PROCEEDINGS OF THE ROYAL SOCIETY B

Proc. R. Soc. B (2010) **277**, 1003–1009
doi:10.1098/rspb.2009.1788
Published online 2 December 2009

Zebra finches exhibit speaker-independent phonetic perception of human speech

Verena R. Ohms[1,*], Arike Gill[1], Caroline A. A. Van Heijningen[1], Gabriel J. L. Beckers[2] and Carel ten Cate[1,3]

[1]*Behavioural Biology, Institute of Biology Leiden (IBL), Sylvius Laboratory, PO Box 9505, 2300 RA Leiden, The Netherlands*
[2]*Behavioural Neurobiology, Max Planck Institute for Ornithology, Eberhard-Gwinner-Strasse, 82319 Seewiesen, Germany*
[3]*Leiden Institute for Brain and Cognition, PO Box 9600, 2300 RC Leiden, The Netherlands*

Humans readily distinguish spoken words that closely resemble each other in acoustic structure, irrespective of audible differences between individual voices or sex of the speakers. There is an ongoing debate about whether the ability to form phonetic categories that underlie such distinctions indicates the presence of uniquely evolved, speech-linked perceptual abilities, or is based on more general ones shared with other species. We demonstrate that zebra finches (*Taeniopygia guttata*) can discriminate and categorize monosyllabic words that differ in their vowel and transfer this categorization to the same words spoken by novel speakers independent of the sex of the voices. Our analysis indicates that the birds, like humans, use intrinsic and extrinsic speaker normalization to make the categorization. This finding shows that there is no need to invoke special mechanisms, evolved together with language, to explain this feature of speech perception.

Keywords: human speech; language evolution; zebra finches; speech perception; formants

1. INTRODUCTION

Human speech is a hierarchically organized coding system. A finite number of meaningless sounds, called phonemes, which are classes of speech sounds that are identified as the same sound by native speakers, are combined into an infinite set of larger units: morphemes or words. These larger units carry meaning and therefore allow linguistic communication (Yule 2006). An important role in the coding process is played by formants—vocal tract resonances that can be altered rapidly by changing the geometrical properties of the vocal tract using articulators such as tongue, lips and soft palate (Titze 2000). Changing the formant pattern of an articulation results in a different vowel produced (figure 1).

It has been argued in the past that many characteristics of speech are uniquely human (e.g. Lieberman 1975, 1984). Therefore it was a revolutionary finding when Kuhl & Miller (1975, 1978) who tested chinchillas on their ability to discriminate between /d/ and /t/ consonant–vowel syllables found that these animals have the same phonetic boundaries as humans, thereby challenged the view that the mechanisms underlying speech perception are uniquely human. A few years later the same phonetic boundary effect has been shown in macaques (Kuhl & Padden 1982). Nevertheless, there is still an ongoing debate about which parameters of human speech production and perception are unique to humans, with the implication that they evolved together with speech or language, and which are shared with other species (Liberman & Mattingly 1985; Hauser *et al.* 2002; Trout 2003; Diehl *et al.* 2004; Pinker & Jackendoff 2005).

One of the most important phenomena in human speech concerns our ability to recognize words regardless of individual variation across speakers. Although human voices differ in acoustic parameters such as fundamental frequency and spectral distribution, we are able to distinguish closely similar words by using the relative formant frequencies in dependence of the fundamental frequency of an utterance. This feature enables the intelligibility of speech (Nearey 1989; Fitch 2000; Assmann & Nearey 2008). But does this mean that the human ability to perceive and normalize formant frequencies in order to develop an abstract formant percept has evolved together with speech and language? Or has the evolution of language exploited a pre-existing perceptual property that allowed formant normalization? An important way to test this question is by examining whether this feature is present in other animals. If so, this suggests that it is not a uniquely evolved faculty.

Here we examined whether zebra finches trained to distinguish two words differing in one vowel only and produced by several same-sex speakers, generalize the distinction to a novel set of speakers of (i) the same sex and (ii) the opposite sex. We chose natural human voices instead of artificial stimuli to confront the animals with a situation humans have to deal with every day when vocally communicating: extracting the relevant sound features from irrelevant ones while listening and building up a percept that allows categorization of these words when originating from novel voices.

2. MATERIAL AND METHODS

(a) *Subjects*

We used three male and five female zebra finches (*Taeniopygia guttata*, aged six months to 2 years) from the Leiden

* Author for correspondence (v.r.ohms@biology.leidenuniv.nl).

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspb.2009.1788 or via http://rspb.royalsocietypublishing.org.

Received 2 October 2009
Accepted 9 November 2009

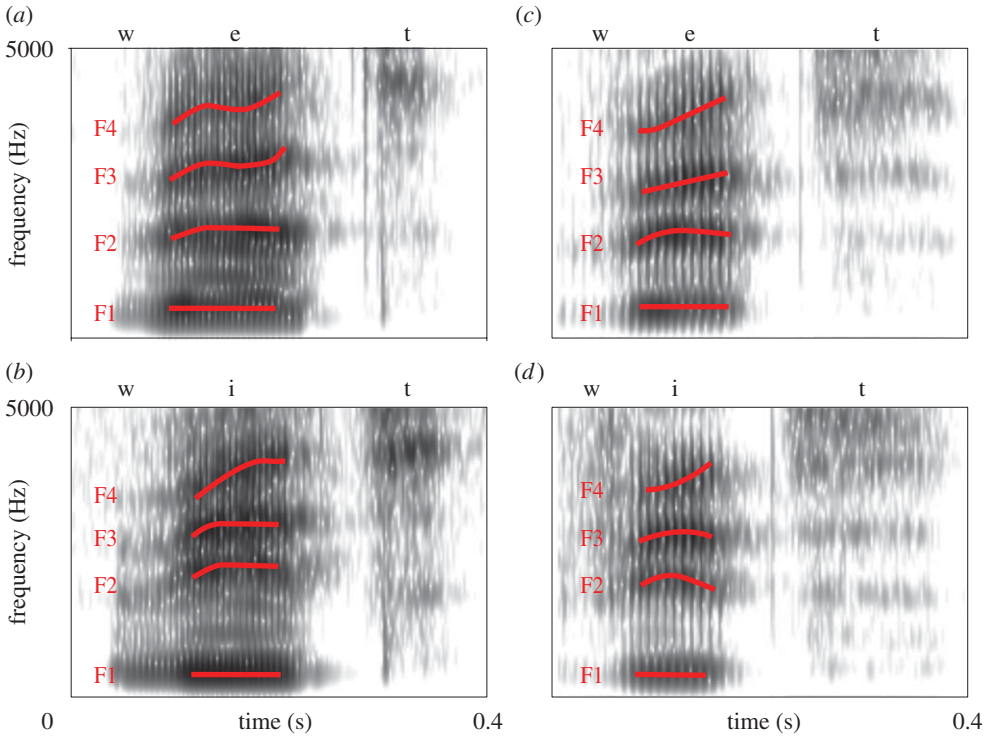

Figure 1. Spectrograms of human voices. (*a*) Female voice saying *wet*; (*b*) female voice saying *wit*; (*c*) male voice saying *wet*; (*d*) male voice saying *wit*. Red lines indicate the formant frequencies. Note the difference in the distance between the first and the second formant frequencies. In (*a*) this distance is smaller than in (*b*) and the same applies for (*c*) and (*d*). F1, 1st formant; F2, 2nd formant; F3, 3rd formant; F4, 4th formant; s, seconds; Hz, Hertz.

University breeding colony. Prior to the experiment, birds were housed in groups of two or three animals and were kept on a 13.5 L : 10.5 D schedule. Food, grit and water were provided ad libitum. None of the birds had previous experience with psychophysical experiments. At the beginning of the study every animal was weighed to allow monitoring of the nutritional state. During the experiment the amount of food eaten by the birds was checked daily. If an animal ate less than necessary it was provided with additional food. In this case the bird was also weighed to ensure that it did not lose more than 20 per cent of its initial body weight. All animal procedures were approved by the animal experimentation committee of Leiden University (DEC number 08054).

(b) *Stimuli*

We obtained naturally spoken Dutch words from second year students at Leiden University. A total of 10 females and 11 male native speakers of Dutch were recorded in the phonetics laboratory of the Faculty of Humanities, Leiden University using a Sennheiser RF Condenser Microphone MKH416T and Adobe Audition 1.5 software with 44.1 kilosamples s^{-1}, at a 16 bit resolution. Every speaker was asked to read a list of Dutch words in which the stimuli *wit* (wIt) and *wet* (wɛt) were embedded to prevent list-final intonation effects. The recordings were processed afterwards using the software Praat (v. 4.6.09) freely available at www.praat.org (Boersma 2001) by cutting out the words *wit* and *wet* and saving both as separate wave files for each voice. To prevent intensity differences between stimuli from playing a role in the discrimination process, the average amplitude of all female and male voices, respectively, was normalized by using the root mean square of the average acoustic energy and equalizing it. During the experiment all stimuli were played back at approximately 70 dB SPL(A).

(c) *Apparatus*

The experiment was conducted in a Skinner box described earlier (Verzijden *et al.* 2007), which was placed in a sound attenuated chamber. Sounds were played through a Vifa MG10SD-09-08 broadband loudspeaker at approximately 70 dB SPL(A) attached 1 m above the Skinner box. A fluorescent lamp (Lumilux De Luxe Daylight, 1150 lm, L 18 W/965, Osram, Capelle aan den IJssel, The Netherlands) served as the light source and was placed on top of the Skinner box. It was switched on automatically every day from 07.00 to 20.30 h, whereby the light was gradually increasing and decreasing in a 15 min time window at the beginning and the end of the light cycle, respectively.

(d) *Discrimination learning*

To train the birds to discriminate between acoustic stimuli we used a 'Go/NoGo' operant conditioning procedure (Verzijden *et al.* 2007). The positive ('Go') stimulus (S^+) was an average zebra finch song, whereas the negative ('NoGo') stimulus (S^-) was a pure tone of 2 kHz constructed in Praat (Boersma 2001). During the training the birds had to learn that responding to S^+ would lead to a 10 s food reward with access to a commercial seed mix, whereas responding to S^- would cause a 15 s punishment interval with the lights in the experimental chamber going out (electronic supplementary material, figure S1).

(e) *Experiment*

The actual experiment consisted of four successive phases. As soon as the birds reached the discrimination criterion ($d' = 1.34$) which we defined as a high response rate to the

Go stimulus (75% or more) and a low response rate to the NoGo stimulus (25% or less) over three consecutive days, they were transferred to the next stage. During the first stage of the experiment every bird had to learn to discriminate the words *wit* and *wet* of a single person (stage 1), whereby every bird started with a different voice. Four groups with two birds per group were formed (electronic supplementary material, figure S2). Two groups started with female voices and the other two groups with male voices. One of the groups that began the experiment with a female voice received *wit* as positive and *wet* as negative stimulus and vice versa for the other group. The birds that started with the male voices were treated accordingly. After the birds had reached the discrimination criterion they were switched to the next stage (stage 2) in which four new minimal pairs of the same sex as the first voice were added. After reaching the discrimination criterion birds were transferred to stage 3 in which the five voices used in stage 2 were replaced by five new voices of speakers of the same sex. In the final stage of the experiment (stage 4) the birds were confronted with five new voices of the opposite sex. The experiment was finished after the birds again fulfilled the discrimination criterion. To prevent pseudoreplication voices were randomly balanced over the four groups.

(f) *Performance evaluation*

To assess performance discrimination between *wit* and *wet*, we calculated the d' and 95% confidence interval (CI) following the procedure used and described by others (Macmillan & Creelman 2005; Gentner *et al.* 2006) for every bird for the first 100, 200 and 300 trials directly after transition between the different phases. This is a sensitivity measure that subtracts the z score of the false-alarm rate (F), which is defined as the proportion of responses to a NoGo stimulus divided by the total number of NoGo-stimulus presentations, from the z score of the hit rate (H), which is the proportion of responses to a Go stimulus divided by the total number of Go-stimulus presentations. This measure allows the evaluation of how well two stimuli are discriminated from each other: $d' = z(\mathrm{H}) - z(\mathrm{F})$. A d' of zero indicates no discrimination, whereas a lower bound of the 95% CI above zero can be considered to indicate significant discrimination (Macmillan & Creelman 2005; Gentner *et al.* 2006). Moreover, this measurement is unaffected by a potential response bias (Macmillan & Creelman 2005).

(g) *Acoustic measurements*

In order to detect acoustic features that might have enabled distinction between *wit* and *wet* we measured word and vowel duration as well as fundamental frequency and the mean first (F1) and second (F2) formant frequencies of both words obtained by the different speakers using Praat software (Boersma 2001). We ran two-tailed Wilcoxon-signed ranks tests separately for male and female voices to detect significant differences of the acoustic characteristics between *wit* and *wet*.

3. RESULTS

In the first phase of the experiment all birds learned to discriminate reliably between the two words *wit* and *wet* and fulfilled the discrimination criterion after an average of 41 blocks (40.72 ± 3.41 s.e.m.) with 100 trials per block.

However, this outcome does not imply generalized categorical discrimination as the birds might have learned the individual features of the training stimuli. In order to show that the birds had developed a generalized percept, their performance should be independent of individual voices. In the next phase we therefore added four additional minimal pairs recorded by same-sex speakers to the first stimulus pair but maintained the same learning criterion. The mean d' (which is a measure of how well two stimuli are discriminated from each other) of the first 100 trial block after this transition was 0.77 ± 0.30 ($d' \pm$ s.e.m.), which is clearly above chance level ($d'=0$). After transition of stimulus sets (figure 2*b*), five out of eight birds immediately performed above chance level and all birds achieved a significant performance within the first three blocks after transition (mean $d' = 0.94 \pm 0.17$ s.e.m. with the lower bound of the 95% CI ranging from 0.14 to 0.94).

It could be argued that these results are biased through the incorporation of an already familiar voice in the stimuli sets. Hence, in the subsequent phase we switched to five completely unknown speakers of the same sex (figure 2*c*). Again, the average d' was already highly above chance level over the first 100 trials after transition ($d' = 1.01 \pm 0.32$ s.e.m.) for six out of eight birds. Within 300 trials after transition all birds showed clear discrimination with a lower bound of the 95% CI ranging from 0.2 to 1.57. Thus, the birds seem to have formed a generalized percept.

So far all voices were of the same sex and overlapped in several features. Therefore, a more critical test is to check whether the birds are able to transfer the discrimination to the same words spoken by the opposite sex, i.e. whether the relevant acoustic features can be transferred to a context with larger differences in pitch and timbre compared to voices within the same sex. Consequently, we switched to five new voices of the opposite sex in the last phase of the experiment (figure 2*d*). This time all birds discriminated well above chance level (average $d' = 0.9 \pm 0.59$ s.e.m.) within the first block after transition, with the lower bound of the 95% CI ranging from 0.02 to 0.59.

We measured various acoustic characteristics that may have allowed discrimination (electronic supplementary material, table S1). It is possible that a consistent difference in either vowel or word duration between *wit* and *wet* enabled distinction, but neither vowel nor word duration differed regarding the male voices. There was a significant difference in vowel duration for the female voices (Wilcoxon signed ranks test: $n = 10$, $\mathrm{T}+ = 47$, $\mathrm{T}- = 8$, $p = 0.048$) with /I/ being shorter than /ɛ/, but as all birds showed a generally high selectivity irrespective of the sex of the voices it can be assumed that vowel duration was not involved in discrimination. Another cue that might have influenced discrimination is the fundamental frequency of the voices that is known to differ between vowels with /ɛ/ having a slightly lower fundamental frequency than /I/ (Peterson & Barney 1952). This observation complies with our measurements although the difference is only significant for the male voices (Wilcoxon-signed ranks test: $n = 11$, $\mathrm{T}+ = 59.5$, $\mathrm{T}- = 6.5$, $p = 0.018$). However, the disparity in fundamental frequency between voices is much larger than within voices, so that this feature alone cannot be sufficient for discrimination.

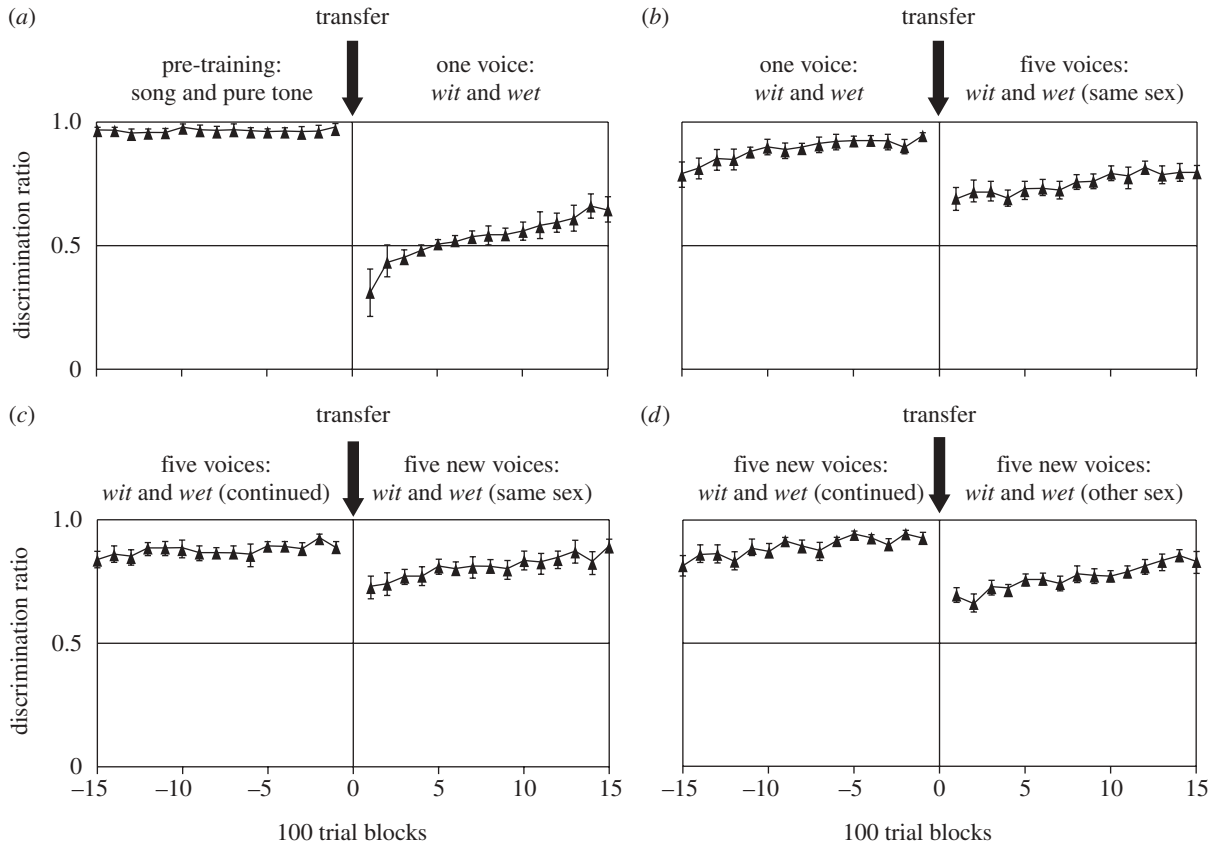

Figure 2. Transitions between discrimination stages. Displayed is the discrimination ratio of the last fifteen 100 trial blocks before and after a transition between two stages. A discrimination ratio of 1.0 reflects perfect discrimination, whereas a discrimination ratio of 0.5 indicates chance performance. The discrimination ratio is calculated as follows: (Go S^+/total S^+)/[(Go S^+/total S^+)/(Go S^-/total S^-)]. (*a*) Transition between the pre-training phase in which all birds had to discriminate a zebra finch song from a 2 kHz pure tone and the training phase in which the animals were confronted with the first minimal pair. (*b*) Shows the transition between the training phase and the subsequent experimental stage in which four additional minimal pairs of the same sex were added to the already familiar voice. (*c*) Transition between minimal pairs of now five familiar voices and five completely unknown voices of the same sex. (*d*) Transition from five voices to five new voices of the other sex. kHz, kilohertz; Go S^+, number of responses to a positive stimulus; total S^+, number of positive stimulus presentations; Go S^-, number of responses to a negative stimulus; total S^-, number of negative stimulus presentations.

On the other hand we found a highly significant difference in the formant frequencies of the first (F1) and second (F2) formant between *wit* and *wet* as expected (figure 3*a*; electronic supplementary material, table S1 and table S2). However, if the birds had only paid attention to the absolute frequency of F1 they should have treated the female *wit* as the male *wet* because of the overlap in F1 frequency (figure 3*a*; electronic supplementary material, table S2), whereas in case they based their discrimination on F2 only they should have treated the male *wit* as the female *wet* as these words overlap in F2 frequency (figure 3*a*; electronic supplementary material, table S2).

From phonetic research we know that humans do not discriminate vowels solely based on their absolute formant frequencies, but rather rely on relative formant ratios in dependence of the fundamental frequency (F0) of a speaker (Assmann & Nearey 2008). A common way of illustrating the relationship between formant frequencies and fundamental frequency as a method of intrinsic speaker normalization (Magnuson & Nusbaum 2007) is plotting the difference between F0 and F1 against the difference of F1 and F2 in 'Bark' (figure 3*b*), which can be regarded as a two-dimensional perceptual similarity measure of different sounds. Applying this method to our stimuli, results in two clearly separate vowel categories despite an extensive overlap between the sexes (figure 3*b*).

4. DISCUSSION

Previous studies on speech perception by non-human animals have suggested that the ability to discriminate human speech sounds based on their formant patterns, such as demonstrated in our study, is not unique to humans, but can be found in other taxa as well. Such studies have been carried out in several mammals, e.g. cats, chinchillas, monkeys and rats (Burdick & Miller 1975; Kuhl & Miller 1975, 1978; Kuhl 1981; Hienz & Brady 1988; Hienz *et al.* 1996; Eriksson & Villa 2006), and birds, such as budgerigars, pigeons, red-winged blackbirds and quail (Hienz *et al.* 1981; Kluender *et al.* 1987; Dooling *et al.* 1989; Dooling & Brown 1990; Dent *et al.* 1997). Most of these experiments used synthesized speech sounds lacking natural variation (Kuhl & Miller 1978; Hienz *et al.* 1981; Kuhl 1981; Hienz & Brady 1988; Dooling *et al.* 1989; Hienz *et al.* 1996; Dent *et al.* 1997; Eriksson & Villa 2006) to demonstrate that the way in which these were discriminated and categorized is equivalent to how humans do so. However, in

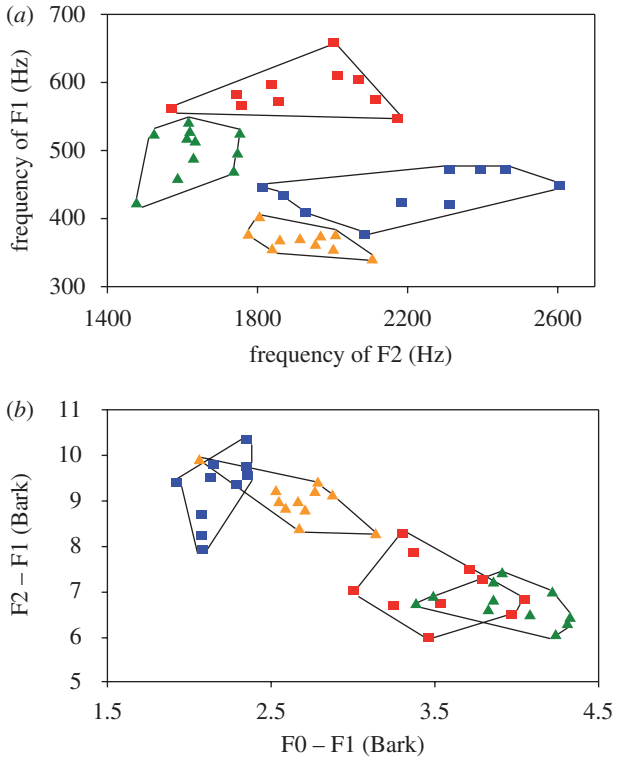

Figure 3. Vowel diagrams. (*a*) Frequencies of the first and second formants of all individual voices saying *wit* and *wet* are plotted against each other. Especially with regard towards the second formant frequencies, the male voices form denser clusters than the female voices, which show more variation. Nevertheless, the vowels /I/ and /ɛ/ can be clearly separated from each other. (*b*) The difference between the fundamental frequency and the first formant (in Bark) is plotted against the difference between the first and the second formant (in Bark) for all recordings used in the experiment. In contrast to the formant scatter plot in (*a*), this figure represents a two-dimensional perceptual concept in which male and female voices clearly overlap, whereas the two vowels /I/ and /ɛ/ are fully separated. F0, fundamental frequency; F1, 1st formant; F2, 2nd formant. Blue squares, female *wit*; red squares, female *wet*; yellow triangles, male *wit*; green triangles, male *wet*.

order to show that animals do use the same mechanisms as humans do when categorizing speech sounds it is crucial to work with natural and varying stimuli, which has been done only in a minority of studies (Burdick & Miller 1975; Kuhl & Miller 1975; Kluender *et al.* 1987; Dooling & Brown 1990). However, these studies either used isolated vowels or speech sounds from a small number of speakers. While definitely instructive none of these studies fulfilled the requirements of testing a phonemic contrast by employing different vowels embedded in a minimal pair of words. This might seem to be a minor detail when studying speech perception by animals, but yet is essential, as humans do not simply make one-bit discriminations between single phonemes (Pinker & Jackendoff 2005), but have to extract relevant information from words that closely match each other in acoustic structure in other respects. Furthermore, it is indispensable to use sufficiently different speakers (Magnuson & Nusbaum 2007).

Our experiment controlled for the above mentioned factors and our results strongly suggest that zebra finches use formants to make phonetically relevant discriminations and, similar to humans, abstract away from irrelevant variation between voices.

For humans, 'intrinsic normalization' theories (Nearey 1989) account for the phenomenon that sounds which are perceived as one phoneme can have several acoustic realizations (Liberman *et al.* 1967) by constituting that every speech sample can be categorized using a normalizing transformation. Our analyses indicate that zebra finches use a similar mechanism. However, these theories cannot explain the learning process also revealed by our data. Although the birds were able to immediately categorize *wit* and *wet* independent of speaker variability their performance dropped when confronted with new voices and then improved constantly (figure 2). Experiments with humans have also shown a clear speaker effect on speech discrimination. In a study by Magnuson & Nusbaum (2007) human subjects were presented with orthographic forms of a target vowel on a computer screen and asked to press the space bar when they heard the target vowel that they saw on the screen. Every subject had to do this task under different conditions, namely 'blocked-talker' condition, which means that all stimuli were from the same talker, and 'mixed-talker' condition, which means that the stimuli were from two different talkers. In most cases the response time was significantly higher in the 'mixed-talker' condition compared with the 'blocked-talker' condition, while the hit rate was significantly lower. The same speaker effect has been demonstrated by other studies in which the human ability to recognize whole words under varying conditions has been tested (Creelman 1957; Mullennix *et al.* 1989). In addition, human subjects also improve their discrimination performance over trial blocks (Mullennix *et al.* 1989) just as the zebra finches in the current study. This outcome indicates the presence of extrinsic normalization in humans and zebra finches, i.e. establishing a reference frame from the vowel distribution of the various speakers as a function of learned formant ranges (Nearey 1989; Magnuson & Nusbaum 2007).

So, because of the design and the results of our study our evidence holds out against arguments that in the past allowed doubts about the universality of the auditory mechanisms underlying speech perception. With respect to speaker normalization our experiment therefore provides very strong evidence that non-human animals use the same perceptual principles as humans do when discriminating speech sounds, by employing a combination of intrinsic and extrinsic speaker normalization and thereby suggesting that the underlying mechanisms originally emerged in a context independent of speech.

It is mainly because of the lowering of the larynx that humans can produce so many distinct speech sounds (Lieberman *et al.* 1969). However, another effect of a lowered larynx is to increase the length of the vocal tract that causes a decrease of formant frequencies. This in turn can be used to exaggerate size, and playback experiments in red deer which possess a lowered larynx too, have shown that stags respond more to roars with lower formant frequencies compared to roars with higher formant frequencies (Reby *et al.* 2005). In humans, formant frequencies are used to correctly estimate age (Collins 2000) and they strongly influence the

perceived height of a speaker (Smith & Patterson 2005) and hence can serve as indexical cues next to their function of coding linguistic information. Rhesus monkeys use formants in species-specific vocalizations as indexical cues as well (Ghazanfar *et al.* 2007) and although not many studies have investigated similar phenomena in bird vocalizations it has been shown that whooping cranes, for example, can perceive changes in formant frequencies in their own species calls and exhibit a different response pattern to calls with higher formants compared with lower formants (Fitch & Kelley 2000). These results led to the speculation that formant perception originally emerged in a wide range of species to assess information about the physical characteristics of conspecifics, and that human speech has exploited the already existing sensitivity for formant perception (Fitch 2000; Ghazanfar *et al.* 2007).

It can, of course, not be ruled out completely that unique perceptual abilities to facilitate speech perception did evolve in humans, or that the observed abilities evolved separately in birds and humans. In the latter case, this would indicate a remarkable convergence. However, our results, in combination with earlier findings, also support the hypothesis that the evolution of the variety of speech sounds in humans might have been shaped by pre-existing perceptual abilities, rather than being the result of coevolution between the mechanisms underlying the production and perception of speech sounds.

All animal procedures were approved by the animal experimentation committee of Leiden University (DEC number 08054).

We thank Vincent J. Van Heuven for advice considering the recording of the stimuli, for permission to use the phonetics laboratory and him, Katharina Riebel, Hans Slabbekoorn and Willem Zuidema as well as two anonymous referees for useful comments on the manuscript. Funding was provided by the Netherlands Organization for Scientific Research (NWO) (grant number 815.02.011).

REFERENCES

Assmann, P. F. & Nearey, T. M. 2008 Identification of frequency-shifted vowels. *J. Acoust. Soc. Am.* **124**, 3203–3212. (doi:10.1121/1.2980456)

Boersma, P. 2001 Praat, a system for doing phonetics by computer. *Glot Int.* **5**, 341–345.

Burdick, C. K. & Miller, J. D. 1975 Speech perception by the chinchilla: discrimination of sustained /a/ and /i/. *J. Acoust. Soc. Am.* **58**, 415–427. (doi:10.1121/1.380686)

Collins, S. A. 2000 Men's voices and women's choices. *Anim. Behav.* **60**, 773–780. (doi:10.1006/anbe.2000.1523)

Creelman, C. D. 1957 Case of the unknown talker. *J. Acoust. Soc. Am.* **29**, 655. (doi:10.1121/1.1909003)

Dent, M. L., Brittan-Powell, E. F., Dooling, R. J. & Pierce, A. 1997 Perception of synthetic /ba/-/wa/ speech continuum by budgerigars (*Melopsittacus undulatus*). *J. Acoust. Soc. Am.* **102**, 1891–1897. (doi:10.1121/1.420111)

Diehl, R. L., Lotto, A. J. & Holt, L. L. 2004 Speech perception. *Annu. Rev. Psychol.* **55**, 149–179. (doi:10.1146/annurev.psych.55.090902.142028)

Dooling, R. J. & Brown, S. D. 1990 Speech perception by budgerigars (*Melopsittacus undulatus*): spoken vowels. *Percept. Psychophys.* **47**, 568–574.

Dooling, R. J., Okanoya, K. & Brown, S. D. 1989 Speech perception by budgerigars (*Melopsittacus undulatus*): the voiced-voiceless distinction. *Percept. Psychophys.* **46**, 65–71.

Eriksson, J. L. & Villa, A. E. P. 2006 Learning of auditory equivalence classes for vowels by rats. *Behav. Process.* **73**, 348–359. (doi:10.1016/j.beproc.2006.08.005)

Fitch, W. T. 2000 The evolution of speech: a comparative review. *Trends Cogn. Sci.* **4**, 258–267. (doi:10.1016/S1364-6613(00)01494-7)

Fitch, W. T. & Kelley, J. P. 2000 Perception of vocal tract resonances by whooping cranes, *Grus americana*. *Ethology* **106**, 559–574. (doi:10.1046/j.1439-0310.2000.00572.x)

Gentner, T. Q., Fenn, K. M., Margoliash, D. & Nusbaum, H. C. 2006 Recursive syntactic pattern learning by songbirds. *Nature* **440**, 1204–1207. (doi:10.1038/nature04675)

Ghazanfar, A. A., Turesson, H. K., Maier, J. X., Van Dinther, R., Patterson, R. D. & Logothetis, N. K. 2007 Vocal-tract resonances as indexical cues in rhesus monkeys. *Curr. Biol.* **17**, 425–430. (doi:10.1016/j.cub.2007.01.029)

Hauser, M. D., Chomsky, N. & Fitch, W. T. 2002 The faculty of language: what is it, who has it and how did it evolve? *Science* **298**, 1569–1579. (doi:10.1126/science.298.5598.1569)

Hienz, R. D. & Brady, J. V. 1988 The acquisition of vowel discriminations by nonhuman primates. *J. Acoust. Soc. Am.* **84**, 186–194. (doi:10.1121/1.396963)

Hienz, R. D., Sachs, M. B. & Sinnott, J. M. 1981 Discrimination of steady-state vowels by blackbirds and pigeons. *J. Acoust. Soc. Am.* **70**, 699–706. (doi:10.1121/1.386933)

Hienz, R. D., Aleszczyk, C. M. & May, B. J. 1996 Vowel discrimination in cats: acquisition, effects of stimulus level, and performance in noise. *J. Acoust. Soc. Am.* **99**, 3656–3668. (doi:10.1121/1.414980)

Kluender, K. R., Diehl, R. L. & Killeen, P. R. 1987 Japanese quail can learn phonetic categories. *Science* **237**, 1195–1197. (doi:10.1126/science.3629235)

Kuhl, P. K. 1981 Discrimination of speech by nonhuman animals: basic auditory sensitivities conducive to the perception of speech-sound categories. *J. Acoust. Soc. Am.* **70**, 340–349. (doi:10.1121/1.386782)

Kuhl, P. K. & Miller, J. D. 1975 Speech perception by the chinchilla: voiced-voiceless distinction in alveolar plosive consonants. *Science* **190**, 69–72. (doi:10.1126/science.1166301)

Kuhl, P. K. & Miller, J. D. 1978 Speech perception by the chinchilla: identification functions for synthetic VOT stimuli. *J. Acoust. Soc. Am.* **63**, 905–917. (doi:10.1121/1.381770)

Kuhl, P. K. & Padden, D. M. 1982 Enhanced discriminability at the phonetic boundaries for the voicing feature in macaques. *Percept. Psychophys.* **32**, 542–550.

Liberman, A. M. & Mattingly, I. G. 1985 The motor theory of speech revised. *Cognition* **21**, 1–36. (doi:10.1016/0010-0277(85)90021-6)

Liberman, A. M., Cooper, F. S., Shankweiler, D. P. & Studdert-Kennedy, M. 1967 Perception of the speech code. *Psychol. Rev.* **74**, 431–461. (doi:10.1037/h0020279)

Lieberman, P. 1975 *On the origins of language. An introduction to the evolution of human speech.* New York, NY: Macmillan.

Lieberman, P. 1984 *The biology and evolution of language.* Cambridge, UK: Harvard University Press.

Lieberman, P., Klatt, D. H. & Wilson, W. H. 1969 Vocal tract limitations on the vowel repertoires of rhesus monkey and other nonhuman primates. *Science* **164**, 1185–1187. (doi:10.1126/science.164.3884.1185)

Macmillan, N. A. & Creelman, C. D. 2005 *Detection Theory. A User's Guide.* Mahwah, NJ: Lawrence Erlbaum Associates.

Magnuson, J. S. & Nusbaum, H. C. 2007 Acoustic differences, listener expectations, and the perceptual

accommodation of talker variability. *J. Exp. Psychol. Hum.* **35**, 391–409.

Mullennix, J. W., Pisoni, D. B. & Martin, C. S. 1989 Some effects of talker variability on spoken word recognition. *J. Acoust. Soc. Am.* **85**, 365–378. (doi:10.1121/1.397688)

Nearey, T. M. 1989 Static, dynamic, and relational properties in vowel perception. *J. Acoust. Soc. Am.* **85**, 2088–2133. (doi:10.1121/1.397861)

Peterson, G. E. & Barney, H. L. 1952 Control methods used in a study of the vowels. *J. Acoust. Soc. Am.* **24**, 175–184. (doi:10.1121/1.1906875)

Pinker, S. & Jackendoff, R. 2005 The faculty of language: what's special about it? *Cognition* **95**, 201–236. (doi:10.1016/j.cognition.2004.08.004)

Reby, D., McComb, K., Cargnelutti, B., Darwin, C., Fitch, W. T. & Clutton-Brock, T. 2005 Red deer stags use formants as assessment cues during intrasexual agonistic interactions. *Proc. R. Soc. B* **272**, 941–947. (doi:10.1098/rspb.2004.2954)

Smith, D. R. R. & Patterson, R. D. 2005 The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. *J. Acoust. Soc. Am.* **118**, 3177–3186. (doi:10.1121/1.2047107)

Titze, I. R. 2000 *Principles of voice production*. Iowa City, IA: National Center for Voice and Speech.

Trout, J. D. 2003 Biological specializations for speech: what can the animals tell us? *Curr. Dir. Psychol. Sci.* **12**, 155–159. (doi:10.1111/1467-8721.t01-1-01251)

Verzijden, M. N., Etman, E., Van Heijningen, C. A. A., Van der Linden, M. & ten Cate, C. 2007 Song discrimination learning in zebra finches induces highly divergent responses to novel songs. *Proc. R. Soc. B* **274**, 295–301. (doi:10.1098/rspb.2006.3728)

Yule, G. 2006 *The study of language*. Cambridge, UK: Cambridge University Press.