Speech across species : on the mechanistic fundamentals of vocal production and perception

Ohms, V.R.

1

General introduction, thesis overview and discussion

Studying the evolution of speech

Human language constitutes one of the most complex behaviours known to date. It allows us to build and communicate an infinite number of conceptual structures independent of modality (Fitch 2000; Brenowitz *et al.* 2010). The origin of language is unclear and there is much debate about its evolution and the particular properties that make human language unique (e.g. Hauser *et al.* 2002; Dunbar 2003; Castro *et al.* 2004; Pinker & Jackendoff 2005; Anderson 2008; Pinker 2010; Fitch 2010).

Speech on the other hand describes the actual physical phenomenon which is used to convey language. It consists of a limited number of meaningless sounds which can be combined to form a potentially infinite set of meaningful larger units (Liberman & Whalen 2000). As such speech and the mechanisms underlying its production and perception can be subjected to acoustic, physiological, anatomical and neurobiological studies. Unfortunately, such studies reveal little about the evolution of speech. Also, the fossil record of structures involved in speech production is, if at all existent, inconclusive (Fitch 2000; Ghazanfar & Rendall 2008; Fitch 2010) and therefore insufficient to reliably trace speech evolution. However, studying vocal communication in other species enables us to detect mechanisms which have evolved convergently and thus can help identify selection pressures or intermediate steps that might have caused the emergence of these mechanisms (Hauser & Fitch 2003; Jarvis 2004). This comparative approach is one of the prime methods of a young research area referred to as biolinguistics (Fitch 2010).

Insights from the comparative approach

In recent years an increasing number of studies have addressed possible similarities between human and animal vocal communication. Being a learned behaviour is one of the core properties of human speech, but vocal learning is rare in the animal kingdom (Janik & Slater 1997). Among mammals it has been found, besides in humans, only in a few distantly related groups including seals, cetaceans, bats and elephants (Janik & Slater 1997; Poole *et al.* 2005) whereas in our closest relatives, the great apes, or other primates for that matter, vocal learning seems to be largely absent (Fitch 2000).

However, vocal learning has also been demonstrated in three orders of birds, namely songbirds (Marler 1976), parrots (reviewed by Pepperberg 2010) and hummingbirds (Baptista & Schuchmann 1990). Additionally, neuroscientific studies suggest that special brain pathways for vocal learning, one posterior and one anterior, are present in all three groups of vocal learning birds and humans, but not in vocal non-learning birds or mammals (Jarvis 2004).

Moreover, several other parallels between human speech and birdsong have established birdsong as the closest animal analogue to human speech that exists and therefore as an excellent model system to study the underlying mechanisms of speech production and perception (Bolhuis *et al.* 2010).

Both humans and songbirds exhibit a sensitive period early in life during which vocal learning is facilitated (White 2001). Auditory feedback by hearing others and themselves is crucial in order to develop normal vocalizations, initially during a sensory learning phase and later during sensorimotor learning (Doupe & Kuhl 1999). It has also been hypothesized that in humans as well as birds innate predispositions and biases guide which sounds will be learned (Whaling *et al.* 1997; Braaten & Reynolds 1999). However, evidence for this is more readily available in songbirds because they can be subjected to cross-fostering (Clayton 1989) and isolation studies (Braaten & Reynolds 1999) which cannot be conducted in humans for apparent ethical reasons.

Recently, thanks to the development of new molecular techniques, another component came into focus: the genetic basis for vocal communication (Bolhuis *et al.* 2010). Mutations in the gene coding for the transcription factor FOXP2, for instance, are associated with a speech disorder in humans, manifested by impaired motor control of orofacial movements and deficits in several aspects of language processing (Lai *et al.* 2001). Interestingly, the same gene when down-regulated in Area X of juvenile zebra finch brains causes these birds to inaccurately copy tutor songs by omitting some song elements and showing more variability between song repetitions (Haesler *et al.* 2007).

The parallels described so far are mainly concentrating on vocal development and learning. However, a growing body of evidence indicates that there are also similarities between human speech and birdsong with respect to vocal production and perception. Both human speech and birdsong are produced by a primary sound source and subsequently filtered in the vocal tract (Fant 1960; Nowicki 1987), but currently vocal tract filtering is less well understood in (song)birds than it is in humans. Regarding vocal perception, which has been extensively studied in humans, it is still unclear whether specialized perceptual abilities are necessary for speech perception and whether these abilities occur in songbirds too. This thesis aims to address both vocal production and perception mechanisms in (song)birds compared to humans in order to shed more light on the similarities and differences of these mechanisms.

Formants and their relevance in vocal communication

Human speech is characterized by a broad frequency spectrum. Voiced speech sounds are produced by the vibrations of the vocal folds in the larynx, generating a fundamental frequency with harmonic overtones. This sound is filtered while traveling through the vocal tract, consisting of the pharyngeal, oral and nasal cavities. Depending on the dimensions of these cavities some frequencies within the broadband spectrum are amplified whereas others are attenuated (Titze 2000). Amplified frequencies appear as dark bands on spectrograms and are referred to as 'vocal tract resonances' or 'formants' (Fig. 1.1). It is important to notice that formants are independent of the sound source and that they are rapidly modulated during speech by moving the articulators such as tongue, lips and soft palate (Fant 1960). Formants are especially relevant in the production and perception of vowels (Ladefoged 2006). The difference between the words 'beat' and 'bit' is based on different formant values, primarily regarding the two lowest formants F1 and F2 which represent resonances of the pharynx and mouth cavity respectively.

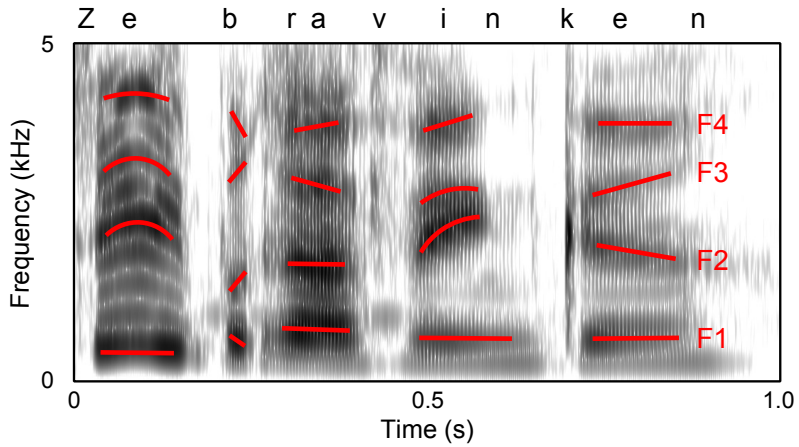

Figure 1.1 Spectrogram of human speech

This figure shows a spectrogram of a female voice saying *'zebravinken'* (zebra finches). Formant frequencies are highlighted in red. kHz, kilohertz; s, seconds; F1, first formant; F2, second formant; F3, third formant; F4, fourth formant.

During the ontogeny of modern humans the larynx descends, enabling the tongue to move both vertically and horizontally thereby allowing the production of a wide variety of formant patterns by shaping the vocal tract in numerous different ways (Lieberman *et al.* 1969). It is commonly believed that this anatomical configuration of the human vocal tract, together with the loss of laryngeal air sacs, was a necessary prerequisite for the evolution of speech (Fitch 2000). However, the idea that the evolution of speech was a driving force in the evolution of this configuration has been questioned by the discovery that several other species exhibit a descended larynx too.

It has been shown, for instance, that males of both red and fallow deer lower their larynges while roaring. This presumably serves to elongate the vocal tract causing it to resonate at lower frequencies which could make these animals sound bigger (Fitch & Reby 2001; Reby *et al.* 2005). This 'size exaggeration hypothesis' gains further support from observations of trumpet birds which exhibit elongated tracheas assumed to lower the pitch of their vocalizations (Clench 1978; Fitch 1999). At the same time this means that the descent of the larynx has been shaped by sexual selection and provided a pre-adaptation for speech evolution (Fitch 2000).

Recent evidence from comparative MRI studies also suggests that the descent of the larynx evolved before the human and chimpanzee lineages separated (Nishimura *et al.* 2006). The authors speculate that facial flattening, instead of lowering the larynx, enabled the typical configuration of the vocal tract and hence played a more important role in the evolution of speech.

However, less is known so far about the production of vocal tract resonances in songbirds. The vocal organ of songbirds, the syrinx, is much more complex than the human larynx. In Oscine songbirds two sets of vibrating labia, located at the cranial end of each bronchus (Goller & Larsen 1997), are involved in vocal production. Each of these sets can be controlled independently (Suthers 1990) and enables the birds to sing with two voices simultaneously or switch between both sets of labia while singing (Suthers 1990; Suthers *et al.* 1994; Suthers *et al.* 2004; Zollinger & Suthers 2004). This complexity has initially led to the hypothesis that acoustic variation in birdsong arises at the sound source and that, contrary to human speech production, vocal tract filtering only plays a minor role (Greenewalt 1968). Newer studies, however, suggest that there might be more parallels in human and avian sound production than originally assumed. Therefore one of the main objectives of this thesis is to identify potential articulators involved in vocal tract filtering and to evaluate their effects on sound production.

The second main objective concerns vocal perception mechanisms. Humans are highly sensitive to speech sound variation and formant patterns, but the question whether this sensitivity coevolved with speech production and is therefore a uniquely human trait or whether it is based on general auditory processing mechanisms remains highly controversial (e.g. Lieberman 1975; Kuhl & Miller 1975, 1978; Pinker & Jackendoff 2005). Some of the perception mechanisms which scientists initially claimed to occur only in humans, like the categorical perception of speech sounds, are shared with a number of different species including both birds (Kluender *et al.* 1987; Dooling & Brown 1990) and mammals (Kuhl & Miller 1975, 1978; Kuhl & Padden 1982). Another significant aspect of speech perception regards our ability to recognize words regardless of speaker identity. We can distinguish words that closely resemble each other by extracting relevant acoustic information while at the same time ignoring speaker-dependent variation. It still has to be tested if this property is uniquely human implying that it has evolved specifically for human speech or if it is, like categorical perception, based on general processing mechanisms of the auditory system.

Thesis overview

This thesis documents four studies of which two are dealing with the production and the other two with the perception mechanisms of vocal tract resonances by birds. I primarily used zebra finches in my experiments because they are the most widely used model species for studies on vocal learning, development and perception while relatively little is known about vocal production in this species. I also chose monk parakeets as another species to study vocal tract filtering since there are indications for significant differences regarding vocal production between parrots and songbirds (e.g. Beckers *et al.* 2004).

In **chapter 2** I describe an experiment which was designed to identify potential vocal tract articulators in zebra finches and to evaluate their significance on vocal tract filtering in this species. First we obtained cineradiographic movies of singing zebra finches, for which the analyses revealed beak gape and the expansion of the oropharyngeal-esophageal cavity (OEC) to be the main articulators involved in vocal production. These results are in line with earlier studies that found positive correlations between beak gape and frequency patterns in several songbird species including zebra finches (Westneat *et al.* 1993; Podos *et al.* 2004; Williams 2001). More interesting is the observation that zebra finches expand their OEC substantially while singing. This has first been demonstrated in northern cardinals (Riede *et al.* 2006) and subsequently in white-throated sparrows (Riede & Suthers 2009). Both of these species, however,

produce rather simple, pure-tone songs with little energy in higher harmonics and it has been hypothesized that OEC expansion tracks the fundamental frequency. Zebra finches on the other hand produce songs consisting of many different element types. Most of these elements are broad-band and exhibit a rich frequency spectrum including harmonic stacks with varying amplitude patterns. Due to the complexity of zebra finch song it is difficult to establish clear relationships between articulator configurations and sound patterns. Nevertheless, when experimentally manipulating beak gape and OEC expansion in the second part of the study, we found a downwards shift in peak frequency with increasing OEC expansion as well as an amplitude increase especially around 1.5 and 4.5 kHz. Beak gape on the other hand seems to emphasize frequencies around 5 kHz and above. These results demonstrate that the upper vocal tract, especially beak gape and OEC expansion, plays a major role in resonance filtering of zebra finch song resulting in elaborate note types exhibiting formant like patterns.

Parrots are another group of birds that produce complex broad-band sounds and they are very well known for their sophisticated ability to imitate human speech. In contrast to songbirds parrots have a prominent tongue with many intrinsic muscles and a fleshy, flexible surface which resembles the human tongue (Homberger 1986). This observation has motivated the hypothesis that tongue movements play a more important role in vocal production in this group of birds compared to songbirds (Patterson & Pepperberg 1994) and observations of a speech-imitating parrot (Warren *et al.* 1996) as well as experimental manipulations of tongue position (Beckers *et al.* 2004) support this claim. However, to date no direct observations of tongue movements in naturally vocalizing parrots exist, nor is it not known what other articulators are involved in vocal production in parrots.

In **chapter 3** of this thesis I therefore address this question by employing cinematographic imaging of naturally vocalizing monk parakeets. On the videos we could identify three main articulatory movements: beak opening, tongue height changes and tracheal shortening. Although earlier studies already indicated the significance of tongue movements, they found main effects in the front/back dimension of tongue position while the parakeets in our study primarily manipulated tongue height. From the nine different vocalization types produced by adult monk parakeets (Martella & Bucher 1990) the birds in our study uttered only three. This leaves the possibility that tongue movements in the front/back dimension are of significant importance in some of the other vocalizations that we could not record. Yet, in greeting calls which exhibit formant changes and which are included in our analysis, manipulations of tongue position in the vertical dimension

seem more prominent than changes in the horizontal plane. Interestingly, we also found evidence for tracheal shortening whereas an earlier study on zebra finches concluded that tracheal length changes are too small to affect vocal production in that species (Daley & Goller 2004). Furthermore we found significant positive correlations between sound amplitude and magnitude of articulator movements in greeting calls and chatter sounds for beak opening, tongue height and tracheal shortening for some of the birds. Since modulations of the fundamental frequency (F0) are very fast in monk parakeet contact calls while articulator movements are comparatively slow it is likely that changes in F0 are generated at the sound source. Formant patterns as occurring in greeting calls, however, are probably the result of the vocal tract filter and as such determined by articulator movements. Unfortunately it was not possible to establish clear relationships between formant changes and articulator configurations since the exact properties of the sound source and its behaviour are largely unknown. Therefore future studies will have to pay attention to the precise nature of these relationships and more data on the anatomical as well as physical properties of the parrot vocal apparatus are needed in order to establish a reliable model of sound production in these birds.

In the second half of this thesis I address formant perception by birds in comparison with humans. As shown in chapters 2 and 3 there is convincing evidence that both songbirds and parrots use various articulators to filter the sound produced in the syrinx. Although there are differences in vocal communication between songbirds, parrots and humans the mechanisms of sound production share the principle of active vocal tract filtering, enabling both humans and birds to increase the variety of sounds that can be produced. Following this observation the question arises whether the mechanisms underlying formant perception in particular and frequency modulation in general are also comparable between birds and humans. If so, there is no reason to assume that special mechanisms enabling formant perception evolved in humans as a result of coevolution between speech production and perception. Instead general auditory processing capabilities might be sufficient to allow discrimination of human speech sounds.

Chapter 4 deals with a study investigating speaker normalization in zebra finches using natural human speech obtained from Dutch speaking young adults of both sexes. One of the most remarkable phenomena in human speech concerns our ability to recognize words independent of speaker and strong variation between speakers. Speech scientists have attributed this to the human capacity for intrinsic and extrinsic speaker normalization. Intrinsic speaker normalization accounts for the fact that sounds

which are perceived as the same phoneme can have different acoustic realizations (Liberman *et al.* 1967) by assuming that every speech sample can be categorized using a normalizing transformation (Nearey 1989). At the same time it is well known that there is a speaker effect on speech discrimination initially hampering discrimination across speakers (Creelman 1957; Mullenix *et al.* 1989). This difficulty however is overcome by establishing a reference frame from different speech sound samples (Nearey 1989; Magnus & Nusbaum 2007). We have applied operant conditioning techniques to train zebra finches to discriminate between two naturally produced words, *'wit'* and *'wet'*, that differ mainly in their formant patterns and later transfer this discrimination to unfamiliar voices of (1) the same sex and (2) the opposite sex. All of the eight birds tested were able to discriminate between the words and categorize them independent of speaker identity. Our analysis revealed that the essential clue enabling categorization were the different formant patterns. Furthermore, the birds employed, just like humans, a combination of intrinsic and extrinsic speaker normalization to accomplish the task. This result indicates that the way formants are perceived is either widely spread in the animal kingdom or evolved convergently in birds and humans.

The last chapter of this thesis, **chapter 5**, describes a direct comparison of acoustic cue-weighting in vowel perception in zebra finches and Dutch adults. It has been shown in the past that both Swedish and Canadian-English babies aged three to fifteen months are more sensitive towards low frequency components, i.e. F1, when discriminating vowels (Lacerda 1993, 1994; Curtin *et al.* 2009). This is somewhat surprising since the general notion assumes that the language environment dictates speech perception starting as early as 6 months of age. This might either indicate a universal human bias towards lower frequencies in vowel perception or be the result of maturation of the auditory system. However, it has yet to be explored if these biases are strictly linked to speech sound perception and hence a uniquely human property or a more general characteristic of auditory perception. In a very first attempt to tackle this question we provided both zebra finches and native speakers of Dutch with a Go/NoGo discrimination task using a highly comparable setup. Both groups first had to learn to discriminate between two synthesized words differing only in the embedded vowel sound. The vowels were chosen to differ in F1 as well as F2. In the next step two synthesized 'probe' sounds were added to the discrimination task. Probe sounds were never reinforced and the reactions of the subjects to the probes allowed us to draw conclusions about the way these were perceived. One of the probe sounds had the same F1 frequency as the first stimulus and the same F2 frequency as the second stimulus and vice versa for the other probe

sound. The responses to the probes were strikingly similar in birds and humans and both exhibited a cue-weighting bias towards high frequency components, i.e. F2. This is exactly opposite to what has been found in human babies and firstly demonstrates that cue-weighting is not a uniquely human property tied to speech perception and secondly suggests that a developmental component, likely in the form of auditory maturation, plays an important role in the emergence of such a bias. The major strength of this experiment lies in the use of a highly similar setup for testing birds and humans and therefore makes the results maximally comparable. Furthermore it emphasizes the value of comparative studies across species and ages which should be taken into account when studying mechanisms of speech perception.

Discussion and conclusion

In this thesis I have shown that both zebra finches and monk parakeets use different vocal articulators to modify the sound produced by the syrinx. While in songbirds beak gape and the expansion of the OEC are most important in vocal production, in parakeets tongue movements seem to be the major source of spectral modulation. This is very comparable to human speech production and might be one of the most important parameters of speech imitation by parrots. Based on these observations it can be concluded that sound production mechanisms between birds and humans are more similar than initially assumed. This might suggest convergence in evolutionary patterns. It is conceivable that some of the structures involved in vocal production, such as tongue and beak, initially evolved as part of the food processing system. In that case ecological adaptations for different diets were the driving forces behind the evolution of these structures. At a later point the already existing articulators might have been exploited by the communication systems in order to increase sound variation.

Nevertheless there are remarkable differences in the anatomy and physiology of the sound producing organs as well as in the articulatory patterns. More research based on the current findings could therefore provide detailed models of sound production in both songbirds and parrots.

With regard to speech perception in humans and songbirds I have shown that zebra finches, just like humans, rely on formant patterns to discriminate between highly similar words while at the same time using both intrinsic and extrinsic speaker normalization to categorize words independent of speaker identity. This is an important finding with strong implications for the evolution of formant perception. As has been speculated earlier formant perception likely emerged in a wide range of species serving

to obtain information about an individual's sex, age, size and identity (Ghazanfar *et al.* 2007). Speech might at a later point have exploited this capacity for formant perception eventually leading to a communication system that makes extensive use of formants in order to code linguistic meaning. The finding that human adults and zebra finches exhibit the same cue-weighting bias in vowel perception is in accordance with this hypothesis. Both rely more on F2 frequencies which fall in the most sensitive frequency range in both species when categorizing ambiguous vowels. Human infants on the other hand seem to initially base their discrimination on those frequency components which are spectrally most prominent, namely F1.

In summary I have shown that (song)birds hold the capacity for formant production and perception and that the underlying mechanisms show more similarities between birds and humans than realized before. Both songbirds and parrots can serve as valuable models to address specific questions on the exact nature of these mechanisms and eventually identify selection pressures that might have shaped the evolution of such elaborate vocal communication systems as are only found in humans and birds. Now that some of the mechanisms underlying formant production and perception have been identified, future studies can build on this knowledge to explore more detailed questions concerning e. g. the function of formant patterns in natural birdsong, or the modeling of vocal production and perception mechanisms. Synthesizing songs and manipulating resonance patterns will be important tools to address these questions.