time model; C, partial dependency plot of SHAP impact versus ferret ID color-coded by target F0; D, permutation feature importance of each factor in the model; E, partial dependency plot of SHAP impact color-coded by the target F0.

Gradient boosted decision tree models reveal systematic false alarms to some non-target words

Our false alarm model implied that false alarms were potentially lapses in concentration related more to timing than to acoustic parameters. However, an alternative possibility is that particular words drive false alarms, independently of the characteristics of the talker. To investigate this, we used gradient-boosted decision trees to ask whether subjects consistently false alarmed to particular non-target words by modeling the animals' response time within a trial based on the word token. We modeled data from the female talker and the male talker separately using only the timing of each word token in a trial, relative to the onset of the trial, to predict the animals' eventual response time (again relative to the onset of the trial rather than the target word as in the previous analysis). The prediction accuracy of this model was high for both talker types, with a train mse of 0.1329s and a test mse of 0.1331s for the female talker, 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).

Reassuringly, in both models, the presence and timing of the target word had the strongest predictive power about when animals would release from the centre port (Fig.5A-D). Nonetheless, some words consistently elicited behavioral responses, suggesting that false alarms are not simply temporary lapses in attention but rather that some words are perceived as more similar to the target. Supporting the idea that subjects are largely invariant to changes in pitch, many of the same non-target words consistently elicited false alarms across subjects Running models on each animal separately (Fig. 5E,F) confirmed that these were repeatable errors across ferrets and talkers. The commonly false alarmed words all shared a mixture of high-frequency consonants and voiced vowel elements and could not be explained by simply duration of power in a particular frequency band (Fig.S5) or in differences in the frequency of occurrence of each word.

DISCUSSION

We describe a novel behavioral task in which animals are trained to recognize a target word embedded in a series of non-target words and employed gradient-boosted decision trees and regression to analyze the subsequent behavior. The results of these models allowed us to understand that, like humans, ferrets are able to form F0-tolerant representations of auditory objects and use F0 to link sounds together into auditory streams. (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).

The data presented here, in which pitch made only a minor contribution to overall performance, supports previous behavioral work in animals, showing that non-human listeners can generalize across variations in F0 for relatively simple sounds. 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, Walker, et al. 2013, Town et al. 2015). Both rats (Engineer et al. 2013) and zebra finches (Ohms et al. 2010) trained to discriminate human speech sounds can generalize across different talkers who naturally vary in their voice pitch. In our models, F0 had only a very small effect on the ability of animals to correctly identify a target word (Fig.2) or on their likelihood of making a false alarm (Fig.3). Together, these results suggest that performance is robust across variations in pitch. Our reaction time models suggest that variation in F0 differently impacts individual animals. One benefit of the models developed in this study is that such individual differences can be explored and potentially taken into account when interpreting and analyzing brain signals.

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 Klump 2004, Hulse, MacDougall-Shackleton, 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 into streams 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. The impact of pitch roving in increasing the likelihood of false alarms and in slowing reaction times is consistent with ferrets using common pitch to link together sounds over time, offering an advantage for subsequent word recognition. We predict that the impact of removing pitch constancy might be more strongly evident in tasks that require separating competing streams. Here, we demonstrate that gradient-boosted decision trees have high predictive power even when incorporating highly correlated variables and are ideally suited for unpicking multiple contributing factors to behavior. Moreover, this gradient-boosted regression tree method allows us to be agnostic to how factors in our metadata are related to each other and thus presents an excellent way to conduct both hypothesis-driven and exploratory data analysis to uncover otherwise hidden trends in behavioral data and drive analysis. Overall, these findings from these sensitive and powerful models could inform later behavioral and neural data studies by giving us an idea of which behavioral factors impact decision-making in individual animals.

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 the behavioral task, using water as a reward. During testing periods, animals were water regulated. Animals were 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, through a combination of task performance and supplementation with a wet mash 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)

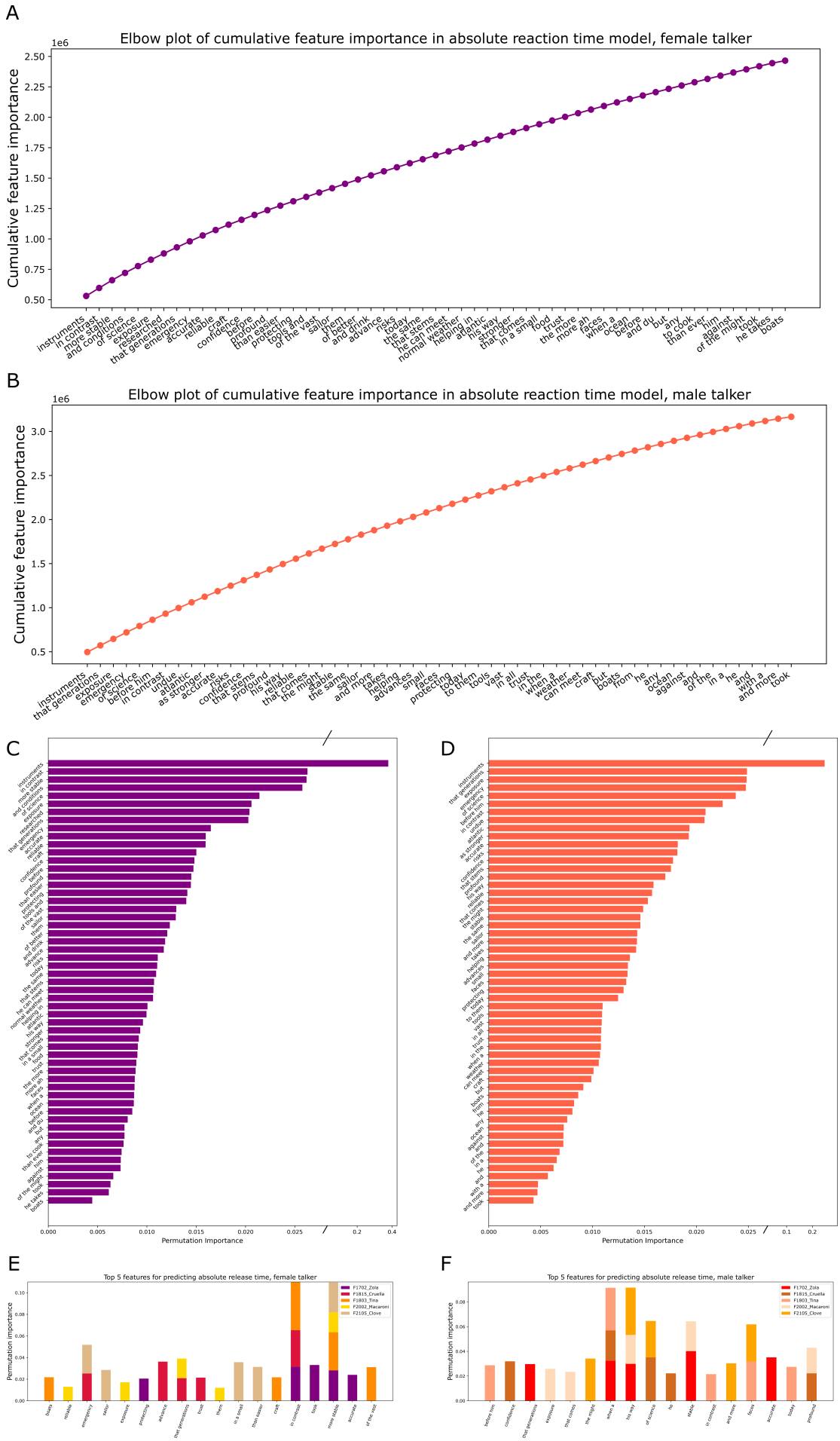


Figure 5. A, elbow plot of cumulative feature importance in the female talker model; B, same as A but for the male talker; C permutation importance (100 shuffles) of features included in the female talker model; D, same as C but for the male talker; E, top 5 permutation importances for each individual animal model for the female talker model; F, same as E but for the male talker.

Equipment

We controlled the task and stimulus presentation through an RZ6 controller (Tucker Davis Technology, Florida, USA) using OpenEx with custom-written software (Town et al. 2015) on a Windows PC. 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

Stimuli were composed of a sequence (or 'stream') of consecutively presented words all of which came from the same talker. Continuous speech from two talkers (1 male, 1 female) reading the same passage from the SCRIBE database was manually segmented into words and linked together with a minimum gap of 0.03s ms between words. The audio files were recorded at 20000 Hz but upsampled to 24,414 Hz for presentation.

Task

In a sound discrimination task, we trained five ferrets to recognize the target stimulus (the word 'instruments') against 54 other non-target stimuli (which were also English words) in a stream. Each stream (or string of words) consisted of a series of non-target words and one occurrence of the target word, which could occur anytime from 500 ms to 6.5 s after the onset of the trial (with the target timing drawn from a uniform distribution). As well as being preceded by non-target words, the target was followed by a sequence of non-target words that exceeded the duration of the response time (2s, see below). Streams were constructed de novo at the start of each trial with non-target words drawn randomly (with replacement) from the pool of 54. Non-target words were chosen at random (from a set of 54 words per talker).

The whole trial was comprised of word tokens from the same talker and presented from either the left or right speaker. Once trained, animals were required to initiate a trial by nose-poking at a center port that contained an infrared sensory and water delivery system. They were required to maintain contact until the target was presented. Once the target sound was presented, they were required to move to the response port on the same side as the stimulus presentation. A correct response required the animal to release the center port within 2s of the target word onset and correctly lateralize the sound stream (although, in practice, animals rarely made localization errors). 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

Initially, ferrets were trained to move between the 3 lick ports (left, center, and right side) by alternating water reward at each port. Once this was accomplished (usually within 1 to 2 sessions) they were trained to lateralize the target sounds ('instruments'). This was achieved by rewarding the initiation of a trial (a response at the center port) and presenting several repetitions of the target sound from one of the lateral locations (either left or right). The ferret would receive a second reward only if they responded at the corresponding location. Once ferrets could perform this target lateralization task at a high rate of performance (>90% correct) over =>2 sessions, the delay between initiating the trial and presenting the target word was systematically increased (from 0 to 5 seconds) between sessions (but only if performance remained above 80% correct for the last two sessions). Once the ferret was capable of waiting 5 seconds at the center port for target presentation and accurately lateralizing the stimulus, we reduced the target presentation to a single-word token. We 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 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 the talker identity was randomly drawn across trials. 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 (the same port 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 trials without an attenuation cue (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 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 shifted up or down, whereas, in intra-trial roving, the F0 value of each word was randomized. As in training, all word tokens within a trial came from the same talker.

Data Analysis

Any trial has four possible outcomes: hit, correct response, miss, and false alarm. A hit was defined as moving away from the center port ('release') and responding at the target location within 2s of the target word presentation. A correct rejection was defined as remaining at the central port for the entire duration of the trial (on a catch trial), a miss as failing to leave the central port within 2s of the target word presentation, and a false alarm as releasing from the center port before target word presentation or the end of a catch trial. False alarms immediately terminated the sound presentation and elicited a time-out (signaled by a modulated noise burst). Time outs lasted 2 seconds, during which the ferret could not reinitiate a 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 (C.R.) 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 release time rather than the lateral response time relative to the timing of the target word. To analyze whether word tokens systematically elicited behavioral responses, we defined the response time as the exit time from the central port relative to trial onset. All data analysis, from behavioral 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 behavioral 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 could 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 behavioral 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 and generalized linear models are commonly used alternatives that allow trial-based analysis of categorical or continuous behavioral data. While powerful, such models can fail to capture non-linear or non-monotonic relationships that might be present in behavioral data. Machine learning approaches offer an alternative model-free approach to uncovering statistical structure in rich behavioral data sets such as those typical of animal behavioral work. Models were generated using LightGBM (Ke et al. 2017). Gradient-boosted regression trees were used to model reaction time data. Gradient-boosted decision trees were used to make classification models for binary trial outcomes (hit vs. miss and false alarm vs. correct rejection). To optimize hyperparameters for this model, we implemented a grid search using optuna (Akiba et al. 2019).

We generated 5 models to address our research questions. Two classification models were developed; one considered determining whether a ferret missed a target word (miss vs. hit model), and the second considered the factors that influenced the likelihood of a false alarm/correct rejection of a non-target word (false alarm/correct reject model). Our reaction time model used gradient-boosted regression to determine the parameters influencing the animals' reaction time to the target word. Our response time models (one each for male and female talker trials) predicted the release time within a trial based on the timing of the words. They were used to assess whether animals made systematic errors with particular words.

We determined which features were significant using permutation testing, which shuffles a feature of our data (e.g., the target F0) and then selects the drop in performance the model has due to that feature being shuffled. We generated permutation importance plots from the sci-kit learn (sklearn) package to extract the model's attributes that would decimate the model's performance if they were shuffled, thereby establishing which features contributed significantly to model performance. 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 minimized binary log loss, whereas, for the regression (reaction time) model, hyperparameter optimization minimized the mean-squared error (l2 loss function).

The regression models' test and train mean-squared error was calculated using 5-fold cross-validation. The train and test accuracy and balanced accuracy were calculated using 5-fold cross-validation for the classification models.

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.

For the regression model that calculated the absolute response time irrespective of whether the response was correct, we used sub-sampling to create a uniform distribution of words. This sub-sampling, or bootstrapping, was done so our gradient-boosted regression tree model wouldn't associate higher-frequency words with a higher likelihood of a false alarm or response just because of its higher frequency. However, this is mathematically impossible to do precisely, as the words were not presented independently. In other words, each trial consisted of multiple word tokens analogous to a sentence, pooling each word from a word bank sampled with replacement. Moreover, in F1702, some of the words were programmed to occur 80% more frequently than other words for neural recordings. Thus, to achieve something close computationally to a mathematically perfect bootstrapping procedure, we created a loop for each of the 54 non-target words, found the trials that contained that non-target word, and placed them into a data frame. We then sub-sampled this resulting data frame to 700 samples (the minimum number of counts across all words in the original data frame) unless the non-target was a naturally high-frequency occurring word, where it was sub-sampled to 50 samples or skipped entirely. After all 54 words were iterated through in order, the resulting sub-sampled data frame was appended to an array. Next, we repeated the same process but went through the non-target words in reverse order to ensure some words wouldn't be over-sampled in the resulting distribution. This whole process of iterating through all the non-target words and flipping the order of iteration was repeated 18 more times (Figure 5S, the results obtained implementing this subsampling were very similar to those obtained using the natural distribution of word occurrences). 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). All code used to perform this analysis is available on GitHub (link will be inserted on publication).

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 JL were responsible for the data collection. CG and JB wrote the manuscript. All authors have read and agreed to the published version.

DATA AVAILABILITY

Behavioral 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).

Ashwood, Zoe C. et al. (Feb. 2022). "Mice alternate between discrete strategies during perceptual decision-making". en. In: *Nature Neuroscience* 25.2. Number: 2 Publisher: Nature Publishing Group, pp. 201–212. ISSN: 1546-1726. DOI: 10.1038/s41593-021-01007-z. URL: https://www.nature.com/articles/s41593-021-01007-z (visited on 09/04/2023).

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. and Yale E. Cohen (Oct. 2013). "The what, where and how of auditory-object perception". en. In: *Nature Reviews Neuroscience* 14.10. Number: 10 Publisher: Nature Publishing Group, pp. 693–707. ISSN: 1471-0048. DOI: 10.1038/nrn3565. URL: https://www.nature.com/articles/nrn3565 (visited on 07/17/2023).

Bizley, Jennifer K., Kerry M. M. Walker, 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.

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.

Griffiths, Timothy D. and Jason D. Warren (Nov. 2004). "What is an auditory object?" en. In: *Nature Reviews Neuroscience* 5.11. Number: 11 Publisher: Nature Publishing Group, pp. 887–892. ISSN: 1471-0048. DOI: 10.1038/nrn1538. URL: https://www.nature.com/articles/nrn1538 (visited on 07/17/2023).

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". eng. In: *Neural Computation* 17.9, pp. 1875–1902. ISSN: 0899-7667. DOI: 10.1162/0899766054322964.

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.

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.

Roy, Nicholas A. et al. (Feb. 2021). "Extracting the dynamics of behavior in sensory decision-making experiments". eng. In: *Neuron* 109.4, 597–610.e6. ISSN: 1097-4199. DOI: 10.1016/j.neuron.2020.12.004.

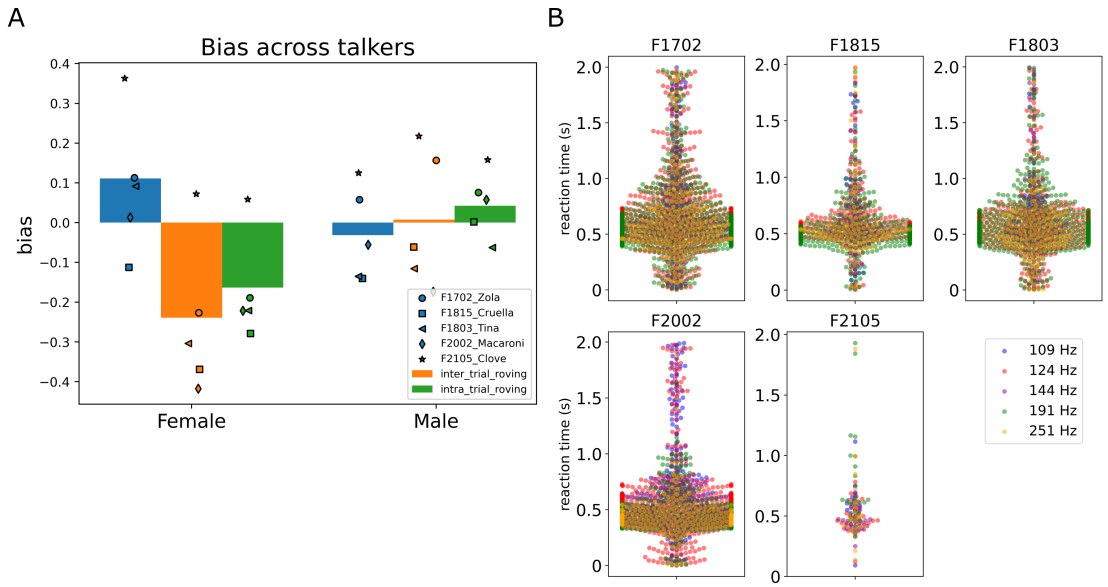
Slonina, Zuzanna A., Katarina C. Poole, and Jennifer K. Bizley (Jan. 2022). "What can we learn from inactivation studies? Lessons from auditory cortex". In: *Trends in Neurosciences* 45.1, pp. 64–77. ISSN: 0166-2236. DOI: 10.1016/j.tins.2021.10.005. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8897194/ (visited on 07/17/2023).

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).

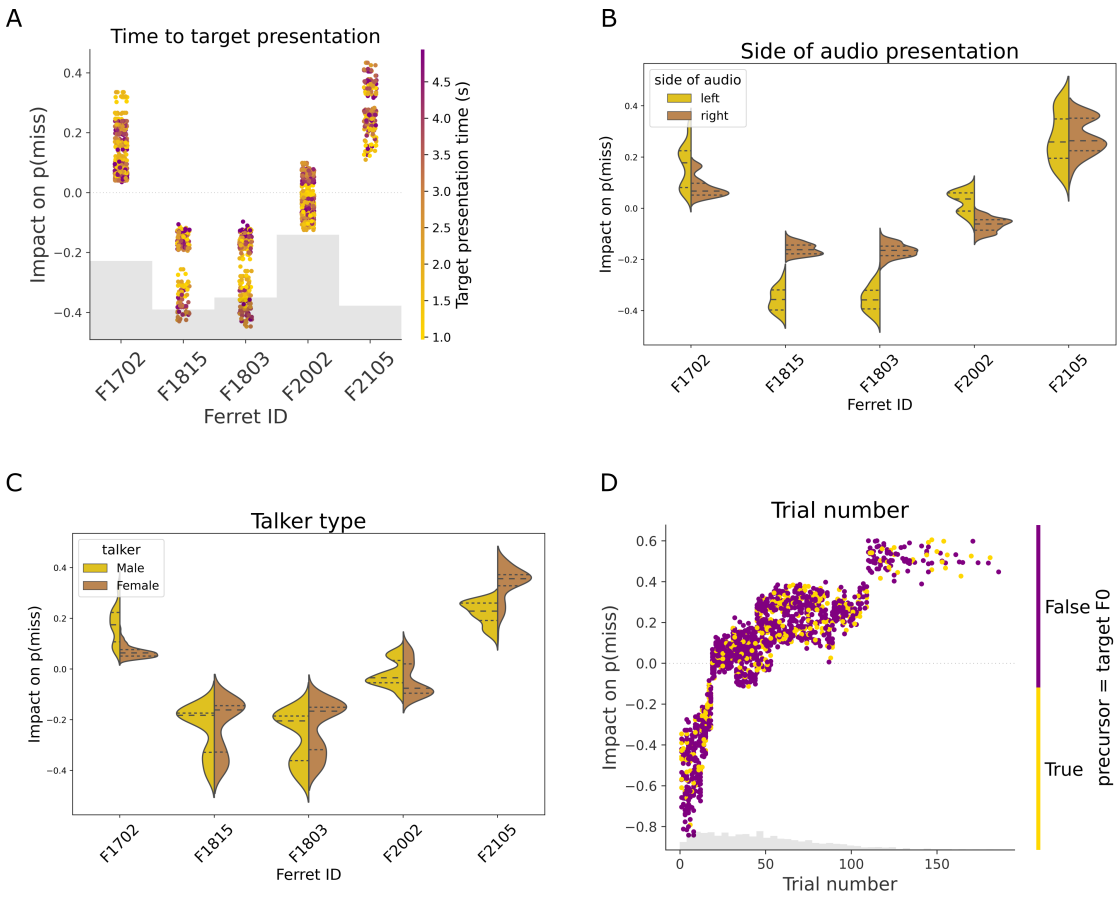
Town, Stephen M. et al. (May 2015). "The role of spectral cues in timbre discrimination by ferrets and humans". In: *The Journal of the Acoustical Society of America* 137.5, pp. 2870–2883. ISSN: 0001-4966. DOI: 10.1121/1.4916690. URL: https://doi.org/10.1121/1.4916690 (visited on 09/21/2023).

SUPPLEMENTARY MATERIAL

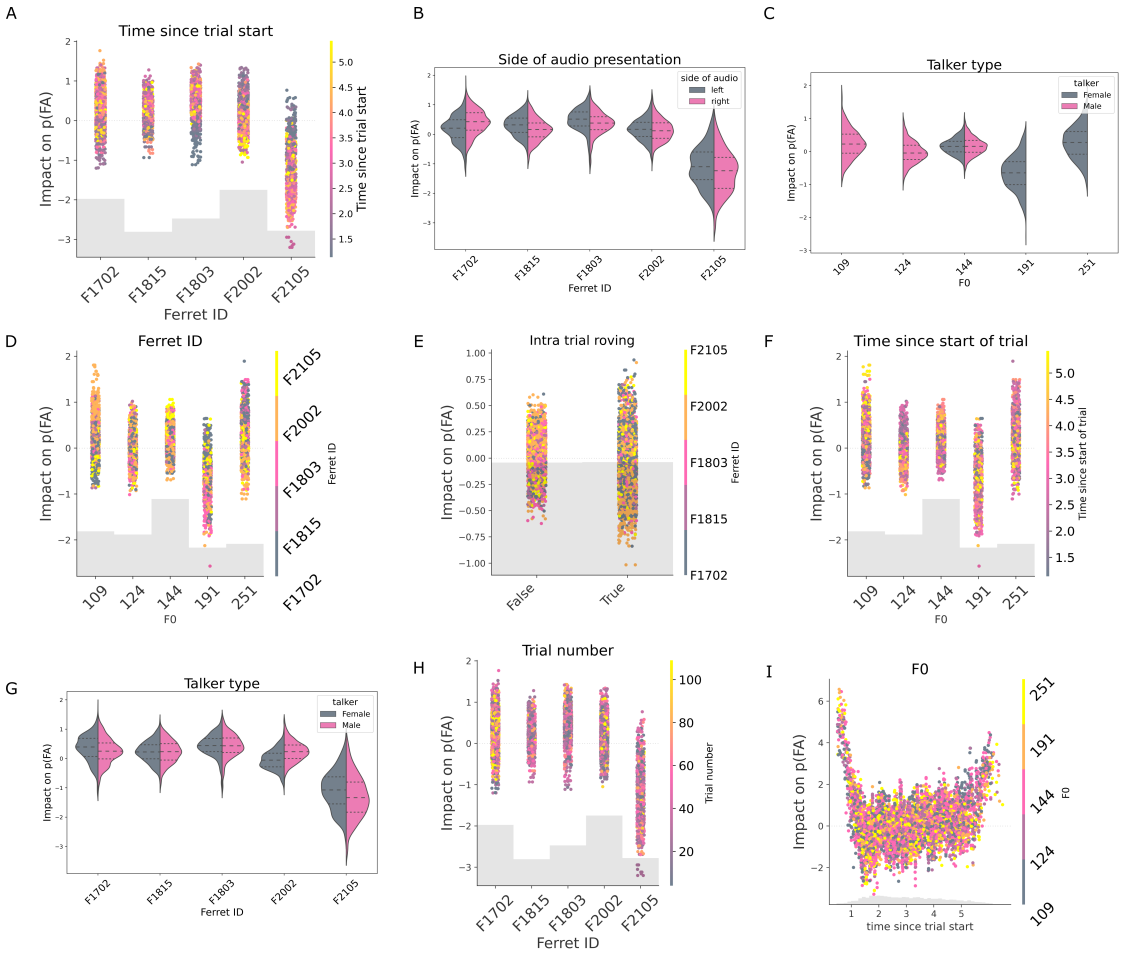
Supplementary Figures



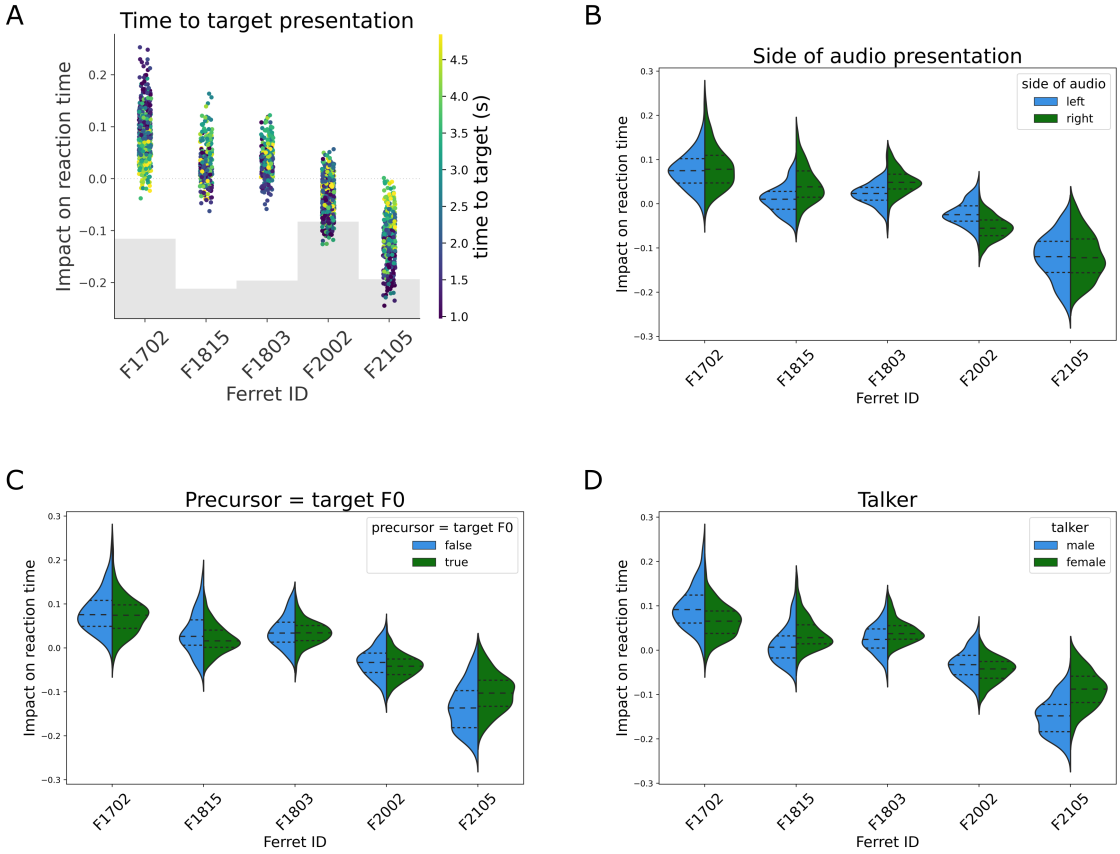
Supplementary Figure S1. A, bia across trial conditions and talker types; B, reaction times of each animal for correct responses color-coded by F0 of the target word.



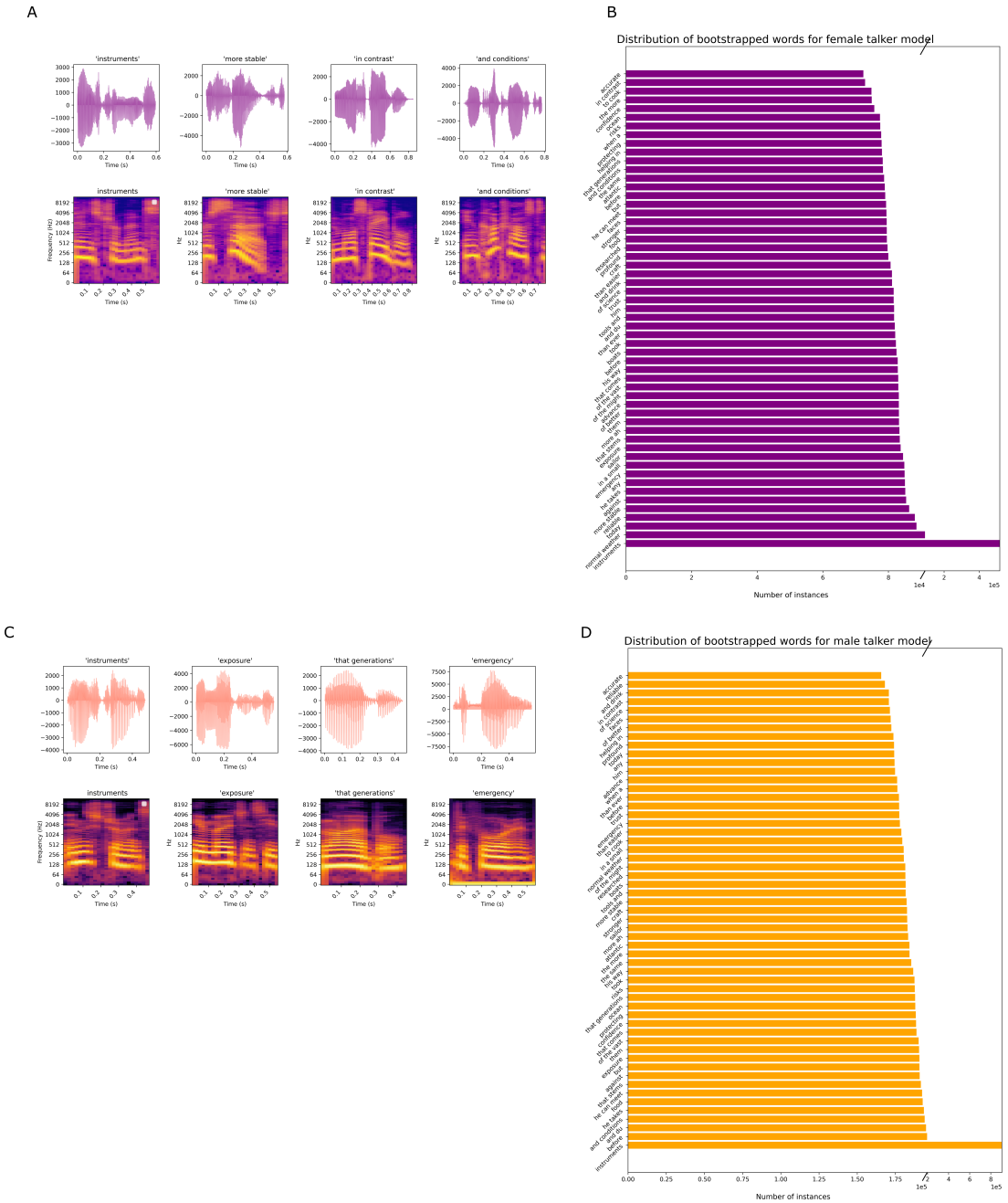
Supplementary Figure S2. Partial dependency plots for the correct hit response/miss response model. A; SHAP values over the ferret ID color-coded by target presentation time; B, SHAP values over ferret ID color-coded by the side of audio presentation; C, same as B but color-coded by talker type; D, SHAP values over trial number color-coded by whether the trial had the precursor word F0 equal to the target F0.s



Supplementary Figure S3. Partial dependency plots for the correct reject/false alarm model. —A, partial dependency plot depicting the mean SHAP impact over the ferret ID color-coded by time within the trial; B, violin plot of the SHAP value over the ferret ID color-coded by the side of audio presentation; C, violin plot of the SHAP values over the F0 of the trial color-coded by talker type; D; swarm plots of SHAP values over F0 color-coded by ferret ID; E, SHAP values over the F0 of the stream color-coded by trial number; F, same as E but color-coded by time since the start of the trial; Note that while the 191Hz F0 is associated with a higher false alarm rate, this should be interpreted in the context of the much lower FA rate associated with the female talker. G, violin plot of the SHAP value over ferret ID color-coded by talker type; H, SHAP value over ferret ID color-coded by trial number; I, SHAP value over trial duration color-coded by F0.



Supplementary Figure S4. Correct hit response reaction time model partial dependency plots. A, SHAP values over the ferret ID color-coded by the time to target presentation; B, violin plot of the SHAP value over ferret ID color-coded by the side of audio presentation; C, same as B but color-coded by whether the precursor word's F0 was the same as the target word's F0; D, same as C but color-coded by the talker type for the trial.



Supplementary Figure S5. A, amplitude waveforms (top) and corresponding spectrograms (bottom) of the top five most frequently false-alarmed to non-target words in the absolute reaction time model; B, distribution of bootstrapped words in the absolute reaction time model; C, same as A but for the male talker absolute reaction time model; D, same as B but for the male talker absolute reaction time model.

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	0.19088470325102014
subsample	0.9304141034109051
learning_rate	0.477510574908984
num_leaves	115
max_depth	18
min_child_samples	53
reg_alpha	0.2609865674187428
reg_lambda	1.3303484138905937
min_split_gain	0.0007545705453046434
bagging_freq	17
feature_fraction	0.8423132245598192
scale_pos_weight	1.113987259898614
min_child_weight	7.651492399061174
max_bin	786
min_data_in_leaf	52
min_sum_hessian_in_leaf	2.9989911144951256

Supplementary Table S2. Hyperparameter values for the false alarm categorical model

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

Hyperparameter	Value
colsample_bytree	0.998 448 361 791 188 9
alpha	10.545 892 165 925 359
n_estimators	120
learning_rate	0.258 529 884 871 212 1
max_depth	20
bagging_fraction	1.0
bagging_freq	23
lambda	0.195 381 053 380 844 05
subsample	0.895 804 443 430 478 9
min_child_samples	20
min_child_weight	9.474 782 393 947 127
gamma	0.157 117 421 509 215 9
subsample_for_bin	6200

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.

Absolute reaction time models - ferret ID specific

Ferret ID	Talker Type	Alpha	Bagging Fraction	Bagging Freq	Colsample bytree	Learning Rate	Max Depth	N Estimators
F1702	Female	14.223322406	0.9	15	0.4376317124943425	0.2950728851506002	17	72
F1702	Male	13.683546599	0.9	27	0.274885701704623	0.2996939725741514	8	92
F1815	Female	7.971900722	0.8	3	0.3400445377182091	0.2767681300147410 4	18	94
F1815	Male	8.423806867	0.9	17	0.2072996759265816 3	0.1561952141128 81	18	70
F1803	Female	12.715761382	0.9	3	0.3900419851696171	0.2999743040074189	17	86
F1803	Male	9.372039068	0.9	26	0.3367889760120146 3	0.2305005540632021	10	94
F2002	Female	14.643599551	0.8	7	0.5003266555294205	0.2493528255372144 7	12	96
F2002	Male	7.975250677	0.9	1	0.4229110961290357	0.2617874794508851	20	70
F2105	Female	13.306184442	0.9	16	0.2449488493150719 6	0.2952666033482773 7	15	100
F2105	Male	5.599860028	0.1	0	0.1294417170969843	0.2411178305650324 1	20	42

Supplementary Table S7. Hyperparameters for the absolute reaction time models for each ferret ID broken down by Female/Male talker type.