# HOW DOGS HEAR US:

# PERCEPTION OF THE HUMAN VOICE

# BY DOMESTIC DOGS (*CANIS FAMILIARIS)*

**Dissertation submitted to the University of Sussex**

**for the degree of Doctor of Philosophy**

**Victoria Frances Ratcliffe**

**August 2015**

## Declaration

This thesis conforms to an 'article format' in which the middle chapters consist of discrete articles written in a style that is appropriate for publication in peer-reviewed journals in the field. The first and final chapters present synthetic overviews and discussions of the field and the research undertaken.

Chapter 3 is published in *Animal Behaviour* as:

> Ratcliffe, V.F., McComb, K. & Reby, D. (2014). Cross-modal discrimination of human gender by domestic dogs, *Animal Behaviour*, *91*, 127-135.

The author contributions are as follows: I was responsible for all aspects of the data collection, data analysis and the writing of the manuscript; Dr. Reby was responsible for the initial conception of the research, as well as contributing to the study design and providing feedback on the manuscript. Prof. McComb also provided feedback on the data analysis and manuscript.

Article II of Chapter 4 is in press in *Multisensory Research* as:

> Ratcliffe, V.F., Taylor, A.M. & Reby, D. (2015). Cross-modal Correspondences in Non-Human Mammal Communication, *Multisensory Research*, in press.

Dr. Reby was responsible for the initial conception of the review. I was responsible for the writing of the manuscript. Dr Reby and Dr Taylor provided feedback and corrections to the manuscript.

Chapter 5 is based on a study published in *Current Biology* as:

> Ratcliffe, V.F., & Reby, D. (2014). Orienting asymmetries in dogs' responses to different communicatory components of human speech, *Current Biology*, *24*, 2908-2912.

I was responsible for all aspects of the study from the initial conception to writing the manuscript; Dr. Reby contributed to the study design, stimuli creation and provided feedback on the manuscript.

I hereby declare that this thesis has not been and will not be, submitted in whole or in part to another University for the award of any other degree.

Signature: Victoria Ratcliffe

**Acknowledgements**

Firstly, I would sincerely like to thank my supervisor David Reby for his expert guidance, constant support and encouragement throughout my PhD. I have learnt a tremendous amount and I am extremely grateful to have had the opportunity to work with him. I could not have asked for a better supervisor.

Many thanks also to Karen McComb for her constructive feedback on our paper on cross-modal human gender discrimination and for her helpful advice in each of my end of year reviews. I would also like to thank Anna Taylor for her collaboration, especially on our review article, as well as all of the assistants who helped me with data collection, particularly Lotte Carlebach, Thibault Chabert, James Day, Solène Derville, Mike Fitzgerald, Meddy Fouquet, Sarah Jeannin, Olessya Zhuravleva and Mariam Zia. I am also extremely grateful to all of the dog owners for their interest in our research and for volunteering their dogs to participate in our studies, particularly Dean and Sara Hart, the owners of The Dog Hut, for their support and assistance with data collection.

I would also like to say a special thanks to my family and friends for their support throughout my PhD. I am immensely grateful to my parents and brother for their love, trust and belief in me throughout my studies. I also want to share my deep appreciation for all of my friends, especially Val Cartei, Mariam Zia, Busra Sultana, Olessya Zhuravleva, Renata Fialho and Emma Baker for their invaluable encouragement and innumerous tea-time chats. Finally, I am thankful for my fantastic dog Kira, who has happily nudged me along for the last two years with enforced breaks, laughter and perspective.

**UNIVERSITY OF SUSSEX**

**Victoria Frances Ratcliffe** **Doctor of Philosophy**

**How dogs hear us: perception of the human voice by domestic dogs (*Canis familiaris*)**

**Summary**

Domestic dogs have co-habited with humans for at least 15000 years. Close social interaction between the two species has promoted inter-specific communication and dogs now show advanced skills in responding to human signals in comparison to wolves. However, research into dogs' abilities to interpret human signals has predominantly focussed on visual gestures, while their responses to vocal signals remain under-investigated. Exploring the perception of human speech by dogs, a phylogenetically distant species, could provide new insights into the evolution of mammal communication. Therefore, the aim of this thesis is to assess human speech perception by dogs. Speech is composed of two main communicative components: the segmental phonemic cues carrying the linguistic content and the supra-segmental cues transmitting information about the speaker such as their gender, age and emotional state. I first explore how dogs perceive supra-segmental cues, determining that they are capable of the cross-modal discrimination of human gender. I then provide a review detailing the mechanisms underlying cross-modal associations in mammal communication, before testing which of these mechanisms may enable dogs to cross-modally associate cues to human age. The results indicate that dogs learn to match some voices to humans according to their age category, while also perceiving more general cross-modal correspondences in the environment. Finally, I investigate how dogs dissociate the main communicatory components of speech during processing, providing evidence that dogs differentially process segmental and supra-segmental cues according to their communicative content. In doing so, dogs appear to express parallel hemispheric biases to those reported in humans. Additionally, the results provide the first clear demonstration that dogs attend to the combinatory structure of the phonemic content in learnt commands. Overall, this thesis extends our knowledge of dogs' perception of human signals, indicating that they are capable of perceiving each of the main components of speech in a functionally relevant manner. Together the results suggest that dogs share some of the cognitive and social processes involved in speech perception with human listeners.

## Table of Contents

# List of Tables

## Chapter 3



## Chapter 4 Article I



## Chapter 5

## List of Figures

### Chapter 3



### Chapter 4 Article II

**Chapter 5**

# CHAPTER 1: INTRODUCTION

## Overview

In the context of animal communication, a 'signal' can be defined as an act or structure that causes a response in other organisms which usually functions to benefit either one or both parties (Maynard-Smith & Harper, 2003; Wilson, 1975). Although communication is often thought of as occurring between members of the same species, signals can also be intended for, or intercepted by, hetero-specific receivers. Interactions between humans and domestic dogs *Canis familiaris* represent a particularly well known case of intentional inter-specific communication, with around nine million dogs living with human families in the UK alone (PFMA Pet Population report, 2014). The social relationship between humans and dogs began somewhere between 30,000 years ago (mitochondrial genome analyses: Skoglund, Ersmark, Palkopoulou & Dalén, 2015; Thalmann et al., 2013) and 12,000 years ago (fossilised shared burial sites; Davis & Valla, 1978), when individual wolves began to move into the human environment, potentially to exploit available resources (Coppinger & Coppinger, 2001). The transition into human settlements resulted in a significant level of physiological and behavioural divergence between dogs and modern grey wolves *Canis lupus* from their ancient wolf ancestor (Wilkins, Wrangham & Fitch, 2014), with dogs eventually integrating into human social groups and becoming the first domesticated species (Savolainen, Zhang, Luo, Lundeberg & Leitner, 2002). Unlike wolves, dogs now appear to be innately predisposed to develop strong attachment bonds with humans (Gácsi et al., 2005; Topál et al., 2005), and humans are important providers and social partners for many individuals. This close social relationship means that the ability to interpret human signals is highly relevant for dogs and over the past 15 years there has been a significant growth in scientific interest in communication between humans and dogs. Evidence has accumulated demonstrating that dogs are unusually adept at responding to human signals in comparison to non-domesticated mammals, inspiring a number of different theories speculating the potential effects various selection pressures may have had on dogs' socio-cognitive abilities during the process of domestication (see Kaminski & Nitzchner, 2013, for a recent review). However, research has largely focussed on dogs' abilities to interpret visual human gestures, with comparatively little attention paid to their perception of human vocal signals, despite the fact that vocal communication constitutes a major aspect of human-dog interactions (Gibson, Scavelli,

Udell & Udell, 2014). The human voice can encode three broad categories of functionally relevant information transmissible to receivers: indexical cues related to physical attributes of the speaker (e.g. body size, age, gender, identity); dynamic cues signalling the emotional and motivational state of the speaker; and phonemic cues which constitute the verbal information in speech (Belin, Fecteau & Bedard, 2004). Because few studies have specifically tested dogs' responses to human vocal cues independently from other communicative cues, little is known about the extent to which they perceive the individual components of human vocal signals. Therefore, the main aim of this thesis is to investigate whether dogs living with humans are spontaneously able to extract information from the three main components of the human voice in a functionally relevant manner. Determining how dogs perceive human vocal signals is necessary before it is possible to fully address the question of whether dogs have undergone specific socio-cognitive adaptations during domestication which facilitate their perception of human signals.

In this introductory chapter, I review our current understanding of how dogs perceive human signals. Although the main focus of the review is related to the communicative content of vocal signals, I first provide a brief overview of dogs' responsiveness to human-given referential signals and ostensive cues (both visual and vocal), as these forms of signalling have received the most attention from previous research. As will be apparent, studies have mainly investigated dogs' responses based on the presence/absence of referential vocal signals rather than on their functional components. I therefore go on to provide a description of human vocal production according to the 'source-filter theory' (Fant, 1960), which offers a practical framework for explaining how anatomically-controlled acoustic variation encodes relevant information in the human voice, as this forms the methodological basis for the current research. This section is followed by a detailed summary of the available evidence to suggest that dogs are able to perceive the main informative components of human vocalisations and speech: indexical vocal cues, dynamic emotional vocal cues and phonemic information in speech. Finally, I discuss the potential existence of cerebral hemispheric asymmetries in the way that dogs process information from these different functional components when they perceive human speech signals.

## Perception of Human Visual Gestures

Dogs' responses to human signals have been primarily investigated using variations of the 'object choice paradigm', which was originally adapted from work with human infants and non-human primates (Povinelli, Reaux, Bierschwale, Allain & Simon, 1997; Tomasello, Call & Gluckman, 1997). In the context of a game where a reward is hidden, the subject is faced with a number of opaque containers and the experimenter indicates towards one of them. The aim of the task is for the subject to infer that the communicative intentions of the experimenter are an attempt to direct their attention in a contextually relevant manner, providing them with the information needed to correctly motion towards the target container themselves and obtain the reward (Tomasello et al., 1997). Studies using this paradigm have demonstrated that dogs are both skilful and flexible in their ability to interpret human referential signals (for reviews, see Kaminski & Nitzschner, 2013; Miklósi & Soproni, 2006). More specifically, dogs have been shown to successfully respond to a variety of visual gestures such as gaze direction (either a head turn or a static head looking towards the location; Hare, Call & Tomasello, 1998; Miklósi, Polgárdi, Topál & Csányi, 1998), nodding (Miklósi et al., 1998), and pointing, including 'distal momentary pointing' where the gesture is made from a greater distance from the target using the cross-lateral arm and extinguished before the subject is released (e.g. Lakatos, Soproni, Dóka & Miklósi, 2009), as well as pointing to the correct container whilst moving towards the incorrect container (Hare & Tomasello, 1999; McKinley & Sambrook, 2000). The flexibility that dogs show in successfully responding to familiar visual gestures also extends to novel communicative signals, such as observing the experimenter place a marker in front of the correct target (Agnetta, Hare & Tomasello, 2000). The potential use of low-level mechanisms to solve these tasks, such as olfactory cues or pure local enhancement, have been ruled out (e.g. Hare et al., 1998; Hare & Tomasello, 1999), confirming that the dogs were in fact responding to the human gestures provided.

Although it has been suggested that similarly to trained non-domesticated animals (e.g. bottlenose dolphins *Tursiops truncatus*: Tschudin, Call, Dunbar, Harris & van der Elst, 2001), dogs may simply learn to associate human gestures with specific responses and outcomes through experience (Bentosela, Barrera, Jakovcevic, Elgier & Mustaca, 2008; Elgier, Jakovcevic, Barrera, Mustaca & Bentosela, 2009; Udell, Dorey & Wynne, 2010), this theory has been brought into question as dogs appear to understand the

communicative nature of the signals. Indeed, dogs ignore similar non-informative arm movements (Kaminski, Schulz & Tomasello, 2012) and use contextual information to determine the relevance of the signals before responding to them (Scheider, Grassmann, Kaminski & Tomasello, 2011). Moreover, across object-choice studies, dogs' responses to many of the gestures have been accurate from the first trial, demonstrating that they were already able to solve the tasks before the experiments began. Prior knowledge of the correct responses to human visual gestures is evident from very early in their development, as 6-week-old puppies also successfully follow human points, suggesting that extensive associative experience with human gestures may not be necessary for dogs to accurately respond to these signals (Riedel, Schumann, Kaminski, Call & Tomasello, 2008, although see Wynne, Udell & Lord, 2008). This hypothesis is further supported by the demonstration that previous training history does not influence adult dogs' accuracy in interpreting human gestures (Cunningham & Ramos, 2014). Further observations that dogs outperform both chimpanzees *Pan troglodytes* (Hare, Brown, Williamson & Tomasello, 2002; Kirchhofer, Zimmermann, Kaminski & Tomasello, 2012) and wolves in following human points (Hare et al., 2002; although see Gácsi et al., 2009 and Udell, Spencer, Dorey & Wynne, 2012), have provided support for the 'domestication hypothesis', which proposed that dogs' social-cognitive abilities may have adapted during domestication, converging with those of humans through a phylogenetic process of enculturation (Hare et al., 2002). However, the precise mechanisms responsible for dogs' aptitude for responding to human gestures still remain unknown, and more recent theories have increasingly emphasised the potential importance of less sophisticated social-cognitive predispositions that dogs have developed during domestication, such as a greater tolerance of humans as social partners (Hare & Tomasello, 2005; Hare et al., 2005; Hare, Wobber & Wrangham, 2012; Range & Viranyi, 2015; Udell et al., 2010), as well as direct human-driven selection for cooperativeness and obedience (e.g. Kaminski & Nitzschner, 2013; Miklósi, Topál & Csányi, 2004). In line with human-driven selection for cooperativeness in dogs, it has been suggested that dogs may simply interpret human gestures as imperative commands directing their movement to a specific location rather than as informative signals that refer to things in the environment (Kaminski, 2009; Topál, Gergely, Erdõhegyi, Csibra & Miklósi, 2009; although see Scheider, Kaminski, Call & Tomasello, 2013). Therefore, whilst it seems likely that adaptations during

domestication have enhanced dogs' skills in responding to human gestures, it is not yet clear whether they interpret these signals in a similar way to humans.

## Do Dogs Show a Comparable Level of Sophistication in their Perception of Human Vocal Signals?

Although the mechanisms underlying dogs' abilities to respond to human gestures remain to be established, it is clear that they are unusually adept at using human-given visual signals in comparison to other species. However, the principle means by which humans communicate with each other is through speech rather than visual gestures, and dog owners also use vocal signals to communicate with their pets (Pongrácz, Miklósi & Csányi, 2001). Based on the large body of evidence demonstrating dogs' skilfulness in responding to human gestures, it is reasonable to predict that dogs will also express comparable interpretive abilities when presented with human vocal signals. Indeed, consistent with the results obtained from investigations using visual gestures, Rossano, Nitzschner and Tomasello (2014) demonstrated that dogs were also able to follow the direction in which human voices were projected to find a hidden reward in an object choice paradigm. In this study, the experimenter positioned themselves behind a screen, in closer physical proximity to the incorrect container but facing the correct container, and verbally expressed excitement towards the correct container. Adult dogs were able to perceive the directionality of the voice and use this cue to choose the correct location. Furthermore, this ability was already established in 10-week-old puppies if they had been sufficiently socialised with humans. The results provided further support to previous demonstrations that dogs can become skilled in responding to human referential signals from an early age, and moreover, that this ability may be independent of the sensory modality through which the signals are perceived.

## Human Voices as 'Ostensive Cues'

Human voices also appear to improve the perception of visual gestures by functioning as 'ostensive cues' for dogs in a comparable way to human infants, serving as primers or occasion setters indicating a person's intention to communicate with them (Csibra, 2003; Sperber & Wilson, 1986). In addition to eye contact, pre-verbal human infants are sensitive to high-pitched infant-directed speech (Csibra, 2010), and learn that these cues indicate that accompanying signals are intended for them from around one year of age (Kaminski et al., 2012). It has been argued that ostensive cues do not merely capture the

infant's attention because they are intrinsically salient and thus increase arousal (as a loud noise would for example), but instead focus attention on the signaller because they are interpreted as cues to communicative intent (Csibra, 2010). Both eye gaze and to a lesser extent vocal addressing may similarly function as ostensive cues for dogs, as these signals improve their responses to subsequent human gestures (e.g. Kis et al., 2012; Téglás, Gergely, Kupán, Miklósi & Topál, 2012), causing independent, additive response enhancement when combined (Kaminski et al., 2012). Furthermore, ostensive cues do not appear to merely increase attention towards the person and consequently to their communicative gestures, as dogs have been observed to ignore such cues if they are directed at a third party (Kaminski et al., 2012). The simple effect of increased attention was also discounted in a study where dogs learnt to follow a maze detour by observing a human demonstrator, as their performance was significantly enhanced if the human provided verbal encouragement during the demonstration rather than remaining silent, despite there being no difference in the amount of attention the dogs paid to the demonstrator in either condition (Pongrácz, Miklósi, Timár-Geng & Csányi, 2004). Kaminski et al., (2012) suggested that similarly to the use of infant-directed speech with pre-verbal human infants, high-pitched vocalisations are likely to function as ostensive cues for dogs. This hypothesis was based on the observation that dogs' performances were equivalently improved in an object choice paradigm when the experimenter said either the dog's own name or an unfamiliar name in a high-pitched tone of voice (Kaminski et al., 2012), indicating that the dogs' responses to the gesture were enhanced when it was preceded by a high-pitched utterance independently of the word spoken. However, further work is still necessary to establish the importance of voice pitch specifically as an ostensive cue for dogs, which could be achieved through a direct comparison of the effectiveness of high- versus low-pitch voices.

Although the importance of specific voice features have not yet been determined, these studies demonstrate that human voices can function as ostensive cues for dogs and increase their performance in responding to information communicated in human signals. It therefore appears that human voices can function equivalently to visual gestures in both priming inter-specific communication with dogs and in directing their responses. However, unlike the visual gestures that humans use to communicate with dogs, which are generally one-dimensional signals transmitting a single message, voices are multi-dimensional signals by nature, encoding a range of additional information

about the characteristics of the signaller and their informative intentions. For humans, vocal communication is a crucial component of intra-specific social interactions, primarily as the principle medium for conveying language, but also as an 'auditory face', transmitting information about the speaker such as their identity, gender and emotional state (Belin et al., 2004). Although it is apparent that dogs pay attention to human voices, the extent to which they are able to obtain relevant information communicated within the vocal signals is less well established. In order to investigate how dogs might extract information from the human voice, it is important to first understand the mechanisms of vocal production, and specifically how information is encoded within the acoustic structure of the signal.

## The Source-Filter Theory of Human Voice Production

Because all mammals share a fundamentally similar vocal apparatus (Titze, 1994; Fitch & Giedd, 1999), the acoustic structure of both human and dog vocalisations can be explained using the 'source-filter theory' of voice production (Fant, 1960), which was first developed to determine the anatomically-related mechanisms engaged in shaping the structure of human speech. According to this theory, sound production involves two independent parts of the vocal apparatus. The first component is the 'source', where the sound is initially formed. This occurs in the glottis, which is positioned at the superior border of the larynx and comprised of soft tissue layers of muscle and vocal ligament (the vocal folds) as well as the spacing between them. During phonation, air is expelled from the lungs and forced through the closed glottis, pushing the vocal folds apart. Biomechanical forces pull the vocal folds back together, resulting in a self-sustaining sequence of opening and closing, known as 'flow induced oscillation' (Chan & Titze, 2006), which causes cyclic variation in air pressure. The rate of oscillation determines the fundamental frequency (F0; perceived as the voice pitch) and associated integer multiple frequencies (harmonics) of the subsequent glottal wave, which forms the source signal. Because the oscillation of the vocal folds can be likened to a simple vibrating string, F0 (in Hz) can be approximated by the following formula:

$$F0 = \frac{1}{2L}\sqrt{\frac{\sigma}{\rho}}$$

where $L$ is the length of the vocal fold, σ is the stress applied (force per unit area) and ρ is the tissue density (Titze, 1989). Based on this equation, the F0 is inversely proportional to the length of the vocal folds and directly proportional to the square root of the tension exerted over the tissue density. Therefore longer, heavier and more relaxed vocal folds vibrate at a slower rate and produce a lower F0. However, the F0 can be manipulated to some extent by flexion/relaxation of the muscles controlling the tension of the vocal folds. Whilst the source signal is generally periodic, 'non-linear phenomena' can also occur, such as subharmonics (additional harmonics beneath the F0), biphonation (two independent F0s), and deterministic chaos (broadband signals with no harmonics) (Fitch, Neubauer & Herzel, 2002). Muscular interactions and changes in subglottal pressure also create additional properties in the source signal, including the amplitude contour and duration of the sound (Titze, 1994).

Once generated in the larynx, the source signal then resonates through the supra-laryngeal vocal tract (the filter), which consists of the air cavities between the larynx and the openings of the mouth and nostrils. The vocal tract acts as a bank of bandpass filters, selectively enhancing or dampening specific harmonic frequencies depending on the resonant properties of its physical structure (primarily length and shape). Resonant frequencies form spectral peaks, or formant frequencies, which are perceived as the timbre of the voice (Fant, 1960; Titze, 1994). The vocal tract can be approximated to a uniform tube shape, enabling formant frequency estimation using the following formulae:

With one end of the tube open: $F_i = \frac{(2i-1)c}{4VTL}$

With both ends of the tube closed: $F_i = \frac{ic}{2VTL}$

Where $i$ is the formant number, $c$ is the speed of sound in the air (350m $s^{-1}$), $VTL$ is the vocal tract length (in m), and $F_i$ is the frequency of the $i$th formant (in Hz) (Titze, 1994). However, the vocal tract is not a uniform tube shape and the frequencies of individual formants can be actively controlled by altering the size and shape (and therefore the resonant properties) of the vocal tract (Fitch & Hauser, 2003), through relative changes to the positioning of the pharynx, velum, tongue and lips. These articulatory movements particularly effect the positioning of the lower formants (Fant, 1960). Therefore,

determining the average spacing between multiple successive formant frequencies (see Reby & McComb, 2003, for a detailed methodology), provides a more precise estimate of the vocal tract length than individual *Fi* (Fitch, 1997) or the *formant dispersion* as defined by Fitch (1997) (see Pisanski et al., 2014, for a review). Independently of the end conditions, the frequency difference between any successive formants, called formant spacing ($\Delta F$), can be given by the formula (Titze, 1994):

$$\Delta F = F_{i+1} - F_i = \frac{c}{2VTL}$$

To summarise, the formant frequencies, or more specifically their average spacing, provide direct information about the length of the tube. In accordance with these formulae, longer vocal tracts produce lower and more closely spaced formants.

The source- and filter-related acoustic properties of mammalian voices are therefore constrained by the relationship between the anatomy of the vocal apparatus and fundamental laws of physics. This results in a certain degree of predictability in the F0 and formant frequencies as a function of the size and shape of the vocal folds and vocal tract respectively (Fitch, 1997; Fitch & Reby, 2001). If there is a strong mapping between the source or filter components of the vocal apparatus and other physical characteristics of the signaller, such as their age or sex, the acoustic properties of the voice have the potential to broadcast accurate information about those characteristics, and are referred to as 'indexical' cues (Ghazanfar et al., 2007).

## Indexical Cues in the Human Voice

Accurate 'acoustic allometry' (the relationship between specific vocal parameters and an organism's size) should be widely conserved, providing a 'cheap' cue to size, in the absence of active selection against it (Fitch, 2000a). In humans, the size of the vocal folds and the length of the vocal tract do broadly co-vary with body size across different age and sex categories (Fitch & Giedd, 1999). Before the onset of puberty, there is a linear correlation between body size and the length of the vocal folds and the vocal tract in both sexes (Titze, 1994; Vorperian et al., 2009; Vorperian et al., 2011). However, because the larynx is cartilaginous and only loosely attached to the skull base, it is not strongly constrained by the size of the surrounding skeletal structures (Fitch, 2000a). This enables it to grow out of proportion to other body parts, facilitating selection for

size adaptations away from its basic scaling with the rest of the body (Fitch, 2000a). Indeed, the growth rate of the human male larynx has almost certainly undergone sexual selection, creating sexual dimorphism in the size of the larynx in adults. Specifically, the substantial increase in male androgen levels during puberty causes the male larynx to become permanently enlarged relative to the female larynx (Kahane, 1982). As a result, the vocal folds of adult men are twice as long as women's vocal folds (Kent & Vorperian, 1995). Corresponding with the anatomical growth of the larynx, the F0 of the voice steadily decreases during development until puberty (Titze, 1994), after which the pubertal androgen-related effects on the larynx result in an average adult male F0 which is 50-80% lower than the adult female F0 (Hollien, Green & Massey, 1994). Therefore the average F0 is lowest in men at around 100 Hz, doubled in women at approximately 200 Hz, and higher still in children at around 260 Hz (Huber, Stathopoulos, Curione, Ash & Johnson, 1999; Lieberman, 1988). However, while the lack of tight constraints on laryngeal growth has allowed the evolution of large categorical differences between the F0s of children, women and men, it also means that the relationship between F0 and body-size breaks down within members of the same age and sex categories, and a recent meta-analysis determined that in adult humans the F0 accounts for less than 2% of the variance in body height or weight (Pisanski et al., 2014).

Pubertal androgen levels also affect the length of the vocal tract. Humans represent an unusual case amongst mammals (but not unique e.g. Fitch & Reby, 2001), as although the larynx is situated in the standard mammalian position at the back of the oral cavity at birth, between three months and three years of age the larynx moves further down into the throat, elongating the vocal tract and producing a permanently 'descended larynx' (Crelin, 1987; Laitman & Crelin, 1976; Lieberman, 1984; Negus, 1949). During puberty, a 7% greater increase in height in males (Gaulin & Boster, 1985) is coupled with a male-specific secondary descent of the larynx to a lower position in the vocal tract (Fitch & Giedd, 1999). As a consequence of these changes, adult male vocal tracts are on average 15-20% longer than in adult females (Fant, 1960). The formant frequencies directly reflect the anatomical differences in the vocal tract length, as men's formants are around 15% lower than women's formants, and both are considerably lower than those of children's voices (Bachorowski & Owren, 1999; Hillenbrand & Clark, 2009). However, although the length of the vocal tract is more tightly constrained

by the surrounding skeletal structures than the larynx, and has been found to account for around 70% of the variance in height and weight for men, women and children individually (Fitch & Giedd, 1999), Pisanski et al.'s (2014) meta-analysis indicated that formant-based measures of the vocal tract length only accounted for around 10% of the variance in the height and weight of adult men and women. It has been suggested that this apparent discrepancy between the formant-based vocal measures and body size could be related to the high level of vocal tract modulation during speech production, which may obscure the allometric relationship between the formant positioning and body size of the speaker (Collins, 2000; Gonzalez, 2004).

To summarise, in accordance with the predictions of the source-filter theory, the size differences in the vocal apparatus between human children and adults, and between adult men and women, generate acoustic differences in the F0 and formant spacing that provide a reliable indication of the categorical size/age and sex of the speaker. Added to these broad differences between speakers, idiosyncratic variation in the precise morphology of the vocal apparatus also causes speakers to have a distinct vocal signature, distinguishing between human voices at an individual level. We will now discuss the perception of these indexical attributes in the voice by both humans and dogs.

## Perception of Human Gender

Although there are a number of additional acoustic differences between adult male and female voices (e.g. women's voices are more breathy and less monotonous than men's voices Assmann, Dembling & Nearey, 2006; Simpson, 2009), the combined effects of the F0 and the formants can classify the sex of adult voices with 98.8% accuracy (Bachorowski & Owren, 1999). It is therefore not surprising that humans rely on the anatomically related differences in the F0 and formants of the voice to judge the gender of unfamiliar speakers (e.g. Smith & Patterson, 2005), where lower-pitched (lower F0) voices with a lower resonance (lower formants with narrower spacing) are perceived to belong to adult men (Hillenbrand & Clark, 2009). The F0 has generally been reported to provide the strongest acoustic cue to the sex of the speaker (e.g. Hillenbrand & Clark, 2009; Lass, Hughes, Bowyer, Waters & Bourne, 1976), which probably reflects the greater level of dimorphism in the F0 than in the formant frequencies of adult voices. Because human faces are also sexually dimorphic (Burton, Bruce & Dench, 1993),

humans develop the ability to associate unfamiliar voices with gender-matching faces at around four months of age (Walker-Andrews, Bahrick, Raglioni & Diaz, 1991).

In contrast to humans and other mammal species that have a highly sexually dimorphic vocal apparatus (e.g. chacma baboons *Papio hamadrayas ursinus*: Rendall, Owren, Weerts & Hienz, 2004; red deer *Cervus elaphus*: Reby & McComb, 2003), dogs have a comparatively low level of anatomical sexual dimorphism (Plotsky, Rendall, Riede & Chase, 2013), which coupled with the exceptionally high degree of morphological variation across breeds (Vilà, Maldonado & Wayne, 1999), means that dog vocalisations do not appear to encode reliable information about the caller's sex (Riede & Fitch, 1999; Taylor, Reby & McComb, 2008). However, dogs can be trained to discriminate between the average F0s and formant frequencies of human male and female vowel sounds (Baru, 1975), indicating that they are perceptually capable of discriminating between the voices of men and women. Humans may also unintentionally exaggerate gender-related differences in their voices when speaking to dogs, as women are more likely to use the 'pet-directed speech register' than men, which is characterised by a higher F0 and larger F0 range than adult-directed speech (Prato-Previde, Fallani & Valsecchi, 2005). Gender-specific differences in attitudes and behaviour towards dogs (e.g. Mariti et al., 2012) may have made human gender discrimination a naturally relevant ability, and indeed, dogs do show different behavioural responses towards unfamiliar humans depending on their gender (Lore & Eisenberg, 1986; Wells & Hepper, 1999). However, although dogs appear to be able to discriminate between men and women on some level, the sensory cues which they use to do so have not yet been determined. To my knowledge, no mammalian species other than humans have been observed to associate cues related to the sex of an unfamiliar individual across different sensory modalities. Therefore I will address the question of whether dogs spontaneously discriminate gender cues in the human voice, and if they are able to use this information to correctly match voices to unfamiliar men and women (Chapter 3).

### Estimation of Age and Body Size

As well as relying on anatomically related differences in the F0 and formant frequencies to judge the sex of adult voices, human listeners also use these cues to distinguish between adult and child speakers. Low-pitched voices with low resonances are judged

to belong to adult males, whilst high-pitched voices with high resonances are judged to belong to pre-pubescent girls, although there can be some confusion between the voices of pre-pubescent boys and adult women (Smith & Patterson, 2005; Smith, Walters & Patterson, 2007). When the difference in height between two speakers is relatively large, as occurs between different age and sex categories, human listeners can judge the taller individual from the values of the F0 and formants in their voice at around 90% accuracy (Rendall, Vokey & Nemeth, 2007). Non-human primates are also capable of discriminating between different age categories, as Rhesus macaques *Macaca mulatta* spontaneously associated the faces of juvenile conspecifics with vocalisations typical of juveniles, and adult conspecific faces with vocalisations typical of mature adults (Ghazanfar et al., 2007). In Chapter 4, I investigate whether dogs show a comparable ability to classify the age of unfamiliar humans, by determining if they spontaneously match adult human voices to adults and child voices to children.

Although humans accurately determine age and sex related differences in body size by attending to anatomically constrained acoustic cues, within the same age and sex categories human listeners' accuracy in judging the taller speaker drops to around 60% for adult voices (Rendall et al., 2007), which reflects the relatively poor correlation between the F0 and formants with body size across speakers within the same age and sex category. Observations that human listeners associate lower formants with larger body sizes in adult men and women separately are not unexpected given that they are weakly correlated, but what is surprising is that listeners actually rely more strongly on the F0 of the voice to estimate body-size within age and sex classes than the formants (Collins, 2000; González, 2003; Rendall et al., 2007), despite the fact that there is no correlation between the F0 and the size of the speaker (Pisanski et al., 2014). It has been suggested that human listeners may associate voice pitch with size because it is perceptually more salient than the resonances, as when both variables are manipulated by perceptually matching amounts, listeners instead prioritise the positioning of the formants over the F0 (Pisanski & Rendall, 2011). Alternatively, it has been proposed that humans over-generalise other correlations between sound frequencies and size, either because of the correspondence between F0 and size in the voice across age and sex categories, or due to more general mappings between auditory pitch and size in the environment (Rendall et al., 2007). Indeed, humans appear to develop a natural tendency to associate simple lower-pitched tones with larger geometric shapes from

around 6-months of age (Fernández-Prieto, Navarra & Pons, 2015). Although it has been theorised that animals also relate lower-pitched vocalisations to larger body-sizes (Morton, 1977), it is not currently known whether any non-human species similarly perceive a general correspondence between lower pitch and larger size, or whether a broad perceptual mapping between auditory pitch and size could be involved in associating vocalisations with signallers. Unlike most other mammals (Taylor & Reby, 2010), the high level of morphological variation in dogs means that the F0 in their vocalisations does actually correlate, albeit weakly, with body size, accounting for 9% of the variance in body weight across breeds (Taylor et al., 2008). Dogs may therefore gain sufficient exposure to the correlation between auditory pitch and visual size to develop the perception of a cross-modal association between these dimensions. I will explore whether dogs do express this form of cross-modal correspondence in relation to their perception of human voices in Chapter 4.

Although it is not known whether vocal pitch influences the assessments that dogs make about conspecific vocalisations, it has been demonstrated that similarly to many other mammals (e.g. red deer: Reby et al., 2005; koalas *Phascolarctos cinereus*: Charlton, Ellis, Brumm, Nilsson & Fitch, 2012; rhesus macaques: Ghazanfar et al., 2007), dogs do perceive the formant spacing in species-specific vocalisations and they are able to use this information to match growls to the appropriately sized signaller. Using a preferential looking paradigm, Taylor, Reby and McComb (2011) visually presented dogs with two differently sized taxidermy models of dogs. The subjects then heard the sound of a conspecific growl that had been artificially manipulated to have either wider or narrower formant spacing. When dogs heard the growl with the wider formant spacing (typical of a smaller sized dog) they looked longer at the smaller model, whilst they looked more towards the larger model when they heard the growl with the smaller formant spacing. Their behavioural responses indicated that they were able to cross-modally associate size information about conspecifics based on the formant positioning in the vocal signals. Dogs are able to rely on the formants to judge size in conspecific vocalisations because their larynx is situated at the back of the oral cavity, so their vocal tract length is constrained to match the length of their skull, producing a strong correlation between the vocal tract length and body size (Riede & Fitch, 1999; Plotsky et al., 2013). Unlike humans, dogs are not known to modulate the shape of their vocal tracts when vocalising (Fitch, 2000b) and the formant spacing remains static,

accounting for around 62% of the variance in their body weight (Taylor et al., 2008). It is possible that dogs' ability to cross-modally match size information in conspecific vocalisations may also extend to their perception of human voices, mirroring the way that humans are able to judge the size of dogs from their growls (Taylor et al., 2008). I investigate if dogs do assess size information in human voices, by determining whether they spontaneously associate unfamiliar voices with adult or child speakers based on cues related to their body size in Chapter 4.

## Individual Recognition

While there have been no previous studies exploring whether dogs perceive functionally relevant information relating the indexical attributes of unfamiliar human voices, it has been demonstrated that dogs recognise the voice of their owner. The co-variation of idiosyncratic characteristics in the vocal apparatus means that humans have individually distinct voices (Bachorowski & Owren, 1999), and both the F0 and the formants influence speaker recognition (Lavner et al., 2000). Humans learn to discriminate between familiar and unfamiliar voices *in utero*, differentially responding to the voice of their mother compared to the voice of a stranger (Kisilevsky et al., 2003), and are able to match voices with other unique traits of familiar individuals, such as their facial features, from infancy (Bahrick, Hernandez-Reif & Flom, 2005). The cross-modal recognition of familiar humans has also been demonstrated in both domestic horses *Equus caballus* (Proops & McComb, 2012) and captive rhesus macaques (Sliwa, Duhamel, Pascalis & Wirth, 2011). Adachi, Kuwahata and Fujita (2007) tested whether dogs similarly learnt to associate the voice of their owner with their owner's face using a violation of expectation paradigm. This methodology tests whether the subject detects any invariance in an unfolding sequence of events, which is evidenced by greater attentional capture relative to an expected event sequence (Baillargeon, Spelke & Wasserman, 1985). In Adachi et al.'s (2007) study, each dog was first presented with a voice recording of either their owner or a stranger of the same gender, saying the subject's name. After hearing the voice, an image of either their owner's face or the stranger's face appeared on a screen in front of the dog. The results showed that the subjects looked longer at the image when it did not match the preceding voice, suggesting that when the dogs heard their owner's voice, they expected to see their owner rather than the stranger, but not when they heard the stranger's voice. This therefore indicated that dogs recognise their owner's voice, and that this recognition

may prime the dogs to expect to see their owner. However, Kriengwatana, Escudero and Cate (2014) outlined two methodological limitations to the study which restrict the strength of these results. Firstly, because the auditory stimulus used was the subject's own name, it is possible that the dogs merely associated the sound of their name spoken in a specific way with their owner's face, rather than their owner's voice independently of the spoken phrase. Secondly, because the dogs were tasked with discriminating between their owner and an unfamiliar person, rather than between two familiar people, it is not clear if dogs can recognise individual people, or if they merely discriminate between familiar and unfamiliar people. Further investigations are thus needed to address these issues before it can be established whether dogs show cross-modal recognition of their owner.

Therefore, to summarise this section, although dogs appear to be capable of perceiving the main acoustic cues that encode indexical information about the physical characteristics of human speakers, it is not known whether dogs spontaneously make functional assessments about human speakers on the basis of these cues. In this thesis I aim to address whether dogs do perceive indexical cues in unfamiliar human voices, by investigating if they are able to associate voices with people according to their gender and age categories.

## Dynamic Cues in the Human Voice

Human voices not only encode anatomically constrained indexical information about the speaker, but also transmit dynamically controlled vocal prosody. At an acoustic production level, the F0 can be manipulated by musculature control within the larynx (Titze, 1994) and through changes in the subglottal pressure, whilst the formant spacing can be modified through vocal tract elongation and pharyngeal constriction (Briefer, 2012). Other vocal parameters such as the amplitude, rate and duration are also controlled by the speaker's respiratory rate. In addition to clarifying the speaker's intentions (e.g. statements versus questions), these vocal cues can also be reliably related to emotional states. Different emotions cause predictable variation in the relative height and modulation of the F0 as well as the voice quality and formant values across speakers, independently of cultural influences, and are universally recognised by human listeners (Sauter, Eisner, Ekman & Scott, 2010). Dogs can also discriminate between different prosodic cues in human voices, as the tone of voice used by an experimenter

has been shown to influence dogs' responses in obedience situations (Mills, Fukuzawa & Cooper, 2005) and search tasks (Scheider et al., 2011). Dogs also appear to be able to use human emotional expressions as socially referential signals, expressing some understanding that the expressions can be directed towards external stimuli in the environment. For example, dogs have been shown to either approach or avoid an unfamiliar object depending on whether their owner reacted positively or negatively towards the object, although they did not rely on the human's expressions to the same extent when the informant was a stranger (Merola, Prato-Previde & Marshall-Pescini, 2012). However, rather than showing an insightful understanding of the referential nature of the emotional expressions of happiness and fear, Yong and Ruffman (2015) demonstrated that dogs appear to simply become confused by the relatively unfamiliar expression of fear and subsequently attend more to the person rather than exploring the environment. To further address whether dogs actually show referential understanding of human emotional expressions, Buttelman and Tomasello (2013) instead used a preferential choice paradigm where dogs were given the choice of between two boxes (only one of which contained a reward) after they had seen the experimenter's emotional reactions to each box. They observed that the dogs were more likely to choose the box that they had seen an experimenter look inside with a happy expression (using both facial and vocal cues) rather than a disgusted expression. The results suggested that the dogs were able to determine both the emotional valence and the directedness of the emotional expressions. However, this ability appeared to be limited to differentiating expressions that strongly differed in valence, as dogs were unable to choose between boxes paired with either happy or neutral expressions. In a follow-up study, Merola, Prato-Previde, Lazzaroni and Marshall-Pescini (2014) determined that dogs were able to correctly choose the box matched with the happy rather than the neutral expression, as well as correctly differentiating between happy and fearful expressions, when the informant was their owner rather than a stranger. However, they struggled to distinguish between the fearful and neutral expressions of their owner, and between the happy and fearful expressions of a stranger. This led Merola et al. (2014) to conclude that dogs learn to associate their owner's positive emotional expressions with positive outcomes, whereas other emotional expressions do not appear to be as clearly recognised. Therefore it appears that dogs learn to respond appropriately to certain human emotional expressions through experience, as they are more likely to gain exposure to their owner's expressions of happiness rather than fear or disgust in natural

learning contexts. Unfortunately, because a combination of visual and vocal cues was used by the informant across these studies, the extent to which the dogs used the vocal cues to discriminate between the emotional expressions cannot be determined.

Although studies investigating how dogs respond to human emotional expressions suggest that they may need to learn to recognise the intended valence of different signals, both dog owners and non dog owners tend to agree that dogs are empathically sensitive to human emotions (Vitulli, 2006). This general assumption recently received empirical support through the demonstration that dogs showed a significant increase in salivary cortisol levels after they had been exposed to audio recordings of human babies crying, but not when they were exposed to babies babbling or to white noise (Yong & Ruffman, 2014). The dogs' hormonal response to the crying was also accompanied by an increase in alertness and submissive behaviour, implying that the dogs perceived the crying sound as aversive. Because dogs that were not experienced with babies showed the same response as those that were, the potential influence of previous reinforcement was ruled out as a possible explanation for their reactions. Instead, as the specific increase in the dogs' attention and cortisol levels in response to the crying matched the hormonal and behavioural responses of the human listeners, the results were interpreted as a demonstration of 'emotional contagion' in dogs, which can be defined as a non-insightful form of empathy. Although it is difficult to say exactly how the dogs interpreted the crying sound, it is possible that emotional contagion could occur due to continuities in the acoustic encoding of arousal and/or valence across different mammalian species' vocalisations (Morton, 1977; Ohala, 1994), which is particularly apparent in infant distress calls (Lingle, Wyman, Kotrba, Teichroeb & Romanow, 2012; Lingle & Riede, 2014). Therefore, although dogs may need to learn to respond to different human emotional expressions, it is possible that some intense emotional vocalisations may cause 'affect-induction' (Owren & Rendall, 1997; 2001), influencing the physiological response of the dog and potentially creating an innate empathic response.

While further research is still needed to determine the extent to which dogs are able to make functional assessments about human emotional expressions, different expressions do appear to modulate their behavioural and physiological responses. Although there is currently only limited evidence specifically relating to vocal emotional expressions,

dogs have shown sensitivity to the tone of voice used in instructional contexts, suggesting that they may discriminate emotional prosody in human speech.

## Phonemic Information in Speech

Though most mammals show some degree of flexible manipulation of their vocal apparatus during vocal production, varying the acoustic structure of their vocalisations (see Taylor & Reby, 2010 and Briefer, 2012, for recent reviews), only humans appear to exercise the fine level of motor control required for speech production. Through precise movements of the articulators, including the tongue, jaw and lips, humans intentionally alter the shape of their vocal tract during speech production. This causes dynamic variation in the positioning of the lower formant frequencies, which creates phonological structure in the vocal signal (Fant, 1960). Vowel sounds depend on the placing of the tongue and can be differentiated by the frequencies and spacing of the first two formants, whereas consonants are formed through energy bursts created by different articulatory constriction gestures (Liberman, 1957). In contrast to humans, but similarly to other mammals studied, when dogs vocalise their tongue remains in a static position and thus no formant variation has been observed within their own species-specific vocalisations (Fitch, 2000b). However, humans frequently use verbal signals to elicit specific responses in dogs, making the ability to perceive formant patterning potentially relevant for dogs to learn to perform appropriate responses (Fukuzawa, Mills & Cooper, 2005). I have already discussed that dogs are perceptually aware of formant scaling, allowing them to assess the size of the signaller from conspecific vocalisations, and they also appear capable of perceiving the dynamically controlled formant positioning used to create different vowel sounds. Baru (1975) demonstrated this ability by training dogs to discriminate between two synthetic vowel sounds. The dogs performed equally well with vowels containing only the first two formants (as opposed to the first four formants), suggesting that similarly to humans, they are able to discriminate vowels on the basis of the first two formant positions. Evidence that dogs may recognise human verbal commands by using phonetic cues was obtained in a study where dogs were first trained to reliably respond to two tape recorded spoken commands: ‘sit’ and ‘come’, after which their responses were measured when part of the phonetic content of each command was changed (Fukuzawa et al., 2005). Independently of which phoneme was altered in either command, the dogs’ recognition scores were significantly reduced, shown by a marked decline in performance. For the

'sit' command, changing the first consonant (to 'chit') had a significantly stronger effect than the last consonant (to 'sik'), whilst changing the vowel sound (to 'sat') had an intermediate effect. In contrast, there was no difference in recognition between the three altered versions the command 'come'. There was also a high degree of individual variability between the dogs across all of the alterations, suggesting that each of the subjects may have relied on different phonemes to recognise the commands. However, whilst these results suggest that dogs do perceive the phonetic content in spoken words, it is not clear whether any other potential differences between the original and altered commands were controlled, such as the F0, amplitude or duration. Therefore, it remains possible that dogs actually used cues other than the phonemic content to recognise the commands. Although, Gibson et al. (2014) determined that dogs' responses to human scolding vocalisations were reduced by an equivalent amount when the voices were manipulated so that either the formants were static (creating unintelligible vocalisations with negative emotional prosody) or the human voice qualities were removed by using computer generated speech (creating intelligible vocalisations with no emotional prosody). The equal decrease in performance when the voices lacked either phonemic cues or emotional prosodic cues suggests that dogs may pay equivalent attention to both elements in the voice. However, the lack of additional control stimuli in this study means that it is difficult to rule out the possibility that the re-synthesised voices merely sounded like novel signals to the dogs instead of familiar signals lacking certain informative content. Therefore, I will further explore how dogs respond to variations in the phonemic content of human speech in Chapter 5.

Although it is not yet known whether dogs perceive the phonetic cues in spoken words, it is clear that they can learn an extensive number of verbal utterances relating to specific behavioural responses through training, showing comparable performances to other language trained animals (e.g. bottlenose dolphins: Herman, Richards & Wolz, 1984; sea lions *Zalophus californianus*: Schusterman & Krieger, 1986; pygmy chimpanzees *Pan paniscus*: Savage-Rumbaugh, McDonald, Sevcik, Hopkins & Rubert, 1986; African grey parrots *Psittacus erithacus*: Pepperberg, 1981). Dogs may even develop some understanding of object word referents, as their apparent ability to rapidly learn new labels was first demonstrated by Kaminski, Call and Fischer (2004). In a single case study, it was demonstrated that a border collie could be given a verbal instruction by his owner to retrieve a specific toy (e.g. 'fetch teddy') from a different

room containing an array of familiar toys, successfully responding to the unique labels of around 200 different objects. This dog also showed evidence of learning new object labels on their first presentation, as when asked to fetch a new toy from a familiar array, he excluded the familiar toys and retrieved the new toy. Furthermore, he appeared to still retain knowledge of the new label one month after testing. The rapid acquisition of new label indicated that rather than simply learning to associate the sound of the utterance with the object, dogs could be capable of 'fast mapping', showing an understanding that human utterances may be used to refer to something in the external environment. Subsequent modifications to this paradigm used different combinations of behavioural responses and object labels (e.g. 'take ball' versus 'paw ball') (Pilley & Reid, 2011; Ramos & Ades, 2012) to successfully exclude Bloom's (2004) suggestion that dogs could have represented the command as a single proposition without understanding that the label for the object was independent from the action of retrieving it. However, Markman and Abelev (2004) also raised a second point, stating that dogs may not necessarily understand the referential nature of human words. Instead, they proposed that dogs could have successfully passed the novel object label tests used by Kaminski et al. (2004) through a combination of neophilia and an extended form of exclusion learning, rather than showing evidence of fast mapping. This hypothesis was supported by evidence obtained from a different dog, which similarly passed the tests used in the original paradigm, but failed to retrieve the new object when it was not paired with other familiar objects, demonstrating that she had not learnt the new object label (Griebel & Oller, 2012). Tempelmann, Kaminski and Tomasello (2014) thus conducted a series of experiments which were designed to determine if dogs could learn human object labels through an understanding of their referential nature instead of the more simple process of associative learning. Dogs with previous experience in learning object names took part in these experiments, which tested whether they were able to use the referential intent of a human demonstrator to form word-object associations without the aid of spatial-temporal congruency. In one of these tests, dogs were tasked with inhibiting the formation of a spatial-temporal association between a verbal label and an object. Specifically, one object was placed in the dog's view whilst another object was hidden, and the human verbally labelled one of the objects. This was then followed by a retrieval phase with both objects. Only one dog showed any evidence of correctly learning the labels, although this ability only appeared to develop in the latter half of the trials. None of the other subjects showed any evidence of learning the labels; therefore it

remains unclear whether dogs can learn object labels through human referential communication alone. Interestingly, the dogs that failed at the task did not show a preference for retrieving the visible object, which was spatio-temporally congruent with the verbal label. One possible explanation for this lack of association is that dogs may be more reliant on non-visual cues when learning to match verbal labels with objects, as the owners stated that their usual method of teaching new labels was to repeat the word whilst allowing the dog to play with the object. Indeed, dogs appear to be more likely to generalise known verbal labels to new shapes if they match the original in size or texture, rather than shape (Van der Zee, Zulch & Mills, 2012), and these cues are more readily available when dogs manipulate objects. Therefore, whilst dogs are capable of learning to match a large number of human utterances with physical objects, there is no current consensus on how they learn to make these associations, or which cues they use from either the voice or the objects to do so. Article I of Chapter 4 provides a detailed review of the potential processing mechanisms involved in the formation of multi-sensory associations in mammals, which may be profitably applied in future investigations of object-label learning in dogs.

## Human Voice Processing

So far we have seen that the strongest evidence obtained for dogs' understanding of specific human vocal features relates to their perception of the emotional prosody, both in speech and non-verbal vocalisations. Although less conclusively demonstrated, their responses suggest that they may also perceive indexical information related to the identity of the speaker, and could conceivably attend to some of the basic phonemic content in learnt verbal commands. It is well established that in humans, these main functional aspects of the vocal signal are processed through partially dissociable neural networks (Scott & Johnsrude, 2003). For most people, the linguistic information present in speech is primarily processed in the left hemisphere of the brain (e.g. Friederici, 2002; Scott & Johnsrude, 2003), whilst the processing of emotional content in the voice is associated with stronger activity in the right hemisphere (e.g. Wildgruber et al., 2005). More specifically, neuroimaging studies have demonstrated dominant left hemispheric activation for processing intelligible segmental (the units or segments that make up the linguistic content) information in speech relative to acoustically matched control stimuli, including phonologically relevant cues (Jacquemot, Pallier, LeBihan, Dehaene & Dupoux, 2003), individual phonemes (Agnew, McGettigan & Scott, 2011),

syllables (Liebenthal, Binder, Spitzer, Possing & Medler, 2005), words (Mummery, Ashburner, Scott & Wise, 1999) and sentences (Scott, Blank, Rosen & Wise, 2000; McGettigan et al., 2012; Narain et al., 2003). Further specialisations of the left hemisphere in relation to speech perception include processing semantic (stored lexical meaning) (Obleser & Kotz, 2009; Obleser, Wise, Dresner & Scott, 2007) and syntactic (grammatical) information (Friederici, 2002; Friederici, Kotz, Scott & Obleser, 2010; Herrmann et al., 2012). In contrast, the right hemisphere appears to be more specialised in processing supra-segmental (continuous across several segmental units) information, including the emotional prosody (e.g. Buchanan et al., 2000; Gandour et al., 2003) and speaker-related indexical cues encoding identity and gender (e.g. Belin & Zatorre, 2003; Lattner, Meyer & Friederici, 2005; von Kreigstein, Eger, Kleinschmidt & Giraud, 2003). In recent models of human vocal perception, the anterior auditory processing pathway, stretching from the auditory cortex to the lateral superior temporal gyrus and sulcus, is thought to represent a 'what' (as opposed to the 'where') pathway for sound identification (Belin et al., 2004; Scott & Johnsrude 2003; Schirmer & Kotz, 2006). Rauschecker and Scott (2009) highlighted similarities in the anatomical and functional organisation of the auditory cortical system between humans and non-human primates, pointing out that although this system supports speech perception in humans, it is likely to be phylogenetically older, meaning that the neural organisation and representation of language in humans has probably been evolutionary constrained by pre-existing mechanisms of vocal processing.

Support for the gradual evolution of left hemispheric language specialisation in humans from pre-existing mammalian mechanisms involved in vocal processing comes from observations that left hemispheric dominance is wide-spread across many different mammals for the perception of species-typical vocalisations (Ocklenburg, Ströckens & Güntürkün, 2013). The most commonly used paradigm to assess hemispheric lateralisation in response to auditory signals in non-human animals has been the behavioural 'head-turn paradigm' (Hauser & Andersson, 1994). Before outlining the details of this paradigm, it is necessary to first provide some contextual background from the human literature. As well as directly evidencing differential hemispheric activation in response to speech, humans also show perceptual asymmetries which have been measured primarily using the 'dichotic listening' technique (Kimura, 1961). Using this procedure, two different sounds are simultaneously presented to each of the

participant's ears, and the participant is usually instructed to report the sound heard most clearly. Most people show a right-ear advantage for reporting linguistic information in speech sounds, and a left-ear advantage for identifying the emotional content (Kimura, 2011). These relative ear advantages have been explained by the fact that although the auditory input from each ear is projected to both of the auditory cortices, the contralateral projections are stronger, which may block or slow the processing of the ipsilateral projections (Bocca, Calearo, Cassinari & Migliavacca, 1955). Therefore, input from the right-ear is mainly transmitted to the left hemisphere, which is more specialised in processing linguistic content, whilst the left-ear input is mainly projected to the right hemisphere, which is more specialised in processing the emotional prosody. Because the predominance of the contralateral auditory pathways appears to be a shared mammalian trait (e.g. dogs: Tunturi, 1946), a similar behavioural paradigm was adapted to test for comparable ear advantages for perceiving conspecific vocalisations in non-human animals. In the head-turn paradigm, a single sound is presented to both of the animal's ears at the same time, by either playing the sound from one loudspeaker positioned directly behind the animal, or from two loudspeakers placed at equal distances to either side of the animal. Similarly to the principle behind the dichotic listening paradigm in humans, the ear that the animal leads with when they orient towards the sound is taken as a behavioural indication of better perception through that ear, and thus dominant processing in the contralateral brain hemisphere. Using this paradigm, the majority of species tested have shown significant right head turning biases in response to conspecific vocalisations, and either left or no orienting biases in response to non-species specific sounds, which have included heterospecific vocalisations, salient environmental sounds and conspecific vocalisations manipulated to fall outside of the species typical range (e.g. rhesus macaques: Ghazanfar, Smith-Rohrberg & Hauser 2001; Hauser & Andersson, 1994; Hauser, Agnetta & Perez, 1998; California sea lions: Böye, Güntürkün & Vauclair, 2005; dogs: Siniscalchi, Quaranta & Rogers, 2008; Siniscalchi, Lusito, Sasso & Quaranta, 2012). However, it should be noted that more confused or contradictory results have also been obtained for other mammal species (see Teufel, Ghazanfar & Fischer, 2010, for a review). Despite some disparaging results, the general left hemispheric bias for processing conspecific vocalisations in mammals suggested by their orienting responses has also been supported by alternative techniques including neuroimaging (rhesus macaques: Joly, Ramus, Pressnitzer, Vanduffel & Orban, 2012; Petkov et al., 2008; Poremba et al.,

2004) and psychophysical studies (rhesus macaques: LePrell, Hauser & Moody, 2002; Japanese macaques *Macaca fuscata*: Petersen, Beecher, Moody & Stebbins, 1978), as well as through more invasive ear plugging (house mice *Mus musculus*: Ehret, 1987) and brain lesioning procedures (Japanese macaques: Heffner & Heffner, 1984; 1986).

Two studies have specifically investigated dogs' orienting responses to conspecific vocalisations. In line with other mammalian species, dogs have been observed to show a right orienting bias when presented with conspecific vocalisations, indicating stronger left hemispheric processing, and a left orienting bias in response to the sound of thunder (Siniscalchi et al., 2008) and temporally reversed conspecific vocalisations (Siniscalchi et al., 2012), indicating stronger right hemispheric processing. These results suggest that the left hemisphere of the dog brain may be specialised in processing conspecific vocalisations, and that similarly to human responses to speech, this specialisation may depend on the functionality of the vocalisations rather than purely their acoustic structure. Interestingly, two dogs that were noted to be extremely fearful during testing showed an opposite response pattern to the other subjects, by consistently turning to their left when presented with conspecific vocalisations (Siniscalchi et al., 2008). Although no firm conclusions can be made due to the small number of fearful dogs, the association between sounds that are perceived to be negative (as dogs are also generally fearful of thunderstorm sounds; Siniscalchi et al., 2012) and greater right hemispheric involvement, suggests that the right hemisphere could be more strongly activated when sounds are perceived to be highly emotionally salient. More detailed evidence for dogs' hemispheric responses to different sounds was obtained through a recent fMRI study, which aimed to determine how dogs process species-specific vocalisations in comparison to human emotional vocalisations (e.g. laughter or crying) and other environmental sounds (Andics, Gácsi, Faragó, Kis & Miklósi, 2014). To do this, the researchers measured the responses of the auditory regions in awake dogs whilst they passively listened to recordings of these sounds. It was determined that similarly to humans, dogs have a specific voice area that responds preferentially to dog vocalisations, which consists of a ventral region close to the temporal pole bilaterally and a left dorsal auditory region. In contrast, no regions were observed to show preferential responses to human vocalisations. However, an auditory region in the right caudal ectosylvian gyri was sensitive to the emotional valence of both dog and human vocalisations. This response was specific to the right hemisphere of the brain, indicating

that dogs show a right hemispheric dominance for processing emotional content in vocal signals, independently of whether they are produced by dogs or humans. Taken together with the behavioural results obtained through the head-turn paradigm, in dogs the left hemisphere appears to be more specialised in processing conspecific vocalisations, whilst the right hemisphere is more responsive to emotionally salient information present in relevant auditory signals, including human vocalisations. However, although human speech is also an important signal for dogs, no studies have examined how dogs process the different communicatory components of speech signals. To address this, Chapter 5 investigates whether dogs show evidence of hemispheric asymmetries in response to different functional information in human speech.

## Research Questions and Thesis Outline

To summarise, while it is already apparent that dogs pay attention to the emotional content in speech and non-verbal vocalisations, their ability to assess indexical and/or phonemic information in human speech has received comparatively little attention from previous research. Developing a clearer understanding of how dogs perceive human vocal signals is important to inform current perspectives on the occurrence of socio-cognitive specialisations in this species which may facilitate intra-specific communication with humans. Therefore, the first main aim of this thesis is to establish whether dogs are capable of perceiving information related to the three main components of the human voice, or if they predominantly rely on the emotional prosody of human vocal signals to formulate their responses. Additionally, the current body of work aims to explore the perceptual mechanisms involved in dogs' perception of human vocal signals, by investigating how dogs associate human vocal cues with individual signallers, and whether they show evidence of differentially processing the individual communicative components of human speech signals.

Specifically, I aim to answer the following questions relating to a) dogs' functional assessments of indexical cues in human voices, and b) cerebral asymmetries when dogs process human speech:

1) Are dogs spontaneously capable of cross-modal human gender discrimination? (Chapter 3)
2) How do non-human animals form cross-modal associations during their perception of multisensory signals? (Chapter 4 – Article I)

3) Are dogs spontaneously capable of the cross-modal discrimination of human age categories, and if so, how do they associate age-related auditory and visual cues? (Chapter 4 – Article II)
4) Do dogs show evidence of hemispheric asymmetries when processing the main communicative components of human speech, and if so, are asymmetries dependent on the acoustic structure of the signals or their functional content? (Chapter 5)

Chapter 2 outlines the materials and methods used to collect the data for the thesis.

Chapter 3 addresses Question 1, by testing dogs' ability to cross-modally match unfamiliar human voices with people according to their gender, using a preferential looking paradigm. Because human gender appears to be relevant to dogs when they assess unfamiliar people, I predict that dogs will express the ability to spontaneously associate human voices with an adult person according to their gender, by looking more quickly and for longer towards the gender-matching person, rather than a person of a different gender, after they are presented with an unfamiliar human voice.

Chapter 4 is divided into two main sections. In *Article I*, I address Question 2 by presenting a detailed review of how mammals associate multi-sensory signal components, with a particular focus on the perceptual mechanisms that enable animals to match vocalisations to signallers. *Article II* addresses Question 3 by investigating if dogs also match unfamiliar voices with people according to their age category (adult or child) using a preferential looking paradigm. Based on previous demonstrations that dogs can match conspecific vocalisations to signallers according to their apparent size (e.g. Taylor et al., 2011), I predict that dogs will successfully associate human voices to people according to their age category, by looking more quickly and for longer towards a human silhouette congruent with the apparent age of the presented voice than an incongruent silhouette. This study will also further explore the perceptual mechanisms that may influence how dogs associate individual visual and vocal attributes related to the age of human speakers, by separately testing cross-modal associations based on low-level cues in the visual (speaker size and height) and auditory (voice pitch) domains.

Chapter 5 focuses on Question 4, whereby a behavioural test of brain hemispheric lateralisation is used (the head-turn paradigm) in order to investigate whether dogs show evidence of hemispheric biases for processing different information encoded in human

speech. Orienting responses to natural speech are compared to speech signals which have been re-synthesised to increase the salience of either the indexical and emotional prosodic cues, or familiar/unfamiliar phonemic cues. In accordance with the observations of a recent neuroimaging study on dogs (Andics et al., 2014), I predict that dogs will show a left orienting bias (indicating stronger right hemispheric activation) in response to speech signals when the salience of the emotional prosody is enhanced.

Chapter 6 provides a final general discussion of the main empirical results presented in the thesis in relation to Questions 1-4. I examine the theoretical implications of the conclusions drawn from each study and offer recommendations for future avenues of research.

# CHAPTER 2: MATERIALS AND METHODS

This section will provide an overview of general methods and materials used across all of the studies included in the thesis. Specific methodological details of each study are also described in full in the relevant chapter.

## Study Animals

Approximately 350 domestic dogs *Canis familiaris* of various breeds took part in the studies (details of the breeds, ages and the sex distribution are given in each study). All of the dogs that contributed were over six months old and healthy, with no known hearing or visual problems and no history of aggression towards humans. The majority of the dogs were privately owned as family pets and volunteered for participation by their owners in response to local area advertisements. In Chapter 5, 41 dogs were recruited from Brighton RSPCA rescue centre. No training or rewards were given to any of the dogs during their participation in the studies.

## Study Locations

In Chapter 3, testing was carried out at two indoor locations in the East Sussex area (UK): The Dog Hut in Barcombe and Hamsey Riding School in Lewes. In Article II of Chapter 4, all of the testing was carried out in a designated room in the School of Psychology at the University of Sussex. Some of the testing for Chapter 5 was also carried out in this room. In addition, two outdoor locations in the East Sussex area were also used in Chapter 5: Stanmer Park in Falmer and the Brighton RSPCA exercise field near Patcham.

## Auditory Stimuli

Human speech recordings were used as auditory stimuli in all of the studies. Unless otherwise stated, the speakers were all adult native British speakers studying at the University of Sussex, UK. In all cases the individuals were recruited through personal contacts and were not paid for their participation. Each person was individually recorded in a sound proof booth. All of the audio recordings were made using a Zoom H4N Handy Recorder. The sampling frequency was always set at 44,100 Hz, with a 32-bit sampling rate. The recordings were then saved as separate .wav files.

To test dogs' discrimination of human gender in Chapter 3, nine men and nine women were recorded pronouncing the following phrases as if speaking to a dog in a positive voice: "hey!", "come on then", "good dog!" and "what's this?". Each speaker pronounced each of the phrases once.

In Article II of Chapter 4, the auditory stimuli used to investigate dogs' discrimination of different human age categories included four adult men pronouncing the word "hod", whilst extending the vowel sound for approximately 1 second. A few recordings were obtained from each participant to ensure that the vowel was correctly pronounced and only the highest quality recording was retained for each participant (giving four recordings in total). Only the steady-state portion of the vowel (i.e. where the formants are static) was subsequently used as a stimulus. The recordings were also re-synthesised to match the average acoustic values of a six-year old boy to explore dogs' categorisation of different human age categories (see the next section for detailed descriptions of all acoustic modifications). In addition to the voices, nine pure sine-wave tones were also created, eight of which matched the F0s of the original and re-synthesised voices, whilst the frequency of the sine-wave fixation tone used was half way between the averages of the adult and re-synthesised 'child' sine-wave tones.

In order to determine if dogs differentially process the main functional components of human speech in Chapter 5, four men and four women were first recorded pronouncing the phrase "come on then" in a happy tone of voice. These recordings were subsequently re-synthesised to provide four additional types of stimuli: 1) speech with neutralised intonation; 2) sine-wave speech; 3) speech prosody with no phonetic content; 4) sine-wave intonation (used in Experiment 1). Eight native French speakers, four men and four women (working at various universities in France and the UK), were also recorded pronouncing the phrase "aller viens le chien" in a happy tone of voice. A pink noise audio-file was also created as a control stimulus. In Experiment 2, four men and four women were recorded pronouncing the pseudo-word phrase "thon om ken" in a happy tone of voice. An additional eight speakers also each produced four different whistles which are commonly used by dog owners. A further eight native French speakers, four men and four women, pronounced the phrase "come on then" in a happy tone of voice with a strong French accent. All of the voices used in Experiment 2 were re-synthesised to neutralise the intonation contours.

## Measurement and Modification of Acoustic Parameters

All of the acoustic analyses and manipulations were conducted using PRAAT v.5.0.3 (http://www.fon.hum.uva.nl/praat/). Because the analyses may not function properly at the start and end of the signal (Wood, 2003), 0.5 s of silence was first added at the beginning and end of each sample. All of the recordings were normalised to -1.0 dB maximum amplitude using Audacity 2.0.0 (http://audacity.sourceforge.net) after any additional acoustic manipulations had been made.

## Acoustic Analyses

The acoustic characteristics of the fundamental frequency (F0) and lowest four formant frequencies were measured in each of the human voice recordings obtained. For all of the analyses the spectrogram settings were as follows: view range, 0–5000 Hz; window length, 0.05 s; dynamic range, 50 dB. In each case the spectrogram was manually inspected to verify the accuracy of the values obtained from the analyses.

## Fundamental Frequency Parameters

The mean, minimum and maximum F0 values were extracted using the PRAAT autocorrelation algorithm 'to Pitch (ac)' which estimates the F0 contour across the utterance. The time step for the analysis was set to 0.01 s and the pitch floor and pitch ceiling parameters were set to 30 Hz and 500 Hz respectively, as this range encompasses the typical frequency range of adult human voices (Titze, 1994). The default values were retained for all of the other parameters. The analysis applies a low pass smoothing filter to remove any rapid variations in the F0 contour before the pitch object is produced. The mean, minimum, maximum and standard deviation values for the F0 across the utterance were then obtained by selecting the pitch object and using the "get mean/minimum/maximum/standard deviation" commands.

## Formant Related Parameters

The centre frequencies of the first four formants were obtained using the PRAAT Linear Predictive Coding (LPC: 'To Formants (Burg)' command) algorithm. The analysis parameters were set as follows: time step, 0.1 s; window length, 0.05 s; maximum number of formants, 5; maximum formant frequency, 5000 Hz or 5500 Hz for male and female voices respectively. The obtained frequency values for the first four formants

were then used to calculate the average formant spacing using the following equation (Fitch, 1997):

(1) $\Delta F = F_{i+1} - F_i$

As discussed in Chapter 1, the vocal tract can be approximated to a uniform tube closed at one end. The centre frequencies of the successive formants are therefore related to the length of the vocal tract as follows:

(2) $F_i = \frac{(2i-1)c}{4VTL}$

where *c* is the speed of sound in the air (350m/s in the human vocal tract) and *VTL* is the length of the vocal tract. Therefore, the spacing between any consecutive formants is the same and equivalent to:

(3) $\Delta F = F_{i+1} - F_i = \frac{c}{2VTL}$

It is thus possible to replace $\frac{c}{2VTL}$ with $\Delta F$ in equation 2, so that individual formants can be related to the formant spacing $\Delta F$ by:

(4) $F_i = \frac{2i-1}{2}\Delta F$

$\Delta F$ can be estimated by determining the best fit for equation (4) to the centre frequency of the first four formants (Reby & McComb, 2003).

## Acoustic Re-Synthesis

### Manipulating the F0 (Article II of Chapter 4 and Chapter 5)

The Pitch-Synchronous Overlap and Add (PSOLA; Moulines & Charpentier, 1990) algorithm was used to re-synthesise the recordings and manipulate the F0 to the required values. This method generates new pitch points whilst leaving all of the other acoustic features unchanged. Manipulations of the median F0 and F0 range were carried out using the 'change gender' command with the following settings: pitch floor, 30 Hz; pitch ceiling, 500 Hz; formant shift, 1 (no change); duration factor, 1 (no change). The desired median pitch value was entered using the 'new pitch median (Hz)' setting. The

pitch range factor was either set at 1 when no changes to the pitch modulation were required, or to 0.001 to create a flat, perceptually monotonous pitch.

**Creating Sine-Wave Tones from the F0 (Article II of Chapter 4 and Chapter 5)**

The pitch contours were first extracted, using the 'To Pitch' command, and then converted into sine-wave tones using the 'To Sound (sine)' command. The onsets and offsets were set to fade gradually using the 'fade in' and 'fade out' commands in Audacity.

**Manipulating the Formant Scaling (Article II of Chapter 4)**

The PSOLA algorithm ('change gender' command) described above was also used to change the formant scaling of adult male voices. The following settings were applied: pitch floor, 30 Hz; pitch ceiling, 500 Hz; new pitch median, 0 Hz (no change); duration factor, 1. The formant shift ratio was calculated by comparing the average percentage difference between the formant spacing of an adult male voice and a six-year old boy (Lee, Potamianos & Narayanan, 1999). The average formant spacing in the children's voices was calculated to be 32% greater than the spacing in the adult male voices, therefore the formant shift ratio for the adult male voice recordings was set at 1.32.

**Removing the Formants (Chapter 5)**

The plosives were first manually cut from each recording. The recordings were then re-sampled to provide a band limit of around 5000 Hz, and a pre-emphasis of 50 Hz was added. The spectral envelope of the recording was then flattened (using LPC synthesis/inverse filtering) to remove the temporal and formant-related phonemic content. To achieve this, the LPC analysis uses linear prediction to estimate the first five formant frequencies and bandwidths, producing a smoothed version of the spectrogram. Ten linear prediction parameters were used, with an analysis window of 25 ms and time steps of 5 ms. This produced an LPC object approximating the formant frequencies. Inverse filtering was then performed on the original sounds using the LPC objects, removing the formant frequencies.

**Creating Sine-Wave Speech (Chapter 5)**

The 'SWS script' written by Chris Darwin (http://www.lifesci.sussex.ac.uk/home/Chris_Darwin/Praatscripts/SWS) was used to

convert original speech recordings into sine-wave speech. This algorithm uses LPC to estimate the first three formant frequencies, and the formant amplitudes are taken from a wideband FFT spectrum. The estimates are then smoothed to produce continuous contours and remove any residual artefacts. This produces three sinusoid curves that track the lowest three formants of the original voice.

**Creating Pink Noise (Chapter 5)**

Audacity was used to create 1 s of Gaussian pink noise (using the 'create noise' function), with the amplitude set to fade gradually at the onset and offset.

**Perceptual Ratings of the Auditory Stimuli**

To verify the validity of the acoustic manipulations carried out on the stimuli used in Chapter 5, all of the stimuli were rated by five volunteers who were naïve to the experimental conditions. A listening experiment was created using PRAAT. The participant first heard one of the sounds through headphones, after which they were presented with two 5-point Likert scales to rate the stimulus. On the first scale, participants were asked if they could understand what the person was saying (1 = very unclear, 5 = very clear) and on the second scale they rated the emotional valence and intensity of the sound (1 = very negative, 5 = very positive). Each sound was scored on both scales and could be replayed multiple times before rating. The participants were asked to rate any speech other than English as unintelligible (1 on the first scale). The stimuli were presented in a pseudo-randomised order so that the participants did not hear the original recordings before the acoustically degraded stimuli, as this could have influenced their perception of the degraded stimuli.

## Visual Stimuli

In Chapter 3, each subject was presented with two unfamiliar people, a man and a woman, who were chosen pseudo-randomly from a pool of five men and five women. Together the individual people provided a range of physical attributes such as age and hairstyle, while their heights were bimodally distributed by gender.

In Article II of Chapter 4, three pairs of images were used, which all appeared on a white background. One pair of images consisted of two equally sized black squares (35 $cm^2$), which appeared at the top and bottom of a projection wall. The second pair of

images consisted of two differently sized black squares (30 $cm^2$ and 60 $cm^2$) placed side by side at an equivalent height half way up the wall. The final pair of images was of two life size black silhouettes, one of a young boy and one of an adult man, which appeared side by side with their feet at the base of the wall.

## Experimental Designs

### Preferential Looking Paradigm (Chapter 3 and Article II of Chapter 4)

In Chapter 3 and Article II of Chapter 4, we used preferential looking paradigms to determine if dogs spontaneously match human voices to unfamiliar speakers by cross-modally associating congruent indexical cues to the person's gender and age. Although originally developed for testing behavioural responses in human infants (Golinkoff, Hirsh-Pasek, Cauley & Gordon, 1987), the preferential looking paradigm has also become an established methodology in non-human animal investigations (e.g. domestic dogs: Taylor, Reby & McComb, 2011; rhesus macaques *Macaca mulatta*: Ghazanfar et al., 2007; domestic horses *Equus caballus*: Proops & McComb, 2012). The premise on which this paradigm is based is that when an association exists between two different perceptual cues, the presence of one will trigger increased attention to the other (Golinkoff et al., 1987). Therefore, one of the major advantages of this paradigm is that it can be used to test ecologically relevant associations that subjects make spontaneously without the need to use inherently artificial training regimes. Specifically, when investigating associations between auditory and visual information, the subject is simultaneously presented with two different visual stimuli and an auditory stimulus is played from a central or otherwise non-biasing location. If the subject perceives that the auditory stimulus matches one of the visual stimuli more than the other in a particular dimension (e.g. they share the same gender cues), then the subject is predicted to preferentially attend to the matching visual stimulus (usually quantified by a shorter response latency and greater amount of looking time towards the matching stimulus relative to the non-matching stimulus; Aslin, 2007). However, in some cases less attendance to the matching image has also been interpreted as demonstrating that the subject has appropriately combined the matching sensory information, where supplementary evidence has suggested that the congruent pairing may have been perceived as negative and therefore visually avoided by the subjects (e.g. Zangenehpour, Ghazanfar, Lewkowicz & Zatorre, 2009). Similarly complex association

preferences observed in human infant studies, where attention appears to shift from familiar to novel stimuli with increasing exposure (Hunter & Ames, 1988), has led to some criticism of the preferential looking paradigm (e.g. Aslin & Fiser, 2005). Although variability in responses can make it more difficult to accurately interpret the results of some studies, the paradigm does proficiently demonstrate when the subject has made a distinction between two different combinations of audio-visual stimuli. When coupled with strong *a priori* hypotheses as well as additional behavioural evidence, the meaning of these distinctions can be reasonably inferred (Houston-Price & Nakai, 2004). The preferential looking paradigm has also been viewed as more limited than alternative behavioural methodologies, such as the violation of expectation paradigm, as while results can show if an animal spontaneously cross-modally associates two stimuli, the preferential looking paradigm cannot be used to reveal the precise nature of any associations formed across the senses, and therefore cannot distinguish between low level and higher level cognitive processes (e.g. Adachi, Kuwahata & Fujita, 2007). However, the reverse argument is also valid, as methodologies such as the violation of expectation paradigm can only identify the presence of higher level cognitive representations because low-level cues such as temporal synchrony are excluded, and can therefore suggest that no associations are made when animals may actually be capable of matching the stimuli by making low-level associations. Because our studies aimed to establish if dogs were able to match human voices with speakers by associating congruent indexical cues, either independently of the cognitive mechanisms involved or based on low-level associations, the preferential looking paradigm was the most appropriate methodology to test our hypotheses.

**Head-Turn Paradigm (Chapter 5)**

Chapter 5 aimed to determine if and how dogs dissociate the main communicative components of human speech, by establishing if they show any evidence of hemispheric asymmetries in response to different functional attributes of speech signals and/or non-speech sounds. The behavioural head-turn paradigm was first developed by Hauser and Andersson (1994) to provide a non-invasive indicator of auditory cerebral laterality in animals. The method consists of presenting a sound equally from both sides of the animal and recording the side to which the subject orients in response. The sound is

either broadcast from directly behind the individual or from two speakers positioned to their left and right, so that each ear receives equal auditory input. Because in mammals the contralateral pathways from the ear to the auditory cortex are stronger than the ipsilateral pathways (e.g. humans: Bocca, Calearo, Cassinari & Migliavacca, 1955; dogs: Tunturi, 1946; domestic cats *Felis catus*: Rosenzweig, 1951; Hall & Goldstein, 1968), the input from the contralateral ear has an advantage in accessing areas in the opposite hemisphere that may be specialised in processing relevant elements of the sound (Grimshaw, Kwasny, Covell & Johnson, 2003). Therefore, similarly to the results shown using the dichotic listening paradigm with human listeners, these elements of the input are perceived more strongly through the contra-lateral ear (Kimura, 2011). In the head-turn paradigm, it is hypothesised that consistent orienting with the right ear leading indicates predominant processing in the left hemisphere, whilst turning with the left ear leading indicates stronger right hemispheric activation (Hauser & Andersson, 1994). Since its conception, the head-turn paradigm has been used to investigate lateralisation in response to species-specific vocalisations in a wide range of species (e.g. dogs: Siniscalchi, Quaranta & Rogers, 2008; rhesus macaques: Hauser & Andersson, 1994; California sea lions *Zalophus californianus*: Böye, Güntürkün & Vauclair, 2005; harpy eagles *Harpia harpyja*: Palleroni & Hauser, 2003). The prevailing demonstrations of right head turn biases in mammals in response to species-specific vocalisations (Ocklenburg, Ströckens & Güntürkün, 2013) are consistent with studies using brain-imaging (Poremba et al., 2004), psychophysical (LePrell, Hauser & Moody, 2002; Petersen, Beecher, Moody & Stebbins, 1978) and lesioning techniques (Heffner & Heffner, 1984), providing validation for the use of the head-turn paradigm as a behavioural indicator of hemispheric asymmetries during processing.

However, while the majority of studies have observed right head turn biases in mammal responses to conspecific vocalisations (Ocklenburg et al., 2013), failure to replicate these results in some studies (Teufel, Hammerschmidt & Fischer, 2007; Gil-da-Costa & Hauser, 2006; Scheumann & Zimmermann, 2008; Basile, Lemasson & Blois-Heulin, 2009; Lemasson et al., 2010; Leliveld, Scheumann & Zimmermann, 2010), has led to concerns about the validity of the head-turn paradigm as an indicator of hemispheric activation (reviewed by Teufel, Ghazanfar & Fischer, 2010). Indeed, in a direct comparison of head turn and fMRI responses to speech in human listeners, Fischer et al., (2009) observed slight left head turn biases in response a range of speech and

nonspeech sounds, while speech stimuli produced stronger left hemispheric activation in the fMRI experiments. In contrast, Marzoli and Tommasi (2009) did obtain right head turn biases in a range of similar naturalistic experiments with human listeners, which was also consistent with right head turn biases previously reported in human infants' responses to female speech (Ecklund-Flores & Turkewitz, 1998). However, the inconsistent results obtained using the head-turn paradigm with both humans and non-human animals could be related to differences in the auditory stimuli, procedures and test conditions used across the studies. Although some idiosyncratic variables such as paw preferences have been directly tested and do not influence orienting responses to sound stimuli (Siniscalchi et al., 2008), other sources of individual variation have been observed to affect the subjects' responses. For example, human hemispheric asymmetries in response to emotional speech can be modulated depending on whether participants are asked to pay attention to the verbal content or emotional content (Mitchell, Elliott, Barry, Cruttenden & Woodruff, 2003). Similarly, the emotional valence of vocalisations has been determined to influence orienting responses in human children and Campbell's monkeys *Cercopithecus campbelli* (Basile et al., 2009), while biases to particular sounds appear to be experience dependent across a range of species (Palleroni & Hauser, 2003; Böye et al., 2005; Hauser & Andersson, 1994). In some cases methodological issues have also been identified such as spatial confounds (Teufel et al., 2010). Therefore, rather than evidencing the invalidity of the head-turn paradigm itself, conflicting response biases may instead be explained by specific experimental effects, which make it difficult to compare and integrate different outcomes across studies. In contrast, a key strength of our study using the head-turn paradigm (detailed in Chapter 5) is that we compared several different acoustic conditions using the same experimental design, making our results directly comparable and therefore more clearly interpretable. Potential individual differences were also limited by testing a larger number of subjects.

**Behavioural Data**

All of the dogs were video recorded in each study. Their behaviour in the video was coded on a frame by frame basis (1 frame = 100 ms) using the digital video analysis software Sportscode Gamebreaker version 7.5.5 (Sportstec, Warriewood, NSW, Australia). In each study, the main behavioural response measure was the looking direction. This provided the response latency and gaze duration towards each of the

visual stimuli in Chapter 3 and Article II of Chapter 4, and in each direction for Chapter 5. Additional behaviours were also coded on a presence/absence basis, such as head tilting, muzzle licking and startle responses. For Chapter 3 and Article II of Chapter 4, all of the videos were coded blind to the correct stimulus position. Across all of the studies, a proportion of the videos were second coded by a research assistant who was naïve to the experimental conditions.

## Ethical Considerations

### Animals

All of the studies complied with the internal University of Sussex regulations on the use of animals and were approved by the University of Sussex Ethical Review Committee. The approval number for each study can be found in the relevant chapter. The data collected was entirely observational, recording dogs' spontaneous behavioural responses to procedures mimicking everyday occurrences in naturalistic settings. There was no manipulation of the animals or use of invasive techniques. In the case of privately owned dogs (which made up the large majority of the subjects), the dog's owner remained with them at all times and was the only person that handled the dog. Dogs were only recruited if they were non-aggressive towards people and had no known health issues. I remained conscious of the welfare considerations at all times, and dogs were monitored for signs of distress. Water was always freely available to the dogs and testing sessions never exceeded 20 minutes. All of the experimental paradigms were well-established and had been used with dogs in previous studies.

### Humans

Approval to record adults' (over 18 years old) voices for the experimental stimuli was obtained from the University of Sussex Life Sciences and Psychology Cluster based Research Ethics Committee. Approval for participants to rate the auditory stimuli used in Chapter 5 was also obtained from the same ethical committee. The approval number for each study can be found in the relevant chapter. Individuals interested in participating were given time to read an information sheet explaining the purpose of the recordings before deciding whether to take part and each person signed a consent form. The participants were also informed that they had the right to withdraw at any time. All of the voice recordings obtained remained anonymous.

# CHAPTER 3: CROSS-MODAL DISCRIMINATION OF HUMAN GENDER BY DOMESTIC DOGS

## Synopsis

---

**Question:** Are dogs spontaneously capable of cross-modal human gender discrimination?

---

**Methods:** In a preferential looking paradigm, dogs were presented with an adult human voice while an unfamiliar man and woman were both stood in front of them. The direction and duration of their gaze responses towards each person were recorded, as well as any additional behavioural responses.

---

**Results:** Dogs living in households with multiple men and women looked more towards the person that matched the gender of the voice, while those dogs living with fewer people avoided responding to the gender-matching person by instead looking more at the person that did not match the gender of the voice.

---

**Conclusions:** Although the expression of the ability to successfully discriminate the gender of adult human voices was dependent on the subjects' level of socialisation with humans, our study demonstrated that dogs are capable of spontaneously categorising human gender by associating information across different sensory modalities.

*Note*. Published as: Ratcliffe, V.F., McComb, K. & Reby, D. (2014). Cross-modal discrimination of human gender by domestic dogs, *Animal Behaviour*, *91*, 127-135.

## Abstract

We spontaneously categorise people as male or female, and when hearing a human voice we expect to see an appropriate gender-matched visual image. The extent to which domesticated species, which share our social environment, spontaneously develop such categorisation abilities remains under-investigated. Here we used a cross-modal preferential looking design to determine if domestic dogs *Canis familiaris*, spontaneously attribute an unfamiliar voice to a person of the corresponding gender. Fifty-one dogs were played a pre-recorded male or female voice in the presence of a man and a woman. The responses were scored as correct or incorrect from both the direction of the first look and the total gaze duration towards each person after the voice presentation. Dogs living with one adult, or one man and one woman, performed significantly below chance as more (71%) of these dogs looked towards the non gender-matched person first. However, dogs living with more than two adults (including at least one man and one woman) performed significantly better, and significantly more (80%) of these dogs looked at the gender-matched person for longer than they looked at the non-matching person. This suggests that while all of the dogs had spontaneously learnt to categorise human gender across sensory modalities, this ability was expressed differently depending on their social experience with humans. Dogs with greater experience, gained through regular exposure to multiple male and female human exemplars, responded by orientating towards the gender-matching person, whilst those with more limited experience avoided looking towards the gender-matching person. We discuss the importance of experience in determining the way that individuals spontaneously form and express categorisation abilities.

## Introduction

Categorisation is a key cognitive mechanism that determines how we perceive and process sensory information. As well as simplifying processing requirements (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976), organising stimuli into categories allows general inferences to be made and applied to new category members. Humans

readily form complex, hierarchical categories representing their environment, using language to create specific referents that can co-ordinate categories between individuals (see Steels & Belpaeme, 2005 for a review). Currently, only a small number of studies have explored spontaneous category formation in other species, focusing on non-human primates (e.g. Murai, Tomonaga, Kamegai, Terazawa, & Yamaguchi, 2004; Murai et al., 2005). Comparative investigations into spontaneous category formation in non-human animals are therefore necessary to determine the functional relevance of this cognitive process in a broader range of species.

Domestic dogs provide an interesting model species to compare natural category formation in animals and humans. Dogs have shared the same environment as humans for at least 15,000 years (Savolainen, Zhang, Luo, Lundeberg, & Leitner, 2002), during which time they are likely to have undergone selection promoting specific socio-cognitive abilities that allow effective co-operation and communication between the two species (Hare, Brown, Williamson & Tomasello, 2002; Bräuer, Kaminski, Riedel, Call, & Tomasello, 2006). Added to this evolutionary predisposition is the effect of experience, as many dogs are extensively socialised with people, often sharing the same living habitat from an early age. As the human environment has become functionally relevant to dogs, this species may be expected to form spontaneous categories that are directly comparable with human categories.

It has already been established that with training, dogs show equivalent categorisation abilities to other mammals and birds. They are able to discriminate between 'dog' and 'non-dog' sounds (Heffner, 1975), images of dogs and landscapes (Range, Aust, Steurer, & Huber, 2008) and images of dogs and other species (Autier-Dérian, Deputte, Chalvet-Monfray, Coulon, & Mounier, 2013), correctly generalising their responses to novel stimuli. Spontaneous, ecologically relevant category formation is also evident in the dog's ability to form cross-modal perceptual associations when responding to familiar people. Using an expectancy violation paradigm, Adachi, Kuwahata and Fujita (2007) presented dogs with a photograph of either their owner or a stranger's face after playing back one of their voices. Dogs looked longer when the face did not match the preceding voice than when the stimuli did match, suggesting that dogs can use cross-modal associative categories when responding to familiar humans. This form of categorical perception is likely to be expressed naturally by dogs as the need to identify familiar humans has a clear function in recognising important social partners and care

providers. The spontaneous use of an inter-specific category representing familiar humans leads to the possibility that it may also be relevant for dogs to form categories about unfamiliar humans, which would allow direct comparisons with our own categories.

One of the predominant ways that we categorise unfamiliar people is by their gender, primarily by associating visual and vocal cues. Because human faces are sexually dimorphic (Burton, Bruce, & Dench, 1993), differing in both shape and texture (Hill, Bruce, & Akamatsu, 1995), face gender classification in adults is close to 100% accuracy (O'Toole et al., 1998). Sexual dimorphism also leads to differences in the vocal tract anatomy of adult men and women. The larger adult male larynx results in a difference of approximately 80 Hz in fundamental frequency (F0) between the voices of adult men and women, with mean values at around 120 Hz and 200 Hz respectively (Titze, 2000). Additionally, adult men have a disproportionally longer vocal tract than women (Vorperian et al., 2009), causing lower first formant (F1) values and formant spacings that are approximately 15-20% lower than women (Fant, 1960; Goldstein, 1980). The relative F0 and formant values classify the gender of adult voices at 98.8% accuracy (Bachorowski & Owren, 1999). The presence of both visual and vocal gender cues enables cross-modal perceptual matching of voices to individuals from an early age (Walker-Andrews, Bahrick, Raglioni, & Diaz, 1991).

Dogs are also likely to be able to perceive these gender differences in the human voice, as they attend to variation in formants to determine size information in conspecific vocalisations (Taylor, Reby, & McComb, 2011) and can be trained to discriminate between average male and female F0 differences in human vowel sounds (Baru, 1975). Gender specific behavioural differences in the way humans interact with dogs have been identified (e.g. Prato-Previde, Fallani, & Valsecchi, 2005) which could have created the need for dogs to categorise human gender in order to adjust their responses appropriately. In support of this, shelter-housed dogs petted by women show more relaxed behaviour and lower cortisol levels than those petted by men (Hennessy, Williams, Miller, Douglas, & Voith, 1998), and are more likely to direct defensive aggressive behaviour towards men than women (Lore & Eisenberg 1986; Wells & Hepper 1999). Although the specific cues to which the dogs were responding cannot be determined from these studies, they do suggest that categorically assessing human gender could be functionally relevant for dogs, influencing their reaction to the

individual person. Therefore the ability to perceive and associate different sensory cues to human gender as categorically equivalent would be a useful ability.

To determine if dogs do categorise human gender using different sensory cues, we tested whether they associate voices with unfamiliar people using gender cues in a cross-modal preferential looking paradigm, where subjects were required to spontaneously match voices to people by their gender. In our study, a man and a woman stood either side of a loud speaker from which a voice recording of a different person was played. Dogs were positioned facing the centre line, and their visual orientation to the person matching the gender of the voice and the non gender-matching person were recorded. If dogs spontaneously combine vocal and visual cues to identify human gender cross-modally, it was predicted that they would look first, and for a longer duration, at the person of the same gender as the voice. The potential effect of social factors on performance was also investigated, as well as possible mechanisms involved in such variation.

## Methods

### Subjects

A total of 51 adult dogs of 17 different breeds were recruited when their owners responded to advertisements in the East Sussex area. Ages ranged from seven months to 11 years old (Mean+SD = 5.03+3.17), including 26 males and 25 females. The selection criteria for subject animals was that they had to be healthy adults (older than six months) with no known sight or hearing problems and no known aggression towards people. Subjects and their owners were naïve to the experimental set-up and had not participated in any previous vocal communication or behavioural research.

### Playback Acquisition

Nine men and nine women, aged between 20 and 52 years (Mean+SD = 30.94+9.75 years), were audio recorded after being instructed to pronounce the following phrases as if speaking to a dog in a positive voice: “Hey!”, “Come on then”, “Good dog!”, “What’s this?”. Each speaker pronounced each phrase once. All recordings were made using a Zoom H4N Handy Recorder in a sound proof booth. The sampling frequency was set at 44 100 Hz, with a 32-bit sampling rate, for each recording. The vocal parameters of the recordings were then checked for a bimodal distribution according to gender using

PRAAT v.5.0.3 (http://www.fon.hum.uva.nl/praat/). The four phrases were analysed together as a single audio-file. The mean, minimum and maximum F0 values were calculated using the PRAAT autocorrelation algorithm "to Pitch (ac)" which estimates the F0 contour across the utterance. The mean F0 for the male voices was between 142.00 Hz and 193.48 Hz (Mean+SD = 166.80+17.64 Hz), whilst the mean F0 for the female voices ranged between 251.13 Hz and 405.99 Hz (Mean+SD = 323.26+61.22 Hz). The F0 ranges (max F0 – min F0) for the male voices were between 109.70 Hz and 164.73 Hz (Mean+SD = 154.16+48.32 Hz), whilst the female F0 ranges were between 269.41 Hz and 528.01 Hz (Mean+SD = 350.20+86.35 Hz). The formant spacing ($\Delta F$) was calculated using the PRAAT Linear Predictive Coding 'Burg' algorithm, which estimates the centre frequencies of the first four formants across the utterance. These values were then used to calculate the average spacing between the formants. The male $\Delta F$s were between 927.98 Hz and 1120.60 Hz (Mean+SD = 1029.15+71.39 Hz), whilst the female $\Delta F$s ranged between 1140.60 Hz and 1241.00 Hz (Mean+SD = 1215.56+45.20 Hz). All of the recordings were normalized to -1.0 dB maximum amplitude in Audacity 2.0.0 (http://audacity.sourceforge.net).

**Experimental Set Up**

Experiments were carried out between June and September 2012 at two indoor testing locations in the East Sussex, U.K., area (The Dog Hut in Barcombe and Hamsey Riding School in Lewes). A cross-modal preferential looking paradigm was used. The design was developed on the basis of pilot trials conducted in April and May 2012 on 20 subjects, who did not take part in the final study trials. The original piloted study included a sequence of six trials per subject; however, we found that habituation to the procedure led to a reduction in responses after the first trial. Therefore in the full study each dog took part in only one trial.

An Anchor LIB-6000H Liberty loud speaker (frequency response: 60 Hz-15 kHz) was mounted onto a 130 cm tall stand and disguised using brown material. The speaker was positioned 500 cm in front of a designated subject area (150 $cm^2$) where subjects could be positioned, with a chair for their owner and a black screen placed directly behind them. A SONY DCR-HC51 Handycam video camera was mounted onto a tripod 30 cm from the floor, and positioned directly in front of the loud speaker, facing and zoomed in towards the subject. Florescent coloured rope was used to delineate a centre line

between the subject and the loud speaker and was clearly visible on the videos, providing a visual determinant of left and right during video coding. A second video camera was mounted on a tripod 100 cm from the floor, and placed behind the subject area, facing the loud speaker. This was to monitor the subject's field of view.

Two assistants, a man and a woman chosen from a pool of 10 people, stood facing the subjects with their nearest foot 150 cm either side of the centre of the loud speaker (Figure 1). Each assistant remained stationary with one hand covering their mouth and gazed straight ahead with a neutral facial expression. The assistants did not make eye contact or interact with the dogs in any way throughout the study. The side that the male and female assistants stood on was counterbalanced across subjects. The two assistants were chosen on a pseudo-randomised basis from a pool of five men and five women, providing variation in physical attributes including age and hair style. The heights of the assistants were bimodally distributed by gender, as the male heights ranged between 180.34-190.50 cm (Mean+SD = 183.27+4.15 cm) whilst the female heights ranged between 153.00-171.00 cm (Mean+SD = 162.53+7.89 cm).

The visual acuity of dogs is typically given to be around 20/75 using the Snellen fraction (Miller & Murphy, 1995), and therefore their visual perception of objects is less detailed than human visual perception at the same distance. However, using a similar paradigm to the current study, Faragó et al., (2010) demonstrated that dogs were able to successfully discriminate between size-matched images of cats and dogs (approximately 30 cm in height) from a distance of 5 m. Therefore it was expected that in the present study, as well as acquiring possible olfactory information, dogs would also be able to obtain sufficient visual information from the humans to discriminate gender-related information from the same distance.

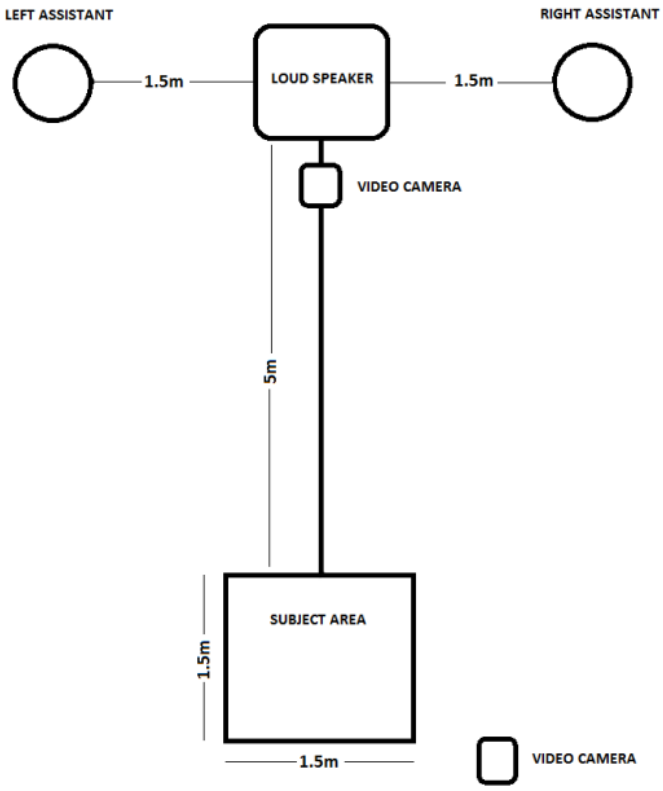



**Fig. 1** Experimental set up with distances between the subject, loud speaker and assistants

**Procedure**

Subjects were held loosely by their lead throughout the experiment and were always handled by their owner. The owners were naïve to the purpose of the experiment and were instructed on entering the test room to allow their dog to familiarise itself with the area, including walking in front of both the assistants and the loud speaker. This was aimed at giving the dogs access to olfactory information from the assistants. The owner was then instructed to sit on the chair provided and place their dog in front of them. The owner was asked to remain silent and still, and avoid interacting with their dog unless necessary to keep their dog inside the subject area. Neither the owner nor the experimenter was in the subject's field of vision during the experiment, both to minimise unconscious cueing and to prevent dogs matching the playback voice to unintended targets. In order to ensure that the owner and assistants were not giving any unintentional cues when they heard the playback voice, half of the tests were conducted with both of the assistants and the owner listening to music from handheld MP3 players, which masked the sound of the playback voices.

After settling in their position the subjects were given 10 s for further visual familiarisation with the assistants. This was followed with 10 s of silence, followed by the presentation of a single playback voice and a further 10 s of silence. The playback voice consisted of one person saying the four phrases outlined above, in the same order, with a 500 ms interval between each phrase. The recording was played at 65 dB (+/- 5 dB), measured by a N05CC Digital Mini Sound Level Meter. The playback exemplar was chosen from the pool of 18 voice recordings in a pseudo-randomised order across subjects, so that half of the subjects heard a female voice and half heard a male voice. The individual playback voices, the gender of the voice and the side that the male and female assistants stood on were counterbalanced across subjects.

**Collection and Coding of Dog Contextual Information**

Following the experiment the owners were asked to fill out a short questionnaire about their dog. This included questions about their dog's breed, age, sex, reproductive status, the number of adult men and women living with the dog (household composition), the average amount of time the dog spent with people per day, behavioural characteristics around unfamiliar men and women and their dog's origination (private or animal shelter).

To explore potential differences between the subjects' performance depending on their human social environment, household composition was also coded as a categorical variable (HC Group) with two levels: subjects living with either one adult person or one man and one woman (HC Group 1, $N = 35$) and subjects living with between three and five adults, including at least one man and one woman (HC Group 2, $N = 15$). The number of men and women living in the household was evenly balanced for the majority of dogs in HC Group 2, as ten of the subjects lived with two men and two women.

**Ethical Note**

The dogs were privately owned and handled by their owner throughout the study, which was designed to replicate a routine interaction between a dog and an unfamiliar person. The study complied with the internal University of Sussex regulations on the use of animals and was approved by the University of Sussex Ethical Review Committee (Approval number: ERC/33/3). Approval to record human voices to be used as stimuli

was also obtained from the University of Sussex Life Sciences & Psychology Cluster based Research Ethics Committee (Approval number: DRVR0312).

**Behavioural Measures and Coding**

Videos were coded in 100 ms intervals using the digital video analysis software Sportscode Gamebreaker version 7.5.5 (Sportstec, Warriewood, NSW, Australia).

The dogs' responses were measured during the 10 s of silence immediately before the playback voice presentation (pre-playback) and 10 ms after the onset of the playback voice for a total duration of 15 s until the end of the trial (trial duration was determined from the maximum response duration during the pilot study). The latency, duration and direction of each look (towards each of the assistants, the loud speaker and away) were recorded in milliseconds.

A look was defined as being at either of the assistants if the dog's head was directed between 15° and 25° from the centre line (delineated by the florescent rope), and was recorded as being at the loud speaker if the dog's head was directed between 0° and 5° from the centre. Finally, a look was recorded as away if it was directed between 6° and 14° or over 26° from the centre point. The orientation of the dog's head was taken by drawing a line from centre of the top of the dog's forehead to the centre of their nose (Figure 2). A protractor was placed along the centre line of each video in order to determine these angles. Although this method does not give an absolute measure of the visual orientation, it does provide a repeatable and standardised index of orientation across the subjects.

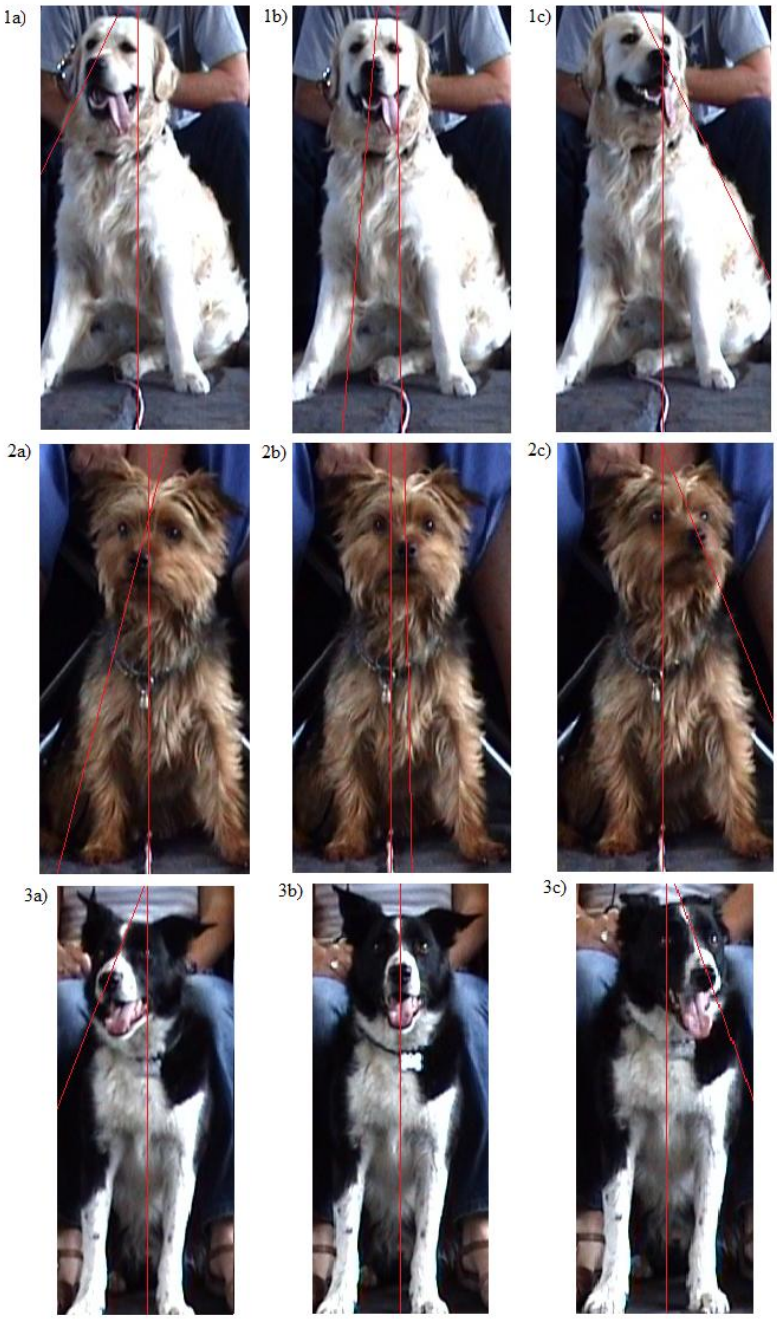



**Fig. 2** Example frames showing the video analysis coding of the gaze orientation of three subjects. Lines illustrate the angle of the dog's head in relation to the centre line. 1a) Golden retriever orientated towards the person stood on their right ($24^{o}$); 1b) the loud speaker ($4^{o}$); 1c) the person stood on their left ($24^{o}$). 2a) Yorkshire terrier orientated towards the person stood on their right ($15^{o}$); 2b) the loud speaker ($1^{o}$); 2c) the person stood on their left ($22^{o}$). 3a) Border collie orientated towards the person stood on their right ($21^{o}$); 3b) the loud speaker ($0^{o}$); 3c) the person stood on their left ($20^{o}$).

From these analyses we derived the following gaze response variables: the total gaze duration (total time looking) and response latency (time to first look) towards each of the assistants, and the total time spent looking elsewhere (at the loud speaker or away). To characterise the dogs' ability to match the voice gender to the correct same gender assistant, we attributed two separate binary 'correct matching scores' of correct or incorrect using both of the gaze response variables directed towards the assistants (total gaze duration and response latency) after the playback voice presentation. The direction of the first look was scored as correct if the subject looked towards the correct gender-matching assistant before the incorrect non gender-matching assistant (First look correct score), while the total duration was scored as correct if the subject looked at the correct gender-matching assistant longer than they looked at the incorrect assistant in total (Total gaze correct score).

To investigate mechanisms underlying the gaze responses, we also recorded the occurrence of appeasement signals, which dogs use to reduce the potential of conflict during social interactions. These behaviours are often produced when a dog feels anxious or threatened, and are likely to co-occur with turning away or gaze aversion (Rugaas, 2005). The frequency of occurrence of each of the following appeasement behaviours was monitored: licking the muzzle, yawning, trembling, scratching, sniffing the floor or attempting to move away. The frequencies for each of the behaviours were then summed to provide the total frequency across all behaviours. Any vocalisations made were also recorded and each call-type (bark, whine, growl or howl) was scored using the following scale (0: no occurrence, 1: produced less than five times, 2: produced five or more times). The scores for each call-type were then added to the total frequency, to provide an overall appeasement behaviour (AB) score for each subject. Two separate AB scores were given to each subject; one score for the time period before (pre-playback) and one score after the playback voice presentation.

The videos were coded in blind order by V.R. A research assistant second-coded 84% of the videos, which resulted in a strong inter-observer correlation for both the response latency (Pearson's R: $r_{40} = 0.80$, $P < 0.001$) and the total gaze duration towards each person (Pearson's R: $r_{40} = 0.87$, $P < 0.001$). A research assistant also second-coded the appeasement signal scores in 66% of the videos, again resulting in a strong inter-observer correlation for the pre-playback AB scores (Spearman's rho: $r = 0.78$, $N = 33$,

$P < 0.001$) and AB scores after the playback presentation (Spearman's rho: $r = 0.96$, $N = 33$, $P < 0.001$).

One male subject was excluded from subsequent analyses as he did not look at either assistant after hearing the playback voice, giving a total of 50 subjects in the statistical analyses.

**Statistical Analysis**

*Pre-Playback Behaviour*

To determine if the subjects showed any orientation biases prior to the presentation of the auditory stimuli, we ran a Mixed Factorial ANOVA to test for effects of the gender of the assistant, and/or the side on which they were stood, on the total gaze duration towards each assistant during the first 10 s before the playback voice was presented.

*Playback Response Scores*

Binomial probability tests were carried out on the Total gaze and First look correct scores to determine if the proportion of correct responses significantly differed from the expected 50% chance level. In order to test the effect of potentially relevant independent variables (IVs) on correct scores, we ran binary logistic regressions with subject's sex, side of the correct gender-matching person, gender of the playback voice, test location and use of headphones as categorical predictors and the subject's age, number of adult people living with the subject (household composition), average number of hours the subject spent with people per day, and the difference in height (cm) between the male and female assistants as continuous predictors. Interactions were included between each of the variables related to the subject with variables related to the experimental procedure (gender of the playback voice, side of the correct gender-matching person and assistant height difference). A forwards stepwise method with a likelihood ratio statistic was used to construct the model by including significant IVs. The same binary logistic regression analyses were also repeated using the categorically coded version of household composition (HC Group). Planned comparisons were then conducted on the significant IVs, using HC Group in order to identify differences at a group level.

To investigate the potential mechanisms underlying differences in behavioural responses, a Mixed Factorial ANOVA was performed to test whether the anxiety levels

displayed by the subjects (measured by the appeasement behaviour (AB) score), differed before or after the presentation of the playback voice (Time), between HC Group 1 and 2 (HC Group), or depended on the side of the correct gender-matching assistant (Side). Finally, in order to test whether gaze aversion, a common appeasement signal produced by dogs (Rugaas, 2005), could explain observed differences in performance, we tested whether dogs from smaller households (HC Group 1) spent more time looking away from either the correct and/or both assistants than dogs from larger households (HC Group 2).

All analyses were conducted using SPSS version 19 (SPSS Inc., Chicago, IL, U.S.A.).

## Results

### Pre-Playback Gazing Behaviour

Analysis of dogs' behaviour during the 10 s prior to playback presentation showed that there was no significant difference in the total gaze duration towards the assistant stood on the subject's left (Mean+SE = 997.25+259.49 ms, $N$ = 50) or right (Mean+SE = 1274.51+218.87 ms, $N$ = 50) (Two-Way Mixed Factorial ANOVA: $F_{1,48}$ = 0.004, $P$ = 0.95), nor any significant differences in total gaze duration towards the man (Mean+SE = 960.00+192.75 ms, $N$ = 50) or woman (Mean+SE = 1291.76+279.04 ms, $N$ = 50) (Two-way Mixed Factorial ANOVA: $F_{1,48}$ = 0.004, $P$ = 0.95). There was also no interaction between the orientation of the assistants (the man on the right or the woman on the right) and side on the total gaze duration (Two-Way Mixed Factorial ANOVA: $F_{1,48}$ = 2.01, $P$ = 0.16).

### Gaze Responses Following Playback

Analysis of the two 'correct matching' scores showed that overall, the proportion of dogs responding correctly did not significantly differ from the expected 50% chance (Binomial test: First look correct score (40%): $N$ = 50, $P$= 0.20; Total gaze correct score: (50%): $N$ = 50, $P$ = 1.00).

However, the binary logistic regression analyses revealed a significant positive correlation between the number of adults living with the subject (household composition) and the proportion of correct responses for both the Total gaze correct scores (Binary logistic regression: $Wald_1$ = 7.76, $P$ < 0.01) and the First look correct

scores (Binary logistic regression: $Wald_1 = 7.36$, $P < 0.01$) (Figure 3). There was also a significant interaction between household composition and the side of the correct gender-matching assistant on the proportion of correct responses for the Total gaze correct scores (Binary logistic regression: $Wald_1 = 4.45$, $P < 0.05$) and the First look correct scores (Binary logistic regression: $Wald_1 = 7.85$, $P < 0.01$). There was no effect of the subject's age or sex, average number of hours spent with people, difference in height between the assistants, gender of the playback voice, test location, use of headphones or any of the other interactions entered on the proportion of correct responses for either response variables and these IVs were not included in the final models. Together household composition and the interaction between this variable and the side of the correct gender-matching assistant accounted for 24% of the variation in the Total gaze correct scores, and 35% of the total variation in the First look correct scores (Cox and Snell $R^2$). Equivalent results were obtained for both response variables when the same analyses were run using the categorical version of household composition (HC Group), as there was a significant main effect of HC Group (Binary logistic regression: First look correct score: $Wald_1 = 6.56$, $P < 0.05$; Total gaze correct score: $Wald_1 = 9.14$, $P < 0.01$) and a significant interaction between HC Group and the side of the correct assistant (Binary logistic regression: First look correct score: $Wald_1 = 5.78$, $P < 0.05$; Total gaze correct score: $Wald_1 = 6.17$, $P < 0.05$) on the proportion of correct responses.

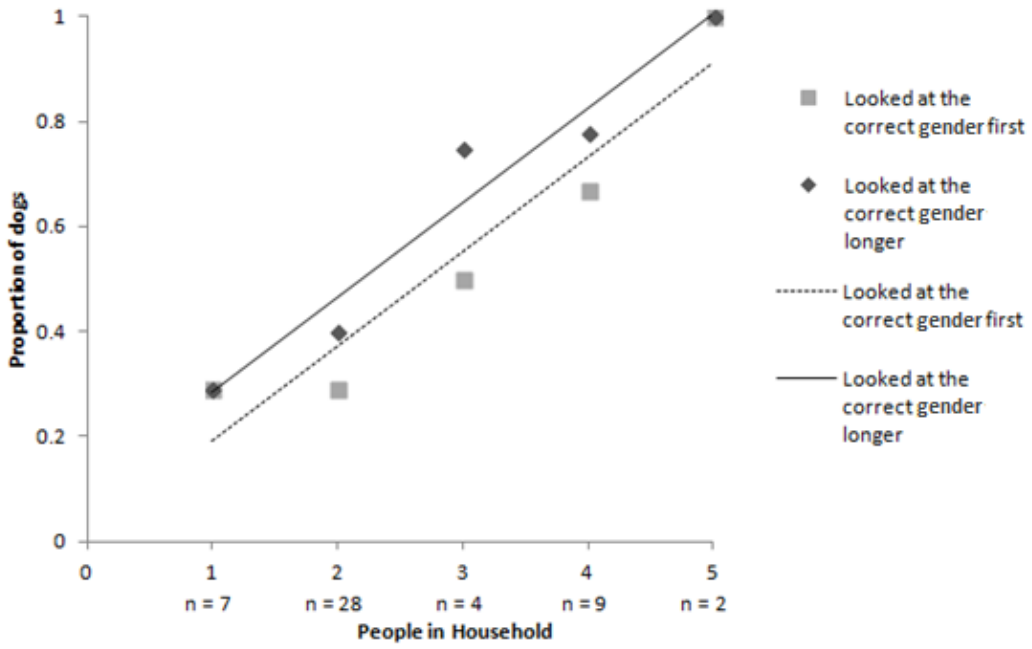



**Fig. 3** The proportion of dogs that looked at the correct gender-matching person first and for longer depending on the number of adult people in their household. N refers to the number of dogs per group.

Planned comparisons showed that dogs living with one adult or one man and one woman (HC Group 1) performed at chance level in their Total gaze correct scores (37% correct, Binomial test: $N = 35$, $P = 0.18$). However significantly more of the dogs living with three to five adults (HC Group 2) responded correctly (Fisher's Exact Test: $N = 15$, $P < 0.05$), which was also significantly more correct responses than expected by chance (80% correct, Binomial test: $N = 15$, $P < 0.05$). Analysis of the First look correct scores revealed that subjects in HC Group 1 performed significantly below chance (29% correct, Binomial test: $N = 35$, $P < 0.05$). Although HC Group 2 performed significantly better than HC Group 1 (Fisher's Exact Test: $N = 15$, $P < 0.05$), here their performance was not significantly above chance (67% correct, Binomial test: $N = 15$, $P = 0.30$).

Separate Binomial tests were then conducted for each HC Group depending on which side was correct. The side of the correct gender-matching assistant was evenly distributed across both groups. For both variables, when the correct gender-matching person was stood on the left side of the subject, HC Group 1 performed significantly below chance, whilst HC Group 2 performed at chance level. In contrast, when the

correct gender-matching person was stood on the right, HC Group 1 performed at chance level, whilst HC Group 2 performed significantly above chance (Tables 1 and 2).

**Table 1:** Percentage of correct responses for Total gaze correct scores depending on the Household Composition Group and the side of the correct gender-matching person.

| **HC Group** | **Side of correct gender-matching person** | ***N*** | **Observed correct responses (%)** | **Binomial Test Significance** |
|---|---|---|---|---|
| 1 (≤ 2 adult people in household) | Right | 17 | 53% | $P = 1.00$ |
| | Left | 18 | 22% | $P < 0.05$ |
| 2 (≥ 3 adult people in household) | Right | 6 | 100% | $P < 0.05$ |
| | Left | 9 | 67% | $P = 0.51$ |

**Table 2:** Percentage of correct responses for First look correct scores depending on Household Composition Group and the side of the correct gender-matching person.

| **HC Group** | **Side of correct gender-matching person** | ***N*** | **Observed correct responses (%)** | **Binomial test Significance** |
|---|---|---|---|---|
| 1 (≤ 2 adult people in household) | Right | 17 | 47% | $P = 1.00$ |
| | Left | 18 | 11% | $P < 0.01$ |
| 2 (≥ 3 adult people in household) | Right | 6 | 100% | $P < 0.05$ |
| | Left | 9 | 44% | $P = 1.00$ |

**Effect of Anxiety/Gaze Aversion on Observed Results**

There was a significant main effect of HC Group on AB scores (Two-way Mixed Factorial ANOVA: $F_{1,46} = 5.11$, $P < 0.05$). Dogs in HC Group 1 had significantly higher AB scores (Mean+SE = 2+0.34, $N = 35$) than dogs in HC Group 2 (Mean+SE = 1+0.27, $N = 15$). There was no significant main effect Time (Two-way Mixed Factorial ANOVA: $F_{1,46} = 0.31$, $P = 0.58$) or Side (Two-way Mixed Factorial ANOVA: $F_{1,46} = 1.01$, $P = 0.32$). None of the interaction terms were significant.

The total duration of time spent looking elsewhere (at the loud speaker and away from both assistants) after the playback presentation was not significantly different (Independent measures *t* test: $t_{48}=0.53$, $P = 0.60$) between dogs in HC Group 1 (Mean+SE =10922.00+522.19 ms, $N = 35$) and dogs in HC Group 2 (Mean+SE = 11402.67+675.75 ms, $N = 15$). There was also no difference between the first response latencies (towards either assistant) (Independent measures *t* test: $t_{48}=0.77$, $P = 0.45$) of dogs in HC Group 1 (Mean+SE = 1650.00+456.77 ms, $N = 35$) and dogs in HC Group 2 (Mean+SE = 1090.00+303.25 ms, $N = 15$). However, the amount of time spent looking away from the correct gender-matching assistant after the presentation of the playback voice (i.e. time looking at either the loud speaker, away from both assistants or at the incorrect non gender-matching assistant) did significantly differ between HC Groups (Independent measures *t* test: $t_{48} = 2.05$, $P < 0.05$): dogs in HC Group 1 spent more time looking away from the correct gender-matching assistant (Mean+SE = 13750.29+260.79 ms, $N = 35$) than dogs in HC Group 2 (Mean+SE = 12570.00+530.05 ms, $N = 15$). Similarly, dogs in HC Group 1 had significantly slower response latencies to the correct gender-matching assistant (Mean+SE = 7641.43+1000.96 ms, $N = 35$) than dogs in HC Group 2 (Mean+SE = 4965.87+1282.18 ms, $N = 15$) (Independent measures *t* test: $t_{48} = 2.81$, $P < 0.01$). Of the total 50 subjects, 10 dogs (29%) in HC Group 1 did not look at the correct gender-matching assistant at all after the presentation of the playback voice, compared to only 2 dogs (13%) in Group 2. Therefore, rather than looking away from both assistants, dogs in HC Group 1 spent less time looking at the correct gender-matching assistant and more time looking at the incorrect non gender-matching assistant than dogs in HC Group 2.

## Discussion

Our results showed that dogs living with more people were significantly more likely to look towards a person of the same gender after hearing an unfamiliar human voice. Significantly more of the dogs living with three or more adults (including at least one man and woman) looked first, and for longer, towards the unfamiliar person that matched the gender of the voice than at the non-matching person. Significantly more of these dogs (80%) also looked longer at the correct gender-matching person than was expected by chance. Conversely, a significantly larger proportion of the dogs living with one or two adult people (71%) looked at the incorrect non gender-matching person first than was expected by chance. Overall performance was not the result of a general response preference to either men or women, as the gender of the playback voice did not influence the number of correct responses. There was also no effect of the subject's age or sex on the proportion of correct responses. Finally, the average amount of time per day the subject spent with people was also not found to predict performance, suggesting that regular exposure to a wider variety of people influenced responses more strongly than the quantity of time spent in human company. However, significant interactions were found between the number of people living with the subject and the side on which the gender-matching person was stood for both the scores for the direction of the first look and the total gaze duration. If the correct gender-matching person was stood on the right side of the subject, dogs living with three or more adult people performed significantly above chance, whilst those living with one or two adults performed at chance level. However, if the correct gender-matching person was stood on their left side, dogs living with three or more adult people performed at chance level, whilst those living with one or two adults performed significantly below chance.

Other studies have also found generally higher performance levels in audiovisual matching tasks when stimuli are viewed on the right side (rhesus macaque (*Macaca mulatta*): Gaffan & Harrison, 1991; bottlenose dolphin (*Tursiops truncatus*): Delfour & Marten, 2006; domestic horse: Proops & McComb, 2012). In these cases it is thought that the left hemisphere may be more strongly recruited for 'matching with sample' tasks and in identifying familiar stimuli (Rogers, 1997; Vallortigara et al., 2008). Furthermore, findings that left hemispheric lateralisation is evident in responses to conspecific or familiar vocalisations in a variety of species (e.g. domestic dog: Siniscalchi, Quaranta & Rogers, 2008; California sea lion (*Zalophus californianus*):

Böye, Güntürkün & Vauclair, 2005; Rhesus macaque (*Macaca mulatta*): Hauser & Andersson, 1994) are consistent with the potential influence of the playback voice on orientation biases. In the current study it is apparent that although this general response bias towards the right side occurred across the subjects, the effect of this bias on their performance differed according to their level of regular exposure to people. We suggest that dogs' previous social experiences with people differentially affect how they respond to an unfamiliar person. Dogs living with a larger number of people tended to look towards the correct gender-matching person, as shown by their accurate performance if the correct gender-matching person was stood on their right, but were also more likely to look towards the person on their right side overall, reducing their performance to chance level if the correct gender-matching person was stood on their left side. In contrast, dogs living with fewer people appeared to avoid looking at the correct gender-matching person, as evidenced by their below chance level performance when this person was stood on their left, but also showed a right side response bias, resulting in their performance at chance level when the correct gender-matching person was stood on the right side.

This interpretation is supported by our finding that after the presentation of the playback voice, dogs living with fewer people looked at the correct gender-matching person for significantly less time than those living with more people, by instead looking more quickly and for longer towards the incorrect non gender-matching person. Gaze avoidance is a coping mechanism used in stressful social situations by both humans and animals (Thompson & Waltz, 2010; Koolhaas et al., 1999). Because direct eye-contact is maintained during dominance displays, dogs use gaze aversion as an appeasement signal to prevent conflict during social interactions with other dogs (Bradshaw & Nott, 1995), and have been shown to avoid making eye contact when approached in a threatening manner by an unfamiliar person (Vas, Topál, Gácsi, Miklósi, & Csányi, 2005; Györi, Gácsi, & Miklósi, 2010). The fact that during our study, dogs living with fewer people also produced significantly more appeasement signals (e.g. licking the muzzle, yawning) than those living with more people suggests that these dogs had greater levels of social anxiety (Rugaas, 2005). Although attempts were made so that the people in the current study did not appear threatening (they did not move or look at the subject), dogs are generally more likely to show a combination of appeasement and defensive behaviour, including gaze aversion, towards unfamiliar people (Rappolt,

John, & Thompson, 1979). Shelter-housed dogs which have had less social experience with people are also more likely to show fear-appeasement behaviour in response to unfamiliar people than dogs with more experience (Barrera, Jakovcevic, Elgier, Mustaca, & Bentosela, 2010). Although human-directed gaze in dogs is strongly affected by previous reinforcement (Bentosela, Barrera, Jakovcevic, Elgier, & Mustaca, 2008), sociability also plays an important role in looking towards unfamiliar people, as after receiving positive reinforcement training for gazing at the experimenter's face, dogs scoring higher in their level of sociability towards an unfamiliar person gazed significantly longer at the experimenter's face during extinction trials, when the behaviour was no longer reinforced (Jakovcevic, Mustaca & Bentosela, 2012). Therefore, we suggest that dogs living in smaller households were more socially anxious during the study, and therefore were more likely to direct their gaze away from the more salient person who they perceived was speaking.

Zangenehpour, Ghazanfar, Lewkowicz and Zatorre (2009) also found this 'reverse effect' in a similar cross-modal paradigm with vervet monkeys, *Cercopithecus aethiops*. Subjects were presented with two videos of rhesus monkeys producing different call-types and heard vocalisations matching one of the videos. The vervet monkeys indicated their ability to match the visual and auditory information by looking significantly longer at the incorrect image, and this result was attributed to gaze aversion due to the higher negative emotional salience of the correct image. This concurs with our interpretation of our observations in the current study, and further demonstrates that the perceived emotional salience of a stimulus can result in significant differences in responses during cognitive tasks. Thus our findings stress the importance of accounting for individual life history when investigating cognitive skills in non-human species.

Despite the strong influence of experience on expression, overall the dogs in the current study showed their ability to correctly discriminate the person that matched the gender of the playback voice. Therefore, our results provide the first demonstration that dogs do spontaneously learn to categorise unfamiliar people as male or female, by associating cues across different sensory modalities. Importantly, these categories are clearly not dependent on any perceptual similarities between cues (as they originate from distinct modalities), which can often explain how animals generalise learning across different stimuli in training paradigms (see Zentall, Wasserman, Lazareva, Thompson and Rattermann (2008) for a review). However, as multiple cues were made available to the

dogs in each sensory modality, we cannot yet determine the specific cues which the dogs associated across modalities, and if these cues are the same as those used by humans to categorise gender. As dogs may rely on different cues to humans when associating objects (Van der Zee, Zulch, & Mills, 2012), it could be that dogs also use different information to categorise human gender. For example, although dogs are perceptually aware of anatomically derived gender-specific cues (formant spacing and F0) in human voices (Baru, 1975), men and women also use different intonation patterns when speaking to dogs (Prato-Previde et al., 2005). This was also seen in the current study, as the F0 range in the female voices was larger than the male range. Therefore gender differences in intonation could provide an alternative means for dogs to discriminate the gender of human voices than through the use of anatomically derived vocal differences.

Alternatively, it may be that dogs recognise more abstract correspondences between voices and people, such as matching a low-pitched voice to a person with a larger body-size. Although the difference in height between the man and woman was not found to influence responses in the current study, the heights were bimodally distributed, and body weight may have also been a contributing factor. Dogs can match growls to conspecifics according to their body-size (Taylor et al., 2011); thus we cannot discount the possibility that dogs also match voices to people based on body-size differences rather than gender-specific cues. Further research is therefore necessary to determine more precisely how dogs learn to categorise human gender, and the extent to which this functionally equivalent category is comparable to the way that we categorise human gender.

Although the bases of this ability remain to be established, our observations suggest that dogs can categorise human gender in both visual/olfactory and auditory modalities. This is consistent with reports that dogs behave differently towards unfamiliar people depending on their gender, often by responding more negatively towards men (Lore & Eisenberg, 1986; Hennessy et al., 1998; Wells & Hepper, 1999), including biting men significantly more often (e.g. Rosado, García-Belenguer, León, & Palacio, 2009). While there do not appear to be gender differences in owner attachment levels towards dogs (Prato-Previde et al., 2005), male and female owners do differ in their interaction style with dogs, as men speak to their dogs less frequently (Prato-Previde et al., 2005) and are less likely to perceive their dog as being stressed (Mariti et al., 2012). It is possible that

gender-specific behavioural differences may create a need for dogs to categorise men and women in order to adapt their responses appropriately. Determining more specifically how and why dogs learn to categorise men and women has important practical implications for understanding their responses to different people. Whilst our study has demonstrated that multisensory cues, including vocal cues, are associated by dogs, we have yet to determine which specific cues are used and how these may influence responses.

**Conclusion**

Our findings illustrate that dogs can spontaneously categorise human gender by associating cues across sensory modalities. The strong influence of the dogs' social experience with humans in the expression of this ability also highlights the important issue of accounting for life history as a source of individual variation in the natural expression of cognitive abilities by non-human species. Investigating how animals perceive and categorise their social environment is a crucial step towards understanding the nature of interactions between domesticated animals and humans.

# CHAPTER 4: AUDIO-VISUAL CORRESPONDENCES IN DOGS' DISCRIMINATION OF HUMAN SPEAKERS

## Article I: Cross-Modal Correspondences in Non-Human Mammal Communication

### Synopsis

---

**Question:** How do non-human animals form cross-modal associations during their perception of multisensory signals?

---

**Background:** Humans perceive a broad range of cross-modal correspondences that provide a more cohesive impression of sensory information from the environment. Sensory percepts are combined with varying levels of automaticity and complexity, from the recognition of general low-level temporal synchrony to the application of stored conceptual multisensory representations.

---

**Review Aims:** Non-human mammals, including domestic dogs, make functional cross-modal associations during inter- and intra-specific communication. The current review aimed to determine what is currently known about the perceptual and cognitive mechanisms underlying these associations, and how they relate to human cross-modal correspondences.

---

**Conclusions:** In addition to matching temporally synchronised stimuli and redundant information across sensory modalities, a range of highly social mammal species appear to form higher level cognitive representations about other individuals. This evidence suggests that some mammals, including dogs, may be capable of forming complex representations about human signallers as well as conspecifics. Because the perception of cross-modal correspondences between basic stimulus dimensions has been under-researched in animals, the final section of the review proposes a number of potential avenues for exploring cross-modal perception in dogs, some of which are subsequently investigated in Article II of this Chapter.

*Note*. In press as: Ratcliffe, V.F., Taylor, A.M. & Reby, D. (2015). Cross-modal Correspondences in Non-Human Mammal Communication, *Multisensory Research*, in press.

## Abstract

For both humans and other animals, the ability to combine information obtained through different senses is fundamental to the perception of the environment. It is well established that humans form systematic cross-modal correspondences between stimulus features that can facilitate the accurate combination of sensory percepts. However, the evolutionary origins of the perceptual and cognitive mechanisms involved in these cross-modal associations remain surprisingly under-explored. In this review we outline recent comparative studies investigating how non-human mammals naturally combine information encoded in different sensory modalities during communication. The results of these behavioural studies demonstrate that various mammalian species are able to combine signals from different sensory channels when they are perceived to share the same basic features, either because they can be redundantly sensed and/or because they are processed in the same way. Moreover, evidence that a wide range of mammals form complex cognitive representations about signallers, both within and across species, suggests that animals also learn to associate different sensory features which regularly co-occur. Further research is now necessary to determine how multisensory representations are formed in individual animals, including the relative importance of low-level feature-related correspondences. Such investigations will generate important insights into how animals perceive and categorise their environment, as well as provide an essential basis for understanding the evolution of multisensory perception in humans.

## Introduction

Similarly to humans, most non-human animals experience the world through different senses, and the ability to combine this perceptual information functions to reduce uncertainty and create more coherent and meaningful representations of objects and events (Lewkowicz and Ghazanfar, 2009). However, because the brain constantly receives a vast array of sensory input from the environment, it must overcome the 'cross-modal binding problem' of identifying when different perceptual information has originated from the same source and should be combined during processing (Ernst,

2007). Systematic mappings between various features or dimensions perceived through different sensory modalities, termed cross-modal correspondences, can promote the combination of information at the perceptual and/or decisional stages of processing (Parise and Spence, 2013). Although a number of different cross-modal correspondences have been identified in humans, our understanding of their evolutionary origins and adaptive function remains very limited (Ludwig *et al.,* 2011). One of the key difficulties that researchers face is differentiating between innate 'hardwired' and experience driven correspondences, as new associations can develop rapidly between different stimulus dimensions with very small levels of exposure to their co-occurrence (Ernst, 2007; Zangenehpour and Zatorre, 2010).

In recent years, the comparative approach has been widely developed to address such questions in other areas of human perception and cognition, providing important advancements, such as furthering our understanding of human language evolution (Fitch, 2010). By establishing the extent to which non-human animals (henceforth animals) perceive cross-modal correspondences, it may be possible to determine the phylogenetic history of hardwired correspondences and the pre-adaptations that were necessary to support their existence in humans. Investigating the functional relevance of cross-modal correspondences for animals can also provide insights into the evolutionary pressures that promote their occurrence. Furthermore, the importance of ontogenetic experience in the formation of cross-modal correspondences can be more directly tested in animals than in humans, either by comparing species across different environments or by controlling the experiences gained by captive animals (Kulahci and Ghazanfar, 2013). Finally, because animals lack language, it is also possible to rule out the influence of linguistic transmission on the development of any shared correspondences, as the use of the same linguistic labels (e.g., the descriptive terms 'low' and 'high' are used for pitch and elevation) can confound attempts to interpret the origins of systematic associations in humans (Spence, 2011). The comparative approach therefore has a strong potential to significantly enhance our current understanding of the origins and function of cross-modal correspondences in humans.

In this review, we outline the range of cross-modal correspondences that are known to be behaviourally expressed by animals when combining different sensory information. We focus on mammals primarily due to their close evolutionary relationship to humans,

but also because correspondences have been more widely studied in mammals than in other taxa. An additional aim of the review is to provide an ecologically relevant framework for the different types of correspondences observed in animals by determining their potential role in multisensory communication. Because a wide range of species use multisensory signals during communication, these signals can be productively used as stimuli when testing cross-modal correspondences to elicit more natural responses from animals, often without the need for inherently artificial training. Our hope is that as future studies continue to contribute to this framework, a clearer understanding of the evolution of cross-modal correspondences will be developed.

More specifically, in the first section of the review we outline the potential that the multisensory signals used in animal communication have to provide receivers with natural opportunities to express the range of correspondences observed in humans. We then discuss how behavioural methodologies have been applied to show that different animal species associate signal components by attending to broadly shared features, ranging from timing and spatial location to quantity (see Appendix 1 for a detailed discussion of the most commonly used experimental paradigms). In the subsequent sections we discuss evidence suggesting that non-human animals do not just depend on mechanically constrained, co-occurring cues, but that they can also respond to correspondences between different signal features. Although there is currently only limited research on the occurrence of correspondences between distinct basic features (such as visual luminance and auditory pitch) in animals, we discuss potentially productive avenues for future study. In the final section, we show that a wide range of mammalian species appear to develop multisensory cognitive representations about signals and signallers, enabling them to form time-independent expectations about the multisensory composition of communicative stimulus features (see Table 1 for a synthesis of studies).

**Table 1.** Synthesis of the cross-modal correspondences that have been demonstrated in mammalian species in relation to multisensory communication.

| | | **Redundant correspondences** | **Structural correspondences** | **Statistical correspondences** | **Categorical representations** |
|---|---|---|---|---|---|
| **Non-human primates** | | | | | |
| great apes | chimpanzee (*Pan troglodytes*) | | luminance and auditory pitch (Ludwig *et al.*, 2011) | | conspecific call types (Izumi and Kojima, 2004; Parr, 2004) |
| | | | | | conspecific identities (Kojima *et al.*, 2003; Martinez and Matsuzawa, 2009) |
| old-world monkeys | rhesus macaque (*Macaca mulatta*) | conspecific call types (Ghazanfar and Logothetis, 2003) | looming/approaching signals (Maier *et al.*, 2004; Ghazanfar and Maier, 2009) | conspecific body size (Ghazanfar *et al.*, 2007) | conspecific identities (Adachi and Hampton, 2011; Sliwa *et al.*, 2011) |
| | | number of conspecific signallers (Jordan *et al.*, 2005) | | | heterospecific identities (Sliwa *et al.*, 2011) |
| | Japanese macaque (*Macaca fuscata*) | | | | species (both their own species and humans) (Adachi *et al.*, 2006, Adachi *et al.*, 2009) |
| | vervet monkey (*Chlorocebus pygerythrus*) | heterospecific call types (Zangenehpour *et al.*, 2009) | | | |

| | | | | |
|---|---|---|---|---|
| | grey cheeked mangabey (*Lophocebus albigena*) | | | conspecific identities (Bovet and Deputte, 2009) |
| new world monkeys | tufted capuchin (*Cebus apella*) | conspecific call type (Evans *et al.*, 2005) | | |
| | squirrel monkey (*Simia sciureus*) | | | heterospecific identities (Adachi and Fujita, 2007) |
| lemurs | ring-tailed lemur (*Lemur catta*) | | | conspecific identities (Kulachi *et al.*, 2014) |
| **Carnivora** | | | | |
| | domestic dog (*Canis familiaris*) | | conspecific body size (Faragó *et al.*, 2010; Taylor *et al.*, 2011) | heterospecific identities (Adachi *et al.*, 2007) |
| | | | | heterospecific gender (Ratcliffe *et al.*, 2014) |
| **Perissodactyla** | | | | |
| | Domestic horse (*Equus caballus*) | | | conspecific identities (Proops *et al.*, 2009) |
| | | | | heterospecific identities (Proops and McComb, 2012) |

## Multisensory Signals in Animal Communication

Obtaining accurate estimations about certain attributes of conspecifics, such as their body size, is essential in mediating the sexual and social interactions of many species (e.g., Davies and Halliday, 1978; Madden *et al.,* 2009; Reby *et al.*, 2005; Tedore and Johnsen, 2014). Because information about individuals can be acquired through different senses, it is functionally relevant for animal receivers to naturally combine sensory information, which can inform our understanding of the evolution of cross-modal correspondences in humans (Kulahci *et al.,* 2014). In animal communication, information about the individual is broadcast through 'signals', which can be defined as an act or structure that has evolved to change the behaviour of other organisms in way that normally functions to benefit the signaller (Maynard-Smith and Harper, 2003). Whilst signals can be transmitted through a single modality (such as visual displays or long distance acoustic signals), multisensory signals are prevalent in the communication systems of a wide range of vertebrates (e.g., California ground squirrel *Spermophilus beecheyi:* Rundus *et al.,* 2007; brown-headed cowbird *Molothrus ater*: Cooper and Goller, 2004; sagebrush lizard *Sceloporus graciosus*: Thompson *et al.,* 2008; dart-poison frog *Epipedobates femoralis*: Narins *et al.,* 2003) and invertebrates (e.g., wolf spiders *Lycosidae:* Uetz and Roberts, 2002; big-clawed snapping shrimp *Alpheus heterochaelis*: Hughes, 1996). Although multi-component signals are typically more costly for animals to produce than single-component signals (Bradbury and Vehrencamp, 1998), they function to overcome production and/or perceptual constraints on transmission (see Bo-Jørgensen, 2009 for a review). For example, redundant (or 'amodal') information is frequently encoded across different sensory components (Partan and Marler, 1999), as some signal properties are not modality specific and can be redundantly sensed via different sensory channels. Redundant features include physical attributes such as the spatial location and temporal duration of events or the size and shape of a physical entity (Spence, 2011). Encoding equivalent information across modalities increases the robustness of the signal, providing signallers with 'backup channels' to ensure transmission through environmental noise (Johnstone, 1996) and improving the reliability of the perceptual estimations obtained by the receiver (Ernst and Bülthoff, 2004). Because sampling these properties through different sensory modalities provides the same metric estimate (Marks *et al.,* 1986), each sensory component should elicit the same response from the receiver when presented alone

(Partan and Marler, 1999). However the multisensory combination of redundant cues in animal signals frequently results in an enhanced response (Hölldobler *et al.*, 1996; Smith and Evans, 2008), improving the signal's efficacy by facilitating its detection, discrimination and memorisation by receivers ('receiver psychology hypothesis', reviewed by Rowe, 1999).

As well as facilitating the transmission of redundant information, animal signals can also contain different non-redundant (or 'modal') components (Moller and Pomiankowski, 1993), increasing the amount of information communicated per unit of time (e.g., multisensory begging signals encode independent indices of nestling condition in European starlings *Sturnus vulgaris*: Jacob *et al.*, 2011). In some cases one non-redundant component can modulate or dominate the effect of another, potentially resulting in the emergence of a new response (see Partan and Marler, 1999, for examples). This combinatorial strategy functions to disambiguate or maximise the amount of information contained in the signal (Ernst and Bülthoff, 2004). Evidence of signal enhancement and modification during multisensory communication indicates that different sensory components are not always processed separately, as interactions can occur between redundant or non-redundant cues. Accordingly, researchers have exploited the ecological validity and salience of such signals to investigate the perceptual and cognitive mechanisms involved in the combination of different sensory information by animals (Kulahci and Ghazanfar, 2013). The majority of studies to date have focussed on the association of auditory and visual information, perhaps because the results can be more directly compared to human speech processing (Ghazanfar, 2013).

In humans, cross-modal correspondences can form between equivalent redundant sensory cues, and also between non-redundant features when they are perceived to be complementary or relatively compatible (Spence, 2011). Congruency effects linking non-redundant features include seemingly arbitrary associations between basic stimulus properties (such as auditory pitch and visual angularity) and can be broadly sub-divided into 'structural' hardwired correspondences associated with the fixed organisation of the perceptual system (Marks, 1978), and learnt 'statistical' correspondences that relate to natural correlations in the environment (Marks, 2000). In addition to perceiving congruency between basic stimulus features, humans also form high-level cognitive

correspondences based on shared semantic attributes between the sensory components (Spence, 2011). These main classes of correspondences can facilitate the combination of different sensory information (Parise and Spence, 2013). Because animal multisensory signals can contain both redundant and non-redundant elements, receivers may also benefit from similarly recognising correspondences in order to efficiently combine sensory elements during processing. We will now consider the extent to which animals also perceive different classes of correspondences linking multisensory signal components, by initially discussing if animals associate different sensory percepts by attending to simple shared (or redundant) cues that co-occur due to mechanical constraints on signal production. We explore the importance of joint timing and spatial location, which have previously been termed 'spatio-temporal correspondences' (Spence, 2007), before discussing other redundancies related to the signal content, such as sensory cues to shape or quantity (which we will term 'redundant feature correspondences').

## Spatio-Temporal Correspondences

Because the different sensory components of animal signals typically co-occur in time and space, receivers can take advantage of this constraint by combining components that originate from the same location and/or occur at the same time. For example, provided that auditory and visual stimuli are temporally aligned (Slutsky and Recanzone, 2001), spatially displaced sounds tend to be automatically 'captured' by visual cues and perceived as originating from a closer location to the visual stimulus, which is known as the 'spatial ventriloquism effect' (Bertelson and Aschersleben, 1998; Howard and Templeton, 1966; Vroomen *et al.*, 2001). Spatial ventriloquism not only occurs in humans (e.g., Bertelson and Radeau, 1981), but can also lead to the mislocalisation of auditory cues in rhesus macaques *Macaca mulatta* (Woods and Recanzone, 2004). Because vocal production mechanisms in vertebrates usually result in the co-occurrence of visual and auditory signals, processing spatial and temporal information can support the receiver's ability to combine the sensory percepts together. The use of low-level temporal redundancies when processing vocal signals appears to be a relatively primitive evolutionary trait in vertebrates. Indeed, the temporal synchronisation of male advertisement vocalisations and air sac inflation influences female mate choice in anuran amphibians (Taylor *et al.*, 2011). Mammals generally

broadcast loud vocalisations orally (e.g., dog barks or goat bleats) (Fitch, 2000b), which means that the acoustic signal is usually accompanied by spatially and temporally corresponding facial movements as the signaller opens and closes their mouth. In an early behavioural study of cross-modal association in mammal communication, Ghazanfar and Logothetis (2003) showed that rhesus macaques could match conspecific vocalisations to the signaller by discriminating between facial gestures associated with different call types. Using a preferential looking paradigm, the subjects were simultaneously presented with two videos showing the same conspecific producing either a 'coo' vocalisation or a 'threat' vocalisation. At the same time, one of these two call types was played from a hidden speaker. The subjects looked longer at the video matching the vocalisation, demonstrating their ability to visually discriminate between the facial expressions and match these gestures to the corresponding auditory cues. Similar results have also been obtained with tufted capuchins *Cebus apella* (Evans *et al.,* 2005), suggesting that the ability to associate conspecific vocalisations with the corresponding facial expression is present in both Old and New World primates.

Because vocalisations and their associated facial expressions have the same temporal characteristics (temporal structure, onset/offset times and duration), the perception of temporal synchronisation was proposed to have enabled the primates' multisensory vocal perception in early preferential looking studies (Izumi and Kojima, 2004; Zangenehpour *et al.,* 2009). The fact that both one- to three-day old human infants (Lewkowicz *et al.,* 2010) and 23–65 week old infant vervet monkeys *Chlorocebus pygerythrus* (Zangenehpour *et al.,* 2009) also matched unfamiliar rhesus macaque vocalisations to corresponding macaque facial expressions gave support to this suggestion. Moreover, both human and vervet monkey infants also consistently associated synthetic tones to the macaque facial gestures. In both studies these complex broadband tones matched the onsets/offsets and durations of the two original call types, but did not include any temporal modulation. The formant frequencies were also removed, whilst the fundamental frequency (F0; perceived as the pitch) of both tones was static and based on the average of the mean F0s of the coo and grunt vocalisations, so that the two tones differed from each other only in duration. Therefore, the human and vervet infants' ability to associate these tones with the corresponding facial gestures strongly suggested that they used temporal synchronisation to match the sounds to the signallers. The young age of the infants, coupled with the novelty of the stimuli, also

suggested that the combination of temporally synchronised sensory cues may be a low-level automatic process in both humans and other primates, potentially allowing receivers to associate information from multiple modalities without any prior experience with their co-occurrence.

Interestingly, the same paradigm had previously been used to show that, while four- and six-month old human infants responded equivalently to neonates by correctly matching the macaque vocalisations with the correct facial expressions, eight to ten-month-old human infants did not (Lewkowicz and Ghazanfar, 2006). The age-related decline in performance supports the theory that whilst humans rely on an innate perception of low-level inter-sensory relations (e.g., temporal synchrony) during their first few months of life, their perceptual sensitivity subsequently narrows to combine only socio-ecologically relevant signals as specific higher-level relations are learnt during development (Lewkowicz and Ghazanfar, 2009). However, unlike in human infants, there was no age-related decline in performance observed in the vervet monkeys, indicating that perceptual narrowing either does not occur in this species, potentially due to the more precocial nature of their neurological system, or that perceptual narrowing does occur but at a much slower rate than in humans (Zangenehpour *et al.*, 2009). The fact that accurate recognition of conspecific call types takes around four years to develop in vervet monkeys (Seyfarth and Cheney, 1986) favours the second hypothesis, leading Zangenehpour *et al.*, (2009) to suggest that mature vervet monkeys should be tested using the same paradigm to determine if they do show evidence of perceptual narrowing through a decrease in reliance on temporal synchrony. Indeed, this could help to determine if the associative mechanism used by the adult rhesus macaques to match different conspecific call types in Ghazanfar and Logothetis (2003)'s original study was related to simple timing or functional differences between the vocalisations.

In non-human primates, temporal synchronisation appears to influence audio-visual signal combination at the early stages of processing. By recording local field potential activity in the auditory cortex in rhesus macaques, Ghazanfar *et al., (*2005) demonstrated that this processing region combined visual and auditory information when subjects were presented with computer generated avatars of conspecifics producing affiliative vocalisations. Whilst voice onset times (VOTs) that were less than 100 ms after the onset of mouth movement caused response enhancement, VOTs longer

than 200 ms instead resulted in response suppression. The importance of VOT in neural responses to multisensory vocal signals was also observed at the behavioural level: whilst macaques predominately focussed on the eye regions of vocalising conspecifics, fixations on the mouth were synchronised with the onset of mouth movements (Ghazanfar *et al.,* 2006). However, although mouth movements appear to be both neurologically and behaviourally relevant during primate vocal perception, changes in the response magnitude of the auditory cortex did not generalise to simple dynamic shapes matching the mouth movements associated with the vocalisations (Ghazanfar *et al.,* 2005). This observation suggested that multisensory processing in the auditory cortex may be specific to biologically relevant faces and not responsive to other temporally synchronised visual and auditory cues. The level of activation was also influenced by the call type, with more extensive enhancement observed in response to grunts rather than coos. The authors speculated that face/voice associations may be more likely to occur in response to grunts because these are generally close range vocalisations directed towards specific individuals, whereas coos are contact calls which are broadcast to the group. The potential role of experience in mediating audiovisual processing provides some support to Zangenehpour *et al.*'s (2009) suggestion that at least in mature primates, higher-level cognitive correspondences such as the functional relevance and production context of multisensory signals may moderate the extent to which different cues are combined together.

The fact that different neurological responses were observed in macaques depending on the nature of the stimuli suggests that higher-level cross-modal correspondences may also affect how non-human primates associate temporally synchronised vocalisations and facial gestures. Such effects have been identified in humans, specifically during the perception of audio-visual speech sounds (Vatakis and Spence, 2007; Vatakis and Spence, 2008). One of the strongest demonstrations of the influence of visual cues on speech perception is the McGurk effect (McGurk and MacDonald, 1976). In this study, participants were asked to repeat the consonant-vowel syllables that they heard whilst watching a video of a person speaking. Though the videos and sounds were temporally synchronised, the syllables produced had different initial consonants that are not formed with the same place of articulation. When presented with an auditory bilabial /ba/, and a visual velar /ga/, participants reported hearing an intermediate alveolar /da/ sound, perceiving a new percept which was a blend of the seen and heard utterance. There is

some mixed evidence suggesting that the magnitude of the McGurk effect may be disrupted if the speaker's voice and face are not identity- or gender-matched (Walker *et al.*, 1995, although see Green *et al.*, 1991). More robust support that gender correspondence can influence the perception of VOT in audio-visual speech comes from studies showing that participants find it easier to judge whether the visual or auditory onsets of speech signals begin first when the stimuli are gender-mismatched (Vatakis and Spence, 2007). Interestingly, the 'unity' effect observed in human responses to congruent audio-visual speech events does not extend to VOT judgements of monkey vocalisations or even to human impersonations of monkey vocalisations, suggesting that higher-order cognitive correspondences may only facilitate multisensory integration for species-specific vocalisations (Vatakis *et al.*, 2008). To date, no studies have tested whether animals' perception of auditory vocalisations can be similarly changed by mismatched, synchronised articulatory cues, or whether they would also differentially perceive the relationship between audio-visual vocal stimuli depending on the availability of additional correspondences.

As the McGurk Effect demonstrates, humans not only attend to the gross temporal synchronisation of visual and auditory stimuli in order to combine different sensory signals (i.e., the similarity between the onset and offset of the signals), but also use the level of cross-correlation between the fine temporal structure within the signals to infer whether they both originated from the same source, even when the signals are not synchronised (Parise *et al.*, 2012). Attending to the fine-scale temporal structure of audiovisual signals is functionally relevant for human communication because speech is a highly rhythmic signal, producing a strong correlation between the movements of the mouth and the acoustic output (Ohala, 1975). Therefore, it is possible that humans may use the fine temporal structure produced by the speech rhythm to match auditory speech to the corresponding signaller if the temporal synchronisation is disrupted. Given that other primates do not produce rhythmic vocalisations (Ghazanfar, 2013), and show a more limited perception of rhythmic sequences (Merchant and Honing, 2013), it is not clear if they would also attend to the detailed temporal structure of audiovisual signals to combine the individual sensory components.

Relatively coarse temporal synchronisation related to the onset and offset of the signal components thus seems to be used generally across vertebrates to associate

vocalisations with signallers during communication. Further work is necessary to determine if other timing-related attributes such as the detailed temporal structure can also influence multisensory perception in animals, as well as to investigate the potential effect of spatial co-occurrence on signal combination. However, despite the evident influence of temporal characteristics on signal processing, it appears that increasing experience with conspecific vocalisations may lead to a reduction in reliance on low-level temporal features in some species. In the following sections, we will explore the extent to which correspondences related to the intrinsic attributes of objects and events may mediate the importance of spatial or temporal co-occurrence for signal combination.

## Redundant Feature Correspondences

Because environmental conditions can impede the transmission of signal components from particular sensory modalities, it is not always possible for receivers to rely solely on the degree of temporal congruency to combine signals. Humans are still able to associate signals even when they do not co-occur, because the perception of additional qualitative or quantitative cross-modal correspondences can bias the brain towards combining certain information together, reducing its sensitivity to inter-sensory conflicts such as spatio-temporal asynchrony (Parise and Spence, 2009).

Before we can determine if animal perceptual systems can be similarly biased towards combining asynchronous signals due to their perceived congruency, we must first establish whether animals also attend to other correspondences that are available during signal production. Indeed, the multisensory signals used by mammals frequently contain additional redundant correspondences that are used to associate individual signal components together. For example, quantitative redundant correspondences can be perceived when the same number of components is simultaneously sensed through different modalities. Rhesus macaques are able to associate the number of conspecific voices they hear with the number of vocalising faces they see, suggesting that they perceive numerosity as a shared redundant attribute across the visual and auditory modalities (Jordan *et al.,* 2005). However, it is yet to be determined if this association was specific to the number of facial gestures or more generally related to the number of conspecifics observed. To investigate this further, future studies could test whether any

species are able to perform this task when some of the conspecifics they can see are not vocalising.

In addition to quantitative dimensions, redundant correspondences may also be perceived using the qualitative features of animal signals. Whilst we will discuss how cues relating to the body-size of the individual are encoded across acoustic and visual percepts at a later stage, differences between the reliability of these cues means that the same metric estimate cannot be obtained across the modalities. Therefore we have not classified the association of size cues in animal signals as a redundant correspondence. Although not related to communication, solid physical bodies also have a size and shape that can be redundantly sensed through vision and touch. Gunderson *et al.,* (1990) observed that normally developing infant pigtailed macaques *Macaca nemestrina* could associate tactile and visual sensory information about object features, and proposed that this ability was potentially related to the discriminability of the outer contours of the objects. The cross-modal congruency of redundant object shape features has also been demonstrated in bottlenose dolphins *Tursiops truncatus* through the association of visual and echoic information (Herman *et al.,* 1998). In a subsequent study, Harley *et al.,* (2003) observed that dolphins found it more difficult to match different novel objects across sensory modalities than to match the same novel object, supporting the hypothesis that dolphins do not simply learn to associate echoic sounds with objects, but instead extract meaningful shape-related characteristics from the echoic and visual information. This suggests that the association of shape-related features may be 'hard-wired', in accord with the observation that 29-day-old human infants are already able to visually recognise the shape of a pacifier after exploring it orally (Meltzoff and Borton, 1979). However the results obtained by Meltzoff and Borton (1979) have not been replicable (Maurer *et al.,* 1999), which coupled with the demonstration that adults newly treated for congenital blindness fail to immediately visually recognise previously handled objects (Held *et al.,* 2011), suggests that the association of shape-related cues may actually be learnt, at least in humans. Further research is needed to clarify the basis of this form of correspondence, and to determine whether shape based associations can be related to the perception of communicative cues. For example, humans tend to systematically match particular nonsense words to simple abstract shapes according to their angularity (e.g., the sound 'kiki' contains sharp phonemic inflections and is usually associated with spiky shapes, whilst 'bouba' contains rounded phonemic inflections and

is mapped onto round shapes — Köhler, 1929; Ramachandran and Hubbard, 2001), independently of cultural influences (Bremner *et al.,* 2013). Consistent pairings between arbitrary sounds and object features, known as the 'sound symbolism' effect, can assist human listeners in guessing the meaning of novel words (Parault and Parkinson, 2008) and facilitates the learning of word-category associations (Monaghan *et al.,* 2012). Japanese mothers also use sound-symbolic words more frequently in speech directed towards their children (Nagumo *et al.,* 2006), which may play a scaffolding role in language acquisition. Consistent with these observations, Ramachandran and Hubbard (2001) suggested that sound symbolism provides a perceptual basis for the sound-referent mappings required for the evolution and acquisition of human language. It is not yet known if this tendency is a linguistic adaptation and unique to humans, or whether other animals would similarly spontaneously associate arbitrary speech sounds with objects according to a perceived correspondence between particular phonemes and physical shape. If sound symbolism effects are present in other species, it could be possible for human speakers to take advantage of such predispositions when training animals.

Together, these studies demonstrate that non-human mammals are able to perceive and associate redundant stimulus features and dimensions that can be encoded within multisensory signals. Although it remains possible that in some cases temporal or spatial synchronisation is necessary for individuals to initially learn that additional sensory redundancies are reliably encoded within certain signals, these redundancies may then moderate the necessity of spatio-temporal synchronisation for signal combination. Further research is now needed to determine how generalised redundant feature correspondences are in animals, and if qualitative associations are applied during communication.

## Structural Correspondences

Besides redundant estimations such as those described previously, it has been suggested that complementary correspondences can also arise between different stimulus properties as a result of the principle of neural economy, whereby shared processing resources respond to multiple stimulus features, resulting in their perceived equivalence (Spence, 2011). In both humans and other animals, magnitude-related, or 'prothetic',

dimensions (e.g., numerosity, area, spatial length, duration, luminance and intensity) are represented using an analogue format, where representations of larger values become increasingly noisy (Cantlon, 2012; Srinivasan and Carey, 2010). Indeed, in most of the species in which quantitative discriminations have been studied, their estimations of 'more' or 'less' appear to obey Weber's law, as their ability to discriminate between two quantities depends on the ratio between them rather than the absolute difference (time: Gibbon, 1977; space: Cheng, 1990; number: Perdue *et al.,* 2012). Because the same estimation principle governs different magnitude-related dimensions, this suggests that they are structurally aligned in the perceptual system, which may facilitate correspondences between different dimensions.

One of the most relevant magnitude dimensions for animal vocal communication is the intensity level of the stimulus, as rising intensity sounds can indicate approaching signallers (Ghazanfar *et al.,* 2002), whilst a greater vocal amplitude generally corresponds with a higher level of arousal across mammals (Briefer, 2012). Stevens (1957) noted that increases in stimulus intensity generally elicit increased neural firing, and Marks (1989) suggested that correspondences between equivalently intense stimuli might arise from the use of a common neurophysiological code, such as the number of impulses per unit of time. In his recent review, Spence (2011) claimed that structurally dependent associations related to intensity coding constitute one of the major forms of cross-modal correspondence in humans. In support of the innate structural basis of intensity relations, human infants are attentive to intensity correspondences very early in development, as they perceive equivalence between the intensity levels of white-lights and white-noise at three weeks of age (Lewkowicz and Turkewitz, 1980). Comparable intensity relations have also been observed in other primates. For example, Ludwig *et al.,* (2011) demonstrated that similarly to human participants, chimpanzees *Pan troglodytes* associated high pitch sounds (which both humans and primates naturally perceive to be more intense/louder than low pitch sounds; Moore, 1989; Stebbens, 1966) with stronger visual luminance, as their performance in classifying squares according to luminance was better when they heard a background tone with a congruent pitch rather than an incongruent pitch. Ludwig *et al.* suggested that because the chimpanzees in this study had not had prior opportunities to learn to associate auditory pitch with brightness, this form of cross-modal association was likely to be innate. However, Spence and Deroy (2012) argued that the chimpanzees could have

internalised correlations in their environment, such as sources of illumination coming from above, and the greater potential tendency for smaller objects or bodies, which generally make higher pitched sounds, to be found in the sky. They also pointed out that the transitive nature of correspondences might have allowed the chimpanzees to acquire new associations on the basis of other learnt regularities in their environment. Marks (1989) bridges these alternative theories by suggesting that whilst some correspondences may be neurologically 'hard-wired', cognitive development could still determine which dimensions correspond. This possibility could be explored by testing infant chimpanzees or by comparing the responses of captive individuals raised in different environments.

Whilst the origin of the correspondence between luminance and pitch in chimpanzees remains unknown, the direction of the association suggests that it may be based on a shared perception of intensity in both dimensions. Indeed, observations that other primate species similarly respond to intensity relations indicates that equivalent intensity perception across sensory modalities may be broadly present across the primate order. For example, Maier *et al., (*2004) showed that rhesus macaques associated complex tones that rose in intensity with expanding circles, which were thought to be perceived as aversive 'looming' or approaching stimuli by the macaques. Furthermore, macaques also associated rising frequency tones with expanding circles (Ghazanfar and Maier, 2009). A related effect known as the 'doppler illusion' is observed in humans: listeners report an increase in the pitch of a sound source moving towards them even though there is no change in the actual frequency of the sound (Neuhoff and McBeath, 1996). However, although the macaques did not have any prior experience with the stimuli used in either study (Ghazanfar and Maier, 2009; Maier *et al.,* 2004), it was not possible to establish whether the association between rising intensity and frequency with increasing size in multisensory looming signals is innately present, or dependent on experience. Therefore, the extent to which intensity-based associations represent fixed structural correspondences remains to be established.

The observations that animals tend to combine signals that share the same level of intensity suggests that other correspondences between magnitude dimensions could similarly influence signal combination. Indeed, although less specifically related to communication, according to the A Theory Of Magnitude (ATOM) framework

proposed by Walsh (2003), time, space and number are equivalently processed by a common analogue magnitude system in the mammalian inferior parietal cortex. The main function of this generalised system is hypothesised to provide an estimate of 'how far, how fast, how much, how long, and how many' with respect to motion. This general magnitude system may be operational in humans from the early stages of development, as Lourenco and Longo (2010) observed that nine-month-old infants mapped arbitrary visual patterns across different dimensions of magnitude, forming an expectation that if a particular pattern was associated with large shapes, then objects with the same pattern should also be more numerous and last longer. Some of these dimensions also appear to correspond in non-human mammals (see Agrillo and Petrazzini, 2013, for a detailed review). For example, rats *Rattus norvegicus* similarly show evidence of perceiving equivalence between estimations of quantity and time (Meck and Church 1983). In this study, rats which were first trained to perform different responses to auditory sequences differing in both the number of elements and the total duration produced identical response curves when they were subsequently tested with stimuli composed of an intermediate number of elements or characterised by an intermediate duration. The results of this study suggest that similarly to human infants, rats may use a general mechanism to represent both time and quantity. Rhesus macaques also show evidence of equivalently processing different magnitude dimensions, as demonstrated by the observation that they naturally confounded the length of lines (space) with how long they were visible for (time) (Merritt *et al.*, 2010).

As well as showing a tendency to associate time and space, humans also represent quantity spatially using a mental number line, with smaller numbers starting from the left, from at least seven months old (De Hevia *et al.*, 2014). Three-day-old domestic chicks *Gallus gallus* similarly appear to associate relatively smaller quantities with their left side and larger quantities with the right space (Rugani *et al.*, 2015). This indicates that in addition to time, numerical magnitude also maps onto spatial cues in both humans and other animals, and may therefore be an ancestral aspect of quantity perception. However, whilst many animals appear to naturally conflate quantity with spatial area (e.g., cats: Pisa and Agrillo, 2009; salamanders *Plethodon*: Krusche *et al.*, 2010), training can lead to a reduction in these effects, as observed in rhesus monkeys (Cantlon and Brannon, 2007) and pigeons *Columbia livia* (Emmerton and Renner,

2006), suggesting that whilst the dimensions of quantity and spatial area are naturally associated, they may not be equivalently processed.

The available research evidence therefore suggests that some aspects of time, space and quantity may be processed by the same mechanism within the mammalian brain, and potentially in more distantly related taxa. The prevalence of similar magnitude-related correspondences across phylogenetically distant species suggests that this potential case of neural economising could be an ancient, conserved adaptation in humans. Whilst the existence of a general magnitude processing system may not be strongly related to associating signals in animal communication, such correspondences could benefit animals in localising and quantifying signals. In contrast, cross-modal correspondences relating to shared stimulus intensities are likely to be functionally relevant in combining the components of multisensory signals, and warrant further investigation in a wider range of species. Future studies are also necessary to establish whether intensity relations are in fact 'hard-wired' structural correspondences in animals, or if they develop as individuals gain experience with regular environmental correlations.

## Statistical Correspondences

Whilst structural correspondences may enable mammals to form associations between complementary stimulus features through the perception of magnitude-related correlations, such 'bottom-up' estimations are inherently noisy, and are therefore likely to lead to ambiguous and unreliable sensory combinations (Ernst and Bülthoff, 2004). Applying a Bayesian integration model, Ernst (2005) suggested that humans act as 'optimal integrators', by combining their prior knowledge that certain stimuli are expected to 'go together' (the coupling prior) with the sensory evidence (the likelihood function) to infer the most reliable interpretation of the environment (Ernst, 2005; Ernst and Bülthoff, 2004). A comparable use of weighted linear estimations, where the weights are proportional to the relative reliability of the cues, has been observed in rhesus macaques (Morgan *et al.*, 2008), suggesting that this strategy may be shared with other mammals.

One way to obtain prior knowledge that stimuli 'belong together' is by attending to their statistical correlation in the environment. Humans can use common environmental

relationships to determine when non-redundant sensory information is likely to have originated from the same source and should be associated. One such statistical correspondence that humans appear to learn is the natural mapping between auditory pitch and visual size, which is likely to occur because there is a strong negative correlation between physical size and acoustic resonance in the environment. For example, larger objects tend to make lower frequency impact sounds when struck or dropped (Gaver, 1993), acoustic waves resonate at lower frequencies when travelling through larger cavities (De Boer, 2008), and the fundamental frequency of a vibrating string is inversely proportional to its length and mass (law of transverse vibrations of a string). Humans consistently generalise this frequency-size relationship, by associating higher-pitched tones with smaller objects and lower-pitched tones with larger objects (e.g., Gallace and Spence, 2006). Although the perceived correspondence between pitch and size could have become genetically hardwired in humans as an adaptation to the environmental correspondence of these variables (Gallace and Spence, 2006), the importance of ontogenetic experience is evidenced by the observation that infants do not form equivalent associations between pitch and size to adults until they are around six-months old (Fernández-Prieto *et al.*, 2015).

The general mapping that humans form between auditory and visual size cues has important functional implications for voice perception. Similarly to the resonances produced by objects in the natural environment, the acoustic parameters in the voice are constrained by the size of the vocal apparatus. According to the 'source-filter theory' (Fant, 1960; Titze, 1994), there are two main sources of size information in the mammalian voice, the fundamental frequency (F0; perceived as the pitch) and the vocal tract resonances or 'formants' (perceived as the timbre). In both humans and other terrestrial mammals, the F0 is produced by the quasi-periodic oscillation of the vocal folds within the larynx. Similarly to the behaviour of a simple vibrating string, longer and denser vocal folds oscillate at a slower rate than shorter and thinner vocal folds under the same level of tension, producing a lower F0 (Titze, 1994; Woods, 1893). Therefore the F0 is inversely proportional to the size of the vocal folds. A second source of size-related information is available from the formants, which are added to the vocal signal when the F0 and associated harmonics (the glottal wave) propagate through the cavities of the supra-laryngeal vocal tract. As the glottal wave passes through it, the vocal tract's resonance properties enhance or dampen the amplitude of certain harmonic

frequencies, producing spectral peaks termed ‘formants’ (Fant, 1960). Because the shape of the mammalian vocal tract is roughly comparable to a uniform cylinder, closed at the glottis at one end and open at the mouth at the other, the primary determinant of the formant frequencies is the vocal tract length, whereby longer vocal tracts produce lower, more closely spaced formants (Titze, 1994).

The pitch of the voice therefore provides listeners with an indication of the size of the vocal folds, whilst information about the vocal tract size is encoded in the vocal timbre. The potential for these acoustic parameters to enable receivers to estimate the signaller’s body size depends on the relationship between either the larynx or vocal tract and the overall body size of the individual. Generally speaking, animals with a larger body size tend to have larger larynges containing longer and thicker vocal folds (Ey *et al.,* 2007; Fitch and Giedd, 1999). However, because the larynx is mostly cartilaginous and only loosely attached to the skull base, it is not strongly constrained by the size of the surrounding skeletal structures (Fitch, 2000a). This allows the larynx to grow out of proportion from other body parts, facilitating selection for size-related adaptations away from a simple scaling ratio with the rest of the body (e.g., male hammerhead bats *Hypsignathus monstrosus*: Kingdon, 1974). Rather than depending on body size, vocal fold growth in humans is believed to be strongly influenced by exposure to androgens, which causes them to thicken and lengthen disproportionally in males during puberty (Harries *et al.,* 1998; Evans *et al.,* 2008). In addition to the weak anatomical association between vocal fold size and body size, the shape of the mammalian vocal folds can be dynamically manipulated both within and between vocalisations by changing their tension through musculature control (see Briefer, 2012, for a recent review), further reducing the relationship between the vocal folds and overall body size. Therefore, due to the relatively unconstrained growth of the vocal folds, as well as their dynamic modulation whilst vocalising, F0 is likely to be a relatively poor correlate with the body size of the signaller.

Although F0 appears to be a limited predictor of individual body size, it generally reflects large size differences across categories of individuals. At the broadest level, across different species, larger animals tend to produce lower F0s, providing an association between size and pitch across all animal vocalisations (Fletcher, 2004). More specifically, within the same species, age-related differences in vocal fold growth

mean that the F0 usually negatively correlates with body size across age categories in mammals (Hillenbrand *et al.,* 1995; Peterson and Barney, 1952). Similarly, in species that have sexually dimorphic body sizes and/or laryngeal sizes, there can be categorical differences in the F0 between adult males and females (e.g., in both humans and baboons *Papio hamadryas,* males are larger than females and have a lower F0; Rendall *et al.,* 2005). However, within members of the same age or sex categories, the relationship between F0 and body-size breaks down for most mammals (e.g., baboons: Rendall *et al.,* 2005; Japanese macaques *Macaca fuscata*: Masataka, 1994; red deer *Cervus elaphus*: Reby and McComb, 2003). Indeed, a recent meta-analysis revealed that in adult humans, the F0 accounted for less than 2% of the variance in height and weight within either sex (Pisanski *et al.,* 2014). Accordingly, F0 has not been observed to influence the size-related judgements of species-specific vocalisations in the two mammalian species which have been studied, and where similarly to humans the F0 does not provide a reliable estimate of body size for adults of the same sex (red deer: Charlton *et al.,* 2008; giant panda *Ailuropoda melanoleuca*: Charlton *et al.,* 2010). The lack of correspondences between pitch and size in animal vocalisations is particularly interesting as it has been hypothesised that animals produce vocalisations with a lower F0 in aggressive contexts as a ritualised exaggeration of body size (Morton, 1977).

Given the lack of reliable correlation between the F0 and body size in human adults, it is surprising that human listeners consistently judge lower pitched adult voices to have a larger body size both within and between the sexes (Feinberg *et al.,* 2005; Pisanski and Rendall, 2011; Rendall *et al.,* 2007; Smith and Pattersen, 2005). Indeed, because of the lack of correlation, listeners are unlikely to learn to map low pitch with large size within adults of the same sex (Pisanski *et al.,* 2014). It has been suggested that similarly to size judgements relating to the resonance of physical objects, the F0 misattribution bias in humans may be the result of a generalisation of statistical pitch-size relationships (Rendall *et al.,* 2007). This generalisation could arise from the actual relationship between voice pitch and body size in humans across age and size categories, whereby adults are lower pitched than children and the average adult female F0, at around 200Hz, is double that of adult males, at approximately 100Hz (Titze, 1994). Alternatively, humans may more generally apply pitch-size correlations learnt from the environment (e.g., object sizes) to human voices. More research is therefore needed to

determine if humans use the same processing mechanisms to judge the pitch-size cues in voices as they do to determine the size of environmental objects.

Animals do not appear to associate pitch and size in vocalisations in the same way as humans do, but instead rely on another vocal parameter that provides a more accurate estimation of size, namely the formants. Indeed, in contrast to the vocal folds, in most mammals the length of the vocal tract is tightly constrained by the skeletal structure (Fitch, 2000a, c), providing in principle a strong positive correlation between the length of the vocal tract and body size in a range of mammals (rhesus macaques: Fitch, 1997; domestic dogs *Canis familiaris*: Plotsky *et al.,* 2013, Riede and Fitch, 1999; humans: Fitch and Giedd, 1999). Although slightly different measures have been used to relate the formant structure to the signaller's body size (e.g., Puts *et al.,* 2012; Reby and McComb, 2003), the majority of studies have shown that the formant structure encodes accurate information about the individual's body-size in a wide range of mammals (e.g., rhesus macaques: Fitch, 1997; red deer: Reby and McComb, 2003; koalas *Phascolarctos cinereus*: Charlton *et al.,* 2011). However, whilst in some mammal species the formant structure can predict a large amount of the variance in body weight (e.g., 62% across dog breeds due to their high level of morphological variation; Taylor *et al.,* 2008), in humans formant related estimates of vocal tract length account for only around 10% of the variance in height and weight for adult men and women (Pisanski *et al.,* 2014), which may be related to the high level of vocal tract flexibility shown during speech production (Cartei *et al.,* 2012; Collins, 2000; Puts *et al.,* 2006). Still, despite their relatively low predictive value, humans do preferentially attend to the formants over the F0 when judging the speaker's body-size if the two variables conflict by equally discriminable amounts (Pisanski and Rendall, 2011).

Animals also assess size-related information from the formant structure of conspecific vocalisations, and some species have been shown to associate this information with the corresponding visual size of unfamiliar individuals. Using a preferential looking paradigm, Ghazanfar *et al.,* (2007) demonstrated that rhesus macaques spontaneously associated conspecific 'coo' vocalisations which had been manipulated to have a smaller formant scaling with images of larger (mature) conspecific faces, whilst they associated vocalisations that had a wider formant scaling with the faces of smaller (juvenile) individuals. The ability to assess size differences between individuals within

the same age category has also been evidenced using the same paradigm in dogs (Faragó *et al.,* 2010; Taylor *et al.,* 2011). The study by Taylor *et al., (*2011) also used resynthesised auditory stimuli where only the scaling of the formant frequencies in the growls were manipulated to change their perceived size, whereas the F0 remained constant across all of the stimuli. Therefore, similarly to the macaques, the dogs used the size-related information encoded within the formants to associate the vocalisations with the different visual stimuli, indicating that they perceived the correspondence of size cues present in each of the sensory modalities. Further investigations are now needed to determine if the ability to associate size cues is innately present in mammals or if it is learnt through regular exposure to the statistical correlation between the formant structure and body size in conspecifics. To investigate this, studies could test whether animals are also able to match vocalisations to body size on the basis of formant frequency spacing in unfamiliar heterospecifics.

In addition to associating auditory pitch and visual size cues, humans also tend to match higher pitched sounds with higher spatial elevations, and lower pitched sounds with lower elevations (e.g., Rusconi *et al.,* 2006) from at least four months of age (Walker *et al.,* 2010). This correspondence appears to automatically influence perception, as low-pitched tones projected from high elevations are actually perceived as originating from low to the ground (known as the Pratt Effect: Pratt, 1930). In a recent study, Parise *et al.,* (2014) observed a consistent mapping between the frequency of sounds in the environment and their source elevation, as high-frequency sounds more frequently originated from higher sources. As well as the frequency-elevation correlation present in the environment, further biases between these dimensions are added during perception for human listeners because the shape of the head and outer ear act as frequency- and elevation-dependent filters (Batteau, 1967), which is known as the head-related transfer function (HRTF). Parise *et al.,* (2014) also analysed the HRTFs produced by the outer ear and determined that sounds coming from high elevations had more energy at high frequencies, accentuating the environmental association between sound frequency and elevation. Human participants were significantly affected by both environmental and head-related elevation biases when localising narrowband sound stimuli, providing strong support for the hypothesis that the pitch-elevation mappings observed in humans develop from natural biases in auditory experience (Parise *et al.,* 2014). To investigate the importance of experience with pitch-elevation correspondences in the environment

in more detail, future studies could determine if there is a difference between the strength of the mappings depending on the elevation. More specifically, it could be hypothesised that because larger physical bodies (producing lower pitched sounds) are normally constrained to low elevations, whilst smaller physical bodies (producing higher pitched sounds) can be found in either high or low elevations (e.g., birds and rodents), the mapping between low pitch sounds and low elevations should be stronger than the mapping between high pitch sounds and high elevations. If this were the case, it would provide additional evidence for the importance of ontogenetic experience in forming this correspondence.

Parise *et al.,* (2014) also suggested that the close association they observed between the anatomically related biases and those present in natural auditory scene statistics could mean that the human ear has adapted to efficiently filter sounds based on natural auditory scene statistics. Whilst to our knowledge pitch–elevation associations have yet to be investigated in animals, differences in pinnae shape and mobility, as well as head shape, between species could be used to test the hypothesis that ear structures adapt to the auditory environment in which the animal lives. However, the possible functional relevance for animals to learn to associate different auditory pitches with specific elevations is currently unclear. Although unrelated to the way that animals match vocalisations to the corresponding signaller, it is interesting to note that some arboreal mammals produce alarm calls which differ in F0 in response to terrestrial or areal predators (e.g., vervet monkeys: Seyfarth *et al.*, 1980a, b; Campbell's monkeys *Cercopithecus campbelli*: Zuberbühler, 2001; red squirrels *Tamiasciurus hudsonicus*: Greene and Meagher, 1998). Although the F0 of alarm calls is not consistently mapped onto terrestrial (low) and areal (high) predators across species, elevation-pitch associations may be functionally relevant in the communication systems of these animals if they can direct receivers' attention to different elevations.

To summarise, currently the only potential statistical correspondence identified in mammals appears to be their ability to associate the formant structure of conspecific vocalisations with the signaller's body size, although the role of experience in the development of this correspondence is yet to be confirmed. However, the lack of research in this area means that additional statistical correspondences may also be present in animals. Moreover, it is possible that some of the associations outlined in

previous sections of this review may be reclassified as statistical correspondences upon further examination. For example, the mapping between luminance and pitch in chimpanzees may reflect either a structural or statistical correspondence, or may even depend on an interaction between the two. The fact that animals can learn more specific correspondences, as we will explore in the next section, implies that they may also learn more general statistical regularities in their environment when it is relevant for them to do so.

## Multisensory Categorical Representations

In addition to learning simple statistical correspondences in the environment, humans also recognise the degree of semantic congruency between stimuli. Higher-level cognitive concepts influence the perceiver's impression of whether signals ought to 'go together' and lead to an assumption of unity between congruent signals. Whilst some degree of awareness of semantic correspondence may be promoted through regular co-occurrences or shared redundant stimulus properties, more complex arbitrary associations between different stimuli can be learnt during development (Spence, 2007). These semantic correspondences contribute to multisensory representations referring to certain physical bodies or events (Doehrmann and Naumer, 2008). Although strongly associated with language in humans, semantic correspondences depend on the perception of shared identity or meaning. Therefore, although they are likely to be qualitatively distinct from semantic correspondences observed in humans, it may also be possible for animals to form semantic correspondences between different sensory information if they also perceive relationships between them.

Semantic correspondences could also be functionally relevant for animals in enabling them to associate signals that occur separately in time or space. However, in order to recognise shared meaning or identity, they would need to be able to access stored information about one modality when another is encountered (Johnston and Bullock, 2001). Storing sensory information could provide some animals with the means to form more complex categorical representations incorporating different sensory information (see Seyfarth and Cheney, 2015, for a recent review). The ability to categorise signal content would convey several advantages over low-level structural or statistical correspondences. Indeed, whilst both mechanisms may help the receiver to locate the

signaller and enhance their perception of information in multisensory signals, categorisation simplifies processing requirements (Rosch *et al.*, 1976) and allows general inferences to be made about the information, which can then be applied to new category members. This would be particularly beneficial in processing multisensory signals when information from all of the sensory modalities is not available, for example in long range vocalisations when the signaller is likely to be out of view.

Returning to the observation that rhesus macaques associate vocalisations with the corresponding signaller depending on the call type produced (Ghazanfar and Logothetis, 2003), although the macaques in the study could have responded correctly by perceiving the temporal synchronisation between the corresponding auditory and visual signals (as inexperienced human and vervet monkey infants appeared to do), it could also be the case that they actually perceived semantic congruency between signals related to the same call type. Investigating vocal perception in a different primate species, Izumi and Kojima (2004) proposed that the multisensory perception of call types in chimpanzees may not be limited to low-level redundant features, but could also depend on a cognitive mechanism enabling them to recognise the categorical congruency of different sensory signals that are related to the same call type. This theory was based on their observation that chimpanzees were able to match vocalisations to videos of vocalising conspecifics according to the call type produced, even when the utterances were not temporally synchronised with the videos. The authors concluded that the chimpanzees had associated the calls to the correct signaller based on the cross-modal semantic congruency of information relating to the same call type. However, because distinct patterns of facial motion are uniquely associated with different call types in primates (Hauser *et al.*, 1993; Partan, 2002), the auditory and visual features systematically co-vary. Therefore, it may be that the chimpanzees merely learnt to associate the visual and auditory cues related to a particular call type through prior exposure to the systematic co-occurrence of these cues, without perceiving their 'semantic' unification. This study illustrates the fact that it is difficult to determine whether animals are capable of forming categorical representations about communicative stimuli using the preferential looking paradigm, because the subject animal is presented with information from both sensory modalities at the same time. The simultaneous availability of both signals could allow the individual to simply associate the related information together based on the statistical correspondence of these cues, without necessarily activating any form of

cognitive representation incorporating the different sensory information (Adachi and Fujita, 2007).

Therefore, whilst studies using the preferential looking paradigm have established that primates do combine different sensory information related to the same call type, they have not been able to fully explain how they do so. To further investigate whether chimpanzees were able to form multisensory categorical representations of different call types, Parr (2004) used a matching-to-sample paradigm that included a time delay between the presentations of the different sensory stimuli, preventing the subjects from merely associating the stimuli that usually co-occurred. The chimpanzees were first shown a video of a vocalising conspecific that had been edited so that it contained only the audio or visual content. This was followed by a blank screen, after which two photographs were presented showing a conspecific producing either the same call type or a different call type, from a different angle to the video. The results showed that the chimpanzees were able to successfully select the photograph that corresponded to the video in both the intra-modal (visual to visual) and cross-modal (auditory to visual) trials. Interestingly, when videos including incongruent audio and visual information (i.e., the audio was changed to a different call type) were presented, the chimpanzees' preferences for matching the audio or visual information to the photographs depended on the type of expression. For example, photographs of play faces tended to be preferentially matched using the auditory modality of the video (laughter), which Parr suggested may be because these call types are usually produced during playful wrestling, when facial expressions are obscured.

Although the subjects were still given a choice of two images to match to the video, the time delay between the video and photograph presentation suggests that the chimpanzees may have activated a cognitive representation of the appropriate expression that incorporated both visual and auditory information. It is therefore possible that the chimpanzees accessed stored knowledge related to specific call types and expected to see the facial expression that was associated with a particular vocalisation. The consistent differences in performance depending on the production context of the call type also suggests that this 'unity effect' may be moderated by the learnt statistical regularity of co-occurring cues, rather than associating the stimuli on the basis of innately equivalent neurological responses. The ability to form multisensory

representations of particular call types is therefore likely to be dependent on consistent, categorical differences between each type of call that primates produce during communication. However, in comparison to other mammals, primates have a greater diversity of facial and vocal expressions (Andrew, 1962). This means that whilst some primates appear to be able to form categorical representations of different call types, non-primate species that have less variability in their facial expressions may be unable to associate call types with facial expressions in this way because of the lack of available visual cues to form correspondences with. The evolutionary origin of this ability may be dependent on the diversity of species-specific facial expressions, which could be determined by investigating whether bimodal categorisation of call types also occurs in non-primate mammalian species.

As well as possessing multisensory representations of the dynamically encoded differences between call types, non-human primates also appear to learn multisensory categories about the static attributes of signallers. These categories can represent a single attribute shared by multiple signallers, as suggested by Adachi *et al.*'s (2006) demonstration that infant Japanese macaques have a multisensory cognitive representation of their own species. Using an expectancy violation paradigm, the subjects were first presented with either a human or conspecific vocalisation, followed by a photograph of an unfamiliar individual's face from either the matching or non-matching species. The subjects looked longer at the photograph of the human face when it was preceded by a conspecific vocalisation, suggesting that they were surprised to see an image of a human and may have instead expected to see a conspecific. This indicated that the conspecific vocalisation had activated a mental representation of the macaques' own species, which included stored corresponding visual information. However, the time spent looking at the conspecific images was the same irrespective of the preceding voice, whilst the time spent looking at the photograph of the human face was equivalent to the conspecific face when it was preceded by a conspecific vocalisation. Therefore it is possible that the macaques only paid attention to conspecific stimuli, which may have then transferred to the subsequently presented human photograph in the non-matching trial.

Because the attentional bias shown toward the conspecific stimuli could have been related to the subjects' lack of prior exposure to humans, the study was replicated using

infant Japanese macaques that had extensive prior experience with humans (Adachi *et al.,* 2009). These macaques looked at the photographs for longer when they were mismatched, irrespective of species, suggesting that they did have multisensory categorical expectations about their own species and the human species. Whilst it therefore appears that Japanese macaques have the capacity to form a cross-modal representation of species, the dependence of the responses of the infant macaques on their previous experience with humans provides further support for the theory that specific cross-modal categorical representations may be learnt and related to the individual's own experiences. This illustrates that the functional relevance of specific representations for individual animals (both within and across species) must always be considered, as this can influence the formation or expression of specific associations.

In addition to forming species-level multisensory representations, animals also appear to associate different sensory signals by perceiving the congruency of sex-related cues. Species such as humans and baboons have a sexually dimorphic vocal apparatus, which results in anatomically-constrained differences in the mean F0 and formant structure between adult males and females (Rendall *et al.,* 2005). Sex-related differences in the acoustic structure of adult human voices enable human listeners to classify adult voices as male or female (e.g., Bachorowski and Owren, 1999). Four-month old human infants expressed the ability to associate unfamiliar voices with corresponding faces according to their gender, by attending more strongly to the congruent image in a preferential looking paradigm (Walker-Andrews *et al.,* 1991). Whilst the ability to match conspecific multisensory signals according to sex has yet to be investigated in other species, dogs have also been shown to associate unfamiliar human voices with a person of the corresponding gender when presented with an unfamiliar woman and man (Ratcliffe *et al.,* 2014). Further investigations are now required to establish whether this ability is learnt via exposure to humans during development, or innately present across dogs as either a shared mammalian mechanism or as a result of their domestication.

Animals thus appear to be capable of forming a variety of multisensory categories about broadly shared signaller attributes, which can be used to associate signals with unfamiliar individuals. Furthermore, the cognitive mechanisms that underlie the categorisation of call types, species and sex also appear to be flexible enough to allow more specific multisensory representations to develop about familiar conspecifics. In

fact, a wide range of phylogenetically distant mammalian species has been shown to form multisensory categorical representations of familiar individual signallers. Using the expectancy violation paradigm, Proops *et al.,* (2009) demonstrated that domestic horses *Equus caballus* form multisensory representations of other individuals in their social group. Subjects first watched as a familiar herd member was led past them and then out of sight, after which a vocalisation produced by either the individual the subject had just seen or a different herd member was played from a loudspeaker. The subject horses looked significantly faster and longer in the direction of the speaker when the vocalisation did not match the individual they had just seen, indicating that they had formed multisensory cognitive representations of individual members of their social group. Similar representations of familiar conspecific individuals have also been reported in rhesus macaques (Adachi and Hampton, 2011); grey-cheeked mangabeys *Lophocebus albigena* (Bovet and Deputte, 2009); chimpanzees (Kojima *et al.,* 2003; Martinez and Matsuzawa, 2009) and even large-billed crows *Corvus macrorhynchos* (Kondo *et al.,* 2012). Whilst these studies have focused on the association of visual and auditory cues, other sensory cues are also usually available, and it has recently been shown that ring-tailed lemurs *Lemur catta* are able to recognise conspecific individuals by associating olfactory and auditory signals (Kulahci *et al.,* 2014). The lemurs' ability to associate scent and vocalisations is especially interesting because these cues are rarely encountered at the same time; therefore lemurs have limited opportunity to learn to associate these cues through temporal or spatial correspondences. Kulahci *et al.* suggested that modality dependent identity information may be learnt separately, and independently linked to generate a multisensory representation of the individual. This observation provided the first evidence that individual identity representations in animals are not necessarily learnt through prior exposure to co-occurring cues.

Many species are therefore able to associate information related to individual conspecifics. Furthermore, there is evidence to suggest that representations of individuals are even flexible enough to extend to familiar heterospecifics. Indeed, squirrel monkeys *Simia sciureus* can form a multisensory representation of their primary human caretaker (Adachi and Fujita, 2007). Similarly, dogs appear to activate a mental representation of their owner's face when they hear their owner's voice (Adachi *et al.,* 2007), further illustrating that the ability to learn functionally relevant multisensory categorical representations can occur between distantly related mammalian

species. Individual identity representations can also be sufficiently detailed to enable animals to distinguish between different, equally familiar heterospecific individuals, as both rhesus macaques and domestic horses can match the vocalisations of different familiar individuals with their visual appearance, either based on a photograph (of either familiar conspecifics or human caretakers; rhesus macaque: Sliwa *et al.,* 2011) or in person (using human handlers; horses: Proops and McComb, 2012). However, because studies investigating animal recognition of individual human voices have used phrases that were highly familiar to the subject, such as the animal's name, it remains possible that the animals associated differences in the pronunciation of those particular phrases with specific human individuals and their recognition may not generalise to unfamiliar utterances (Kriengwatana *et al.,* 2014). Further experiments should therefore use unfamiliar phrases to clarify the whether these animals have the ability to recognise the voices of familiar humans independently of the verbal content of the speech utterance.

Although further confirmation remains necessary, observations of cross-modal heterospecific recognition suggest that multisensory identity representations might be widely present across mammals and highly flexible in their formation. Alternatively, it is possible that both primates and domesticated mammals may have different innate predispositions that facilitate the categorisation of individual humans, which are not necessarily present in other species. Similarities between identity cues in more closely related species might allow non-human primates to generalise the same associations used to form conspecific identity categories to familiar humans, and this could similarly apply across other closely related species. This form of generalisation may not be possible in the case of phylogenetically distant domesticated animals such as dogs and horses, where the recognition of individual conspecifics is more likely to involve different identity cues to those used to recognise individual humans. However, both species might have adapted to be able to form representations of familiar humans during the process of domestication. To test these hypotheses, heterospecific identity representations of familiar humans could be investigated in non-domesticated, phylogenetically distant species. Inter-specific identity representations could also be tested between two distantly related non-human species, such as if horses recognise familiar dogs.

Although evidence of multisensory categorical representations in mammals is currently limited to species that live in relatively complex social groups, it is clear that a range of distantly related animals are capable of forming detailed categories about various static attributes of signallers, and in non-human primates multisensory representations have also been observed to extend to dynamic expressions. The ability to form complex cognitive representations indicates that the evolutionary precursors of concept formation in humans may be present in these species (Barsalou, 2005). Indeed, in a neuro-imaging study in which rhesus macaques heard conspecific vocalisations and unfamiliar non-biological sounds, Gil-da-Costa *et al.,* (2004) demonstrated that the vocalisations generated activation in a distributed neural circuit including higher-order visual cortical areas associated with the visual perception of object form and motion. The amygdala and hippocampus (areas associated with emotional processing) also selectively responded to affectively salient scream vocalisations. This pattern of activation showed striking parallels with the neural circuits underlying conceptual representations in humans, leading Gil-da-Costa *et al., (*2004) to suggest that this system may have played an important role in the evolution of human concept formation.

However, whilst the current research suggests that natural categorical representations in mammals may be learnt, our limited knowledge of the relative importance of specific signal components in different cognitive representations means that at present it is difficult to establish how these representations are acquired in non-human mammals. It remains possible that the perception of low-level structural and statistical correspondences other than temporal synchrony, such as size or shape, may contribute to the formation and application of some of the multisensory categorical representations involved in mammalian communication.

## Conclusion

In this review, we have attempted to address how animals might solve the ‘cross-modal binding problem’ in order to combine the individual sensory components of multisensory signals used in their communication systems. Although the range of mammalian species in which mechanisms influencing multisensory perception have been investigated is predominantly limited to the primate order, it is apparent that more distantly related species naturally associate information across sensory modalities in

order to perceive the functional content encoded within their signals. Similarly to humans, many other mammals show a tendency to associate sensory signals that co-occur in time or space, which can be beneficial when they lack prior experience with the signals. However, other basic features can also facilitate signal combination when they are perceived as equivalent, either because they represent the same feature or if they are estimated using the same underlying neural mechanisms. Whilst it remains to be seen if animals can learn to apply prior knowledge of whether basic features usually co-occur in the environment to mediate associations, observations that a wide range of mammals learn to combine different sensory cues about other individuals, and that this system is flexible enough to support representations of other species, suggests that they may also learn to use more general statistical correlations in the environment.

Together, the observations that many species share some of the cross-modal correspondences observed in humans implies that they are likely to possess perceptual and cognitive mechanisms that parallel some of the processes present in humans. Although it is not known whether such processes have arisen through convergent evolution or whether they are present in other animals due to their shared evolutionary history, it is perhaps unsurprising that non-human primates in particular have been observed to display more homologous correspondences, demonstrating their perception of temporal synchronisation between conspecific vocalisations and facial movements, as well as the ability to form detailed cognitive representations of individuals and expressions. Because complex categorical representations have only been investigated in highly social mammalian species, it has not yet been determined whether the ability to form such representations is a specific adaptation to greater sociality, or if solitary species similarly have a capacity to form complex categorical representations if they are functionally relevant for the individual. Further research will also be necessary to investigate potential interactions between different perceptual and cognitive mechanisms in the formation of cross-modal associations in mammals, particularly the relative importance of more general statistical correspondences. Determining how, and to what extent, different associations are acquired across a wider range of mammalian species will be an essential step in developing our understanding of the evolutionary origins and function of cross-modal correspondences in multi-sensory perception.

**Future Directions: Investigating Cross-Modal Correspondences in Domestic Dogs**

From the outcomes of our review it is clear that cross-modal correspondences are largely under-researched in non-human animals, and that comparative studies are vital to develop a clearer understanding of the function and evolution of these perceptual mechanisms in humans. Although comparative studies have traditionally explored multisensory perception in primates, future research would be likely to benefit from examining the perception of cross-modal correspondences in domestic dogs. Dogs are a particularly well suited model species for such investigations primarily because they live in anthropogenic environments, and are therefore naturally exposed to many of the same multisensory correspondences as humans. Dogs have already been shown to form functionally comparable multisensory categorical representations to humans in inter-specific communicative contexts, by naturally associating speech cues to identity (Adachi *et al.,* 2007) and gender (Ratcliffe *et al.,* 2014) with the corresponding human speaker. Due to the strong relevance of human signals for dogs, it is likely that they are similarly able to perceive further correspondences between human vocal and visual cues, such as the age of the speaker or their emotional expression. If dogs are able to associate human voices with the appropriate signaller using either age or emotionally related cues, it would provide the first demonstration of the cross-modal perception of these functionally relevant features outside of the primate order. Further work is also necessary to establish the underlying mechanisms enabling the expression of these abilities, particularly whether apparently complex associations are based on the perception of low-level or high-level correspondences, and the extent to which they are dependent on experience.

As well as demonstrating the ability to associate human voices with people by perceiving correspondences between speaker-related characteristics, dogs' strong responsiveness to human speech could also facilitate investigations into the mechanisms underlying object-label correspondences, such as the 'bouba-kiki' effect (Köhler, 1929). Demonstrations that pre-linguistic human infants map novel words to their likely referents using sound symbolism as an inferential cue (e.g. Peña *et al.,* 2011) has led to the suggestion that sound symbolism could be the vestige of a protolanguage, and may have played a key role in the evolution of language (Imai *et al.,* 2015). However, the existence of such a pre-adaptation for human language evolution has not yet been

investigated in non-human animals. Researchers could take advantage of the fact that dogs can readily learn a variety of verbal object labels (e.g. Kaminski *et al.,* 2004), by testing whether the presence of sound symbolism cues similarly scaffolds their acquisition of new labels. Determining the extent to which the perception of sound symbolism is shared across mammals would provide important insights into the evolution of human language.

Dogs' established ability to differentiate between human spoken commands as well as speaker-related cues could also be fruitfully employed in studies investigating the interaction between different cross-modal correspondences during perception. To our knowledge, it is currently unknown whether similarly to humans, animals' perception of auditory vocalisations can be manipulated by presenting additional correspondences, such as mismatched articulatory cues (the McGurk effect; McGurk and MacDonald, 1976). It also remains unclear whether animals undergo a comparable process of perceptual narrowing to humans, where higher-order correspondences are prioritised with increased experience. Because dogs are one of the few species where their natural ontogenetic development can be closely monitored without disrupting the process, it would be possible to document the emergence of higher-order cross-modal associations, such as the multisensory perception of human gender cues. As well as discovering whether perceptual narrowing occurs in other mammals, such studies would provide important insights into the mechanisms underlying complex representations in animals by establishing which cues are associated and how this occurs.

In addition to the cross-modal correspondences involved in communication, many pet dogs are exposed to the same environmental correspondences as humans through their co-habitation with us. Therefore dogs could be a key comparative species for studying the perception and use of statistical regularities in the environment. For example, as discussed earlier in the review, Parise *et al.,* (2014) determined that the human tendency to map sounds to different spatial elevations based on their frequency was dependent on an interaction between both environmental experience and frequency-elevation filter effects caused by the shape of the human head and outer-ear. While dogs gain the same environmental experiences as humans, their pinnae shape differs markedly from humans, and also between individual breeds. Therefore it would be interesting to test whether dogs also express statistical correspondences, and whether the correlation

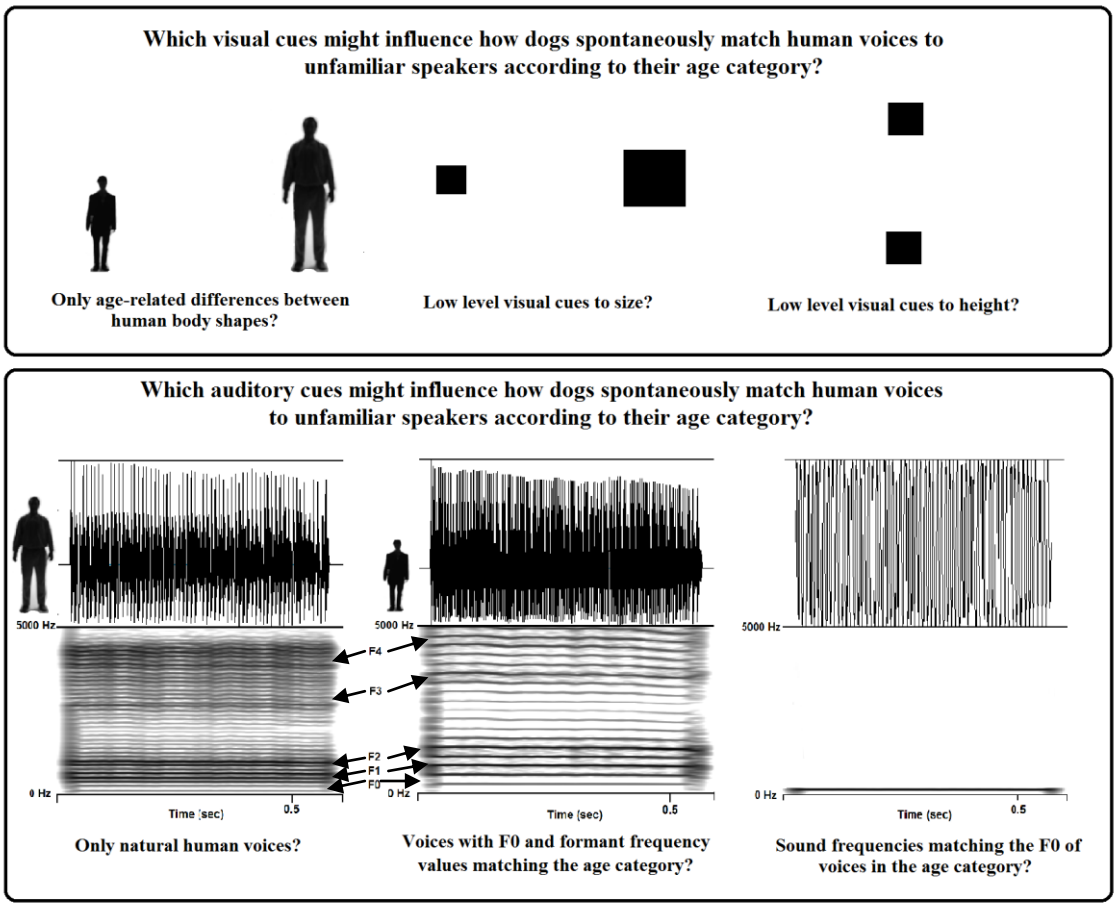
between values differs according to their ear structure. Furthermore, there has been some debate as to whether some of the observed basic feature correspondences, such as the association between auditory pitch and visual luminance, are learnt due to their statistical correspondence in the environment, or if they are structurally dependent, occurring because both features are equivalently processed in the brain. If basic feature correspondences are observed in domestic dogs, comparisons with their closest ancestor, the wolf, would help to answer whether such correspondences are environmentally-related or structurally dependent.

Therefore, investigating the perception of cross-modal correspondences by dogs could provide key insights into the evolution, mechanisms and development of multisensory perception in mammals. Dogs' co-habitation with humans means that comparable associations are not only functionally relevant for this species, but that they can be readily tested using ecologically valid scenarios. Article II of this Chapter presents an experimental study aiming to explore the potential mechanisms involved in dogs' multisensory perception of human signals.

## Article II: Cross-modal Discrimination of Human Age Categories by Domestic Dogs *Canis familiaris:* an Exploration of the Mechanisms Involved in Multisensory Perception

### Synopsis

**Question:** Are dogs spontaneously capable of the cross-modal discrimination of human age categories, and if so, how do they associate age-related auditory and visual cues?




**Results and conclusions:** Dogs visually matched the adult human body shape to an adult male voice, but the association between visual stimuli and adult male voices did not generalise to simple shapes with the congruent low level visual cues of a relatively large size or high elevation. Dogs also did not match either of the human body shapes to re-synthesised 'child' voices, suggesting that while the F0 and formants may be used to categorise adult speakers, dogs had not learnt to use these cues to categorise child speakers. However, 'child' voices were matched to lower positioned shapes, indicating that human voices with the F0 and formant values of a child were associated with lower elevations. Finally, unlike humans, dogs showed no evidence of mapping low level visual cues to tones matching the voice pitch frequencies.

*Note.* This article is formatted in the style of *Animal Behaviour*.

**Abstract**

Domestic dogs discriminate the familiarity and gender of humans across different sensory modalities, enabling them to match adult human voices to speakers on the basis of these attributes. To determine if the ability to cross-modally categorise humans using indexical cues also extends to the person's age, we investigate whether dogs are capable of spontaneously matching human silhouettes with human voices by associating visual and acoustic cues to their categorical age group (adult versus child). Furthermore, in order to explore the potential mechanisms behind this ability, we also investigate if dogs' multisensory perception of human age is achieved through more general associations between low-level features within either the visual domain (size and height) or the auditory domain (voice pitch). In a preferential-looking paradigm with three trials, we simultaneously presented dogs with human voices and different image pairs: the silhouettes of a man and a child, a large and a small square positioned at the same elevation, or two same-sized squares positioned at a high and a low elevation. Dogs successfully matched the corresponding adult body shape to the adult male voices, as significantly more of the subjects looked faster and longer towards the silhouette of the adult male rather than the child when presented with a man's voice. In contrast, they did not match any of the square sizes or positions with the men's voices, indicating an absence of perceptual association between adult male voices and low-level visual features of size or height. However, when the main acoustic cues that human listeners use to differentiate adult and child voices (the pitch and vocal tract resonances) were manipulated to re-synthesise the voices to sound like a 6-year-old boy, dogs failed match the child body shape to these voices, as they looked equally towards both silhouettes. While no associations were made between the differently-sized squares and the re-synthesised 'child' voices either, the majority of dogs presented with equally-sized squares positioned at different elevations looked faster and longer at the low square when they heard a 'child' voice, indicating that they perceived an association between a low elevation and vocal resonance and/or pitch values characterising the voice of a human child. Finally, when dogs were presented pure tones matching the voice pitches they did not associate either of the silhouettes, nor any of the square sizes or elevations, with these tones, indicating that none of the cross-modal associations observed in response to the human voices were related to the perception of low-level

auditory correspondences between the voice pitch and any of the visual cues. Together the results suggest an experience-dependent process where dogs may initially associate human voices from different age categories with average projection heights before learning to discriminate between specific age-related human body shapes. Further research is now necessary to confirm the development of human age perception in this species by testing dogs with more controlled and varied levels of exposure to people from different age categories.

## Introduction

Recent evidence has demonstrated that domestic dogs, like human listeners, spontaneously discriminate both the identity (Adachi, Kuwahata & Fujita, 2007) and gender of human adults across different sensory modalities (Ratcliffe, McComb & Reby, 2014), enabling them to match speech signals to the corresponding human speaker according to these key indexical characteristics. However, although discriminating between human age categories is likely to be ecologically relevant for dogs in order to modify their behaviour towards different people (Reisner & Shofer, 2008), their ability to similarly combine multisensory cues related to the person's age category has not yet been investigated, and more generally, the perceptual mechanisms underlying dogs' categorisation of human voices remain unclear. Human listeners predominantly categorise unfamiliar voices according to the speaker's age and gender by attending to the pitch (fundamental frequency; F0) and vocal tract resonances (formant frequencies) (Smith & Patterson, 2005). The strong reliance on these two acoustic parameters arises because they directly reflect anatomical structures which are sexually dimorphic in humans (Titze, 1994). Age and sex related differences in the size of the larynx culminate to produce the lowest average F0 in adult men at around 100 Hz, which is doubled in adult women at approximately 200 Hz, and is even higher in children at around 260 Hz (Lieberman, 1988; Huber, Stathopoulos, Curione, Ash & Johnson, 1999). Additionally, increasing vocal tract length causes the formant frequencies to decrease by around 32% from age four to adulthood in men, and by around 20% in women (e.g. Huber et al., 1999). Corresponding growth and pubertal changes in body shape and size enable human listeners to readily match unfamiliar voices to speakers according to their age group and gender.

Although the acoustic parameters that dogs use to perceive the gender of adult human voices have not yet been determined (Ratcliffe et al., 2014), dogs can be successfully trained to discriminate between synthesised vowel sounds when the F0 and lower formant frequencies match the average values of either an adult man or woman (Baru, 1975). This indicates that dogs also have the perceptual capacity to differentiate between human voice-age categories if the differences in the F0 and formant frequencies are at least as large as those between the voices of adult men and women. Similarly, dogs' ability to match adult male and female voices to unfamiliar people of the corresponding gender (Ratcliffe et al., 2014) suggests that they are also likely to perceive related visual cues to the person's age. The most salient visual difference between human adults and children is their body size (height and mass). Dogs are already known to cross-modally assess the body size of conspecifics, as they naturally use the formant frequency values of conspecific vocal signals to match growls with corresponding signallers (Taylor, Reby & McComb, 2011), and the ability to use these cues to discriminate between different age categories has been identified in mammals, as rhesus macaques *Macaca mulatta* spontaneously attend to the formant frequency spacing in species-specific vocalisations in order to associate these signals with the faces of unknown adult or juvenile conspecifics (Ghazanfar et al., 2007). However, it remains undetermined whether the ability to assess size cues in vocalisations is limited to conspecific signals or if it also extends to hetero-specific signals. With sufficient exposure, humans generalise the correlation that exists between the formant spacing and body size encoded in human voices (Pisanski et al., 2014) to the vocalisations of other mammals, accurately judging the size of dogs from the formant positioning of their growls (Taylor, Reby & McComb, 2008). This raises the possibility that animals may similarly discriminate conservative size cues in hetero-specific vocalisations, potentially enabling them to match vocalisations to human signallers according to age-related differences in their body size.

Although dogs appear to rely on variation in the formant frequencies of species-specific vocalisations to assess body size, humans attend to both the formant frequencies and F0 of people's voices in order to categorise their age (Smith & Patterson, 2005). Indeed, there is a larger average difference between the F0s of child and adult voices (around 75%) than between their formant frequency values (around 26%) (Huber et al., 1999), enabling listeners to take advantage of the greater natural variation in the F0 to

discriminate between differently aged voices. While human listeners may simply learn that specific F0 and formant frequency values relate to different age categories, it has also been suggested that the perception of basic low-level correspondences between the voice pitch and rudimentary visual dimensions could influence the way that vocal and visual signals are associated (Rendall, Vokey & Nemeth, 2007). One potentially relevant correspondence is the tendency that humans show to automatically map lower frequency tones to larger simple shapes and higher frequency tones to smaller shapes (Parise & Spence, 2009; Evans & Treisman, 2010), which is likely to arise as individuals perceive generic frequency-size correspondences present in the environment (Spence, 2011). Morton's (1977) observation that across many species of mammals and birds, low pitch vocalisations tend to be used in agonistic contexts where projecting a larger impression of body size is likely to be advantageous, whilst high pitch vocalisations usually characterise appeasement or affiliative contexts (where animals can benefit from signalling a smaller, less threatening size) suggests that animals may similarly perceive broad correspondences between the frequency structure of vocalisations and the size of the signaller, which could be beneficial in categorising the signaller as an adult or juvenile. In addition to perceiving low-level correspondences between auditory frequencies and visual size, humans also appear to map lower frequency tones to lower spatial elevations and higher frequency tones to higher spatial elevations, which has been linked to both the regularity of the co-occurrence of these dimensions in the environment and their emphasised co-perception due to the shape of the human ear (Parise, Knorre & Ernst, 2014). While the statistical correspondence between sound frequency and spatial elevation has also been shown to influence vocal size judgements in humans, as listeners actually judge voices as belonging to larger people when they are projected from lower spatial positions (Pisanski, 2014), the perception of statistical frequency-elevation correspondences by non-human animals has not yet been explored.

In the present study, we investigate whether dogs spontaneously associate human voices to human body shapes by matching acoustic and visual cues to the speaker's categorical age (adult or child). In addition, we aim to explore whether dogs' multisensory perception of human age groups can be explained by more general associations made between low-level sensory cues in the visual domain (either size or height) and/or in the auditory domain (the pitch of the voice). Using a preferential looking paradigm, in

Experiment 1 we first test if dogs are able to match natural human voices of adult men with a corresponding adult body shape, by simultaneously presenting the subjects with the silhouette of an average sized adult man and the silhouette of a 6-year-old boy while an unfamiliar adult male voice is played. Given that dogs are able to perceive pictures in an ecologically valid manner (Adachi et al., 2007; Kaminski, Tempelmann, Call & Tomasello, 2009; Faragó et al., 2010; Somppi, Törnqvist, Hänninen, Krause & Vainio, 2012) we use two-dimensional silhouettes as visual models to exclude any additional cues to the age and sex of the person other than their physical body size and shape. Because dogs are capable of cross-modal human gender discrimination, which for human listeners also involves assessing the relative positions of the F0 and formants (Bachorowski & Owren, 1999), we predict that dogs will successfully match the adult male voices with the corresponding silhouette, by looking at the adult silhouette faster and for longer than at the child silhouette (Ghazanfar et al., 2007; Taylor et al., 2011). In order to determine whether dogs' cross-modal human age assessments can be accounted for by their perception of more general low-level visual cues than specific human body shapes, we also test whether the subjects associate simple square shapes with the adult male voices according to their visual size or elevation. We predict that when dogs are simultaneously presented with two differently sized squares positioned at the same elevation, they will associate the larger square with the adult male voice. Similarly, when presented with two equally-sized squares in different spatial elevations, we predict that dogs will associate the higher positioned square with the men's voices.

As well as exploring the role of low-level visual cues, we investigate the relative contribution of two key vocal cues differentiating adult and child voices, F0 and formant frequencies, in dogs' multisensory perception of human age. To do this we present dogs with re-synthesised adult male voices where F0 and formant values are typical of a 6-year-old boy, while all of the other acoustic parameters remain unchanged across the two acoustic conditions. Similar re-synthesis methods have successfully altered the perceived age of human voices in previous studies (Smith & Patterson, 2005) and our re-synthesised voices were of equally high quality. Given the previous demonstrations that dogs can learn to discriminate between the synthetic voices of adult men and women based on the formant frequencies and F0 alone (Baru, 1975) as well as accurately judge the size of conspecific signallers after the formant frequencies of the vocalisations have been re-synthesised (Taylor et al., 2011), we predict that the subjects

will also assess the apparent age of human voices using these key anatomically-related age cues. As the F0 and formant frequency values of the re-synthesised voice stimuli match those of a young child, we expect dogs to match the child silhouette, as well as the smaller sized square and the lower positioned square, to these 'child' voices.

In Experiment 2 we further degrade the range of acoustic cues available to the subjects to test whether the perception of general cross-modal correspondences involving low-level features in the acoustic domain is involved in dogs' assessments of human age. Using the same three pairs of visual stimuli as the first experiment, we measure the gaze responses of a new subject group when they are presented with pure tones matching the F0s of the original and re-synthesised vocal stimuli. If dogs predominantly categorise human age groups by associating the more pronounced F0 differences with corresponding age-related visual cues, we expect that the subjects will express the same associations as those presented with the full human voices in Experiment 1. Specifically, we hypothesise that dogs presented with the lower frequency pure tones (matching the F0s of the adult male voices) will look more quickly and for longer towards the adult male silhouette, and similarly associate the larger square and higher positioned square shape with these low frequency tones. In contrast, we predict that dogs presented with the high frequency pure tones (matching the F0s of the re-synthesised 'child' voices) will look more at the child silhouette, as well as the smaller sized square. Although humans have been observed to associate high frequency tones with higher visual elevations (e.g. Parise et al., 2014), the fact that this low-level correspondence remains untested in non-human animals led us to predict that dogs would instead match the lower square with the high frequency tone, due to their greater experience with child voices projecting from relatively low elevations.

## Methods

### Subjects

All of the dogs were privately owned pets recruited from the East Sussex area when their owners responded to local advertisements. The owners confirmed that their dog was healthy, with no known visual or hearing problems and no known aggression towards humans. The subjects and their owners were also naïve to the experimental setup and only participated in one of the following experiments. A total of 27 dogs (12 females and 15 males) took part in Experiment 1, including 13 different pure breeds.

Ages ranged between 6 months and 9 years old (mean + SD = 4.12 + 2.67 years). In Experiment 2, a further 26 dogs (11 females and 15 males) were tested, including 16 different pure breeds. Subjects were aged between 6 months and 14 years old (mean + SD = 4.81 + 3.27 years).

**Auditory Stimuli**

Four men, aged between 21 and 46 (mean + SD = 31.00 + 11.52 years), were recorded in a sound proof room after being instructed to pronounce the word 'hod' with a sustained vowel sound three times. The recordings were made using a Zoom H4N Handy Recorder with a sampling frequency of 44100 Hz and a 32-bit sampling rate. The recordings were then manipulated using PRAAT v.5.0.3 (http://www.fon.hum.uva.nl/praat/). One recording was chosen from each speaker based on the quality and length of the sustained vowel. The recordings were then cut so that only the central 0.5 sec of the vowel sounds were retained. The mean F0 of each recording was then measured using the PRAAT autocorrelation algorithm 'to Pitch (ac)', giving values between 109.3 Hz and 123.9 Hz (mean + SD = 116.5 + 6.1 Hz) which is within the average adult male F0 range (Titze, 2000). The centre frequencies of the first four formants were estimated using PRAAT's Linear Predictive Coding 'Burg' algorithm. The average spacing between the formants was then calculated using the method described in Reby and McComb (2003), returning values between 996.4 and 1040.8 Hz across recordings (mean + SD = 1011.0 + 20.2 Hz), falling within the average range for adult males (Pisanski et al., 2014). Each voice was then re-synthesised by changing the F0 and formant spacing to the average values of a 6-year-old boy pronouncing the same vowel sound (F0 = 273.0 Hz; spacing = 1266.9 Hz; Lee, Potamianos & Narayanan, 1998). Using the 'change gender' command in PRAAT, the formant spacing in each recording was increased by a factor of 1.32, producing formant spacings between 1154.1 and 1336.7 Hz across recordings (mean + SD = 1251.7 + 90.8 Hz), and the F0 was raised to 273 Hz. These manipulations created a total of eight auditory stimuli used for Experiment 1 (Figure 1).

For Experiment 2, pure sine-wave tones were created which precisely matched the F0 of each of the voice stimuli used in Experiment 1. This was carried out in PRAAT using the 'to Pitch' and 'to Sound (sine)' commands (Figure 1). The amplitudes were then standardised to 65dB using Audacity 2.0.0 (http://audacity.sourceforge.net).

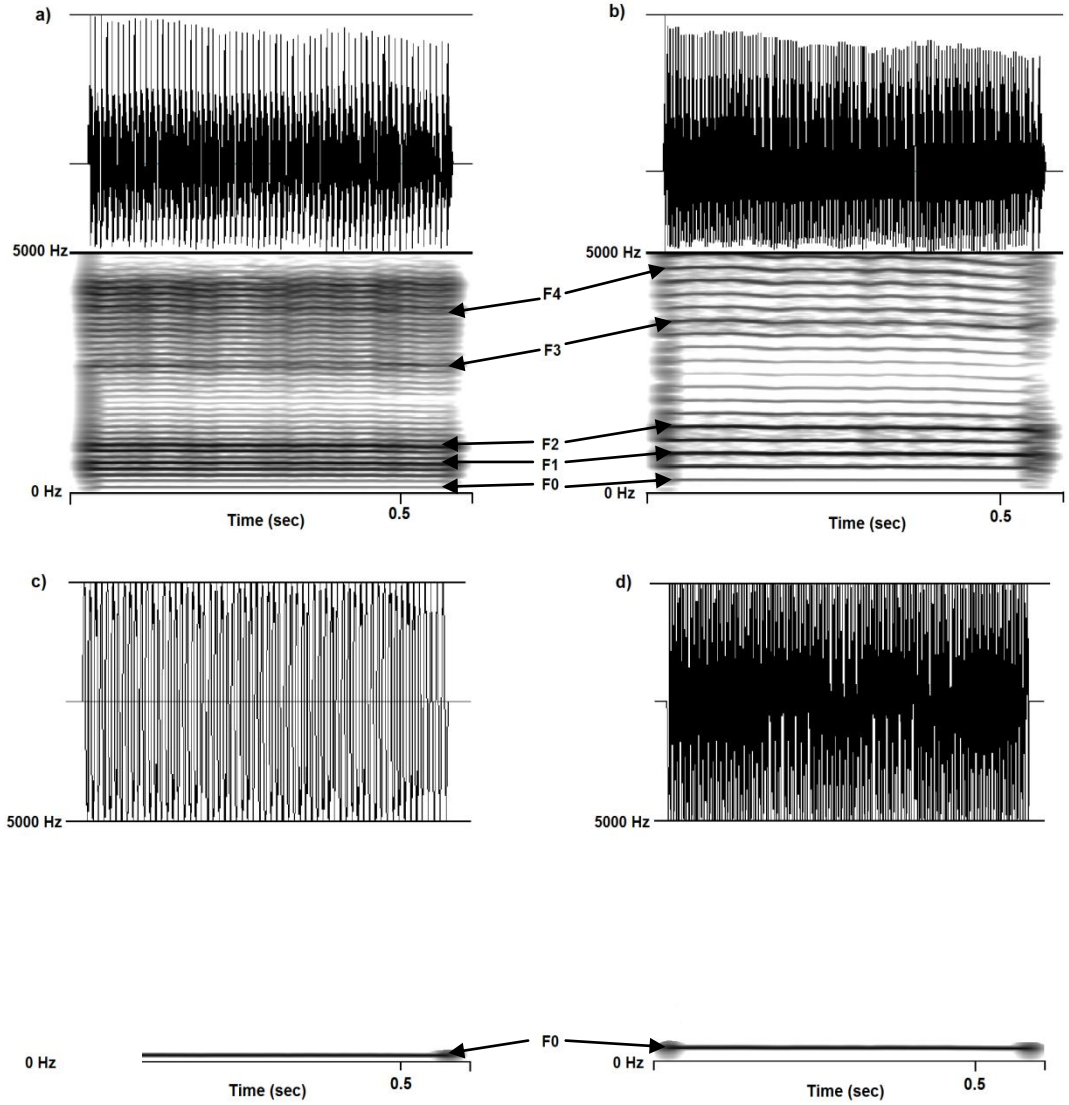



**Fig 1.** Spectrograms detailing the acoustic manipulations from a) the original /o/ vowel spoken by an adult man to b) the re-synthesised 'child' voice in which the F0 and formant frequencies (F1-F4) have been increased to match the average values of a 6-year-old child's voice (Experiment 1). Spectrograms c) and d) represent the pure sine-wave tones created to match the F0s of the original and re-synthesised voices (Experiment 2).

A pure tone (fixation tone) was also created in PRAAT with a frequency of 194.9 Hz (the midpoint between the average F0s of the original and re-synthesised recordings) and duration of 0.5 sec. All of the stimuli were normalised to -1.0dB maximum amplitude in Audacity.

**Visual Stimuli**

Three different pairs of images were created using PowerPoint and projected onto a white wall (240 cm high by 300 cm wide). All of the images were animated to help to attract and retain the dogs' attention. The first pair of images tested the association between voices and people: black silhouettes of an average size adult man and 6-year-old boy (175 and 115 cm tall respectively) were placed 165 cm apart (from the centres of the shapes) with their feet touching the bottom of the screen. Both of the silhouettes gradually increased in size by 5% over 5 sec and then decreased back to their original size over the following 5 sec to give the impression of forward and backward motion (People condition). Both silhouettes increased and decreased in size together to maintain their absolute size difference, and the percentage change in size was minimal relative to the difference in size between the silhouettes. Visual size was tested with the second pair of images using two black squares; one of which was 30 $cm^2$, while the other was 60 $cm^2$. These shapes were positioned at the centre left and centre right of the screen, with 165 cm between the centres of the shapes (Size condition). The final pair of images consisted of two equally sized black squares (40$cm^2$), with one at the top centre and the other at the bottom centre of the screen, 135 cm apart (from the centre of the shapes) providing a test of visual elevation (Elevation condition). All of the squares slowly rotated clockwise, completing one full rotation every 5 sec. There was no change in the size or elevation of the squares during these animations. A black oblong (20 cm x 35 cm) fixation stimulus was also created which appeared in the centre of the screen and rotated 720$^o$ clockwise every 5 sec. All of the images were surrounded by a plain white background (Figure 2).

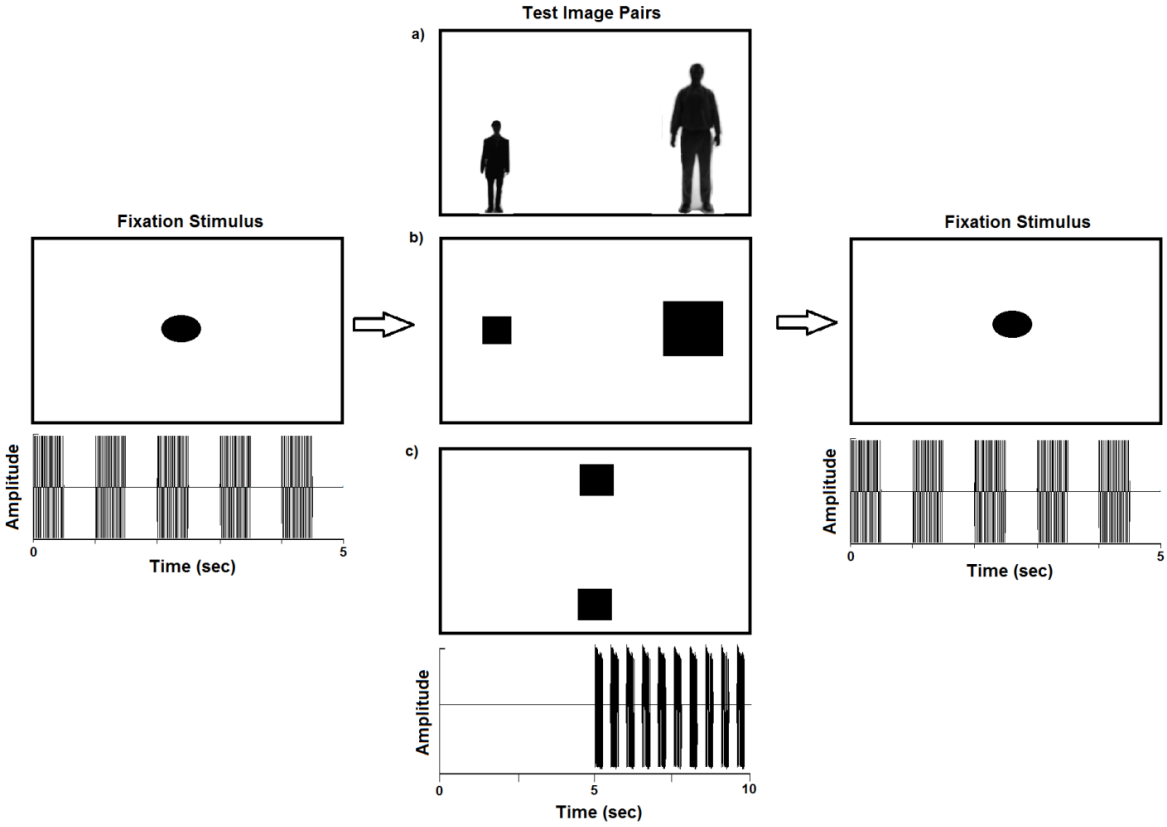



**Fig 2.** Visual stimuli sequence beginning with the fixation screen and accompanying fixation tone, to one of the three test image pairs with an auditory test stimulus, followed by the fixation screen again before the next test image pair. Test image pairs: a) People condition; b) Size condition; c) Elevation condition.

## Experimental Set-up

Testing was carried out between November 2013 and March 2014 in a designated room at the University of Sussex, U.K. Two Sony SRS-A60 loud speakers were mounted 150 cm high at the far left and right sides of the wall, so that they were directly in line with the height of the fixation point which appeared in the centre of the screen. A 15 cm high white board was placed along the bottom of the screen to conceal both the wire connecting the two speakers and the SONY DCR-HC51 Handycam video camera, which was mounted on a small tripod and placed on the floor at the centre of the wall. Equally spaced holes were made in three sections of the board, and the video camera was positioned behind the centre hole. The video camera was set to night vision mode as the lights were turned off during testing and a black out screen covered the window.

Florescent coloured rope was used to delineate the centre line from the video camera to the opposite wall. The subject was positioned 300 cm away from the projection wall, and lined up along the centre line. A chair was placed directly behind the subject's position for the dog's owner to sit on. The visual stimuli were projected onto a wall covered with white projector screen paint so that the screen filled the entire wall.

**Procedure**

Subjects remained on a loose lead and were placed in front of their owner, who was seated on the chair provided. The owners were instructed not to speak to their dog during the test and to only touch them if necessary to keep them situated in front of the chair. The lights were then turned off and the test sequence was started. A blank white screen appeared for 20 sec whilst the experimenter left the room. After this the visual fixation stimulus appeared in the centre of the white screen. At the same time the fixation tone played 10 times with 0.5 sec of silence between each repetition. The visual fixation stimulus then disappeared and the first pair of test images appeared. After 10 secs of silence, one of the auditory stimuli (voice or sine-wave tone for Experiment 1 and 2 respectively) was repeated 10 times with 0.5 sec of silence between each repetition. The pair of images then disappeared and the fixation stimulus reappeared and rotated again with the accompanying pure tone for 5 sec. The second and third pairs of images then followed in the same way. When the final pair of images disappeared they were replaced by a blank white screen. Each pair of images was accompanied by a different auditory stimulus. All of the sounds were played at 65dB, measured by an N05CC Digital Mini Sound Level Meter. The congruent position was counterbalanced so that the same side was not congruent in consecutive image pairs, and the order of image pairs, congruent positions, congruent images, apparent voice ages and auditory stimulus exemplars were each counterbalanced across subjects.

**Ethical Note**

The study complied with the internal University of Sussex regulations on the use of animals and was approved by the University of Sussex Ethical Review Committee (Approval number: ERC/33/3). Approval was also obtained to record human voices to use as stimuli from the University of Sussex Life Sciences & Psychology Cluster based Research Ethics Committee (Approval number: DRVR0312).

**Behavioural Coding**

The digital video analysis software Sportscode Gamebreaker version 7.5.5 (Sportstec, Warriewood, NSW, Australia) was used to code the videos. The videos were analysed in 100 ms intervals for the 10 sec of voice presentation, during which the direction, latency and duration of each look towards each stimulus image was recorded in milliseconds. For the People and Size trials where the images were placed side by side, a look was defined as being at either of the images if the dog's head was angled between $10^{o}$ and $20^{o}$ from the centre point. Because in the Elevation trials the angle of the dogs' head when viewing the shapes depended on their body size, a look was defined as being at the lower shape if their head was angled straight down towards the camera, and at the higher shape if their head was angled straight upwards, as indicated by the position of their nose relative to the rest of the face. The videos were coded by V.R. From these visual analyses we obtained two response variables: the total gaze duration (total time looking) and response latency (time to first look) towards each image. A research assistant naïve to the purpose of the study then second-coded 42% of the videos for Experiment 1, which resulted in a strong inter-observer correlation for both the response latencies (Spearman: $r = 0.86$, $N = 66$, $P < 0.001$) and total gaze durations (Spearman: $r = 0.84$, $N = 66$, $P < 0.001$). The same assistant also second coded 54% of the videos from Experiment 2, where there was also a strong inter-observer correlation for both the response latency (Spearman: $r = 0.99$, $N = 86$, $P < 0.001$) and total gaze duration (Spearman: $r = 0.98$, $N = 86$, $P < 0.001$).

To determine the dogs' ability to associate the voices with the matching images, we calculated two binary 'congruency scores' of congruent or incongruent for both response variables. The response latency was coded as congruent if the dog looked towards the matching image before the non-matching image (First look congruent score), while the total gaze duration was coded as congruent if the dog looked towards the matching image longer than the mismatching image (Total gaze congruent score). The shapes judged to match the adult male voices in Experiment 1 and low pitched tones in Experiment 2 were the adult male silhouette in the People trials, the large square in the Size trials and the highest square in the Elevation trials. In contrast, when the re-synthesised child voices or high pitched tones were played, the matching images were the opposite image in each pair: the child silhouette in the People trials, the small square in the Size trials and the lowest square in the Elevation trials.

In Experiment 1, four trials were discarded prior to statistical analysis because the subjects did not look at either image during the voice presentation due to unsettled behaviour (four different subjects did not look at either of the images in one of their trials; three of which were Elevation trials and one was a Size trial; adult voices were played in two of the trials). In Experiment 2, a total of 12 trials were discarded prior to analysis because eight subjects did not look at either image. Six of the discarded trials were Elevation trials (four of which were accompanied by high pitch tones), two were People trials (both were accompanied by high pitch tones), and four were Size trials (two were accompanied by high pitch tones).

### Statistical Analysis

Separate binary logistic regression analyses were carried out on the First Look Congruent scores and Total Gaze Congruent scores for each experiment to test the effects of potential independent variables on the dogs' responses. The following variables were entered as categorical predictors: the subject's sex, the subject's nervousness towards people (reported as yes/no by their owner), the image pair, position of the congruent shape and the apparent age of the auditory stimulus (adult or child). The subject's age, the number of men currently living with the dog and number of children currently living with the dog were included as continuous predictors. Finally, an interaction term between the apparent age of the auditory stimulus and the image pair was also tested. A forwards stepwise entry method with a likelihood ratio statistic was used to construct the model by including only significant variables. Separate binomial tests were then conducted for the significant predictors to determine if they influenced performance significantly from 50% chance levels.

All analyses were conducted using SPSS version 22 (SPSS Inc, Chicago, IL, U.S.A.).

## Results

### Experiment 1. Cross-modal Perception of Human Age Cues using Human Voices

The binary logistic regression analyses identified a significant interaction between the image pair and the apparent age of the playback voice on the proportion of congruent responses for both the First Look Congruent scores ($Wald_2 = 9.96$, $P = .007$) (Figure 3a) and the Total gaze congruent scores ($Wald_2 = 10.63$, $P = .005$) (Figure 3b). For the First Look Congruent scores, there were also significant main effects of the position of the

congruent image ($Wald_3$ = 17.25, $P$ = .001) and the subject's sex ($Wald_1$ = 4.74, $P$ = .03). The apparent age of the playback voice, the subject's age, their nervousness towards people and the number of men and children currently living with the subject did not significantly affect the proportion of congruent responses for either dependent variable (all $P > .05$).

Planned comparisons showed that when the People image pair was presented and an adult voice was played, significantly more dogs looked faster (85.7% congruent; binomial test: $N$ = 14, $P$ = .01) and for longer (85.7% congruent; binomial test: $N$ = 14, $P$ = .01) towards the silhouette of the man than was expected by chance. In contrast, when the People image pair was presented and a re-synthesised 'child' voice was played, dogs looked at both silhouettes equivalently for both the First Look Congruent scores (61.5% congruent; binomial test: $N$ = 13, $P$ = .58) and the Total Gaze Congruent scores (61.5% congruent; binomial test: $N$ = 13, $P$ = .58).

For the Size image pair, dogs looked at both sized-squares equally for both the First Look Congruent scores (30.8% congruent; binomial test: $N$ = 13, $P$ = .27) and Total gaze congruent scores (23.1% congruent; binomial test: $N$ = 13, $P$ = .09) in response to the adult voice. Equally, dogs performed at chance levels for both the First look congruent scores (46.2% congruent; binomial test: $N$ = 13, $P$ = 1.00) and Total gaze congruent scores (30.8% congruent; binomial test: $N$ = 13, $P$ = .27) in response to the re-synthesised 'child' voice. In both the People and Size trials, significantly more subjects looked towards the image that was to their left side before they looked at the image to their right (73.6%; binomial test: N = 53, P > .001).

For the Elevation image pair, dogs looked significantly faster at the lower square independently of whether an adult male voice (9.1% congruent, binomial test: $N$ = 11, $P$ = .01) or a child's voice was presented (84.6% congruent; binomial test: $N$ = 13, $P$ = .02). Thus overall, significantly more of the subjects looked at the low shape before they looked at the high shape during the voice presentation (87.5%; binomial test: $N$ = 24, $P < .001$). However, whilst all of the dogs looked longer at the lower square when they heard a re-synthesized 'child' voice (Total Gaze Congruent Score: 100% congruent; binomial test: $N$ = 13, $P < .001$), dogs were equally likely to look longer at either the low and high square when the voice belonged to an adult male (Total Gaze Congruent Score: 27.3% congruent; binomial test: $N$ = 11, $P$ = .23).

Finally, although male dogs responded to the matching image first more often than females, neither the males (First Look Congruent Score: 55.8% congruent; binomial test: $N$ = 43, $P$ = .54), nor females (First Look Congruent Score: 44.1% congruent; binomial test: $N$ = 34, $P$ = .61) performed significantly differently from chance across the three trials.

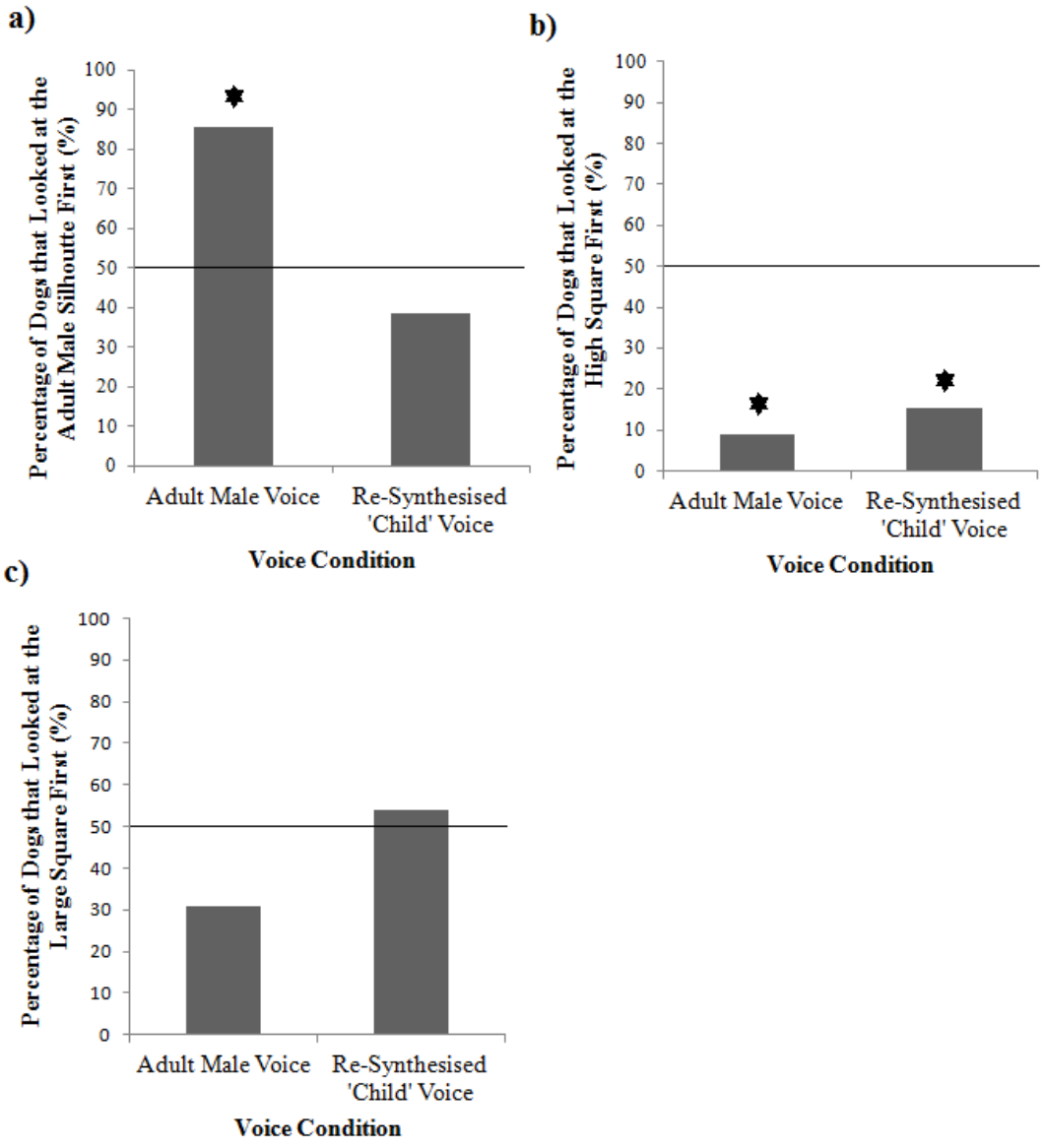



**Fig 3a**. The percentage of dogs that looked at the image matching the adult male voices first depending on the voice condition. a) People images: adult male voices matched the adult male silhouette; b) Elevation images: adult male voices matched the high square; c) Size images: adult male voices matched the large square. Asterisks indicate conditions in which the proportions were significantly different from chance (50%) at $p < .05$.

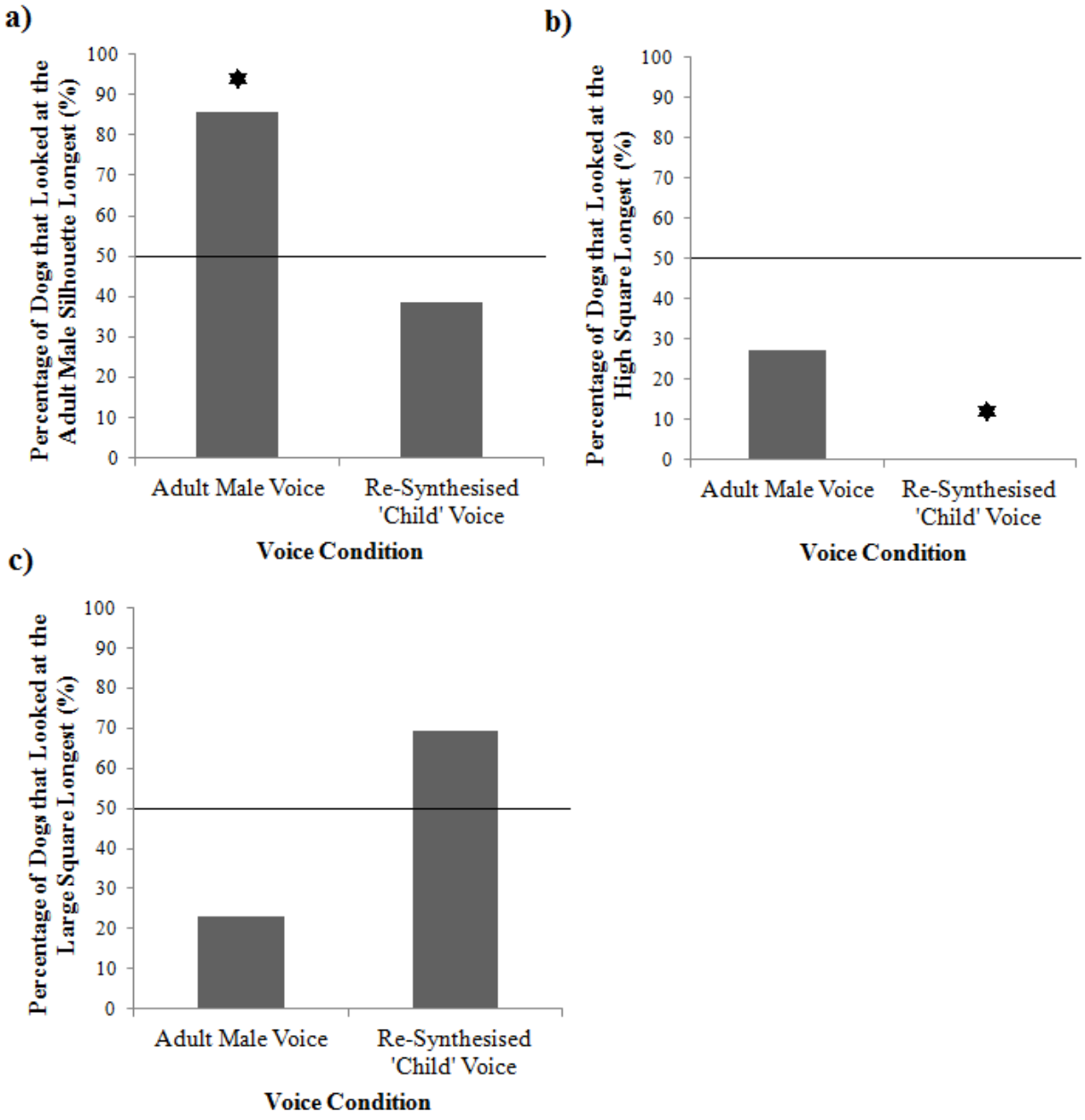



**Fig 3b**. The percentage of dogs that looked at the image matching the adult male voices longest depending on the voice condition. a) People images: adult male voices matched the adult male silhouette; b) Elevation images: adult male voices matched the high square; c) Size images: adult male voices matched the large square. Asterisks indicate conditions in which the proportions were significantly different from chance (50%) at $p < .05$.

**Experiment 2. Cross-modal Perception of Human Age Cues using Pure Tones**

The results of the binary logistic regression analyses indicated that none of the independent variables were significant predictors of either the First Look Congruent scores (all P > .06; Figure 4a) or the Total Gaze Congruent scores (all P > .08; Figure 4b).

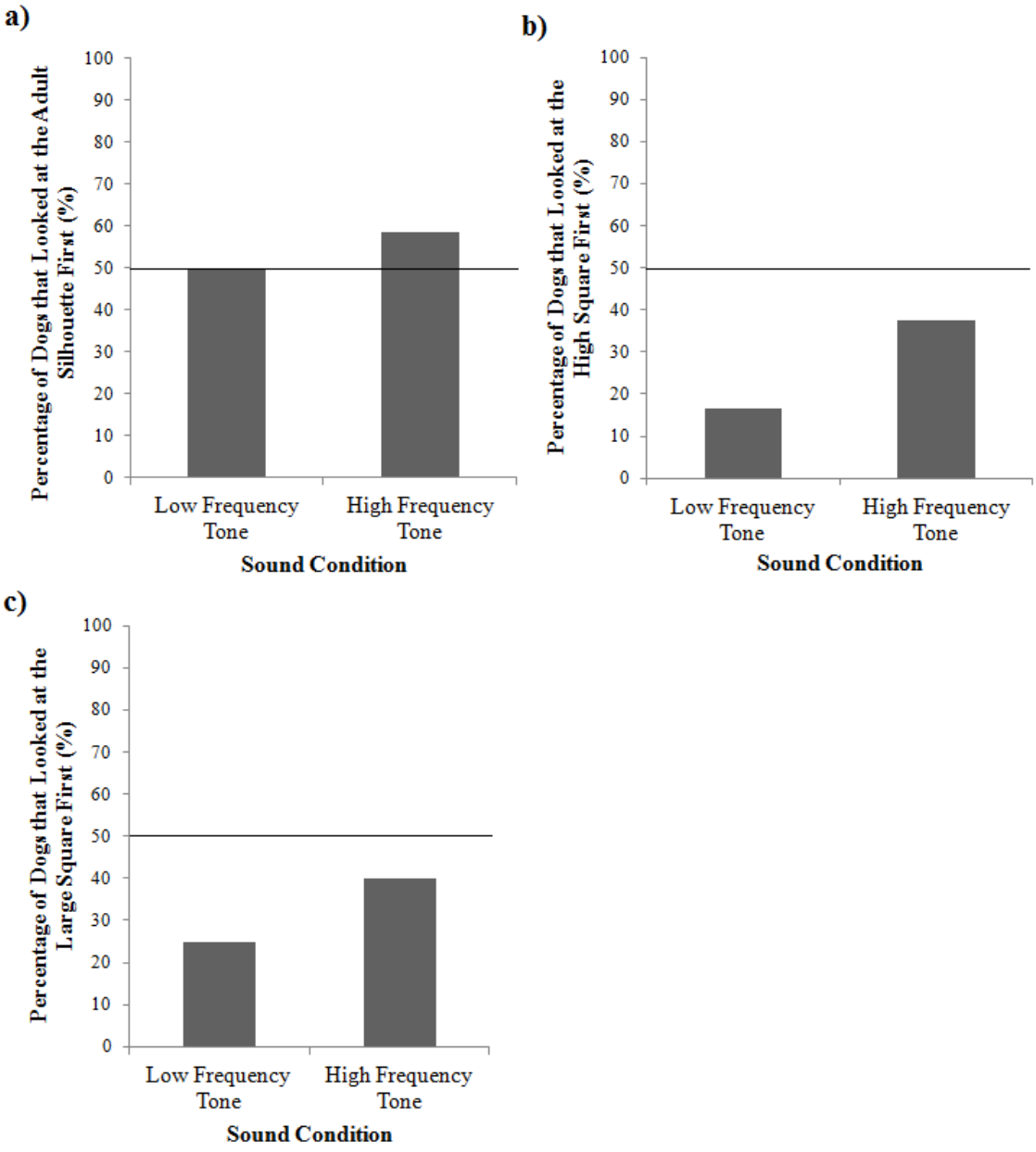



**Fig 4a**. The percentage of dogs that looked at the image matching the low frequency tones first depending on the sound condition. a) People images: low frequency tones matched the adult male silhouette; b) Elevation images: low frequency tones matched the high square; c) Size images: low frequency tones matched the large square.

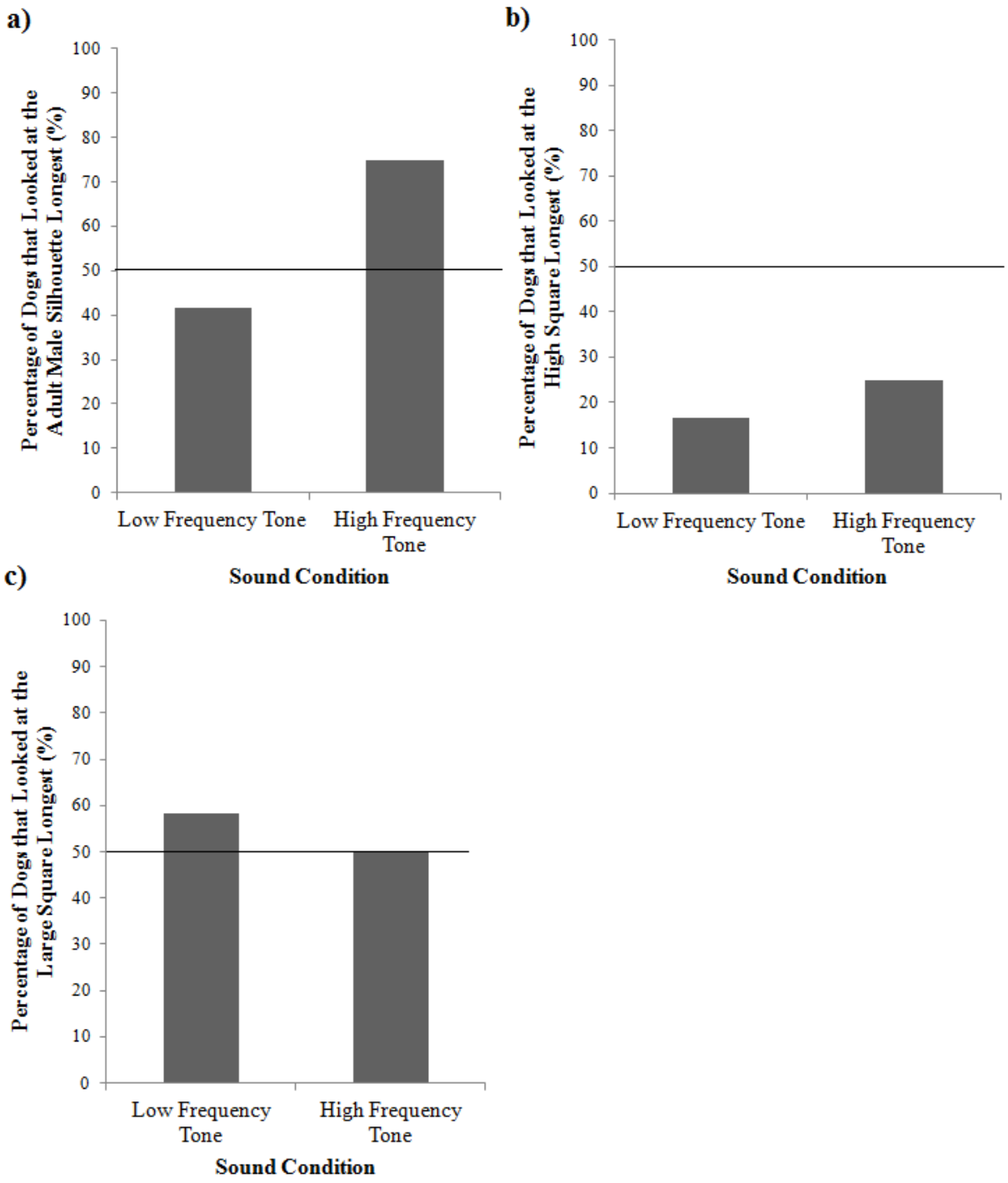



**Fig 4b.** The percentage of dogs that looked at the image matching the low frequency tones longest depending on the sound condition. a) People images: low frequency tones matched the adult male silhouette; b) Elevation images: low frequency tones matched the high square; c) Size images: low frequency tones matched the large square.

## Discussion

The current study aimed to determine if dogs spontaneously associate human silhouettes with human voices according to their age category (adult versus child), and to explore the perceptual mechanisms that may be involved in dogs' multisensory human age assessments. The results from Experiment 1 revealed that dogs did express the ability to associate the body shape of an adult with an adult voice, as the majority of the dogs tested looked faster and longer towards the silhouette of a man, rather than the silhouette of a young boy, when they heard a sustained vowel sound spoken by an unfamiliar man. The observation that subjects associated the silhouette of a man with adult male human voices directly builds on previous evidence that dogs match articulated phrases with unfamiliar human adults according to their gender (Ratcliffe et al., 2014), by indicating that dogs are capable of associating adult male voices with men using only their age-related body shape and the more limited acoustic information available in single vowel sounds. Dogs' ability to match adult male voices with the corresponding silhouette did not appear to be dependent on the perception of low-level size-related features in the visual domain, as the subjects did not associate the larger of two simple square shapes with the men's voices when presented with a pair of differently-sized squares. Similarly, the subjects did not associate simple height-related cues with the voices, as they did not match the shape positioned in a higher elevation with the adult male voices when presented with two same-sized squares in different spatial elevations. In fact, in contrast to our predictions, dogs that heard the adult male voice in the Elevation trials instead looked more quickly towards the lower positioned square than the higher positioned square, although they were equally likely to look longer at either square over the course of the trial. A related effect has been observed in humans, where listeners judge adult male voices as belonging to larger individuals when they are projected from a lower position (Pisanski, 2014), potentially due to an over-generalisation of the perceived correspondence between low frequency sounds and objects occupying low elevations present in the environment (Parise et al., 2014). However, because our subsequent tests with higher frequency sounds also produced greater gaze biases towards the shape in the lower position, as detailed below, the dogs' responses are unlikely to be similarly attributable to the over-generalisation of a perceived correspondence between sound frequencies and visual elevations.

Interestingly, the associations dogs made in response to the adult male voices were not reversed when we re-synthesised the voices so that the main anatomically-related acoustic cues to the age of the speaker, the formants and F0, mimicked the average values for a 6-year-old boy. As predicted, dogs no longer matched the adult male silhouette to the re-synthesised voices, suggesting that similarly to human listeners (e.g. Smith & Patterson, 2005), dogs rely on these key anatomically-related acoustic cues to discriminate adult male voices. However the subjects did not associate the silhouette of the child with the re-synthesised 'child' voices either. The lack of a matching response to the child stimuli indicates that the dogs in our study did not perceive the correspondence between formant and F0 values typical of children's voices and the body shape of a child. Therefore, although the subjects successfully discriminated and combined adult multisensory cues from those of a child, they appeared to be unable to similarly combine multisensory cues related to the child age category. The discrepancy between the dogs' responses to the adult and re-synthesised 'child' stimuli could be related to the subjects' generally low level of experience with children, as although the number of children living in the same household as the subject was not identified as a significant predictor of their performance, only three of the 27 subjects in Experiment 1 currently lived with children. In contrast, the majority of the dogs lived in the same household as at least one man (n = 21), providing the subjects with more regular exposure to adult male voices than to children's voices. Dogs may therefore need sufficient experience with a specific human age category in order to learn to combine multisensory cues relating to that age group. To determine the precise effect of this limitation in our sample, additional testing should be conducted with dogs that do currently live with children.

However, rather than relating to a lack of experience with children, it is also possible that the subjects did not recognise the child silhouette, but instead perceived it as an adult human stood at a further distance away from them. A wide variety of species have been shown to estimate the size of objects based on size consistency mechanisms (whereby more distant entities are judged as relatively larger despite their smaller retinal image) (Boycott & Young, 1956; Ingle & Cook, 1977; Locke, 1938; Pastore, 1958). Animals also perceive depth cues in two-dimensional pictures, as demonstrated by their susceptibility to visual illusions (e.g. 'the Ponzo illusion': Fujita, 1997; Gunderson, Yonas, Sargent & Grant-Webster, 1993; Timney & Keil, 1996). The slight expansion

and contraction of the silhouettes used in our study could have been interpreted as depth cues by the dogs, as animals can perceive expanding shapes as looming (Ghazanfar & Maier, 2009; Maier, Neuhoff, Logothetis & Ghazanfar, 2004; Schiff, Caviness & Gibson, 1962). Therefore some of the dogs might have actually perceived the child silhouette as a man stood further away, creating confusion when the child voice was presented. Although, if dogs did perceive both images as men stood at different distances, we may have also expected equally variable responses to the silhouettes when they heard the adult male voice, whereas they correctly matched this voice to the adult silhouette. An alternative potential explanation is that the subjects may not use the F0 and formant values to discriminate children's voices. The re-synthesis process in itself is unlikely to have strongly influenced the dogs' responses to the stimuli, as dogs have appropriately interpreted similarly re-synthesised dog growls in previous studies (e.g. Taylor et al., 2011). Yet, in their investigation of the perception of human voices by African elephants *Loxodonta africana*, McComb, Shannon, Sayialel and Moss (2014) observed that the elephants discriminated between the speech signals of men and women even when the F0 and formants were re-synthesised to mimic the values of the opposite gender. Although in our study the subjects did not respond to the re-synthesised voices in the same way as to the original voices (as they did not associate the child voices with the adult silhouette) it is possible that similarly to elephants, dogs use different acoustic cues to human listeners to distinguish children's voices. However, unlike McComb et al.'s (2014) study, our stimuli were very short vowel sounds, excluding any prosodic cues to age such as intonation and stress patterns. Because the availability of additional cues was so limited, it is more probable that the dogs did attend to the formant frequencies and/or F0 of the voices. Further testing could be carried out with natural children's voices in order to confirm this.

Although they did not express the ability to match the child silhouette with the re-synthesised 'child' voices, in accordance with our predictions the dogs did associate a shape placed at a low visual elevation with these voices, as significantly more subjects looked faster and for longer at the lowest square in the Elevation trials. Because subjects looked more quickly towards the low shape in response to the re-synthesised 'child' as well as the original adult male voices, this suggests that rather than over-generalising a perceived correlation between low frequency sounds and low spatial elevations, the dogs' initial tendency to look at the low shape first was not actually related to the voice

presented. Instead, the greater likelihood of looking at the lower shape first can probably be accounted for by the fact that the lower square was closer to eye level than the higher square for many of the subjects. However, in contrast to their initial gaze responses, all of the dogs exposed to the re-synthesised 'child' voice looked longer at the lowest shape, whilst the subjects presented with the adult voice were equally likely to look longer at either shape. This implies that dogs may perceive human voices with a greater formant spacing and higher pitch as being more likely to originate from lower to the ground. Therefore, it seems that whilst the dogs in our study had not learnt to associate children's voices with the typical body shape of a child based on the formant or F0 values, they did recognise a general correspondence between human voices with these acoustic values and lower elevations. While we may have similarly expected the subjects to map the shape positioned at a high visual elevation to the adult male human voices if dogs do initially learn to match human voices to either children or adults according to low-level projection height cues, humans show evidence of perceptual narrowing as multisensory associations develop, initially relying on low-level inter-sensory relations before combining only socio-ecologically relevant signals as specific higher-level relations are learnt (Lewkowicz & Ghazanfar, 2009). Although comparable evidence of perceptual narrowing has not been observed in non-human animals (e.g. infant vervet monkeys *Cercopithecus aethiops*: Zangenehpour, Ghazanfar, Lewkowicz & Zatorre, 2009), it remains possible that narrowing may still occur but at a much slower rate than in humans (Zangenehpour et al., 2009). This leads to the possibility that dogs may initially associate child and adult human voices with different visual elevations, but cease to make general associations using this basic visual cue as they learn to combine only ecologically relevant information relating specifically to human bodies. This hypothesis could be tested by replicating the current procedure with either puppies or adult dogs that have different levels of experience with human age and sex groups.

While dogs show some evidence of associating acoustic age-related cues in human voices with low-level features in the visual domain, the suggestion that similarly to humans (Rendall et al., 2007), animals may be influenced by general low-level cross-modal correspondences during their perception of age-related information in vocalisations, particularly by mapping the F0 with corresponding visual cues (Morton, 1977), is not supported by the results of Experiment 2. Unlike previous demonstrations

with human listeners (Rendall et al., 2007; Pisanski, 2014), the dogs in the current study did not match high and low visual elevations (e.g. Rusconi, Kwan, Giordano, Umilta & Butterworth, 2006; Walker et al., 2010) or small and large sizes (e.g. Parise & Spence, 2009) with the high and low frequency pure tones respectively. The perception of broad cross-modal correspondences between auditory pitch and different visual dimensions remains largely unexplored in non-human animals, although an association between higher frequency tones and greater visual luminance has been demonstrated in chimpanzees *Pan troglodytes* (Ludwig, Adachi & Matsuzawa, 2011). However, even though the dogs in the present study did not express any perception of comparable cross-modal correspondences to humans, this lack of effect may be related to the comparatively small differences between the high and low frequency sound conditions used in our study (see Spence & Deroy, 2013). To determine if dogs do perceive more general cross-modal correspondences between auditory pitch and visual size or elevation, our procedure in Experiment 2 could be replicated with a greater frequency difference between the auditory conditions.

Furthermore, because the adult and re-synthesised child voice categories used in Experiment 1 differed only in their F0 and formant values, the lack of any associations between the image pairs and tones matching the F0 of the voices suggests that that the vocal pitch alone may not be sufficient for dogs to attribute human voices to age-related visual cues, and instead implies that the formant frequencies may also be a necessary component for them to make these associations. The potential importance of formants for dogs to categorise the age of human voices is supported by evidence that dogs attend to the spacing of the formant frequencies to associate growls with size-matching conspecifics (Taylor et al., 2011). The relative contributions of the F0 and formants for dogs' perception of human voices could be examined by adapting the procedure used by Smith and Patterson (2005), where by independently manipulating the F0 and formants of human voices, the authors determined that human judgements of speaker sex and age were equivalently influenced by both cues.

Finally, as well as providing initial insights into the mechanisms involved in associating human voices with visual cues, the results also indicated that in addition to showing a general tendency to look at the lower shape first in the Elevation trials during Experiment 1, dogs were also more likely to look at the image on their left side before the image on their right in both the People and Size trials, independently of which

image was in that position. Side biases in the opposite direction (towards the right side) have previously been observed in preferential looking studies when domesticated animals have been faced with two people and presented with recordings of human speech (dogs: Ratcliffe et al., 2014; horses *Equus caballus*: Proops & McComb, 2012). The opposite left side bias obtained specifically in response to the human voices in Experiment 1 may be related to the fact that we used simple vowel sounds with no meaningful phonemic content, whilst familiar phrases were used as stimuli in the previous studies. Indeed, in agreement with the direction of the biases observed, dogs express right orienting biases in response to meaningful phonemic content in human speech, and left orienting biases when this information is absent from speech signals (Ratcliffe & Reby, 2014). Because all of the visual stimuli were counterbalanced in our study, the general side bias shown in response to the human voices could not have confounded any of the other results obtained.

**Conclusion**

Our results demonstrate that the dogs in the current study spontaneously matched the silhouettes of adult male humans with men's voices. This association does not seem to be based solely on the perception of low-level visual features of size or spatial elevation, indicating instead that dogs specifically match the average body shape of an adult man to adult male voices. However, although dogs appear to attend to the formant frequencies and/or F0 positioning in adult male voices to match them with the corresponding speakers, they appeared to be unable to use these anatomically related acoustic cues to associate the body shape of a child with children's voices. Interestingly, although they did not express the ability to match the correspondingly-aged silhouette to the re-synthesised 'child' voices using the formant frequencies and/or F0 positioning, potentially due to a lack of sufficient prior exposure to children, the subjects did appear to associate lower spatial elevations with voices with a relatively wider formant spacing and higher F0 (typical of children's voices). Finally, none of the dogs' responses appeared to be related to a generalisation of low-level frequency-based cross-modal correlations present in the environment. Together, these observations provide initial evidence of the specific visual and vocal cues that dogs use in order to make cross-modal assessments about human age categories, suggesting that dogs may initially learn the anatomically-related vocal parameters typifying different human age-categories and match these cues with different projection heights, before their perception narrows to

associate formant frequency and/or F0 values with specific human age groups. However, due to the lack of consistency between the subjects' responses to the child and adult stimuli, it is not currently possible to confirm whether dogs are capable of forming multisensory representations of multiple human age categories. Further investigations with dogs that have had varying levels of exposure to different human age and sex groups are therefore necessary to develop a more complete understanding of this potential learning process.

# CHAPTER 5: ORIENTING ASYMMETRIES IN DOGS' RESPONSES TO DIFFERENT COMMUNICATORY COMPONENTS OF HUMAN SPEECH

## Synopsis

---

**Question:** Do dogs show evidence of hemispheric asymmetries when processing the main communicative components of human speech, and if so, are asymmetries dependent on the acoustic structure of the signals or their functional content?

---

**Methods:** Using a head-orienting paradigm, dogs were positioned between two loud speakers, and either a human voice or control sound was presented simultaneously from both sides. Each dog heard one sound from one of ten possible conditions. Eight of these were human voices that had been re-synthesised to increase the salience of either the segmental or supra-segmental cues: meaningful (a familiar command) or meaningless (unfamiliar speech) phonemic cues; speaker-related cues (accent, age, gender); or emotional prosody (intonation). The final two conditions were non-vocal controls. The direction of the dogs' head-turn to the left or right in response to the sound provided a behavioural indication of hemispheric processing biases.

---

**Results and conclusions:** Similarly to humans, dogs showed a right head turn/left hemispheric bias when the salience of meaningful segmental/phonemic cues in speech was increased, whilst they expressed a left head turn/right hemispheric bias when the salience of the supra-segmental cues was increased. This suggests that dogs dissociate and process the communicative cues in human speech in a way that broadly parallels speech perception by human listeners.

*Note.* Based on the published article: Ratcliffe, V.F., & Reby, D. (2014). Orienting asymmetries in dogs' responses to different communicatory components of human speech, *Current Biology*, *24*, 2908-2912. Formatted in the style of *Animal Behaviour*.


## Abstract

It is well established that in speech perception the left hemisphere of the human brain is more specialised in processing intelligible phonemic (segmental) content, whilst the right hemisphere is more sensitive to prosodic (supra-segmental) cues. Human speech can also be a familiar and relevant signal for domesticated animals; however, despite evidence that a range of mammal species show a left hemispheric specialisation when processing conspecific vocalisations, the presence of hemispheric biases in domesticated animals' responses to the communicative components of human speech has never been investigated. Using the head-orienting paradigm, we presented domestic dogs *Canis familiaris* with manipulated speech and tones differing in segmental or supra-segmental content and recorded their orienting responses. We determined that dogs showed a significant left hemispheric bias when presented with a familiar spoken command in which the salience of meaningful phonemic (segmental) cues was artificially increased, but a significant right hemispheric bias in response to commands where the salience of intonational or speaker-related (supra-segmental) vocal cues was increased. Our results provide insights into mechanisms of inter-specific vocal perception in a domesticated mammal, and suggest that dogs may share ancestral or convergent hemispheric specialisations for processing the different functional components of speech with human listeners.


## Introduction

Human speech is a complex vocal signal, transmitting information about the physical and motivational attributes of the speaker in addition to the linguistic content. During speech perception, humans show functional asymmetries in neurological activity in response to specific aspects of the acoustic signal. There is robust evidence, obtained using a broad range of behavioural and neurological experimental techniques, demonstrating that the left hemisphere of the human brain is typically more specialised in processing meaningful linguistic (segmental) content, whilst the right hemisphere

responds more strongly to speaker-related information, including the emotional prosody, voice identity and gender cues (supra-segmental content) (see Belin, Fecteau & Bedard, 2004 and McGettigan & Scott, 2012 for reviews). Indications of comparable parallel, hierarchical processing streams in the auditory cortical regions of non-human primates has led to the suggestion that the brain structures which support speech perception may not be uniquely human, and could have adapted from a phylogenetically older system (Rauschecker & Scott, 2009; Scott & Johnsrude, 2003). Indeed, consistent with the human literature, studies using brain lesioning (Heffner & Heffner, 1984) and PET imaging procedures have demonstrated that in macaque monkeys, the left cerebral hemisphere is more strongly recruited during the perception of conspecific vocalisations (Kikuchi, Horwitz & Mishkin, 2010; Poremba et al., 2004), while Petkov et al., (2008) determined that right hemisphere of the macaque brain is sensitive to the familiarity of the caller. Hemispheric asymmetries for processing conspecific vocalisations appear to be widespread across mammals (see Ocklenburg, Ströckens & Güntürkün, 2013, for a review), predominantly demonstrated using the behavioural 'head orienting' paradigm, which was first developed by Hauser and Andersson (1994) to investigate lateralised orienting in rhesus macaques *Macaca mulatta* in response to conspecific vocalisations. Sounds were played from directly behind the subjects and the direction in which the animal turned towards the sound source was recorded. Adult macaques showed a significant right head-turn bias in response to a range of species-specific calls, whilst familiar hetero-specific (bird alarm) calls elicited a left turning bias. Given that in mammals auditory information entering each ear is processed primarily in the contra-lateral hemisphere of the brain via the dominant contra-lateral auditory pathways (Grimshaw, Kwasny, Covell & Johnson, 2003; Rosenzweig, 1951; Tunturi, 1946), it was assumed that the macaques expressed a left hemispheric specialisation for processing species-specific vocalisations, corresponding with the results obtained from neuro-imaging studies. Using the head orienting paradigm, right head-turn biases in response to conspecific vocalisations have since been observed in more phylogenetically distant mammal species, including California sea lions *Zalophus californianus* (Böye, Güntürkün & Vauclair, 2005) and domestic dogs (Siniscalchi, Quaranta & Rogers, 2008), suggesting that a left hemispheric specialisation for processing conspecific vocalisations may be broadly shared across mammals.

Hemispheric specialisations in the perception of conspecific vocalisations are perhaps unsurprising given their ecological significance to receivers (Poremba, Bigelow & Rossi, 2013). For domesticated species, human speech can be a similarly familiar and functionally relevant vocal signal, and may also have the potential to elicit neurological processing biases. Determining hemispheric activation in non-human animal responses to the different communicative components of speech could provide key insights into how speech lateralisation evolved in the human brain. However, the presence of processing asymmetries during domesticated animals' perception of speech remains largely unexplored. The domestic dog is a particularly well suited model species to investigate this, as there is evidence to suggest that dogs may respond to segmental phonemic cues (Baru, 1975; Fukuzawa, Mills & Cooper, 2005) in addition to supra-segmental speaker-related (Adachi, Kuwahata & Fujita, 2007; Ratcliffe, McComb & Reby, 2014) and emotional (Scheider, Grassmann, Kaminski & Tomasello, 2011) prosodic information in speech signals. As well as exhibiting a left hemispheric bias in response to conspecific vocalisations (Siniscalchi, Lusito, Sasso & Quaranta, 2012; Siniscalchi et al., 2008), a recent fMRI study demonstrated that auditory regions in the dog's right cerebral hemisphere were sensitive to the emotional valence of both dog and human non-verbal vocalisations (Andics, Gácsi, Faragó, Kis, & Miklósi, 2014), suggesting that hemispheric lateralisation also extends to human vocal signals in this species. Although Reinholz-Trojan, Włodarczyk, Trojan, Kulczyński, and Stefańska (2012) observed no lateralised head-turning response when dogs were presented with a learnt spoken command, natural speech stimuli contain both segmental and supra-segmental cues, which are known to produce different biases in hemispheric activation in humans and might also have influenced the subjects' responses. Therefore, the aim of the current study was to investigate whether dogs would show hemispheric asymmetries in response to the segmental and/or supra-segmental components of human speech separately, by manipulating the acoustic content of natural speech signals. Using a between-subject head-orienting design, in Experiment 1 dogs were presented with human spoken commands in which the relative salience of either the segmental (phonemic) or supra-segmental (speaker-related) content was artificially increased. Experiment 2 then aimed to establish if the responses obtained in Experiment 1 were related to the communicative content of the signals or to their acoustic structure. In each experiment we recorded the presence of any orienting response biases to the different

auditory conditions, which provides a behavioural indication of hemispheric processing asymmetries.

## Experiment 1: Dogs' Orienting Responses to Human Speech Commands when the Salience of the Segmental or Supra-segmental Cues is Manipulated

Five different human speech-related sound conditions were used to compare the subjects' responses to the segmental versus supra-segmental cues. For the first two conditions, we enhanced the salience of familiar segmental content in the signals. In Test 1, dogs were presented with a recording of a familiar learnt command in which the original positive intonational cues had been artificially degraded to increase the salience of the phonemic content ("come on then" with a flat intonation; Meaningful Speech with Neutralised Intonation). The same command was then further degraded in Test 2 by replacing the first three formants with sine wave tones (Meaningful Sine-Wave Speech), strongly reducing the supra-segmental cues (both emotional and speaker-related) but preserving the segmental phonemic information.

In contrast, in the subsequent two conditions the salience of the supra-segmental content was enhanced. Both speaker-related (indexical) and emotional (dynamic) cues are encoded in the supra-segmental content of speech signals. Dogs' responses to speaker-related indexical cues were investigated by exposing them to a comparable phrase with neutralised intonation, but spoken in an unfamiliar language (Test 3: Meaningless (Foreign) Speech with Neutralised Intonation). Here the phonemic cues were unfamiliar and the intonational prosodic cues were removed, whereas the indexical speaker-related cues remained intact. We also tested dogs' responses to emotional prosodic cues by presenting them with a version of the original command in which the phonemic components had been removed by extracting the formants and plosives, creating unintelligible speech-like vocal stimuli with reduced speaker cues, but retained positive emotional prosody (Test 4: Meaningless Voice with Positive Intonation). Finally, in Test 5 dogs were exposed to natural meaningful speech containing both segmental phonemic and supra-segmental prosodic cues ("come on then" with happy intonation; Meaningful Speech with Positive Intonation) (see Figure 4 for example spectrograms of each of the acoustic stimuli used in Experiment 1 and 2).

## Method

### Subjects

Subject animals were over six months old, healthy with no known hearing or sight problems and not aggressive towards people. Owners of dogs exposed to the 'meaningful' speech conditions confirmed that their dog responded to the command 'come on then' or a similar variant, whilst only dogs with no previous exposure to French were presented with 'meaningless' speech. An *a priori* power analysis conducted using G*Power (Faul, Erdfelder, Lang & Buchner, 2007) with power $(1 - \beta)$ set at 0.80 and $\alpha = 0.05$, two-tailed, showed that a minimum sample size of 20 was required in each condition for detecting a medium effect size in a binomial test. We included the first 25 dogs that reacted to the stimuli in each condition. A small proportion of subjects (n=19) failed to react to the stimuli (with an even distribution of failed responses across conditions ($\chi^2_{(4)} = 4.95$, $p = 0.29$)), and were excluded from the study at the time of testing. The 125 dogs retained in the analysis included 61 females and 64 males from 44 different breeds. Ages ranged from six months to 13 years old (mean ± SD = 4.12 ± 3.00 years). One hundred and five dogs were privately owned pets whilst 20 dogs were housed in a local animal shelter.

### Stimuli Acquisition

*Voice Recordings*

Four men and four women, who were native British speakers and aged between 20 and 58 years old (mean ± SD = 30.25 ± 13.68 years), were audio recorded after being instructed to pronounce the phrase 'come on then' once in a happy tone of voice (Meaningful Speech with Positive Intonation). Acoustic analyses determined that the spoken phrases were produced with a relatively high mean fundamental frequency (F0; perceived pitch) (men: mean = 206.76 Hz; women: mean = 321.29 Hz) and large F0 range (men: mean = 111.56 Hz; women: mean = 186.61 Hz), which is consistent with previous observations that the expression of happiness in speech is characterised by a raised mean F0 and higher F0 variability in relation to neutral speech (Banse & Scherer, 1996; Juslin & Laukka, 2003). These original recordings were used as stimuli in Test 5 (Meaningful Speech with Positive Intonation).

Eight native French speakers, four men and four women aged between 22 and 56 years old (mean ± SD = 35.50 ± 15.17 years), were also recorded pronouncing the phrase 'aller viens le chien' in a happy tone of voice (used to create Meaningless (Foreign) Speech with Neutralised Intonation; Test 3). All of the recordings were made using a Zoom H4N Handy Recorder in a sound proof booth. The sampling frequency was set at 44 100 Hz, with a 32-bit sampling rate, across recordings. Each recording was then normalised to -1.0 dB maximum amplitude using Audacity 2.0.0. (http://audacity.sourceforge.net).

*Acoustic Manipulations*

The content of the recordings was manipulated in three different ways, using PRAAT v.5.0.3 (http://www.fon.hum.uva.nl/praat/) (see Figure 4 for example spectrograms).

Neutralising the intonation contour (Tests 1 and 3)

The first manipulation was carried out on all of the recordings and aimed to reduce the presence of emotionally related prosodic cues, whilst retaining intact phonemic content. Although the vocal portrayal of happiness also involves other acoustic parameters, F0 related cues are the most perceptually prominent features for human listeners (Juslin & Laukka, 2001). Therefore to create perceptually neutral stimuli PSOLA re-synthesis was used to lower the F0 in each recording to a value typical of neutral speech for an average man or woman (110 or 220 Hz respectively; Titze, 1994), and all F0 variation was removed. This manipulation preserved clearly intelligible speech, but reduced the emotional prosodic content. The resulting re-synthesised English stimuli were used in Test 1 (Meaningful Speech with Neutralised Intonation), whilst the French stimuli were used in Test 3 (Meaningless (Foreign) Speech with Neutralised Intonation).

Creating sine-wave speech (Test 2)

The second manipulation was carried out only on the English recordings, and aimed to create sine-wave speech signals (Meaningful Sine-Wave Speech). This was achieved using the 'SWS script' written by Chris Darwin (http://www.lifesci.sussex.ac.uk/home/Chris_Darwin/Praatscripts/SWS). The algorithm uses LPC to estimate the first three formant frequencies, and the formant amplitudes are taken from a wideband FFT spectrum. The estimates are then smoothed to produce continuous contours and remove residual artefacts from the signal. This produces three

sinusoid curves which track the first three formants of the original voice; therefore the re-synthesised sound is still intelligible but not voice-like. The resulting stimuli were used in Test 2 (Meaningful Sine-Wave Speech).

Removing the phonemic cues (Test 4)

The third manipulation was also carried out only on the English recordings, and aimed to remove the phonemic information, whilst retaining all other acoustic features of the sound, therefore preserving the emotional content. To do this the plosives were manually cut from the signal, and the spectral envelope of the speech utterance was flattened (using LPC synthesis/inverse filtering) in order to remove the temporal and formant-related phonemic information whilst retaining the F0-related modulation associated with the speech prosody. The LPC analysis uses linear-prediction to estimate the first five formant frequencies and bandwidths in the waveform, producing a smoothed version of the spectrogram. To perform this analysis the recordings were first re-sampled to either 10 kHz or 11 kHz (for a male or female voice respectively) to provide a band limit of either 5000 Hz or 5500 Hz. A pre-emphasis of 50 Hz was also added prior to the analysis. Ten linear-prediction parameters were used, with an analysis window of 25 ms and time steps of 5 ms. This produced an LPC object for each recording approximating the formant frequencies and the source signal (residue). Inverse filtering was then performed on the original sounds using the LPC objects, to remove the formant frequencies. These stimuli were used in Test 4 (Meaningless Voice with Positive Intonation).

*Perceptual Ratings*

To verify the validity of the intended manipulations, five volunteers (two men, three women) who were naïve to the experimental conditions rated each stimulus (in addition to recordings of speech and non-verbal vocalisations with angry emotional prosody which were not used in the current study) in a listening experiment run using PRAAT. Each sound was scored on both scales and could be replayed multiple times before rating. Volunteers were asked to rate any speech other than English as unintelligible. Two 5-point Likert scales were used to score each sound. On the first scale, participants were asked if they could understand what the person was saying, providing a rating of intelligibility (1 = very unclear, 5 = very clear), and on the second scale they rated the emotional valence of the sound (1 = very negative, 5 = very positive) (Table 1).

**Table 1.** Mean ratings of emotional content and intelligibility for the stimuli used in each auditory condition.

| Auditory Condition | Mean Score for Intelligibility (3 = medium clarity) | Mean Score for Emotional Content (3 = neutral) |
|---|---|---|
| Meaningful Speech with Neutralised Intonation | 4.63 | 3.13 |
| Meaningful Sine-Wave Speech | 3.24 | 3.19 |
| Meaningless (Foreign) Speech with Neutralised Intonation | 1.00 | 2.94 |
| Meaningless Voice with Positive Intonation | 1.90 | 4.13 |
| Meaningful Speech with Positive Intonation | 4.79 | 4.21 |

**Experimental Set-up**

Tests were carried out between May 2013 and April 2014 in one indoor location (a designated experimental room at the University of Sussex) and two outdoor locations (Brighton RSPCA exercise field and Stanmer Park) in the local East Sussex area. Outdoor trials were only conducted on days without wind or rain, in quiet open areas away from pedestrian and road traffic. Trials were only initiated when no other people or animals were in close vicinity to the test site.

Two speakers (SONY SRS-A60) were placed 1.5 m to the right and left of a centre point. The side of each speaker was counter-balanced across subjects. The speakers were connected to a laptop placed on a table 3 m from the centre point. A video camera was positioned underneath the table to record the dog's response (Figure 1). A N05CC Digital Mini Sound Level Meter was used to ensure that the speakers broadcast at the same volume.

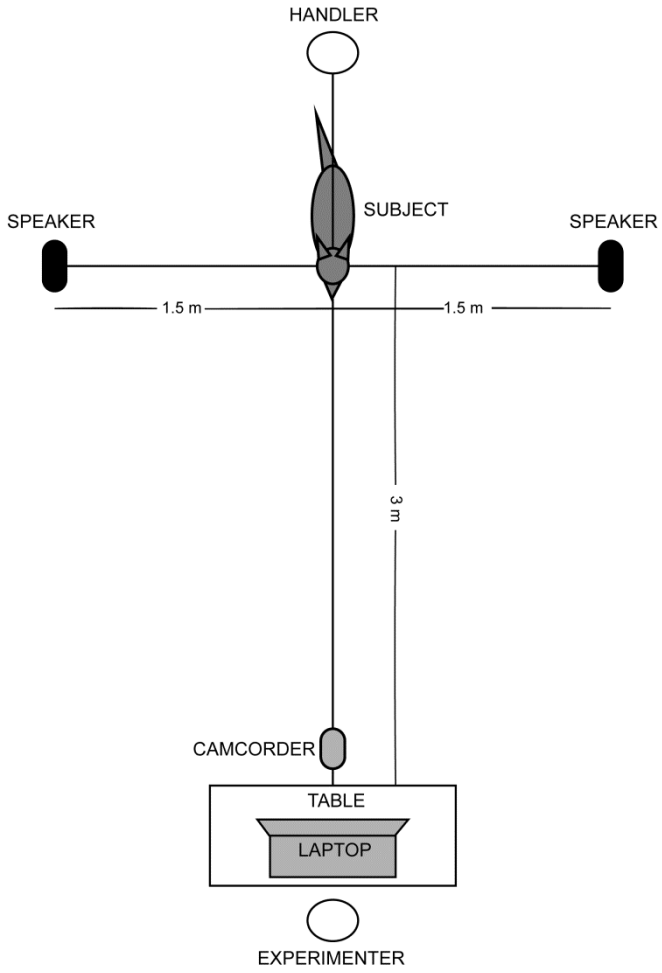



**Fig 1.** Experimental set up with distances between the subject, loud speakers and experimenter

**Procedure**

The dog was held on a loose lead by their owner (or a research assistant for shelter housed dogs) who was naïve to the experimental conditions. Owners positioned their dog at the centre point, facing the table, and then stood still directly behind their dog. The experimenter stood behind the table facing the dog and attracted the dog's attention by saying their name. When the dog was stationary and facing directly forwards the experimenter looked down at the laptop (to avoid providing any gaze cues) and played the stimulus once. Stimuli were presented at 65 dB in pseudo-randomised order across trials, with equal numbers of male and female voices until 25 subjects responded in each condition. Trials ended when the dog was no longer oriented towards one of the speakers. Dogs that did not react to the sound between the stimulus onset and two seconds after the offset were recorded as non-responsive. Each dog took part in only

one trial in which they were presented with a single sound stimulus from one of the five speech conditions.

**Ethical Note**

The study was approved by the University of Sussex ethics committees (certificates ERC/33/3 and DRVR0312).

**Behavioural Measures and Statistical Analysis**

Videos were coded using the digital video analysis software Sportscode Gamebreaker version 7.5.5 (Sportstec, Warriewood, NSW, Australia). The first direction of movement of the dog's head to the left or right immediately after the onset of the playback stimulus was the main response variable. A research assistant naïve to the auditory condition second coded 10% of the videos with 100% inter-observer agreement for the orienting direction.

To test the effect of potentially meaningful independent variables (IVs) on orientation biases, we ran a binary logistic regression with the following IVs: subject's sex, age, breed type and current residence (animal shelter or private home), test condition, stimulus voice gender, stimulus exemplar and test location (inside or outside). The dependent variable was the side of the response (left or right). A forwards entry stepwise method with a likelihood ratio statistic was used so that only significant IVs were included in the final model. Binomial probability tests were then carried out on significant IVs.

## Results and Discussion

A binary logistic regression analysis identified a significant overall effect of auditory condition on head-turn direction (Wald(4) = 28.59, $p < 0.001$), indicating that the content of the acoustic signals affected the direction of the dogs' orienting responses during sound perception (Figure 2). In contrast, there was no significant effect of subject sex ($p = 0.88$), age ($p = 0.34$), breed type ($p = 0.53$), current residence (animal shelter or private home) ($p = 0.10$), stimulus exemplar ($p = 0.32$), stimulus voice gender ($p = 0.56$) or test location ($p = 0.31$) on responses. Separate binomial tests were subsequently performed to establish whether there were significant orienting response biases for each of the auditory conditions.

It was determined that dogs presented with Meaningful Speech with Neutralised Intonation (Test 1) showed a significant right head-turn response bias (binomial test: (80% Right head-turn), p = 0.004). Because orienting biases are assumed to result from stronger processing in the contra-lateral hemisphere of the brain, the right head-turn bias observed suggests that when supra-segmental intonation is neutralised and segmental phonemic cues become more salient, dogs display a left hemispheric advantage for processing human speech. The right orienting bias shown in response to salient phonemic cues was further verified in Test 2, as the subjects exposed to Meaningful Sine-wave Speech also demonstrated a significant right head-turn bias (Binomial test: (76% Right head-turn), p = 0.015). The dogs' responses to sine-wave versions of the speech signals reinforced the interpretation that in dogs the left hemisphere of the brain is sensitive to segmental phonemic information, and that this sensitivity is independent from the nature and naturalness of the acoustic elements composing the signal. These observations parallel the stronger left hemispheric sensitivity observed in humans when processing intelligible phonemic content in natural speech (e.g. Jerger & Martin, 2004; Kimura, 1961; McGettigan et al., 2012) and sine-wave speech signals (Möttönen et al., 2006). This conspicuous cross-species similarity could be related to the lateralised processing of conspecific vocal signals, as dogs also express a right-head turn bias in response to conspecific vocalisations (Siniscalchi et al., 2012; Siniscalchi et al., 2008), suggesting that the same processing mechanism may also respond to the phonemic content in human speech.

In contrast to the responses obtained for meaningful phonemic cues, when dogs were exposed to Meaningless (Foreign) Speech with Neutralised Intonation (Test 3), which aimed to increase the salience of the speaker-related supra-segmental content, they instead showed a significant left head-turn bias (Binomial test: 24% right head turn, p = .015). In this test, phonemic cues were still present in the signals but were unfamiliar to the dogs, whilst the indexical cues remained intact, suggesting that dogs may show a right hemispheric advantage when processing salient speaker-related supra-segmental content in speech. Dogs are known to perceive speaker-related human vocal cues such as identity (Adachi et al., 2007) and gender (Ratcliffe et al., 2014), and the observed right hemispheric advantage is consistent with human lateralisation when processing these features (Belin & Zatorre, 2003; Lattner et al., 2005; von Kriegstein, Eger, Kleinschmidt & Giraud, 2003).

The suggestion that the left orienting bias in Test 3 was generated by greater attention to the supra-segmental content of the speech signal is further supported by our observation that dogs also showed a significant left head-turn bias when they were presented with a Meaningless Voice with Positive Intonation (Test 4) (Binomial test: 28% right head turn, $p = .04$). In this condition, there was no phonemic content in the signals and the speaker-related cues were strongly degraded, leaving only the emotional prosody intact, indicating that when segmental phonemic cues are neutralised and supra-segmental emotional prosodic cues become more salient, dogs again display a right hemispheric advantage. This result furthers recent neuro-imaging evidence that auditory regions in the dog's right hemisphere are sensitive to the emotional valence in both conspecific calls and human non-verbal vocalisations, with increased activation in response to calls with greater positive valence (Andics et al., 2014). This is consistent with the observation that humans show stronger right hemispheric activation not only in response to emotional speech prosody and vocalisations, but also when exposed to animal vocalisations with strong affective content, independently of their familiarity with the species (Belin et al., 2008). This suggests that the perception of emotional content in vocalisations, and its lateralisation to the right hemisphere, may be conserved across mammals.

Finally, in Test 5, dogs were exposed to Meaningful Speech with Positive Intonation, containing intact segmental phonemic and supra-segmental prosodic cues, and no significant head-turn bias was observed (Binomial test: 48% right head turn, $p = 1.00$). While directing dogs' attention to either of these components using manipulated speech was found to produce opposite hemispheric biases in previous tests, the simultaneous presence of salient segmental and supra-segmental cues that characterise neutral speech resulted in the absence of a bias at the population level (Figure 2). Interestingly, in humans, different hemispheric activation can be produced in response to speech depending on the attentional focus of the listener to the segmental or supra-segmental content (Grimshaw et al., 2003; Mitchell, Elliott, Barry, Cruttenden & Woodruff, 2003). This raises the possibility that differential attentional focus, and thus orientation biases, in the individual dogs' responses to the stimuli may also have resulted in an absence of any overall bias across the subjects.

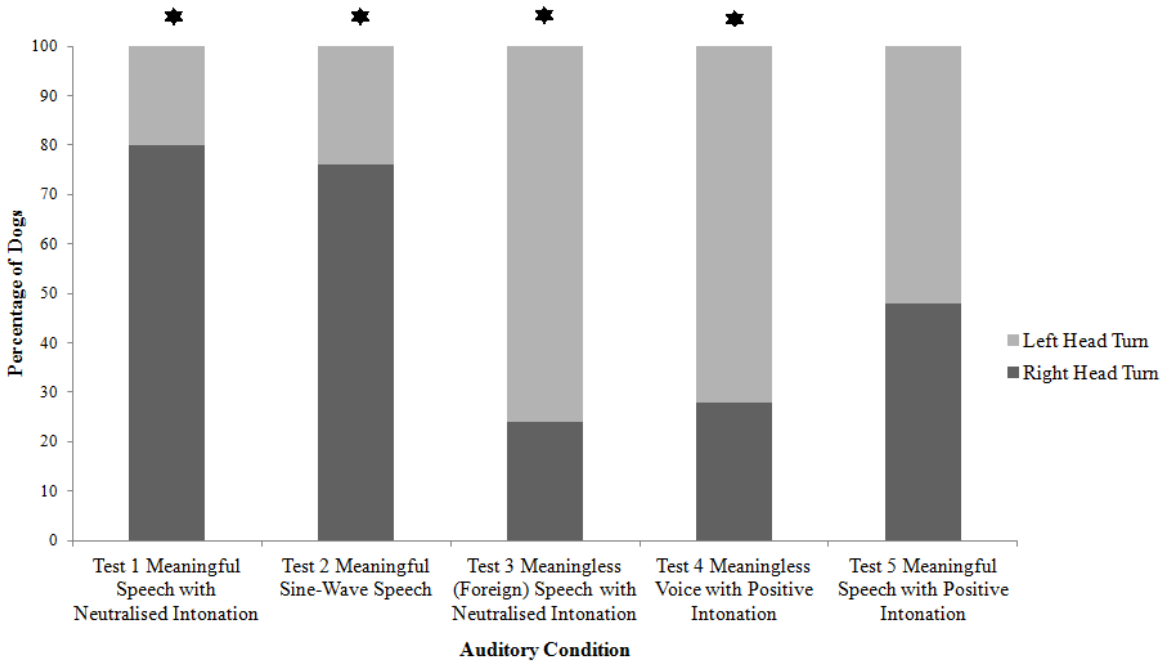



**Fig 2.** Percentage of dogs that oriented to their left or right in each condition after the playback presentation. Asterisks indicate conditions in which the proportions were significantly different from chance (50%) at $p < 0.05$.

Taken together, our results indicate that dogs appear to differentially process the segmental and supra-segmental components of human speech. Furthermore, the direction of their orienting biases suggests that dogs preferentially process phonemic cues in the left hemisphere of the brain, whilst the right hemisphere appears to be more strongly involved in processing the supra-segmental content of the signal. These biases directly converge with those observed in human listeners during behavioural dichotic listening tasks (reviewed by Kimura, 2011) and through functional neuro-imaging techniques (e.g. Mitchell et al., 2003). In humans, it is unclear whether the hemispheric specialisations related to speech perception are speech-specific or if general auditory mechanisms are responsible for processing particular acoustic features (McGettigan & Scott, 2012). Acoustic (cue dependent) hypotheses of speech perception, such as the Asymmetric Sampling in Time hypothesis (Poeppel, 2003), propose that in humans, auditory processing areas in the right hemisphere operate at a lower temporal resolution to those of the left hemisphere, resulting in a greater preference for processing slow acoustic modulation including the supra-segmental cues in speech, whereas the left hemisphere is more specialised in analysing rapidly changing auditory information such as phonemic cues (Zatorre & Belin, 2001; Poeppel, 2003). Support for the cue

dependent interpretation has been obtained through demonstrations that the right hemisphere displays stronger activation in response to sounds with longer durations and pitch modulation (Belin et al., 1998; Boemio, Fromm, Braun & Poeppel, 2005; Schönwiesner, Rübsamen & Von Cramon, 2005; Zatorre & Belin, 2001), although there is currently only limited evidence to suggest that the left hemisphere is more specialised in processing sounds with fast temporal modulation (see McGettigan & Scott, 2012 for a recent review). Applying this interpretation to our observations would suggest that the right hemisphere may also operate at a lower temporal resolution than the left hemisphere in dogs, preferentially processing the supra-segmental components of the speech signal, whilst the left hemisphere is more responsive to the fast temporal modulation which characterises the phonemic content.

However, the dogs in our study expressed opposite response biases to speech signals with equivalent spectrotemporal complexity if the phonemic content differed in functional relevance (Meaningful and Meaningless (foreign) Speech with Neutralised Intonation; Tests 1 and 3), suggesting that the left hemispheric bias in the dogs' responses to meaningful phonemic cues was not purely dependent on the increased salience of the rapidly modulated components of the signal, but also on the meaningfulness of these cues. Our results therefore appear more consistent with the functional interpretation of lateralisation, which proposes that hemispheric specialisation is dependent on the communicatory function and value of the acoustic content rather than the basic acoustic structure (e.g. McGettigan et al., 2012; Narain, et al., 2003; Rosen, Wise, Chadha, Conway & Scott, 2011). Indeed, the observation that the left hemisphere is preferentially recruited when dogs process the phonemic cues of the highly familiar and learnt command "come on then" is consistent with reports that the left hemisphere tends to respond to familiar or learnt patterns across mammals (Vallortigara et al., 2008). To further clarify whether the observed biases were related to the acoustic structure of speech components or to their functional communicative content, we conducted a second experiment testing an additional five sound conditions with new subject dogs.

## Experiment 2: Do Dogs' Orienting Responses to Human Speech Relate to the Communicative Content or the Acoustic Structure of the Signal?

Firstly, to determine whether the left-head turn bias observed in response to supra-segmental cues could be explained by a general preference for slow acoustic modulation, we presented dogs with a sine-wave tone matching the intonation contour of the original command (Test 6: Sine-Wave Intonation). While it was unclear from Experiment 1 whether the right hemisphere preferentially responded to the acoustic composition of the manipulated signals independently from their content, the results suggested that dogs' responses to the phonemic content of speech signals depended on their communicative function rather than the acoustic structure. However it was unclear whether the obtained biases were generated due to the familiarity of the command (which could be related to the familiarity of the speakers' accents and/or familiarity with the phonemes independently of their meaning) or whether this bias was dependent on the learned functional relevance of the phonemic content in the signal. To investigate the importance of the familiarity of the signals, dogs were presented with the original command with degraded prosodic cues, but spoken by a non-native British speaker in a strong, unfamiliar accent (Test 7: Meaningful Speech in an Unfamiliar Accent with Neutralized Intonation). Based on the significant left hemispheric response bias obtained in the meaningful sine-wave speech condition, in which the speaker-related cues were degraded, we predicted that reducing the familiarity of the speaker's accent would not influence responses. We also assessed whether right-head turn biases were dependent on the presence of meaningful phonemic cues, or merely familiar phonemic cues, by presenting dogs with a pseudo-word phrase using the same phonemes as the original command ("thon om ken"; Meaningless Phonemes with Neutralised Intonation; Test 8). In this condition, both the phonemes and the speakers' accent were familiar, but the phrase itself was meaningless.

In the final two conditions, we investigated whether dogs' orienting biases would extend to non-vocal signals. Firstly, to determine if the left hemispheric response bias to meaningful phonemic cues would generalise to non-vocal stimuli with learnt communicative value, we exposed dogs to a Meaningful Human Whistle (Test 9). It was also possible that, because the stimuli used in all of the conditions eliciting a right hemispheric response bias in Experiment 1 were resynthesised, the perceived novelty of these sounds could have generated stronger right hemispheric activation (Vallortigara et

al., 2008). To test this, dogs were exposed to a novel artificial control sound (Test 10: Pink Noise).

## Method

### Subjects

None of the subjects had previously participated in Experiment 1, to ensure that they were naïve to the experimental procedure. All of the subject animals were healthy, over 6 months old, with no known history of aggression towards people. Owners of dogs exposed to the speech conditions were native British speakers and confirmed that their dog responded to the command 'come on then' or a similar variant, with no previous exposure to any other languages. We included the first 25 dogs that reacted to the stimuli in each condition. A total of sixteen subjects failed to react to the stimuli (with an even distribution of failed responses across conditions ($\chi^2_{(4)} = 6.50$, $p = 0.17$)), and were excluded from the study at the time of testing. The 125 dogs retained in the analysis included 51 females and 74 males from 41 different breeds. Ages ranged from six months to 14 years old (mean ± SD = 4.14 ± 2.86 years). One hundred and sixteen dogs were privately owned pets, whilst nine dogs were housed in a local animal shelter.

### Stimuli Acquisition

*Voice Recordings*

Four men and four women aged between 20 and 58 years old (mean ± SD = 30.25 ± 13.68 years) were recorded producing four different whistles commonly used by dog owners. Whistles were produced with a mean frequency of 1741.01 Hz ± 488.73 Hz and a frequency range of 825.73 Hz ± 577.69 Hz. These recordings made up the stimuli in Test 9 (Meaningful Whistle).

A further eight native British speakers, four men and four women aged between 23 and 33 years old (mean ± SD = 27.65 ± 3.62 years), were recorded pronouncing the pseudo-word phrase 'thon om ken' with a happy intonation (used to create Meaningless Phonemes with Neutralised Intonation; Test 8).

Additionally, eight native French speakers, four men and four women aged between 24 and 51 (mean ± SD = 38.14 ± 10.82 years) were recorded pronouncing the phrase 'come

on then' in a happy tone of voice with a strong French accent (used to create Meaningful Speech in an Unfamiliar Accent with Neutralised Intonation; Test 7).

All of the recordings were made using a Zoom H4N Handy Recorder in a sound proof booth (sampling frequency: 44 100 Hz; sampling rate: 32-bit). Each recording was normalised to -1.0 dB maximum amplitude in Audacity 2.0.0.

*Acoustic Manipulations*

The same procedure for neutralising the intonation described in Experiment 1 was applied to all of the recordings, providing the stimuli for Tests 7 and 8.

Creating a sine-wave tone matching the positive intonation contour (Test 6)

This manipulation was carried out only on the English recordings of 'come on then' from Experiment 1, and aimed to create a sine-wave tone based on the pitch contour (Sine-Wave Intonation). The pitch contour was first extracted, using the 'To Pitch' command in PRAAT, and then converted into a sine-wave sound using the 'To Sound (sine)' command. The signal was then set to fade in at the onset and out to the offset using Audacity.

Creating pink noise (Test 10)

The final stimulus type was Gaussian pink noise with a duration of 1 s, set to fade in at the onset and out to the offset (Pink Noise). This stimulus was created in Audacity using the 'Generate Noise…' command.

*Perceptual Ratings*

The validity of the intended manipulations was also verified by the five volunteers in the same listening experiment used to check the stimuli for Experiment 1. Each sound was again scored on both 5-point rating scales, where 1 represented 'very unclear' and 'very negative' on the intelligibility and emotional valence scales respectively (Table 2).

**Table 2.** Mean ratings of emotional content and intelligibility for the stimuli used in each auditory condition.

| Auditory Condition | Mean Score for Emotional Content (3 = neutral) | Mean Score for Intelligibility (3 = medium clarity) |
|---|---|---|
| Sine-wave Intonation | 3.19 | 1.00 |
| Meaningful Speech in an Unfamiliar Accent with Neutralised Intonation | 2.94 | 3.13 |
| Meaningless Phonemes with Neutralised Intonation | 3.04 | 1.00 |
| Meaningful Whistle | 3.25 | 1.00 |
| Pink Noise | 3.00 | 1.00 |

**Experimental Set-up and Procedure**

Experiment 2 was carried out and analysed identically to Experiment 1 (see Methods section for details). Testing occurred between May 2013 and September 2014.

## Results and Discussion

A binary logistic regression analysis identified a significant overall effect of auditory condition on head-turn direction (Wald(4) = 9.65, p = 0.02), indicating that the orienting responses differed between the experimental conditions (Figure 2). There was no significant effect of subject sex (p = 0.73), age (p = 0.33), breed type (p = 0.70), current residence (animal shelter or private home) (p = 0.79), stimulus exemplar (p = 0.32), stimulus voice gender (p = 0.19) or test location (p = 0.35) on responses. Binomial tests were then used to investigate the presence of head-turn biases for each sound condition.

Dogs showed no significant orienting bias in response to sine-wave tones matching the intonation patterns of the original commands (Test 6: Sine-wave Intonation) (binomial test: 56% right head turn, p = .69), signifying that the observed left-head turn bias for supra-segmental cues in speech demonstrated in Experiment 1 does not generalise to

slow frequency modulation across all acoustic signals. This observation provides further support for the hypothesis that dogs' hemispheric asymmetries in response to speech depend on the communicative content of the signals rather than their basic acoustic structure. Our results are in agreement with the previous demonstration that auditory regions in the right hemisphere of the dog's brain show valence-dependent activation in response to human non-verbal vocalisations, which could suggest a functional basis to this specialisation (Andics et al., 2014). In humans, there is clearer evidence indicating that right hemispheric activation is related to the content of the signals rather than the acoustic composition, as the right hemisphere preferentially responds to non-speech vocal sounds but not to their frequency scrambled counterparts (Belin, Zatorre & Ahad, 2002). Further studies are necessary determine if stronger right hemispheric activation also occurs in dogs during the perception of voice-like stimuli only when it contains supra-segmental content of communicative value.

The pattern of results obtained across Tests 1-6 is therefore most consistent with the hypothesis that in mammals, the left hemisphere of the brain is specialised in categorising information and selectively responds to familiar and learnt patterns, whilst the right hemisphere responds preferentially to emotionally-related cues (Andrew & Rogers, 2002; MacNeilage, Rogers & Vallortigara, 2009; Vallortigara & Rogers, 2005). Additionally, similarly to humans (Lattner et al., 2005; von Kriegstein et al., 2003; Belin & Zatorre, 2003) and rhesus macaques (Petkov et al., 2008), our observations indicate that the right hemisphere of the dog brain may also be specialised in processing speaker-related supra-segmental cues such as gender and identity.

More specifically, our results suggest that the dogs' left hemispheric bias to salient phonemic cues in the familiar command was dependent on the learnt relevance of these cues rather than their familiarity, as dogs exposed to the original command with degraded emotional prosodic cues, but spoken by a non-native British speaker in a strong accent (Test 7: Meaningful Speech in an Unfamiliar Accent with Neutralised Intonation), still showed a significant right head-turn bias (Binomial test: 72% right head turn, $p = .04$), demonstrating that the left hemispheric response bias observed in response to the neutral familiar command was not dependent on the familiarity of the speaker's accent. Furthermore, in Test 8 dogs showed the opposite orienting bias in response to the meaningless phonemes with neutralised intonation (binomial test: 20% right head turn, $p = 0.004$), which confirms that increasing the salience of segmental

phonemic content in speech only generates a left hemispheric response bias in dogs if it is functionally meaningful – i.e. if it is known to trigger a specific learned response from the animal. This is in agreement with speech perception in humans, as only intelligible speech has been found to generate a left hemispheric processing bias (McGettigan et al., 2012). Whilst previous findings have suggested that dogs do recognise the phonemic content in learnt spoken commands, supra-segmental cues have also been available to the subjects, which could have been used to recognise the commands instead (Fukuzawa et al., 2005). Therefore our observations provide the first clear demonstration that the phonemic content in human speech is naturally meaningful to dogs, and that they have some understanding of the combinatorial structure of spoken words. The subjects' differential responses to learnt versus unfamiliar phonemic cues further supports the conclusion that in dogs, the left hemisphere preferentially responds to phonemic content with meaningful communicative value, whereas voice or speech-like stimuli lacking this information generate right hemispheric biases.

Interestingly, no significant orienting bias was obtained when we presented dogs with a Meaningful Whistle in Test 9 (binomial test: 60% right head turn, $p = 0.42$), suggesting that the left hemispheric advantage for meaningful phonemic content in speech may not extend to other familiar or communicatively relevant non-vocal sounds. While this result may seem in opposition with the left hemispheric advantage that characterises the perception of articulated whistled language by experienced human listeners (Carreiras, Lopez, Rivero & Corina, 2005), such languages encode phonological segmental information (Meyer, 2008) and are therefore more comparable to the meaningful sine-wave speech used in Experiment 1, which also triggered a left hemispheric bias. In contrast, the simple command whistles used in our study did not contain segmental information (they did not result from the combination of phonological units) and were more comparable to the intonation contours used in Test 6, which also failed to trigger a bias. It therefore appears that auditory signals must not only be meaningful, but also voiced, to elicit stronger left hemispheric activation in dogs.

Finally, dogs also showed no orienting biases in response to the unfamiliar control sound (Pink Noise; Test 10), which contained neither segmental nor supra-segmental frequency modulation (binomial test: 48% right head turn, $p = 1.00$). This result confirms that the orienting asymmetries observed across the conditions did not arise

from the perceived novelty or intrinsic unnaturalness which is associated with re-synthesised stimuli (Figure 3).

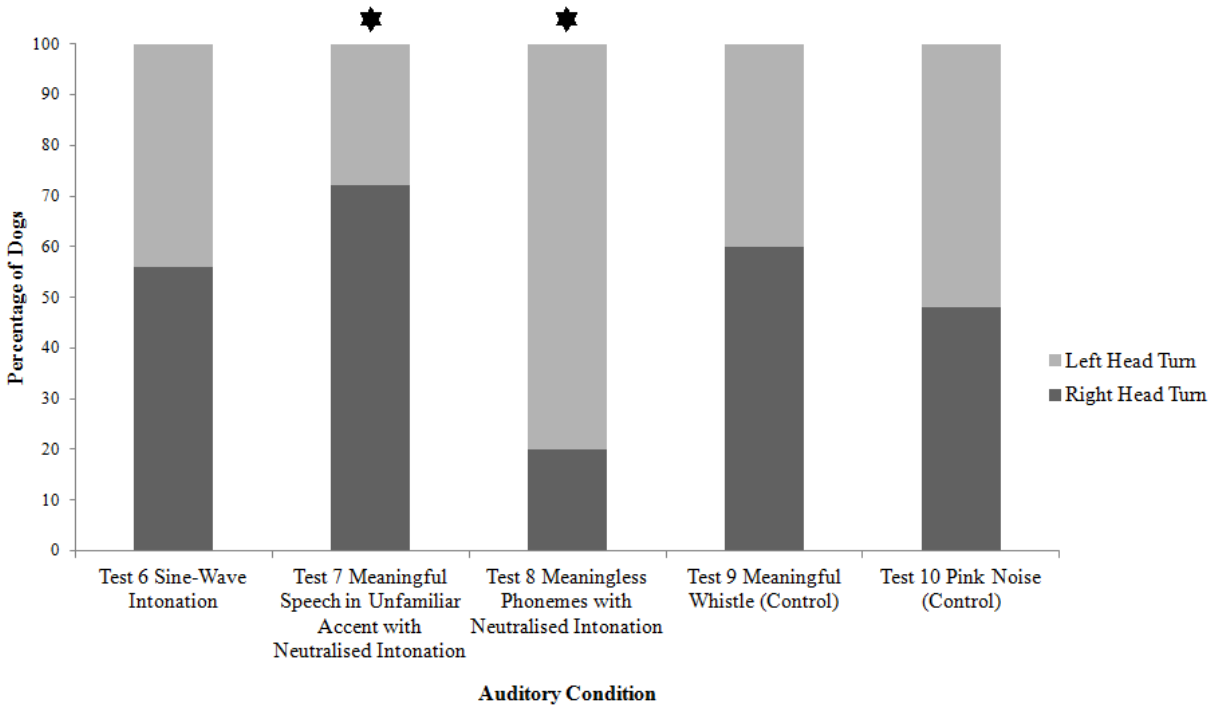



**Fig 3.** Percentage of dogs that oriented to their left or right in each condition after the playback presentation. Asterisks indicate conditions in which the proportions were significantly different from chance (50%) at $p < 0.05$.

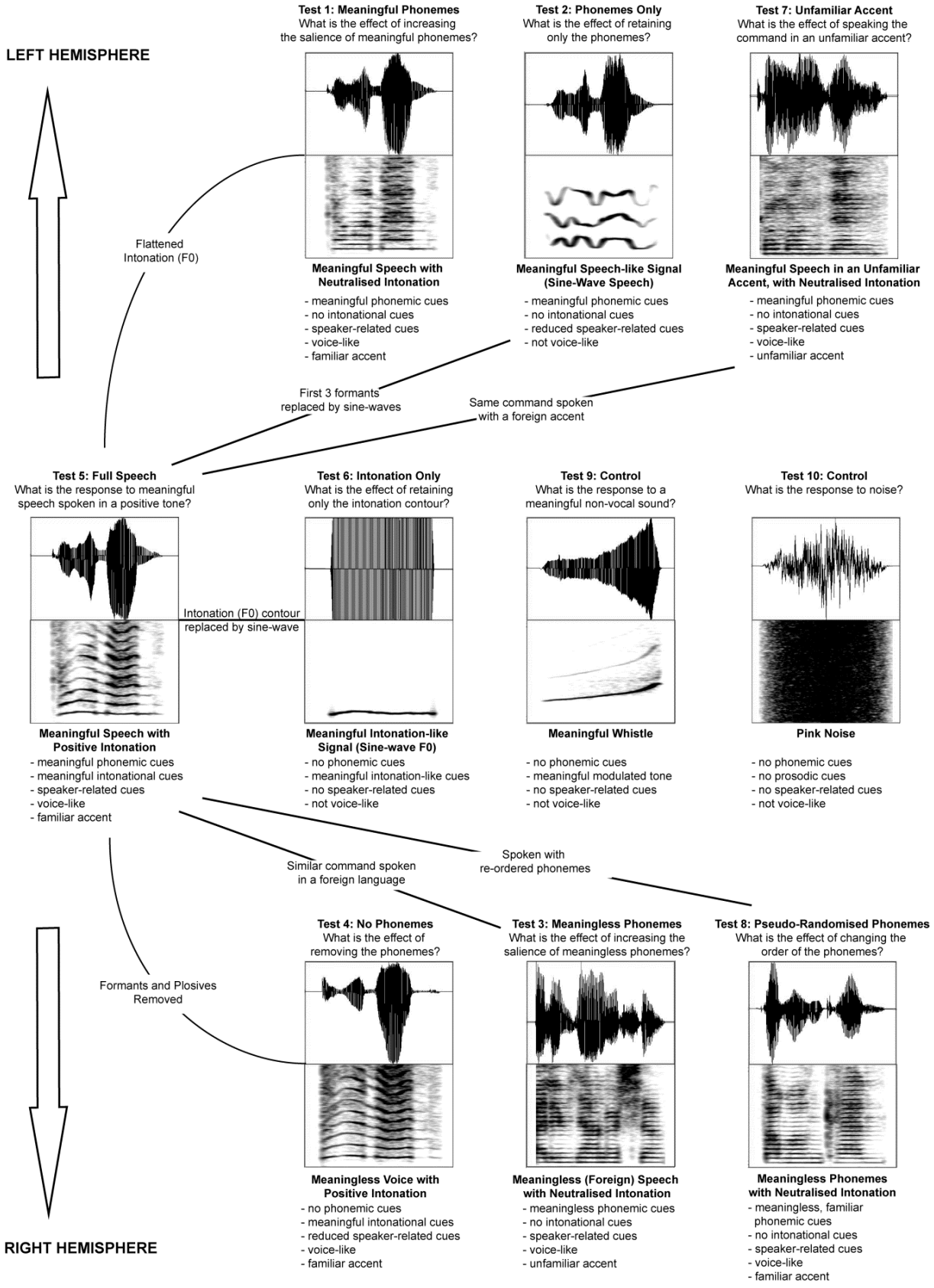



**Fig 4.** Example spectrograms and brief descriptions of each of the auditory conditions used in Experiments 1 and 2, organised by hemispheric response bias.

## General Discussion

Our study demonstrates that dogs show a significant right-sided orienting bias in response to meaningful segmental phonemic information in speech, suggesting that these cues are preferentially processed by the left hemisphere of the dog's brain. In contrast, human voices lacking this information (increasing the salience of prosodic and/or speaker-related cues) produce significant left head-turn biases, indicating stronger right hemispheric activation. It therefore appears that while dogs perceive each of the main communicative components of human speech signals, meaningful segmental cues generate opposite response biases to supra-segmental information, implying that these components are distinguished and processed separately. Furthermore, dogs differentiate at least some speech cues according to their functional relevance, as their orienting biases depended on whether salient phonemic cues were meaningful to them. This indicates that the observed processing divisions cannot be explained by simple differences in the acoustic characteristics of the individual components of speech, but that they are also related to the functional content. Although dogs' perception of specific speech cues remains largely unexplored, they are known to attend to the emotional prosody (Mills, Fukuzawa & Cooper, 2005; Scheider et al., 2011) as well as discriminating gender (Ratcliffe et al., 2014) and identity-related information in the human voice (Adachi et al., 2007). Because dogs showed the opposite head-turning bias when they heard a familiar command if the phonemes were re-ordered, our results suggest that they also learn the combinatorial structure of the verbal content in familiar spoken phrases. This provides the first definitive demonstration that in addition to perceiving relevant information encoded in the supra-segmental components of human speech, dogs also naturally perceive phonemic cues in a functionally relevant manner, consolidating previous evidence that dogs fail to respond appropriately to learnt commands when individual phonemes are substituted (Fukuzawa et al., 2005).

Taken together, our results reveal a striking parallel between the hemispheric biases indicated by the dogs' responses and those reported in humans, suggesting that both species may dissociate and process the communicatory components of speech in a broadly comparable way. Although further research is necessary to determine the precise brain mechanisms involved, the strong correspondence between dog and human hemispheric asymmetries for speech processing may reflect convergent evolution if

dogs have been selected to respond to human vocal signals during domestication. Indeed, although the precise effect of domestication on dogs' socio-cognitive abilities remains unclear (see Kaminski & Nitzschner, 2013, for a recent review), they do exhibit enhanced skills in responding to human visual gestures (e.g. Kaminski, Schulz & Tomasello, 2012; Scheider et al., 2011), outperforming equivalently socialised wolves *Canis lupus* in some tasks (Gásci et al., 2009; Hare, Brown, Williamson & Tomasello, 2002). It is possible that dogs' responsiveness to human vocal signals has similarly improved to facilitate inter-specific communication with humans, leading to the evolutionary development of comparable hemispheric activation during speech processing.

Alternatively, the similarities observed between dog and human responses to speech may be indicative of shared hemispheric specialisations that are present across phylogenetically distant mammal species, and expressed when the individual is exposed to functionally meaningful speech signals. A wide range of mammalian species, including dogs, show stronger left hemispheric activation when attending to conspecific vocalisations (Ocklenburg et al., 2013). The specialisation of the left hemisphere for processing species-specific calls appears to be similarly dependent on the functional relevance of the signals rather than spectrally-related acoustic cues, as left hemispheric biases no longer occur in response to conspecific calls if they are modified to fall outside of the species-typical range (Hauser, Agnetta & Perez, 1998; Siniscalchi et al., 2012). Juveniles also commonly fail to show equivalent hemispheric asymmetries to adults in response to conspecific calls (California sea lions: Böye et al., 2005; Rhesus macaques: Hauser & Andersson 1994), which is consistent with evidence that individuals must learn to discriminate between conspecific and hetero-specific vocal signals, and develop specific responses to different call types (e.g. Mateo, 1996). Therefore, the left hemisphere may preferentially respond to species-specific acoustic content in vocal signals that has learnt relevance to the individual, which extends to include the phonemic cues in human commands in dogs due to the equivalent relevance of these hetero-specific vocal signals.

In contrast to the apparent left hemispheric preference for learnt cues to discriminate species-specific vocal signals from other sounds, emotionally salient human and animal vocalisations are processed primarily in the right hemisphere of the human brain, independently of the listener's familiarity with the species or their ability to identify the

emotional valence of the calls (Belin et al., 2008). Dogs also show a right hemispheric specialisation for processing affective cues in both conspecific and human vocalisations (Andics et al., 2014), suggesting that they may perceive the emotional content in human vocal signals similarly to conspecific vocalisations. Equivalent acoustic encoding of emotionally-related information has been observed across a range of mammal vocal signals (Briefer, 2012), and it has recently been determined that humans use the same cues to judge emotional valence and intensity in human and dog vocalisations (Faragó et al., 2014). It is therefore possible that the same processing mechanism, associated with the right hemisphere, is involved in the assessment of the emotional content of vocalisations independently of the caller's species, at least across mammals.

Although the neurological mechanisms involved in processing the caller-related characteristics of vocal signals remain largely unknown in mammals, rhesus macaques have been found to show a right hemispheric bias when discriminating the familiarity of conspecific callers (Petkov et al., 2008), suggesting that a right hemispheric specialisation during the perception of indexical supra-segmental cues may also be shared across a wider range of mammals. Therefore, similarities in the functional content of vocalisations may enable mammals to process relevant hetero-specific vocalisations equivalently to conspecific calls, implying that the hemispheric biases generated in humans and dogs in response to speech may be conservative across mammalian species. However, it has been proposed that both evolutionary and epigenetic processes generate a socio-cognitive predisposition specifically in dogs to show a greater acceptance of humans as social partners, increasing dogs' latent potential to perceive human signals similarly to conspecific signals (Gácsi et al., 2009; Hare & Tomasello, 2005; Hare, Wobber & Wrangham, 2012; Range & Viranyi, 2015; Udell, Dorey & Wynne, 2010). This would suggest that although mammals may generally share the same hemispheric organisation in response to conspecific vocalisations, dogs may have adapted to process the communicative content of hetero-specific human vocalisations equivalently to conspecific signals. To test these hypotheses more directly, further experiments could replicate our study with other domesticated (e.g. horses) versus non-domesticated (e.g. captive wolves) species that are regularly exposed to human speech. If comparable hemispheric biases are observed in non-domesticated species, this would indicate that the evolutionary origins of lateralisation during speech perception could have developed earlier in our evolutionary history than previously

thought. Alternatively, domestication may have resulted adaptations affecting how dogs process speech signals which are not shared with other species, facilitating their inter-specific communication with humans.

# CHAPTER 6: GENERAL DISCUSSION

## Introduction

A concerted body of research has determined that domestic dogs *Canis familiaris* are highly adept at inter-specific communication with humans, to the extent that it has been hypothesised that selection pressures during domestication may have (directly or indirectly) enhanced their ability to perceive human signals (e.g. Gácsi et al., 2009; Hare & Tomasello, 2005; Kaminski & Nitzschner, 2013; Miklósi, 2009; Range & Virányi, 2014; Udell, Dorey & Wynne, 2010). However, evidence supporting dogs' advanced abilities to interpret human signals has predominantly derived from their responses to visual gestures, with comparatively little focus on their interpretation of vocalisations, thus providing only a partial perspective on how they perceive human signals. To address this issue, the research comprising this thesis aimed to further our current knowledge of how dogs interpret human signals by investigating their perception of information encoded in the human voice. While the results of previous studies have suggested that dogs discriminate dynamic supra-segmental information related to the emotional prosody of human speech (Marshall-Pescini, Prato-Previde & Valsecchi, 2011; Mills, Fukuzawa & Cooper, 2005; Scheider, Grassmann, Kaminski & Tomasello, 2011) and vocalisations (Andics, Gácsi, Faragó, Kis & Miklósi, 2014; Yong & Ruffman, 2014), prior to the current studies there had only been limited indications that dogs can perceive relevant indexical cues related to the physical characteristics of the speaker (as evidenced by the recognition of their owner's voice; Adachi, Kuwahata & Fujita, 2007), or the segmental phonemic content in human speech (Fukuzawa, Mills & Cooper, 2005). Therefore, the present research was designed to clarify whether dogs are able to perceive relevant information encoded within the segmental phonemic or supra-segmental indexical content of speech, or if their perception is predominantly limited to variation in the emotional prosody. In addition, this body of work also aimed to explore the perceptual mechanisms involved when dogs associate cues encoded in human voices with corresponding visual information, and the extent to which such mechanisms enable dogs to match unfamiliar human voices with different speakers. The final general aim of the thesis was to determine whether similarly to humans, dogs show evidence of dissociating different functional components of speech during processing. In the following sections, I summarise the main empirical results and conclusions of each article in relation to the key research questions outlined in the introduction, discuss

the theoretical implications of these observations and provide recommendations for future research.

**Are Dogs Spontaneously Capable of Cross-Modal Human Gender Discrimination?**

Although it had previously been established that dogs recognise highly familiar human voices (Adachi et al., 2007), suggesting that they are able to perceive speaker-related indexical cues encoding identity in speech, it was possible that their recognition could have instead been related to the perception of the intonational or phonemic cues characterising a highly familiar phrase rather than the identity of the voice (Kriengwatana, Escudero & Cate, 2014). Therefore, to further clarify whether dogs do perceive indexical information encoded in the supra-segmental content of human speech, Chapter 3 investigated if dogs could express the ability to discriminate the gender of adult human voices by visually attending to a man when they heard an adult male voice and a woman when they heard an adult female voice. Using a preferential looking paradigm, we demonstrated that family owned dogs are spontaneously capable of cross-modal human gender discrimination, although the expression of this ability was dependent on the dog's degree of socialisation with people. Specifically, when faced with an unfamiliar man and woman, highly socialised dogs looked more at the person matching the gender of the voice they heard, while less socialised dogs appeared to avoid looking at the matching person, by instead looking more at the person that did not match the voice gender. Despite showing these opposite response patterns, all of the dogs expressed their ability to associate human voices with people according to their gender.

The observation that dogs are able to discriminate between unfamiliar male and female human voices provides the first clear demonstration that dogs spontaneously perceive speaker-related indexical cues in human speech without any prior exposure to the person's voice. Furthermore, because dogs are able to discriminate between individual human voices of the same gender on some level (Adachi et al., 2007), their differentiation of male and female voices represents a natural ability to categorise human voices according to their gender. Although individual animals from a wide range of mammal and bird species have been successfully trained to learn human given categories (see Jitsumori & Delius, 2001 for a review), our results add to the currently small number of studies demonstrating spontaneous, ecologically relevant category

formation in non-human animals (e.g. Murai et al., 2005; Murai, Tomonaga, Kamegai, Terazawa, & Yamaguchi, 2004), and to our knowledge give the first indication of spontaneous category formation in a non-primate mammal. Furthermore, the human gender categories formed by dogs are not limited to a generalisation of perceptual similarities between voices of the same gender, as the subjects were able to associate the voices they heard with the corresponding person, demonstrating that they had instead formed multisensory human gender categories encoding distinct perceptual cues.

However, it is difficult to establish the precise nature of dogs' human gender categories due to the limitations of the preferential looking paradigm used in our study. Because the visual stimuli (the man and woman) were simultaneously presented, it is not possible to determine whether the subjects matched the corresponding cross-modal cues according to the person's gender merely because they had learnt the greater likelihood of their co-occurrence, or if they actually formed a functional association between the cues. The results of previous studies suggest that dogs do in fact assess human gender in a functionally relevant manner, as they behave differently towards unknown men and women, particularly showing a general tendency to perceive men as more threatening (Hennessy, Williams, Miller, Douglas & Voith, 1998; Lore & Eisenberg, 1986; Rosado, García-Belenguer, León, & Palacio, 2009; Wells & Hepper, 1999). The need for dogs to differentiate people according to their gender could be explained by gender-specific human interaction styles (Mariti et al., 2012; Prato-Previde, Fallani & Valsecchi, 2005), promoting associations that enable dogs to tailor their responses to men and women separately. To further establish the nature of the cognitive representations formed by dogs relating to human gender, it would be beneficial to build on our present observations by testing dogs' responses when they are not able to simultaneously compare a man and a woman. For example, in the violation of expectation paradigm, animals are presented with a sequence of congruent or incongruent cues, and are expected to pay greater attention to the incongruent sequence. Similarly to Adachi et al.'s (2007) investigation of human identity discrimination in dogs using violation of expectation paradigm, dogs would first be presented with the voice of a man or woman, after which the face of either a man or woman would appear. If dogs possess a multisensory cognitive representation of adult human males, this stored representation should be activated by the presentation of the male voice, generating a stronger level of

attention if the greater expectation of a corresponding male face is violated by the subsequent presentation of a female face instead.

While dogs are likely to benefit from the ability to classify humans using multisensory cues, and therefore combine cross-modal human gender cues in a functionally relevant manner, it is not clear whether they have undergone any specific adaptations during domestication to facilitate human gender categorisation, or if their responses purely relate to extensive ontogenetic exposure to men and women. Although we found no experience-dependent effects on our subjects' abilities to match the voices to the correct humans, it is probable that even the subjects living with a single person had some degree of prior exposure to both genders. Therefore, we cannot rule out the potential effects of individual experience on the dogs' responses. Replicating our procedure with puppies may provide clearer evidence of how dogs acquire the ability to discriminate human gender, while comparisons with individuals from non-domesticated species that have been equivalently socialised with people (e.g. human-socialised grey wolves *Canis lupus*) would help to determine whether dogs have undergone specific cognitive adaptations to enable them to form multisensory categories about human groups.

As well as further exploring the evolutionary origins human gender categorisation in dogs, in order to fully understand this ability it is also important to establish the specific cues that dogs associated across their different sensory modalities. Because the aim of Chapter 3 was to determine if dogs are capable of spontaneously matching voices to men and women, the subjects were provided with the full range of gender cues across three sensory modalities (acoustic, visual and olfactory). Although it was necessary to provide the subjects with a sufficient range of information to ensure that they could express their ability to make the association between the voices and the people, because of this provision it was not possible to determine the specific cues that the dogs matched together. Even though dogs are known to be perceptually aware of the F0 and formant positioning in human voices (Baru, 1975), which provide the strongest anatomically-related cues to the voice gender, women are also more likely to use exaggerated intonation patterns than men when speaking to dogs (Prato-Previde et al., 2005), which was also apparent in the voice recordings used as stimuli in Chapter 3. This could have enabled dogs to discriminate the voice gender differently to human listeners (e.g. Smith & Patterson, 2005), by using the intonation pattern instead of anatomically derived vocal cues. Similarly, regarding the visual cues available to the subjects, although the

difference in height between the male and female assistants was not found to influence the dogs' responses, the male and female heights were bimodally distributed, potentially allowing the subjects to match the voices to the people according to broad body-size differences rather than gender-specific differences. Finally, the subjects may have also used olfactory cues to associate the voices with the corresponding person, due to sex specific differences in human body odour (Penn et al., 2007). These potential cues illustrate that fact that there are a number of different ways that dogs may have cross-modally discriminated human gender which do not necessarily parallel the perceptual mechanisms used by humans. By reducing and controlling the wide range of sensory information available in our experiment, future studies could establish the precise cues that dogs use to cross-modally discriminate human gender.

Chapter 3 demonstrated that dogs categorise human gender by combining auditory cues with different sensory information, enabling them to match unfamiliar human voices to individuals of the corresponding gender. However, while it is clear that dogs express this ability, additional questions remained concerning the perceptual and associative mechanisms that enable dogs to combine multisensory human indexical cues. Therefore, the aim of Chapter 4 was to explore the range of cross-modal associations that mammals are known to naturally form during communication, before investigating the most likely correspondences that influence dogs during their association of multisensory indexical cues about human signallers.

## How do Non-Human Animals form Cross-Modal Associations during their Perception of Multisensory Signals?

Article I in Chapter 4 provided a broad review of the range of perceptual and cognitive mechanisms that are known to be expressed by mammals as they combine different sensory information in communicative signals. In a direct comparison with cross-modal perceptual mechanisms that have been identified in humans, we first outlined evidence that other mammals also naturally perceive multisensory signals by matching generic low-level cues across different sensory channels, such as temporal synchronisation, from a relatively early stage in their development (e.g. Zangenehpour, Ghazanfar, Lewkowicz & Zatorre, 2009). Furthermore, it was apparent that a range of highly social mammalian species also store complex cognitive representations about specific group members, encoding multisensory indexical characteristics of well known individuals

(e.g. Adachi & Hampton, 2011; Kojima, Izumi & Ceugniet, 2003; Proops, McComb & Reby, 2009). These multisensory representations enable animal receivers to match vocalisations with familiar signallers without the need for temporal synchronisation between sensory cues (Proops et al., 2009), and can be flexible enough to allow some species, including dogs (Adachi et al., 2007), to combine multisensory information about familiar humans as well as conspecifics (Adachi & Fujita, 2007; Proops & McComb, 2012; Sliwa, Duhamel, Pascalis & Wirth, 2011). Although associations between multisensory indexical cues related to the same individual may be learnt through repeated exposure to their temporal and spatial co-occurrence, such prior synchronisation does not always seem to be necessary for animals to successfully match different signals produced by the same individual (e.g. Kulahci, Drea, Rubenstein & Ghazanfar, 2014), indicating higher order cognitive mechanisms may also be involved in this process.

In addition to forming multisensory representations about familiar individuals, primates that have had sufficient exposure to heterospecifics also appear to learn to categorise different species across sensory modalities (Adachi, Kuwahata, Fujita, Tomonaga & Matsuzawa, 2009), and it is possible that our demonstration of dogs' ability to match multisensory human gender cues in Chapter 3 exemplifies a similarly learnt form of cognitive categorisation. However, the perception of relatedness or equivalence between different sensory cues encoding information about the same feature is also evident in more general correspondences expressed by mammals. Such broad associations include the ability to match the number of vocalisations to the corresponding number of signallers (Jordan, Brannon, Logothetis & Ghazanfar, 2005), and associate vocalisations with conspecifics by assessing multisensory cues related to their body size (Ghazanfar et al., 2007; Taylor, Reby & McComb, 2011). Similarly, evidence that primates cross-modally associate information about specific call types (Parr, 2004) indicates that animals may perceive correspondences between dynamic multisensory cues to motivation or emotion. It is possible that the perception of these general correlations could either account for, or facilitate the acquisition of, more specific categories that animals appear to form which enable them to match signals with the appropriate signaller. For example, vocalisations produced by the same individual, or by the same class of individuals, also often correspond more broadly by encoding similar size-related cues. Therefore, in relation to our observations in Chapter 3, it is possible that

rather than specifically learning to associate human gender cues, dogs may match human voices to people of the same gender by attending to the lower-level mapping between body size cues.

Humans also perceive a further class of cross-modal correspondences between low-level stimulus features, some of which are thought to develop as people learn statistically prevalent correlations between different sensory dimensions in the environment (see Spence, 2011 for a review). The learnt probability that different physical entities will produce certain sound frequencies has been suggested to account for the automatic tendencies that humans have to map lower frequency sounds to larger shapes (e.g. Gallace & Spence, 2006) as well as lower spatial elevations (e.g. Rusconi, Kwan, Giordano, Umilta & Butterworth, 2006). There is evidence to suggest that these statistical correspondences are used in human language to facilitate associations between words and their referents (e.g. Evans & Treisman, 2010; Imai, Kita, Nagumo & Okada, 2008; Köhler, 1947; Parise & Pavani, 2011; Sapir, 1929), in an effect known as 'sound symbolism'. Furthermore, Pisanski (2014) recently demonstrated that the perception of these low-level correspondences can influence human listeners' assessments of speaker size, as in accordance with the perceived mapping between low frequency sounds and low visual elevations, participants actually judged adult male human voices to belong to larger individuals when they were projected from a speaker placed at a lower spatial elevation. This raises the question of whether broad statistical correspondences between particular sound frequencies and visual cues influence the way that audio-visual communicative cues are combined in both humans and other species. However, Article I in Chapter 4 highlighted the fact that the potential formation and influence of statistical correspondences on the spontaneous associations made by non-human animals is largely unexplored, with very little known about whether non-human animals actually perceive comparable low-level mappings to humans, although initial indications have suggested that some may do so (Ludwig, Adachi & Matsuzawa, 2011). The perception of comparable broad cross-modal correspondences between low-level features by non-human animals represents a substantial area for future research in order to establish the potential influence of such associations in both communicative and non-communicative contexts.

Therefore, it was apparent from Article I in Chapter 4 that there are a number of perceptual mechanisms that mammals could apply to match signals with the

corresponding signallers, ranging from the perception of generic low-level redundancies between signals to storing complex cognitive representations of signaller categories. The majority of the current evidence concerning mammals' abilities to combine audio-visual information is related to the indexical characteristics of the signaller, although it is still unclear which mechanisms are naturally used by non-human animals to acquire and express these functionally relevant associations. Article II in Chapter 4 thus aimed to determine which, if any, of the main perceptual mechanisms identified in our review are used by dogs to match human voices to speakers according to their indexical characteristics.

## Are Dogs Spontaneously Capable of the Cross-Modal Discrimination of Human Age Categories, and if so, How do they Associate Age-Related Auditory and Visual Cues?

Coupled with prior evidence that dogs may differentiate between highly familiar and unfamiliar human speakers (Adachi et al., 2007), our demonstration of cross-modal human gender discrimination by dogs in Chapter 3 confirmed that they spontaneously perceive indexical features of the human voice in a functionally relevant manner. A further key indexical attribute encoded in the human voice is the age category that the speaker belongs to (as an adult or child). In Article II of Chapter 4, we investigated whether dogs are also spontaneously capable of cross-modally discriminating human age categories by combining audio-visual information from unfamiliar speakers. Additionally, based on the theoretical conclusions of Article I in Chapter 4, we explored some of the potential mechanisms that might influence how dogs match human voices to speakers according to their age category.

We first determined that dogs were successfully able to associate an adult male human body shape with a correspondingly aged male voice, which directly expanded on the observations of Chapter 3 by demonstrating that dogs can match adult male voices to the appropriate speaker by associating single vowel sounds (which lack any prosodic cues) with their visual body shape (without any additional visual or olfactory cues). While dogs appear to be capable of making this discrimination using only a small number of distinctive features, their multisensory perception of adult male humans does not appear to be dependent on general associations between more general low-level visual cues. Specifically, the subjects did not associate simple shapes with the voices of

adult men according to their relative compatibility in size, suggesting that in contrast to theories of animal vocal perception, dogs may not broadly perceive a sound-size "frequency code" (Ohala, 1994) in relation to human voices, whereby low frequency vocalisations are perceived to be large and high frequency vocalisations are perceived to be small (Morton, 1977). Similarly to their responses to generic size cues, the subjects also failed to match simple height-related cues with the voices, as they did not associate shapes in a relatively higher spatial position with the adult male voices either.

Although no evidence was obtained to suggest that dogs match adult male human voices to the corresponding speakers on the basis of low-level visual cues related to age-dependent variation in their size or height, dogs' perception of adult male voices was connected to the values of two key anatomically-related vocal cues, the F0 and formant frequencies. Indeed, when the F0 and formant frequencies were re-synthesised to match the average values of a 6-year-old boy, the dogs no longer associated the adult silhouette with these voices. Dogs' reliance on the F0 and formant frequencies in discriminating human adult voices directly parallels human listeners' use of these anatomically-related cues to categorise the voices of men, women and children (Smith & Patterson, 2005), suggesting comparable underlying mechanisms of human voice age perception in both humans and dogs. However, contrary to our expectations, dogs did not associate the child silhouette with these child-like voices either. Therefore, while the subjects associated audio-visual cues related to adult men according to the relative values of the two key anatomical cues in their voices, they did not appear to perceive a similar correspondence between the body shape of a child and F0 and formant frequency values typical of children's voices. The most likely explanation for the discrepancy in their performance towards the child stimuli is that the subjects lacked sufficient prior exposure to children to have learnt to match their vocal parameters with children's body shapes. Consequently, the subjects' responses suggest that dogs must learn that certain anatomically-related F0 and formant frequency values in human voices are generally associated with specific age- and/or sex-related size categories (i.e. adult men tend to have a certain body shape and their voices have specific F0 and formant values). However, dogs did associate simple shapes placed at relatively low spatial elevations with children's voices, indicating that while they did not perceive a correspondence between the body shape of a child and children's voices, they may have learnt that children's voices tend to project from lower spatial positions. We propose

that our subjects may have previously learnt that human voices with a greater formant spacing and higher pitch tend to originate from lower to the ground than other human voices, and that perceptual narrowing may occur to only match children's voices with human body shapes (as indicated by their responses to adult male voices) as they acquire further experience with this specific age category (e.g. Lewkowicz & Ghazanfar, 2009). By replicating our study with subjects that have had specifically monitored or controlled levels of exposure to different human age and sex groups, it would be possible to test these predictions of how dogs acquire the ability to match human voices to speakers according to their age category.

Although dogs appear to associate some low-level features in the visual domain with the main age-related cues in human voices, they did not show any evidence of being influenced by general cross-modal correspondences involving the most perceptually salient cue in the auditory domain, the voice pitch. Firstly, dogs did not express their perception of any general statistical correspondences between either the child or adult male human silhouettes and simple pure tones matching the F0 in the child and adult voices. Furthermore, in contrast to human listeners, the subjects did not match shapes placed in high and low visual elevations (e.g. Rusconi et al., 2006; Walker et al., 2010) or shapes of small and large sizes (e.g. Parise & Spence, 2009) with the high and low frequency pure tones respectively. The lack of association between any of the available visual cues and the pure tones suggests that dogs do not apply general statistical correspondences between simple sound frequencies and basic age-related visual features in order to match human voices with unfamiliar speakers. Therefore, dogs may not perceive comparable audio-visual statistical correspondences to humans. As discussed in Article I of Chapter 4, the perception of statistical correspondences between basic visual and auditory features is largely unexplored in non-human animals. However, studies testing cross-modal correspondences in humans generally use larger frequency differences between the low and high tones than the stimuli used in our study (Spence & Deroy, 2013), which could explain why dogs did not show comparable responses. Furthermore, it has been suggested that the human perception of some cross-modal correspondences, such as between sound frequency and visual elevation, may be related to the structure of the human ear (Parise, Knorre & Ernst, 2014), and therefore may not be shared with more distantly related species like dogs that have very different pinnae shapes.

From Article II Chapter 4 it was possible to conclude that unlike human listeners (Pisanski, 2014), statistical cross-modal correspondences between basic stimulus features do not influence dogs' combination of multisensory indexical cues relating to human speakers. Furthermore, dogs do not appear to apply a broad "frequency code" to human voices, where high-frequency vocalisations are habitually matched to smaller sized individuals due to general correspondences between size-related audio-visual cues (Morton, 1977; Ohala, 1997). Instead, the results presented in Article II of Chapter 4 suggest that dogs learn to match specific values of the main anatomically-related cues in human voices with different age and sex categories, which might initially develop in dogs by associating the main age-related cues in human voices with different projection heights. This interpretation would also suggest that dogs' ability to match adult human voices to individuals of the corresponding gender, as demonstrated in Chapter 3, may also be acquired by learning to match the formant frequencies and F0 values characterising male and female adult voices with specific human body shapes. The fact that sexual dimorphism has led to categorical differences between the physical sizes of men, women and children, as well as categorical differences in their anatomically-related vocal cues (Fitch & Giedd, 1999), is likely to facilitate dogs' ability to learn to associate human voices to unfamiliar people according to these physical attributes. Further work is now needed to determine if dogs' perceptual or cognitive capacities have converged with humans to enable them to learn to combine multisensory indexical cues to categorise people, as such adaptations could have been promoted in dogs as their social interactions with humans became more complex during domestication. To test whether the ability to perceive indexical cues in unfamiliar human voices is specific to dogs as a result of domestication, or even specific to domesticated species, the studies presented in Chapters 3 and 4 should be replicated with individual animals from a wider range of domesticated and non-domesticated species that have been strongly exposed to humans.

Although important questions still remain regarding precisely how dogs discriminate indexical information, the results of Chapter 3 and Article II of Chapter 4 clearly demonstrate that dogs do attend to indexical information in human voices, perceiving relevant physical characteristics about unfamiliar adult human speakers relating to their gender and age. These observations build on previous evidence that dogs perceive the emotional and motivational prosody of human speech (Marshall-Pescini et al., 2011;

Mills et al., 2005; Scheider et al., 2011) and vocalisations (Andics et al., 2014; Yong & Ruffman, 2014), signifying that they process both dynamic and indexical supra-segmental content in human speech. Interestingly, in both of our studies dogs also showed significant side biases in their orientation responses to human voices, as they were more likely to look at the visual stimulus to their right after they were presented with human speech in Chapter 3, while the reverse bias was expressed in Article II of Chapter 4, as the subjects were more likely to look towards the image on their left when they were presented with simple vowel sounds. Human listeners generally show a right ear bias when asked to report linguistic information in speech, and a left ear bias when they are asked to report non-linguistic content such as the emotional tone (Kimura, 2011). These relative ear advantages are believed to occur because the contralateral projections from each ear to the auditory cortex are dominant over the ipsilateral projections (Bocca et al., 1995), and directly relate to evidence that for most people, the left hemisphere of the brain responds more strongly to meaningful segmental content in speech, whilst the right hemisphere is relatively more specialised in processing supra-segmental cues including the emotional prosody (e.g. Beaucousin et al., 2007; Buchanan et al., 2000; Friederici & Alter, 2004; McGettigan et al., 2012; Zatorre, Belin & Penhune, 2002). The contralateral auditory pathways are similarly dominant in dogs (Tunturi, 1946), which led to the question of whether our subjects' orientation responses differed because the familiar human phrases were processed differently to the simple vowel sounds presented across the two studies. However, due to the differential placing of the loud speakers between our two studies (one central speaker was used in Chapter 3 while two speakers were placed at either edge of the viewing wall in Article II of Chapter 4) it was not possible to directly compare the dogs' responses to the two types of vocal stimuli. Therefore to establish whether similarly to humans, dogs' orienting responses relate to the attended content of the speech signal, Chapter 5 investigated whether dogs show evidence of differently processing specific information transmitted in human speech.

**Do Dogs show Evidence of Hemispheric Asymmetries when Processing the Main Communicative Components of Human Speech, and if so, are Asymmetries Related to the Acoustic Structure of the Signals or their Functional Content?**

In the human brain, left hemispheric specialisations have been identified for processing intelligible segmental content in human speech, from individual phonemes (Agnew, McGettigan & Scott, 2011) through to meaningful sentences (McGettigan et al., 2012; Narain et al., 2003). In contrast, the right hemisphere of the brain responds more strongly to the supra-segmental content, preferentially processing the dynamic emotional prosody (e.g. Buchanan et al., 2000; Gandour et al., 2003) as well as speaker-related indexical cues encoding identity and gender (e.g. Belin, Zatorre, Lafaille, Ahad & Pike, 2000; Belin & Zatorre, 2003; Kreigstein & Girauld, 2004). Dogs' behavioural (Marshall-Pescini et al., 2011; Mills et al., 2005; Scheider et al., 2011), physiological (Yong & Ruffman, 2014) and neurological (Andics et al., 2014) responses had previously indicated that they perceive dynamic emotional content in human voices, while our investigations in Chapters 3 and 4 consolidated prior evidence that they also perceive indexical cues related to the physical characteristics of the speaker, including the person's familiarity (Adachi et al., 2007), gender and age. Some indications were also available to suggest that dogs may perceive segmental phonemic cues in speech (e.g. Fukuzawa et al., 2005; Gibson, Scavelli, Udell & Udell, 2014), although this evidence was not definitive.

The results obtained in Chapter 5 provided the first clear demonstration that dogs do perceive phonemic cues in human speech, as in our head-orienting paradigm the subjects showed opposite orienting responses to speech with salient segmental/phonemic content depending on whether the phonemes represented a learnt command or a novel phrase. When dogs were presented with the same learnt command after either the intonation pattern was neutralised, or all of the other acoustic cues were removed (increasing the salience of the segmental phonemic content), through two loud speakers on their left and right, they were more likely to orient towards the sound source on their right. In line with other studies that have used the head-orienting paradigm with animals (e.g. Hauser & Andersson, 1994; Siniscalchi, Quaranta & Rogers, 2008), as well as human dichotic listening studies (Kimura, 2011), we interpreted the dogs' orienting direction as suggesting stronger processing in the opposite hemisphere of the brain because of the dominance of the contra-lateral auditory

pathways (Tunturi, 1946). Therefore, the right head turn bias observed in response to human speech signals in which the salience of meaningful segmental phonemic cues had been increased indicates that dogs may predominantly process this information in the left hemisphere of their brain. In contrast, human speech lacking meaningful segmental phonemic content, either because the phonemic content was present but not meaningful or because it was removed altogether, generated opposite (left) head turn biases, indicating stronger right hemispheric involvement in processing these voices. Furthermore, dogs also showed a left head turn bias when the individual phonemes of the original command were re-ordered (i.e. from 'come on then' to 'thon om ken'), suggesting that they not only perceive the phonemic content of speech, but also understand its combinatorial structure to some extent. Because these speech signals differed only in the phonemic content, while all other acoustic parameters were held constant, the opposite response biases indicated that dogs perceive phonemic cues and differentiate between relevant and irrelevant verbal information in human speech. We hypothesised that because dogs attend to supra-segmental cues in a functionally relevant manner, removing any meaningful segmental cues may have increased the salience of the supra-segmental content for the dogs, and this information might be processed primarily in the right hemisphere of their brain. Evidence in support of stronger right hemispheric involvement for processing supra-segmental cues in human voices had previously been obtained by Andics et al., (2014), who observed stronger right hemispheric activation in alert dogs' brains in response to emotional content in both dog and human vocalisations. Therefore the subjects' opposite orienting biases in response to meaningful segmental and supra-segmental cues in human speech suggested that while they perceived both components, the segmental content of a familiar command appeared to be dissociated from the supra-segmental content and separately processed. The direction of the dogs' orienting responses further suggested that in parallel to humans, the left hemisphere of their brain preferentially responded to learnt segmental cues, while the right hemisphere showed stronger activation in response to the supra-segmental content.

Two general approaches have been developed in an effort to explain the lateralisation of speech perception in humans (Zatorre & Gandour, 2008). Acoustic, cue-dependent theories (Poeppel, 2003; Zatorre et al., 2002) propose that there are domain general differences in the acoustic sensitivities of the two brain hemispheres. Zatorre et al.

(2002) suggested that the left auditory regions code the detail of auditory signals with fine temporal resolution but poor spectral resolution, while the opposite occurs in the right hemisphere. In the related Asymmetric Sampling in Time model, Poeppel (2003) proposed that the left hemisphere of the brain samples smaller temporal windows than the right hemisphere, preferentially responding to more rapid acoustic transitions. Both accounts posit that hemispheric biases during speech perception are purely due to the basic acoustic structure of the information encoded in speech signals, as fast spectral variation characterises the segmental phonemic content while slow spectral variation characterises other speaker-related cues such as the emotional prosody. However, cue dependent theories have been criticised due to a lack of supporting evidence for rapid temporal processing in the left hemisphere of the human brain (e.g. McGettigan & Scott 2012; Scott & McGettigan, 2013). Such criticisms have given increasing support for more domain specific approaches, which propose that speech lateralisation is dependent on the communicative function of the acoustic content in vocal signals rather than their basic structure (e.g. Belin, Fecteau & Bedard, 2004; Scott & Wise, 2004). In our study presented in Chapter 4, dogs showed opposite orienting biases to acoustically equivalent speech stimuli depending on whether the phonemic content was relevant or meaningless to them. Furthermore, no orientation biases were obtained in response to simple sine-wave tones matching the intonation patterns of the learnt commands. Therefore, the basic acoustic structure of the sounds could not account for the dogs' responses, indicating that in dogs, hemispheric biases in response to human speech are not related to domain general acoustic biases as proposed by cue-dependent approaches (Poeppel, 2003; Zatorre et al., 2002). Instead, our results are more consistent with the functionally dependent accounts of human hemispheric specialisations in response to speech. Although, as outlined in the human literature (Scott & McGettigan, 2013), the suggested biases indicated by the dogs' responses may not be specifically related to speech processing, but could be accounted for by more general processes. For example, the left hemisphere is thought to be specialised in the formation of learnt patterns in mammals (Vallortigara et al., 2008), which may not be mutually exclusive with observations that in humans, the right hemisphere is more strongly activated by unattended speech and other sounds (Scott, Rosen, Bearman, Davis & Wise, 2009).

The left hemispheric advantage suggested by the dogs' orienting responses to salient meaningful segmental phonemic cues, and the right hemispheric advantage suggested

by the dogs' responses to salient supra-segmental prosodic and speaker-related cues, is strikingly parallel to the hemispheric asymmetries reported in humans. The similarities suggest that dogs not only perceive each of the main communicative components of speech, but that they may also process speech in a way which is broadly comparable to humans. Although more in depth exploration using neuro-imaging techniques is needed before we can determine the precise degree of similarity between how humans and dogs process relevant information in speech, it is possible that correspondences between the hemispheric asymmetries across the two species may be the result of convergent evolution, if dogs have adapted to perceive human vocal signals to facilitate inter-specific communication during the process of domestication. Such cognitive convergence has previously been suggested to have occurred in relation to dogs' perception of human visual gestures (e.g. Hare & Tomasello, 2005), as they outperform their closest wild relative, the grey wolf, in responding to visual human signals (Gásci et al., 2009; Hare, Brown, Williamson & Tomasello, 2002). Alternatively, rather than indicating convergent evolution, the similarities observed between dog and human responses to speech may be explained by conserved mammalian hemispheric specialisations that are expressed when the individual is exposed to functionally meaningful speech signals. Support for this hypothesis derives from demonstrations that a wide range of phylogenetically distant mammal species, including dogs (Siniscalchi, Lusito, Sasso & Quaranta, 2012; Siniscalchi et al., 2008), show hemispheric asymmetries in response to conspecific vocalisations, primarily by expressing stronger left hemispheric activation (Ocklenburg, Ströckens, & Güntürkün, 2013). In dogs, right hemispheric lateralisation has also been observed in response to the emotional content in conspecific vocalisations (Andics et al., 2014). Therefore, across mammals the left hemisphere may preferentially process species-specific features of vocal signals that have learnt relevance to the individual, which extends to include the phonemic cues in familiar spoken commands in dogs due to the comparably high relevance of human signals for them. Because dogs may show a greater predisposition to accept humans as social partners (Gácsi et al., 2009; Hare et al., 2005; Hare & Tomasello, 2005; Hare, Wobber & Wrangham, 2012; Range & Viranyi, 2014; Udell et al., 2010), a slight variation on this hypothesis could also be put forward, whereby although mammals may generally share the same hemispheric organisation in response to conspecific vocalisations, dogs could have adapted a stronger latent potential to process the communicative content of hetero-specific human vocalisations equivalently to their own

vocal signals. To test these hypotheses, the experiments in Chapter 5 could be replicated with subjects from other domesticated (e.g. domestic horses *Equus caballus*) and non-domesticated species (e.g. captive wolves) that have had a comparable level of exposure to relevant human speech as dogs in their lifetime. If the same orienting biases are observed in non-domesticated species, this would suggest that the evolutionary origins of speech lateralisation occurred at an earlier point in human evolutionary history than previously thought. Alternatively, failure to replicate our results in any other species would imply some degree of specific adaptation in dogs which has altered their perception of information encoded in human speech.

Further avenues for relevant future research include applying the orienting biases displayed by dogs to determine if their attention to the different components of human speech is context specific, such as when responding to verbal commands, discriminating between speakers or interpreting emotional expressions. As well as furthering our understanding of how dogs perceive speech in relation to human listeners, investigating the extent to which dogs preferentially rely on segmental or supra-segmental cues when responding to different human vocal signals could provide important practical insights for human-dog interactions, particularly if differences in reliance on specific speech components are breed, age or individually specific. Technologies such as near-infrared spectroscopy, which has been successfully applied in human infant studies investigating hemispheric biases in response to speech (e.g. Grossmann, Oberecker, Koch & Friederici, 2010), could potentially facilitate such investigations, as direct readings of hemispheric activation can be taken whilst the subject is in motion.

To summarise, Chapter 4 determined that dogs showed right sided orienting biases when their attention was drawn to meaningful segmental cues in human speech, while left sided orienting biases were produced in response to human speech lacking meaningful segmental cues. These biases indicate that in dogs the left hemisphere may preferentially process learnt phonemic cues encoded in the segmental content of human speech signals, while the right hemisphere may show stronger activation in response to supra-segmental cues related to the emotional and physical characteristics of the speaker, directly paralleling the hemispheric biases which have been observed in human listeners during speech perception.

## Conclusion

The primary aim of the thesis was to determine if dogs living with humans spontaneously perceive the main communicatory components of human speech, as although they were already known to respond to the emotional/motivational prosody of human voices, it had not previously been established whether dogs perceived indexical cues related to the physical characteristics of the speaker or the phonemic verbal content of speech signals. In addition, a further aim of the current work was to explore the perceptual and cognitive mechanisms underlying dogs' functional discrimination of speech components in relation to those documented in humans. We first confirmed that dogs do spontaneously perceive physical indexical cues in human speech, enabling them to categorise adult human speakers according to their gender and to discriminate adult male speakers from children. Furthermore, dogs showed evidence of using the same anatomically-related acoustic cues to human listeners in order to make these associations, suggesting that the perception of indexical characteristics in human voices may be achieved through shared mechanisms that are present in both humans and dogs. Additional indications of comparable perceptual mechanisms between the two species were obtained from dogs' orienting responses to different components of speech signals, which suggested that in parallel with humans, dogs may also show a left hemispheric specialisation for processing meaningful phonemic segmental content, and stronger right hemispheric activation in response to supra-segmental emotional and indexical cues. Furthermore, similarly to evidence obtained for humans, dogs' orienting responses indicated that the potential underlying hemispheric biases are more likely to be related to the communicative content of the signals and rather than their basic acoustic composition. Determining the functional basis of the expressed orienting biases also revealed that dogs do attend to the phonemic cues in human speech, and moreover that they show some understanding of the combinatory structure of the phonemes in familiar commands. Therefore, together the studies comprising the current thesis provide the first clear demonstration that dogs spontaneously perceive each of the main communicative components of human speech in a functionally relevant manner, confirming that they are not limited to extracting information from only the emotional prosody of human voices. The additional evidence which we obtained suggesting that dogs may perceive some human voice characteristics through comparable perceptual and cognitive mechanisms to human listeners provides the first step in establishing

whether dogs have adapted to extract information from human vocalisations during the process of domestication, and whether any such adaptations represent a case of convergent evolution with humans. As well as further investigating dogs' perception of different speech cues, future studies should also be carried out with human-socialised individuals from other domesticated and non-domesticated species in order to test the level of specialism in dogs' perception of human vocal signals.

## References


Adachi, I., & Fujita, K. (2007). Cross-modal representation of human caretakers in squirrel monkeys, *Behavioural Processes, 74*, 27–32.

Adachi, I., & Hampton, R. (2011). Rhesus monkeys see who they hear: Spontaneous cross-modal memory for familiar conspecifics, *PLoS One, 6*, e23345.

Adachi, I., Kuwahata, H., & Fujita, K. (2007). Dogs recall their owner's face upon hearing the owner's voice. *Animal Cognition*, *10*(1), 17-21.

Adachi, I., Kuwahata, H., Fujita, K., Tomonaga, M., & Matsuzawa, T. (2006). Japanese macaques form a cross-modal representation of their own species in their first year of life, *Primates, 47*, 350–354.

Adachi, I., Kuwahata, H., Fujita, K., Tomonaga, M., & Matsuzawa, T. (2009). Plasticity of ability to form cross-modal representations in infant Japanese macaques, *Developmental Science, 12*, 446–452.

Agnetta, B., Hare, B., & Tomasello, M. (2000). Cues to food location that domestic dogs (Canis familiaris) of different ages do and do not use. *Animal Cognition*, *3*(2), 107-112.

Agnew, Z. K., McGettigan, C., & Scott, S. K. (2011). Discriminating between auditory and motor cortical responses to speech and nonspeech mouth sounds. *Journal of Cognitive Neuroscience*, *23*(12), 4038-4047.

Agrillo, C. & Petrazzini, M. E. M. (2013). Glimpse of ATOM in non-human species? *Frontiers in Psychology, 4*, 1–4.

Andics, A., Gácsi, M., Faragó, T., Kis, A., & Miklósi, Á. (2014). Voice-sensitive regions in the dog and human brain are revealed by comparative fMRI. *Current Biology*, *24*(5), 574-578.

Andrew, R. J. (1962). The origin and evolution of the calls and facial expressions of the primates, *Behaviour, 20*, 1–109.

Andrew, R. J., & Rogers, L. J. (2002). The nature of lateralization in tetrapods. In L. J. Rogers & R. Andrew (Eds.). *Comparative Vertebrate Lateralization* (pp 94-125). Cambridge: Cambridge University Press.

Aslin, R. N. (2007). What's in a look? *Developmental Science, 10*, 48–53.

Aslin, R. N., & Fiser, J. (2005). Methodological challenges for understanding cognitive development in infants. *Trends in Cognitive Science, 9*, 92–98.

Assmann, P. F., Dembling, S., & Nearey, T. M. (Eds.). (2006). Effects of frequency shifts on perceived naturalness and gender information in speech. Proceedings from Ninth International Conference Spoken Language Processing. Pittsburgh, USA.

Autier-Dérian, D., Deputte, B. L., Chalvet-Monfray, K., Coulon, M., & Mounier, L. (2013). Visual discrimination of species in dogs (*Canis familiaris*). *Animal Cognition*, *16,* 637-651.

Bachorowski, J. A., & Owren, M. J. (1999). Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech. *The Journal of the Acoustical Society of America, 106*(2), 1054-1063.

Bahrick, L. E., Hernandez-Reif, M., & Flom, R. (2005). The development of infant learning about specific face-voice relations. *Developmental Psychology*, *41*(3), 541-552.

Baillargeon, R., Spelke, E. S., & Wasserman, S. (1985). Object permanence in five-month-old infants. *Cognition* 20, 191–208.

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, *70*(3), 614-636.

Barrera, G., Jakovcevic, A., Elgier, A. M., Mustaca, A., & Bentosela, M. (2010). Responses of shelter and family dogs to an unknown human. *Journal of Veterinary Behavior: Clinical Applications and Research*, *5*(6), 339-344.

Barsalou, L. W. (2005). Continuity of the conceptual system across species. *Trends in Cognitive Science, 9*, 309–311.

Baru, A. V. (1975). Discrimination of synthesized vowels [a] and [i] with varying parameters (fundamental frequency, intensity, duration and number of formants) in dog. In G. Fant & M. A. A. Tatham (Eds.), *Auditory Analysis and Perception of Speech* (pp. 91-101). New York: Academic Press.

Basile, M., Lemasson, A., & Blois-Heulin, C. (2009). Social and emotional values of sounds influence human (*Homo sapiens*) and non-human primate (*Cercopithecus campbelli*) auditory laterality. *PLoS One, 4*, e6295.

Batteau, D. W. (1967). The role of the pinna in human localization, *Proceedings of the Royal Society of London B. Biological. Sciences, 168*, 158–180.

Beaucousin, V., Lacheret, A., Turbelin, M. R., Morel, M., Mazoyer, B., & Tzourio-Mazoyer, N. (2007). FMRI study of emotional speech comprehension. *Cerebral Cortex, 17*(2), 339-352.

Belin, P., Fecteau, S., & Bedard, C. (2004). Thinking the voice: neural correlates of voice perception. *Trends in Cognitive Sciences, 8*(3), 129-135.

Belin, P., Fecteau, S., Charest, I., Nicastro, N., Hauser, M. D., & Armony, J. L. (2008). Human cerebral response to animal affective vocalizations. *Proceedings of the Royal Society of London B: Biological Sciences, 275*(1634), 473-481.

Belin, P., & Zatorre, R. J. (2003). Adaptation to speaker's voice in right anterior temporal lobe. *Neuroreport, 14*(16), 2105-2109.

Belin, P., Zatorre, R. J., & Ahad, P. (2002). Human temporal-lobe response to vocal sounds. *Cognitive Brain Research, 13*(1), 17-26.

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature, 403*(6767), 309-312.

Belin, P., Zilbovicius, M., Crozier, S., Thivard, L., Fontaine, A., Masure, M. C., & Samson, Y. (1998). Lateralization of speech and auditory temporal processing. *Journal of Cognitive Neuroscience, 10*(4), 536-540.

Bentosela, M., Barrera, G., Jakovcevic, A., Elgier, A. M., & Mustaca, A. E. (2008). Effect of reinforcement, reinforcer omission and extinction on a communicative response in domestic dogs (Canis familiaris). *Behavioural Processes, 78*(3), 464-469.

Bertelson, P. & Aschersleben, G. (1998). Automatic visual bias of perceived auditory location, *Psychonomic Bulletin & Review, 5*, 482–489.

Bertelson, P. & Radeau, M. (1981). Cross-modal bias and perceptual fusion with auditory–visual spatial discordance, *Perception & Psychophysics, 29*, 578–584.

Bocca, E., Calearo, C., Cassinari, V., & Migliavacca, F. (1955). Testing "cortical" hearing in temporal lobe tumours. *Acta Otolaryngologica, 45*(4), 289-304.

Boemio, A., Fromm, S., Braun, A., & Poeppel, D. (2005). Hierarchical and asymmetric temporal sensitivity in human auditory cortices. *Nature Neuroscience, 8*(3), 389-395.

Bo-Jørgensen, J. (2009). Dynamics of multiple signalling systems: Animal communication in a world of flux, *Trends Ecol. Evol*. **25**, 292–300.

Bovet, D. & Deputte, B. L. (2009). Matching vocalizations to faces of familiar conspecifics in grey-cheeked mangabeys (*Lophocebus albigena*). *Folia Primatologica, 80*, 220–232.

Boycott, B. B., & Young, J. Z. (1956). Reactions to shape in *Octopus vulgaris Lamarck*. *Proceedings of the Zoological Society of London*, *126*(4), 491-547.

Böye, M., Güntürkün, O., & Vauclair, J. (2005). Right ear advantage for conspecific calls in adults and subadults, but not infants, California sea lions (*Zalophus californianus*): hemispheric specialization for communication? *European Journal of Neuroscience*, *21*(6), 1727-1732.

Bradbury, J. W. & Vehrencamp, S. I. (1998). *Principles of Animal Communication*. Sinauer Press, Sunderland, MA, USA.

Bradshaw, J. W. S., & Nott, H. M. R. (1995). Social and communication behaviour of companion dogs. *The domestic dog: its evolution, behaviour and interactions with people*, 115-130.

Bräuer, J., Kaminski, J., Riedel, J., Call, J., & Tomasello, M. (2006). Making inferences about the location of hidden food: social dog, causal ape. *Journal of Comparative Psychology*, *120*(1), 38–47.

Bremner, A. J., Caparos, S., Davidoff, J., de Fockert, J., Linnell, K. J. & Spence, C. (2013). "Bouba" and "Kiki" in Namibia? A remote culture make similar shape–sound matches, but different shape–taste matches to Westerners. *Cognition*, *126*, 165–172.

Briefer, E. F. (2012). Vocal expression of emotions in mammals: mechanisms of production and evidence. *Journal of Zoology*, *288*(1), 1-20.

Buchanan, T. W., Lutz, K., Mirzazade, S., Specht, K., Shah, N. J., Zilles, K., & Jäncke, L. (2000). Recognition of emotional prosody and verbal components of spoken language: an fMRI study. *Cognitive Brain Research*, *9*(3), 227-238.

Burton, M. A., Bruce, V., & Dench, N. (1993). What's the difference between men and women? Evidence from facial measurement. *Perception*, *22*(2), 153-176.

Buttelmann, D., & Tomasello, M. (2013). Can domestic dogs (*Canis familiaris*) use referential emotional expressions to locate hidden food? *Animal Cognition*, *16*(1), 137-145.

Carreiras, M., Lopez, J., Rivero, F., & Corina, D. (2005). Linguistic perception: neural processing of a whistled language. *Nature*, *433*(7021), 31-32.

Chan, R. W., & Titze, I. R. (2006). Dependence of phonation threshold pressure on vocal tract acoustics and vocal fold tissue mechanics. *The Journal of the Acoustical Society of America*, *119*(4), 2351-2362.

Cantlon, J. F. (2012). Math, monkeys, and the developing brain. *Proceedings of the National Academy of Sciences*. *109,* 10725–10732.

Cantlon, J. F. & Brannon, E. M. (2007). Basic math in monkeys and college students. *PLoS Biology, 5*, e328.

Cartei, V., Cowles, H. W. & Reby, D. (2012). Spontaneous voice gender imitation abilities in adult speakers. *PLoS One, 7,* e31353.

Charlton, B. D., Reby, D. & McComb, K. (2008). Effect of combined source (F0) and filter (formant) variation on red deer hind responses to male roars. *The Journal of the Acoustical Society of America. 123,* 2936–2943.

Charlton, B. D., Ellis, W. A., Brumm, J., Nilsson, K., & Fitch, W. T. (2012). Female koalas prefer bellows in which lower formants indicate larger males. *Animal Behaviour*, *84*(6), 1565-1571.

Charlton, B. D., Ellis, W. A., McKinnon, A. J., Cowin, G. J., Brumm, J., Nilsson, K. & Fitch, W. T. (2011). Cues to body size in the formant spacing of male koala (*Phascolarctos cinereus*) bellows: Honesty in an exaggerated trait, *Journal of Experimental Biology*. *214*, 3414–3422.

Charlton, B. D., Zhihe, Z. & Snyder, R. J. (2010). Giant pandas perceive and attend to formant frequency variation in male bleats. *Animal Behaviour, 79,* 1221–1227.

Cheng, K. (1990). More psychophysics of the pigeon's use of landmarks. *Journal of Comparative Physiology A, 166,* 857–863.

Collins, S. A. (2000). Men's voices and women's choices. *Animal Behaviour*, *60*(6), 773-780.

Cooper, B. G. & Goller, F. (2004). Multimodal signals: Enhancement and constraint of song motor patterns by visual display. *Science, 303*(5657), 544–546.

Coppinger, R., & Coppinger, L. (2001). *Dogs: A startling new understanding of canine origin, behavior & evolution*. New York: Simon and Schuster.

Crelin, E. S. (1987). *The human vocal tract: Anatomy, function, development, and evolution*. London: Vantage Press.

Csibra, G. (2003). Teleological and referential understanding of action in infancy. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences*, *358*(1431), 447–458.

Csibra, G. (2010). Recognizing communicative intentions in infancy. *Mind & Language*, *25*(2), 141-168.

Cunningham, C. L., & Ramos, M. F. (2014). Effect of training and familiarity on responsiveness to human cues in domestic dogs (Canis familiaris). *Animal Cognition*, *17*(3), 805-814.

Davies, N. B. & Halliday, T. R. (1978). Deep croaks and fighting assessment in toads *Bufo bufo*. *Nature, 274*, 683–685.

Davis, S. J., & Valla, F. R. (1978). Evidence for domestication of the dog 12,000 years ago in the Natufian of Israel. *Nature, 276*, 608-610

De Boer, B. (2008). The acoustic role of supralaryngeal air sacs. *The Journal of the Acoustical Society of America. 123*, 3779–3779.

De Hevia, M. D., Girelli, L., Addabbo, M. & Cassia, V. M. (2014). Human infants' preference for left-to-right oriented increasing numerical sequences. *PLoS One, 9*, e96412.

Delfour, F., & Marten, K. (2006). Lateralized visual behavior in bottlenose dolphins (*Tursiops truncatus*) performing audio-visual tasks: The right visual field advantage. *Behavioural Processes*, *71*(1), 41-50.

Doehrmann, O. & Naumer, M. J. (2008). Semantics and the multisensory brain: How meaning modulates processes of audio-visual integration. *Brain Research, 1242*, 136–150.

Ecklund-Flores, L., & Turkewitz, G. (1996). Asymmetric headturning to speech and nonspeech in human newborns. *Developmental Psychobiology*, *29*(3), 205-217.

Ehret, G. (1987). Left hemisphere advantage in the mouse brain for recognizing ultrasonic communication calls. *Nature*, *325*(6101), 249-251.

Elgier, A. M., Jakovcevic, A., Barrera, G., Mustaca, A. E., & Bentosela, M. (2009). Communication between domestic dogs (*Canis familiaris*) and humans: dogs are good learners. *Behavioural Processes*, *81*(3), 402-408.

Emmerton, J. & Renner, J. C. (2006). Scalar effects in the visual discrimination of numerosity by pigeons. *Learning & Behavior, 34*, 176–192.

Ernst, M. O. (2005). A Bayesian view on multimodal cue integration, in: *Perception of the Human Body From the Inside Out*, G. Knoblich, I. Thornton, M. Grosejan and M. Shiffrar (Eds), pp. 105–131, Oxford University Press, Oxford, UK.

Ernst, M. O. (2007). Learning to integrate arbitrary signals from vision and touch. *Journal of Vision, 7*, 1–14.

Ernst, M. O. & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Science, 8,* 162–169.

Evans, S., Neave, N., Wakelin, D. & Hamilton, C. (2008). The relationship between testosterone and vocal frequencies in human males. *Physiology & Behavior, 93,* 783–788.

Evans, T. A., Howell, S. & Westergaard, G. C. (2005). Auditory-visual cross-modal perception of communicative stimuli in tufted capuchin monkeys (*Cebus apella*). *Journal of Experimental Psychology: Animal Behavior Processes, 31,* 399–406.

Evans, K. K., & Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features. *Journal of Vision*, *10*(1), 6.

Ey, E., Pfefferle, D. & Fischer, J. (2007). Do age-and sex-related variations reliably reflect body size in non-human primate vocalizations? A review. *Primates, 48,* 253–267.

Fant, G. (1960). *The Acoustic Theory of Speech Production*. The Hague: Mouton.

Faragó, T., Andics, A., Devecseri, V., Kis, A., Gácsi, M., & Miklósi, Á. (2014). Humans rely on the same rules to assess emotional valence and intensity in conspecific and dog vocalizations. *Biology Letters*, *10*(1), 20130926.

Faragó, T., Pongrácz, P., Miklósi, Á., Huber, L., Virányi, Z., & Range, F. (2010). Dogs' expectation about signalers' body size by virtue of their growls. *PLoS One*, *5*(12), e15175.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175-191.

Feinberg, D. R., Jones, B. C., Little, A. C., Burt, D. M., & Perrett, D. I. (2005). Manipulations of fundamental and formant frequencies influence the attractiveness of human male voices, *Animal Behaviour, 69,* 561–568.

Fernández-Prieto, I., Navarra, J., & Pons, F. (2015). How big is this sound? Crossmodal association between pitch and size in infants. *Infant Behavior and Development*, *38*, 77-81.

Fischer, J., Teufel, C., Drolet, M., Patzelt, A., Rusamen, R., von Cramon, D. Y., & Schubotz, R. I. (2009). Orienting asymmetries and lateralized processing of sounds in humans. *BMC Neuroscience*, *10*(1), 14.

Fitch, W. T. (1997). Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. *The Journal of the Acoustical Society of America, 102*(2), 1213-1222.

Fitch, W. T. (2000a). The phonetic potential of nonhuman vocal tracts: comparative cineradiographic observations of vocalizing animals. *Phonetica, 57*(2-4), 205-218.

Fitch, W. T. (2000b). Skull dimensions in relation to body size in nonhuman mammals: The causal bases for acoustic allometry. *Zoology-Analysis of Complex Systems, 103*(1-2), 40-58.

Fitch, W. T. (2000c). The evolution of speech: A comparative review, *Trends in Cognitive Science, 4*, 258–267.

Fitch, W. T. (2010). *The Evolution of Language*. Cambridge University Press, Cambridge, UK.

Fitch, W. T. & Giedd, J. (1999). Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *Journal of the Acoustical Society of America, 106*(3), 1511-1522.

Fitch, W. T., & Hauser, M. D. (2003). Unpacking "honesty": vertebrate vocal production and the evolution of acoustic signals. In A. M. Simmons, Fay, R. R. & A. N. Popper (Eds.), *Acoustic communication* (pp. 65-137). New York: Springer.

Fitch, W. T., Neubauer, J., & Herzel, H. (2002). Calls out of chaos: the adaptive significance of nonlinear phenomena in mammalian vocal production. *Animal Behaviour, 63*(3), 407-418.

Fitch, W. T., & Reby, D. (2001). The descended larynx is not uniquely human. *Proceedings of the Royal Society of London B: Biological Sciences, 268*(1477), 1669-1675.

Fletcher, N. H. (2004). A simple frequency-scaling rule for animal communication. *The Journal of the Acoustical Society of America, 115*, 2334–2338.

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences, 6*(2), 78-84.

Friederici, A. D., & Alter, K. (2004). Lateralization of auditory language functions: a dynamic dual pathway model. *Brain and Language, 89*(2), 267-276.

Friederici, A. D., Kotz, S. A., Scott, S. K., & Obleser, J. (2010). Disentangling syntax and intelligibility in auditory language comprehension. *Human Brain Mapping*, *31*(3), 448-457.

Fujita, K. (1997). Perception of the Ponzo illusion by rhesus monkeys, chimpanzees, and humans: Similarity and difference in the three primate species. *Perception & Psychophysics*, *59*(2), 284-292.

Fukuzawa, M., Mills, D. S., & Cooper, J. J. (2005). The effect of human command phonetic characteristics on auditory cognition in dogs (*Canis familiaris*). *Journal of Comparative Psychology*, *119*(1), 117.

Gácsi, M., Győri, B., Miklósi, Á., Virányi, Z., Kubinyi, E., Topál, J., & Csányi, V. (2005). Species-specific differences and similarities in the behavior of hand-raised dog and wolf pups in social situations with humans. *Developmental Psychobiology*, *47*(2), 111-122.

Gácsi, M., Gyori, B., Virányi, Z., Kubinyi, E., Range, F., Belényi, B., & Miklósi, Á. (2009). Explaining dog wolf differences in utilizing human pointing gestures: selection for synergistic shifts in the development of some social skills. *PLoS One*, *4*(8), e6584.

Gaffan, D., & Harrison, S. (1991). Auditory-visual associations, hemispheric specialization and temporal-frontal interaction in the Rhesus monkey. *Brain*, *114*(5), 2133-2144.

Gallace, A., & Spence. C. (2006). Multisensory synesthetic interactions in the speeded classification of visual size, *Perception & Psychophysics, 68*, 1191–1203.

Gandour, J., Wong, D., Dzemidzic, M., Lowe, M., Tong, Y., & Li, X. (2003). A cross-linguistic fMRI study of perception of intonation and emotion in Chinese. *Human Brain Mapping*, *18*(3), 149-157.

Gaulin, S., & Boster, J. (1985). Cross-cultural differences in sexual dimorphism: Is there any variance to be explained? *Ethology and Sociobiology*, *6*(4), 219-225.

Gaver, W. W. (1993). What in the world do we hear?: An ecological approach to auditory event perception. *Ecological Psychology, 5*, 1–29.

Ghazanfar, A. A. (2013). Multisensory vocal communication in primates and the evolution of rhythmic speech. *Behavioral Ecology and Sociobiology, 67*, 1441–1448.

Ghazanfar, A. A., & Logothetis, N. K. (2003). Facial expressions linked to monkey calls. *Nature, 423*, 937–938.

Ghazanfar, A. A., & Maier, J. X. (2009). Rhesus monkeys (*Macaca mulatta*) hear rising frequency sounds as looming. *Behavioral Neuroscience, 123,* 822–827.

Ghazanfar, A. A., Neuhoff, J. G., & Logothetis, N. K. (2002). Auditory looming perception in rhesus monkeys, *Proceedings of the National Academy of Sciences, 99,* 15755–15757.

Ghazanfar, A. A., Maier, J. X., Hoffman, K. L., & Logothetis, N. K. (2005). Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *Journal of Neuroscience, 25*, 5004–5012.

Ghazanfar, A.A., Nielsen, K., & Logothetis, N.K. (2006). Eye movements of monkey observers viewing vocalizing conspecifics. *Cognition, 101,* 515–529.

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. *Current Biology*, *17*(5), 425-430.

Ghazanfar, A. A., Smith-Rohrberg, D., & Hauser, M. D. (2001). The role of temporal cues in rhesus monkey vocal recognition: orienting asymmetries to reversed calls. *Brain, Behavior and Evolution*, *58*(3), 163-172.

Gibbon, J. (1977). Scalar expectancy theory and Weber's law in animal timing, *Psychological Review, 84*, 279–325.

Gibson, J. M., Scavelli, S. A., Udell, C. J., & Udell, M. A. R. (2014). Domestic dogs (Canis lupus familiaris) are sensitive to the "human" qualities of vocal commands. *Animal Behavior & Cognition*, 1(3), 281-295.

Gil-da-Costa, R., Braun, A., Lopes, M., Hauser, M. D., Carson, R. E., Herscovitch, P., & Martin, A. (2004). Toward and evolutionary perspective on conceptual representation: Species-specific calls activate visual and affective processing systems in the macaque, *Proceedings of the National Academy of Sciences, 101*, 17516–17521.

Gil-da-Costa, R., & Hauser, M. D. (2006). Vervet monkeys and humans show brain asymmetries for processing conspecific vocalizations, but with opposite patterns of laterality. *Proceedings of the Royal Society B—Biological Sciences, 273,* 2313–2318.

Goldstein, U. G. (1980). An articulatory model for the vocal tracts of growing children. (Doctoral dissertation). Retrieved from Massachusetts Institute of Technology, http://hdl.handle.net/1721.1/16118

Golinkoff, R. M., Hirsh-Pasek, K., Cauley, K. M., & Gordon, L. (1987). The eyes have it: Lexical and syntactic comprehension in a new paradigm, *Journal of Child Language, 14*, 23–45.

Golinkoff, R. M., Ma, W., Song, L., & Hirsh-Pasek, K. (2013). Twenty-five years using the intermodal preferential looking paradigm to study language acquisition what have we learned? *Perspectives on Psychological Science, 8*, 316–339.

González, J. (2004). Formant frequencies and body size of speaker: a weak relationship in adult humans. *Journal of Phonetics, 32*(2), 277-287.

Green, K. P., Kuhl, P. K., Meltzoff, A. N., & Stevens, E. B. (1991). Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Percept. Psychophysics, 50*, 524–536.

Greene, E., & Meagher, T. (1998). Red squirrels, *Tamiasciurus hudsonicus*, produce predator-class specific alarm calls. *Animal Behavior, 55*, 511–518.

Griebel, U., & Oller, D. K. (2012). Vocabulary learning in a Yorkshire terrier: slow mapping of spoken words. *PLOs one, 7*(2), e30182.

Grimshaw, G. M., Kwasny, K. M., Covell, E., & Johnson, R. A. (2003). The dynamic nature of language lateralization: effects of lexical and prosodic factors. *Neuropsychologia, 41*(8), 1008-1019.

Grossmann, T., Oberecker, R., Koch, S. P., & Friederici, A. D. (2010). The developmental origins of voice processing in the human brain. *Neuron, 65*(6), 852-858.

Gunderson, V. M., Rose, S. A., & Grant-Webster, K. S. (1990). Cross-modal transfer in high- and low-risk infant pigtailed macaque monkeys. *Developmental Psychology, 26*, 576–581.

Gunderson, V. M., Yonas, A., Sargent, P. L., & Grant-Webster, K. S. (1993). Infant macaque monkeys respond to pictorial depth. *Psychological Science, 4*(2), 93-98.

Györi, B., Gácsi, M., & Miklósi, A. (2010). Friend or foe: Context dependent sensitivity to human behaviour in dogs. *Applied Animal Behaviour Science, 128*, 69-77.

Hall II, J. L., & Goldstein Jr, M. H. (1968). Representation of binaural stimuli by single units in primary auditory cortex of unanesthetized cats. *Journal of the Acoustical Society of America, 43*(3), 456-461.

Hare, B., Brown, M., Williamson, C., & Tomasello, M. (2002). The domestication of social cognition in dogs. *Science, 298*(5598), 1634-1636.

Hare, B., Call, J., & Tomasello, M. (1998). Communication of food location between human and dog (*Canis familiaris*). *Evolution of Communication*, *2*(1), 137-159.

Hare, B., & Tomasello, M. (1999). Domestic dogs (Canis familiaris) use human and conspecific social cues to locate hidden food. *Journal of Comparative Psychology*, *113*(2), 173-177.

Hare, B., & Tomasello, M. (2005). Human-like social skills in dogs?. *Trends in Cognitive Sciences*, *9*(9), 439-444.

Hare, B., Plyusnina, I., Ignacio, N., Schepina, O., Stepika, A., Wrangham, R., & Trut, L. (2005). Social cognitive evolution in captive foxes is a correlated by-product of experimental domestication. *Current Biology*, *15*(3), 226-230.

Hare, B., Wobber, V., & Wrangham, R. (2012). The self-domestication hypothesis: evolution of bonobo psychology is due to selection against aggression. *Animal Behaviour*, *83*(3), 573-585.

Harley, H. E., Putman, E. A., & Roitblat, H. L. (2003). Bottlenose dolphins perceive object features through echolocation. *Nature*, *424*, 667–669.

Harries, M., Hawkins, S., Hacking, J., & Hughes, I. (1998). Changes in the male voice at puberty: vocal fold length and its relationship to the fundamental frequency of the voice. *Joural of Laryngology & Otology*, *112*, 451–454.

Hauser, M. D., Agnetta, B., & Perez, C. (1998). Orienting asymmetries in rhesus monkeys: the effect of time-domain changes on acoustic perception. *Animal Behaviour*, *56*(1), 41-47.

Hauser, M. D., & Andersson, K. (1994). Left hemisphere dominance for processing vocalizations in adult, but not infant, rhesus monkeys: field experiments. *Proceedings of the National Academy of Sciences*, *91*(9), 3946-3948.

Heffner, H. (1975). Perception of biologically meaningful sounds by dogs. *Journal of the Acoustical Society of America*, *58*, S124.

Hauser, M. D., Evans, C. S., & Marler, P. (1993). The role of articulation in the production of rhesus-monkey, *Macaca mulatta*, vocalisations, *Animal Behaviour*, *45*, 423–433.

Heffner, H. E., & Heffner, R. S. (1984). Temporal lobe lesions and perception of species-specific vocalizations by macaques. *Science*, *226*(4670), 75-76.

Heffner, H. E., & Heffner, R. S. (1986). Effect of unilateral and bilateral auditory cortex lesions on the discrimination of vocalizations by Japanese macaques. *Journal of Neurophysiology, 56*(3), 683-701.

Held, R., Ostrovsky, Y., de Gelder, B., Gandhi, T., Ganesh, S., Mathur, U., & Sinha, P. (2011). The newly sighted fail to match seen with felt. *Nature Neuroscience, 14*, 551–553.

Hennessy, M.B., Williams, M.T., Miller, D.D., Douglas, C.W., & Voith, V.L. (1998). Influence of male and female petters on plasma cortisol and behaviour: can human interaction reduce the stress of dogs in a public animal shelter? *Applied Animal Behaviour Science, 61*, 63-77.

Herman, L. M., Pack, A. A., & Hoffmann-Kuhnt, M. (1998). Seeing through sound: Dolphins *(Tursiops truncatus)* perceive the spatial structure of objects through echolocation, *Joural of Comparative Psychology, 112*, 292–305.

Herman, L. M., Richards, D. G., & Wolz, J. P. (1984). Comprehension of sentences by bottlenose dolphins. *Cognition, 16*(2), 129-219.

Herrmann, B., Obleser, J., Kalberlah, C., Haynes, J. D., & Friederici, A. D. (2012). Dissociable neural imprints of perception and grammar in auditory functional imaging. *Human Brain Mapping, 33*(3), 584-595.

Hill, H., Bruce, V., & Akamatsu, S. (1995). Perceiving the sex and race of faces: the role of shape and colour. *Proceedings of the Royal Society of London. Series B: Biological Sciences, 261*, 367-373.

Hillenbrand, J. M., & Clark, M. J. (2009). The role of f0 and formant frequencies in distinguishing the voices of men and women. *Attention, Perception, & Psychophysics, 71*(5), 1150-1166.

Hillenbrand J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Joural of the Acoustical Society of America, 97*, 3099–3111.

Hölldobler, B., Janssen, E., Bestmann, H. J., Kern, F., Leal, I. R., Oliveira, P. S., & König, W. A. (1996). Communication in the migratory termite-hunting ant *Pachycondyla (= Termitopone) marginata* (Formicidae, Ponerinae). *Journal of Comparative Physiology A, 178*, 47–53.

Hollien, H., Green, R., & Massey, K. (1994). Longitudinal research on adolescent voice change in males. *The Journal of the Acoustical Society of America, 96*(5), 2646-2654.

Houston-Price, C., & Nakai, S. (2004). Distinguishing novelty and familiarity effects in infant preference procedures, *Infant Child Development, 13*(4), 341–348.

Howard, I. P., & Templeton, W. B. (1966). *Human Spatial Orientation*. Wiley, New York, NY, USA.

Huber, J. E., Stathopoulos, E. T., Curione, G. M., Ash, T. A., & Johnson, K. (1999). Formants of children, women, and men: The effects of vocal intensity variation. *Journal of the Acoustical Society of America, 106*(3), 1532–1542.

Hughes, M. (1996). The function of concurrent signals: Visual and chemical communication in snapping shrimp. *Animal Behaviour, 52*, 247–257.

Hunter, M. A., & Ames, E. W. (1988). A multifactor model of infant preferences for novel and familiar stimuli. *Advances in Infancy Research*, *5*, 69–95.

Imai, M., & Kita, S. (2014). The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 369*(1651), 20130298.

Imai, M., Kita, S., Nagumo, M., & Okada, H. (2008). Sound symbolism facilitates early verb learning, *Cognition, 109*, 54–65.

Ingle, D., & Cook, J. (1977). The effect of viewing distance upon size preference of frogs for prey. *Vision research, 17*(9), 1009-1013.

Izumi, A., & Kojima, S. (2004). Matching vocalisations to vocalizing faces in a chimpanzee (*Pan troglodytes*). *Animal Cognition, 7*, 179–184.

Jacob, S., Rieucau, G., & Heeb, P. (2011). Multimodal begging signals reflect independent indices of nestling condition in European starlings. *Behavioral Ecology, 22*, 1249–1255**.**

Jacquemot, C., Pallier, C., LeBihan, D., Dehaene, S., & Dupoux, E. (2003). Phonological grammar shapes the auditory cortex: a functional magnetic resonance imaging study. *The Journal of Neuroscience*, *23*(29), 9541-9546.

Jakovcevic, A., Mustaca, A., & Bentosela, M. (2012). Do more sociable dogs gaze longer to the human face than less sociable ones? *Behavioural Processes*, *90*, 217-222.

Jerger, J., & Martin, J. (2004). Hemispheric asymmetry of the right ear advantage in dichotic listening. *Hearing Research, 198*(1), 125-136.

Jitsumori, M., & Delius, J. D. (2001). Object recognition and object categorization in animals. In T. Matsuzawa (Ed.), *Primate Origins of Human Cognition and Behavior* (pp. 269-293). Japan: Springer.

Johnstone, R. A. (1996). Multiple displays in animal communication: Backup signals and multiple messages. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 351*(1337), 329–338.

Johnston, R. E., & Bullock, T. A. (2001). Individual recognition by use of odours in golden hamsters: the nature of individual representations. *Animal Behaviour, 61*(3), 545–557.

Joly, O., Ramus, F., Pressnitzer, D., Vanduffel, W., & Orban, G. A. (2012). Interhemispheric differences in auditory processing revealed by fMRI in awake rhesus monkeys. *Cerebral Cortex, 22*(4), 838-853.

Jordan, K. E., Brannon, E. M., Logothetis, N. K., & Ghazanfar, A. A. (2005). Monkeys match the number of voices they hear to the number of faces they see, *Current Biology, 15*, 1034–1038.

Juslin, P. N., & Laukka, P. (2001). Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion. *Emotion, 1*(4), 381-412.

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code?. *Psychological Bulletin, 129*(5), 770-814.

Kahane, J. C. (1982). Growth of the human prepubertal and pubertal larynx. *Journal of Speech, Language, and Hearing Research*, *25*(3), 446-455.

Kaminski, J. (2009). Dogs (Canis familiaris) are adapted to receive human communication. In A. Berthoz & Y. Christen (Eds.), *Neurobiology of "Umwelt"* (pp. 103-107). Berlin: Springer.

Kaminski, J., Call, J., & Fischer, J. (2004). Word learning in a domestic dog: evidence for "fast mapping". *Science*, *304*(5677), 1682-1683.

Kaminski, J., & Nitzschner, M. (2013). Do dogs get the point? A review of dog–human communication ability. *Learning and Motivation*, *44*(4), 294-302.

Kaminski, J., Schulz, L., & Tomasello, M. (2012). How dogs know when communication is intended for them. *Developmental Science, 15*(2), 222-232.

Kaminski, J., Tempelmann, S., Call, J., & Tomasello, M. (2009). Domestic dogs comprehend human communication with iconic signs. *Developmental Science, 12*(6), 831-837.

Kent, R. D., & Vorperian, H. K. (1995). *Development of the craniofacial-oral-laryngeal anatomy.* San Diego: Singular Publishing Group.

Kikuchi, Y., Horwitz, B., & Mishkin, M. (2010). Hierarchical auditory processing directed rostrally along the monkey's supratemporal plane. *The Journal of Neuroscience*, *30*(39), 13021-13030.

Kimura, D. (1961). Cerebral dominance and the perception of verbal stimuli. *Canadian Journal of Psychology*, *15*(3), 166-171.

Kimura, D. (2011). From ear to brain. *Brain and Cognition*, *76*(2), 214-217.

Kingdon, J. (1974). *East African Mammals. Vol II, Part A: Insectivores and Bats*. Academic Press, New York, NY, USA.

Kirchhofer, K. C., Zimmermann, F., Kaminski, J., & Tomasello, M. (2012). Dogs (*Canis familiaris*), but not chimpanzees (*Pan troglodytes*), understand imperative pointing. *PloS one*, *7*(2), e30913.

Kis, A., Topál, J., Gácsi, M., Range, F., Huber, L., Miklósi, Á., & Virányi, Z. (2012). Does the A-not-B error in adult pet dogs indicate sensitivity to human communication?. *Animal Cognition*, *15*(4), 737-743.

Kisilevsky, B. S., Hains, S. M., Lee, K., Xie, X., Huang, H., Ye, H. H., Zhang, K. & Wang, Z. (2003). Effects of experience on fetal voice recognition. *Psychological Science*, *14*(3), 220-224.

Köhler, W. (1929). *Gestalt Psychology*. Liveright, New York, NY,USA.

Köhler, W. (1947). *Gestalt Psychology* (2nd ed.). New York: Liveright.

Kojima, S., Izumi, A., & Ceugniet, M. (2003). Identification of vocalizers by pant hoots, pant grunts and screams in a chimpanzee, *Primates, 44*, 225–230.

Koolhaas, J.M., Korte, S.M., De Boer, S.F., Van Der Vegt, B.J., Van Reenen, C.G., Hopster, H., . . . Blokhuis, H.J. (1999). Coping styles in animals: current status in behavior and stress-physiology. *Neuroscience and Biobehavioral Reviews, 23*, 925-935.

Kondo, N., Izawa, E-I., & Watanabe, S. (2012). Crows cross-modally recognise group members but not non-group members, *Proceedings of the Royal Society B: Biological Sciences, 279*, 1937–1942.

Kriegstein, K. V., & Giraud, A. L. (2004). Distinct functional substrates along the right superior temporal sulcus for the processing of voices. *Neuroimage*, *22*(2), 948-955.

Kriengwatana, B., Escudero, P., & ten Cate, C. (2014). Revisiting vocal perception in non-human animals: a review of vowel discrimination, speaker voice recognition, and speaker normalization. *Frontiers in Psychology, 5*, 1543.

Krusche, P., Uller, C., & Dicke, U. (2010). Quantity discrimination in salamanders. *Journal of Experimental Biology, 213*, 1822–1828.

Kulahci, I. G., & Ghazanfar, A. A. (2013). Multisensory recognition in vertebrates (especially primates), in: *Integrating Face and Voice in Person Perception*, P. Belin, S. Campanella & T. Ethofer (Eds), pp. 3–27, Springer, New York, NY, USA.

Kulachi, I. G., Drea, C. M., Rubenstein, D. I., & Ghazanfar, A. A. (2014). Individual recognition through olfactory–auditory matching in lemurs, *Proceedings of the Royal Society B: Biological Sciences, 281*, 20140071.

Laitman, J. T., & Crelin, E. S. (1976). Postnatal development of the basicranium and vocal tract region in man. In *Symposium on Development of the Basicranium* (pp. 206-219). Washington DC: US Government Printing Office.

Lakatos, G., Soproni, K., Dóka, A., & Miklósi, Á. (2009). A comparative approach to dogs'(Canis familiaris) and human infants' comprehension of various forms of pointing gestures. *Animal Cognition*, *12*(4), 621-631.

Lass, N. J., Hughes, K. R., Bowyer, M. D., Waters, L. T., & Bourne, V. T. (1976). Speaker sex identification from voiced, whispered, and filtered isolated vowels. *The Journal of the Acoustical Society of America*, *59*(3), 675-678.

Lattner, S., Meyer, M. E., & Friederici, A. D. (2005). Voice perception: sex, pitch, and the right hemisphere. *Human Brain Mapping*, *24*(1), 11-20.

Lavner, Y., Gath, I., & Rosenhouse, J. (2000). The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels. *Speech Communication*, *30*(1), 9-26.

Lee, S., Potamianos, A., & Narayanan, S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America*, *105*(3), 1455-1468.

Leliveld, L. M. C., Scheumann, M., & Zimmermann, E. (2010). Effects of caller characteristics on auditory laterality in an early primate (*Microcebus murinus*). *PLoS One*, *5*(2), e9031.

Lemasson, A., Koda, H., Kato, A., Oyakawa, C., Blois-Heulin, C., & Masataka, N. (2010). Influence of sound specificity and familiarity on Japanese macaques' (*Macaca fuscata*) auditory laterality. *Behavioural Brain Research, 208,* 286–289.

LePrell, C. G., Hauser, M. D., & Moody, D. B. (2002). Discrete or graded variation within rhesus monkey screams? Psychophysical experiments on classification. *Animal Behaviour, 63*(1), 47-62.

Lewkowicz, D. J., & Ghazanfar, A. A. (2006). The decline of cross-species intersensory perception in human infants. *Proceedings of the National Academy of Sciences, 103*, 6771–6774.

Lewkowicz, D.J., & Ghazanfar, A. A. (2009). The emergence of multisensory systems through perceptual narrowing. *Trends in Cognitive Sciences, 13*, 470–478.

Lewkowicz, D. J., & Turkewitz, G. (1980). Cross-modal equivalence in early infancy: Auditory–visual intensity matching. *Developmental Psychology, 16*, 597–607.

Lewkowicz, D. J., Leo, I., & Simion, F. (2010). Intersensory perception at birth: Newborns match nonhuman primate faces and voices. *Infancy, 15*, 46–60.

Liebenthal, E., Binder, J. R., Spitzer, S. M., Possing, E. T., & Medler, D. A. (2005). Neural substrates of phonemic perception. *Cerebral Cortex*, *15*(10), 1621-1631.

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

Lieberman, P. (1988). *Speech physiology, speech perception, and acoustic phonetics.* Cambridge: Cambridge University Press.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*(5), 358.

Lingle, S., & Riede, T. (2014). Deer mothers are sensitive to infant distress vocalizations of diverse mammalian species. *The American Naturalist*, *184*(4), 510-522.

Lingle, S., Wyman, M. T., Kotrba, R., Teichroeb, L. J., & Romanow, C. A. (2012). What makes a cry a cry? A review of infant distress vocalizations. *Current Zoology*, *58*(5), 698-726.

Locke, N. M. (1938). Some Factors in Size-Constancy. *The American Journal of Psychology*, 514-520.

Lore, R. K., & Eisenberg, F. B. (1986). Avoidance reactions of domestic dogs to unfamiliar male and female humans in a kennel setting. *Applied Animal Behaviour Science, 15*(3), 262–266.

Lourenco, S. F., & Longo, M. R. (2010). General magnitude representation in human infants, *Psychological Science, 21*, 873–881.

Ludwig, V. U., Adachi, I., & Matsuzawa, T. (2011). Visuoauditory mappings between high luminance and high pitch are shared by chimpanzees (*Pan troglodytes*) and humans, *Proceedings of the National Academy of Sciences, 108*, 20661–20665.

MacNeilage, P. F., Rogers, L. J., & Vallortigara, G. (2009). Origins of the left & right brain. *Scientific American, 301*(1), 60-67.

Madden, J. R., Kunc, H. J. P., English, S., & Clutton-Brock, T. H. (2009). Why do meerkat pups stop begging? *Animal Behaviour, 78*, 85–89.

Maier, J. X., Neuhoff, J. G., Logothetis, N. K., & Ghazanfar, A. A. (2004). Multisensory integration of looming signals by rhesus monkeys. *Neuron, 43*, 177–181.

Mariti, C., Gazzano, A., Moore, J. L., Baragli, P., Chelli, L., & Sighieri, C. (2012). Perception of dogs' stress by their owners. *Journal of Veterinary Behavior: Clinical Applications and Research*, *7*(4), 213-219.

Markman, E. M., & Abelev, M. (2004). Word learning in dogs?. *Trends in Cognitive Sciences*, *8*(11), 479-481.

Marks, L. E. (1978). *The Unity of the Senses: Interrelations among the Modalities*. Academic Press, New York, NY, USA.

Marks, L. E. (1989). On cross-modal similarity: The perceptual structure of pitch, loudness, and brightness. *Journal of Experimental Psychology: Human Perception & Performance, 15*, 586–602

Marks, L. E. (2000). Synesthesia, in: *Varieties of Anomalous Experience: Examining the Scientific Evidence*, E. Cardeña, S. J. Lynn and S. C. Krippner (Eds), pp. 121–149, American Psychological Association, Washington, DC, USA.

Marks, L. E., Szczesiul, R., & Ohlott, P. (1986). On the cross-modal perception of intensity. *Journal of Experimental Psychology: Human Perception & Performance, 12*, 517.

Marshall-Pescini, S., Prato-Previde, E., & Valsecchi, P. (2011). Are dogs (Canis familiaris) misled more by their owners than by strangers in a food choice task?. *Animal Cognition, 14*(1), 137-142.

Martinez, L., & Matsuzawa, T. (2009). Auditory-visual intermodal matching based on individual recognition in a chimpanzee (*Pan troglodytes*), *Animal Cognition, 12*, 71–85.

Marzoli, D., & Tommasi, L. (2009). Side biases in humans (*Homo sapiens*): three ecological studies on hemispheric asymmetries. *Naturwissenschaften, 96*(9), 1099-1106.

Masataka, N. (1994). Lack of correlation between body size and frequency of vocalizations in young female Japanese macaques (*Macaca fuscata*). *Folia Primatologica, 63,* 115–118.

Mateo, J. M. (1996). The development of alarm-call response behaviour in free-living juvenile Belding's ground squirrels. *Animal Behaviour, 52*(3), 489-505.

Maurer, D., Stager, C.L., & Mondloch, C. J. (1999). Cross-modal transfer of shape is difficult to demonstrate in one-month-olds, *Child Development, 70,* 1047–1057.

Maynard-Smith, J., & Harper, D. (2003). *Animal signals*. New York: Oxford University Press.

McComb, K., Shannon, G., Sayialel, K. N., & Moss, C. (2014). Elephants can determine ethnicity, gender, and age from acoustic cues in human voices. *Proceedings of the National Academy of Sciences, 111*(14), 5433-5438.

McGettigan, C., Evans, S., Rosen, S., Agnew, Z. K., Shah, P., & Scott, S. K. (2012). An application of univariate and multivariate approaches in fMRI to quantifying the hemispheric lateralization of acoustic and linguistic processes. *Journal of Cognitive Neuroscience*, *24*(3), 636-652.

McGettigan, C., Faulkner, A., Altarelli, I., Obleser, J., Baverstock, H., & Scott, S. K. (2012). Speech comprehension aided by multiple modalities: Behavioural and neural interactions. *Neuropsychologia, 50*(5), 762-776.

McGettigan, C., & Scott, S. K. (2012). Cortical asymmetries in speech perception: what's wrong, what's right and what's left?. *Trends in Cognitive Sciences, 16*(5), 269-276.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746–748.

McKinley, J., & Sambrook, T. D. (2000). Use of human-given cues by domestic dogs (*Canis familiaris)* and horses (*Equus caballus*). *Animal Cognition, 3*(1), 13-22.

Meck, W. H., & Church, R. M. (1983). A mode control model of counting and timing processes. *Journal of Experimental Psychology: Animal Behavior Processes, 9,* 320–334.

Meltzoff, A. N., & Borton, R. W. (1979). Intermodal matching by human neonates. *Nature, 282,* 403–404.

Merchant, H., & Honing, H. (2013). Are non-human primates capable of rhythmic entrainment? Evidence for the gradual audiomotor evolution hypothesis. *Frontiers in Neuroscience, 7*, 274.

Merola, I., Prato-Previde, E., Lazzaroni, M., & Marshall-Pescini, S. (2014). Dogs' comprehension of referential emotional expressions: familiar people and familiar emotions are easier. *Animal Cognition*, *17*(2), 373-385.

Merola, I., Prato-Previde, E., & Marshall-Pescini, S. (2012). Dogs' social referencing towards owners and strangers. *PLoS ONE*, e47653.

Merritt, D. J., Casasanto, D., & Brannon, E. M. (2010). Do monkeys think in metaphors? Representations of space and time in monkeys and humans. *Cognition, 117*, 191–202.

Meyer, J. (2008). Typology and acoustic strategies of whistled languages: Phonetic comparison and perceptual cues of whistled vowels. *Journal of the International Phonetic Association*, *38*(01), 69-94.

Miklósi, Á. (2009). Evolutionary approach to communication between humans and dogs. *Veterinary Research Communications*, *33*(1), 53-59.

Miklósi, Á., Polgárdi, R., Topál, J., & Csányi, V. (1998). Use of experimenter-given cues in dogs. *Animal Cognition*, *1*(2), 113-121.

Miklósi, Á., & Soproni, K. (2006). A comparative analysis of animals' understanding of the human pointing gesture. *Animal Cognition*, *9*(2), 81-93.

Miklósi, Á., Topál, J., & Csányi, V. (2004). Comparative social cognition: what can dogs teach us?. *Animal Behaviour*, *67*(6), 995-1004.

Miller, P.E., & Murphy, C.J. (1995). Vision in dogs. *Journal-American Veterinary Medical Association*, *207*(12), 1623-1634.

Mills, D. S., Fukuzawa, M., & Cooper, J. J. (2005). The effect of emotional content of verbal commands on the response of dogs. In *Current issues and research in veterinary behavioural medicine—papers presented at the 5th international veterinary behavior meeting, Purdue University Press, West Lafayette* (pp. 217-220).

Mitchell, R. L., Elliott, R., Barry, M., Cruttenden, A., & Woodruff, P. W. (2003). The neural response to emotional prosody, as revealed by functional magnetic resonance imaging. *Neuropsychologia*, *41*(10), 1410-1421.

Moller, A. P., & Pomiankowski, A. (1993). Why have birds got multiple sexual ornaments? *Behavioral Ecology & Sociobiology, 32*, 167–176.

Monaghan, P., Mattock, K., & Walker, P. (2012). The role of sound symbolism in language learning, *Joural of Experimental Psychology: Learning, Memory & Cognition, 38,* 1152–1164.

Moore, B. C. J. (1989). *An Introduction to the Psychology of Hearing.* Academic Press, London, UK.

Morgan, M. L., DeAngelis, G. C., & Angelaki, D. E. (2008). Multisensory integration in macaque visual cortex depends on cue reliability, *Neuron, 59*, 662–673.

Morton, E. S. (1977). On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *American Naturalist, 111*(981), 855-869.

Möttönen, R., Calvert, G. A., Jääskeläinen, I. P., Matthews, P. M., Thesen, T., Tuomainen, J., & Sams, M. (2006). Perceiving identical sounds as speech or non-speech modulates activity in the left posterior superior temporal sulcus. *Neuroimage*, *30*(2), 563-569.

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, *9*(5), 453-467.

Mummery, C. J., Ashburner, J., Scott, S. K., & Wise, R. J. (1999). Functional neuroimaging of speech perception in six normal and two aphasic subjects. *The Journal of the Acoustical Society of America*, *106*(1), 449-457.

Murai, C., Kosugi, D., Tomonaga, M., Tanaka, M., Matsuzawa, T., & Itakura, S. (2005). Can chimpanzee infants (*Pan troglodytes*) form categorical representations in the same manner as human infants (*Homo sapiens*)? *Developmental Science*, *8*(3), 240-254.

Murai, C., Tomonaga, M., Kamegai, K., Terazawa, N., & Yamaguchi, M.K. (2004). Do infant Japanese macaques (*Macaca fuscata*) categorize objects without specific training? *Primates*, *45*(1), 1-6.

Nagumo, M., Imai, M., Kita, S., Haryu, E., & Kajikawa, S. (2006). Sound iconicity bootstraps verb meaning acquisition, in: *XVth International Conference of Infant Studies*, Kyoto, Japan. [Cited in: Imai, M., Kita, S., Nagumo, M., & Okada, H. (2008). Sound symbolism facilitates early verb learning. *Cognition, 109*, 54–65.]

Narain, C., Scott, S. K., Wise, R. J., Rosen, S., Leff, A., Iversen, S. D., & Matthews, P. M. (2003). Defining a left-lateralized response specific to intelligible speech using fMRI. *Cerebral Cortex*, *13*(12), 1362-1368.

Narins, P. M., Hödl, W., & Grabul, D. S. (2003). Bimodal signal requisite for agonistic behavior in a dart-poison frog, *Epipedobates femoralis*. *Proceedings of the National Academy of Science, 100*, 577–580.

Negus, V.E. (1949). *The comparative anatomy and physiology of the larynx*. New York: Hafner Publishing Company.

Neuhoff, J. G., & McBeath, M. K. (1996). The Doppler illusion: The influence of dynamic intensity change on perceived pitch. *Journal of Experimental Psychology: Human Perception & Performance, 22*, 970–985.

Obleser, J., & Kotz, S. A. (2009). Expectancy constraints in degraded speech modulate the language comprehension network. *Cerebral Cortex*, bhp128.

Obleser, J., Wise, R. J., Dresner, M. A., & Scott, S. K. (2007). Functional integration across brain regions improves speech perception under adverse listening conditions. *Journal of Neuroscience*, *27*(9), 2283-2289.

Ocklenburg, S., Ströckens, F., & Güntürkün, O. (2013). Lateralisation of conspecific vocalisation in non-human vertebrates. *Laterality: Asymmetries of Body, Brain and Cognition*, *18*(1), 1-31.

Ohala, J. J. (1975). The temporal regulation of speech., in: *Auditory Analysis and Perception of Speech*, G. Fant and M. A. A Tatham (Eds), pp. 431–453, Elsevier, Amsterdam, The Netherlands.

Ohala, J. J. (1994). "The frequency codes underlies the sound symbolic use of voice pitch," In L. Hinton, J. Nichols, & J. J. Ohala (Eds.), *Sound symbolism* (pp. 325-347). Cambridge: Cambridge University Press.

O'Toole, A.J., Deffenbacher, K.A., Valentin, D., McKee, K., Huff, D., & Abdi, H. (1998). The perception of face gender: The role of stimulus structure in recognition and classification. *Memory & Cognition*, *26*(1), 146-160.

Owren, M. J., & Rendall, D. (1997). An affect-conditioning model of nonhuman primate vocal signaling. In D. H. Owings, M. D. Beecher & N.S. Thompson (Eds.), *Communication* (pp. 299-346). New York: Springer.

Owren, M. J., & Rendall, D. (2001). Sound on the rebound: bringing form and function back to the forefront in understanding nonhuman primate vocal signaling. *Evolutionary Anthropology: Issues, News, and Reviews*, *10*(2), 58-71.

Palleroni, A., & Hauser, M. (2003). Experience-dependent plasticity for auditory processing in a raptor. *Science*, *299*(5610), 1195-1195.

Parault, S. J., & Parkinson, M. (2008). Sound symbolic word learning in the middle grades. *Contemporary Educational Psychology, 33*, 647–671.

Parise, C. V., Knorre, K., & Ernst, M. O. (2014). Natural auditory scene statistics shapes human spatial hearing, *Proceedings of the National Academy of Sciences, 111*, 6104–6108.

Parise, C. V., & Pavani, F. (2011). Evidence of sound symbolism in simple vocalizations. *Experimental Brain Research*, *214*(3), 373-380.

Parise, C. V., & Spence, C. (2009). 'When birds of a feather flock together': Synesthetic correspondences modulate audiovisual integration in non-synesthetes, *PLoS One, 4*, e5664.

Parise, C. V., & Spence, C. (2013). Audiovisual cross-modal correspondences in the general population, in: *The Oxford Handbook of Synesthesia*, J. Simner and E. M. Hubbard (Eds), pp 790–815, Oxford University Press, Oxford, UK.

Parise, C. V., Spence, C., & Ernst, M. O. (2012). When correlation implies causation in multisensory integration, *Current Biology, 22*, 46–49.

Parr, L. A. (2004). Perceptual biases for multimodal cues in chimpanzee (*Pan troglodytes*) affect recognition, *Animal Cognition, 7*, 171–178.

Partan, S. R. (2002). Single and multichannel facial composition: Facial expressions and vocalizations of rhesus macaques (*Macaca mulata*). *Behaviour, 139*, 993–1027.

Partan, S., & Marler, P. (1999). Communication goes multimodal. *Science, 283*(5406), 1272–1273.

Pascalis, O., & De Haan, M. (2003). Recognition memory and novelty preference: What model, in: *Progress in Infancy Research, Vol. 3*, H. Hayne & J. Fagen (Eds), pp. 95–120, Psychology Press, New York, NY, USA.

Pastore, N. (1958). Form perception and size constancy in the duckling. *Journal of Psychology*, *45*(2), 259-261.

Peña, M., Mehler, J., & Nespor, M. (2011). The role of audiovisual processing in early conceptual development. *Psychological Science, 22*(11), 1419-1421.

Penn, D. J., Oberzaucher, E., Grammer, K., Fischer, G., Soini, H. A., Wiesler, D., Novotny, M.V., Dixon, S.J., Xu, Y., & Brereton, R. G. (2007). Individual and gender fingerprints in human body odour. *Journal of the Royal Society Interface*, *4*(13), 331-340.

Pepperberg, I. M. (1981). Functional vocalizations by an African Grey parrot (Psittacus erithacus). *Zeitschrift für Tierpsychologie, 55*(2), 139-160.

Perdue, B. M., Talbot, C. F., Stone, A., & Beran, M. J. (2012). Putting the elephant back in the herd: Elephant relative quantity judgments match those of other species. *Animal Cognition,15,* 955–961.

Pet Food Manufacturers Association. (2014). *Pet Population Report*. Retrieved from http://www.pfma.org.uk/pet-population-2014/

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24,* 175–184.

Petersen, M. R., Beecher, M. D., Moody, D. B., & Stebbins, W. C. (1978). Neural lateralization of species-specific vocalizations by Japanese macaques (*Macaca fuscata*). *Science, 202*(4365), 324-327.

Petkov, C. I., Kayser, C., Steudel, T., Whittingstall, K., Augath, M., & Logothetis, N. K. (2008). A voice region in the monkey brain. *Nature Neuroscience, 11*(3), 367-374.

Pilley, J. W., & Reid, A. K. (2011). Border collie comprehends object names as verbal referents. *Behavioural Processes, 86*(2), 184-195.

Pisa, P. E., & Agrillo, C. (2009). Quantity discrimination in felines: a preliminary investigation of the domestic cat (*Felis silvestris catus*), *Journal of Ethology, 27,* 289–293.

Pisanski, K. (2014). *Human vocal communication of body size* (Doctoral dissertation).

Pisanski, K., Fraccaro, P. J., Tigue, C. C., O'Connor, J. J., Röder, S., Andrews, P. W., Fink, B., DeBruine, L. M., Jones, B. C. & Feinberg, D. R. (2014). Vocal indicators of body size in men and women: a meta-analysis. *Animal Behaviour, 95,* 89-99.

Pisanski, K., & Rendall, D. (2011). The prioritization of voice fundamental frequency or formants in listeners' assessments of speaker size, masculinity, and attractiveness. *The Journal of the Acoustical Society of America, 129*(4), 2201-2212.

Plotsky, K., Rendall, D., Riede, T., & Chase, K. (2013). Radiographic analysis of vocal tract length and its relation to overall body size in two canid species. *Journal of Zoology, 291*(1), 76-86.

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. *Speech Communication, 41*(1), 245-255.

Pongrácz, P., Miklósi, Á., & Csányi, V. (2001). Owner's beliefs on the ability of their pet dogs to understand human verbal communication: A case of social understanding. *Current Psychology of Cognition*, *20*(1/2), 87-108.

Pongrácz, P., Miklósi, Á., Timár-Geng, K., & Csányi, V. (2004). Verbal attention getting as a key factor in social learning between dog (Canis familiaris) and human. *Journal of Comparative Psychology*, *118*(4), 375.

Poremba, A., Bigelow, J., & Rossi, B. (2013). Processing of communication sounds: contributions of learning, memory, and experience. *Hearing Research*, *305*, 31-44.

Poremba, A., Malloy, M., Saunders, R. C., Carson, R. E., Herscovitch, P., & Mishkin, M. (2004). Species-specific calls evoke asymmetric activity in the monkey's temporal poles. *Nature*, *427*(6973), 448-451.

Povinelli, D. J., Reaux, J. E., Bierschwale, D. T., Allain, A. D., & Simon, B. B. (1997). Exploitation of pointing as a referential gesture in young children, but not adolescent chimpanzees. *Cognitive Development*, *12*(4), 423-461.

Prato-Previde, E., Fallani, G., & Valsecchi, P. (2005). Gender differences in owners interacting with pet dogs: an observational study. *Ethology*, *112*(1), 64-73.

Pratt, C. C. (1930). The spatial character of high and low tones. *Journal of Experimental Psychology, 13*, 278–285.

Proops, L., & McComb, K. (2012). Cross-modal individual recognition in domestic horses (*Equus caballus*) extends to familiar humans. *Proceedings of the Royal Society of London B: Biological Sciences*, *279*(1741), 3131-3138.

Proops, L., McComb, K., & Reby, D. (2009). Cross-modal individual recognition in domestic horses (*Equus caballus*). *Proceedings of the National Academy of Sciences, 106*, 947–951.

Puts, D. A., Gaulin, S. J., & Verdolini, K. (2006). Dominance and the evolution of sexual dimorphism in human voice pitch. *Evolution & Human Behavior, 27*, 283–296.

Puts, D. A., Apicella, C. L., & Cardenas, R. A. (2012). Masculine voices signal men's threat potential in forager and industrial societies. *Proceedings of the Royal Society of London B: Biological Sciences, 279*, 601–609.

Ramachandran, V. S., & Hubbard, E. M. (2001). Synaesthesia — A window into perception, thought and language. *Journal of Consciousness Studies, 8*, 3–34.

Ramos, D., & Ades, C. (2012). Two-item sentence comprehension by a dog (*Canis familiaris*). *PloS One*, *7*(2), e29689.

Range, F., Aust, U., Steurer, M., & Huber, L. (2008). Visual categorization of natural stimuli by domestic dogs. *Animal Cognition*, *11*(2), 339–347.

Range, F., & Virányi, Z. (2014). Tracking the evolutionary origins of dog-human cooperation: the "Canine Cooperation Hypothesis". *Frontiers in Psychology*, *5*, 1582, 1-10.

Rappolt, G.A., John, J., & Thompson, N.S. (1979). Canine Responses to Familiar and Unfamiliar Humans. *Aggressive Behavior*, *5*(2), 155-161.

Ratcliffe, V.F., McComb, K. & Reby, D. (2014). Cross-modal discrimination of human gender by domestic dogs, *Animal Behaviour, 91*, 127-135.

Ratcliffe, V.F., & Reby, D. (2014). Orienting asymmetries in dogs' responses to different communicatory components of human speech, *Current Biology, 24*, 2908-2912.

Rauschecker, J. P., & Scott, S. K. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nature Neuroscience*, *12*(6), 718-724.

Reby, D., & McComb, K. (2003). Anatomical constraints generate honesty: acoustic cues to age and weight in the roars of red deer stags. *Animal Behaviour, 65*(3), 519-530.

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. *Proceedings of the Royal Society of London B: Biological Sciences*, *272*(1566), 941-947.

Reinholz-Trojan, A., Włodarczyk, E., Trojan, M., Kulczyński, A., & Stefańska, J. (2012). Hemispheric specialization in domestic dogs (*Canis familiaris*) for processing different types of acoustic stimuli. *Behavioural processes*, *91*(2), 202-205.

Reisner, I. R., & Shofer, F. S. (2008). Effects of gender and parental status on knowledge and attitudes of dog owners regarding dog aggression toward children. *Journal of the American Veterinary Medical Association*, *233*(9), 1412-1419.

Rendall, D., Kollias, S., Ney, C., & Lloyd, P. (2005). Pitch (F0) and formant profiles of human vowels and vowel-like baboon grunts: The role of vocalizer body size and

voice-acoustic allometry. *Journal of the Acoustical Society of America*, *177*, 944–955.

Rendall, D., Owren, M. J., Weerts, E., & Hienz, R. D. (2004). Sex differences in the acoustic structure of vowel-like grunt vocalizations in baboons and their perceptual discrimination by baboon listeners. *Journal of the Acoustical Society of America*, *115*(1), 411-421.

Rendall, D., Vokey, J. R., & Nemeth, C. (2007). Lifting the curtain on the Wizard of Oz: biased voice-based impressions of speaker size. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(5), 1208.

Riede, T., & Fitch, T. (1999). Vocal tract length and acoustics of vocalization in the domestic dog (Canis familiaris). *Journal of Experimental Biology*, *202*(20), 2859-2867.

Riedel, J., Schumann, K., Kaminski, J., Call, J., & Tomasello, M. (2008). The early ontogeny of human–dog communication. *Animal Behaviour*, *75*(3), 1003-1014.

Rogers, L.J. (1997). Early experiential effects on laterality: research on chicks has relevance to other species. *Laterality*, *2*(3-4), 199–219.

Rosado, B., García-Belenguer, S., León, M., & Palacio, J. (2009). A comprehensive study of dog bites in Spain, 1995-2004. *The Veterinary Journal*, *179*(3), 383-391.

Rosch, E., Mervis, C.B., Gray, W.D., Johnson, D.M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382-439.

Rosen, S., Wise, R. J., Chadha, S., Conway, E. J., & Scott, S. K. (2011). Hemispheric asymmetries in speech perception: sense, nonsense and modulations. *PLoS One*, *6*(9), e24672.

Rosenzweig, M. R. (1951). Representations of the two ears at the auditory cortex. *American Journal of Physiology--Legacy Content*, *167*(1), 147-158.

Rossano, F., Nitzschner, M., & Tomasello, M. (2014). Domestic dogs and puppies can use human voice direction referentially. *Proceedings of the Royal Society of London B: Biological Sciences*, *281*(1785), 20133201.

Rowe, C. (1999). Receiver psychology and the evolution of multicomponent signals. *Animal Behaviour*, *58*, 921–931.

Rugaas, T. (2005). *On talking terms with dogs: calming signals*. Dogwise Publishing.

Rugani, R., Vallortigara, G., Priftis, K., & Regolin, L. (2015). Number-space mapping in the newborn chick resembles humans' mental number line. *Science*, *347*(6221), 534–536.

Rundus, A.S., Owings, D.H., Joshi, S.S., Chinn, E., & Giannini, N. (2007). Ground squirrels use an infrared signal to deter rattlesnake predation. *Proceedings of the National Academy of Sciences, 104,* 14372–14376.

Rusconi, E., Kwan, B., Giordano, B. L., Umilta, C., & Butterworth, B. (2006). Spatial representation of pitch height: the SMARC effect. *Cognition, 99,* 113–129.

Sapir, E. (1929). A study in phonetic symbolism. *Journal of Experimental Psychology*, *12*(3), 225-239.

Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences*, *107*(6), 2408-2412.

Savage-Rumbaugh, S., McDonald, K., Sevcik, R. A., Hopkins, W. D., & Rubert, E. (1986). Spontaneous symbol acquisition and communicative use by pygmy chimpanzees (Pan paniscus). *Journal of Experimental Psychology: General*, *115*(3), 211-235.

Savolainen, P., Zhang, Y. P., Luo, J., Lundeberg, J., & Leitner, T. (2002). Genetic evidence for an East Asian origin of domestic dogs. *Science*, *298*(5598), 1610-1613.

Scheider, L., Grassmann, S., Kaminski, J., & Tomasello, M. (2011). Domestic dogs use contextual information and tone of voice when following a human pointing gesture. *PLoS One*, *6*(7), e21676.

Scheider, L., Kaminski, J., Call, J., & Tomasello, M. (2013). Do domestic dogs interpret pointing as a command?. *Animal Cognition*, *16*(3), 361-372.

Scheumann, M., & Zimmermann, E. (2008). Sex-specific asymmetries in communication sound perception are not related to hand preference in an early primate. *BMC Biology, 6*(1), 3.

Schiff, W., Caviness, J. A., & Gibson, J. J. (1962). Persistent fear responses in rhesus monkeys to the optical stimulus of" looming". *Science, 136*(3520), 982-983.

Schirmer, A., & Kotz, S. A. (2006). Beyond the right hemisphere: brain mechanisms mediating vocal emotional processing. *Trends in Cognitive Sciences*, *10*(1), 24-30.

Schönwiesner, M., Rübsamen, R., & Von Cramon, D. Y. (2005). Hemispheric

asymmetry for spectral and temporal processing in the human antero-lateral auditory belt cortex. *European Journal of Neuroscience*, *22*(6), 1521-1528.

Schusterman, R. J., & Krieger, K. (1986). Artificial language comprehension and size transposition by a California sea lion (Zalophus californianus). *Journal of Comparative Psychology*, *100*(4), 348-355.

Scott, S. K., Blank, C. C., Rosen, S., & Wise, R. J. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain*, *123*(12), 2400-2406.

Scott, S. K., & Johnsrude, I. S. (2003). The neuroanatomical and functional organization of speech perception. *Trends in Neurosciences*, *26*(2), 100-107.

Scott, S. K., & McGettigan, C. (2013). Do temporal processes underlie left hemisphere dominance in speech perception?. *Brain and Language*, *127*(1), 36-45.

Scott, S. K., Rosen, S., Beaman, C. P., Davis, J. P., & Wise, R. J. (2009). The neural processing of masked speech: evidence for different mechanisms in the left and right temporal lobes. *The Journal of the Acoustical Society of America*, *125*(3), 1737-1743.

Scott, S. K., & Wise, R. J. (2004). The functional neuroanatomy of prelexical processing in speech perception. *Cognition*, *92*(1), 13-45.

Seyfarth, R. M., & Cheney, D. L. (1986). Vocal development in vervet monkeys. *Animal Behaviour*,*34*, 1640–1658.

Seyfarth, R. M., & Cheney, D. L. (2015). Social cognition. *Animal Behaviour, 103*, 191–202.

Seyfarth, R. M., Cheney, D. L., & Marler, P. (1980a). Vervet monkey alarm calls: Semantic communication in a free-ranging primate. *Animal Behaviour, 28*, 1070–1094.

Seyfarth, R. M., Cheney, D. L., & Marler, P. (1980b). Monkey responses to three different alarm calls: evidence of predator classification and semantic communication. *Science, 210*(4471), 801–803.

Simpson, A. P. (2009). Phonetic differences between male and female speech. *Language and Linguistics Compass*, *3*(2), 621-640.

Siniscalchi, M., Lusito, R., Sasso, R., & Quaranta, A. (2012). Are temporal features crucial acoustic cues in dog vocal recognition?. *Animal Cognition*, *15*(5), 815-821.

Siniscalchi, M., Quaranta, A., & Rogers, L. J. (2008). Hemispheric specialization in dogs for processing different acoustic stimuli. *PLoS One, 3*(10), e3349.

Skoglund, P., Ersmark, E., Palkopoulou, E., & Dalén, L. (2015). Ancient Wolf Genome Reveals an Early Divergence of Domestic Dog Ancestors and Admixture into High-Latitude Breeds. *Current Biology*, 25(11), 1515-1519

Sliwa, J., Duhamel, J. R., Pascalis, O., & Wirth, S. (2011). Spontaneous voice–face identity matching by rhesus monkeys for familiar conspecifics and humans. *Proceedings of the National Academy of Sciences*, *108*(4), 1735-1740.

Slutsky, D. A., & Recanzone, G. H. (2001). Temporal and spatial dependency of the ventriloquism effect. *Neuroreport, 12*(1), 7–10.

Smith, C. L., & Evans, C. S. (2008). Multimodal signaling in fowl, *Gallus gallus*. *Journal of Experimental Biology, 211*, 2052–2057.

Smith, D. R., & Patterson, R. D. (2005). The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. *The Journal of the Acoustical Society of America*, *118*(5), 3177-3186.

Smith, D. R., Walters, T. C., & Patterson, R. D. (2007). Discrimination of speaker sex and size when glottal-pulse rate and vocal-tract length are controlled. *The Journal of the Acoustical Society of America*, *122*(6), 3628-3639.

Sokolov, E. (1963). *Perception and Conditioned Reflex*. Pergamon, New York, NY, USA.

Somppi, S., Törnqvist, H., Hänninen, L., Krause, C., & Vainio, O. (2012). Dogs do look at images: eye tracking in canine cognition research. *Animal Cognition*, *15*(2), 163-174.

Spence, C. (2007). Audiovisual multisensory integration. *Acoustical Science and Technology, 28*, 61–70.

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception & Psychophysics*, *73*, 971–995.

Spence, C., & Deroy, O. (2012). Crossmodal correspondences: Innate or learned? *I-Perception, 3*, 316–318.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and Cognition*. Oxford: Blackwell.

Srinivasan, M., & Carey, S. (2010). The long and the short of it: On the nature and origin of functional overlap between representations of space and time. *Cognition, 116*, 217–241.

Stebbens, W. C. (1966). Auditory reaction times and the derivation of equal loudness contours for the monkey. *Journal of the Experimental Analysis of Behavior, 9,* 135–142.

Steels, L., & Belpaeme, T. (2005). Coordinating perceptually grounded categories through language: a case study for colour. *Behavioral and Brain Sciences, 28*(04), 469-529.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review, 64,* 153–181.

Taylor, R. C., Klein, B. A., Stein, J., & Ryan, M. J. (2011). Multimodal signal variation in space and time: How important is matching the signal with its signaler? *Journal of Experimental Biology,214,* 815–820.

Taylor, A. M., & Reby, D. (2010). The contribution of source–filter theory to mammal vocal communication research. *Journal of Zoology, 280*(3), 221-236.

Taylor, A. M., Reby, D., & McComb, K. (2008). Human listeners attend to size information in domestic dog growls. *The Journal of the Acoustical Society of America, 123*(5), 2903-2909.

Taylor, A. M., Reby, D., & McComb, K. (2011). Cross modal perception of body size in domestic dogs (Canis familiaris). *PLoS One, 6*(2), e17069-e17069.

Tedore, C., & Johnsen, S. (2014). Visual mutual assessment of size in male *Lyssomanes viridis* jumping spider contests. *Behavioral Ecology, 26,* 510–518.

Téglás, E., Gergely, A., Kupán, K., Miklósi, Á., & Topál, J. (2012). Dogs' gaze following is tuned to human communicative signals. *Current Biology, 22*(3), 209-212.

Tempelmann, S., Kaminski, J., & Tomasello, M. (2014). Do Domestic Dogs Learn Words Based on Humans' Referential Behaviour. *PloS One, 9*(3), e91014.

Teufel, C., Ghazanfar, A. A., & Fischer, J. (2010). On the relationship between lateralized brain function and orienting asymmetries. *Behavioral Neuroscience, 124*(4), 437-445.

Teufel, C., Hammerschmidt, K., & Fischer, J. (2007). Lack of orienting asymmetries in Barbary macaques: Implications for studies of lateralized auditory processing. *Animal Behaviour, 73,* 249–255.

Thalmann, O., Shapiro, B., Cui, P., Schuenemann, V. J., Sawyer, S. K., Greenfield, D. L., ... & Wayne, R. K. (2013). Complete mitochondrial genomes of ancient canids suggest a European origin of domestic dogs. *Science, 342*(6160), 871-874.

Thompson, J. T., Bissell, A. N., & Martins, E. P. (2008). Inhibitory interactions between multimodal behavioural responses may influence the evolution of complex signals. *Animal Behaviour, 76*, 113–121.

Thompson, B.L., & Waltz, J. (2010). Mindfulness and experiential avoidance as predictors of posttraumatic stress disorder avoidance symptom severity. *Journal of Anxiety Disorders*, *24*(4), 409-415.

Timneylf, B., & Keil, K. (1996). Horses are sensitive to pictorial depth cues. *Perception, 25*, 1121-1128.

Titze, I. R. (1989). On the relation between subglottal pressure and fundamental frequency in phonation. *The Journal of the Acoustical Society of America*, *85*(2), 901-906.

Titze, I.R. (1994). *Principles of voice production*. Englewood Cliffs, NJ: Prentice Hall.

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

Tomasello, M., Call, J., & Gluckman, A. (1997). Comprehension of novel communicative signs by apes and human children. *Child Development*, *68*(6), 1067-1080.

Topál, J., Gácsi, M., Miklósi, Á., Virányi, Z., Kubinyi, E., & Csányi, V. (2005). Attachment to humans: a comparative study on hand-reared wolves and differently socialized dog puppies. *Animal Behaviour*, *70*(6), 1367-1375.

Topál, J., Gergely, G., Erdőhegyi, Á., Csibra, G., & Miklósi, Á. (2009). Differential sensitivity to human communication in dogs, wolves, and human infants. *Science*, *325*(5945), 1269-1272.

Tschudin, A., Call, J., Dunbar, R. I., Harris, G., & van der Elst, C. (2001). Comprehension of signs by dolphins (Tursiops truncatus). *Journal of Comparative Psychology, 115*(1), 100-105

Tunturi, A. R. (1946). A study on the pathway from the medial geniculate body to the acoustic cortex in the dog. *American Journal of Physiology--Legacy Content*, *147*(2), 311-319.

Udell, M. A., Dorey, N. R., & Wynne, C. D. (2010). What did domestication do to dogs? A new account of dogs' sensitivity to human actions. *Biological Reviews*, *85*(2), 327-345.

Udell, M. A., Spencer, J. M., Dorey, N. R., & Wynne, C. D. (2012). Human-socialized wolves follow diverse human gestures… and they may not be alone. *International Journal of Comparative Psychology, 25*(2) 97-117.

Uetz, G. W., & Roberts, J. A. (2002). Multisensory cues and multimodal communication in spiders: insights from video/audio playback studies. *Brain, Behavior & Evolution, 59,* 222–230.

Vachon, F., Hughes, R. W., & Jones, D. M. (2012). Broken expectations: Violation of expectancies, not novelty, captures auditory attention. *Journal of Experimental Psychology: Learning, Memory & Cognition, 38,* 164–177.

Vallortigara, G., & Rogers, L. J. (2005). Survival with an asymmetrical brain: advantages and disadvantages of cerebral lateralization. *Behavioral and Brain Sciences, 28*(4), 575-588.

Vallortigara, G., Snyder, A., Kaplan, G., Bateson, P., Clayton, N. S., & Rogers, L. J. (2008). Are animals autistic savants? *PLoS Biology, 6,* 208–214.

Van der Zee, E., Zulch, H., & Mills, D. (2012). Word generalization by a dog (Canis familiaris): is shape important?. *PLoS One,* e49382.

Vas, J., Topál, J., Gácsi, M., Miklósi, A., & Csányi, V. (2005). A friend or an enemy? Dogs' reaction to an unfamiliar person showing behavioural cues of threat and friendliness at different times. *Applied Animal Behaviour Science, 94*(1-2), 99-115.

Vatakis, A., & Spence, C. (2007). Crossmodal binding: Evaluating the "unity assumption" using audiovisual speech stimuli. *Perception & Psychophysics, 69,* 744–756.

Vatakis, A., & Spence, C. (2008). Evaluating the influence of the 'unity assumption' on the temporal perception of realistic audiovisual stimuli. *Acta Psychologica, 127,* 12–23.

Vatakis, A., Ghazanfar, A. A., & Spence, C. (2008). Facilitation of multisensory integration by the "unity effect" reveals that speech is special. *Journal of Vision, 8,* 14, 1–11.

Vilà, C., Maldonado, J. E., & Wayne, R. K. (1999). Phylogenetic relationships, evolution, and genetic diversity of the domestic dog. *Journal of Heredity, 90*(1), 71-77.

Vitulli, W. F. (2006). Attitudes towards empathy in domestic dogs and cats 1, 2. *Psychological Reports, 99*(3), 981-991.

Von Kriegstein, K., Eger, E., Kleinschmidt, A., & Giraud, A. L. (2003). Modulation of neural responses to speech by directing attention to voices or verbal content. *Cognitive Brain Research, 17*(1), 48-55.

Vorperian, H. K., Wang, S., Chung, M. K., Schimek, E. M., Durtschi, R. B., Kent, R. D., Ziegert, A.J. & Gentry, L. R. (2009). Anatomic development of the oral and pharyngeal portions of the vocal tract: An imaging study. *The Journal of the Acoustical Society of America, 125*(3), 1666-1678.

Vorperian, H. K., Wang, S., Schimek, E. M., Durtschi, R. B., Kent, R. D., Gentry, L. R., & Chung, M. K. (2011). Developmental sexual dimorphism of the oral and pharyngeal portions of the vocal tract: An imaging study. *Journal of Speech, Language, and Hearing Research, 54*(4), 995-1010.

Vroomen, J., Bertelson, P., & De Gelder, B. (2001). The ventriloquist effect does not depend on the direction of automatic visual attention. *Perception & Psychophysics, 63*, 651–659.

Walker, P., Bremner, J. G., Mason, U., Spring, J., Mattock, K., Slater, A., & Johnson, S. P. (2010). Preverbal infants' sensitivity to synaesthetic cross-modality correspondences, *Psychological Science, 21*, 21–25.

Walker, S., Bruce, V., & O'Malley, C. (1995). Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect, *Perception & Psychophysics, 57*, 1124–1133.

Walker-Andrews, A.S., Bahrick, L.E., Raglioni, S.S., & Diaz, I. (1991). Infants' bimodal perception of gender. *Ecological Psychology, 3*(2), 55-75.

Walsh, V. (2003). A theory of magnitude: Common cortical metrics of time, space and quantity. *Trends in Cognitive Science, 7*, 483–488.

Wang, S. H., & Baillargeon, R. (2008). Detecting impossible changes in infancy: A three-system account. *Trends in Cognitive Science, 12*, 17–23.

Wang, S. H., Baillargeon, R., & Brueckner, L. (2004). Young infants' reasoning about hidden objects: Evidence from violation-of-expectation tasks with test trials only. *Cognition, 93*, 167–198.

Wells, D.L., & Hepper, P.G. (1999). Male and female dogs respond differently to men and women. *Applied Animal Behaviour Science, 61*(4), 341-349.

Wildgruber, D., Riecker, A., Hertrich, I., Erb, M., Grodd, W., Ethofer, T., & Ackermann, H. (2005). Identification of emotional intonation evaluated by fMRI. *Neuroimage, 24*(4), 1233-1241.

Wilkins, A. S., Wrangham, R. W., & Fitch, W. T. (2014). The "domestication syndrome" in mammals: A unified explanation based on neural crest cell behavior and genetics. *Genetics, 197*(3), 795-808.

Wilson, E.O. (1975). *Sociobiology: The New Synthesis*. Cambridge: Harvard University Press.

Wood, S. (2003). Praat for beginners [Manual]. Retrieved May 12, 2014, from http://person2.sol.lu.se/SidneyWood/praate/wavformedform.html

Woods, R. H. (1893). Law of transverse vibrations of strings applied to the human larynx. *Journal of Anatomy & Physiology, 27*, 431–435.

Woods, T. M., & Recanzone, G.H. (2004). Visually induced plasticity of auditory spatial perception in macaques. *Current Biology, 14*, 1559–1564.

Wynne, C. D., Udell, M. A., & Lord, K. A. (2008). Ontogeny's impacts on human–dog communication. *Animal Behaviour, 76*(4), e1-e4.

Yong, M. H., & Ruffman, T. (2014). Emotional contagion: Dogs and humans show a similar physiological response to human infant crying. *Behavioural Processes, 108*, 155-165.

Yong, M. H., & Ruffman, T. (2015). Is that fear? Domestic dogs' use of social referencing signals from an unfamiliar person. *Behavioural Processes, 110*, 74-81.

Zangenehpour, S., & Zatorre, R.J. (2010). Cross-modal recruitment of primary visual cortex following brief exposure to bimodal audiovisual stimuli. *Neuropsychologia, 48*, 591–600.

Zangenehpour, S., Ghazanfar, A.A., Lewkowicz, D.J., & Zatorre, R.J. (2009). Heterochrony and cross-species intersensory matching by infant vervet monkeys. *PLoS One. 4*, e4302.

Zatorre, R. J., & Belin, P. (2001). Spectral and temporal processing in human auditory cortex. *Cerebral Cortex, 11*(10), 946-953.

Zatorre, R. J., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: music and speech. *Trends in Cognitive Sciences, 6*(1), 37-46.

Zatorre, R. J., & Gandour, J. T. (2008). Neural specializations for speech and pitch: moving beyond the dichotomies. *Philosophical Transactions of the Royal Society B: Biological Sciences, 363*(1493), 1087-1104.

Zentall, T.R., Wasserman, E.A., Lazareva, O.F., Thompson, R.K.R., & Rattermann, M.J. (2008). Concept learning in animals. *Comparative Cognition and Behavior Reviews*, *3*, 13-45.

Zuberbühler, K. (2001). Predator-specific alarm calls in Campbell's monkeys, *Cercopithecus campbelli*. *Behavioral Ecology & Sociobiology, 50*(5), 414–422.

# Appendices

## Appendix 1. Key Experimental Paradigms

To determine whether mammalian species form cross-modal associations about information encoded in their signals, researchers have commonly used two different behavioural experimental paradigms, both of which were originally designed for developmental research with human infants: preferential looking, as described by Golinkoff *et al., (*1987), and the violation of expectation method outlined by Baillargeon *et al.,* (1985). Because researchers face fundamentally similar methodological challenges when investigating the perceptual and cognitive abilities of both preverbal human infants and non-human animals, such as limited attention and communication skills, paradigms initially developed for human infants can usually be adapted to explore comparable traits in non-human animals.

The preferential looking paradigm is based on the observation that when an association exists between two perceptual cues, the presence of one will trigger increased attention to the other (see Golinkoff *et al.,* 1987). Additional attention to the congruent pairing can also be obtained for ecologically valid stimuli as human infants generally prefer to fixate on familiar socially or emotionally relevant stimuli (Houston-Price and Nakai, 2004). Since its introduction, the preferential looking paradigm has become a well-established methodology to study associative knowledge and memory in nonverbal populations such as human infants (Golinkoff *et al.*, 2013). When investigating associations between visual and auditory information in animals, the subject is presented with two visual stimuli, and a sound matching one of the visual stimuli in a specific dimension is played. Similarly to the human infant research, preferentially attending to the visual image that best matches the sound (e.g., faster response latency, longer looking duration, or more looks in total; Aslin, 2007) is usually taken to provide a behavioural indication that the animal has combined the different sensory information according to the shared dimension. However, in some cases shorter attendance to the congruent image has also been interpreted as showing that the animal has associated the audiovisual stimuli, where additional evidence has suggested that the congruent pairing may have been perceived as negative and therefore visually avoided (e.g., Zangenehpour *et al.,* 2009). The association pattern is even more complex in human

infant studies, as according to the 'dynamic attentional preference model', attention can shift from familiar to novel stimuli with increasing levels of exposure (Hunter and Ames, 1988). The attentional shift to novel stimuli is thought to occur after the familiar stimuli have been encoded, or when there is no discrepancy between the familiar stimuli presented and the infant's internal representation of those stimuli (Pascalis and De Haan, 2003; Sokolov, 1963). Therefore, whilst differential looking times to the visual stimuli can enable researchers to conclude that animals have made a distinction between stimuli, and that (usually) the most strongly attended stimulus is perceived to be more salient, *a priori* hypotheses are necessary to infer whether the behavioural responses reflect a familiarity or novelty preference (Houston-Price and Nakai, 2004). A further limitation of the preferential looking paradigm is that because stimuli from both modalities are simultaneously presented, it is possible for animals to match the congruent cues simply on the basis of their previous co-occurrence, and so it cannot be determined whether the subjects form a functional association between the stimuli. Therefore, a major shortcoming of the preferential looking paradigm is that it does not reveal the nature of the processes that underlie associations across the senses, and can limit the ability of studies using this paradigm to distinguish between low level and higher level cognitive processes.

The main alternative research methodology is the violation of expectation paradigm, which was originally designed to test the understanding of object permanence by presenting human infants with a possible and an impossible physical event (Baillargeon *et al.*, 1985). The authors proposed that if infants possess a concept of object permanence, then they will attend more to the impossible event, as attentional capture occurs when there is an invariance detected in an unfolding sequence of events. Similarly to the preferential looking paradigm, stronger attentional capture is suggested by longer looking times (Aslin and Fiser, 2005). The two methodologies initially appear to be conflicting, as stronger attendance to the matching stimulus is usually predicted from the preferential looking paradigm, whilst stronger attendance towards the non-matching stimuli is predicted in the violation of expectation paradigm. However, this contradiction can be explained by the way that the stimuli are presented. Unlike the preferential looking paradigm, the violation of expectancy design does not test if the subject has formed a prior association between the stimuli or not (stimulus novelty), but

rather whether they perceive that the sequence of events which they are presented with fit together (stimulus deviance) (Vachon *et al.*, 2012).

Although there has been some controversy in the interpretation of infant responses in this paradigm (Wang *et al.*, 2004), the violation of expectation method has since been used to test conceptual understanding in many areas of developmental and cognitive psychology (Wang and Baillargeon, 2008). When investigating multisensory abilities in animals, the key advantage of the violation of expectation paradigm over the preferential looking paradigm is that it enables researchers to determine not just whether information can be associated across the senses, but also whether subjects possess a functional cognitive representation of the dimension being investigated. The most common experimental procedure applying the violation of expectation paradigm with non-human animals involves presenting the subject with a stimulus from one sensory modality (e.g., visual) to prime a representation and thereby set up an expectation of what should follow. The first stimulus is then removed before a second stimulus from a different sensory modality (e.g., auditory) is presented. The second stimulus either matches a specific dimension of the first stimulus, or does not match it in any way. When non-matching stimuli are presented the animal is predicted to pay more attention to the second stimulus as it has not been primed to expect that stimulus and should be 'surprised' by its appearance. As in studies that have used this paradigm with human infants, surprise is usually inferred by higher levels of attention to the incongruent stimulus (e.g., response latency, duration of first look, number of looks and total look duration; Proops *et al.*, 2009).

Both paradigms have been successfully applied within the field of multisensory research to determine how animals associate relevant biological information transmitted through different sensory modalities. The preferential looking paradigm has been most frequently used to investigate how animals associate stimuli using basic redundant features, such as temporal synchrony (e.g., Zangenehpour *et al.*, 2009), whilst the advantages of the violation of expectation paradigm in identifying cognitive representations has led to its greater application in exploring the occurrence of more complex correspondences which can be related to multisensory categorical representations (e.g., Adachi *et al.*, 2007).