Perceptual Interplay of Pitch and Formant Contours in Melody Recognition by Zebra Finch

Chapter 5

Zhi-Yuan Ning · Henkjan Honing · Carel ten Cate

This chapter is prepared for submission

Chapter 5

ABSTRACT

Auditory perception of complex acoustic sequences involves the integration of multiple perceptual attributes, such as pitch and formant contours. While both attributes contribute to speech and music perception, the relative importance of each and their potential interactions remain underexplored. Here, we investigate how zebra finches (*Taeniopygia guttata*) discriminate harmonic complex tone sequences, which were characterized by pitch and formant contours that were either both increasing or decreasing in the frequency domain or were going in the opposite direction, thus probing the interplay between pitch and formant contours, and evaluating the influence of training conditions. After being trained in a Go-left/Go-right paradigm, we next manipulate the pitch and formant contours of the tone sequences in test sounds to assess their role in sound sequence recognition and the presence of perceptual interactions. Zebra finches demonstrate remarkable sensitivity to both attributes, detecting variations across harmonic tones in pitch and formant contours. In most cases the responses of the two training groups to modified stimulus versions are the same, indicating training conditions have only a limited impact on the birds' attention given to pitch and formant contours. The current study adds to an expanding body of literature supporting cognitive flexibility in songbirds and highlights a holistic approach using harmonic complex tone sequences to provide a comprehensive perspective on auditory discrimination in zebra finches.

INTRODUCTION

Music and speech share certain characteristics: both consist of sequences of acoustic units that are systematically ordered, and the continuous acoustic dimension is partitioned by attending to perceptual attributes (e.g., pitch and formant) (Patel, 2008). The basic encoding of acoustic features underlying these attributes may involve largely overlapping subcortical circuits (Patel, 2011). The cognitive processing of a certain perceptual attribute can be quite different in speech and music, reflecting the different patterns and functions the attribute has in the two domains (Patel, 2008). However, recent studies have demonstrated that experience with a perceptual attribute (e.g., pitch) in one domain can affect the perception in another domain. For instance, lexical pitch perception may have an influence on musical pitch perception, and vice versa (e.g., Sadakata *et al*., 2020; Choi W, 2021). One example of this concerns the perception of harmonic complex tones, i.e., tones that have a rich harmonic spectrum, that are present in both human music and speech. The perception of such complex tones has been shown to depend on whether they act as a musical tone (emphasizing pitch) or a speech syllable (emphasizing the formant structure) (Sadakata *et al*., 2020; Albouy *et al*., 2023). The perception of pitch and formant might also play a crucial role in the vocalization and communication of nonhuman animal (hereafter: animal) species (Hoeschele, 2017).

Pitch is conventionally defined as the perceptual correlate of a sound's fundamental frequency (f0) (Dowling & Harwood, 1986). However, what has been considered "pitch perception" in humans is mediated by several different mechanisms, not all of which involve estimating f0. A spectral-pattern tracking mechanism (irrespective of harmonic or inharmonic) that registers the direction of pitch shifts (i.e., contour) by tracking shifts in general spectral patterns, appears to operate for both musical tones and for speech (McDermott *et al*., 2008; McPherson & McDermott, 2018). On the other hand, the f0-estimation mechanism (e.g., based on harmonic patterns in the spectrum of the sound) plays an important role in tasks that required judgments of pitch intervals (the magnitude of pitch shifts) or voice identity (McPherson & McDermott, 2018). Human listeners perceive harmonic or quasi-harmonic sounds as a coherent entity, rather than as a simultaneous collection of unrelated pure tones, which suggests that the human auditory system tends to "group" or "bind" together components that are presented simultaneously and are harmonically related (Micheyl & Oxenham, 2010). Noticeably, contour representations in other dimensions besides pitch (such as loudness and

brightness contours) are also recognizable (by humans) (McDermott *et al*., 2008). Nevertheless, the f0-based pitch quality is one of the most important characteristics of human auditory experience and plays a central role in human music cognition (Patel, 2003). Pitch perception is also a focus in animal auditory perception, as studying other species can provide a more comprehensive perspective to understand the evolutionary history of pitch perception (Honing *et al*., 2015; Hoeschele *et al*., 2015; Walker *et al*., 2019).

Pitch perception has been studied in several mammal species, such as Japanese macaque (*Macaca fuscata*) (Izumi, 2001), crab-eating macaque (*Macaca fascicularis*) (Brosch *et al*., 2006) and ferret (Walker *et al*., 2009), as well as in many avian species, including European starling (*Sturnus vulgaris*) (Hulse & Cynx, 1985; MacDougall-Shackleton & Hulse, 1996), white-throated sparrow (*Zonotrichia albicollis*) (Hurly *et al*., 1990), black-capped chickadee (*Parus atricapillus*) (Weary & Weisman, 1991), and zebra finch (*Taeniopygia guttata*) (Weisman *et al*., 1994). Vocal learning avian species (especially songbirds) are more accurate than most of the mammals, including humans, when tested with absolute pitch (Weisman *et al*., 1998; Weisman *et al*., 2004; Friedrich *et al*., 2007), and they do this more readily than identifying relative pitch relationships (Hulse *et al*., 1984; Page *et al*., 1989; Weisman *et al*., 1994; MacDougall-Shackleton & Hulse, 1996; Bregman *et al*., 2012). However, humans appear to rely on "octave equivalence" to solve an absolute pitch perception task (Hoeschele, 2017; See ten Cate & Honing, 2022 for a more elaborate discussion).

Nonetheless, pitch may need not to be the primary acoustic cue for the perception of sound patterns in animals. Rats (*Rattus norvegicus*) take variations in pitch (f0) to be less psychologically distant than changes in timbre (i.e., spectral quality or "sound color") in an operant conditioning task (Crespo-Bojorque *et al*., 2022). In a series of operant conditioning tasks, Bregman *et al*. (2016) examined how a songbird, the European Starling (*Sturnus vulgaris*), perceives tone sequences in which one of three particular attributes (pitch, timbre, and spectral envelope) varies systematically over a sequence of four tones. Surprisingly, the starlings do not use pitch but the "acoustic spectral shape" (the overall distribution of the spectral energy for each tone: spectral envelope) to recognize successive tonal stimuli (Bregman *et al*., 2016). The way starlings gravitate towards spectral envelope for complex tone sequence recognition contrasts the human bias to pitch in perceiving tone sequences. While the spectral envelope provides sufficient information for accurate tone sequence recognition in starlings, it's important to note that it is not the only cue they use for sound

recognition. Starlings are capable of perceiving missing fundamentals in individual tones containing harmonic complexes (Cynx & Shapiro, 1986) and recognizing frequency-shifted conspecific songs in which the spectral envelope has been altered (Bregman *et al*., 2012).

The "formant", defined as "a characteristic peak in the spectral envelope" of vocal or musical sounds (e.g., the definition in the standards for acoustical terminology by Acoustical Society of America, 1994), has been studied in animals, suggesting that many species, including mammals and songbirds, can perceive formant information in acoustic signals. Rhesus macaques (*Macaca mulatta*) can spontaneously respond to a change in formant frequencies in their own species-typical vocalizations (Fitch & Fritz, 2006), or be trained to discriminate diverse sounds (human vowel, conspecific and heterospecific vocalizations, and artificial sounds) based on morphs and formants (Melchor *et al*., 2021). Mice (*Mus musculus*) were found to share the same perceptual mechanism as humans, which combines specific formants and temporal patterns, for detecting auditory objects and sound event streams with biological communication functions (Geissler & Ehret, 2002). In birds, studies have shown that European starlings (*Sturnus vulgaris*) (Kluender *et al*., 1998) and budgerigars (*Melopsittacus undulatus*) (Henry *et al*., 2017) can attend to formant in vowel discrimination. Similarly, zebra finches (*Taeniopygia guttata*) were found sensitive to different formant patterns in human speech (Ohms *et al*., 2012; Kriengwatana *et al*., 2015; Burgering *et al*., 2018).

Thus, to date, pitch and formant have been well documented as two of the most crucial perceptual attributes involved in speech and music. Nevertheless, the synchronous presence of both perceptual attributes may lead to perceptual interaction (concordance/competition), affecting how humans and other species organize sensory information and make perceptual judgments. Perceptual concordance refers to the state in which two or more perceptual attributes or cues in a sensory stimulus align or cooperate in the same direction or pattern, enhancing the perceptual grouping effect. On the other hand, perceptual competition denotes a situation where the same perceptual attributes or cues in a stimulus compete or work in different directions or patterns, potentially weakening the overall perceptual grouping effect. This competition doesn't imply that one attribute be completely suppressed by another but both cues may have a noticeable impact on perception, even when in conflict. Until very recently, however, the majority of research on the topic of perceptual interaction has been influenced by cognitive studies of orienting responses in human perceptual systems. Much of the research on non-human animals has employed stimuli with relatively simple attribute

patterns, such as single-tone strings, vowel-like units, or natural vocalizations. Consequently, the investigation of perceptual interaction has generally taken a secondary role in the field of auditory sensory perception research. Nevertheless, those interested in comprehending how the auditory perceptual system, typically in humans, handles and organizes more complex combinations of multiple attribute inputs into coherent perceptual objects, as well as how it interprets ambiguous inputs related to these attribute patterns, may find inspiration in the empirical research conducted in the domain of visual perception. Perceptual organization studies from human visual field show that the grouping effect is stronger and more stable when two cues concord in the same direction/pattern, while the grouping effect is weaker and more unstable when they compete in different directions/patterns (Kubovy & van den Berg, 2008; Luna & Montoro, 2011; Schmidt & Schmidt, 2013; Luna *et al*., 2016; Montoro *et al*., 2017; Villalba-García *et al*., 2018). In fact, in situations of perceptual competition between two visual cues, the non-dominant cue is perceived to a certain extent, so it's not completely suppressed by the dominant cue (Luna *et al*., 2016; Rashal *et al*., 2017). Additionally, dominance dynamics between perceptual grouping cues were found in visual competition (Palmer & Beck, 2007; Luna *et al*., 2016; Villalba-García *et al*., 2021). While it's reasonable to consider that the perceptual mechanisms demonstrated to influence perceptual organization in the visual sensory domain when multiple attribute patterns are presented within a same stimulus might also apply to auditory sensory processing, caution is warranted when conducting such examinations. In the auditory domain, a recent study by McPherson & McDermott (2023) examined the effects of timbral differences on relative pitch judgments and suggested that relative pitch judgments are not completely invariant to timbre, even in naturalistic conditions, and even when such judgments are based on representations of the fundamental frequency (f0). However, the literature has paid less attention to the potential effect of perceptual interaction (concordance/competition) between auditory-perceptual attributes compared to visual-perceptual attributes so far, especially in cross-species studies. Hence, it is interesting to examine whether the above grouping mechanisms from visual perception are also present for auditory perception when there's concordance/competition between two acoustic attributes, in a cross-species comparison paradigm.

Previous studies with zebra finches (*Taeniopygia guttata*), a popular avianmodel species for investigating the cognitive basis of auditory perception, have demonstrated that they can perceive both pitch and formant in the sound discrimination task (e.g., Burgering *et al*., 2019). The acoustic units used by Burgering *et al*.'s study (2019) resemble the structure of zebra finch

distance calls (single syllable units with a harmonic structure). These vowel-like sounds can be regarded as a relatively simple acoustic stimulus in spectral structure, consisting of one unit with a single formant peak. Another study showed that zebra finches used a particular harmonic (the 2nd harmonic located between the frequency region from 2kHz to 3kHz) as the main discriminative cue and the fundamental frequency as a secondary discriminative cue when trained to perceive the harmonic structure from a male distance call (Uno *et al*., 1997). It is likely that they are still quite sensitive to pitch (fundamental frequency), but pitch information needs to be prominent in their best hearing range. Perceptually, the pitch cues can be assessed from the harmonics, so the fundamental frequency need not be present to still be abstracted (Cynx & Shapiro, 1986). However, the potential interaction of both pitch and formant in zebra finch auditory perception has not been thoroughly examined. Specifically, the perception of changes over a series of units and the impact of Same-direction / Crossed-direction between these acoustic attributes remains unexplored.

In the current study, zebra finches were trained to perform an auditory discrimination task using a Go-left/Go-right paradigm with corrective feedback. The stimuli used consisted of sequences of five complex tones, with some small silent gaps between them. Each sequence was characterized by pitch and formant patterns (referred to as *pitch contour* and *formant contour*) that were either both increasing/decreasing in frequency or were going in the opposite direction.

In the *Same-direction condition*, the contours of pitch and formant are arranged in the same direction over the tone sequences (both ascending or both descending in frequency). In the *crossed-direction condition*, the contours of pitch and formant are in opposite directions (for example, when the pitch contour is ascending, the formant contour is descending), and the sound sequence can be perceived as two crossing contours. With this experimental paradigm, we were able to directly assess the relative contribution of pitch and formant pattern to a complex tone sequence, and determine how they contribute to the identification of the sequence.

METHODS

Subjects

Twenty-three zebra finches (12 males and 11 females) completed the task in this experiment. They were tested at the age of 334 ± 47 days post-hatching (dph), (age males: M= 366, SD= 37, age females: M= 299, SD= 26). All birds originated from the in-house breeding colony at Leiden University. Before the experiment, the birds lived in single-sex groups of about 15 to 30 individuals in aviaries (2m × 2m × 1.5m), in which food and water were available ad libitum. The housing rooms were kept at 20–22°C and 40–60% humidity and illuminated with artificial lights (Philips Master TL5 HO 49W/830) from 07:00–20:30 (13.5h light : 10.5h dark) with a 15 min twilight phase with the light fading in and out at the beginning and the end of each day. A week before the operant test, birds were caught and transferred from the aviary to standard laboratory cages (two birds of equal sex in one cage) in order to acclimatize (cage size: length × width × height = 80 × 40 × 40 cm) and reduce stress from catching in aviary. The birds were divided randomly in two experimental groups: twelve of the birds were assigned to the Same-direction group, and the other eleven birds to the Crossed-direction group (6 males and 6 females in Same-direction group, 6 males and 5 females in Crossed-direction group; age Same-direction group: M=316, SD=33, age Crossed-direction group: M=353, SD=51).

Operant conditioning cage

Zebra finches were trained and tested individually in an operant conditioning cage (Skinner Box) (70x30x45 cm). The cage was built from wire mesh walls and one foamed PVC back wall and was containing 3 pecking sensors with a red LED light at the top of each sensor (Fig. 1A). Each operant cage was situated in a separate sound-attenuated chamber. The chamber was illuminated by a fluorescent lamp (Phillips Master TL-D 90 DeLuxe 18W/ 965, The Netherlands), which emitted a daylight spectrum following a 13.5-h/10.5-h light/dark schedule. Sound stimuli were played through a speaker (Vifa MG10SD09–08, Vifa, Viborg, Denmark) 1 meter above the Skinner Box. The volume of the speaker was adjusted to ensure that the sound amplitude in the Skinner Box was approximately 65 dB (measured by an SPL meter - RION NL 15, RION). Sensors (S1, S2, S3), lamp, food hatch and speaker were connected to operant conditioning controller that also registered all sensor pecks.

Stimuli

Training stimuli

Each experimental group was trained to discriminate between a pair of tonal sequences consisting of five complex tones, and within each group part of the birds got one pair of training stimuli ("sequence series 1") and the rest another pair of training stimuli ("sequence series 2") (see Table 1). The training stimuli were synthesized, normalized and filtered using Praat (version 6.1.12) and Audacity (version 2.3.0). The stimuli have not been heard before by the birds. To synthesize a sound unit, the first step was to construct a complex tone with a defined f0 value (e.g., 150Hz as the f0) by choosing the function "Create Sound as tone complex" (to create a sound combining the f0 and its constituent harmonics occurring at integer multiples of the f0) in Praat. Secondly, the complex tone was manipulated by the Effect "Fade in/Fade out" and its peak amplitude was Normalized to "-50 dB" by Audacity. Additionally, to create a formant for the unit, the amplitude located in a particular frequency range (e.g., the Formant peak located in 2.6kHz) was amplified (Width 1kHz, Gain 15dB x2) by applying the "Parametric EQ" function (an effect plugin) in Audacity. Afterwards these complex tones were combined into a single sequence using Praat, with each tone (duration: 0.21s) separated by a silence of 0.05s, resulting in a sequence with a duration of 1.25s. All stimuli were low-pass filtered (below 8kHz).

The training stimuli in this experiment were 4 stimulus pairs (2 sequence series/pairs for each training group, as shown in Table 1), each pair was consisting of two different tone sequences that differ in both pitch and formant direction. Two training groups were categorized by the training stimuli the birds trained with: for the Crossed-direction group, the training stimuli was a pair of complex tones that within which both the pitch contour and formant contour of a single training stimulus were going in the opposite direction: for instance, the pitch contour of training stimulus A was decreasing (in the frequency domain over an entire tone sequence), while the formant contour of training stimulus A was increasing (Fig. 1B). For the Same-direction group, the training stimuli was a pair of complex tones that within which both the pitch contour and formant contour of a single training stimulus were going in the same direction (both contours of two attributes were either increasing or decreasing) (Fig. 1D).

In each pair of training stimuli, one stimulus (e.g., training stimulus A) featured a formant-ascending pattern, for example, “2.6kHz- 3.1kHz- 3.6kHz- 4.1kHz- 4.6kHz”. This formant-ascending pattern increased the amplitude (+30 dB) within specific frequency bands of each complex tone. These amplitude-enhanced frequency bands fell within the zebra finches’ hearing threshold, as determined by Okanoya & Dooling (1987), typically peaking around 3.5-4.0 kHz. In contrast, the complex tones of the other stimulus (e.g., training stimulus B) followed a formant-descending pattern, for instance, “4.6kHz- 4.1kHz- 3.6kHz- 3.1kHz- 2.6kHz”. Similarly, one sequence featured a pitch-ascending pattern, such as “150Hz- 220Hz- 290Hz- 360Hz- 430Hz”, while the other exhibited a pitch-descending pattern, for instance, “430Hz- 360Hz- 290Hz- 220Hz- 150 Hz”. The stimuli from different sequence series of the same training group were arranged in the same way, but with different f0 and formant values. This was done to prevent the selected value set from accidentally coinciding with frequencies that might hold specific biological significance or relevance for the birds. When played, the sequences were normalized such that the average intensity (RMS, calculated over the total duration of the stimulus) was the same for the two sequences within a pair to avoid amplitude differences affecting the responses to the stimuli.

Table 1 Overview of the training groups

Training group	Sequence series	Subjects (n)	Sequence of tonal units
Crossed-direction	Series 1	3 ♂ + 3 ♀	430Hz(2.6kHz)-360Hz(3.1kHz)-290Hz(3.6kHz)-220Hz(4.1kHz)-150Hz(4.6kHz)
Crossed-direction	Series 2	3 ♂ + 2 ♀	405Hz(2.1kHz)-330Hz(2.7kHz)-255Hz(3.3kHz)-180Hz(3.9kHz)-105Hz(4.5kHz)
Same-direction	Series 1	3 ♂ + 3 ♀	150Hz(2.6kHz)-220Hz(3.1kHz)-290Hz(3.6kHz)-360Hz(4.1kHz)-430Hz(4.6kHz)
Same-direction	Series 2	3 ♂ + 3 ♀	105Hz(2.1kHz)-180Hz(2.7kHz)-255Hz(3.3kHz)-330Hz(3.9kHz)-405Hz(4.5kHz)

Note: In each training group, the tone sequences were arranged in the same way (the pitch and formant contours of a stimulus were either going in the same or crossed direction) but differed in f0 and Formant values of their complex tones. The numerical values between the "-" indicates the f0 and Formant values of each complex tone (e.g., "430Hz(2.6kHz)" stands for f0 = 430Hz and Formant peak = 2.6kHz). The values of tonal units for tone sequence are shown in this table, with training stimulus A presented as an exemplar. While training stimulus B is not displayed in this table, note that the values of tonal units for training stimulus B were identical to those of training stimulus A, with the only difference being that its tonal units were arranged in the reverse order of training stimulus A. See text for details.

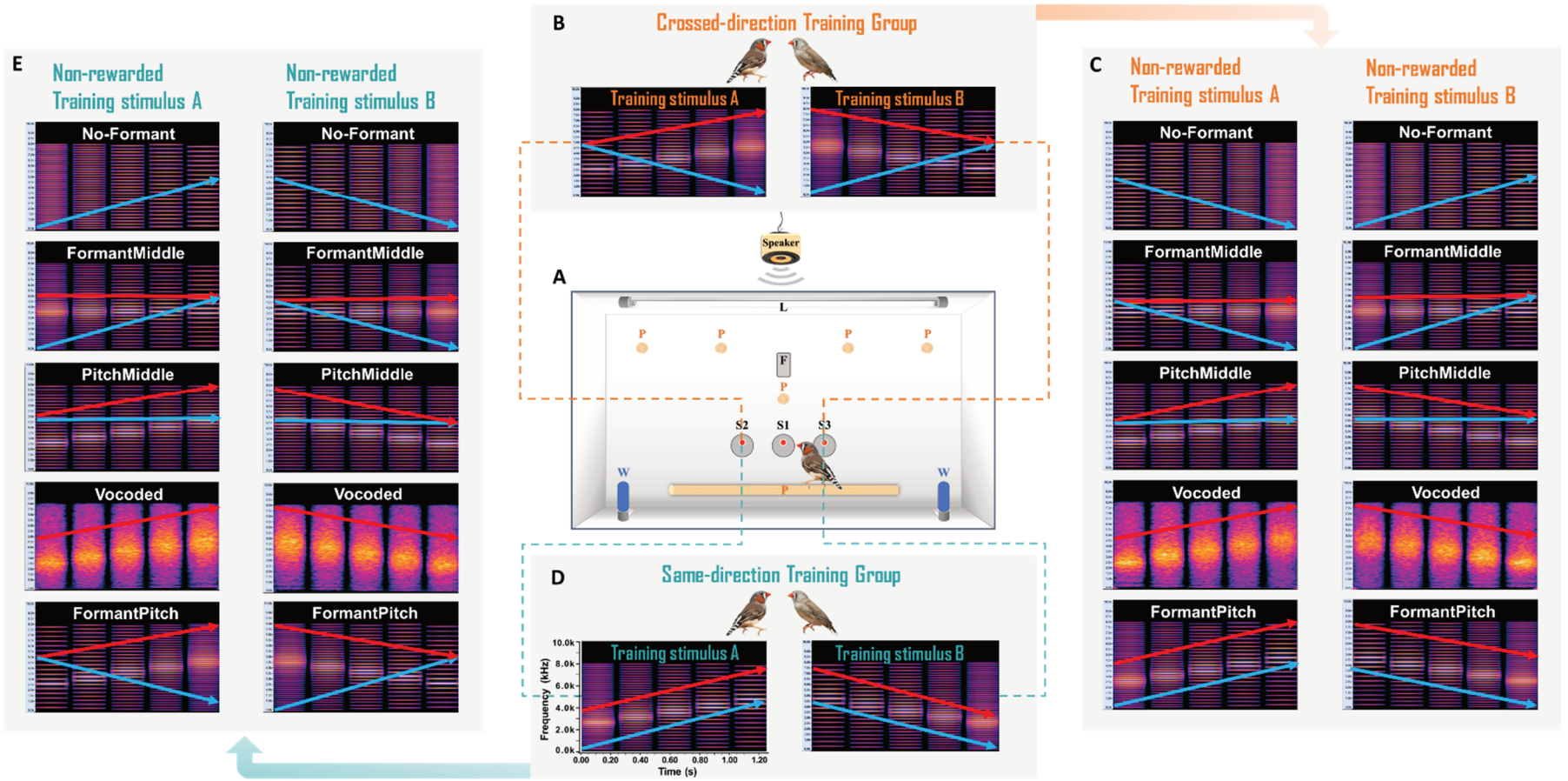


Figure 1. Operant conditioning apparatus (Skinner box) and stimuli used for the experiment: **(A)** Schematic front view of the Skinner box. A speaker (top of figure) is suspended from the ceiling above the cage. Within the cage, there are several perches (P) for the bird to sit on, a food hatch (F) located in the upper middle of the back panel, a lamp (L) is placed at the top of the cage. Two tubes with ad libitum water (W) are placed symmetrically on two sides of the cage, and three sensors (S1, S2, S3) with red LEDs are lined horizontally in the lower middle of the back panel. **(B)** An example of a pair of training stimuli for Crossed-direction group. **(C)** Modified stimuli used in the testing phase for Crossed-direction group. The birds of Crossed-direction group were tested with 5 modified versions of each training stimulus after completion of the training — see text for a description of these manipulations. **(D)** A pair of training stimuli for Same-direction group. **(E)** Modified stimuli in the testing phase for Same-direction group. The birds of Same-direction group were also tested with 5 similarly modified versions of each training stimulus. In the spectrogram of each tone sequence, the direction (but not the exact value) of the pitch contour is indicated by a blue line, and formant contour is indicated by a red line.

Test stimuli

To test the relative importance of the pitch contour and formant contour in discrimination of the training stimuli, the birds were tested with modified versions of the training stimuli (Fig. 1C & 1E). Praat and Audacity were used to edit each original training stimulus to produce a version in which either pitch contour or the formant contour was changed. For both the Crossed-direction and the Same-direction training group, the test stimuli were always modified from the training stimuli in an identical way (some examples of the training and test stimuli are provided as supplementary material):

- No-Formant – The purpose of this manipulation was to create stimuli where all frequency bands have the same energy, while keeping the f0 values identical to training stimuli. The construction of the No-Formant version followed the same procedure as that of the training stimuli, except for omitting the Formant synthesis step. This ensured that the No-Formant version maintained the same characteristics (tone sequences with increasing/decreasing pitch contour) as the training stimuli, yet without any formant.

- FormantMiddle – In this stimulus a single, fixed formant was added to the training stimulus, using the formant value of the middle unit of the sequences. As such it preserved the pitch contour, while all elements have the same spectral envelope. This manipulation was accomplished by assigning a same "Frequency (Hz)" value (the "Parametric EQ" function in Audacity) to tonal units of different f0.

- PitchMiddle – In this manipulation, all pitches were equalized to the f0 of the middle unit, while preserving the formant contour of the initial training stimulus. For example, a sequence with five tonal units featuring "f0=290Hz" and arranged in a formants-ascending pattern of "2.6kHz- 3.1kHz- 3.6kHz- 4.1kHz- 4.6kHz". This adjustment was achieved by modifying the "Frequency (Hz)" parameter within the "Parametric EQ" function in Audacity.

- Vocoded – This modification maintains the spectral envelope (both its shape and position) of the training stimulus, but averages the energy within specific frequency bands, thus removing any harmonic structure. For this we used the Matt Winn's Praat vocoded script (http://www.mattwinn.com/praat/vocode_all_selected_v45.txt) to synthesize a vocoded version of training stimuli. The script was set to divide cut-off frequency bandwidths equally

for 30 bands contiguous with smooth transitions (From lowCornerFreq 100Hz to highCornerFreq 8000Hz).

- FormantPitch – In this modification, the formant contour of the training stimulus was altered by adjusting the formant frequencies of each tonal unit and arranging them in a reversed order compared to the initial training stimulus. This modification aimed to combine the formant contour of one training stimulus with the pitch contour of the other training stimulus from the same training pair. For example, a modified sequence would share the pitch contour of training stimulus A while sharing the formant contour of training stimulus B. In the analysis, the responses to this ambiguous "FormantPitch" stimulus are compared to the training stimulus that shares the same formant contour (see Fig. 1C & 1E).

Procedure

A Go-Left/Go-Right paradigm was employed for both training and testing. The experimental procedure consisted of five phases, namely acclimation, pre-training, discrimination training, transition, and probe testing.

Acclimation phase

In the acclimation phase, each of the birds was introduced to a Skinner box, with the food hatch left open. The pecking sensors' LED lights were illuminated to attract attention from the bird. The primary objective of this phase was to familiarize the birds with the cages and the location of the food source. Pecking the central sensor, S1, resulted in the playback of either sound A or sound B with a 50% probability for each. Pecking one of the side sensors, S2 or S3, triggered the playback of one of the two sounds. After a period of several hours or overnight, the food hatch was closed, marking the transition to the next phase.

Pre-training phase

This phase aimed to acquaint the birds with the training procedures. In this phase, the food hatch was closed so that the birds had to learn to peck all three sensors. Pecking each sensor had specific consequences: pecking S1 resulted in sound A or sound B playback without food reinforcement, pecking S2 led to sound A playback accompanied by a 15-seconds food hatch opening, and pecking S3 triggered sound B playback along with a 15-seconds food hatch opening. This training continued until the birds consistently pecked all sensors and associated

sensor pecking with access to food. Some of the birds may also learn the association between specific sounds and the corresponding sensors in this phase. In cases when the birds did not spontaneously peck the sensors, the experimenter could activate or deactivate the LED lights to attract the bird's attention. Once the birds consistently pecked all sensors for several days, the discrimination training phase commenced.

Discrimination training

During the discrimination training phase, the birds were trained to peck the middle sensor (S1) to elicit sound playback, and then to subsequently peck either the left or right sensor, depending on the played sound triggered by the middle sensor. Correct responses, where the bird pecked the sensor associated with the played sound, were rewarded with a 15-second access to food hatch as the positive feedback. If an incorrect sensor was pecked, the light of the lamp was turned off for 1 second as a signal of negative feedback for the bird. Prior to any pecking, only the LED light for S1 was illuminated. For instance, when sound A was played, pecking S2 opened the food hatch, while pecking S3 resulted in a preset time of darkness, and vice versa. If a bird failed to respond within 15 seconds, the trial ended without a food reward or a light-off hint. The duration of this phase varied among individual birds. The discrimination rate for each bird, representing the proportion of correct responses out of all trials, was calculated on a daily basis. Once a bird achieved a discrimination score greater than 0.75 for the training stimuli for three consecutive days (with an accuracy rate of each sensor pecking exceeding 0.60 for three consecutive days), it was considered to have successfully discriminated the trained sequence pair, and the training transitioned to the next phase.

Transition phase

During the transition phase, the training stimuli remained the same as in the discrimination training phase, but the frequency of reinforcement by food or darkness was reduced to occur randomly on 80% of the trials (instead of 100% during the discrimination training phase). In the remaining 20% of trials (with stimuli identical to the training sounds), the subjects did not receive reinforcement in the form of food or darkness. If the birds maintained the same level of discrimination for two consecutive days during this phase, the test phase began.

Probe testing phase

In this phase, 20% of the pecks on S1 resulted in presenting one of twelve probe stimuli. These twelve probe stimuli were never reinforced and were randomly interspersed between training stimuli. Ten of these were modified versions of the training stimuli (five modified versions of training stimulus A and five of training stimulus B). The other two probe stimuli were non-reinforced training stimuli. The remaining 80% were training stimuli with reinforcement. Testing continued until each probe stimulus had been presented 40 times to a bird. After reaching this, the bird was transferred back to its aviary. The order of stimulus presentation was randomized across the subjects.

Analysis

To assess potential differences in the speed of discrimination learning between the two training groups, we analyzed the cumulative number of trials until reaching the learning criterion, including the day when the criterion was achieved. As the distribution of trial numbers did not conform to a normal distribution, a Mann-Whitney-Wilcoxon Test (R Core Team, 2016) was employed to examine any significant differences in learning speed (i.e., the number of required training trials) between the two training groups.

The reactions to the various test stimuli were classified into three categories: a "correct response" (i.e., the bird identifies the modified version of training stimulus A as A, and the modified version of training stimulus B as B), an "incorrect response" (responding with pecking the sensor for B if the stimulus was a modification of sound A and vice versa), and a "nonresponse" (not pecking a sensor). For the statistical analyses, we examined the proportion of "correct responses" out of "correct + incorrect responses" (Correct rate = Number_CorrectResp / (Number_CorrectResp + Number_IncorrectResp)), as well as the "response rate", calculated as "correct + incorrect responses" to modifications of sound A plus those to modification of sound B, as the proportion of the 40 presentations of each test stimulus (Response rate = (Number_CorrectResp + Number_IncorrectResp) / (Number_CorrectResp + Number_IncorrectResp + Number_NoResp)).

We used Generalized Linear Mixed-effects Models (GLMMs) to examine the discrimination of various test sounds by the birds. All model analyses were conducted in Rstudio (R Core Team, 2016 & lme4; Bates *et al*., 2015). We calculated the "Correct rate" and the "Response

rate" based on the counts of "correct response", "incorrect response", and "no response", combining the response counts to (variants of) Training stimuli A and B, (using the function cbind, R package mice; Van Buuren & Groothuis-Oudshoorn, 2011), and used these two proportions rates as response variables in GLMMs in R (using the function glmer, R package lme4; Bates *et al*., 2015). We used "Training_Group" (Same-direction or/ Crossed-direction training pairs), "Test_Treatment", "Sequence_Series" and the interaction between "Training_Group" and "Test_Treatment" as covariates in the full model with "Bird_ID", "Age", "Number_of_Training_Trials" as the random factors and a binomial error structure of the "Correct rate" and the "Response rate". The best model was chosen based on corrected Akaike criterion (AICc) provided by dredge model selection (using the function Dredge, R package MuMIn; Bartoń, 2020). The model with the smallest value of AICc was considered to be the best model by default, but if "Training_Group", "Test_Treatment" and the interaction between these two were not part of the best model, we kept them in the final model anyway because these were variables of our interest. To determine the effect and significance of the covariates, we ran the final models and, if applicable, used Post hoc Tukey's HSD tests to make pairwise comparisons of the test treatments (using the emmeans function, R package lsmeans; Lenth, 2016), with false discovery rate (FDR) correction of p-values (Benjamini & Hochberg, 1995) for multiple comparisons. In the above models, the counts of the responses to (modifications of) both sequence A and sequence B were combined in all tests. In the above models, the counts of the reactions to modifications of both sound A and sound B were combined.

Additionally, to determine whether the individual test stimuli were discriminated above chance (50%), the ratio of "Number_CorrectResp / Number_IncorrectResp" was assessed (specifically, whether this ratio differed from 1). We did so by applying the log (Number_CorrectResp / Number_IncorrectResp) (indicated as "Log (Cor/Inco)" from now on as the response variable against a log (Odds-ratio) = 0 in a GLM analysis. If correct/incorrect = 1, then the probability of observing a correct response is as large as the probability of observing an incorrect response, representing both probabilities are 0.5, then log (Odds- ratio) = log (1) = 0. Therefore, comparing the outcomes of the Binomial GLM to 0 is comparing the results to the 50% chance for a correct response.

RESULTS

Learning speed

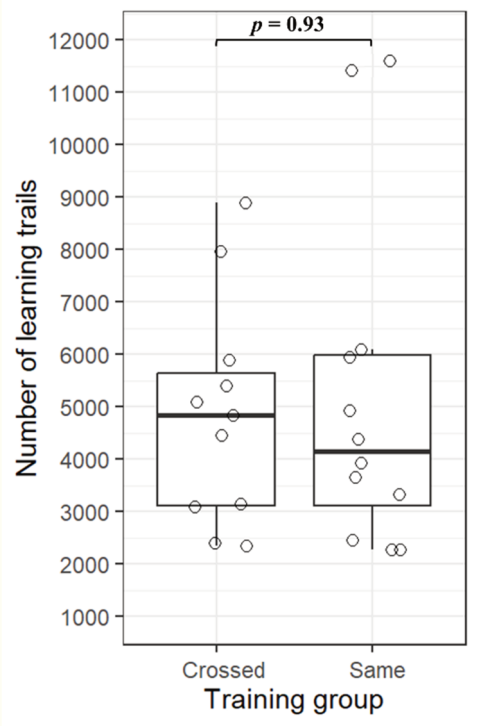


Figure 2. Number of learning trials needed to reach the learning criterion. Individual zebra finch results are shown with open circles. There is no significant difference between the Different-syllables group and the Same-syllables group in learning speed. Box plots show median, 1st and 3rd quartile, and whiskers the 1.5 interquartile range.

The discrimination training lasted until the zebra finches reached the learning criterion of over 75% correct responses to both sound A and sound B for three days. All twenty-three birds finished the training and learned the discrimination on a median value of 4465 (IQR = 2857) trials to reach the criterion. No significant difference ($p = 0.93$, z = 0.12) was found between the Crossed-direction group (Median = 4841, IQR = 2808) and the Same-direction group (Median = 4158, IQR = 3126). It suggests that birds from two training groups learn approximately equally fast in both training conditions.

The impact of pitch and formant on stimulus classification

Do training groups differ in responses to test stimuli?

We compared the Correct rates and Response rates for both experimental groups in response to the training and various test stimuli (Fig. 1). We chose the two factors, "Training_Group" and "Test_Treatment", along with their interaction effects, which were used as fixed factors in the statistical models for the response variables "Correct rates" and "Response rates" (see Table 2). Although the two selected models were not the most recommended ones based on the dredge model selection, they included the variables of interest and were still close to the most recommended models (based on AICc).

All modifications of the training stimuli resulted in a strong reduction of the correct rate indicating that both formant and pitch were used to distinguish the training stimuli, irrespective of the training group. Most test stimuli did not exhibit significant differences in the correct rate between the two training groups (see Fig. 3A), with the exception of the "No-Formant" version, which showed a significant distinction. In this case, the Crossed-direction group achieved a higher Correct rate compared to the Same-direction group (Crossed – Same = 0.347 ± 0.114, $p = 0.014$, as indicated in Table 3).

There were no significant differences in Response rates for any of the stimuli between the two training groups (Fig. 3B). Notably, the variation in Response rate for all five modified stimuli in the Same-direction group was more prominent compared to the Crossed-direction group. This suggests that the Same-direction training condition, rather than the Crossed-direction, might affect the consistency of individual responses to the modified stimuli, or that some individuals within the Same-direction group consistently respond more often than others.

Table 2 Summary of the GLMs selection for a) the proportion of correct responses if birds respond to one of two sounds; and b) the proportion of trials that birds respond with pecking A or B.

Model	df	logLik	AICc	Δi	wi
a. Correct rate of responses					
1 Training_Group + Test_Treatment + Test_Treatment:Training_Group + Sequence_Series + (1|Bird_ID) + (1|Age) + (1|Number_of_Training_Trials)	16	-408.692	853.9	2.80	0.073
2* Training_Group + Test_Treatment + Test_Treatment:Training_Group + (1|Bird_ID) + (1|Age) + (1|Number_of_Training_Trials)	15	-408.694	851.3	0.24	0.263
3 Training_Group + Test_Treatment + (1|Bird_ID) + (1|Age) + (1|Number_of_Training_Trials)	10	-415.021	851.8	0.69	0.210
4 Test_Treatment + Sequence_Series + (1|Bird_ID) + (1|Age) + (1|Number_of_Training_Trials)	10	-415.835	853.4	2.32	0.093
5 Test_Treatment + (1|Bird_ID) + (1|Age) + (1|Number_of_Training_Trials)	9	-415.839	851.1	0.00	0.296
null (1|Bird_ID) + (1|Age) + (1|Number_of_Training_Trials)	4	-764.165	1536.6	685.55	0.000
b. Response rate of trials					
1* Training_Group + Test_Treatment + Test_Treatment:Training_Group + (1|Bird_ID) + (1|Age) + (1|Number_of_Training_Trials)	15	-441.024	916.0	3.17	0.075
2 Training_Group + Test_Treatment + Sequence_Series + (1|Bird_ID) + (1|Age) + (1|Number_of_Training_Trials)	11	-445.658	915.4	2.60	0.100
3 Training_Group + Test_Treatment + (1|Bird_ID) + (1|Age) + (1|Number_of_Training_Trials)	10	-446.210	914.2	1.34	0.187
4 Test_Treatment + Sequence_Series + (1|Bird_ID) + (1|Age) + (1|Number_of_Training_Trials)	10	-445.983	913.7	0.89	0.235
5 Test_Treatment + (1|Bird_ID) + (1|Age) + (1|Number_of_Training_Trials)	9	-446.703	912.8	0.00	0.366
null (1|Bird_ID) + (1|Age) + (1|Number_of_Training_Trials)	4	-715.047	1438.4	525.58	0.000

Note: Best five models of the model selection (ranked by AICc and logLik) and the null models. The Akaike weight (wi) indicates the probability of a better model in the model candidates set, and Delta AICc (Δi) was used to show the difference in AICc score between the best model and the model being compared. A * indicates the model we choose.

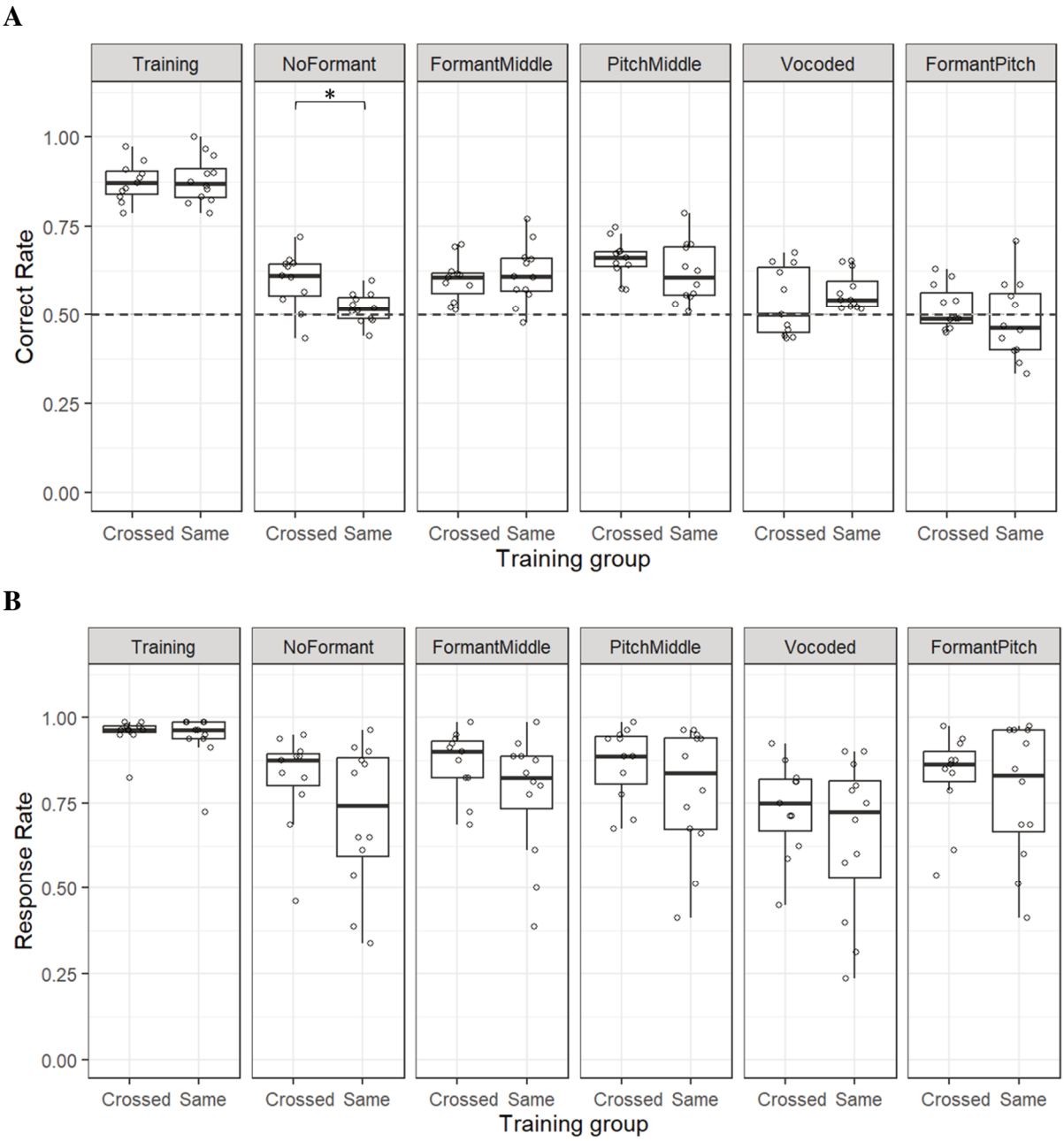


Figure 3. Correct rate of responses and Response rate of trials: (A) the proportion of correct responses (Correct rate) to the training and modified stimuli for the two training groups; **(B)** the Response rates to the training and modified stimuli for the two training groups. "Crossed" refers to the Crossed-direction training group, and "Same" refers to the Same-direction training group. Significant differences between the responses between the training groups are indicated: * refers to a significant difference of $0.01 < p \leq 0.05$, and for non-indicated comparisons p value is > 0.05. Box plots show median, 1st and 3rd quartile, and whiskers the 1.5 interquartile range. The dashed line represents chance level, which was 50% for both tasks.

Do different test stimuli give rise to different responses?

Post hoc Tukey's HSD tests (Table 3a, b) revealed significantly higher Correct rates and Response rates for the non-rewarded training stimuli compared to all five modified stimuli within each training group (both $p < 0.0001$).

Both the Crossed-direction and Same-direction groups exhibited higher correct rates in response to the "PitchMiddle" stimuli compared to the "No-Formant" (Crossed-direction group: $p < 0.05$; Same-direction group: $p = 0.0001$), "Vocoded" (Crossed-direction group: $p < 0.0001$; Same-direction group: $p < 0.05$), and "FormantPitch" (Crossed-direction group: $p < 0.0001$; Same-direction group: $p < 0.0001$) stimuli. Additionally, both groups showed higher correct rates in response to the "FormantMiddle" stimuli compared to "FormantPitch" stimuli (Crossed-direction group: $p < 0.01$; Same-direction group: $p < 0.0001$). Moreover, the Crossed-direction group responded with significantly higher correct rates to the "PitchMiddle" stimuli than to the "FormantMiddle" stimuli ($p < 0.05$) and responded with higher correct rates to the "No-Formant" stimuli than to the "Vocoded" ($p < 0.05$) and "FormantPitch" ($p < 0.01$) stimuli. The Same-direction group exhibited significantly higher correct rates when responding to the "FormantMiddle" stimuli than to the "No-Formant" ($p < 0.001$). Additionally, a discernible trend towards differentiation between the "No-Formant" and "Vocoded" stimuli, as well as between the "FormantMiddle" and "Vocoded" stimuli, was observed in the Same-direction group (both $p = 0.08$).

Both groups showed a higher response rate in responding to the "PitchMiddle" and the "FormantMiddle" stimuli than to the "No-Formant" (Crossed-direction group: both $p < 0.01$; Same-direction group: both $p < 0.001$) and the "Vocoded" (Crossed-direction group: both $p < 0.0001$; Same-direction group: both $p < 0.0001$). Additionally, both groups responded with a significantly lower response rate to the "Vocoded" stimuli than to other four modified stimuli (Crossed-direction group: both $p < 0.0001$; Same-direction group: both $p < 0.01$). Moreover, the Crossed-direction group responded with a significantly higher response rate to the "PitchMiddle" and the "FormantMiddle" stimuli than to the "FormantPitch" (both $p < 0.05$) stimuli. The Same-direction group responded with a significantly higher response rate to the "FormantPitch" stimuli than to the "No-Formant" ($p < 0.001$) and "Vocoded" ($p < 0.0001$) stimuli.

Overall, the response rates of birds in both groups shows a pattern that is somewhat similar to their correct rate for most of the modified stimuli. In both groups, the birds predominantly responded to the "FormantMiddle" and "PitchMiddle" stimuli, while responding least to the "Vocoded" stimuli. However, noteworthy is the relatively high response rate to the "FormantPitch" stimuli in both groups, even though the correct rate for this modified version was relatively low.

Table 3 Post hoc test results of Binomial GLM for the interaction of Test_Treatment & Training_Group

Stimuli	Training_Group	estimate	SE	z.ratio	*p*.value
a. Correct rate of responses					
Training	Crossed - Same	-0.032	0.150	-0.214	0.8302
No-Formant	**Crossed - Same**	**0.347**	**0.114**	**3.039**	**0.0144**
FormantMiddle	Crossed - Same	-0.081	0.112	-0.721	0.5650
PitchMiddle	Crossed - Same	0.148	0.113	1.305	0.4869
Vocoded	Crossed - Same	-0.117	0.119	-0.985	0.4869
FormantPitch	Crossed - Same	0.128	0.111	1.155	0.4869
Training - No-Formant	**Crossed**	**1.522**	**0.129**	**11.807**	**<.0001**
Training - FormantMiddle	**Crossed**	**1.544**	**0.128**	**12.109**	**<.0001**
Training - PitchMiddle	**Crossed**	**1.281**	**0.129**	**9.929**	**<.0001**
Training - Vocoded	**Crossed**	**1.781**	**0.131**	**13.636**	**<.0001**
Training - FormantPitch	**Crossed**	**1.866**	**0.128**	**14.596**	**<.0001**
No-Formant - FormantMiddle	Crossed	0.022	0.106	0.210	0.8333
No-Formant - PitchMiddle	**Crossed**	**-0.241**	**0.108**	**-2.233**	**0.0320**
No-Formant - Vocoded	**Crossed**	**0.259**	**0.110**	**2.361**	**0.0249**
No-Formant - FormantPitch	**Crossed**	**0.344**	**0.106**	**3.236**	**0.0023**
FormantMiddle - PitchMiddle	**Crossed**	**-0.263**	**0.106**	**-2.477**	**0.0199**
FormantMiddle - Vocoded	**Crossed**	**0.237**	**0.108**	**2.190**	**0.0329**
FormantMiddle - FormantPitch	**Crossed**	**0.322**	**0.105**	**3.073**	**0.0035**
PitchMiddle - Vocoded	**Crossed**	**0.500**	**0.110**	**4.549**	**<.0001**
PitchMiddle - FormantPitch	**Crossed**	**0.585**	**0.107**	**5.489**	**<.0001**
Vocoded - FormantPitch	Crossed	0.085	0.108	0.786	0.4626
Training - No-Formant	**Same**	**1.901**	**0.127**	**14.935**	**<.0001**
Training - FormantMiddle	**Same**	**1.496**	**0.126**	**11.825**	**<.0001**
Training - PitchMiddle	**Same**	**1.461**	**0.126**	**11.562**	**<.0001**
Training - Vocoded	**Same**	**1.696**	**0.130**	**13.063**	**<.0001**

Training - FormantPitch	**Same**	**2.026**	**0.125**	**16.184**	**<.0001**
No-Formant - FormantMiddle	**Same**	**-0.405**	**0.108**	**-3.762**	**0.0003**
No-Formant - PitchMiddle	**Same**	**-0.440**	**0.107**	**-4.092**	**0.0001**
No-Formant - Vocoded	Same	-0.205	0.111	-1.837	0.0811
No-Formant - FormantPitch	Same	0.125	0.106	1.182	0.2543
FormantMiddle - PitchMiddle	Same	-0.035	0.107	-0.327	0.7440
FormantMiddle - Vocoded	Same	0.200	0.111	1.810	0.0811
FormantMiddle - FormantPitch	**Same**	**0.530**	**0.105**	**5.041**	**<.0001**
PitchMiddle - Vocoded	**Same**	**0.235**	**0.110**	**2.128**	**0.0455**
PitchMiddle - FormantPitch	**Same**	**0.565**	**0.105**	**5.381**	**<.0001**
Vocoded - FormantPitch	**Same**	**0.330**	**0.109**	**3.026**	**0.0037**
b. Response rate of trials					
Training	Crossed - Same	-0.036	0.365	-0.098	0.9964
No-Formant	Crossed - Same	0.525	0.316	1.663	0.5294
FormantMiddle	Crossed - Same	0.507	0.321	1.579	0.5294
PitchMiddle	Crossed - Same	0.416	0.321	1.296	0.5294
Vocoded	Crossed - Same	0.322	0.311	1.034	0.5576
FormantPitch	Crossed - Same	0.116	0.318	0.364	0.8542
Training - No-Formant	**Crossed**	**1.591**	**0.187**	**8.509**	**<.0001**
Training - FormantMiddle	**Crossed**	**1.202**	**0.193**	**6.239**	**<.0001**
Training - PitchMiddle	**Crossed**	**1.212**	**0.192**	**6.299**	**<.0001**
Training - Vocoded	**Crossed**	**2.130**	**0.182**	**11.694**	**<.0001**
Training - FormantPitch	**Crossed**	**1.557**	**0.187**	**8.313**	**<.0001**
No-Formant - FormantMiddle	**Crossed**	**-0.389**	**0.137**	**-2.844**	**0.0067**
No-Formant - PitchMiddle	**Crossed**	**-0.378**	**0.136**	**-2.773**	**0.0076**
No-Formant - Vocoded	**Crossed**	**0.539**	**0.121**	**4.459**	**<.0001**
No-Formant - FormantPitch	Crossed	-0.033	0.129	-0.259	0.8525
FormantMiddle - PitchMiddle	Crossed	0.010	0.144	0.072	0.9423
FormantMiddle - Vocoded	**Crossed**	**0.928**	**0.130**	**7.145**	**<.0001**
FormantMiddle - FormantPitch	**Crossed**	**0.355**	**0.137**	**2.589**	**0.0120**
PitchMiddle - Vocoded	**Crossed**	**0.917**	**0.130**	**7.080**	**<.0001**
PitchMiddle - FormantPitch	**Crossed**	**0.345**	**0.137**	**2.518**	**0.0136**
Vocoded - FormantPitch	**Crossed**	**-0.572**	**0.122**	**-4.709**	**<.0001**
Training - No-Formant	**Same**	**2.149**	**0.164**	**13.140**	**<.0001**
Training - FormantMiddle	**Same**	**1.743**	**0.166**	**10.516**	**<.0001**
Training - PitchMiddle	**Same**	**1.663**	**0.166**	**9.998**	**<.0001**

Training - Vocoded	**Same**	**2.484**	**0.163**	**15.273**	**<.0001**
Training - FormantPitch	**Same**	**1.707**	**0.166**	**10.282**	**<.0001**
No-Formant - FormantMiddle	**Same**	**-0.406**	**0.116**	**-3.509**	**0.0006**
No-Formant - PitchMiddle	**Same**	**-0.486**	**0.117**	**-4.162**	**0.0001**
No-Formant - Vocoded	**Same**	**0.335**	**0.110**	**3.060**	**0.0028**
No-Formant - FormantPitch	**Same**	**-0.442**	**0.116**	**-3.805**	**0.0002**
FormantMiddle - PitchMiddle	Same	-0.080	0.120	-0.663	0.5853
FormantMiddle - Vocoded	**Same**	**0.741**	**0.114**	**6.507**	**<.0001**
FormantMiddle - FormantPitch	Same	-0.036	0.120	-0.300	0.7641
PitchMiddle - Vocoded	**Same**	**0.821**	**0.115**	**7.140**	**<.0001**
PitchMiddle - FormantPitch	Same	0.044	0.121	0.363	0.7641
Vocoded - FormantPitch	**Same**	**-0.777**	**0.114**	**-6.794**	**<.0001**

Note: Response variables in GLMMs: (**a**) the proportion of correct responses if birds respond to one of two sounds; and (**b**) the proportion of trials that birds respond with pecking A or B. "Crossed" refers to the Crossed-direction training group, and "Same" refers to the Same-direction training group. Each of the corrected pairwise multiple comparison tests is separated by borders within the table. Bold indicates significant differences <0.05.

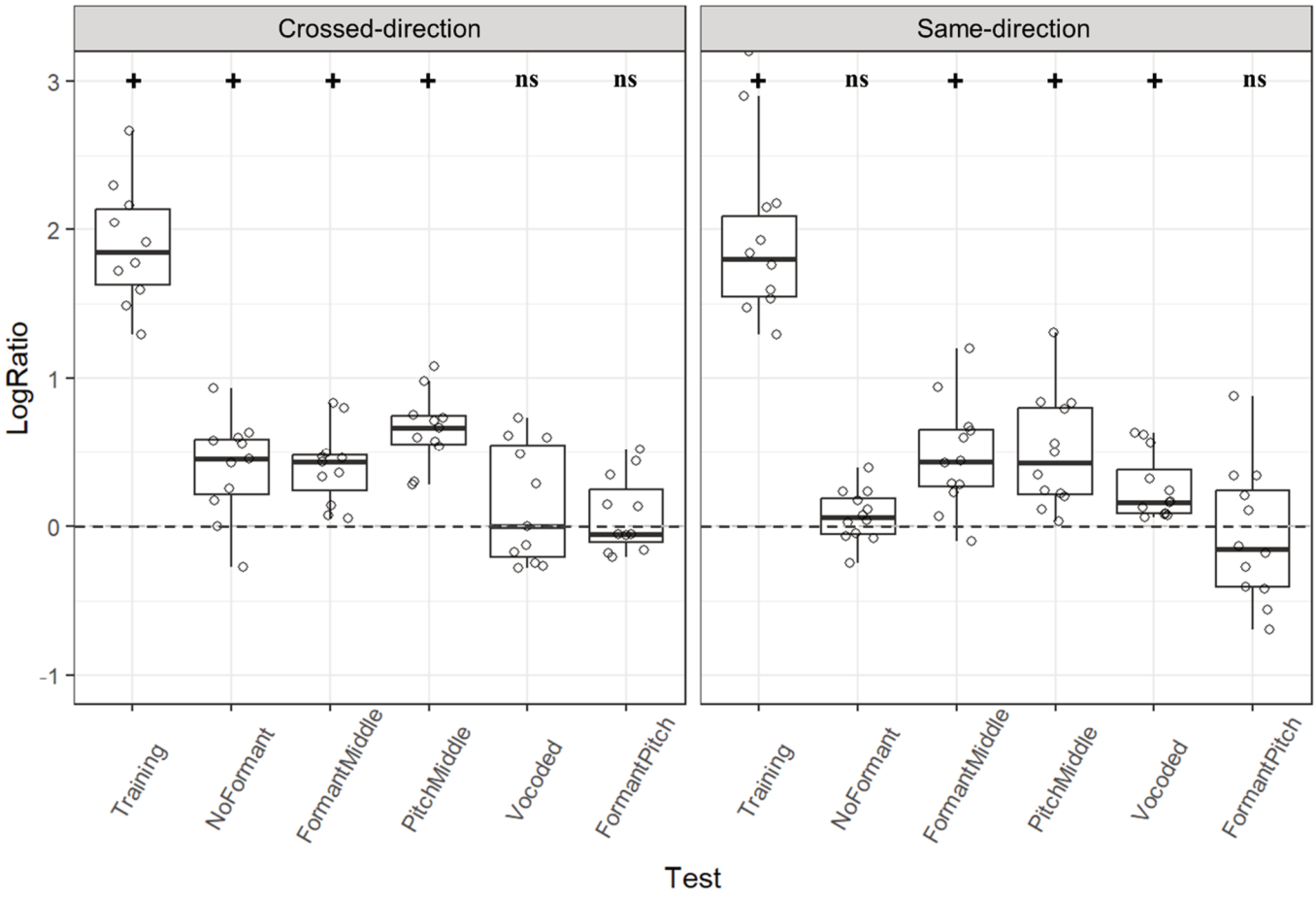
Are modified stimuli still discriminated?

The previous analyses primarily focused on disparities in Correct rates among training groups and across test stimuli. However, these analyses did not show whether a low Correct rate means that birds are no longer able to discriminate between the modified versions of training sound A and training sound B. If the birds are still capable of associating the test stimuli with the respective training stimuli, the proportion of correct responses to the test stimuli should be higher than the proportion of incorrect responses. For the Crossed-direction group, two modified versions (the "Vocoded" and the "FormantPitch" versions) were statistically similar to 0, suggesting that the birds responded to these two modified versions by chance. In contrast, the rest test stimuli significantly differed from 0, indicating that these modified versions still showed resemblance to the training stimuli from which they were derived. In the Same-direction group, two modified versions (the "No-Formant" and the "FormantPitch" versions) were statistically similar to 0, with the remaining test versions showing a significant difference from 0, favouring correct responses (Table 4 & Fig. 4). In conclusion, both groups of birds maintained the ability to discriminate the "FormantMiddle" and "PitchMiddle" versions of the training stimuli, but their discrimination diminished for the "FormantPitch" version. Interestingly, the Crossed-direction group still differentiated the "NoFormant" version but lost

discrimination for the "Vocoded" version, whereas the Same-direction group exhibited the opposite pattern for these two versions. These results suggest that different training conditions had some effect on birds' attention to pitch and formant in the training sequences.

Figure 4. Visualisation of logRatios = log (Correct/Incorrect). The Log (Cor/Inco) for two training groups responding to the various test stimuli. A + indicates that the Log (Cor/Inco) of a Test treatment is significantly above 0. A ns indicates that the Log (Cor/Inco) of a Test treatment is overlapping with 0. Box plots show median, 1st and 3rd quartile, and whiskers the 1.5 interquartile range. Horizontal dashed lines show the discrimination boundaries in which the proportion of correct responses is equal to the proportion of incorrect responses. The calculation of logRatios was based on the counts of "correct response" and "incorrect response" from the same data set that was also used for Fig.3.

Table 4 Estimates and 95% confidence intervals for the correct identification of test stimuli



Note: Lower CL and Upper CL represent the lower and upper 95% confidence limits (CL) of the confidence interval. If zero is part of that confidence interval, the treatment combination Training_Group and Stimuli is not significantly different from 0. If both confidence levels are positive, then there is a bias toward correct responses. If they are both negative, then they are biased toward incorrect responses. Bold indicates significance.

DISCUSSION

The present study examined the perceptual interaction between pitch and formant cues in zebra finches' auditory discrimination, employing a Go-left/Go-right paradigm. Through a systematic manipulation of the pitch and formant contours of tone sequences, our study investigated the relative contributions of these attributes to the recognition of sound sequences, as well as examining the presence of any perceptual interaction between them. Below, we discuss key findings concerning the effects of pitch and formant contour directions on birds' discrimination learning, the influence of training conditions on their discrimination of modified stimuli, and the interplay between pitch and formant contours in zebra finches' auditory discrimination.

Both training groups demonstrated similar learning speeds, suggesting that the perceptual interactions (if there were any) between pitch and formant contours, whether going in the same or opposite direction, did not affect the difficulty of acquiring discrimination. However, when analyzing Correct and Response rates for various test stimuli, distinctions between the training groups emerged. It then becomes evident that the relative importance of pitch and formant contours shows some effect of training conditions.

Among the modified versions tested, the "PitchMiddle" and "FormantMiddle" versions were consistently well-recognized by the birds, indicating that both formants and pitch, respectively, were attended for tone sequence recognition. In contrast, the responses to the "FormantPitch" version, despite its relatively high response rates, were at chance level. This suggests that the conflicting information presented by pitch and formant contours in the "FormantPitch" version led the birds to perceive it as ambiguous. On the other hand, among the five modified versions, the "Vocoded" version proved to be the most easily detected as differing from the training sounds, as the birds respond least to this version. The manipulation involving noise-vocoding not only disrupts the harmonic attributes (hence also removes pitch information) of the tones but also renders the spectrum of the stimulus "noise-like". Such "noise-like spectrum" alterations may likely capture the birds' attention, making the "Vocoded" stimuli distinguishable from the training stimuli when perceived by the birds. In addition, both groups showed a pronounced distinction in responding between training stimuli and their modified versions. This suggests that zebra finches excel in detecting spectral structures in either pitch or formant contours no matter the manipulation was on pitch or formant cues.

Remarkably, the "No-Formant" version had a higher impact on stimulus recognition within the same-direction group, which resulted in the same-direction group losing discrimination of this version, while remaining distinguishable in the crossed-direction group. This suggests that the crossed-direction group tends to focus slightly more on pitch contour than on formant contour for stimulus identification, although such a bias is not visible in the responses of both groups to the "FormantMiddle" and "PitchMiddle" versions. Zebra finches trained with crossed-direction sequences exhibited elevated correct rates for the "No-Formant" version, suggesting that the crossed-direction group paid relatively less attention to changes in formants compared to the same-direction group, although here also no difference is present between the groups in their responses towards the "FormantMiddle" and "PitchMiddle" versions. Such a difference would be expected if the groups really differed in their relative attention for pitch and formants.

A key question addressed in our study pertains to whether zebra finches when presented with stimuli containing two salient parameters prioritize one parameter for discrimination while disregarding the other, or whether they consider both parameters in their discrimination process. Our findings indicate that the latter strategy is adopted by these birds. Moreover, one might anticipate that differentiating between stimuli A and B could be more straightforward for the birds when both pitch and formant contours are oriented in the same direction, and as a result rely more on one parameter rather than taking both into account. However, based on our results, there is no evidence that this alignment had any impact on the birds' discrimination learning, apart from the small bias observed in the response to the "No-Formant" version in the crossed-direction group. These results together demonstrate that with these stimuli both parameters play comparable roles in zebra finches' tone sequence recognition. The variation in response to different modified versions can be explained by "additive effects" rather than more complex interactions between attribute contours or training conditions: both attributes are assessed and used to distinguish stimuli, and if one attribute remains constant throughout the stimulus sequence, the other suffices to keep discriminating the sequences.

With respect to the significance of our findings it is worthwhile to compare our study with the research of Bregman *et al*. (2016) and Burgering *et al*. (2019) on the role of various acoustic features in songbird's auditory perception. Bregman *et al*. (2016) investigated the ability of starlings to discriminate a sequence of synthetic harmonic tones. This investigation revealed that starlings were attending to spectral shape (i.e., spectral envelope) over absolute pitch in

tonal sound discrimination tasks – they still showed recognition of vocoded tone sequences. This is in contrast with our finding that zebra finches could not identify vocoded versions of the tone sequences, which are maintaining the spectral shape, but lack pitch information. Our finding also differs from an earlier study by Burgering *et al*. (2019) on zebra finches' perception of vowel-like sounds, examining the roles of pitch and spectral envelope. This research revealed that in this case the zebra finches were responding to vocoded stimuli, hence attending to the spectral shape. The discrepancy between the results of our study and those of Bregman *et al*. (2016) and Burgering *et al*. (2019) may have different causes. It could be that starlings and zebra finches are sensitive to different vocal parameters. In relation to the study by Burgering *et al*. (2019), the finding that the zebra finches attend to other parameters (i.e., pitch and spectral envelope) than in the current study indicates that what gets attention may depend on the nature of the stimulus. It is still unclear what causes this disparity in zebra finches' responses between shifts in artificial tone sequences and in vowel-like sound elements. One possible reason could be that the nature of the training stimuli influences the birds' attention to specific acoustic features during discriminating tasks, as well as their future generalization of learning to novel stimuli. There might also be a methodological factor affecting the difference between the earlier studies and ours. In the current study, we utilize true "tone sequences" with silent gaps to partition the tone units. This stimuli design serves as a valuable complement to the starling experiment conducted by Bregman *et al*. (2016), which used similar "tone sequences" but without silent gaps. Previous studies, including those conducted by Bregman *et al*. (2016) and Burgering *et al*. (2019), have primarily focused on local features, whether in isolated units or in tone sequences lacking silent gaps. The absence of silent gaps raises the question whether birds perceive the entire "tone sequence" as a single acoustic object or as a sequence of tonal units processed sequentially. Our way of arranging stimuli prompted zebra finches to engage with the comprehensive contour of the overall tone sequence. This methodology is distinct from previous investigations that concentrated on localized features, such as the pitch or spectral attributes of individual acoustic units.

It is interesting to compare our results on zebra finches' attention to pitch and formant contours with a study on how humans attend to pitch and timbre (McPherson & McDermott, 2023). McPherson & McDermott (2023) demonstrated that judgments of harmonic sounds in humans relied on f0 representations, while relative pitch judgments were influenced by timbral differences, leading to biases in discrimination tasks. Comparatively, our findings highlight zebra finches' ability to integrate pitch and formant contours for discrimination. The fact that

zebra finches are capable of attending to both attributes in sound recognition is noteworthy, as it differs from the human tendency to prioritize one attribute (i.e., pitch) over the other (i.e., timbre) in perceiving tonal sounds. Moreover, the stimuli used in our study, consisting of sequences of units, required broader attention to contour attributes rather than local features like the pitch or timbre of individual units. However, it's premature to determine the similarities or differences between our current study and these human studies. This is due to the limited number of studies on the perceptual interplay between attributes in human auditory perception, aside from a few studies (e.g., Shinn-Cunningham *et al*., 2007; McPherson & McDermott, 2023). Conducting a similar experiment as presented here with human subjects would enable a direct comparison (Ning *et al*., in prep).

CONCLUSION AND OUTLOOK

Our study investigated the interplay between pitch and formant attributes in zebra finches. The findings demonstrate that when tone sequences exhibit variations in both pitch and formant across a series of tones, zebra finches attend to both pitch and formant contours when distinguishing the series. This observation holds true regardless of whether the changes in pitch and formant across the tones occur in the same or opposite directions, indicating a limited impact of the direction of these changes on tone sequence discrimination. Furthermore, our study, in combination with earlier ones (e.g., Ning *et al*., 2023) reaffirms the remarkable perceptual flexibility exhibited by zebra finches. This enhanced understanding of avian auditory perception prompts consideration of how the attention to acoustic attributes extends across species. It also indicates the relevance of future cross-species experiments to elucidate the differences between humans and songbirds in attending to pitch and formant cues. This line of inquiry holds promise for uncovering the underlying mechanisms of auditory perception and contributes to the broader field of cognitive research.

References

Albouy P., Mehr S.A., Hoyer R.S., Ginzburg J., Zatorre R.J. (2023). Spectro-temporal acoustical markers differentiate speech from song across cultures. *BioRxiv*, 2023.01.29.526133. Preprint. https://doi.org/10.1101/2023.01.29.526133

Bartoń K. (2020). MuMIn: Multi-Model Inference. R package version 1.43.17. https://CRAN.R-project.org/package=MuMIn

Bates D., Maechler M., Bolker B., Walker S. (2015). Fitting linear mixed-effects models using lme4. *J Stat Softw*., 67(1): 1-48. https://doi.org/10.18637/jss.v067.i01

Benjamini Y., Hochberg Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B*., 57(1): 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bregman M.R., Patel A.D., Gentner T.Q. (2012). Stimulus-dependent flexibility in non-human auditory pitch processing. *Cognition*, 122(1): 51–60. https://doi.org/10.1016/j.cognition.2011.08.008

Bregman M.R., Patel A.D., Gentner T.Q. (2016). Songbirds use spectral envelope, not pitch, for sound pattern recognition. *Proc Natl Acad Sci*., 113(6): 1666–1671. https://doi.org/10.1073/pnas.1515380113

Brosch M., Selezneva E., Bucks C., Scheich H. (2004). Macaque monkeys discriminate pitch relationships. *Cognition*, 91(3): 259–272. http://dx.doi.org/10.1016/j.cognition.2003.09.005

Burgering M.A., Ten Cate C., Vroomen J. (2018). Mechanisms underlying speech sound discrimination and categorization in humans and zebra finches. *Anim Cogn*., 21(2): 285-299. https://doi.org/10.1007/s10071-018-1165-3

Burgering M.A., Vroomen J., Ten Cate C. (2019). Zebra finches (*Taeniopygia guttata*) can categorize vowel-like sounds on both the fundamental frequency ("pitch") and spectral envelope. *J Comp Psychol*., 133(1): 106-117. https://doi.org/10.1037/com0000143

Choi W. (2021). Musicianship Influences Language Effect on Musical Pitch Perception. *Front Psychol*., 12: 712753. https://doi.org/10.3389/fpsyg.2021.712753

Crespo-Bojorque P., Celma-Miralles A., Toro J.M. (2022). Detecting surface changes in a familiar tune: exploring pitch, tempo and timbre. *Anim Cogn*., 25(4): 951-960. https://doi.org/10.1007/s10071-022-01604-w

Cynx J., Shapiro M. (1986). Perception of missing fundamental by a species of songbird (Sturnus vulgaris). *J Comp Psychol*., 100(4): 356–360. https://doi.org/10.1037/0735-7036.100.4.356

Dowling W.J., Harwood D.L. (1986). Music cognition Cognitive processing. Orlando, FL: Academic Press.

Fitch W.T., Fritz J.B. (2006). Rhesus macaques spontaneously perceive formants in conspecific vocalizations. *J Acoust Soc Am*., 120(4): 2132-2141. https://doi.org/10.1121/1.2258499

Friedrich A., Zentall T., Weisman R. (2007). Absolute pitch: frequency-range discriminations in pigeons (*Columba livia*): comparisons with zebra finches (*Taeniopygia guttata*) and humans (*Homo sapiens*). *J Comp Psychol*., 121(1): 95–105. https://doi.org/10.1037/0735-7036.121.1.95

Geissler D.B., Ehret G. (2002). Time-critical integration of formants for perception of communication calls in mice. *Proc Natl Acad Sci.*, 99(13): 9021-9025. https://doi.org/10.1073/pnas.122606499

Henry K.S., Amburgey K.N., Abrams K.S., Idrobo F., Carney L.H. (2017). Formant-frequency discrimination of synthesized vowels in budgerigars (Melopsittacus undulatus) and humans. *J Acoust Soc Am.*, 142(4): 2073. https://doi.org/10.1121/1.5006912

Hoeschele M., Merchant H., Kikuchi Y., Hattori Y., ten Cate C. (2015). Searching for the origins of musicality across species. *Philos Trans R Soc Lond B Biol Sci.*, 370(1664): 20140094. https://doi.org/10.1098/rstb.2014.0094

Hoeschele M. (2017). Animal Pitch Perception: Melodies and Harmonies. *Comp Cogn Behav Rev.*, 12: 5-18. https://doi.org/10.3819/CCBR.2017.120002

Honing H., ten Cate C., Peretz I., Trehub S.E. (2015). Without it no music: cognition, biology and evolution of musicality. *Philos Trans R Soc Lond B Biol Sci.*, 370(1664): 20140088. https://doi.org/10.1098/rstb.2014.0088

Hulse S.H., Cynx J. (1985). Relative pitch perception is constrained by absolute pitch in songbirds (*Mimus*, *Molothrus*, and *Sturnus*). *J Comp Psychol.*, 99(2): 176–196. http://dx.doi.org/10.1037/0735-7036.99.2.176

Hulse S.H., Cynx J., Humpal J. (1984). Absolute and relative pitch discrimination in serial pitch perception by birds. *J Exp Psychol Gen.*, 113(1): 38–54. https://doi.org/10.1037/0096-3445.113.1.38

Hurly T.A., Ratcliffe L., Weisman R. (1990). Relative pitch recognition in white-throated sparrows (Zonotrichia albicollis). *Anim Behav.*, 40(1): 176–181. http://dx.doi.org/10.1016/S0003-3472(05)80677-3

Izumi A. (2001). Relative pitch perception in Japanese monkeys (*Macaca fuscata*). *J Comp Psychol.*, 115(2): 127–131. http://dx.doi.org/10.1037/0735-7036.115.2.127

Kluender K.R., Lotto A.J., Holt L.L., Bloedel S.L. (1998). Role of experience for language-specific functional mappings of vowel sounds. *J Acoust Soc Am.*, 104(6): 3568–3582. https://doi.org/10.1121/1.423939

Kriengwatana B., Escudero P., Kerkhoven A.H., ten Cate C. (2015). A general auditory bias for handling speaker variability in speech? Evidence in humans and songbirds. *Front Psychol.*, 6: 1243. https://doi.org/doi:10.3389/fpsyg.2015.01243

Kubovy M., van den Berg M. (2008). The whole is equal to the sum of its parts: a probabilistic model of grouping by proximity and similarity in regular patterns. *Psychol Rev.*, 115(1): 131-154. https://doi.org/10.1037/0033-295X.115.1.131

Lenth R.V. (2016). Least-Squares Means: The R Package lsmeans. *J Stat Softw.*, 69(1): 1-33. https://doi.org/10.18637/jss.v069.i01

Luna D., Montoro P.R. (2011). Interactions between intrinsic principles of similarity and proximity and extrinsic principle of common region in visual perception. *Perception*, 40(12): 1467-1477. https://doi.org/10.1068/p7086

Luna D., Villalba-García C., Montoro P.R., Hinojosa J.A. (2016). Dominance dynamics of competition between intrinsic and extrinsic grouping cues. *Acta Psychol (Amst).*, 170: 146-154. https://doi.org/10.1016/j.actpsy.2016.07.001

MacDougall-Shackleton S.A., Hulse S.H. (1996). Concurrent absolute and relative pitch processing by European starlings (*Sturnus vulgaris*). *J Comp Psychol.*, 110(2): 139–146. https://doi.org/10.1037/0735-7036.110.2.139

McDermott J.H., Lehr A.J., Oxenham A.J. (2008). Is relative pitch specific to pitch? *Psychol Sci*., 19(12): 1263–1271. https://doi.org/10.1111/j.1467-9280.2008.0223

McPherson M.J., McDermott J.H. (2018). Diversity in pitch perception revealed by task dependence. *Nat Hum Behav*., 2(1): 52-66. https://doi.org/10.1038/s41562-017-0261-8

Melchor J., Vergara J., Figueroa T., Morán I., Lemus L. (2021). Formant-Based Recognition of Words and Other Naturalistic Sounds in Rhesus Monkeys. *Front Neurosci*., 15: 728686. https://doi.org/10.3389/fnins.2021.728686

Micheyl C., Oxenham A.J. (2010). Pitch, harmonicity and concurrent sound segregation: psychoacoustical and neurophysiological findings. *Hear Res*., 266(1-2): 36-51. https://doi.org/10.1016/j.heares.2009.09.012

Montoro P.R., Villalba-García C., Luna D., Hinojosa J.A. (2017). Common region wins the competition between extrinsic grouping cues: Evidence from a task without explicit attention to grouping. *Psychon Bull Rev*., 24(6): 1856-1861. https://doi.org/10.3758/s13423-017-1254-3

Ning Z.Y., Honing H., ten Cate C. (2023). Zebra finches (*Taeniopygia guttata*) demonstrate cognitive flexibility in using phonology and sequence of syllables in auditory discrimination. *Anim Cogn*., 26: 1161–1175 https://doi.org/10.1007/s10071-023-01763-4

Page S.C., Hulse S.H., Cynx J. (1989). Relative pitch perception in the European starling (Sturnus vulgaris): further evidence for an elusive phenomenon. *J Exp Psychol Anim Behav Process*., 15(2): 137–146. https://doi.org/10.1037/0097-7403.15.2.137

Palmer S.E., Beck D.M. (2007). The repetition discrimination task: an objective method for studying perceptual grouping. *Percept Psychophys*., 69(1): 68-78. https://doi.org/10.3758/bf03194454

Patel A.D. (2003). Language, music, syntax and the brain. *Nat Neurosci*., 6(7): 674-681. https://doi.org/10.1038/nn1082

Patel A.D. (2008). Music, Language, and the Brain. New York, NY: Oxford University Press.

Patel A.D. (2011). Why would Musical Training Benefit the Neural Encoding of Speech? The OPERA Hypothesis. *Front Psychol*., 2: 142. https://doi.org/10.3389/fpsyg.2011.00142

Ohms V.R., Escudero P., Lammers K., ten Cate C. (2012). Zebra finches and Dutch adults exhibit the same cue weighting bias in vowel perception. *Anim Cogn*., 15(2): 155-161. https://doi.org/10.1007/s10071-011-0441-2

Okanoya K., Dooling R.J. (1987). Hearing in passerine and psittacine birds: a comparative study of absolute and masked auditory thresholds. *J Comp Psychol*., 101(1): 7-15. https://doi.org/10.1037/0735-7036.101.1.7

Rashal E., Yeshurun Y., Kimchi R. (2017). The time course of the competition between grouping organizations. *J Exp Psychol Hum Percept Perform*., 43(3): 608-618. https://doi.org/10.1037/xhp0000334

Sadakata M., Weidema J.L., Roncaglia-Denissen M.P.M., Honing H. (2020). Parallel pitch processing in speech and melody: A study of the interference of musical melody on lexical pitch perception in speakers of Mandarin. *PLoS ONE*, 15(3): e0229109. https://doi.org/10.1371/journal.pone.0229109

Schmidt F., Schmidt T. (2013). Grouping principles in direct competition. Vision Res., 88: 9-21. https://doi.org/10.1016/j.visres.2013.06.002

Shinn-Cunningham B.G., Lee A.K., Oxenham A.J. (2007). A sound element gets lost in perceptual competition. *Proc Natl Acad Sci USA*., 104(29): 12223–12227. https://doi.org/10.1073/pnas.0704641104

Standards Secretariat, Acoustical Society of America (1994). ANSI S1.1-1994 (R2004) American National Standard Acoustical Terminology, (12.41) Acoustical Society of America, Melville, NY.

ten Cate C., Honing H. (2022). Precursors of music and language in animals. *PsyArXiv*, Preprint. https://doi.org/10.31234/osf.io/4zxtr

Uno H., Maekawa M., Kaneko H. (1997). Strategies for harmonic structure discrimination by zebra finches. *Behav Brain Res*., 89(1-2): 225-228. https://doi.org/10.1016/s0166-4328(97)00064-8

van Buuren S., Groothuis-Oudshoorn K. (2011). Mice: Multivariate Imputation by Chained Equations in R. *J Stat Softw*., 45(3): 1-67. https://doi.org/10.18637/jss.v045.i03

Villalba-García C., Santaniello G., Luna D., Montoro P.R., Hinojosa J.A. (2018). Temporal brain dynamics of the competition between proximity and shape similarity grouping cues in vision. *Neuropsychologia*., 121: 88-97. https://doi.org/10.1016/j.neuropsychologia.2018.10.022

Villalba-García C., Jimenez M., Luna D., Hinojosa J.A., Montoro P.R. (2021). Competition between perceptual grouping cues in an indirect objective task. *Q J Exp Psychol (Hove)*., 74(10): 1724-1736. https://doi.org/10.1177/17470218211010486

Walker K.M.M., Schnupp J.W.H., Hart-Schnupp S.M.B., King A.J., Bizley J.K. (2009). Pitch discrimination by ferrets for simple and complex sounds. *J Acoust Soc Am*., 126(3): 1321-1335. http://dx.doi.org/10.1121/1.3179676

Walker K.M., Gonzalez R., Kang J.Z., McDermott J.H., King A.J. (2019). Acrossed-species differences in pitch perception are consistent with differences in cochlear filtering. *Elife*, 8: e41626. https://doi.org/10.7554/eLife.41626

Weary D.M., Weisman R.G. (1991). Operant discrimination of frequency ratio in the black-capped chickadee (*Parus atricapillus*). *J Comp Psychol*., 105(3): 253–259. http://dx.doi.org/10.1037/0735-7036.105.3.253

Weisman N., Njegovan M., Ito S. (1994). Frequency ratio discrimination by zebra finches (*Taeniopygia guttata*) and humans (*Homo sapiens*). *J Comp Psychol*., 108(4): 363–372. https://doi.org/10.1037/0735-7036.108.4.363

Weisman R.G., Njegovan M.G., Sturdy C.B., Phillmore L., Coyle J., Mewhort D. (1998). Frequency-range discriminations: special and general abilities in zebra finches (*Taeniopygia guttata*) and humans (*Homo sapiens*). *J Comp Psychol*., 112(3): 244–258. https://doi.org/10.1037/0735-7036.112.3.244

Weisman R.G., Njegovan M.G., Williams M.T., Cohen J.S., Sturdy C.B. (2004). A behavior analysis of absolute pitch: sex, experience, and species. *Behav Process*., 66(3): 289–307. https://doi.org/10.1016/j.beproc.2004.03.010