Published in final edited form as:
J Acoust Soc Am. 2015 May 1; 137(5): 2870–2883. doi:10.1121/1.4916690.

The role of spectral cues in timbre discrimination by ferrets and humans

Stephen M Town, **Huriye Atilgan**, **Katherine C Wood**, and **Jennifer K Bizley**
Ear Institute, University College London, 332 Gray's Inn Road, London, WC1X 8EE, United Kingdom

Abstract

Timbre distinguishes sounds of equal loudness, pitch and duration; however little is known about the neural mechanisms underlying timbre perception. Such understanding requires animal models such as the ferret in which neuronal and behavioral observation can be combined. The current study asked what spectral cues ferrets use to discriminate between synthetic vowels. Ferrets were trained to discriminate vowels differing in the position of the first (F1) and second formants (F2), inter-formant distance and spectral centroid. In Experiment One, ferrets responded to probe trials containing novel vowels in which the spectral cues of trained vowels were mismatched. Regression models fitted to behavioral responses determined that F2 and spectral centroid were stronger predictors of ferrets' behavior than either F1 or inter-formant distance. Experiment Two examined responses to single formant vowels and found that individual spectral peaks failed to account for multi-formant vowel perception. Experiment Three measured responses to unvoiced vowels and showed that ferrets could generalize vowel identity across voicing conditions. Experiment Four employed the same design as Experiment One but with human participants. Their responses were also predicted by F2 and spectral centroid. Together these findings further support the ferret as a model for studying the neural processes underlying timbre perception.

I Introduction

Timbre is the perceptual feature that distinguishes sounds of equal pitch, loudness and duration. Timbre plays an important role in the discrimination of musical and environmental sounds (McAdams, 1999; McAdams *et al.*, 2010), animal vocalizations (Fitch and Kelley, 2000; Reby *et al.*, 2005; Fitch and Fritz, 2006) and phonetic components of human speech such as vowels (Liberman *et al.*, 1967; Darwin, 2008; Moore, 2008; Town and Bizley, 2013). In the case of vowels, timbre is related to the spectral envelope of vowel sounds and, in particular, peaks in the spectral envelope known as formants. A formant is defined as a resonance introduced by the filtering properties of the vocal tract and vocalizations often contain multiple formants, with each formant numbered in ascending order according to the frequency position of the peak. As such, formants are not restricted to human vocalizations and formant-related timbre perception may play a significant role in the behaviors of birds and mammals (Fitch and Kelley, 2000; Reby *et al.*, 2005; Charlton *et al.*, 2009).

In the case of vowel timbre perception by humans, the frequency positions of the first, second and third formant peaks in the spectrum (F1, F2 and F3 respectively) differ between spoken vowels (Potter and Peterson, 1948; Peterson and Barney, 1952) and provide valuable

Europe PMC Funders Author Manuscripts

cues for vowel identification and discrimination (Delattre *et al.*, 1952; Plomp *et al.*, 1967; Pols *et al.*, 1969; Klatt, 1982; Molis, 2005; Swanepoel *et al.*, 2012). Models of vowel discrimination and timbre perception have also been proposed that rely upon the distance between adjacent formants (Miller, 1989) or the spectral centroid (spectral center of gravity) of sounds (Chistovich and Lublinskaya, 1979; McAdams, 1999). There are thus a variety of acoustic cues that may contribute to our perception of vowel timbre. It is worth noting however that models of vowel perception based on formant positions and other spectral parameters are not completely independent, and so the perception of vowel timbre may ultimately involve a number of inter-related acoustic cues that offer redundancy in different listening conditions (Kiefte *et al.*, 2013).

Despite existing knowledge about the acoustic cues that influence vowel timbre, relatively little is understood about the neural mechanisms underlying timbre perception. To elucidate such physiological processes, animal models are needed in which it is possible to study single neuron activity during behavior. Many animals are capable of discriminating vowels (see Town and Bizley, 2013 for review) and several studies in non-human species have demonstrated the importance of formants as cues in discrimination (Kojima and Kiritani, 1989; Sinnott *et al.*, 1997; Ohms *et al.*, 2012). Recently the study of vowel discrimination has been extended to the ferret (Bizley *et al.*, 2013b) – a model organism for examining neural processing in the auditory system (Fritz *et al.*, 2003; Bizley *et al.*, 2009; Sumner and Palmer, 2012; Bizley *et al.*, 2013a). Here we build upon this work by examining the contribution of formants and other acoustic cues to vowel discrimination by ferrets.

Previously Bizley *et al.* (2013b) demonstrated that ferrets can be trained to discriminate multi-formant vowels such as [u] and [ɛ]. However such sounds contain redundant cues for discrimination as several acoustic cues (e.g. F1 and F2) are available to distinguish between vowels, and so it is unclear which cues guide ferrets' behavior. Here, Experiment One attempts to extend our understanding of spectral cue use by ferrets: Animals were trained to discriminate two vowels that differed in both F1 and F2 ([u] and [ɛ] or [a] and [i]) in a two-alternative forced choice (2AFC) task. Following acquisition of the task, ferrets were presented with a variety of probe sounds in which spectral cues (F1, F2 and inter-formant distance) from the trained vowels were systematically mismatched in order to create conflicts between acoustic cues. How ferrets respond to such probe sounds and resolve such conflicts should indicate whether any particular cue was disproportionately important for vowel discrimination and thus timbre perception.

Experiment Two measured the responses of ferrets to vowel sounds with only a single formant. In human listeners single formant vowels can evoke similar percepts to those elicited by multi-formant vowels with closely positioned first and second formants (e.g. /a u/), suggesting that listeners perform a spectral averaging process (Delattre *et al.*, 1952). This averaging may integrate across higher frequency formants (e.g F2, F3 and F4; Fant and Risberg, 1963; Carlson *et al.*, 1970; Xu *et al.*, 2004) and depend on relative formant amplitude, leading to a center-of-gravity effect (Chistovich and Lublinskaya, 1979). A previous study, in which ferrets were trained to discriminate multi-formant vowels ([u] and [ɛ]), demonstrated that appropriate single formant vowels can evoke similar behavioral responses to the trained vowels (Bizley et al., 2013b). In this study, the effective formants

Europe PMC Funders Author Manuscripts

that elicited such responses were positioned at the F1 and F2 of [u] and [ɛ] respectively, whereas formants at the alternative positions (i.e. F2 of [u] and F1 of [ɛ]) were apparently classified randomly. These findings suggest that each formant may contribute significantly to ferrets' perception of timbre, but it is unclear whether this contribution matches that revealed using multi-formant mismatch vowels as in Experiment One. The aim of Experiment Two was to resolve this ambiguity by measuring the single formant responses (i.e. repeating the experiment of Bizley and colleagues (2013b)) of ferrets for which spectral cue use was also characterized using mismatch vowels in Experiment One.

Experiment Three examined the importance of harmonic structure to ferret timbre perception by measuring animals' discrimination of voiced from voiceless vowels. When uttered, vowels can be voiced when the vocal cords oscillate (giving rise to periodic sounds that are perceived as having a pitch) or can be unvoiced as when whispered. Human listeners can recognise spectral timbre across voicing conditions but it is unclear whether other animals, including ferrets, also show this form of perceptual invariance.

The study of timbre perception in ferrets (Experiments 1-3) necessarily requires the design of psychophysical tasks that deviate from classical paradigms developed for humans. This is because animals cannot be verbally instructed how to perform the task and thus must learn through trial and error. It is also beneficial to use relatively simple stimuli (at least in the first instance) such as steady-state artificial vowels that the animal will rapidly learn to discriminate. Task demands can influence behavior and so confound comparisons between species. To minimise such methodological confounds it is important to characterize human timbre perception in a task matched as closely as possible to that in which ferrets were trained and tested. In Experiment Four, human listeners were trained to discriminate vowels in a 2AFC task and subsequently presented probe sounds with mismatched spectral cues, as ferrets were in Experiment One. This enabled characterization of spectral cue use by humans under comparable conditions to the animal model in question.

II Methods

A Experiments 1-3 Ferret Psychophysics

1 Subjects—Subjects were eight female adult pigmented ferrets, housed in groups of between two and eight with free access to high-protein food pellets and water. During training, access to water was restricted to training / testing sessions, in which it served as a positive reinforcer for task performance. Training runs lasted five days or less with water being removed on the day before the beginning of a run and being returned on the last day of the run. Training runs were separated by two or more days in which ferrets had free access to water. On each day of training and testing, subjects received a minimum of 60 ml/kg of water either during task performance or supplemented as a wet mash made from water and ground high-protein pellets. The weight and water consumption of all animals was measured throughout the experiment. Regular otoscopic examinations were made to ensure the cleanliness and health of ferrets' ears. All experimental procedures were approved by a local ethical review committee and performed under licence from the UK Home Office and in accordance with the Animals (Scientific Procedures) Act 1986.

2 Apparatus—Ferrets were trained to discriminate sounds in a customized pet cage (80 cm x 48 cm x 60 cm, length x width x height) in which the walls were made from wire surrounded by sound-attenuating foam. The floor of the cage was made from plastic, with an additional plastic skirting into which three spouts were inserted. Each spout contained an infra-red sensor (OB710, TT electronics, UK) that detected nose-pokes and an open-ended tube through which water could be delivered (Fig. 1A).

Sound stimuli were presented through two loud speakers (Visaton FRS 8) positioned on the left and right sides of the head at equal distance and approximate head height. These speakers produce a flat response (±2 dB) from 200Hz to 20 kHz, with an uncorrected 20 dB drop-off from 200 to 20 Hz when measured in an anechoic environment using a microphone positioned at a height and distance equivalent to that of the ferrets in the testing chamber. An LED was also mounted above the center spout and flashed (flash rate: 3 Hz) to indicate the availability of a trial. The LED was continually illuminated whenever the animal successfully made contact with the IR sensor within the center spout until a trial was initiated. The LED remained inactive during the trial to indicate the expectation of a peripheral response and was also inactive during a time-out following an incorrect response.

The behavioral task, data acquisition, and stimulus generation were all automated using custom software running on personal computers, which communicated with a TDT RM1 real-time signal processor (Tucker-Davis Technologies, Alachua, FL).

3 Acoustic stimuli—Stimuli were artificial vowel sounds synthesized in MATLAB (Mathworks, USA) based on an algorithm adapted from Malcolm Slaney's Auditory Toolbox (https://engineering.purdue.edu/~malcolm/interval/1998-010/). The adapted algorithm simulates vowels by passing a sound source (either a click train, to mimic a glottal pulse train in Experiments One and Two, or broadband noise in Experiment Three) through a biquad filter with appropriate numerators such that formants are introduced in parallel (Smith, 2007). In the current study, four formants (F1-4) were modelled and the resulting spectra of artificial vowel sounds are illustrated in Figure 2 for an F0 of 200 Hz. In experiments with both ferrets and naïve human listeners, half of subjects were trained to discriminate [u] (F1-4: 460, 1105, 2857, 4205 Hz) from [ɛ] (730, 2058, 2857, 4205 Hz) while the other half of subjects were trained to discriminate [a] (936, 1551, 2975, 4263 Hz) from [i] (437, 2761, 2975, 4263 Hz). Selection of formant frequencies was based on previously published data (Peterson and Barney, 1952; Bizley *et al.*, 2009) and synthesis produced sounds consistent with the intended phonetic identity. The parameters used to synthesize training stimuli differed only in the center frequencies of the first and second formants while the positions of the third and fourth formants were fixed at 2857 and 4205 Hz for subjects discriminating [u] and [ɛ], and 2975, 4263 Hz for subjects discriminating [a] and [i]. These values reflect the midpoints between the formant positions for each pair of vowels as used in previous work (Bizley *et al.*, 2009). Formant bandwidths were kept constant at 80, 70, 160 and 300 Hz (F1-4 respectively) as in previous work (Bizley *et al.*, 2009; Bizley *et al.*, 2013b). Vowel fundamental frequency was determined by the repetition rate of the click train and was set at 200 Hz unless otherwise stated. All sounds were ramped on and off with 5 ms cosine ramps and all sound tokens were presented for 250 ms. Sound levels were calibrated using a Brüel & Kjær (Norcross, USA) sound level meter and free-

field [1/2] inch microphone (4191). Vowel sounds were presented at 70 dB SPL. During testing, the level of trained vowels was varied randomly over a 6 dB range (i.e. 70 ± 3 dB SPL).

In Experiment One, probe sounds (mismatch vowels) were created by swapping the F1 and F2 values of training vowels (Fig. 2Bi). For example, subjects trained with [a] (F1, F2: 936 Hz, 1551 Hz) and [i] (437, 2761) were presented with two probe sounds: (936, 2761) and (437, 1551). Additional probe stimuli were generated by swapping F1 and inter-formant distance (IFD) (pegged to F1 e.g. 437, 1052) (Fig. 2Bii) or, F2 and inter-formant distance (pegged to F2 e.g. 2146, 2761) (Fig. 2Biii). Mismatch vowels in which F1 would be negative were excluded, leaving a total of 5 probe stimuli for each pair of learned vowels. Formant frequencies for the full set of mismatch stimuli for [u] and [ɛ] are shown in Fig. 3. In Experiment Two, single formant vowels were synthesized by selecting a single value (either the F1 or F2 value of each trained vowel) at which to introduce a spectral peak. In both Experiments 1 and 2, probe sounds were presented at a fundamental frequency of 200 Hz. In Experiment Three, voiceless vowels were synthesized in the same way as voiced vowels with the exception that the sound source was changed from a click train to broadband noise. For both single formant and voiceless vowels, sound level was roved over the same 6 dB range as for trained vowels (i.e. 70 ± 3 dB SPL). The spectra of stimuli in each experiment were confirmed from Fourier analysis of sound recordings in the experimental chamber as shown in the Appendix.

4 Training—Subjects were trained to discriminate a pair of vowels ([u] and [ɛ] or [a] and [i]) through a series of stages of increasing difficulty. When first introduced to the training apparatus, animals were rewarded with water if they visited any spout. Once subjects had associated the spouts with water, a contingency was introduced in which the subject was required to hold the head at the central spout for a short time (501 – 1001 ms) before receiving a reward. The central spout activation initiated a trial period in which a nose-poke at either peripheral spout was rewarded.

Following acquisition of the basic task structure (typically 2-3 sessions), sounds were introduced. On each trial, two repeats of a single vowel sound (each 250 ms in duration with a 250 ms interval) were played at a variable delay (0 – 500 ms) after the animal first made contact with the spout. A trial was initiated if the subject's head remained at the spout for the required hold time, plus an additional 500 ms in which the first repeat of the sound and the interval were played. Following trial initiation, the pair of sounds were repeated until the ferret completed the trial by visiting the "correct" peripheral spout (e.g. left) to receive a reward. Nose-pokes at the "incorrect" peripheral spout (e.g. right) were not rewarded or punished at this stage and incorrect responses did not terminate trials. If the animal failed to visit the correct spout within a specified period after initiating a trial (25 – 60 seconds), that trial was aborted and the animal could begin the next trial.

Once animals were completing trials frequently, the consequences of incorrect responses were altered so that incorrect responses terminated the current trial. Subjects were then required to return to the center spout in order to initiate a correction trial in which the same stimulus was presented. Correction trials were included to prevent animals from biasing their

responses to only one spout, and were repeated until the animal made a correct response. For each session, performance was quantified as the percentage of trials (excluding correction trials) in which subjects made a correct response. After a minimum of two sessions in which errors terminated trials, a timeout (5-15 seconds) punishment was added to incorrect responses. Timeouts were signalled by a burst of white noise (100 ms) and the center spout was disabled for the duration of the timeout, preventing initiation of further trials.

Once subjects could discriminate looped sounds on consecutive sessions with a performance of ≥80%, looping of sounds was removed so that subjects were presented with only two vowel sounds during the initiation of the trial at the center spout. When ferrets correctly identified ≥80% of vowel pairs in two consecutive sessions, the fundamental frequency (F0) of vowels was roved across trials (F0 = 149, 200, 263, 330, 409 and 499 Hz). Initially the range of F0s used was limited to a subset of this range (e.g. 149 and 200 Hz) and gradually increased over sessions. When performance with all F0s exceeded 80% on two or more consecutive sessions, subjects moved on to testing.

5 Testing—In testing, subjects discriminated vowels across F0 as in training but probe sounds were presented on 20% of trials. In Experiment One, all subjects were presented with probe sounds consisting of mismatch vowels (n = 5) in which the first and second formants or inter-formant distance of vowels were mismatched.

In Experiment Two, two ferrets were presented with single formant vowels as probe sounds. Single formant vowels have only one peak in the spectral envelope set either at the frequency of the first or second formant of training vowels. As both subjects tested were trained to discriminate [u] and [ɛ], single formant peaks were positioned at 460 Hz, 730 Hz, 1105 Hz and 2058 Hz. Probe sounds in Experiments 1 and 2 were always generated with a fundamental frequency of 200 Hz and all responses to probe sounds were rewarded.

In Experiment Three, probe sounds were voiceless versions of the two trained vowels, presented to six ferrets. Five animals were initially reward for any response to probe voiceless vowels and in four of these animals, probe trials revealed accurate discrimination. However, for one subject it was necessary to reward responses according to the stimulus-response contingency used during training. It was also necessary to present probe sounds on 40% of trials to this subject in order to overcome a strong response bias of the animal to respond at the right spout on virtually all probe trials. For the sixth animal, accurate discrimination was observed when responses to probe sounds were unrewarded.

6 Analysis—In all testing sessions, correction trials were excluded from the analysis and only data from test sessions in which ferrets discriminated trained vowels with ≥ 80% accuracy were included in the analysis. While testing in Experiment One included a range of F0s for the trained vowels, only responses on trials on which vowels were generated with an F0 of 200 Hz were analysed as probe sounds were always generated with an F0 of 200 Hz.

In Experiment One, two lines of analysis were pursued: The first used signal detection theory (Green and Swets, 1966) to calculate the discriminability (d′) between all pairs of vowels (trained and mismatch probes) presented: For a given vowel pair X and Y, a response

at the right spout to sound X was considered as a hit while a response left was considered as a miss; a response at the left spout to sound Y was considered a correct rejection while a response right was considered a false positive. The probabilities of hits and false alarms were then normalized and subtracted to give d′. In this case, discriminability thus summarizes the degree to which a subject responded differently to two sounds but not whether the subject responded correctly. Since the assignment of sound to response side was arbitrary, the absolute (unsigned) discriminability |d′| was considered as our metric of discrimination. Values of |d′| near zero indicate that an animal responded similarly for two given sounds, whereas increasing values of |d′| indicate that subjects responded more frequently at different spouts for given sounds and so |d′| was large for trained vowel.

Values of |d′| were calculated for each pair of vowels presented and compared to the difference in F1, F2, inter-formant distance and spectral centroid between the vowels in question. Inter-formant distance was calculated simply as F2-F1 expressed in Hz on a linear scale, in keeping with previous studies of the neural representation of vowel spectra (Ohl and Scheich, 1997). Spectral centroid (SC) was defined as the integral across frequency bands (x) of the product of the band frequency ($f(x)$; measured on a linear scale in Hz) and the normalized energy in that band ($p(x)$; the steady-state spectrum [E, in dB] divided by the total sound energy across frequency bands):

$$SC = \int x \cdot p(x)\, dx \quad (1)$$

where

$$x = f(x) \quad (2)$$

$$p(x) = \frac{E(x)}{\sum_x E(x)} \quad (3)$$

This calculation was performed using the MIRtoolbox (Lartillot and Toiviainen, 2007). Linear regression was then used to summarize the degree to which each cue was related to vowel discrimination.

In a second line of analysis in Experiment One, the responses of each subject (left / right) were fitted with a multiple logistic regression model, assuming binomial distributions and using the logit link function (Matlab R2014a; glmfit function). The model was fitted with a variable number of predictors ranging from zero (constant model), to individual acoustic cues (e.g. F1), up to three predictors (e.g. F1+F2+SC). As F1 and F2 directly determine IFD, models combining all these parameters were over-parameterized and so not included in the analysis. Each model provided a measure of deviance from which it was possible to compare nested models (e.g. F1+F2 vs. F1) and test whether addition of a predictor significantly

improved model fit (analysis of deviance, χ^2 distribution, $p < 0.05$ adjusted for multiple comparisons using the Bonferroni correction). For each subject, we identified the most parsimonious models (i.e. those with fewest predictors) for which inclusion of additional predictors did not significantly improve model fit.

B Experiment Four Human Psychophysics

1 Participants—Participants were 10 naïve adults (5 male, mean age 26.3 years, SD 3.77) that had no previous experience of the specific synthetic vowel sounds used in the study, and 4 experienced adults (1 male, mean age 28.8 years, SD 3.06) that were familiar with the stimuli and task as they trained ferrets (and are authors). Of the naïve subjects, 4 were native English speakers whilst two were bilingual (English and Urdu) and four spoke English as a second language. Of the experienced subjects, three were native English speakers. Similar findings were observed for both native and non-native English speakers and so results were pooled across linguistic background. No subjects reported hearing loss or impairment. All subjects provided written informed consent in accordance with the Declaration of Helsinki. Ethical approval was provided by the UCL Research Ethics Committee.

2 Testing Procedure—Training and testing of subjects was designed to match as closely as possible the procedures used for ferrets. Subjects were seated in a sound isolating booth with the same speakers (Visaton FRS 8) positioned on the table top, to the subject's left and right. In contrast to experiments with ferrets, human listeners initiated trials and responded using a keyboard rather than IR sensors, and were given visual feedback on task performance rather than water reward or timeout punishments. As with the ferret task, data acquisition and stimulus generation were automated using the same custom software and TDT RM1 real-time signal processor.

On each trial, subjects initiated a trial by making a key press and were presented with two tokens of a vowel sound (2 x 250 ms with a 250 ms interval). Subjects were then required to respond using the left and right arrow keys and feedback was given to indicate whether the response was correct or incorrect (Fig. 1B). Incorrect responses led to correction trials in which the previous trial was repeated until a correct response was given. Half of naïve subjects (n = 5) and all experienced subjects were required to discriminate [u] from [ɛ]; the remaining naïve subjects discriminated [a] from [i]. To minimize differences between human and non-human experiments, subjects were given no explicit instructions beyond details of the trial structure described above. Like ferrets, human participants therefore had to learn by trial and error which sound was associated with which response.

Training consisted of blocks of 25 (non-correction) trials in which each vowel was presented at one of six F0s and sound level was varied between trials over a 6 dB range. Performance was measured as the percentage of correct responses (excluding correction trials) and was assessed at the end of each block. Training was considered complete when block performance exceeded 85% correct. This criterion was reached quickly by both naïve and experienced subjects (Naïve: median number of blocks to criterion = 2; Experienced: median number of blocks to criterion = 1).

Europe PMC Funders Author Manuscripts

Once subjects could discriminate between vowels, probe trials were introduced in the testing session in which the first and second formants or inter-formant distance of the training vowels were mismatched as described for ferrets. Briefly, subjects continued to discriminate the training vowels with roving F0 and sound level whilst probe trials were introduced on 20% of trials. Each mismatch vowel (n = 5) was presented on 25 trials over a 6 dB sound level range with a constant fundamental frequency (200 Hz). The testing session was split into five blocks of 121 trials (not including correction trials) and the order of probe trial presentation was randomized. All responses to probe stimuli were reported as correct to the subject. During testing, participants were required to continue discriminating the training vowels with 80% accuracy or greater. One subject failed to meet this performance criterion and so was excluded from the analysis. Human subjects were not presented with single formant or voiceless vowels described in the ferret studies in Experiments Two and Three.

3 Analysis—Task performance by human subjects was analysed using the same approach described previously for ferrets (Section II.A.6). Additionally, to compare across species a repeated-measures analysis of variance was conducted on regression coefficients obtained from measuring |d′| with cue (*Δ*F1, *Δ*F2, *Δ*IFD and *Δ*SC) as a within-subject factor (SPSS v. 22, IBM). Mauchly's test was conducted to test for sphericity and the Greenhouse-Geisser correction applied when the assumption of sphericity was violated.

III Results

A Contribution of spectral cues to vowel discriminability

Ferrets successfully learnt to discriminate artificial vowel sounds and to generalize across vowels differing in their fundamental frequencies. Fig. 3A illustrates the responses of a single ferret trained to discriminate [u] (left responses rewarded) and [ɛ] (right responses rewarded). Responses are shown to the training vowels and all mismatch probe sounds presented. In accordance with the training contingency, the ferret responded to [u] at the left spout and [ɛ] at the right spout on the majority (88.1 %) of trials. In the case of mismatch vowels, the animal responded more frequently in the direction of the training vowel with the same F2 value. For example, she responded as frequently at the right spout when presented with mismatch stimuli containing an F2 value of 2058 Hz as she did when presented with the training stimulus [ɛ] with an F2 value of 2058 Hz. This was regardless of whether the first formant was higher (1413 Hz) or lower (460 Hz) in frequency than the trained F1 (730 Hz).

Ferrets' responses to mismatch vowels suggest that the position of the second formant is an important determinant of behavior and that perceptual similarity may depend heavily on F2 position. To visualize this, discriminability matrices were generated to illustrate how differently each subject responded to vowels across the stimulus set. Figure 3B illustrates an example based on the responses of one ferret. Each comparison within the discriminability matrix represents the magnitude of the d′ (|d′|) calculated from the responses an animal made to a given pair of vowels. In this calculation |d′| will be high for vowels that the ferret perceives to be different (for example the trained vowels, where the animal responds at the left and right spouts for [u] and [ɛ] respectively) and low for vowels that the ferret perceives

to be similar (for example [ɛ] and the mismatch vowels that share F2 with the trained [ɛ]). For each comparison, discriminability was related to the difference in F1 and F2 frequencies of the compared vowels (Fig. 3Ci and 3Cii), as well as the difference in inter-formant distance (Fig. 3Ciii) and spectral centroid (Fig. 3Civ). For the ferret in Figure 3C, discriminability between vowel pairs was strongly positively correlated with differences in F2 positions between vowels ($R^2 = 0.953$), moderately correlated with differences in spectral centroid ($R^2 = 0.365$) but not differences in F1 positions or IFD ($R^2 < 0.1$). The same pattern was observed across ferrets (Fig. 3D): positive correlations were found between discriminability of vowels and differences in F2 frequency (6/8 ferrets, $R^2 > 0.4$), and also differences in spectral centroids (6/8 ferrets, $R^2 > 0.25$) but not differences in F1 frequency or inter-formant distance (8/8 ferrets, $R^2 < 0.1$).

To assess the contribution of each acoustic cue to ferrets' behavior, we used multiple logistic regression to model behavioral responses (left or right) on each trial using one or more acoustic cues as predictors. Figure 4 illustrates the deviance that resulted when models were generated with individual cues or combinations of cues as predictors. When considering models in which only a single cue is chosen as a predictor of behavioral response, models based on F2 fit better (measured qualitatively by reduction in deviance) than models based on any other cue and F2 models appear to fit as well as models considering multiple cues. To test this statistically, an analysis of deviance compared the deviance resulting from the F2 only model with deviance resulting from models incorporating additional cues (e.g. F2 and F1). A significant drop in deviance indicates that the model fit is improved by the addition of extra parameters. Since the underlying acoustic cues do not vary independently (e.g. F2 is always higher than F1, and SC is strongly correlated with F2), fitting a model with as few parameters as possible is desirable in order to minimize the possible influence of multi-collinearity. For the majority of ferrets (6/8), addition of further predictors did not significantly reduce deviance ($p > 0.05$). For one other ferret, use of spectral centroid alone as predictor provided a robust and parsimonious model of behavior. For the remaining ferret, any two-predictor model was found to significantly reduce deviance when compared to any single-predictor model ($p < 0.05$). However inclusion of a third predictor did not further improve model fit for this subject. For all subjects, models using only F1 or only IFD as predictor were significantly improved by inclusion of either F2 or spectral centroid ($p < 0.001$). In all cases, the model identified as being robust to inclusion of additional factors was found to fit significantly better than the constant model (no predictors, $p < 0.001$).

In summary, Experiment One revealed that when discriminating multi-formant vowels, the majority of ferrets base their discrimination on F2, with spectral centroid being important for a small number of ferrets. This finding appears to conflict with a previous experiment (Bizley et al., 2013b) in which ferrets tested with single formant vowels treated a single formant vowel with a peak at the F1 frequency of [u] in the same way as the complete multi-formant vowel. This difference may indicate that animals integrate spectral information in multi-formant vowels in a non-linear fashion. However, since both studies test different groups of ferrets it is possible that individuals show variability in their discrimination strategy. In order to explore this further we tested three ferrets that had participated in Experiment One with the single formant vowels used in Bizley et al., (2013b).

B Discrimination of single formant vowels

Experiment Two investigated whether the spectral cues underlying discrimination of multi-formant vowels also play a role in responding to single formant vowels (Fig. 5A). This experiment therefore repeated the single formant testing from Bizley et al. (2013b) but now with three animals for whom the weighting of acoustic cues was known from Experiment One. The performance of three subjects in the current study (ferrets 1-3) and two subjects from the previous study by Bizley et al. (ferrets 4 and 5) is illustrated in Fig. 5B. In agreement with earlier findings, the ferrets in the current study responded to single formant vowels with peaks at the first formant of [u] and the second formant of [ɛ] as they did for the original training vowels. Furthermore, subjects in both studies did not respond to single formant vowels with the second formant of [u] or the first formant of [ɛ] as if they were the trained multi-formant vowels, tending to either categorize randomly (F1 of [ɛ]), or systematically mis-classify the vowel (F2 of [u]). To summarise, subjects for whom responses to multi-formant vowels were strongly associated with F2 frequency were presented with single formant vowels at the locations of the F1 and F2 of the trained multi-formant vowels. However, single formant vowels centered at F2 did not always elicit accurate behavioral discrimination.

For subjects in the current study (ferrets 1-3), there were only weak correlations between discriminability and spectral centroid differences between multi-formant vowels in Experiment One (R^2 = 0.23, 0.27 and 0.37) suggesting that these subjects do not rely heavily on spectral centroids in their judgement of sounds. To determine whether responses to single formant vowels reflected decision-making based on spectral centroid, ferret's responses were compared with the spectral centroid of single formant vowels (Fig. 5C). In the earlier study by Bizley et al. (2013b) single formant vowels were presented at a number of fundamental frequencies and therefore with a wide range of spectral centroids. This data set provides a richer sampling of the relationship between centroid and behavior and so is also included in Fig. 5C. The data from both studies show that ferrets' responses to single formant vowels are non-linearly related to the spectral centroid of sounds: Single formants with centroid values below the spectral centroid of [u] (845 Hz at a fundamental frequency of 200 Hz) are classified as [u] whereas single formants with centroids above that of [ɛ] (1636 Hz at a fundamental frequency of 200 Hz) are classified as [ɛ]. In between these boundaries, prediction of an animal's response based on spectral centroid is difficult and linear correlation coefficients are small ($R^2 < 0.1$) when calculated between response and centroid for the animals studied by Bizley and colleagues.

C Discrimination of voiceless vowels

Experiment Three assessed the importance of harmonic structure for discrimination of the vowel sounds. Human listeners are able to estimate the identity of a vowel from the spectral envelope whether the vowel is whispered or spoken. Therefore, the aim of Experiment Three was to determine whether ferrets were able to extract the spectral envelope from voiceless vowels by presenting ferrets with probe sounds generated by switching the sound source from a click train to broadband noise. This preserved the spectral envelope of vowel sounds, including the position of formant peaks, but eliminated the harmonic structure that underlies vowel pitch (Fig. 6A).

Europe PMC Funders Author Manuscripts

When presented with voiceless vowels as probe stimuli, most (5/6) ferrets discriminated accurately and automatically without the need for reward contingency (i.e. responses to probe stimuli were rewarded on all trials [five ferrets] or never rewarded [one ferret]). For one ferret, whose responses to probe stimuli were heavily biased to the right spout, performance was initially near chance (53.4%). A reward contingency matching that of the trained voiced vowels was therefore introduced during presentation of voiceless vowels on probe trials. Introduction of the reward contingency improved performance (58.9%), indicating that while this ferret did not automatically generalize in the same manner as other ferrets, she was nonetheless capable, if poorly so, of discriminating voiceless vowels at a level significantly better than chance (binomial test; $n = 236$ trials, $p < 0.01$).

Across all subjects ($n = 6$), the median accuracy for the discrimination of voiceless vowels was 69.5% compared to a performance level of voiced vowels of 87.4% (Fig. 6B). For all ferrets, discrimination was poorer for voiceless than voiced vowels (Wilcoxon signed rank test: $W = 21$, $p = 0.03$). At present it is unclear whether this reflects a contribution of harmonic structure to task performance, or differences between voiced and voiceless vowels such as reward contingency and explicit training.

D Comparison of Human and Ferret spectral timbre discrimination

Experiment One demonstrated that ferrets are heavily reliant on F2 for vowel discrimination. In contrast, when identifying vowels humans are known to base their judgments on multiple acoustic cues including both F1 and F2 (see Introduction). The aim of Experiment Four was to determine what acoustic cues human listeners utilize when tested under similar task constraints to the ferrets tested in Experiment One. Subjects were either naïve ($n = 10$, subjects had not participated in a psychoacoustic study of vowels or listened to the ferrets' task and were naïve to the purpose of the experiment) or experienced listeners ($n = 4$, subjects regularly tested ferrets in this task and are authors).

Experiment Four was designed to replicate Experiment One as closely as possible. In contrast to most psychoacoustic studies therefore subjects were not instructed in how to complete the task beyond being told that they had two response options and would receive feedback on each trial as to whether they had selected the correct option. Like our ferret subjects they therefore had to learn by trial and error what the appropriate response contingency was. When discriminating the trained vowels and probe trials both naïve and experienced subjects showed similar behavior to each other, and to that observed in Experiment 1 in trained ferrets. Figure 7A illustrates the behavior of a naïve subject from whom discriminability between vowels was correlated strongly with differences in F2 frequency between vowels ($R^2 = 0.865$) and differences in spectral centroid ($R^2 = 0.728$) but not differences in F1 frequency or inter-formant distance ($R^2 < 0.1$). The distribution of correlation coefficients and regression parameters (Fig. 7B) across all humans resembled the distribution observed for ferrets (Fig. 3D) although the human data was more variable – possibly due to the variety of linguistic backgrounds of listeners or the lack of instructions that may have left listeners free to form different expectations about stimuli or use different discrimination strategies.

Participants' responses were modelled using multiple logistic regression as in Experiment One. Figure 8 illustrates the proportion of deviance accounted for by models in which individual cues or combinations of cues are taken as predictors of subjects' key press (left or right). For most subjects (11/13), models using a single acoustic cue as a predictor were robust to the addition of further predictors (i.e. addition of further predictors did not reduce deviance significantly). For five of these eleven subjects, identified models were based on F2 frequency whereas another five, models were based on spectral centroid. For the eleventh subject, a model based on either F2 or spectral centroid was robust to addition of extra parameters. For the twelfth subject, a model incorporating both F2 and spectral centroid significantly reduced deviance when compared to either the F2-only or spectral centroid only model ($p < 0.001$) and was robust to addition of a third parameter. For the final subject, any two-predictor model better fitted the data than any single predictor model from which it was composed. Each two-predictor model was robust to the addition of a third predictor. In all cases, the model identified for each subject as parsimonious and robust to inclusion of additional predictors was significantly better fitted to the data than the constant model ($p < 0.001$). In no subject was a model based only on F1 frequency or inter-formant distance robust to inclusion of additional factors.

The results of linear regression (Figures 3 and 7) illustrated that human behavior was more variable in this task than ferret behavior. Nevertheless, the overall trends in which F2 and spectral centroid were implicated most strongly in the determination of subjects' responses were consistently observed across species. In order to quantitatively compare across species, a repeated-measures ANOVA was performed on the regression coefficients obtained from linear regression models illustrated in Fig. 3D and 7B. Unsurprisingly there was a large effect of cue type on regression coefficient values (b, which determines discriminability on the y-axis of Figs 3D and 7B, $F_{1.96, 37.3} = 39.5$, $p < 0.001$) but there was no main effect of species or interaction between cue and species ($p > 0.05$). Similarly, a repeated measure ANOVA on correlation coefficients (R^2, x-axis in Figures 3D and 7B) revealed a significant main effect of cue ($F_{1.79, 34.0} = 34.5$, $p < 0.001$), but no main effect of species or interaction between cue and species ($p > 0.05$). Thus both the dependence of discriminability on cue separation and the strength of the association between the discriminability and magnitude of cue separation varied with cue type (F1, F2 etc.) but this variation did not differ significantly between species.

IV Discussion

The results presented in Experiment One demonstrate that timbre perception in ferrets is strongly associated with the position of the second formant and spectral centroid of multi-formant vowels. However single formant responses in Experiment Two illustrate that perception of multi-formant vowels may not always be mimicked by spectral peaks at F2 positions alone. Experiment Three showed that ferrets could generalize across voiced and voiceless (whispered) vowels, further emphasizing the ability of ferrets to extract information about the spectral envelope of vowels. Experiment Four illustrated that humans performing an analogous task as in Experiment One behave similarly to ferrets.

In Experiment One and Four, differences in F2 position and spectral centroid were shown to be key predictors of the discriminability of vowel pairs. The position of the second formant is strongly correlated with spectral centroid for the vowels presented (Fig. 2; $R^2 = 0.895$) and thus it is unsurprising that both are implicated in discriminability. From the current data it is unclear which, if either, of the two parameters is more important in ferrets' vowel discrimination. Results from Experiment Two suggest that responses to single formant vowels are related to their spectral centroid, but that there is not a simple linear relationship between the ferrets' classification of probe trials and spectral centroid position. Future experiments in which the position of the second formant and spectral centroid of vowels are dissociated may therefore be helpful. Additionally it may be interesting to investigate whether variation in spectral centroid correlates with discriminability of other sounds such as those produced by musical instruments.

The involvement of particular spectral cues such as F2 or spectral centroid in timbre perception is not unique to ferrets. Ohms and colleagues (2012) demonstrated that zebra finches trained in a go/no-go task will respond to vowels on the basis of second and third formant combinations rather than the position of the first formant. In contrast, Chimpanzees (Kojima and Kiritani, 1989) and Old World monkeys (Sinnott *et al.*, 1997) place equal or greater weight on the position of F1 than F2. All of these studies contrasted human and non-human performance in matched tasks (i.e. using the same stimuli and methods of analysis) and found a greater reliance of human listener's on F2 than F1. The performance of humans in Experiment Four presented here is thus consistent with previous reports from comparative psychoacoustics. The importance of F2 and spectral centroid in our findings also resemble those made in many studies focussing on human vowel identification (Delattre *et al.*, 1952; Singh and Woods, 1971; Rakerd, 1984) and timbre perception (Pols *et al.*, 1969; McAdams *et al.*, 1995; Caclin *et al.*, 2005). However, when relating the current findings to the established literature on the perception of vowels by humans, caution must be exercised. The discrimination task used here may not require the use of cues that may be important in more complex behaviors such as phonetic identity. It may therefore be that F1 frequencies, inter-formant distances and other cues not considered here such as formant bandwidth (de Cheveigné, 1999) or spectral shape (Zahorian and Jagharghi, 1993; Ito *et al.*, 2001) may play greater roles in everyday human vowel perception.

In comparing the results of ferrets (Experiment One) and humans (Experiment Four) it is notable that human behavior was more variable across subjects than that of ferrets'. This variability in human performance likely stems from the additional experience that listeners (who were drawn from a range of linguistic backgrounds) had in discriminating speech sounds, coupled with the minimization of task instructions that were required to match the task to that performed by ferrets. Although the same pattern of results was observed when considering all humans or only those who were native English speakers, it is possible that without specific instruction, subjects employed different discrimination strategies or had different expectations about the stimuli. The variability across listeners observed in Experiment Four may be minimised by providing subjects with greater instruction – although this would have implications for our ability to perform comparisons between species.

In this study, ferrets and humans appear to use spectral cues in a similar manner. Yet such similarities between human and non-human subjects are not always observed. One example from is the demonstration that human and non-human primates appear to place different weight on F1 and F2 in a task that required subjects to detect a change in vowel sound (Kojima and Kiritani, 1989; Sinnott *et al.*, 1997). What determines the likelihood of two species using the same strategy? One explanation might be the pattern of auditory sensitivity of each species. Humans, ferrets and zebra finches have relatively similar audiograms in which thresholds for tones in the typical F2 frequency range (1-3 kHz) are lower than for tones in the F1 frequency range (0.3 – 1 kHz) (Kelly *et al.*, 1986; Okanoya, 1987; ISO:226, 2003) whereas other primates have equal or lower sensitivity within the F2 than F1 frequency range (Behar et al., 1965; Kojima, 1990; Coleman, 2009). Such explanations must be treated tentatively as it is unclear how accurately frequency thresholds measured using individual tones reflect representations of supra-threshold, wide-band sounds – particular in non-human species in which audiogram measurements are highly sensitive to methodology (Coleman, 2009). Additionally such an explanation relies on comparison of studies using not only different species, but also different stimuli with different formant frequencies and methods of vowel synthesis. Therefore in future it may be beneficial to compare formant weighting across species using standardized stimuli in order to better understand the mechanisms underlying phylogenetic differences.

Experiment Two measured the behavioral discrimination of single formant vowels by ferrets, and confirmed a previously demonstrated interaction between the position of a single formant and ferrets' responses (Bizley *et al.*, 2013b). Specifically, that subjects responded to the first but not second formant of [u] and the second but not first formant of [ɛ] as if they were the original training vowels. This outcome apparently contrasts with the findings from Experiment One in which the same subjects were found to discriminate multi-formant vowels on the basis of second but not first formant positions. This inconsistency might arise due to formant interactions in which the presence of a second formant affects perception of the first formant. A simple example is provided by ferrets' responses to the mismatch vowel combining the first formant of [u] (460 Hz) and the second formant of [ɛ] (2058 Hz). When these formants are presented individually as single formant sounds, ferrets respond most often at the spouts associated with [u] and [ɛ] respectively (e.g. left and right). However when these cues are placed in conflict in the mismatch vowel (460, 2058 Hz; Fig. 2), F2 position dominates behavior as animals consistently respond at the spout associated with [ɛ] (right) rather than equally at both spouts. Thus animals do not behave as one might predict if responses were determined by a linear combination of F1 and F2. In future it will be important to determine whether such formant interactions arise at the level of the acoustic properties of the stimulus (e.g. through alterations in existing acoustic cues such as formant bandwidth that are introduced when formants are added or introduction of additional cues such as relative formant amplitude) or within the auditory system itself (discussed further below).

Experiment Three demonstrated that ferrets were able to access the underlying spectral envelope of vowel sounds because voiceless vowels lacking a harmonic structure were successfully discriminated. However, harmonic structure may contribute to vowel discrimination as performance of all subjects was poorer for voiceless than voiced vowels.

Although contributions of harmonic structure may not be surprising given previous work (Darwin and Gardner, 1986; Chalikia and Bregman, 1989; de Cheveigné and Kawahara, 1999) additional considerations must be made when drawing such conclusions from the current study: Performance deficits may be an equally likely result of presentation of voiceless vowels as probe sounds (20% of trials) without prior training or reward contingency. To establish the importance of harmonic structure in vowel perception more thoroughly, it would be necessary to train ferrets to discriminate voiceless vowels from the beginning of the study and employ voiced vowels as probe sounds after training is complete. A demonstration that behavioral discrimination of voiceless vowels remains poorer than voiced vowels in voiceless trained animals would provide a more robust demonstration of the contribution of harmonic structure to vowel discrimination by ferrets.

Ferrets' ability to generalize their discrimination behavior across sound sources provides another example of perceptual invariance in this species. In the current study and previous work (Bizley *et al.*, 2013b), ferrets also generalize across fundamental frequency (pitch) and sound level (loudness) of vowels and can also discriminate vowels in background noise. Similarly, other animals can recognize vowels across individual speakers of the same (chimpanzee; Kojima and Kiritani, 1989) and different genders (zebra finch; Ohms *et al.*, 2010), and across vocal tract length (gerbil; Schebesch *et al.*, 2010). Thus like humans, non-human subjects can discriminate vowels across a number of task-irrelevant perceptual dimensions.

A key question stemming from this research is how acoustic cues such as formant frequencies or spectral centroid are extracted by the auditory system and how neural computations underpin vowel discrimination, formant interactions and generalization of vowel timbre across other perceptual dimensions, such as F0 and voicing. The extraction of acoustic cues begins in the auditory nerve, where the bandwidths of auditory nerve fibers (ANFs) are sufficiently narrow (Sumner and Palmer, 2012) that both formant peaks are likely to be resolved. However fiber thresholds are lower within the F2 than F1 frequency range for vowels used in the current study. Therefore fibers representing the second formant may have higher firing rates during sound presentation than those representing the first formant. This could in turn encourage an F2-based neural representation further on in the auditory system in regions such as the inferior colliculus or auditory cortex. While many studies have demonstrated encoding of vowel timbre in the central auditory system (Ohl and Scheich, 1997; Versnel and Shamma, 1998; Bizley *et al.*, 2009; Walker *et al.*, 2011), few have addressed neural sensitivity to acoustic components of vowels such as individual formants. Ohl and Scheich (1997) described cortical neurons in the gerbil that were sensitive to changes in the F2 frequency of vowels with a constant F1 frequency. This suggests that either F2 frequency or inter-formant distance are important features in the responses of cortical neurons. However additional tests in which F1 and F2 are both varied are necessary to fully establish how such acoustic cues are represented at the cellular level in the cortex.

Experiment Two demonstrated that interactions between formants can influence vowel perception and it seems likely that central mechanisms may also contribute to this behavior. Neurons in field L of the Mynah bird (Langner *et al.*, 1981) and in auditory cortex of the gerbil (Ohl and Scheich, 1997) are sensitive to the number of formants in vowel sounds and

Europe PMC Funders Author Manuscripts

their relative frequencies. Langner *et al.* (1981) suggested that this neural behavior may arise, at least in part, as a result of pure tone frequency tuning of units and this is supported by additional evidence demonstrating the involvement of spectral tuning in influencing neural response to vowels (Versnel and Shamma, 1998). The patterns of synaptic connectivity that determine frequency tuning may thus contribute not only to a neuron's response to formant cues but also provide sensitivity to formant-interactions or indeed underlie the non-linear integration of formant representations. How neural responses in the auditory system then go on to influence behavior and timbre perception remains to be determined.

The cues necessary for the generation of voicing-invariant neural representations appear to be represented in the auditory nerve: ANFs are capable of representing spectral features of both voiced and voiceless vowels in the anaesthetized cat (Voigt *et al.*, 1982) and chinchilla (Stevens and Wickesberg, 2005). Similarly neurons in primary auditory cortex of anaesthetized ferrets respond similarly to voiced and unvoiced vowels (Versnel and Shamma, 1998). It may therefore be that physiological representations of vowels are, to a degree, invariant to voicing from the periphery. In future it will be important to determine whether this invariance to voicing also exists when animals are actively engaged in vowel discrimination.

In conclusion, ferrets and humans use similar cues to discriminate artificial vowel sounds, placing greater emphasis on the frequency of the second formant. In addition to accurately discriminating vowels across fundamental frequency and sound level, ferrets were also able to generalize vowel timbre across harmonic and whispered stimuli. Together with earlier work, this study provides a growing framework within which to study the neuronal basis of vowel discrimination and timbre perception.

Acknowledgements

This work was supported by a grant from the Biotechnology and Biological Sciences Research Council (grant BB/H016813/1), a Royal Society Dorothy Hodgkin Fellowship and a Royal Society / Wellcome Trust Sir Henry Dale Fellowship to J.K.B. (WT098418MA).

Appendix

The spectral properties of vowels presented to ferrets were confirmed by measuring the spectra of stimuli recorded from a microphone located in the experimental testing chamber at the animal's head position if it were performing the discrimination task. Figure 9 illustrates a close correspondence between recorded and schematic spectra for sounds presented in Experiments One, Two and Three.

References

Behar I, Cronholm JN, Loeb M. Auditory sensitivity of the rhesus monkey. J Comp Physiol Psychol. 1965; 59:426–428. [PubMed: 14313786]

Bizley JK, Walker KM, Nodal FR, King AJ, Schnupp JW. Auditory cortex represents both pitch judgements and the corresponding acoustic cues. Curr Biol. 2013a; 23:620–625. [PubMed: 23523247]

Bizley JK, Walker KM, Silvermann BW, King AJ, Schnupp JW. Interdependent encoding of pitch, timbre and spatial location in auditory cortex. J Neurosci. 2009; 29:2064–2075. [PubMed: 19228960]

Bizley JK, Walker KMM, King AJ, Schnupp JWH. Spectral timbre perception in ferrets: Discrimination of artificial vowels under different listening conditions. J Acoust Soc Am. 2013b; 133:365–376. [PubMed: 23297909]

Caclin A, McAdams S, Smith BK, Winsberg S. Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones. J Acoust Soc Am. 2005; 118:471–482. [PubMed: 16119366]

Carlson R, Granstrom B, Fant G. Some studies concerning perception of isolated vowels. STL-QPRS. 1970; 2:19–35.

Chalikia MH, Bregman AS. The perceptual segregation of simultaneous auditory signals: Pulse train segregation and vowel segregation. Percept Psychophys. 1989; 46:487–496. [PubMed: 2813035]

Charlton BD, Zhihe Z, Snyder RJ. The information content of giant panda, Ailuropoda melanoleuca, bleats: acoustic cues to sex, age and size. Anim Behav. 2009; 78:893–898.

Chistovich LA, Lublinskaya VV. The ‘center of gravity’ effect in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli. Hear Res. 1979; 1:185–195.

Coleman MN. What do primates hear? A meta-analysis of all known nonhuman primate behavioral audiograms. Int J Primatol. 2009; 30:55–91.

Darwin CJ. Listening to speech in the presence of other sounds. Phil Trans R Soc B. 2008; 363:1011–1021. [PubMed: 17827106]

Darwin CJ, Gardner RB. Mistuning a harmonic of a vowel: Grouping and phase effects on vowel quality. J Acoust Soc Am. 1986; 79:838–845. [PubMed: 3958326]

de Cheveigné A. Formant bandwidth affects the identification of competing vowels. ICPhS. 1999; 2093

de Cheveigné A, Kawahara H. Missing-data model of vowel identification. J Acoust Soc Am. 1999; 105:3497–3508. [PubMed: 10380672]

Delattre P, Liberman AM, Cooper FS, Gerstman LJ. An experimental study of the acoustic determinants of vowel color; observations on one- and two-formant vowels synthesized from spectrographic patterns. Word. 1952; 8:195–210.

Fant G, Risberg A. Auditory matching of vowels with two formant synthetic. STL-QPSR. 1963; 4:7–11.

Fitch WT, Fritz JB. Rhesus macaques spontaneously perceive formants in conspecifics vocalizations. J Acoust Soc Am. 2006; 120:2132–2141. [PubMed: 17069311]

Fitch WT, Kelley JP. Perception of vocal tract resonances by Whooping Cranes Grus americana. Ethology. 2000; 106:559–574.

Fritz J, Shamma S, Elhilali M, Klein D. Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. Nat Neurosci. 2003; 6:1216–1223. [PubMed: 14583754]

Green, DM, Swets, JA. Signal detection theory and psychophysics. Wiley; New York: 1966. 1–505.

ISO: 226. Normal equal-loudness level contours. International Organization for Standardization; Geneva: 2003. 1–18.

Ito M, Tsuchida J, Yano M. On the effectiveness of whole spectral shape for vowel perception. J Acoust Soc Am. 2001; 110:1141–1149. [PubMed: 11519581]

Kelly JB, Kavanagh GL, Dalton JC. Hearing in the ferret (Mustela putorius): thresholds for pure tone detection. Hear Res. 1986; 24:269–275. [PubMed: 3793642]

Kiefte, M, Nearey, TM, Assmann, PF. Vowel perception in normal speakersHandbook of Vowels and Vowel Disorders. Ball, MJ, Gibbon, FE, editors. Psychology Press; New York: 2013. 160–185.

Klatt, DH. Prediction of perceived phonetic distance from critical-band spectra: A first step. Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP ' 82; New York: IEEE; 1982. 1278–1281.

Kojima S. Comparison of auditory functions in the chimpanzee and human. Folia Primatol (Basel). 1990; 55:62–72. [PubMed: 2227723]

Kojima S, Kiritani S. Vocal-Auditory Functions in the Chimpanzee: Vowel Perception. Int J Primatol. 1989; 10:199–213.

Langner G, Bonke D, Scheich H. Neural discrimination of natural and synthetic vowels in Field L of trained Mynah Birds. Exp Brain Res. 1981; 43:11–24. [PubMed: 6265257]

Lartillot, O; Toiviainen, P. A matlab toolbox for musical feature extraction from audio. 10th International Conference on Digital Audio Effects; Bordeaux: University of Bordeaux; 2007. 1–8.

Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M. Perception of the speech code. Psychol Rev. 1967; 74:431–461. [PubMed: 4170865]

McAdams S. Perspectives on the contribution of timbre to musical structure. Comput Music J. 1999; 23:85–102.

McAdams S, Roussarie V, Chaigne A, Giordano BL. The psychomechanics of simulated sound sources: Material properties of impacted thin plates. J Acoust Soc Am. 2010; 128:1401–1413. [PubMed: 20815474]

McAdams S, Winsberg S, Donnadieu S, De Soete G, Krimphoff J. Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. Psychol Res. 1995; 58:177–192. [PubMed: 8570786]

Miller JD. Auditory-perceptual interpretation of the vowel. J Acoust Soc Am. 1989; 85:2114–2134. [PubMed: 2659639]

Molis MR. Evaluating models of vowel perception. J Acoust Soc Am. 2005; 118:1062–1071. [PubMed: 16158661]

Moore BC. Basic auditory processes involved in the analysis of speech sounds. Phil Trans R Soc B. 2008; 363:947–963. [PubMed: 17827102]

Ohl FW, Scheich H. Orderly cortical representation of vowels based on formant interaction. Proc Natl Acad Sci U S A. 1997; 94:9440–9444. [PubMed: 9256501]

Ohms VR, Escudero P, Lammers K, ten Cate C. Zebra finches and Dutch adults exhibit the same cue weighting. Anim Cogn. 2012; 15:155–161. [PubMed: 21761144]

Ohms VR, Gill A, Van Heijningen CA, Beckers GJ, ten Cate C. Zebra finches exhibit speaker-independent phonetic perception of human speech. Proc R Soc B. 2010; 277:1003–1009.

Okanoya K, D RJ. Hearing in Passerine and Psittacine birds: A comparative study of absolute and masked auditory thresholds. J Comp Psychol. 1987; 101:7–15. [PubMed: 3568610]

Peterson GE, Barney HL. Control methods used in a study of vowels. J Acoust Soc Am. 1952; 24:175–184.

Plomp R, Pols LCW, van der Geer JP. Dimensional analysis of vowel spectra. J Acoust Soc Am. 1967; 41:707–712.

Pols LCW, van der Kamp LJT, Plomp R. Perceptual and physical space of vowel sounds. J Acoust Soc Am. 1969; 46:458–467. [PubMed: 5804118]

Potter RK, Peterson GE. The Representation of Vowels and their Movements. J Acoust Soc Am. 1948; 20:528–535.

Rakerd B. Vowels in consonantal context are perceived more linguistically than are isolated vowels: Evidence from an individual differences scaling study. Percept Psychophys. 1984; 35:123–136. [PubMed: 6718209]

Reby D, McComb K, Cargnelutti B, Darwin C, Fitch WT, Clutton-Brock T. Red deer stags use formants as assessment cues during intrasexual agonistic interactions. Proc R Soc B. 2005; 272:941–947.

Schebesch G, Lingner A, Firzlaff U, Wiegrebe L, Grothe B. Perception and neural representation of size-variant human vowels in the Mongolian gerbil (Meriones unguiculatus). Hear Res. 2010; 261:1–8. [PubMed: 20004713]

Singh S, Woods DR. Perceptual structure of 12 American English vowels. J Acoust Soc Am. 1971; 49:1861–1866. [PubMed: 5125734]

Sinnott JM, Brown CH, Malik WT, Kressley RA. A multidimensional scaling analysis of vowel discrimination in humans and monkeys. Percept Psychophys. 1997; 59:1214–1224. [PubMed: 9401456]

Smith, JO. [Last accessed: 8 March 2015] Introduction to digital filters with audio applications. 2007. https://ccrma.stanford.edu/~jos/filters/

Stevens HE, Wickesberg RE. Auditory nerve representations of naturally-produced vowels with variable acoustics. Hear Res. 2005; 205:21–34. [PubMed: 15953512]

Sumner CJ, Palmer AR. Auditory nerve fibre responses in the ferret. Eur J Neurosci. 2012; 36:2428–2439. [PubMed: 22694786]

Swanepoel R, Oosthuizen DJJ, Hanekom JJ. The relative importance of spectral cues for vowel recognition in severe noise. J Acoust Soc Am. 2012; 132:2652–2662. [PubMed: 23039458]

Town SM, Bizley JK. Neural and behavioral investigations into timbre perception. Front Syst Neurosci. 2013; 7:88. [PubMed: 24312021]

Versnel H, Shamma SA. Spectral-ripple representation of steady-state vowels in primary auditory cortex. J Acoust Soc Am. 1998; 103:2502–2514. [PubMed: 9604344]

Voigt HF, Sachs MB, Young ED. Representation of whispered vowels in discharge patterns of auditory-nerve fibers. Hear Res. 1982; 8:49–58. [PubMed: 7142032]

Walker KMM, Bizley JK, King AJ, Schnupp JWH. Multiplexed and robust representations of sound features in auditory cortex. J Neurosci. 2011; 31:14565–14576. [PubMed: 21994373]

Xu Q, Jacewicz E, Feth LL, Krishnamurthy A. Bandwidth of spectral resolution for two-formant synthetic vowels and two-tone complex signals. J Acoust Soc Am. 2004; 115:1653–1664. [PubMed: 15101644]

Zahorian SA, Jagharghi AJ. Spectral-shape features versus formants as acoustic correlates for vowels. J Acoust Soc Am. 1993; 94:1966–1982. [PubMed: 8227741]

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

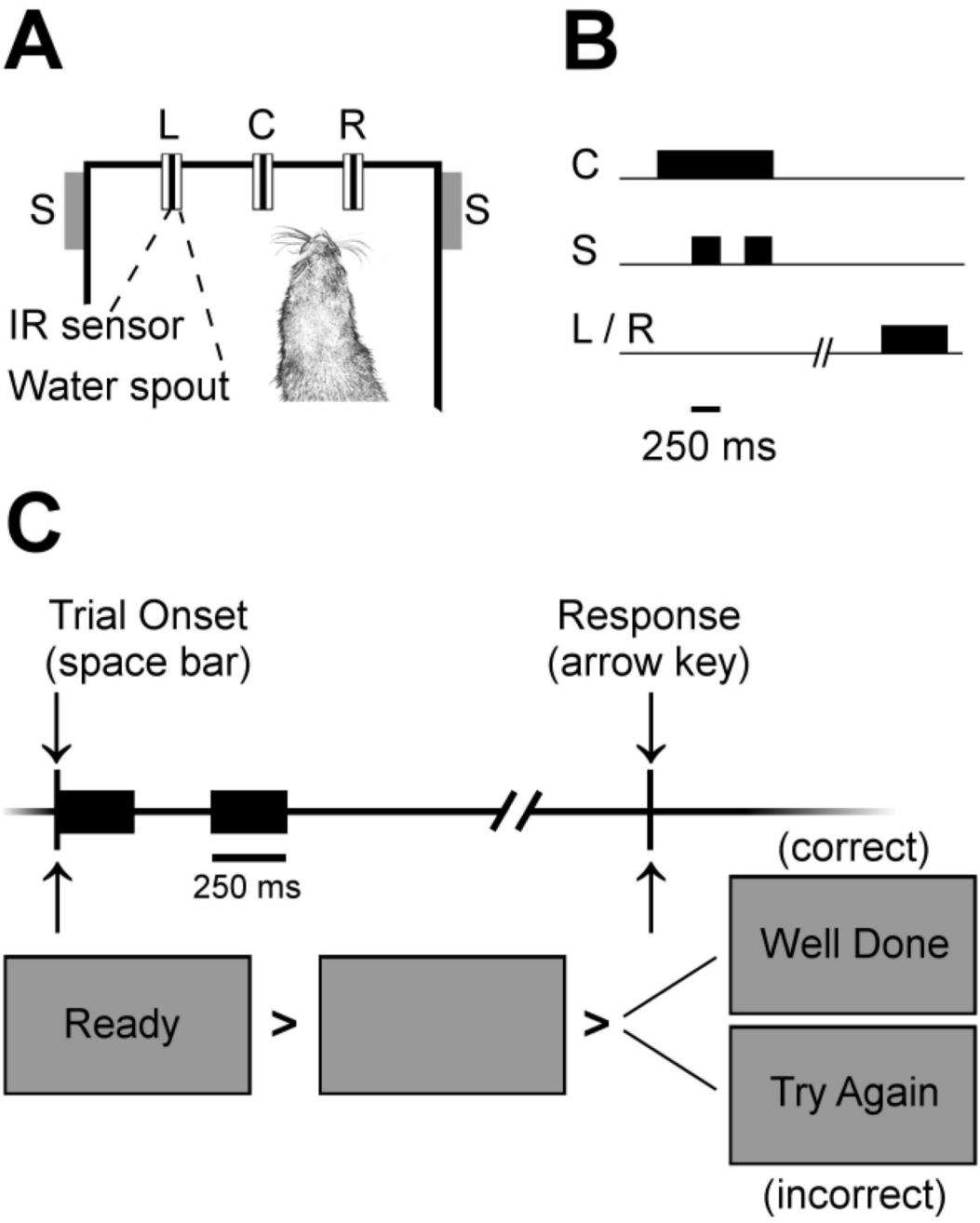

Fig. 1.

(**A**) Behavioral training apparatus in Experiments 1-3 in which ferrets were presented with two vowel tokens from speakers (S) to the right and left of a center spout (C). A subsequent response at a left or right spout (L or R) would be rewarded / punished depending on sound identity. (**B**) Trial structure showing the timing of the acoustic stimulus (S) and responses measured at the infer-red lick detectors (C, L/R). (**C**) Task design for human study in Experiment Four. Subjects initiated trials by pressing the space bar and were presented with

two vowel tokens. Responses were made using the arrow keys and feedback was given on-screen.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

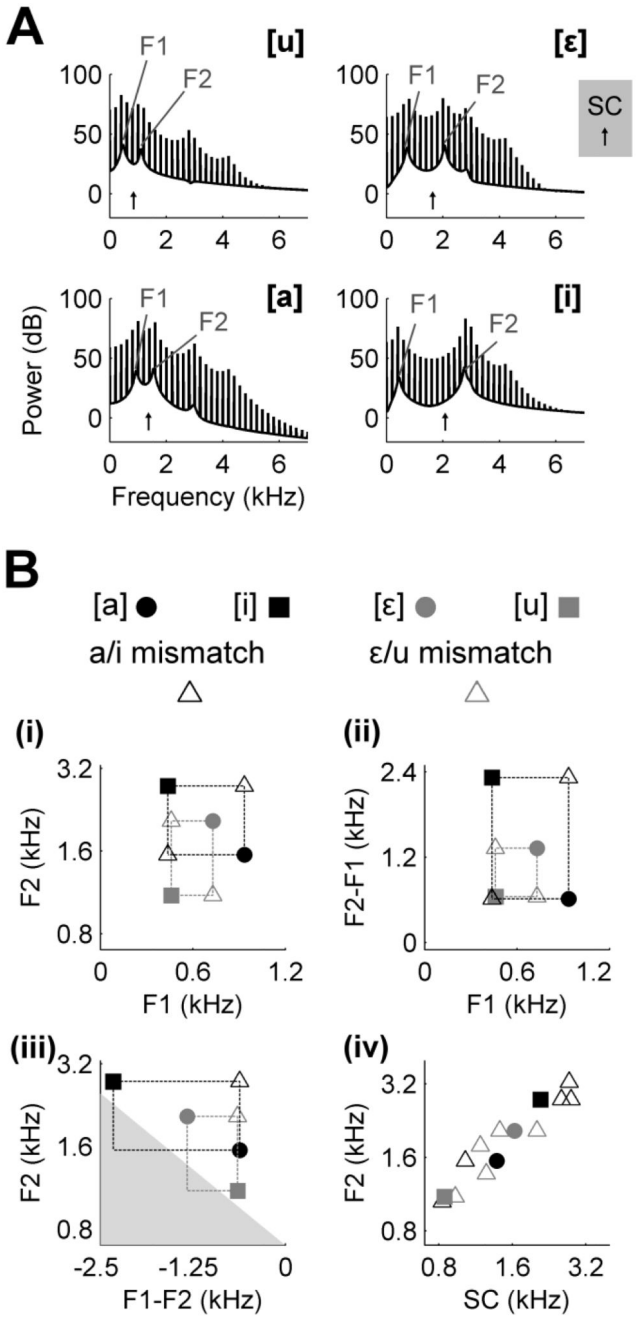

Fig. 2.

Stimulus spectra and parameter distribution (**A**) Spectra of artificial vowels with which ferrets were trained (F0 = 200 Hz) highlighting the first (F1) and second (F2) formant frequencies. Upward arrows indicate spectral centroid (SC). (**B**) The positions of training vowels within parameter spaces defined by spectral cues. Filled symbols indicate the positions of training vowels for ferrets trained with [a] and [i] (black) and [ɛ] and [u] (grey). Mismatch stimuli (open triangles) were synthesized by selecting positions in F1 vs. F2 (i), F1 vs. F2-F1 (ii) and F2 vs. F1-F2 (iii) parameter spaces that created a 2 x 2 experimental

design in each plane. Filled grey region in (iii) indicates region of parameter space in which F1 < 0 Hz and therefore stimuli could not be generated. (iv) Comparison of spectral centroid vs. F2 for all stimuli.

Europe PMC Funders Author Manuscripts

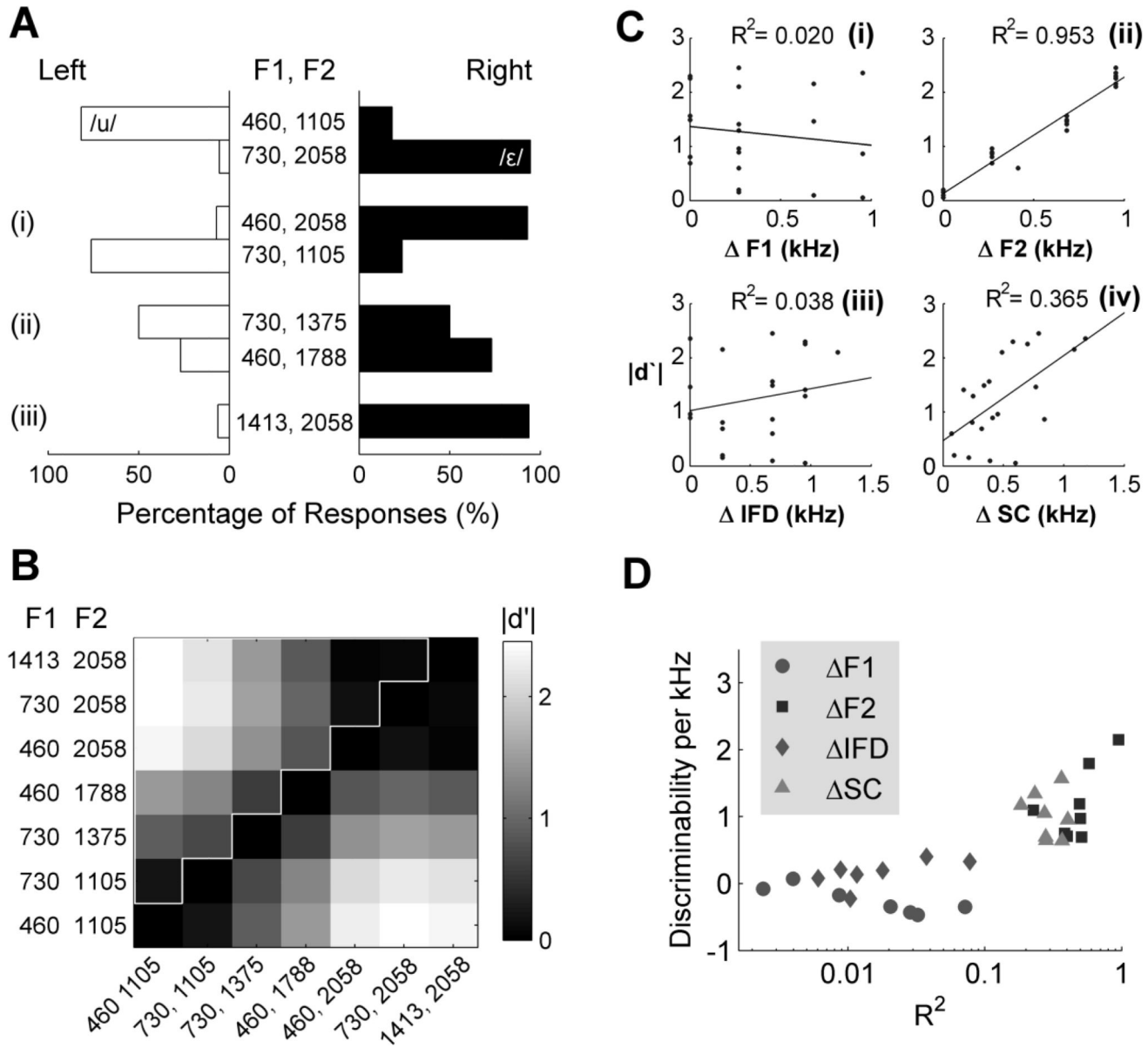

Fig. 3.

(**A**) Behavior of a ferret trained to respond to [u] and [ɛ] at the left and right spouts respectively. Top rows indicate the responses to training vowels ([u]: 82 trials; [ɛ]: 92 trials). (i) Responses to probe sounds in which F1 and F2 were mismatched (460, 2058 Hz: 43 trials; 730, 1105 Hz: 42 trials). (ii) Responses to probe sounds in which F1 and inter-formant distance were mismatched (730, 1375 Hz: 36 trials; 460, 1788 Hz: 41 trials). (iii) Responses to probe sounds in which F2 and inter-formant distance were mismatched (1413, 2058 Hz: 48 trials). (**B**) Discriminability matrix indicating how differently the ferret responded to all vowels presented. Each cell indicates comparison for one pair of vowels. Border indicates the set of unique comparisons used to correlate discriminability with spectral cues (see Figs 4 and 5). (**C**) Correlations between absolute discriminability (|d′|) of vowel pairs and

differences between vowels in F1 (*Δ*F1)(i), F2 (*Δ*F2)(ii), inter-formant distance (*Δ*IFD)(iii) and spectral centroid (*Δ*SC)(iv). Each point represents one of the unique comparisons (cells above the diagonal) in Fig. 3B. (**D**) Population scatterplots illustrating the correlation coefficients (X-axis) and regression gradients (Y-axis) for all ferrets.

Europe PMC Funders Author Manuscripts

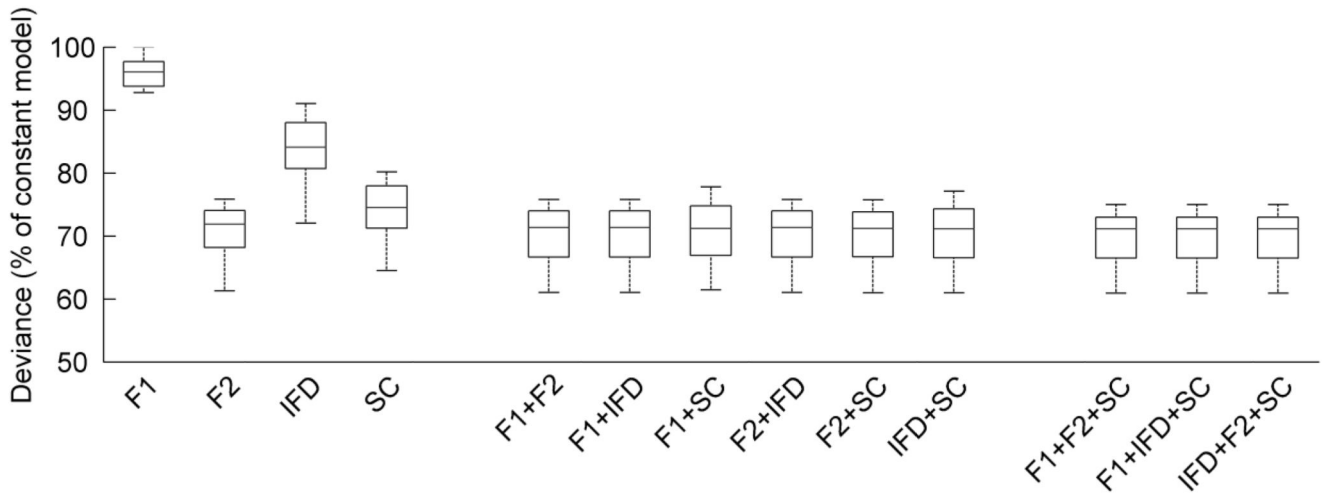

Fig. 4.

Boxplots indicating median deviance (and interquartile range) of models based on individual acoustic cues and combinations of cues (n = 8 ferrets). Deviance values are normalized relative to the constant model fitted without predictors. Lower deviance indicates better model fit.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

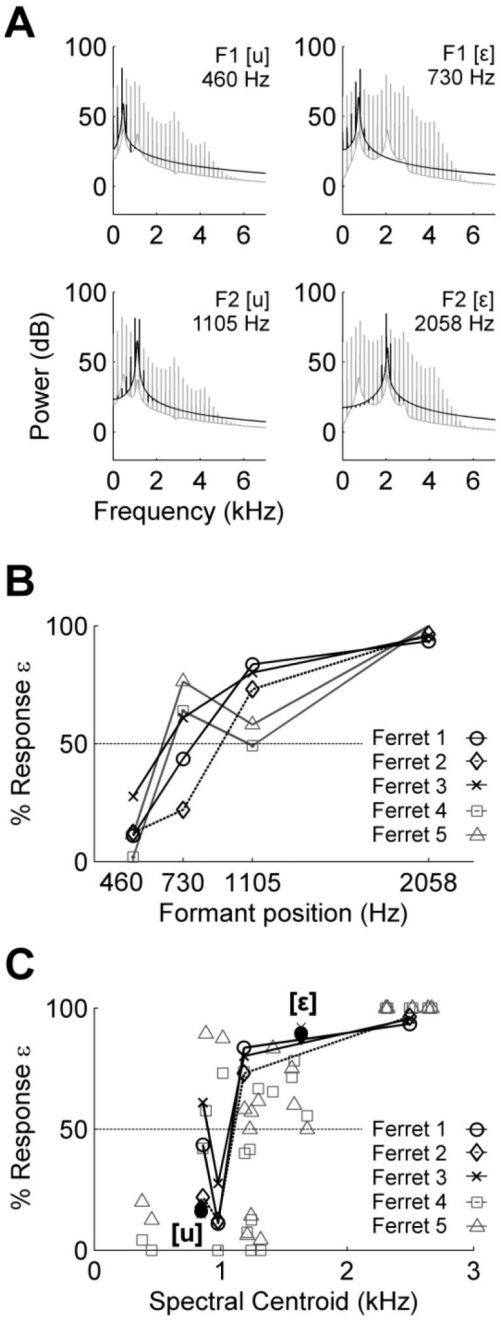

Fig. 5.

Discrimination of single formant vowels. (**A**) Spectra of single formant vowels (black) superimposed on the original training vowels (grey). (**B**) Performance of ferrets plotted against frequency position of formant peak when discriminating single formant vowels. Ferret 1, 2 and 3 are taken from the current study and were tested at an F0 of 200 Hz. Data for ferret 4 and 5 taken from Bizley et al. (2013b) and are the mean performance across all fundamental frequencies tested. (**C**) Performance of subjects in the current study plotted against spectral centroid of single formant vowels (unfilled) and the original training vowels

(filled). For ferrets in the study by Bizley et al. (2013b), single formant vowels were presented at a range of fundamental frequencies (n≥ 5) with each having its own spectral centroid.

Europe PMC Funders Author Manuscripts

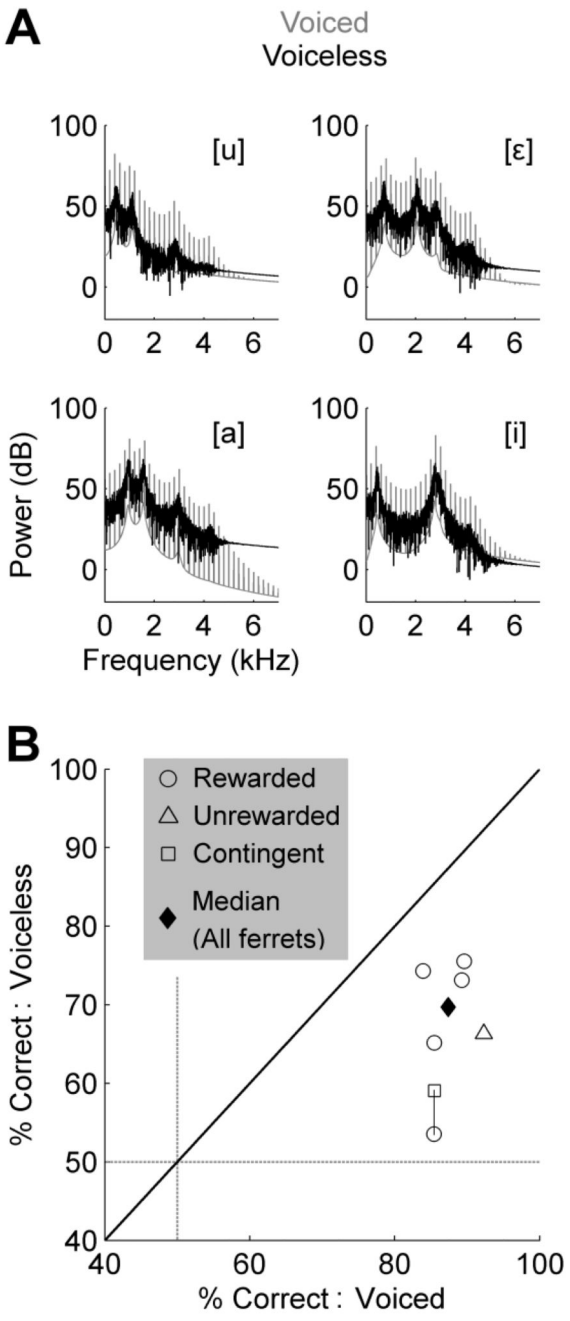

Fig. 6.

Vowel discrimination across sound sources. (**A**) Spectra of training vowels (voiced) and probe sounds (voiceless) illustrating the loss of harmonic structure (pitch) but retention of spectral envelope. (**B**) Comparison of performance (n = 6 ferrets) when discriminating voiced and unvoiced vowels. Voiced vowels were presented with an F0 of 200 Hz. Symbol legend indicates whether probe trials were all rewarded, never rewarded or rewarded contingent upon the vowel presented. Line from circle to square indicates increase in performance after adding reward contingency for one subject.

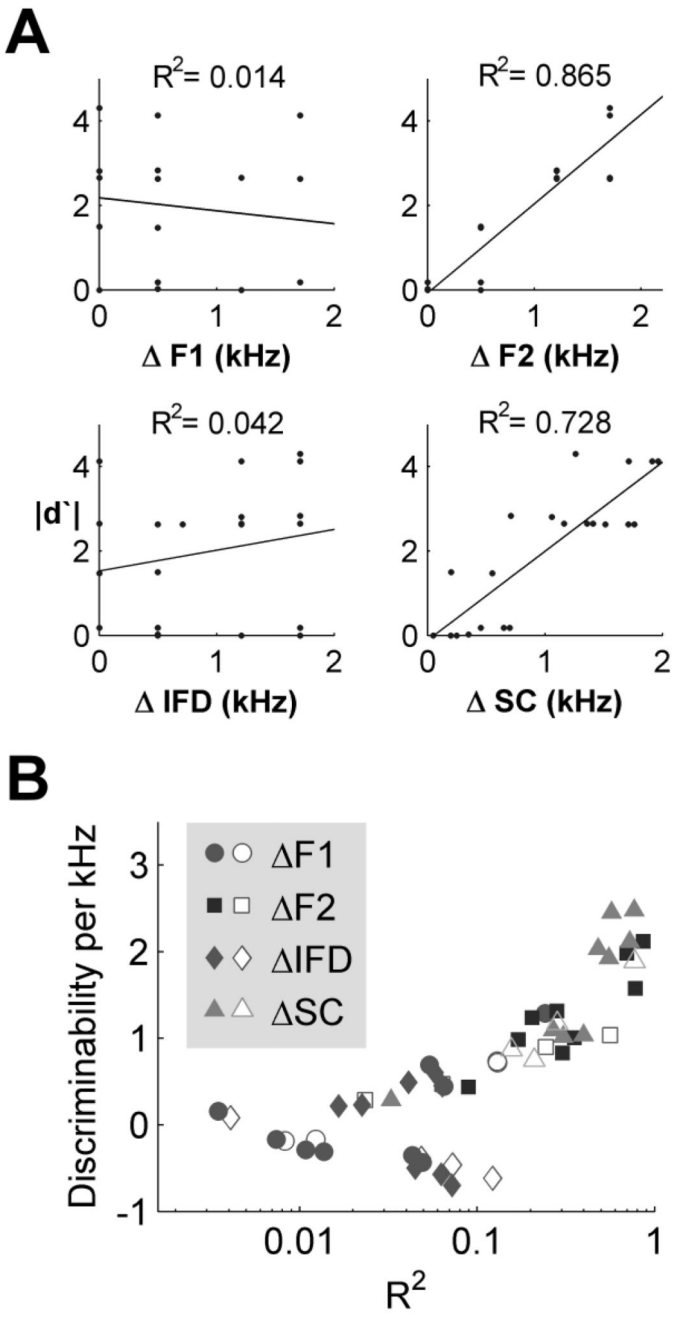

Fig. 7.

(**A**) Correlations between discriminability (|d′|) of vowel pairs and differences between vowels in F1 (*Δ*F1), F2 (*Δ*F2), inter-formant distance (*Δ*IFD) and spectral centroid (*Δ*SC) for one human listener trained to discriminate [a] and [i]. Correlation coefficients indicated by R^2 values. (**B**) Distribution of correlation coefficients and regression gradients for all participants (n = 13). Naïve and experienced subjects denoted by filled and unfilled markers respectively.

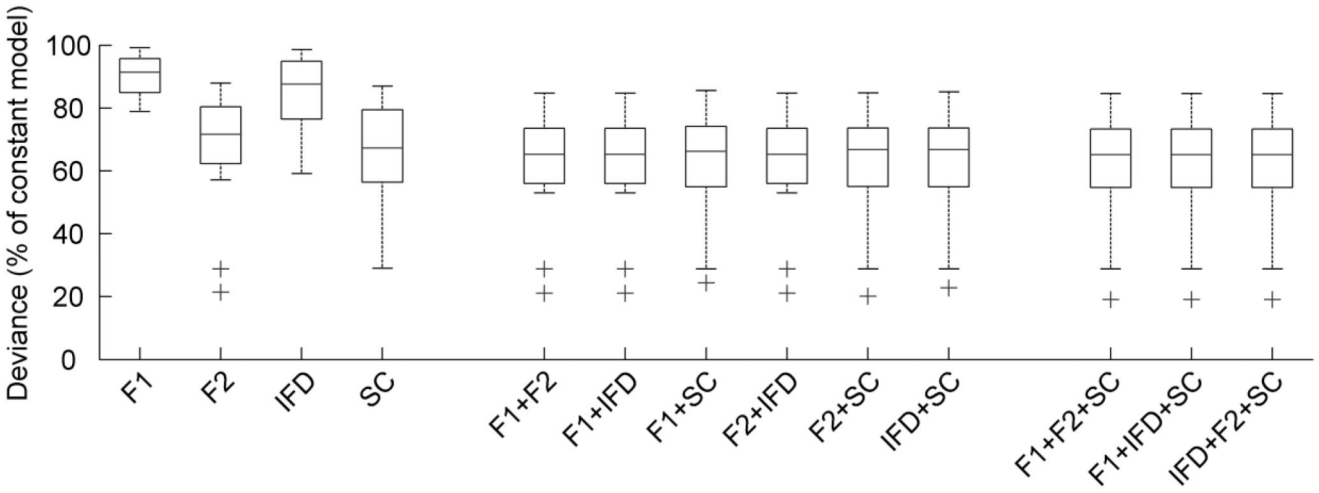

Fig. 8.

Boxplots indicating median (and inter-quartile range) deviance (n = 13 humans) of models based on individual acoustic cues and combinations of cues. Deviance values are normalized relative to the constant model fitted without predictors. Lower deviance indicates better model fit. Marker indicates outlier (defined as points beyond the 75th percentile ± 1.5 × interquartile range).

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

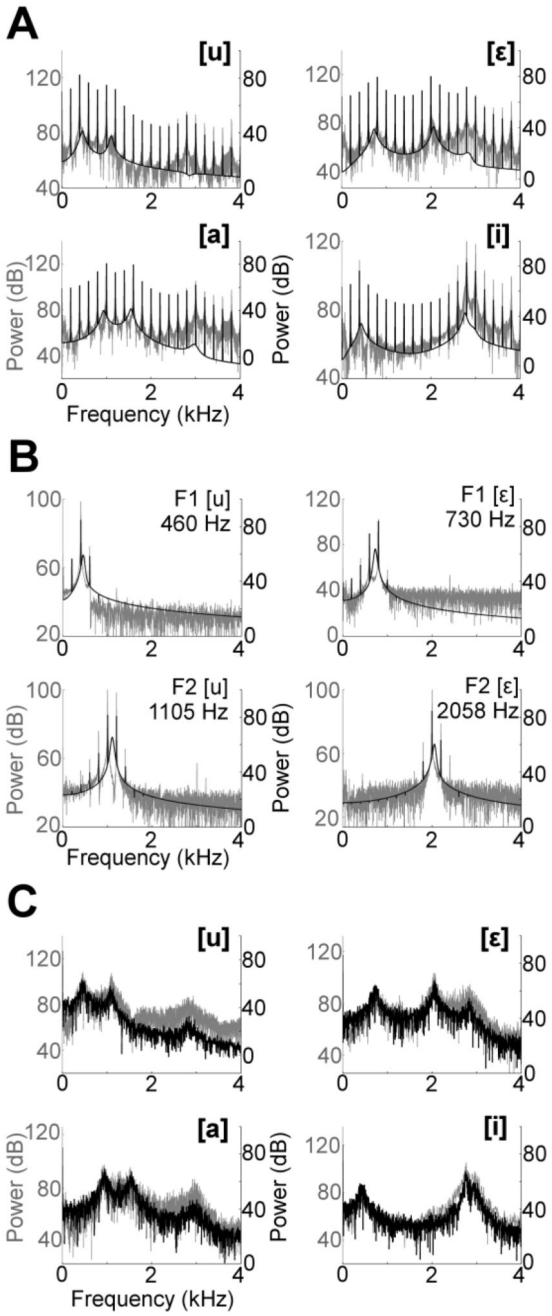

Fig. 9.

Comparison of schematic (black) and actual (grey) stimulus spectra recorded from ferret testing chamber. (**A**) Trained vowels with multiple formants presented in Experiment One. (**B**) Single formant vowels presented in Experiment Two. (**C**) Voiceless vowels presented in Experiment Three.