Cross-modal correspondences in non-human mammal communication

Article (Accepted Version)

Cross-modal Correspondences in Non-Human Mammal Communication

Victoria F. Ratcliffe*, Anna M. Taylor and David Reby

School of Psychology, University of Sussex, Falmer BN1 9QH, UK

* To whom correspondence should be addressed. E-mail: v.ratcliffe@sussex.ac.uk

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.

Keywords

Animal, mammal, communication, perception, cross-modal correspondence, multisensory

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, 2000a), 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, 2000b). 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, 2000b, 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.

References

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

Adachi, I. and Hampton, R. (2011). Rhesus monkeys see who they hear: Spontaneous cross-modal memory for familiar conspecifics, *PLoS One* **6**, e23345. doi: 10.1371/journal.pone.0023345

Adachi, I., Kuwahata, H., Fujita, K., Tomonaga, M. and 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. and Fujita, K. (2007). Dogs recall their owner's face upon hearing their owner's voice, *Anim. Cogn*. **10**, 17–21.

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

Agrillo, C. and Petrazzini, M. E. M. (2013). Glimpse of ATOM in non-human species? *Front. Psychol.* **4**, 1–4. doi: 10.3389/fpsyg.2013.00460

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

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

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

Bachorowski, J. A. and Owren, M. J. (1999). Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech, *J. Acoust. Soc. Am.* **106**, 1054–1063.

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

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

Batteau, D. W. (1967). The role of the pinna in human localization, *Proc. R. Soc. Lond. B. Biol. Sci.* **168**, 158–180.

Bertelson, P. and Aschersleben, G. (1998). Automatic visual bias of perceived auditory location, *Psychon. Bull. Rev.* **5**, 482–489.

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

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

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

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

Bremner, A. J., Caparos, S., Davidoff, J., de Fockert, J., Linnell, K. J. and 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, *J. Zool.* **288**, 1–20.

Cantlon, J. F. (2012). Math, monkeys, and the developing brain, *Proc. Natl Acad. Sci*. **109**, 10725–10732.

Cantlon, J. F. and Brannon, E. M. (2007). Basic math in monkeys and college students, *PLoS Biol.* **5**, e328. doi: 10.1371/journal.pbio.0050328

Cartei, V., Cowles, H. W. and Reby, D. (2012). Spontaneous voice gender imitation abilities in adult speakers, *PLoS One* **7**, e31353. doi: 10.1371/journal.pone.0031353

Charlton, B. D., Reby, D. and McComb, K. (2008). Effect of combined source (F0) and filter (formant) variation on red deer hind responses to male roars, *J. Acoust. Soc. Am.* **123**, 2936–2943.

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

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

Cheng, K. (1990). More psychophysics of the pigeon's use of landmarks, *J. Comp. Physiol. A* **166**, 857–863.

Collins, S. A. (2000). Men's voices and women's choices, *Anim. Behav*. **60**, 773–780.

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

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

De Boer, B. (2008). The acoustic role of supralaryngeal air sacs, *J. Acoust. Soc. Am.* **123**, 3779–3779.

De Hevia, M. D., Girelli, L., Addabbo, M. and Cassia, V. M. (2014). Human infants' preference for left-to-right oriented increasing numerical sequences, *PLoS One* **9**, e96412. doi: 10.1371/journal.pone.0096412

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

Emmerton, J. and Renner, J. C. (2006). Scalar effects in the visual discrimination of numerosity by pigeons, *Learn. Behav.* **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, *J. Vis*. **7**, 7, 1–14. doi: 10.1167/7.5

Ernst, M. O. and Bülthoff, H. H. (2004). Merging the senses into a robust percept, *Trends Cogn. Sci.* **8**, 162–169.

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

Evans, T. A., Howell, S. and Westergaard, G. C. (2005). Auditory-visual cross-modal perception of communicative stimuli in tufted capuchin monkeys (*Cebus apella*), *J. Exp. Psychol. Anim. Behav. Process.* **31**, 399–406.

Ey, E., Pfefferle, D. and 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*. Mouton, The Hague, The Netherlands.

Faragó, T., Pongrácz, P., Miklósi, Á., Huber, L., Virányi, Z. and Range, F. (2010). Dogs' expectation about signalers' body size by virtue of their growls, *PLoS One* **5**, e15175. doi: 10.1371/journal.pone.0015175

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

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

Fitch, W. T. (1997). Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques, *J. Acoust. Soc. Am*. **102**, 1213–1222.

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

Fitch, W. T. (2000b). Skull dimensions in relation to body size in nonhuman mammals: The causal bases for acoustic allometry, *Zoology (Jena)* **103**, 40–58.

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

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

Fitch, W. T. and Giedd, J. (1999). Morphology and development of the human vocal tract: A study using magnetic resonance imaging, *J. Acoust. Soc. Am.* **106**, 1511–1522.

Fletcher, N. H. (2004). A simple frequency-scaling rule for animal communication, *J. Acoust. Soc. Am.* **115**, 2334–2338.

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

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

Ghazanfar, A. A. (2013). Multisensory vocal communication in primates and the evolution of rhythmic speech, *Behav. Ecol. Sociobiol.* **67**, 1441–1448.

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

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

Ghazanfar, A. A., Neuhoff, J. G. and Logothetis, N. K. (2002). Auditory looming perception in rhesus monkeys, *Proc. Natl Acad. Sci*. **99**, 15755–15757.

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

Ghazanfar, A.A., Nielsen, K. and 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. and Logothetis, N. K. (2007). Vocal-tract resonances as indexical cues in rhesus monkeys, *Curr. Biol*. **17**, 425–430.

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

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

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

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

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

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

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

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

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

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

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

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

Hillenbrand J., Getty, L. A., Clark, M. J. and Wheeler, K. (1995). Acoustic characteristics of American English vowels, *J. Acoust. Soc. Am.* **97**, 3099–3111.

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

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

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

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

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

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

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

Johnstone, R. A. (1996). Multiple displays in animal communication: Backup signals and multiple messages, *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **351**(1337), 329–338.

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

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

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

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

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

Kondo, N., Izawa, E-I. and Watanabe, S. (2012). Crows cross-modally recognise group members but not non-group members, *Proc. Biol. Sci.* **279**, 1937–1942.

Kriengwatana, B., Escudero, P. and Ten Cate, C. (2014). Revisiting vocal perception in non–human animals: A review of vowel discrimination, speaker voice recognition, and speaker normalization, *Front. Psychol.* **5**(1543), 1–13. doi: 10.3389/fpsyg.2014.01543

Krusche, P., Uller, C. and Dicke, U. (2010). Quantity discrimination in salamanders, *J. Exp. Biol.* **213**, 1822–1828.

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

Kulachi, I. G., Drea, C. M., Rubenstein, D. I. and Ghazanfar, A. A. (2014). Individual recognition through olfactory–auditory matching in lemurs, *Proc. Biol. Sci*. **281**, 20140071.

Lewkowicz, D. J. and Ghazanfar, A. A. (2006). The decline of cross-species intersensory perception in human infants, *Proc. Natl Acad. Sci*. **103**, 6771–6774.

Lewkowicz, D.J. and Ghazanfar, A. A. (2009). The emergence of multisensory systems through perceptual narrowing. *Trends Cogn. Sci*. **13**, 470–478.

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

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

Lourenco, S. F. and Longo, M. R. (2010). General magnitude representation in human infants, *Psychol. Sci*. **21**, 873–881.

Ludwig, V. U., Adachi, I. and Matsuzawa, T. (2011). Visuoauditory mappings between high luminance and high pitch are shared by chimpanzees (*Pan troglodytes*) and humans, *Proc. Natl Acad. Sci*. **108**, 20661–20665.

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

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

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, *J. Exp. Psychol. Hum. Percept. Perform.* **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. and Ohlott, P. (1986). On the cross-modal perception of intensity, *J. Exp. Psychol. Hum. Percept. Perform.* **12**, 517.

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

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

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

Maynard-Smith, J. and Harper, D. (2003). *Animal Signals*. Oxford University Press, New York, NY, USA.

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

Meck, W. H. and Church, R. M. (1983). A mode control model of counting and timing processes, *J. Exp. Psychol. Anim. Behav. Process.* **9**, 320–334.

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

Merchant, H. and Honing, H. (2013). Are non-human primates capable of rhythmic entrainment? Evidence for the gradual audiomotor evolution hypothesis, *Front. Neurosci.* **7**, 274. doi: 10.3389/fnins.2013.00274

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

Moller, A. P. and Pomiankowski, A. (1993). Why have birds got multiple sexual ornaments? *Behav. Ecol. Sociobiol.* **32**, 167–176.

Monaghan, P., Mattock, K. and Walker, P. (2012). The role of sound symbolism in language learning, *J. Exp. Psychol. Learn.* **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. and 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, *Am. Nat.* **111**, 855–869.

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

Narins, P. M., Hödl, W. and Grabul, D. S. (2003). Bimodal signal requisite for agonistic behavior in a dart-poison frog, *Epipedobates femoralis*, *Proc. Natl Acad. Sci*. **100**, 577–580.

Neuhoff, J. G. and McBeath, M. K. (1996). The Doppler illusion: The influence of dynamic intensity change on perceived pitch, *J. Exp. Psychol. Hum. Percept. Perform.* **22**, 970–985.

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.

Parault, S. J. and Parkinson, M. (2008). Sound symbolic word learning in the middle grades, *Contemp. Educ. Psychol.* **33**, 647–671.

Parise, C. V. and Spence, C. (2009). 'When birds of a feather flock together': Synesthetic correspondences modulate audiovisual integration in non-synesthetes, *PLoS One*, **4**, e5664. doi: 10.1371/journal.pone.0005664

Parise, C. V. and 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. and Ernst, M. O. (2012). When correlation implies causation in multisensory integration, *Curr. Biol*. **22**, 46–49.

Parise, C. V., Knorre, K. and Ernst, M. O. (2014). Natural auditory scene statistics shapes human spatial hearing, *Proc. Natl Acad. Sci*. **111**, 6104–6108.

Parr, L. A. (2004). Perceptual biases for multimodal cues in chimpanzee (*Pan troglodytes*) affect recognition, *Anim. Cogn*. **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. and Marler, P. (1999). Communication goes multimodal, *Science* **283**(5406), 1272–1273.

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

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

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

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

Pisanski, K. and Rendall, D. (2011). The prioritization of voice fundamental frequency or formants in listeners' assessments of speaker size, masculinity, and attractiveness, *J. Acoust. Soc. Am.* **129**, 2201–2212.

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. and Feinberg, D. R. (2014). Vocal indicators of body size in men and women: A meta-analysis. *Anim. Behav*. **95**, 89–99.

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

Pratt, C. C. (1930). The spatial character of high and low tones, *J. Exp. Psychol.* **13**, 278–285.

Proops, L. and McComb, K. (2012). Cross-modal individual recognition in domestic horses (*Equus caballus*) extends to familiar humans, *Proc. Biol.. Sci*. **279**, 3131–3138.

Proops, L., McComb, K. and Reby, D. (2009). Cross-modal individual recognition in domestic horses (Equus caballus). *Proc. Natl. Acad. Sci*. **106**, 947–951.

Puts, D. A., Gaulin, S. J. and Verdolini, K. (2006). Dominance and the evolution of sexual dimorphism in human voice pitch, *Evol. Hum. Behav.* **27**, 283–296.

Puts, D. A., Apicella, C. L. and Cardenas, R. A. (2012). Masculine voices signal men's threat potential in forager and industrial societies, *Proc. Biol. Sci.* **279**, 601–609.

Ramachandran, V. S. and Hubbard, E. M. (2001). Synaesthesia — A window into perception, thought and language, *J. Conscious. Stud*. **8**, 3–34.

Ratcliffe, V. F., McComb, K. and Reby, D. (2014). Cross-modal discrimination of human gender by domestic dogs, *Anim. Behav*. **91**, 126–134.

Reby, D. and McComb, K. (2003). Anatomical constraints generate honesty: Acoustic signals to age and weight in the roars of red deer stags, *Anim. Behav*. **65**, 519–530.

Reby, D., McComb, K., Cargnelutti, B., Darwin, C., Fitch, W. T. and Clutton-Brock, T. (2005). Red deer stags use formants as assessment cues during intrasexual agonistic interactions, *Proc. Biol. Sci.* **272**(1566), 941–947.

Rendall, D., Kollias, S., Ney, C. and 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, *J. Acoust. Soc. Am.* **117**, 944–955.

Rendall, D., Vokey, J. R. and Nemeth, C. (2007). Lifting the curtain on the Wizard of Oz: Biased voice-based impressions of speaker size, *J. Exp. Psychol. Hum. Percept. Perform.* **33**, 1208–1219.

Riede, T. and Fitch, T. (1999). Vocal tract length and acoustics of vocalization in the domestic dog (*Canis familiaris*), *J. Exp. Biol.* **202**, 2859–2867.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M. and Boyes-Braem, P. (1976). Basic objects in natural categories, *Cogn. Psychol*. **8**, 382–439.

Rowe, C. (1999). Receiver psychology and the evolution of multicomponent signals, *Anim. Behav*. **58**, 921–931.

Rugani, R., Vallortigara, G., Priftis, K. and 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. and Giannini, N. (2007). Ground squirrels use an infrared signal to deter rattlesnake predation, *Proc. Natl Acad. Sci*. **104**, 14372–14376.

Rusconi, E., Kwan, B., Giordano, B. L., Umilta, C. and Butterworth, B. (2006). Spatial representation of pitch height: the SMARC effect *Cognition* **99**, 113–129.

Seyfarth, R. M. and Cheney, D. L. (1986). Vocal development in vervet monkeys, *Anim. Behav*. **34**, 1640–1658.

Seyfarth, R. M. and Cheney, D. L. (2015). Social cognition. *Anim. Behav***103**, 191–202.

Seyfarth, R. M., Cheney, D. L. and Marler, P. (1980a). Vervet monkey alarm calls: Semantic communication in a free-ranging primate, *Anim. Behav*. **28**, 1070–1094.

Seyfarth, R. M., Cheney, D. L. and Marler, P. (1980b). Monkey responses to three different alarm calls: evidence of predator classification and semantic communication, *Science* **210**(4471), 801–803.

Sliwa, J., Duhamel, J.-R., Pascalis, O. and Wirth, S. (2011). Spontaneous voice-face identity matching by rhesus monkeys for familiar conspecifics and humans, *Proc. Natl Acad. Sci*. **108**, 1735–1740.

Slutsky, D. A. and Recanzone, G. H. (2001). Temporal and spatial dependency of the ventriloquism effect, *Neuroreport* **12**(1), 7–10.

Smith, C. L. and Evans, C. S. (2008). Multimodal signaling in fowl, *Gallus gallus*, *J. Exp. Biol.* **211**, 2052–2057.

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

Sokolov, E. (1963). *Perception and Conditioned Reflex*. Pergamon, New York, NY, USA.

Spence, C. (2007). Audiovisual multisensory integration, *Acoust. Sci. Technol.* **28**, 61–70.

Spence, C. (2011). Crossmodal correspondences: A tutorial review, *Atten. Percept. Psychophys*. **73**, 971–995.

Spence, C. and Deroy, O. (2012). Crossmodal correspondences: Innate or learned? *I-Perception* **3**, 316–318.

Srinivasan, M. and 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, *J. Exp. Anal. Behav.* **9**, 135–142.

Stevens, S. S. (1957). On the psychophysical law, *Psychol. Rev.* **64**, 153–181.

Taylor, R. C., Klein, B. A., Stein, J. and Ryan, M. J. (2011). Multimodal signal variation in space and time: How important is matching the signal with its signaler? *J. Exp. Biol.* **214**, 815–820.

Taylor, A. M., Reby, D. and McComb, K. (2008). Human listeners attend to size information in domestic dog growls, *J. Acoust. Soc. Am.* **123**, 2903–2909.

Taylor, A.M., Reby, D. and McComb, K. (2011). Cross modal perception of body size in domestic dogs (*Canis familiaris*), *PLoS One* **6**, e17069. doi: 10.1371/journal.pone.0017069

Tedore, C. and Johnsen, S. (2014). Visual mutual assessment of size in male *Lyssomanes viridis* jumping spider contests, *Behav. Ecol.* **26**, 510–518.

Thompson, J. T., Bissell, A. N. and Martins, E. P. (2008). Inhibitory interactions between multimodal behavioural responses may influence the evolution of complex signals, *Anim. Behav*. **76**, 113–121.

Titze, I. R. (1994). *Principles of Vocal Production*. Prentice Hall, Englewood Cliffs, NJ, USA.

Uetz, G. W. and Roberts, J. A. (2002). Multisensory cues and multimodal communication in spiders: insights from video/audio playback studies, *Brain Behav. Evol*. **59**, 222–230.

Vachon, F., Hughes, R. W. and Jones, D. M. (2012). Broken expectations: Violation of expectancies, not novelty, captures auditory attention, *J. Exp. Psychol. Learn.* **38**, 164–177.

Vatakis, A. and Spence, C. (2007). Crossmodal binding: Evaluating the "unity assumption" using audiovisual speech stimuli, *Percept. Psychophys.* **69**, 744–756.

Vatakis, A. and Spence, C. (2008). Evaluating the influence of the 'unity assumption' on the temporal perception of realistic audiovisual stimuli, *Acta Psychol.* **127**, 12–23.

Vatakis, A., Ghazanfar, A. A. and Spence, C. (2008). Facilitation of multisensory integration by the "unity effect" reveals that speech is special, *J. Vis*. **8**, 14, 1–11. doi: 10.1167/8.9.14

Vroomen, J., Bertelson, P. and De Gelder, B. (2001). The ventriloquist effect does not depend on the direction of automatic visual attention, *Percept. Psychophys.* **63**, 651–659.

Walker, S., Bruce, V. and O'Malley, C. (1995). Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect, *Percept. Psychophys.* **57**, 1124–1133.

Walker, P., Bremner, J. G., Mason, U., Spring, J., Mattock, K., Slater, A. and Johnson, S. P. (2010). Preverbal infants' sensitivity to synaesthetic cross-modality correspondences, *Psychol. Sci*. **21**, 21–25.

Walker-Andrews, A. S., Bahrick, L. E., Raglioni, S. S. and Diaz, I. (1991). Infants' bimodal perception of gender, *Ecol. Psychol*. **3**, 55–75.

Walsh, V. (2003). A theory of magnitude: Common cortical metrics of time, space and quantity, *Trends Cogn. Sci*. **7**, 483–488.

Wang, S. H. and Baillargeon, R. (2008). Detecting impossible changes in infancy: A three-system account,. *Trends Cogn. Sci.* **12**, 17–23.

Wang, S. H., Baillargeon, R. and Brueckner, L. (2004). Young infants' reasoning about hidden objects: Evidence from violation-of-expectation tasks with test trials only, *Cognition* **93**, 167–198.

Woods, R. H. (1893). Law of transverse vibrations of strings applied to the human larynx. *J. Anat. Physiol.* **27**, 431–435.

Woods, T. M. and Recanzone, G.H. (2004). Visually induced plasticity of auditory spatial perception in macaques. *Curr. Biol*. **14**, 1559–1564.

Zangenehpour, S., and 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. and Zatorre, R.J. (2009). Heterochrony and cross-species intersensory matching by infant vervet monkeys. *PLoS ONE.* **4**, e4302.

Zuberbühler, K. (2001). Predator-specific alarm calls in Campbell's monkeys, *Cercopithecus campbelli*. *Behav. Ecol. Sociobiol.* **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).