Behavioural Processes 163 (2019) 99–112

Contents lists available at ScienceDirect

Behavioural Processes

journal homepage: www.elsevier.com/locate/behavproc

Vocal tract constancy in birds and humans

Cleopatra Diana Pike*, Buddhamas Pralle Kriengwatana

School of Psychology and Neuroscience, University of St Andrews, St Mary's Quad, South Street, St Andrews, Fife, KY16 9JP, UK

ARTICLE INFO

Keywords:
Vocal tract normalisation
Auditory constancy
Bird song
Speech
Categorical perception
Spectral contrast
Spectral compensation effect
Auditory afterimage

ABSTRACT

Humans perceive speech as being relatively stable despite acoustic variation caused by vocal tract (VT) differences between speakers. Humans use perceptual 'vocal tract normalisation' (VTN) and other processes to achieve this stability. Similarity in vocal apparatus/acoustics between birds and humans means that birds might also experience VT variation. This has the potential to impede bird communication. No known studies have explicitly examined this, but a number of studies show perceptual stability or 'perceptual constancy' in birds similar to that seen in humans when dealing with VT variation. This review explores similarities between birds and humans and concludes that birds show sufficient evidence of perceptual constancy to warrant further research in this area. Future work should 1) quantify the multiple sources of variation in bird vocalisations, including, but not limited to VT variations, 2) determine whether vocalisations are perniciously disrupted by any of these and 3) investigate how birds reduce variation to maintain perceptual constancy and perceptual efficiency.

1. Introduction

Perceptual constancy describes the perceived stability of an object, or its properties, despite physical changes that occur when the object is produced in different contexts. For example, the colour of an apple viewed under sunlight remains stable despite changes in the spectrum of illumination over the course of the day (Foster, 2011). This perceptual process is necessary for consistently recognising objects: colour constancy might have been particularly helpful during human evolution for foraging and recognising the ripest food under different light conditions. Auditory constancy also occurs in humans and appears to be useful for speech recognition. The acoustic properties of speech are modified by the individual speaker's vocal tract (the vocal apparatus lying above the larynx including the mouth, tongue and lips), neighbouring speech (the 'phonetic context'), accent, and the environment (e.g. room reverberation). It is well known that listening machines (e.g. speech recognition devices) require processes to deal with these variations to perform accurate speech recognition. Evidence suggests that humans also engage perceptual constancy mechanisms to recognise speech and other sounds (Pisoni et al., 1997; Cohen et al., 1995).

In this article we focus on acoustic variation caused by differences in vocal tracts (VT). Many studies have examined how humans deal with VT variation during speech perception. However, this has received relatively little attention in the avian literature. Bird vocalisations share certain acoustical features with speech (Doupe and Kuhl, 1999) and the acoustic realization of any sung note may be subject to VT differences between birds (Samuels, 2015; Yip, 2006; Lachlan et al., 2014, 2016), neighbouring vocalisations (the 'phonetic context'), dialect, and environmental effects (e.g. reverberation and sound reflections from rocks and plants). As for humans, VT variation in birds has the potential to prevent the recognition of messages contained in vocalisations. Therefore, birds might also need a constancy mechanism. However, birds might benefit from hearing VT differences to a greater extent than humans (e.g. as an indicator of singer quality) and so require sensitivity to singer characteristics contained in VT acoustics.

Studying perception in birds is interesting in its own right but where there is evidence of homologous or convergent evolution it can also provide insight into human perception. The auditory periphery (the hearing system up to and including the auditory nerve) is suggestive of homologous evolution between humans and birds because some of its primary features, e.g. the transduction of air pressure by a tympanic middle ear and tonotopic representation at the auditory nerve, are seen in our common ancestor, tetrapods (Fritzsch et al., 2013, but see Clack, 2002, for evidence of convergent evolution). There is even evidence of similar efferent projections to cochlear hair cells in birds and humans, which may be especially important for VT constancy (Elgoyhen and Katz, 2012; Beeston et al., 2014). As a result, VT constancy mechanisms that involve the auditory periphery are likely to be shared. However, there is also evidence of divergence at the periphery, for example, the owl is similar to humans in its encoding of interaural time and level

Abbreviations: CP, categorical perception; LTAS, long term average spectrum; VT, vocal tract; VTN, vocal tract normalization; PME, perceptual magnet effect

* Corresponding author.
E-mail address: cdp5@st-andrews.ac.uk (C.D. Pike).

https://doi.org/10.1016/j.beproc.2018.08.001
Received 11 June 2017; Received in revised form 30 July 2018; Accepted 10 August 2018
Available online 23 August 2018
0376-6357/

differences at the auditory nerve (Köppl, 1997) but has developed additional capabilities for sound localisation through the use of asymmetric ears. Therefore, common peripheral mechanisms cannot be assumed.

Above the periphery between-species divergence in perceptual mechanisms is increasingly likely due to the complexity of the hearing system. Generally, there is evidence that some high-level vocal communication mechanisms are shared between birds and humans (for example the learning of vocalisations may arise from analogous processes (Warren et al., 2011) and/or shared genes such as FOXP2 (White et al., 2006). However, little is currently known about whether high-level *constancy* mechanisms are shared. Ohms posits that VT constancy may be homologous because the mechanisms appear not to be speech specific (Ohms et al., 2010). Additionally, there is potential homology in vocal production mechanisms (both species use a periodic vocal source and a tube-like vocal tract filter – Fitch and Hauser, 2001) and VT constancy appears to deal with a common problem (within-species variation in vocal tracts), so analogies, if not homologies, are likely. However, more research is needed to establish the evolutionary origins of the mechanisms discussed in this paper.

Due to the potential usefulness of comparing human and bird perception, the current review examines whether, and how, both species maintain constancy across VT variation. Section 2 of this paper describes human speech perception, human vocal tract normalisation (VTN) and other human 'non-normalisation' methods for dealing with VT variation. Section 3 examines bird vocalisations and the evidence for similar constancy mechanisms in bird perception. We conclude by summarising the case for VT constancy in birds and providing suggestions for future experiments.

2. Vocal tract constancy in humans

The main acoustical features necessary for recognising speech units ('phonemes') can be divided into: spectral (the location of vocal energy along the frequency spectrum); spectral-transitional (the frequency regions that the energy traverses during the phoneme); and temporal (the exact timing of changes in this energy) (Nearey, 1989). To produce these features, speakers change the size and shape of their vocal tract. Specifically, they configure the vocal tract to produce narrow peaks in energy at distinct points in the frequency spectrum known as 'formants', which they then vary over time. Formants are numbered according to their frequency order on the spectrogram (Fig. 1). F0 represents the lowest or 'fundamental' frequency in the speech – this is related to voice pitch - and formants are located further along the spectrum and labelled F1, F2, etc.

The frequency location of formants is important for defining the character or 'timbre' of the phoneme and creating acoustic distinctiveness (e.g. creating an "ah" sound rather than "eh"). The first two formants (F1 and F2) appear particularly important for creating different sounding speech sounds. The simplest speech sounds, the 'monophthongal vowels', can be produced by simply altering the location of F1 and F2 (Fig. 1): higher formants and the wide-band spectrum can contribute to identification of these vowels, but F1 and F2 are sufficient for perception (Kiefte and Kluender, 2005; Carlson et al., 1975; Helmholtz, 1863; Johnson, 2005).

The frequency location of F1 and F2 is also central to the perception of most other phonemes. However, for other phonemes additional cues are important. Formant change during the phoneme and timing cues (e.g. voice onset time) are essential cues for some speech sounds (e.g. stop consonants, Delattre et al., 1955; Liberman et al., 1967, 1957, Ainsworth, 1988; Bennett, 1968) and it should be noted that formants are not relevant for some sounds, but other spectral cues are (e.g. noise bursts, and their average frequency, are cues for fricative consonants such as /s/ and /f/).

It is evident that spectral cues, such as formants, are key to speech recognition and traditional models of speech perception state that correct and exact cues ('canonical' cues) need to occur for correct perception (Strange, 1989; Nearey, 1989). However, a prominent source of variation to these cues occurs at the production stage due to differences in VT size and shape between speakers: for example, the VT of men is on average 30% larger than that of women (Fant, 2001; Ainsworth, 1988). As a result, the formant values used by different speakers to produce the same phoneme vary. Peterson and Barney (1952) describe between-speaker variation in F1 and F2 for various vowels (Fig. 2). Fig. 2 shows that the canonical formants do not occur for most speakers. Importantly, as well as general variation, the formants produced for a particular vowel by some speakers are the same as those produced by other speakers for an entirely different vowel. This is known as 'overlap in format space'. In regions of overlap it is not possible to determine the vowel based on F1 and F2 alone. Given the supposed importance of these cues, this indicates a problem for speech recognition (overlap also affects consonants as well as vowels and can affect higher formants and other spectral attributes).

2.1. Normalisation methods

Despite this acoustic variation, humans are good at perceiving speech. Even in regions of overlap the correct vowel is almost always heard. This may be because listeners normalise formants to reduce variation and remove overlap. Normalisation involves the perceptual shifting of modified cues back to the canonical form. VTN, specifically, is the shifting of cue variations that have occurred due to variation in VT dimensions (Pisoni et al., 1997). The normalisation process can result in accurate recognition but implies a perception that is no longer sensitive to the original variation (Pisoni et al., 1997; Cohen et al., 1995).

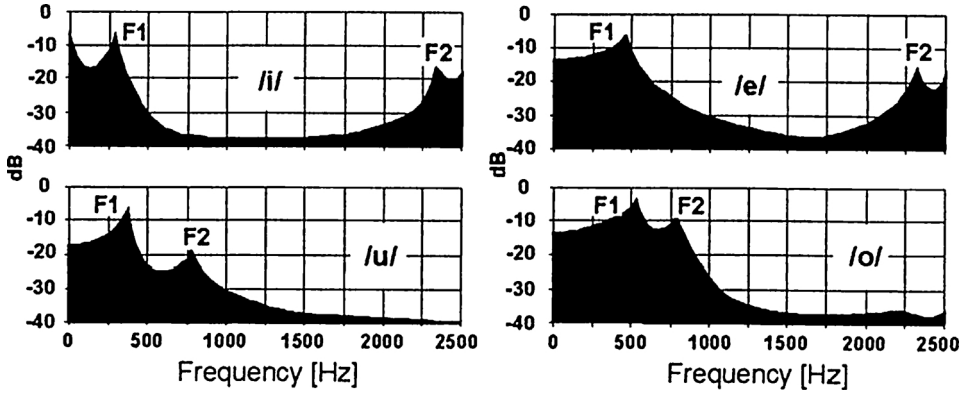

Fig. 1. Frequency domain representation of formant frequencies for the monophthongal vowels, /i/, /e/, /u/ and /o/. The formants are labelled F1, F2 according to their order on the spectrogram. From Ohl and Scheich (1997).

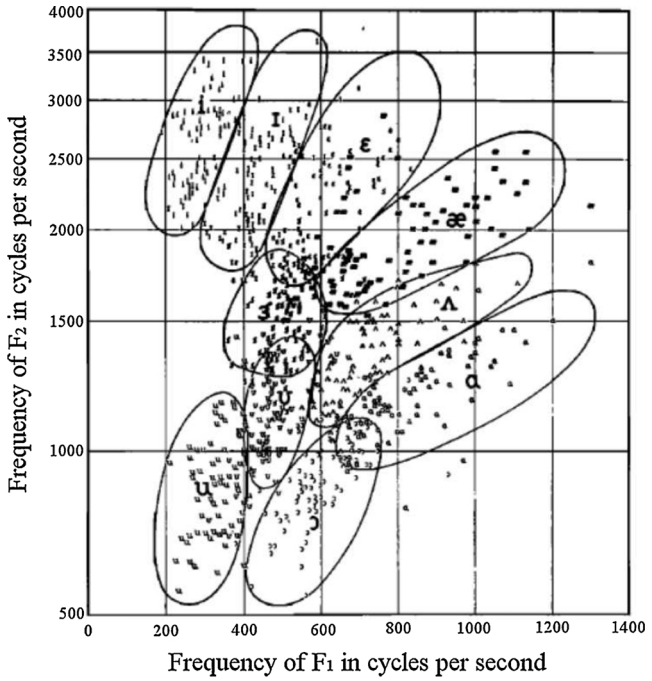

Fig. 2. From Peterson and Barney (1952). Formant frequency values for F1 and F2 in American English vowels – Ellipses represent formant space. Each point within an ellipse represents one utterance by a different speaker of the vowels labelled. The Koenig frequency scale (linear to 1000 Hz and logarithmic above) is used. Copied verbatim from original article but with axis values enhanced.

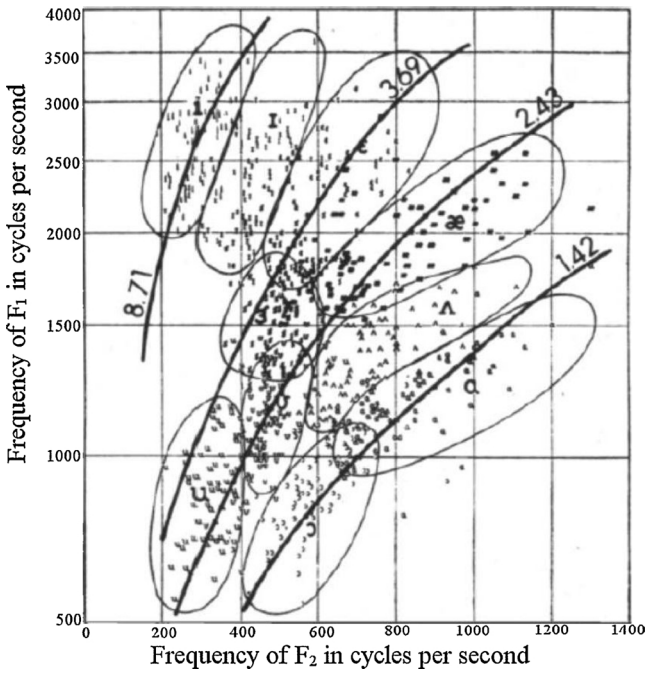

Fig. 3. From Miller (1989) after Peterson and Barney (1952) [SIC]. F1 and F2 locations for American English Vowels. Ellipses represent formant space. Each point within an ellipse represents one utterance by a different speaker of the vowels labelled. Equal F1/F2 ratios are depicted as curved lines. The Konig scale is used. Copied verbatim from original article but with axis values enhanced.

2.1.1. Phoneme intrinsic normalisation

Normalisation requires cues to signal the VT dimensions and therefore the kind of perceptual shift required to bring the cues back to cannonical form. Such cues may exist within the phoneme being normalised (intrinsic cues) or within wider speech by the same speaker (extrinsic cues).

Specific phoneme intrinsic cues might be used. Fundamental frequency (F0) is a correlate of VT size and shape and is usually present within each phoneme. Evidence that perceived F0 is used for normalisation comes from studies showing: improved speech identification with phonated speech (F0 present) compared to whispered speech (F0 absent – Nusbaum and Morin, 1992; Halberstam and Raphael, 2004); identification errors with a discrepancy between F0 and formants (Lehiste and Meltzer, 1973; Peterson and Barney, 1952); and shifts in vowel categorisation with F0 alterations, particularly in the central area of Peterson and Barney's vowel space where there is more overlap (Miller, 1989). However, the relationship between F0 and higher formants can be unreliable and recognition is sometimes good for whispered or 'sine wave' speech without F0 (Remez et al., 1981). It is noted that F0 can vary somewhat between phonemes as well as speakers, so this may not provide a reliable phoneme intrinsic cue to VT size (although it could provide a good cue over a longer speech segment if the listener is able to establish an average F0 of the speaker). F3 is also correlated with vocal tract size and is a distinctive VT cue as it varies more between speakers than between phonemes. Miller (1953) found that overlap within the [u] region in Peterson and Barney's study was no longer a perceptual problem when the primary cues, F1 and F2, were heard in light of their relationship to perceived F3.

While formants may shift between speakers, the relationship between formants is more constant. Formant ratio theories of perception state that the listener listens directly for these relationships (Lloyd, 1890a,b; Potter and Steinberg, 1950). This theory is aligned with the influential Gestalt theory – the perception of the whole occurs before the parts (Johnson, 2005; Koffka, 1935; Traunmüller, 1984) and Sussman (1989), Sussman et al. (1997) found 'combination sensitive neurons' that appear to detect formant relationships like feature detection mechanisms for lines and edges in vision (Hubel and Wiesel, 1959; Shapley and Tolhurst, 1973). Fig. 3, from Miller (1989), shows Peterson and Barney's data and displays constant ratios of F1 and F2 as lines. However, it is evident from this figure that only a little variation is explained by ratios between F1 and F2 alone. Accurate perception could come about from listening for specific relationships between additional cues but this begins to describe 'whole phoneme' methods of perception such as 'categorisation' (see below), rather than an approach that uses one or two specific cues.

2.1.2. Phoneme extrinsic normalisation

Other research suggests that experience with the speaker is necessary for normalisation because cues for normalisation reside outside of the phoneme. There is significant evidence of improved speech recognition with speaker experience (Strange et al., 1976; Mullennix et al., 1989) which may show 'phoneme extrinsic' normalisation.

A seminal study by Ladefoged and Broadbent (1957) demonstrated perceptual shifts in formants that were suggestive of VTN. Listeners were presented with the sentence "please say what this word is…" followed by an ambiguous test word between "bit" and "bet", (F1 altered vowels were used to create test words – a high F1 created a canonical /e/ vowel in "bet", a low F1 created /i/ in "bit", and values in between were used to create ambiguous test words). Listeners had to identify test words after hearing the prior sentence with all F1 cues shifted to be higher or lower than in the original recording. This mimicked between speaker VT differences (e.g. someone with a larger VT would display lower F1 cues). Perceptual shifts in the test word depending on F1 in the prior sentence occurred in the manner expected if VTN had taken place. The authors suggested that gaining information about the speakers 'formant space' from prior listening (the typical position of the formants for that speaker) allowed for 'relational processing', whereby the relationship of test sound cues to the speaker's

formant space is relevant, rather than their absolute value (Helson, 1948; Joos, 1948). Specifically, if the immediately preceding sentence was altered to contain low F1 values, a test vowel with a mid-F1 value was more likely to be perceived as a having high F1 value ("bet") rather than its acoustic value because, relative to the prior context, it was heard as 'high'.

As well as between-speaker variation, variation to formants is also caused by speech immediately adjacent to the phoneme. In normal running speech, a steady-state phoneme nucleus containing the relevant formants is not reached (Assmann et al., 1982; Macchi, 1980). Cues for each phoneme are assimilated and shifted toward the location of those in the adjacent phonemes (Liberman et al., 1967; Lindblom and Studdert-Kennedy, 1967; Verbrugge et al., 1976). This process is due to co-articulation caused by the sluggishness of articulators (Lindblom, 1963; Nearey and Assmann, 1986). The shifting of cues towards the adjacent values is known as acoustical "undershoot" because the formant values are not quite reached during this process. Undershoot has a similar effect as VT variation: undershoot involves a pulling of formants towards the immediately prior VT configuration and VT variation involves a pulling towards the general VT configuration of the speaker (as is necessary due to the confines of the size and shape of the speaker). Undershoot causes overlap in formant space like that caused by VT variation (Lindblom and Studdert-Kennedy, 1967; House and Fairbanks, 1953). However, undershoot does not result in perceptual problems. It appears to be dealt with by a similar relational perception as was demonstrated for between-speaker variation. In the perception of a test phoneme with a middle acoustic F1 value, if F1 in the immediately prior phoneme is low then this sets the context: the mid F1 in the current phoneme will be perceived in light of the prior phoneme context as relatively high. This creates normalisation to counteract undershoot but also an apparent shifting apart of phoneme cues from each other, which has been named the "overshoot" effect (Lindblom and Studdert-Kennedy, 1967).

2.1.3. Phoneme extrinsic normalisation – mechanisms

Similar perceptual shifts occur to compensate for the similar problems of between-speaker VT variation and undershoot. Therefore, as was suggested by Summerfield et al. (1984, 1989), the underlying mechanism of compensation might be the same and researchers can look to mechanisms of overshoot to determine mechanisms of VTN. Both overshoot and VTN appear to be normalised via relational processing, or simply "listening in context". However, the mechanisms behind listening in context may not be simple. Possible mechanisms are described below and the potential for birds to use these is discussed.

A 'learning' of cues that map out the speaker's vowel space might be necessary for the VTN seen in Ladefoged and Broadbent's study. Common experience tells us that the act of adjusting to a new speaker (or removing undershoot) is automatic and unconscious. 'Explicit learning' (Dienes and Berry, 1997), whereby we can describe the acoustical cues used to identify phonemes does not seem to occur. Most listeners have little overt knowledge regarding which cues they use during phoneme identification. This appears to be the case whether listening to one's first or a second language, even though acoustical learning during second language acquisition is more deliberate (Ellis, 1994). Supporting this is the fact that gaining explicit knowledge of the nature of distortion (e.g. being told of a phoneme shift by orthographic representation) does not help the listener undo its effect (Summerfield and Assmann, 1989). This calls into question whether normalisation is 'learning'. Some form of short-term learning in the sense of 'identifying' useful information from immediately prior speech and preserving this memory might be involved in the normalisation seen, but any such learning if it occurs, appears to occur implicitly. If normalisation is an implicit process, this increases the chances that it is a lower-level process that birds might share, rather than a cognitive process.

An example of cue 'learning' is seen in studies examining the usefulness of 'point vowels'. Point vowels (e.g. /i/ /a/ /u/) represent extremes of articulation space for the speaker and can act as reference points for the VT area (Verbrugge et al., 1976). These cues may be learnt through experience with the speaker to build a formant-space map. Evidence for such learning comes from studies that show that experience with point vowels enhances the identification of target vowels (Liberman, 1973) and Gerstman (1968) accurately classified the data of Peterson and Barney using calibration based on point vowels. However, point vowels are not always useful for identification (Verbrugge et al., 1976) and were not a favoured explanation by Ladefoged and Broadbent (1957). It is therefore assumed that better recognition after hearing prior speech by the same speaker results from extracting different specific cues or wider information about the speaker's spectral space.

Rather than hypothesising that their effects were due to a learning of specific cues, Ladefoged and Broadbent observed that their results are like those in colour perception, where normalisation is caused by neural adaptation (Foster, 2011). Neural adaptation occurs in single cells and populations of neurons in lower and higher brain regions (Antunes and Malmierca, 2014). Importantly, neural adaptation can provide effects of a similar nature to the perceptual shifts in overshoot and VTN - it appears to directly explain 'relative perception' (Smith, 1979). If normalisation involves neural adaptation this may mean a more basic mechanism, rather than a higher-level cognitive process. Such a process is more likely to be shared with birds. Lindblom and Studdert-Kennedy (1967) specifically hypothesised that overshoot involves peripheral neural adaptation. If the process has a peripheral source, it is even more likely to be a basic process that might be shared with birds, as the peripheral neural system is generally simpler.

Experimental evidence indicating that VTN could be a result of peripheral neural adaptation comes from studies investigating overshoot. The overshoot effect, described above, appears to be a manifestation of the 'enhancement effect' (also known as the 'negative auditory after-image') (Summerfield et al., 1984, 1989). The enhancement effect describes the perceptual enhancement of energy change between short (phoneme-length) sounds heard in sequence: if a short sound with more energy in one frequency region and less in another is heard immediately prior to a sound with equal energy in all regions, the regions where there were peaks in the prior sound will be diminished in perception of the current sound and the areas where there were troughs will be enhanced. For example, in speech perception, where there is a peak in energy in a phoneme (e.g. a formant), this energy is relatively diminished in the immediately following phoneme. However, areas where there was previously lower energy (e.g. frequency regions adjacent to the formant) this will be relatively enhanced.

A number of studies appear to show this enhancement effect directly explaining overshoot. In a study by Mann (1980), real speech sounds from a ga-da continuum (where F3 varies between the two phonemes) were more likely to be perceived as /da/ (F3 energy at a higher frequency) after /ar/ (F3 energy at a slightly lower frequency) and /ga/ (F3 energy at a lower frequency) after /al/ (F3 at a slightly higher frequency). The same effect has been observed for vowels (Holt, 1999). In this type of research the effects are termed 'spectral contrast' effects rather than the 'enhancement' effects or 'overshoot' but Holt and Lotto (2002) explicitly note the similarity of spectral contrast effects and the enhancement effect, and it is clear from spectral contrast studies with speech, that the same process is probably behind the overshoot effect.

There is evidence that these shifts are caused by peripheral neural adaptation. Studies on the enhancement effect show that if the prior sound is presented to one ear it does not affect the sound at the other (Summerfield et al., 1984). This is evidence that the process might occur at the periphery before sounds from both ears combine. The effect has also been shown to have a very short time course (about the length of a phoneme), which is indicative of peripheral adaptation (Summerfield et al., 1984). 'Simple neural adaptation' (a decreased firing rate with continuous stimulation in frequency sensitive channels) at the level of the auditory nerve (Smith, 1979; Kiang et al., 1965) or

adaptation of suppression (Viemeister and Bacon, 1982) have been proposed explain the effect (Summerfield et al., 1987) and neurophysiological evidence of peripheral adaptation during enhancement has been reported (Palmer et al., 1995). Spectral contrast studies also confirm the short time course of spectral contrast and its (largely) non-crossaural nature (Holt and Lotto, 2002). These studies also show the same sort of shifts with non-speech sounds, which means that the process is not speech-specific and therefore potentially available for bird vocalisations (Lotto et al., 1997 – See Section 2 for a description of this study).

Given the above research, there is good evidence that peripheral adaptation explains overshoot in humans. Further, this appears to be a low-level process that is not speech specific. Therefore, this mechanism is potentially available to birds. However, it is not clear that birds experience problems similar to co-articulation/undershoot when listening to birdsong. Therefore, birds may not possess such an 'overshoot' mechanism. Importantly to the topic of this paper, it was suggested that the same mechanism might be responsible for VTN in humans, and therefore also available for birds for VTN. However, contrary to suggestions by Summerfield and Assmann (1989) it is not certain that this process can in fact explain VTN. The short time course of peripheral adaptation means that the perceptual shifts can only be in response to an immediately prior sound (which cannot fully contain the VT characteristics). The VTN shifts in Ladefoged and Broadbent's study appear to be based on the global spectrum of a whole sentence rather than an immediately prior phoneme. Therefore, it is not clear that studies on overshoot show a non-speech-specific peripheral process behind VTN. VTN might require a mechanism of greater complexity which would be less likely to occur in birds.

Other potential mechanisms of human VTN have been examined. Specifically, the mechanisms behind shifts that occur after a longer (sentence length) segment of prior speech have been investigated. Watkins (1991) used precursors similar to Ladefoged and Broadbent's sentences and showed that a central rather than peripheral process causes perceptual shifts suggestive of VTN. In Watkin's study, precursor sentences were filtered by the inverse spectrum of an /e/ vowel or the inverse of an /i/ vowel to mimic speakers with different VTs. Test words drawn randomly drawn from an "Itch" - "Etch" continuum followed the precursors. Increased 'Etch' perceptions were observed when the precursor sentence was filtered with the inverse spectrum of /e/ and increased 'Itch' perceptions occurred when the precursor was filtered for inverse /i/. Shifts were the same in nature as seen in Ladefoged and Broadbent's study – low energy regions in precursors were heard as enhanced in test sounds and vice versa. The author described these shifts as showing a "restoration of perception of a sound filtered by a speaker to that of an unfiltered context" – i.e. VTN. Additionally, they noted that the shifts were in response to the long-term average spectrum (LTAS) of the precursor sentence (Watkins and Makin, 1994), rather than just the immediately prior segment of sound. This shows that the general spectral characteristics of the speaker are normalised for, as would be expected by a VTN process. Importantly, unlike with overshoot effects, the shift was largely cross-aural. This cross-aural shift must be caused by a central mechanism because peripheral adaptation effects cannot manifest cross-aurally. Further, maintaining a picture of LTAS requires a 'sampling' of the speaker spectrum over the course of the sentence, which is likely to be beyond the time window of peripheral adaptation. The authors offered auditory memory as an explanation for this effect (Watkins and Makin, 1996a,b). While memory of some kind may be required to establish a 'spectral average' and short-term timbral memory can store detailed spectral representations of sounds for short periods (Cowan, 1984), the memory explanation is called into question by a number of features of this effect: 1) it was found to require spectro-temporal variation in sounds (Watkins, 1991). It is far from certain that memory requires spectro-temporal variation in sounds for them to be remembered (though see Pike et al., 2014); 2) shifts do not occur when the prior sentence and test sound appear to come from different direction: memory is not affected by the direction of the stimuli being remembered. However, this aspect might be explained by co-occurring perceptual ungrouping effects (Bregman, 1990), leaving a memory explanation intact.

Other studies support the apparent centrally produced VTN seen in Watkins' work. These studies have further elaborated on mechanisms behind this process (Holt, 2005, 2006; Laing et al., 2012). Holt's (2005, 2006) studies show that precursors consisting of a sequence of 21, 70 ms sine tones with an average frequency centred on the either F3 offset frequencies of /al/ or /ar/, caused shifts in /ga/ to /da/ test sounds, as seen in Mann's (1980) study described above. It was clear that shifts in response to the mean frequency of the whole precursor occurred, rather than just the frequency of the tones immediately prior to the test sound. Further, the effect did not break with a more than 1.3 s gap between precursor and test, which appears to rule out peripheral adaptation as this is usually shorter lived. This effect was not tested for its cross-aural nature but was claimed to be central due to is long time course. Holt put forward adaptation at central sites – specifically 'Stimulus Specific Adaptation', which occurs in the primary auditory cortex (Ulanovsky et al., 2003, 2004) as the mechanism behind this effect. Such high-level adaptation is more suited to explaining VTN than auditory memory, as adaptation has a nature that perfectly explains relational processing and the shifts seen. Again, the non-speech sounds used in Holt's studies mean that such a mechanism may occur with birds and further research has supported this with musical sounds (Stilp et al., 2010). However, such higher-level processes may be less likely to be found in birds.

Additional evidence of higher-level mechanisms causing VTN comes from studies looking at VTN mechanisms more generally. Wong et al. (2004) showed that 'central' speech processing areas of the cortex are engaged when new speakers are presented. In Nusbaum and Morin (1992)'s study participants were measured on memory for words. Mixed speaker conditions showed more error and this was concluded to be due increase attentional demands due to variability. They stated that 'speech perception requires 'active' processing to reduce the set of possible responses to a single response'.

2.1.4. Normalisation section conclusion

In humans, perceptual constancy across VT variation among speakers may occur via intrinsic or extrinsic normalisation processes. Perceptual normalisation involves the removal of variation, resulting in cues being perceived as the same across speakers and 'cannoical'. However, this process implies a loss of perception regarding the speaker VT characteristics.

Little is known about the precise mechanisms involved in intrinsic normalisation but if listeners can find reference cues to VT dimensions within the phoneme in question then instant normalisation of shifted cues may be possible. Normalisation appears to benefit from information outside of the phoneme. Either particular reference cues are extracted from prior speech for this purpose, or a more complex spectral average for the speaker is captured using auditory memory and/or higher-level neural adaptation. These processes appear not to be speech specific but applicable to all sounds so they may occur with bird song. However, the high-level nature of some of these may mean that they are not shared with birds. Compensation for undershoot (the overshoot effect) is more likely to be a process which uses a low-level peripheral neural adaptation. This makes it more likely to be available to birds, however, it may not be effective in reducing between-speaker or between-bird VT variation due to an inappropriate sampling duration. It might only function to reduce any 'co-articulatory' effects that humans and birds experience.

2.2. Non-normalisation methods

Non-normalisation methods describe other mechanisms by which the listener can correctly identify phonemes despite variation. Unlike

normalisation these methods either do not appear to involve a reduction in perceived cue variation or involve only a partial reduction. Therefore, these methods may be useful for maintaining perceived variation in the signal while still accurately categorising speech sounds in order to recognise them. However, the extent to which they fully allow for accurate recognition is unclear. These methods are 'non-analytical' (Pisoni et al., 1997; McClelland and Elman, 1986) and may instead involve a 'statistical' assessment of the correct category of the sound for recognition.

2.2.1. Categorical perception/PME

Categorical perception (CP) is a general perceptual process that explains variance reduction in speech. CP describes the effects seen when listeners are asked to identify speech sounds taken from an acoustical continuum, using distinct category names (e.g. /b/ or /d/). With strict or 'traditional' CP, listeners hear only one type of sound and thus label the sound consistently (e.g. /b/) until a distinct point in the continuum where they hear another sound (e.g. /d/). Additionally, for traditional CP it must be shown that listeners are not sensitive to any variation in sounds that are assigned same category label in discrimination tasks (e.g. they cannot hear the difference between two phonemes that are variations of /b/). Therefore, via this process within-category variation appears to be lost but between-category variation is maintained. Examples of traditional CP are most readily seen in the perception of consonants. Consonants that vary continuously in a parameter used to distinguish between two different consonants e.g. (F2 transition) result in listeners hearing one consonant until a distinct point where they hear another (Ainsworth, 1988; Liberman et al., 1957). Categorical perception of vowel sounds however tends to be less strict with perception of within-category variation remaining alongside categorical labelling (Liberman et al., 1957; Pisoni, 1973).

The mechanisms of traditional CP are not certain. It was originally believed that CP was unique to speech sounds because it was created by a psychological link between speech perception and the articulatory processes used to produce speech ("motor theory of speech" Liberman et al., 1967; Galantucci et al., 2016). This theory is no longer in favour as CP has been found to occur for non-speech sounds. For example, Cutting and Rosner (1974) show that sawtooth waves are heard as plucked when a continuously varying rise time is below 40 ms but as bowed, when it is greater than 40 ms. Miller et al. (1976) has also shown CP for noise/no-noise sounds where the time delay between a noise and buzz was continuously varied. Traditional CP may instead come about from natural discontinuities in the auditory system, or in other aspects of the perceptual system. Such discontinuities have not been found within the peripheral hearing system thus far. For example, the basilar membrane does not treat sounds discretely. However, Fujisaki and Kawashima (1970)'s memory model proposes that CP may be due to the differential use of particular memory stores. Specifically, they explain why there is more CP for consonants compared to vowels: for constants, it is the temporal variation in formants and other temporal cues (e.g. voice onset time) that tend to be important. The time varying nature of these cues means that each portion of the cue is only present a short time. An auditory memory store that stores a detailed representation of cues appears to exist but it is proposed that this store cannot grasp these time varying features for long enough to preserve them. At the same time, a less detailed categorical or 'phonetic' store exists. This appears to have a quicker temporal resolution being able to grasp and store short/changing cues but in less detail. Therefore, for consonants with their time varying cues, only less detailed categorical perception remains but vowels with more static formants can be grasped by the detailed store as well as the phonetic store, allowing for better within-category discrimination of two acoustical similar vowels, (at least if presented in close succession – Pisoni, 1973). Therefore, auditory memory may be a good explanation of traditional CP.

Normalisation appears to be another candidate explanation for CP because during normalisation cues appear to get perceptually shifted towards a canonical location, which would result in the perception of speech belonging to unique categories. However, there is no known research on whether normalisation processes explain CP. Further, it is posited that there is a conceptual difference between the two processes. In categorical perception, it's not clear that acoustical cues are perceptually *shifted* before the category judgement. It appears that all cues remain in place, category judgements are made based on the most appropriate category given the cues, and once this occurs, the perception of within-category variation is lost. Under such a process, if one or many cues are shifted by distortion to a location which puts them in a frequency region suggestive of a different (incorrect) phoneme, CP will not move those cues back to the correct location, but categorical perception will still occur and incorrect, but categorical, identification will occur (Pike, 2015). Normalisation, on the other hand will shift cues back to the canonical location, after which a categorical like perception of the correct phoneme can occur because cues have been shifted to their canonical location. Therefore, normalisation appears to do more than CP alone – it both corrects the distortion and results in categorical and canonical perceptions. It can explain correct perception in regions of ambiguity (e.g. vowel space overlap).

Less strict versions of CP also exist and these are thought to come about through learning to categorise sounds over time. The Perceptual Magnet Effect (PME) describes CP with softer boundaries, whereby category labels are not so readily assigned, and within-category variation remains accessible. More specifically, PME describes a warping of perceptual space where within-category variation is gradually more difficult to discern the closer the sounds are to the centre of the category (the prototype). Evidence of this being due to learning comes from infants at 6 months who after language learning, show PME (Kuhl, 1991) and the fact that within-category variation perception appears reduced after learning language but, it has been shown that categories are not fixed and immersion in a second language can create new categories (Flege et al., 1999; Werker and Tees, 1984).

2.2.2. Redundancy in cues and categorisation

In spite of VT variation, it is possible that no or little adjustment needs to occur. It may be that sufficient spectral/non-spectral cues within the phoneme remain to allow for accurate identification.

Firstly, cues that are particularly robust to VT distortion may be more heavily relied on. Temporal cues are robust to distortions affecting spectral attributes. For example, cue length usually has a limited role in identification but may be weighted more heavily when other cues are distorted. This effect was illustrated by Ainsworth (1972) who showed that duration effects were most prominent when a target vowel was in the centre of F1 and F2 space, where it is subject to more overlap. Further spectral-transitional cues may not be distorted to the same extent by general VT variation, which is a static distortion, and may play a larger role in identification in the case of VT differences.

Secondly, there may be sufficient cue redundancy in each phoneme. 'Extended target theories' of speech perception state that the overlap in formant space is not a problem because cues other than lower formants can identify the sound. For example, if F1 and F2 are distorted the listener can use other cues such as higher formants (Strange, 1989) or spectral tilt (Kiefte and Kluender, 2005). Therefore, there may be overlap in F1 and F2 space but not in the categorisation of phonemes. Some of the strongest examples of cue redundancy can be seen in the accurate perception of "sine-wave speech" (harmonically reduced speech, produced by tracking the frequency and amplitude of the first 3 formants over the course of a sentence and replacing all information with sine waves representing this tracking – Remez et al., 1981), telephone speech with a bandwidth of 300–3000 Hz, and in studies showing that speech is perceived with 90% accuracy where only information below 800 Hz and above 4000 Hz remains (Lippman, 1996).

Elaborating on redundancy theories, Repp's trading relations theory describes how speech cues can be weighed against each other (Repp, 1982). Repp states that speech forms categories and cues to speech

categories have different weights but no cue is essential – a cue can be moved so that the phoneme favours one category and this can be offset by moving another cue to favour another category. For example, a shift in a cue (e.g. F1) from the canonical location to a higher value could result in the perception of a different phoneme (i.e. one with generally higher formants) but this would not result in this perception if a different cue (e.g. F2) was shifted in the opposite direction (potentially signalling a phoneme with lower formants) as both shifts would counter-balance each other. This demonstrates robustness to distortion of some cues, as long as the distortion is balanced. It is not clear how such a process would deal with VT distortion as this is not balanced – cues would be shifted in a particular direction. However, this finding demonstrates flexibility of cues and appears to describe a statistical weighting process is behind placing speech in categories.

"Categorisation" theory is a further elaboration of the cue weighing described by Repp and the redundancy described by extended target models. Categorisation theory implies that all cues are relevant and speech is recognised by statistical pattern matching to 'exemplars' or 'prototypes' of speech stored in memory (Samuel, 1982). Holt and Lotto (2010) states that: "speech stimuli are represented by continuous values, as opposed to binary values of the presence or absence of some feature. Speech perception is the process that maps from this space onto representations of phonemes or linguistic features that subsequently define the phoneme. This is an example of categorisation, in that potentially discriminable sounds are assigned to functionally equivalent classes." Holt implies something other than a loss of sensitivity to within-category differences within her description of categorisation. In fact, it appears that in order to distinguish categorisation from CP and PME, an assumption of no or little perceptual loss of variation is necessary. Studies showing that some exemplars are identified as better exemplars of phonemes than others may be evidence of within-category discrimination ability remaining in speech perception (Kuhl and Iverson, 1995). Further, physiological studies and eye-tracking studies show variation in the speech signal remains perceptible during recognition tasks (Holt and Lotto, 2010).

The mechanisms for speech categorisation are likely to be similar to categorisation in other cognitive domains. For example, researchers looking at implicit learning have researched categorisation more generally and have shown good identification of objects when specific examples of the objects have been stored from previous tasks but worse performance with more novel stimuli. This would be expected if speech recognition takes place by matching to new speech to previously heard examples. Nygaard and Pisoni (1998) show that experience with a particular speaker means enhanced word recognition for familiar words in noise, but not novel words, suggesting that pattern matching of new sounds to stored examples takes place. Such an effect cannot be explained by experience with the speaker's vowel space during the test as this occurred in both conditions. Maddox et al. (2002) found processing in the striatum was involved in unconsciously storing past examples and matching of new speech to these.

3. Vocal tract constancy in songbirds

As the mechanisms of constancy do not appear to be specific to speech, they are potentially processes shared with other animals, especially those that produce vocalisations which bear some similarity to human speech, like songbirds (Doupe and Kuhl, 1999). Maintaining constancy of vocal signals may be critical for survival and reproduction in birds as vocal signals are the key source of communication in birds. Similarities in constancy mechanisms between humans and birds may be evident if these mechanisms are basic processes of the auditory system that are evolutionarily conserved or because birds have separately evolved these mechanisms for their own vocalisations. Bird vocalisations are broadly divided into two categories; songs and calls. Songs and speech are comparable in the sense that both need to be learned from conspecifics, mostly during early life (Doupe and Kuhl, 1999). Speech and songs are also both composed of hierarchically structured units. In birdsongs, notes (also called elements) are the smallest unit of continuous sound that are concatenated to form syllables, phrases, and motifs – similar to how phonemes in human speech are combined into syllables and words (though this similarity is somewhat superficial – Berwick et al., 2011). Critically, both songs and calls are used to communicate a message to the receiver. Songs are predominantly sung to attract mates, repel rivals, and defend territories, so the information contained in songs is limited to these contexts (e.g. "I am good at keeping this territory from others so I am probably also good mate"; Searcy and Andersson, 1986). For songs to effectively serve their purpose they should convey information about a singer's characteristics, such as species identity, quality as a mate, and strength as a competitor. Song rate, syllable or song-type repertoire, and song pitch have been found to correlate with singer condition, learning ability, and body size, respectively (Saino et al., 1997; Nowicki and Searcy, 2005; Linhart and Fuchs, 2015). Songbird calls may be more similar to human speech in terms of the diversity of messages that they can relate. Calls – which are also composed of notes and syllables – are used in a variety of contexts such as maintaining social contact, pair bonding, and signalling predator presence and therefore have different functions. It has been shown that modifications of calls are sometimes made to affect the message that is being conveyed: for example, black capped chickadees convey information about the size and threat of approaching predators by varying the notes within the call (Templeton et al., 2005). This is contrary to song, where modifications are usually not used to change their message (but sometimes messages regarding level of aggression are made via altering amplitude and song performance – Searcy et al., 2006, DuBois. et al., 2009). Both types of communication may vary between birds less deliberately, via the acoustical modifications caused by the environment or via VT differences. This can affect how well messages are recognised by other conspecifics as a song, or how well they understand the message contained within a call. Hence, we explore the possible utility of VT constancy for the recognition of bird songs and calls; even though calls are not conventionally viewed as being learned and consequently not compared to speech, constancy may still be required for recognition of the message. Finally, in additional to functional similarities, songbird vocalisations can also be acoustically similar to human speech. At a very general level, they are time-varying, complex, frequency-modulated sounds. At a more specific level, some songbird vocalisations, such as the zebra finch, exhibit formant-like peaks that look like formants in human vowels (Elie and Theunissen, 2016).

Like human phoneme recognition, syllable recognition in songs and calls must be robust to a range of modifications to allow songbirds to make appropriate behavioural responses. This includes acoustic modifications caused by variation in signaller characteristics, which could potentially stem from variation in VTs similar to human speech. Compared to other causes of inter-individual variation in acoustic realisations of notes and syllables within a species (such as social environments that produce regional song dialects; Marler and Tamura, 1964), the physical factors related to the mechanics of sound production have received much less attention yet are likely to be important (e.g. see Podos et al., 2009). Here, we argue that one physical factor that is expected to significantly contribute to inter-individual variation is VT size and shape. That is, individual differences in VT dimensions should increase the acoustic variability of songs and calls by systematically distorting the acoustic realisations of notes and syllables. In birds the VT consists of the trachea, larynx, and beak (Podos, 2001). A number of studies have shown that there is considerable variation in the length and size of the VT of individual birds of the same species, in both songbirds (zebra finch, Riede et al., 2010; European starling, Prince et al., 2011) and non-songbirds (whooping crane, Fitch and Kelley, 2000; herring gulls, Hardouin et al., 2014; oilbirds, Suthers, 1994). For instance, the VT lengths of whooping cranes can range from around 10 cm in juveniles up to 147 cm in an adult, with older individuals

having longer VTs (Fitch and Kelley, 2000). Of particular interest to the question in this paper is the fact that between-adult variation is also notable. There is between species variation – for example species with larger beaks will produce songs emphasizing lower frequencies (Podos, 2001). But there is also within species variation – for example, Geospiza fortis show bimodal variation in beak size, with birds with larger beaks showing more limited bandwidth and reduced vocal performance (Podos, 2001; Huber and Podos, 2006). It has previously been suggested that the vocal tract served to supress overtones contributing to the pure tone of most avian species (Huber and Podos, 2006; Riede et al., 2006; Nowicki, 1987; Nowicki and Marler, 1988) However, work by Ohms et al. (2010) describes increasing evidence that vocal tract filtering is a relevant dimension in bird song as well as speech. In Ohm's study, linear sound sweeps where passed through the VTs of euthanized zebra finches with varying beak gape and oropharyngeal-esophageal cavity expansion. The LTAS was calculated and change in spectrum of the sound with narrower beak gape (there was filtering below 6 kHz) and narrower OEC (amplitude decrease around 5 kHz), was observed.

Additionally to similarities between humans and birds regarding the purpose of vocalisations and the acoustical structure of vocalisations, there are similarities in terms of vocal production problems. Blurring of boundaries between syllables and overlap of syllables in acoustic space reminiscent of overlap in vowel categories in human speech (e.g. Peterson and Barney, 1952) has also been observed in songbird vocalisations (Williams et al., 1989; Elie and Theunissen, 2016), but it is not known how much of this is due to inter-individual variation in VT dimensions per se. During song learning, the songs that juveniles hear are influenced by tutor characteristics, which could be difficult or impossible to reproduce exactly given that the juvenile and tutor have different VT shape and size. As a result, juveniles may seek to learn songs in a "tutor free manner", storing songs in a canonical form from which they can more easily reproduce. Similarly, VT differences between individuals may make song recognition more difficult because the same syllable produced by different singers will sound different. Thus, for the sake of accurate recognition, the listener may want to remove variation caused by the singer.

3.1. Normalisation methods

Birds may have similar ways of dealing with these problems as humans and there are good reasons to believe that songbirds may be able to perceive VT dimensions from vocalisations. In this section we will explore the methods that songbirds could use to compensate for acoustic variability caused by VT dimensions. For the methods of normalisation and non-normalisation discussed in the human section, we will discuss whether birds may be using these mechanisms in the perception of own-species vocalisations.

3.1.1. Normalisation by vocal tract length

Normalising individual differences in VT length may be one mechanism that songbirds use to maintain constancy of notes and syllables. To show that songbirds could use VT normalisation as a mechanism, it would first be necessary to show that VT length contributes significantly to acoustic variation of vocal units. This has been demonstrated in mammals and whooping cranes (Reby and McComb, 2003; Riede and Fitch, 1999; Fitch and Kelley, 2000), but not songbirds. Subsequently, it would be necessary to show that within-category scatter caused by inter-individual differences in VT length can be eliminated by scaling the category-defining properties of songbird vocal units to fundamental frequency (F0), or an equivalent to third formant frequency (F3), or formant dispersion – these factors have been correlated or causally related to VT length in past studies (Potter and Steinberg, 1950; Nordström and Lindblom, 1975; Fitch, 1997). Several studies have shown that songbirds are sensitive to F0 and whooping cranes notice when formant frequencies of their calls were shifted to mimic a different VT length, suggesting that they perceive information about an individual's VT length through formant dispersion (Fitch and Kelley, 2000). Many songbirds produce vocalisations that contain harmonic overtones, which makes the acoustic correlates of VT length within these (e.g. formant dispersion; Fitch and Hauser, 2002; Fitch and Hauser, 2001; Williams et al., 1989) potentially available to perception and use by the receiver.

3.1.2. Normalisation by formant ratios

Perception of formant ratios is a normalisation mechanism also conceivably employed by some songbirds. Zebra finches may be good species to study formant ratio normalisation, as both humans and zebra finches are capable of categorising sounds that differ in frequency ratios (Weisman et al., 1994). In addition to being sensitive to relative frequency ratios (which are important for perceiving speech from different speakers), zebra finch vocalisations exhibit formant-like peaks, with relative differences in these formant-like peaks present in different vocalisations, suggesting that they may distinguish formant ratios in their own songs and calls (Elie and Theunissen, 2016). If zebra finches use formant ratios for perception of their own vocalisations, then their transferral of this process for use with human vowels could explain why zebra finches appear capable of spontaneously adjusting their perception of vowels to different speakers (Kriengwatana et al., 2015; Ohms et al., 2010).

Williams et al. (1989) has also shown in zebra finches that i) multiple renditions of a given song syllable show consistency in the suppression and emphasis of the amplitude of specific harmonic frequencies. This produces varying peaks and dips in the power spectrum or 'timbres', similar to formants in human speech; ii) these patterns differ between song syllables; iii) these patterns were learned, as juveniles copied these patterns from adult models. Thus, given the importance of formants and complex spectral structure in speech, this research suggests that patterns of harmonic frequencies "are among the song characteristics important for zebra finches". However, as Williams et al. note, while differences in the amplitude of a signal harmonic (e.g. 2nd harmonic) are perceived, the perception of more complex timbres has not been directly measured in zebra finches or other birds, and the use made of timbres is not well understood. However, the fact that this timbre is learnt by juveniles implies that complex timbres are perceived and may indeed be useful in conveying a message.

Interestingly, zebra finches are capable of detecting a single missing harmonic in their vocalisations (Cynx et al., 1990; Uno et al., 1997; Lohr and Dooling, 1998), suggesting that they may be paying attention to the amplitude of particular harmonics rather than the relationship between harmonic amplitudes. This could enable accurate recognition without the need for formant ratio normalisation. Hence, if they recognize and discriminate syllables using this strategy, then future work needs to determine the function of formant-like peaks their vocalisations (Elie and Theunissen, 2016) and what role, if any, formant perception and formant ratio perception plays in zebra finch syllable recognition.

3.1.3. Normalisation by "point vowels"

Instead of scaling perception of vocal units to VT length using intrinsic cues, songbirds might scale their perception to an individual's vocal-articulatory range, calibrating their perception to vocal units that occupy the most extreme positions in articulatory space (such as point vowels in speech). Although there are no investigations into the use of this mechanism in songbirds, data from other studies suggest that it is unlikely that this mechanism is required for recognition. In operant conditioning experiments songbirds can classify isolated vocal units produced by unfamiliar individuals (Sturdy et al., 1999, 2000), which indicates that constancy is feasible even without calibration to an individual's articulatory space via experience with extreme VT positions. However, it is possible that the songbird equivalent of point vowels was among the vocal units that were tested experiments, so at the moment this mechanism cannot be ruled out entirely.

Despite the results from operant studies going against the idea that information extrinsic to the vocal unit aids classification, there is evidence that familiarity with a singer's song influences perception. Cynx and Nottebohm (1992) reported that for zebra finches, songs that were more familiar were easier to discriminate. Specifically, male zebra finches were faster to discriminate between two songs when one was their own; more trials were required for zebra finches to discriminate between the songs of familiar birds, and even more were required for the songs of unfamiliar birds. Unexpectedly, females did not show a familiarity effect, a finding which is not predicted if birds normalise based on an individual's coordinates. While this study shows that familiarity is helpful, it is not clear whether the facilitation is due to the use of "point vowels" - songbirds could be normalising inter-individual differences by using an external reference system based on a singer's vocal articulatory coordinates obtained via different cues - or other information such as the use of the whole spectral range or rate of information. Additional research to support the idea that birds create perceptual frames of reference for each singer would be if they showed faster or more accurate recognition of a syllable if they had previously heard another syllable from the same singer, as well as the existence of scaling methods that can successfully shift critical acoustic features in songs into a singer-specific coordinate system (see Johnson, 2005).

3.1.4. Normalisation by spectral contrast mechanisms

In 1997, Lotto and colleagues showed that a non-songbird (Japanese quail) appeared to show effects indicative of compensation for co-articulation when listening to speech, just as was shown by humans in 'spectral contrast' studies. Quail were trained to respond to /da/ or /ga/ , which were synthesised as a ten-step series that varied in F3-onset frequency (/da/ with high and /ga/ with low F3-onset). These syllables were preceded by one of three synthesised syllables /a/, /al/, or /ar/. Like in human experiments, quail perceived syllables with intermediate F3-onset frequencies as more similar to /ga/ (low F3-onset) if the syllable was preceded by /al/, which had high F3-offset frequencies. In contrast, they perceived the same syllables as more similar to /da/ (high F3-onset) if it was preceded by /ar/, which had a low F3-offset frequency. Thus, 'spectral contrast'/overshoot between neighbouring phonemes appears to occur and may be sufficient for quail maintain constancy when perceiving human speech. As explained earlier, this mechanism could be the same as that used to compensate for speaker differences in speech, so it is possible that spectral contrast plays a role in maintaining constancy in songbird vocalisations, across VT variation as well as allowing songbirds to deal with any co-articulation effects. The study by Lotto used human speech and to the best of our knowledge, researchers have not directly tested whether such contrast effects are used by songbirds to maintain constancy of their own vocal units. To find support for spectral contrast effects in songbirds, the spectral envelope of a preceding song syllable should influence classification of a following target syllable in species with low to moderate levels of syntactical structure of song syllables where the order of song notes and syllables are not highly predictable, such as the zebra finch or Bengalese finch (Lachlan et al., 2016; Wohlgemuth et al., 2010).

3.1.5. Recognition by extended experience

Extended experience may also facilitate recognition by providing context for which to evaluate the sound and the opportunity for a learning of phonological sequences or syntax. For example, zebra finches were better at detecting changes in timbre in a syllable if it was embedded in a whole song compared to if the syllable was presented in isolation (Nottebohm et al., 1990). Lachlan et al. (2014) showed that the position an ambiguous syllable occupied in a song influenced how likely it was to be treated by swamp sparrows as belonging to one of two distinct categories. The use of phonetic context to maintain constancy of vocal units may be widespread in species where there is some predictability in how syllables are ordered, including in humans, who use phonological rules that govern the sound sequences permissible in a language to segment words in running speech (Saffran et al., 1996). Most probably, the use of context will not be as useful for species where syllables can be arranged in any order, suggesting that species may vary in which constancy mechanisms – including the ones discussed here – are likely to be employed.

3.2. Non-normalisation methods

Normalisation implies that listeners "clean" the signal by removing individual differences in how the signal is produced. However, singer-related cues provide important information that is worth retaining in speech and song. For example, singing is a demonstration of the ability to very precisely control and coordinate VT movements and breathing and is more demanding for songs with syllables that cover a large range of frequencies and are rapidly repeated. Thus, individual differences in syllable production can reflect differences in motor ability (Podos, 1996). Syllables that are difficult to produce have been found to be especially attractive to females (Vallet and Kreutzer, 1995, 1998), which suggests that individual differences in vocal production are important indicators of signaller quality and potentially sexually selected (Podos et al., 2009). Body size is another cue that songbirds may use to decide whether to engage with the signaller in competition or reproduction and is inherently tied to the dimensions of the VT which are proposed to be removed during normalisation (Hinds and Calder, 1971; Hall et al., 2013; Linhart and Fuchs, 2015; Riede and Goller, 2014). Hence, we examine evidence for other methods of maintaining constancy in recognition that do not assume that signaller characteristics in vocalisations are disregarded.

3.2.1. Recognition by categorical perception

Although once thought as a human and speech-specific characteristic (Liberman et al., 1957), categorical perception of sounds is now known to be widespread in the animal kingdom, occurring both in response to human speech (e.g. rodents, Kuhl and Miller, 1975; birds, Dooling et al., 1989) as well as conspecific vocalisations (e.g. frogs, Baugh et al., 2008; crickets, Wyttenbach et al., 1996; monkeys, May et al., 1989). Categorical perception of song note duration has been convincingly demonstrated in a songbird (swamp sparrows) at the behavioural and neural level (Nelson and Marler, 1989; Prather et al., 2009; Lachlan and Nowicki, 2015).

Although categorical perception could occur as a result of innate discontinuities in the auditory system, a study in European starlings showed that songbirds can also learn to exhibit categorical perception of human phoneme boundaries in a human-like fashion. Specifically, birds learned to respond to tokens of synthetic vowels within a category as more similar than vowel tokens from other categories, even though all vowel tokens were separated by the same acoustic distance (Kluender et al., 1998). Recent work supports the view that these perceptual learning effects are present in natural conditions, as swamp sparrow song syllable categories exhibit internal structure and conspecifics respond more strongly to syllable renditions that are good exemplars of that syllable type (Lachlan et al., 2014). However, the studies by Kluender et al. (1998) and Lachlan et al. (2014) are only suggestive of categorical perception and not direct evidence for it because animals were not tested in the way categorical perception is classically done (i.e. vary a stimulus that can be categorised along a single dimension and determine the steepness of the stimulus boundary). It is important to note that despite the evidence of categorical perception in animals, not all vocal units are perceived categorically by songbirds. For instance, unlike swamp sparrows, great tits do not perceive song syllable duration categorically (Weary, 1989). Thus, categorical perception may work together with other perceptual processes to maintain constancy, depending on the type of vocal unit and also perhaps the species.

3.2.2. Recognition by comparing to stored exemplars

In the 1990s, researchers suggested that birds may recognise the songs of conspecifics by determining how closely an incoming song matches stored representations of their own song (Williams, 1989; Cynx, 1993; Pytte and Suthers, 1999). We will refer to it here as the "own-song hypothesis". Young songbirds learn to sing by creating an internal auditory song template through a combination of innate predispositions and experience (i.e. listening to adult tutor conspecifics), and then learn to match their own vocal output to the memory of the song template (reviewed by Soha, 2016). This view that songbirds form a representation of tutor song is supported by neurobiological studies (reviewed in Bolhuis and Moorman, 2015). The caudomedial nidopallium (NCM) is an auditory forebrain region where processing and storage of auditory memories of conspecific vocalisation occurs (Chew et al., 1996; Gobes and Bolhuis, 2007; Mello et al., 1992), and distinct patterns of NCM activity are observed in response to different syllable types in the canary brain (Ribeiro et al., 1998). These results suggest that NCM could be the site where exemplary representations of vocal units are stored and matched to during recognition. For females that do not sing, the template that is used could be that of their father or mate's song (Cynx, 1993). Neurons in an auditory forebrain region critical for song learning respond selectively to a bird's own song, as well as other conspecific songs including tutor song (albiet less strongly, e.g. McCasland and Konishi, 1981; Margoliash, 1983; Theunissen et al., 2004), which provides neurobiological evidence for song template matching. Damage to this region also impaired conspecific song recognition (Gentner et al., 2000; Brenowitz, 1991). Chirathivat et al. (2015) also found that neural activity in response to song has also been found in young birds that have had no prior song exposure.

One prediction from this hypothesis is that discrimination between sounds should gradually worsen the farther they deviate from the bird's own song. This is supported by studies showing that songbirds respond differently to songs from different dialect and songs of neighbours versus strangers (Searcy and Andersson, 1986). It is not clear, however, how the own-song hypothesis accounts for how songbirds recognise conspecific syllables that are improvised (and thus often unique to individual birds; e.g. North American sedge wrens; Kroodsma et al., 1999), how birds that only sing a subset of song syllable types rather than all possible syllable types can recognise conspecific syllables that they themselves do not produce (e.g. zebra finches; Sturdy et al., 1999), or how songbirds can discriminate between sounds that completely different from their own song (such as human speech) and potentially unrecognisable as a song syllable. To resolve some of these shortcomings, the own-song hypothesis might imply specialised processing of conspecific songs by songbirds, such that sounds with spectral-temporal characteristics that fall within the range of a bird's own song is compared to the bird's song template, whereas sounds that depart significantly are processed by other auditory mechanisms unrelated to song perception. European starlings from the study on phonetic learning described earlier by Kluender et al. (1998) seemed to be categorising human phonemes based on multiple exemplars rather than a single exemplar (which would be the case in own-song comparisons), so evaluating how starlings categorise their own song syllables would be useful for clarification and development of the own-song hypothesis. Another example of matching to prior experienced song is seen in compensation for degradation of acoustic vocal signals by distance and habitat whereby it is thought that listeners deal with this distortion by comparing the incoming signal to an "undegraded" version of the vocalisation stored in memory (Morton, 1986). Support for this hypothesis include work by McGregor and Krebs (1984) showing that male great tits differentiated between degraded and undegraded songs of familiar neighbours but not of songs of unfamiliar individuals.

4. Discussion

Perceptual constancy allows humans and birds to recognise vocalisations despite modifications during production and transmission through the environment. We focused on how, in humans, perceptual constancy mechanisms ameliorate modifications caused by differing size/shape vocal tracts between speakers. We observed that bird vocalisations are also likely to be affected by VT variation and may also benefit from constancy. An aim of this project was to determine whether constancy mechanisms are shared between birds and humans and whether human VT constancy research can offer avenues for better understanding bird vocalisations and vice versa. We found a number of similarities that suggest that this is the case. Similarities in mechanisms could be due to convergent evolution or properties of the auditory system inherited from a common ancestor. We noted that shared mechanisms of constancy are more likely if they are peripheral in origin.

For both human language and bird song we described how vocalisations are used to a) send messages containing specific information, such as 'there is a threat over there' and b) provide information regarding signaller characteristics. Cleaning the signal of variation is clearly useful for receiving informational messages, but for perceiving signaller characteristics the situation is less clear. Birds and humans use vocalisations to convey signaller identity ('who they are' or 'what species they belong to') via what is being said (i.e. the pattern of notes in the song or saying one's name) but also via how something is said: humans use general VT characteristics between speakers (pitch and timbre), as well as time varying characteristics (accent) to identify speakers by their voice. Voice characteristics are also used for recognising the signaller in some bird species (e.g. Weary and Krebs, 1992; Vignal et al., 2008). Additionally, birds and humans use VT characteristics to give information about the 'physiological quality' of the signaller. A healthy human has as a pleasing timbre to their voice that is attractive to other humans (Bruckert et al., 2010). Birds can use information contained in pitch and timbre to judge body size (e.g. Fitch and Kelley, 2000; Linhart and Fuchs, 2015) and we suspect that the pureness of tone may also be used when judging traits such as physiological symmetry and quality. Differences in the relative importance of conveying messages vs conveying signaller characteristics may indicate differences in VT constancy mechanisms between humans and birds: for humans the main purpose of speech appears to be communicating via accurate delivery of linguistic messages (e.g. making a command, saying one's name) rather than communicating their identity/fitness through VT characteristics. Therefore, achieving constancy is more important than maintaining sensitivity to VT variation. For birds, vocal communication is more often used for the purpose of attracting a mate or repelling rivals so the fitness and body size information that is contained within VT characteristics would be more useful. This could mean that VT constancy is reduced in birds. However, it is noted that even for birds fitness information is probably primarily conveyed in aspects of the vocalisation other than general VT characteristics, such as the time varying VT movements (e.g. 'vocal performance'; Podos et al., 2009) and the song complexity (e.g. Nowicki and Searcy, 2005; Catchpole and Slater, 2008). Normalising for VT differences across speakers would not impede this.

This review examined how VT variation, while potentially useful, could disrupt the building blocks of messages. For both species, vocal building blocks are distinct segments which contain similar spectral cues to recognition: pitch cues and higher harmonics/formants. For songbirds pitch appears to be the most important cue (e.g. Lohr, 2008; Weary, 1989) and for humans energy at higher harmonics (e.g. formants) is more important. For both species the effect of varying VT dimensions between signallers can disrupt these cues.

We showed a variety of methods by which humans might maintain constancy across VT variation. Normalisation processes remove perception of the variation by a perceptual 'shifting back' of speech cues distorted by the VT. Intrinsic normalisation does this by using cues to VT dimensions contained in the same phoneme. Intrinsic normalisation methods are difficult to test in humans and birds, primarily because the normalisation happens within the same phoneme – so the time course is near immediate and the process and mechanisms are not easily

accessible (Pike, 2015). Further, Peterson and Barney's (1952) study and studies following this, show that perception of phoneme variation and ambiguity can remain where there is no prior experience with the speaker, so any intrinsic normalisation, if it occurs, is not sufficient to remove all overlap. It is extrinsic methods that provide the necessary additional normalisation to help to overcome this and these methods may be more interesting to investigate. Therefore we recommend that new studies initially focus on extrinsic rather than intrinsic normalisation.

Extrinsic normalisation involves the shifting back of distorted cues using information contained in a short segment of prior speech. This prior speech may provide the opportunity for learning of specific reference cues for normalisation (e.g. point vowels) but evidence suggests that extrinsic normalisation occurs via neural adaptation in response to the prior speech. A low-level peripheral adaptation (enhancement/spectral contrast) was suggested to provide VTN but this is unlikely because, while this mechanism shows normalisation in response to the prior spectral context, it has a short sampling frame: any normalisation via this mechanism is only in response to the spectrum of the immediately prior vocal unit and might not contain all the information about the speaker's vocal tract area, and its corresponding spectral profile, that is needed for VTN. This peripheral mechanism was concluded to provide overshoot effects, which allow for normalisation in response to co-articulation, rather than VTN (however, it is acknowledged here, that this process could still result in VTN via enhancing spectral change between individual phonemes (Pike, 2015)). As well as this short-time course process, a higher-level process also appears to exist, which has a time course more appropriate to sampling the whole range of the speaker spectrum (the long-term average spectrum). This more clearly provides perceptual shifts in response to the general spectral characteristics of the speaker (LTAS) and therefore may provide between-speaker VTN. The research discussed suggests that this is due to a memory process or higher-level neural adaptation.

Birds show perceptual effects that suggest that they use similar normalisation methods. Firstly, Japanese quail (a non-songbird) also showed typical 'spectral contrast' shifts as seen in human speech (Lotto et al., 1997). Therefore, the act of perception of spectral cues in light of prior spectrum does occur in birds, although it has yet to be explicitly demonstrated with their own vocalisations. Further, this study is evidence that birds undergo short-time course normalisation in response to the immediately adjacent sound (i.e. compensate for co-articulation effects) but not necessarily to the whole speaker. If birds experience co-articulation effects when listening to their own song, it seems they have a mechanism to compensate for this. However, it is not clear birds do experience co-articulation as their specialised vocal organ (the syrinx) allows for greater precision of vocal motor control and insertion of silent gaps (i.e. minibreaths) between syllables (Suthers, 2001). Regarding compensation for the general VT characteristics, we did not find studies that aimed to determine whether shifts in phoneme/song perception in response to the LTAS of a longer segment of speech occur in birds, and whether they might utilise a higher-level adaptation process to achieve this. Studies which extend the research of Holt et al using longer tone sequences (Holt, 2005, 2006) could be conducted in birds. If enhancement/spectral contrast-like shifts are found and these are: a) shown to be cross-aural and therefore central, and b) shown to change the perception of the phoneme relative to the long-term average spectrum of the prior sounds rather than the immediately prior spectrum, then this is suggestive of high-level VT constancy processes in birds. These studies would be simple to conduct by extending the paradigm used in Lotto et al's., 1997 study.

Humans might also use non-normalisation methods to compensate for VT differences. Categorical perception (CP) and Perceptual Magnet Effects (PME) remove a certain amount of variation due to the speaker (and other factors) but possibly do not produce the cue shifting necessary to reduce overlap in formant space. Therefore, they would not be a normalisation process, but might be called variance reduction processes. Categorisation is also used, whereby all cues available in the sound are used to identify the sound and there is reliance on there being sufficient undistorted information to correctly classify despite VT variation (and other distorting factors). This process leaves perception of the variation, but listeners are able to perceive sufficient relevant features of the sound to correctly place it in a category. This process is the most similar to that used by listening machines, whereby the whole listening context including the syntax and semantics of prior speech, as well as acoustical cues, is used to predict individual speech elements as part of machine-learning approaches (Cambria and White, 2014).

Regarding non-normalisation methods, songbirds have shown an ability to place sounds in categories like humans, but the extent to which these studies demonstrate traditional categorical perception of song syllables (showing distinct boundaries, and no within-category discrimination) or perceptual magnet effects (Kuhl, 1991) is often not clear. The perception of note duration by swamp sparrows is an exception, as it provides a clear demonstration of CP, as traditionally defined (Nelson and Marler, 1989). Studies which more specifically aim to test whether perception of bird vocalisations by songbirds is traditionally categorical may show that the apparent 'placing in categories' by birds is due to strict CP, possibly arising from discontinuities in the auditory system (or perhaps, as has been suggested for humans, in the memory system). Additionally, studies showing CP at a very young age would provide support for CP with an innate mechanism in songbirds. Otherwise, if songbird perceptual categories show looser boundaries (and more discrimination within categories) then this would be evidence of weak CP or PME, which is more likely to be a result of learning either at a juvenile stage or later. Repeating Lachlan et al.'s (2014) experiment – which showed that adult swamp sparrows responded differently to syllables depending on whether they were more or less typical examples of their category - with juveniles could provide useful insights into whether PME/CP-like effects are in place at an early age or must be learned during the course of song learning. Finally, if songbirds show a tendency to identify sounds as belonging to categories but can show very good sensitivity to within category variation, it may be that they are employing general "categorisation" processes. They may perform this categorisation by using statistical pattern matching to exemplars stored in memory, such as their song template, acquired from listening to adult conspecifics during development. In humans, categorisation appears to involve high-level processes. Therefore, it is not certain that songbirds perform this. However, evidence of categorisation might be suggestive of potentially sophisticated pattern matching mechanisms in songbirds.

Further, the potential for songbirds to exhibit accurate perception due to a tie between hearing song and motor commands (see the "motor theory of speech perception" Liberman et al., 1967) was not discussed in this review. However, it is recommended that the reader explores comparisons between humans and birds regarding the motor theory proposals for VT constancy (see Galantucci et al., 2016 for a review).

In the process of conducting this review we observed that there are number of differences between humans and birds that mean that VT variation for birds may not be the problem it is for humans. One issue, discussed above, is the increased utility of VT information for birds. Another issue is whether VT variation has as large an effect on bird vocalisations as it does on human speech. The human VT has more degrees of freedom – the mouth is capable of more movement and sound modification than a beak that is ridged and fixed in shape. Therefore, the potential for between bird VT variation is smaller.

Another reason birds may not require normalisation to the same extent as humans is that, for humans VT variation has a notable effect in shifting formants in different ways between people. However, if birds rely more on pitch cues than higher harmonics, any shifting of VT dimensions would not shift these as readily and there would be less disruption to recognition (Fant, 2001). Further, any pitch disruptions that do occur may be possible to fix by simpler processes than the adaptation to the whole spectrum described for humans. The reliance on pitch in

songbirds is relevant for the applicability of Lotto et al.'s (1997) results to bird song. Their study tested bird perception using formant rich human speech, so it is not clear if the process tested would be used by bids to normalise their own vocalisations, which is primarily pitch information. However, in a similar study (Huang and Holt, 2009), showed that humans normalise pitches as well as formants so spectral contrast effects may extend to pitch perception in birds. Further, Lotto's test was also conducted with non-songbirds, thus the study could be duplicated using song stimuli used by songbirds.

Future work should bear in mind that speech is a more complex signal and may require more/different mechanisms of constancy. Further, while other constancies (e.g. colour) might have been useful to humans since the beginning of human evolution, speech is a later evolving process so it is more likely that mechanisms have diverged between species. However, on balance this review concludes that are sufficient similarities regarding the purpose of vocal communication, the physics of the VT and the need for constancy between humans and birds to suggest that further research in this area would be fruitful. We firstly recommend that i) the effect of individual VT dimensions on vocalisations should be ascertained for birds. We also recommend that the relative importance of this source of variation in relation to other sources of variation (e.g. environmental) is established. Researchers could then ii) determine the extent to which VT variation is disruptive to messages within vocalisations and the extent to which it provides a source of useful information iii) if it is found to be disruptive, investigate mechanisms of VT constancy, starting with extrinsic methods as these appear to be most pertinent to constancy.

Statement of author contributions

The work in this review was divided as follows: Sections 1 and 2 were written by C Pike, Section 3, was written by P. Kriengwatana. Both authors contributed to Section 4.

Acknowledgements

Cleopatra Pike was funded by the Biotechnology and Biological Sciences Research Council (BBSRC) grant BB/N010108/1. Pralle Kriengwatana was funded by BBSRC Grant BB/L002264/1.

References

Ainsworth, W., 1972. Duration as a cue in the recognition of synthetic vowels. J. Acoust. Soc. Am. 51 (2B), 648–651.
Ainsworth, W., 1988. Speech Recognition by Machine. Billing and Sons Ltd.
Antunes, F.M., Malmierca, M.S., 2014. An overview of stimulus-specific adaptation in the auditory thalamus. Brain Topogr. 27, 480–499.
Assmann, P., Nearey, T., Hogan, J., 1982. Vowel identification: orthographic, perceptual, and acoustic aspects. J. Acoust. Soc. Am. 71 (4), 975–989.
Baugh, A.T., Akre, K.L., Ryan, M.J., 2008. Categorical perception of a natural, multivariate signal: mating call recognition in túngara frogs. Proc. Natl. Acad. Sci. U. S. A. 105 (26), 8985–8988.
Beeston, A., Brown, G., Watkins, A., 2014. Perceptual compensation for the effects of reverberation on consonant identification. J. Acoust. Soc. Am. 136 (6), 3072–3084.
Bennett, D.C., 1968. Spectral form and duration as cues in the recognition of English and German vowels. Lang. Speech 11 (2), 65–81.
Berwick, R., Okanoya, K., Beckers, G., Bolhuis, J., 2011. Songs to syntax: the linguistics of birdsong. Trends Cognit. Sci. 15 (3), 113–121.
Bolhuis, J.J., Moorman, S., 2015. Birdsong memory and the brain: in search of the template. Neurosci. Biobehav. Rev. 50, 41–55.
Bregman, A., 1990. Auditory Scene Analysis. MIT Press, Cambridge, Massachusetts.
Brenowitz, E.A., 1991. Altered perception of species-specific song by female birds after lesions of a forebrain nucleus. Science 251 (4991), 303–305.
Bruckert, L., Bestelmeyer, P., Latinus, M., Rouger, J., Charest, I., Rousselet, G., Kawahara, H., Belin, P., 2010. Vocal attractiveness increases by averaging. Curr. Biol. 20 (2), 116–120.
Cambria, E., White, B., 2014. Jumping NLP curves: a review of natural language processing research. Comput. Intell. Mag. 9, 48–57.
Carlson, R., Fant, G., Granström, B., 1975. Two formant models, pitch and vowel perception. In: Fant, G., Tatham, M. (Eds.), Auditory Analysis and Perception of Speech. Academic Press, London, pp. 55–82.
Catchpole, C.K., Slater, P.J.B., 2008. Bird Song: Biological Themes and Variations. Cambridge University Press, Cambridge.
Chew, S.J., Vicario, D.S., Nottebohm, F., 1996. A large-capacity memory system that recognizes the calls and songs of individual birds. Proc. Natl. Acad. Sci. U. S. A. 93, 1950–1955.
Chirathivat, N., Raja, S.C., Gobes, S.M.H., 2015. Hemispheric dominance underlying the neural substrate for learned vocalizations develops with experience. Sci. Rep. 5, 11359.
Clack, J.A., 2002. Patterns and processes in the early evolution of the tetrapod ear. J. Neurobiol. 53, 251–264.
Cohen, J., Kamm, T., Andreou, A., 1995. Vocal tract normalization in speech recognition: compensating for systematic speaker variability. J. Acoust. Soc. Am. 97, 3246.
Cowan, N., 1984. On short and long autitory stores. Psychol. Bull. 96, 341–370.
Cutting, J., Rosner, B., 1974. Categories and boundaries in speech and music. Percept. Psychophys. 16 (1974), 564.
Cynx, J., 1993. Conspecific song perception in zebra finches (*Taeniopygia guttata*). J. Comp. Psychol. 107 (4), 395–402.
Cynx, J., Nottebohm, F., 1992. Role of gender, season, and familiarity in discrimination of conspecific song by zebra finches (*Taeniopygia guttata*). Proc. Natl. Acad. Sci. U. S. A. 89, 1368–1371.
Cynx, J., Williams, H., Nottebohm, F., 1990. Timbre discrimination in zebra finch (*Taeniopygia guttata*) song syllables. J. Comp. Psychol. 104 (3), 303–308.
Delattre, P.C., Liberman, A.M., Cooper, F.S., 1955. Acoustic loci and transitional cues for consonants. J. Acoust. Soc. Am. 27 (4), 769–773.
Dienes, Z., Berry, D., 1997. Implicit learning: below the subjective threshold. Psych. Bull. Rev. 4 (1), 3–23.
Doupe, A.J., Kuhl, P.K., 1999. Birdsong and human speech: common themes and mechanisms. Annu. Rev. Neurosci. 22, 567–631.
DuBois, A., Nowicki, S., Searcy, W., 2009. Swamp sparrows modulate vocal performance in an aggressive context. Biol. Lett. 5 (2), 163–165.
Elgoyhen, A., Katz, E., 2012. The efferent medial olivocochlear-hair cell synapse. J. Physiol. 106 (1–2), 47–56.
Elie, J.E., Theunissen, F.E., 2016. The vocal repertoire of the domesticated zebra finch: a data-driven approach to decipher the information-bearing acoustic features of communication signals. Anim. Cogn. 19, 285–315.
Ellis, 1994. Implicit and Explicit Learning of Languages. Academic Press Inc.
Fant, G., 2001. T Chiba and M Kajyama: pinoneers in speech acoustics. Speech Music Hear. Q. Prog. Status Rep. (TMH-QPSR) 42 (1), 59–60.
Fitch, W., 1997. Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. J. Acoust. Soc. Am. 102 (2), 1213–1222.
Fitch, W., Hauser, M., 2001. Unpacking "honesty": vertebrate vocal production and the evolution of acoustic signals. In: Simmons, A., Fay, R.R., Popper, A.N. (Eds.), Acoustic Communication. Springer, New York, pp. 275–323.
Fitch, W., Hauser, M., 2002. Unpacking "honesty": vertebrate vocal production and the evolution of acoustic signals. In: Simmons, A.M., Fay, R.R., Popper, A.N. (Eds.), Acoustic Communication. Springer, New York, pp. 65–137.
Fitch, W.T., Kelley, J.P., 2000. Perception of vocal tract resonances by whooping cranes *Grus americana*. Ethology 106, 559–574.
Flege, J., Yeni-Komshian, G., Liu, S., 1999. Age constraints on second-languageacquisition. J. Mem. Lang. 41, 78–104.
Foster, D.H., 2011. Color constancy. Vis. Res. 51, 674–700.
Fritzsch, B., Pan, N., Jahan, I., Duncan, J.S., Kopecky, B.J., Elliott, K.L., Kersigo, J., Yang, T., 2013. Evolution and development of the tetrapod auditory system: an organ of corti-centric perspective. Evol. Dev. 15, 63–79.
Fujisaki, H., Kawashima, T., 1970. Some Experiments in Speech Perception and a Model for the Perceptual Mechanisms, vol. 29. Annual Report of the Engineering Research Institute, Faculty of Engineering, University of Tokyo, pp. 207–214.
Galantucci, B., Fowler, C., Turvey, M., 2016. The motor theory of speech reviewed. Psych. Bull. Rev. 13 (3), 361–377.
Gerstman, L., 1968. Classification of self-normalized vowels. IEEE Trans. Audio Electroacoust. AU-16, 78–80.
Gobes, S.M.H., Bolhuis, J.J., 2007. Birdsong memory: a neural dissociation between song recognition and production. Curr. Biol. 17, 789–793.
Halberstam, B., Raphael, L., 2004. Vowel normalization: the role of fundamental frequency and upper formants. J. Phon. 32, 423–434.
Hall, M.L., Kingma, S.A., Peters, A., 2013. Male songbird indicates body size with low-pitched advertising songs. PLoS One 8 (2), e56717.
Hardouin, L.A., Thompson, R., Stenning, M., Reby, D., 2014. Anatomical bases of sex- and size-related acoustic variation in herring gull alarm calls. J. Avian Biol. 45 (2), 157–166.
Helmholtz, H., 1863. On the Sensations of Tone. Dover Books.
Helson, H., 1948. Adaptation-level as a basis for a quantitative theory of frames of reference. Psychol. Rev. 55 (6), 297–313.
Hinds, D.S., Calder, W.A., 1971. Tracheal dead space in the respiration of birds. Soc. Study Evol. 25 (2), 429–440.
Holt, L.L., 1999. Auditory Constraints on Speech Perception: An Examination of Spectral Contrast', Doctoral Dissertation. University of Wisconsin, Madison.
Holt, L.L., 2005. Temporally non-adjacent non-linguistic sounds affect speech characterization. Psychol. Sci. 16, 305–312.
Holt, L.L., 2006. The mean matters: effects of statistically defined nonspeech spectral distributions on speech categorisation. J. Acoust. Soc. Am. 120, 2801–2817.
Holt, L.L., Lotto, A.J., 2002. Behaviorial examinations of the neural mechanisms of speech context effects. Hear. Res. 167, 156–169.
Holt, L.L., Lotto, A.J., 2010. Speech perception as categorization. Atten. Percept. Psychophys. 72, 1218–1227.
House, A., Fairbanks, G., 1953. The influence of consonant environment upon the secondary acoustical characteristics of vowels. J. Acoust. Soc. Am. 25 (1), 105–113.
Huang, J., Holt, L.L., 2009. General perceptual contributions to lexical tone

normalization. J. Acoust. Soc. Am. 125 (6), 3983–3994.
Hubel, D., Wiesel, T., 1959. Receptive fields of single neurones in the cat's striatecortex. J. Physiol. 148, 574–591.
Huber, S.K., Podos, J., 2006. Beak morphology and song features covary in a population of Darwin's finches (Geospiza fortis). Biol. J. Linn. Soc. 88, 489–498.
Johnson, K.A., 2005. Speaker normalization in speech perception. In: Pisoni, D.B., Remez, R.E. (Eds.), The Handbook of Speech Perception. Blackwell, Oxford, pp. 363–389.
Joos, M., 1948. Acoustic phonetics: supplement to language. J. Linguist. Soc. Am. 24 (2) Language monograph 23.
Kiang, N., Watanabe, T., Thomas, E., Clark, L., 1965. Response patterns of single fibers in the cat's auditory nerve. MIT Res. Monogr. 35.
Kiefte, M., Kluender, K., 2005. The relative importance of spectral tilt in monophthongs and diphthongs. J. Acoust. Soc. Am. 117, 1395–1404.
Kluender, K.R., Lotto, A.J., Holt, L.L., Bloedel, S.L., 1998. Role of experience for language-specific functional mappings of vowel sounds. J. Acoust. Soc. Am. 104 (6), 3568–3582.
Koffka, K., 1935. Principles of Gestalt Psychology. Routledge.
Köppl, C., 1997. Phase locking to high frequencies in the auditory nerve and cochlear nucleus magnocellularis of the barn owl, Tyto alba. J. Neurosci. 17 (9), 3312–3321.
Kriengwatana, B., Escudero, P., Kerkhoven, A.H., ten Cate, C., 2015. A general auditory bias for handling speaker variability in speech? Evidence in humans and songbirds. Front. Psychol. 6, 1234.
Kroodsma, D.E., Liu, W.-C., Goodwin, E., Bedell, P.A., 1999. The ecology of song improvisation as illustrated by North American sedge wrens. Auk 116 (2), 373–386.
Kuhl, P., 1991. Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories, monkeys do not. Percpetion Psychophys. 50, 93–107.
Kuhl, P., Iverson, P., 1995. Linguistic experience and the perceptual magnet effect. In: Strange, W. (Ed.), Speech Perception and Linguistic Experience: Issues in Cross Language Research. York Press, Baltimore.
Kuhl, P.K., Miller, J.D., 1975. Speech perception by the Chinchilla: voiced-voiceless distinction in alveolar plosive consonants. Science 190, 69–72.
Lachlan, R.F., Nowicki, S., 2015. Context-dependent categorical perception in a songbird. Proc. Natl. Acad. Sci. U. S. A. 112, 1892–1897.
Lachlan, R.F., Anderson, R., Peters, S., Searcy, W., Nowicki, S., 2014. Typical versions of learned swamp sparrow song types are more effective signals than are less typical versions. Proc. R. Soc. B: Biol. Sci. 281 (1785).
Lachlan, R.F., van Heijningen, C.A.A., ter Haar, S.M., ten Cate, C., 2016. Zebra finch song phonology and syntactical structure across populations and continents—a computational comparison. Front. Psychol. 7, 980.
Ladefoged, P., Broadbent, D., 1957. Information conveyed by vowels. J. Acoust. Soc. Am. 29 (1), 98–104.
Laing, E., Liu, R., Lotto, A., Holt, L., 2012. Tuned with a tune: talker normalisation via general auditory processes. Front. Psychol. 3, 1–9.
Lehiste, I., Meltzer, D., 1973. Vowel and speaker identification in natural synthetic speech. Lang. Speech 16, 356–364.
Liberman, P., 1973. On the evolution of language. Cognition 2, 59–94.
Liberman, A., Harris, K., Hoffman, H., Griffith, B., 1957. The discrimination of speech sounds within and across phoneme boundaries. J. Exp. Psychol. 54, 358–368.
Liberman, A., Cooper, F., Shankweiler, D., Studdert-Kennedy, M., 1967. Perception of the speech code. Psychol. Rev. 74, 431–461.
Lindblom, B., 1963. Spectrographic study of vowel reduction. J. Acoust. Soc. Am. 35 (11), 1773–1781.
Lindblom, B., Studdert-Kennedy, M., 1967. On the role of formant transitions in vowel recognition. J. Acoust. Soc. Am. 42 (4), 830–843.
Linhart, P., Fuchs, R., 2015. Song pitch indicates body size and correlates with males' response to playback in a songbird. Anim. Behav. 103, 91–98.
Lippman, R., 1996. Accurate consonant perception without mid-frequency speech energy. IEEE Trans. Speech Audio Process. 4, 66–69.
Lloyd, R., 1890a. Some Research into the Nature of Vowel-Sound. Turner and Dunnett, Liverpool, England.
Lloyd, R., 1890b. Speech sounds their nature and causation. Phonetische Studien 3, 251–278.
Lohr, B., 2008. Pitch-related cues in the songs of sympatric mountain and black-capped chickadees. Behav. Process. 77, 156–165.
Lohr, B., Dooling, R.J., 1998. Detection of changes in timbre and harmonicity in complex sounds by zebra finches (Taeniopygia guttata) and budgerigars (Melopsittacus undulatus). J. Comp. Psychol. 112 (1), 36–47.
Lotto, A.J., Kluender, K.R., Holt, L.L., 1997. Perceptual compensation for coarticulation by Japanese quail (*Coturnix coturnix japonica*). J. Acoust. Soc. Am. 102 (2), 1134–1140.
Macchi, M., 1980. Identification of vowels spoken in isolation versus vowels spoken in consonantal context. J. Acoust. Soc. Am. 68 (6), 1636–1642.
Maddox, W., Molis, M., Diehl, R., 2002. Generalizing a neuropsychological model of visual categorization to auditory categorization of vowels. Percept. Psychophys. 64, 584–597.
Mann, V., 1980. Influence of preceding liquid on stop-consonant perception. Percept. Psychophys. 28, 407–412.
Margoliash, D., 1983. Acoustic parameters underlying the responses of song-specific neurons in the white-crowned sparrow. J. Neurosci. 3 (5), 1039–1057.
Marler, P., Tamura, M., 1964. Culturally transmitted patterns of vocal behavior in sparrows. Science 146 (3650), 1483–1486.
May, B., Moody, D.B., Stebbins, W.C., 1989. Categorical perception of conspecific communication sounds by Japanese macaques, *Macaca fuscata*. J. Acoust. Soc. Am. 85, 837.
McCasland, J.S., Konishi, M., 1981. Interaction between auditory and motor activities in an avian song control nucleus. Proc. Natl. Acad. Sci. U. S. A. 78 (12), 7815–7819.
McClelland, J.L., Elman, J.L., 1986. The TRACE model of speech perception. Cogn. Psychol. 18 (1), 1–86.
McGregor, P.K., Krebs, J.R., 1984. Sound degradation as a distance cue in great tit (Parus major) song. Behav. Ecol. Sociobiol. 16, 49–56.
Mello, C.V., Vicario, D.S., Clayton, D.F., 1992. Song presentation induces gene expression in the songbird forebrain. Proc. Natl. Acad. Sci. U. S. A. 89 (15), 6818–6822.
Miller, R., 1953. Auditory tests with synthetic vowels. The Journal of the AcousticalSociety of America 25 (1), 114–121.
Miller, J.D., 1989. Auditory-perceptual interpretation of the vowel. J. Acoust. Soc. Am. 85 (5), 2114–2134.
Miller, J., Wier, C., Pastore, R., Kelly, W., Dooling, R., 1976. Discrimination and labeling of noise–buzz sequences with varying noise-lead times: an example of categorical perception. JASA 60 (2), 410–417.
Morton, E.S., 1986. Predictions from the ranging hypothesis for the evolution of long distance signals in birds. Behaviour 99 (1/2), 65–86.
Mullennix, J.W., Pisoni, D.B., Martin, C.S., 1989. Some effects of talker variability on spoken word recognition. J. Acoust. Soc. Am. 85 (1), 365–378.
Nearey, T.M., 1989. Static, dynamic, and relational properties in vowel perception. J. Acoust. Soc. Am. 85 (5), 2088–2113.
Nearey, T., Assmann, P., 1986. Modeling the role of inherent spectral change in vowel identification. J. Acoust. Soc. Am. 80 (5), 1297–1308.
Nelson, D.A., Marler, P., 1989. Categorical perception of a natural stimulus continuum: birdsong. Science 244, 976–978.
Nordström, P.E., Lindblom, B., 1975. A normalization procedure for vowel formant data. Proceedings of the 8th International Congress of Phonetic Sciences.
Nottebohm, F., Alvarez-Buylla, A., Cynx, J., Kirn, J., Ling, C.Y., Nottebohm, M., Williams, H., 1990. Song learning in birds: the relation between perception and production. Philos. Trans. R. Soc. Lond. Ser. B: Biol. Sci. 329 (1253), 115–124.
Nowicki, S., 1987. Vocal tract resonances in oscine bird sound production: evidence from birdsongs in a helium atmosphere. Nature 325, 53–55.
Nowicki, S., Marler, P., 1988. How do birds sing? Music Percept.: Interdiscip. J. 5 (4), 391–426.
Nowicki, S., Searcy, W.A., 2005. Song and mate choice in birds: how the development of behavior helps us understand function. Auk 122 (1), 1–14.
Nusbaum, H., Morin, T., 1992. Paying attention to differences among talkers. In: Tohkura, Y., Vatikiotis-Bateson, E., Sagisaka, Y. (Eds.), Speech perception, production and linguistic structure. IOS Press, pp. 113–123.
Nygaard, L., Pisoni, D., 1998. Talker-specific learning in speech perception. Percept. Psychophys. 60, 355–376.
Ohl, F., Scheich, H., 1997. Orderly cortical representations of vowels based on formant interaction. Proc. Natl. Acad. Sci. U. S. A. 94, 9440–9444.
Ohms, V.R., Snelderwaard, P.C., ten Cate, C., Beckers, G.J.L., 2010. Vocal tract articulation in Zebra finches. PLoS One 5 (7) e11923.
Palmer, A., Summerfield, Q., Fantini, D.A., 1995. Responses of auditory nerve fibers to stimui producing psychophysical enhancement. J. Acoust. Soc. Am. 97, 1786–1799.
Peterson, G.E., Barney, H.L., 1952. Control methods used in a study of the vowels. J. Acoust. Soc. Am. 24 (2), 175–184.
Pike, C., 2015. Timbral Constancy and Compensation for Spectral Distortion Caused by Loudspeaker and Room Acoustics. Doctoral Thesis. University of Surrey.
Pike, C., Brookes, T., Mason, R., 2014. The effect of auditory memory on the perception of timbre. Audio Engineering Society 136th Convention Preprint 9028.
Pisoni, D., 1973. Auditory and phonetic memory codes in the discrimination of consonants and vowels. Percept. Psychophys. 13, 253–260.
Pisoni, D., 1997. Some thoughts on "normalisation" in speech perception. In: Johnson, K., Mullennix, J. (Eds.), Talker Variability in Speech Processing. Academic Press, California, pp. 9–32.
Podos, J., 1996. Motor constraints on vocal development in a songbird. Anim. Behav. 51, 1061–1070.
Podos, J., 2001. Correlated evolution of morphology and vocal signal structure in Darwin's finches. Nature 409, 185–188.
Podos, J., Lahti, D.C., Moseley, D.L., 2009. Vocal performance and sensorimotor learning in songbirds. Adv. Study Behav. 40, 159–193.
Potter, R., Steinberg, J., 1950. Toward the specification of speech. J. Acoust. Soc. Am. 22 (6), 807–820.
Prather, J.F., Nowicki, S., Anderson, R.C., Peters, S., Mooney, R., 2009. Neural correlates of categorical perception in learned vocal communication. Nat. Neurosci. 12 (2), 221–228.
Prince, B., Riede, T., Goller, F., 2011. Sexual dimorphism and bilateral asymmetry of syrinx and vocal tract in the European starling (Sturnus vulgaris). J. Morphol. 272, 1527–1536.
Pytte, C.L., Suthers, R.A., 1999. A bird's own song contributes to conspecific song perception. Neuroreport 10 (8), 1773–1778.
Reby, D., McComb, K., 2003. Anatomical constraints generate honesty: acoustic cues to age and weight in the roars of red deer stags. Anim. Behav. 65, 519–530.
Remez, R., Rubin, P., Personi, D., Carrell, T., 1981. Perception without traditional speech cues. Science 212, 947–950.
Repp, B.H., 1982. Phonetic trading relations and context effects: new experimental evidence for a speech mode of perception. Psychol. Bull. 92 (1), 81–110.
Ribeiro, S., Cecchi, Ga, Magnasco, M.O., Mello, C.V., 1998. Toward a song code: evidence for a syllabic representation in the canary brain. Neuron 21, 359–371.
Riede, T., Fitch, T., 1999. Vocal tract length and acoustics of vocalization in the domestic dog (*Canis familiaris*). J. Exp. Biol. 202, 2859–2867.
Riede, T., Goller, F., 2014. Morphological basis for the evolution of acoustic diversity in oscine songbirds. Proc. R. Soc. B 281, 20132306.
Riede, T., Suthers, R.A., Fletcher, N.H., Blevins, W.E., 2006. Songbirds tune their vocal

tract to the fundamental frequency of their song. Proc. Natl. Acad. Sci. U. S. A. 103 (14), 5543–5548.
Riede, T., Fisher, J.H., Goller, F., 2010. Sexual dimorphism of the zebra finch syrinx indicates adaptation for high fundamental frequencies in males. PLoS One 5 (6) e11368.
Saffran, J.R., Aslin, R.N., Newport, E.L., 1996. Statistical learning by 8-month-old infants. Science 274, 1926–1928.
Saino, N., Galeotti, P., Sacchi, R., Moeller, A.P., 1997. Song and immunological condition in male barn swallows (Hirundo rustica). Behav. Ecol. 8 (4), 364–371.
Samuel, A., 1982. Phonetic prototypes. Percept. Psychophys. 31, 307–314.
Samuels, B.D., 2015. Can a bird brain do phonology? Front. Psychol. 6 (July), 1–10.
Searcy, W.A., Andersson, M., 1986. Sexual selection and the evolution of song. Annu. Rev. Ecol. Syst. 17, 507–533.
Searcy, W., Anderson, R., Nowicki, S., 2006. Bird song as a signal of aggressive intent. Behav. Ecol. Sociobiol. 60 (2), 234–241.
Shapley, R., Tolhurst, D., 1973. Edge detectors in human vision. J. Physiol. 229, 165–183.
Smith, R., 1979. Adaptation, saturation, and physiological masking in single auditory-nerve fibers. J. Acoust. Soc. Am. 65, 166–178.
Soha, J., 2016. The auditory template hypothesis: a review and comparative perspective. Anim. Behav. 124, 247–254.
Stilp, C., Alexander, J., Kiefte, M., Kluender, K., 2010. Auditory color constancy: calibration to reliable spectral properties across speech and non speech contexts and targets. Atten. Percept. Psychophys. 72, 470–480.
Strange, W., 1989. Evolving theories of vowel perception. J. Acoust. Soc. Am. 85 (5), 2081–2087.
Strange, W., Verbrugge, R.R., Shankweiler, D.P., Edman, T.R., 1976. Consonant environment specifies vowel identity. J. Acoust. Soc. Am. 60 (1), 213–224.
Sturdy, C.B., Phillmore, L.S., Price, J.L., Weisman, R.G., 1999. Song-note discriminations in Zebra finches (*Taeniopygia guttata*): categories and pseudocategories. J. Comp. Psychol. 113 (2), 204–212.
Sturdy, C.B., Phillmore, L.S., Weisman, R.G., 2000. Call-note discriminations in black-capped chickadees (Poecile atricapillus). J. Comp. Psychol. 114 (4), 357–364.
Summerfield, Q., Assmann, P., 1989. Auditory enhancement and the perception of concurrent vowels. Atten. Percept. Psychophys. 45, 529–536.
Summerfield, Q., Haggard, M., Foster, J., Gray, S., 1984. Perceiving vowels from uniform spectra: phonetic exploration of an auditory aftereffect. Atten. Percept. Psychophys. 35, 203–213.
Summerfield, Q., Sidwell, A., Nelson, T., 1987. Auditory enhancement of changes in spectral amplitude. J. Acoust. Soc. Am. 81 (3), 700–708.
Sussman, H., 1989. Neural coding of relational invariance in speech: Human language analogs to the barn owl. Psychol. Rev. 96 (4), 631–642.
Sussman, H., Fruchter, D., Hilbert, J., Sirosh, J., 1997. Linear correlates in the speech signal: the orderly output constraint. Brain Behav. Sci. 21, 260–299.
Suthers, R.A., 1994. Variable asymmetry and resonance in the avian vocal tract: a structural basis for individually distinct vocalizations. J. Comp. Physiol. A 175, 457–466.
Suthers, R., 2001. Peripheral vocal mechanisms in birds: Are songbirds special? Neth. J. Zool. 51 (2), 217–242.
Templeton, C.N., Greene, E., Davis, K., 2005. Allometry of alarm calls: black-capped chickadees encode information about predator size. Science 308, 1934–1937.
Theunissen, F.E., Amin, N., Shaevitz, S.S., Woolley, S.M., Fremouw, T., Hauber, M.E., 2004. Song selectivity in the song system and in the auditory forebrain. Ann. N. Y. Acad. Sci. 1016, 222–245.
Traunmüller, H., 1984. Articulation and perceptual factors controlling the age and sex conditioned variability in formant frequencies of vowels. Speech Commun. 3, 49–61.
Ulanovsky, N., Las, L., Nelken, I., 2003. Processing of low-probability sounds by cortical neurons. Nat. Neurosci. 6, 391–398.
Ulanovsky, N., Las, L., Farkas, D., Nelken, I., 2004. Multiple time scales of adaptation in auditory cortex neurons. J. Neurosci. 24, 10440–10453.
Uno, H., Maekawa, M., Kaneko, H., 1997. Strategies for harmonic structure discrimination by zebra finches. Behav. Brain Res. 89, 225–228.
Vallet, E., Kreutzer, M., 1995. Female canaries are sexually responsive to special song phrases. Anim. Behav. 49 (6), 1603–1610.
Vallet, E., Kreutzer, M., 1998. Two-note syllables in canary songs elicit high levels of sexual display. Anim. Behav. 55, 291–297.
Verbrugge, R.R., Shankweiler, D.P., Strange, W., 1976. Shifts in vowel perception as a function of speaking rate. J. Acoust. Soc. Am. 59 (S1) S5–S5.
Viemeister, N., Bacon, S., 1982. Forward masking by enhanced components in harmonic complexes. J. Acoust. Soc. Am. 71, 1502–1507.
Vignal, C., Mathevon, N., Mottin, S., 2008. Mate recognition by female zebra finch: analysis of individuality in male call and first investigations on female decoding process. Behav. Process. 77 (2), 191–198.
Warren, T., Tumer, E., Charlesworth, J., Brainard, M., 2011. Mechanisms and time course of vocal learning and consolidation in the adult songbird. J. Neurophysiol. 106 (4), 1806–1821.
Watkins, A.J., 1991. Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion. J. Acoust. Soc. Am. 90 (6), 2942–2955.
Watkins, A.J., Makin, S.J., 1994. Perceptual compensation for speaker differences and for spectral-envelope distortion. J. Acoust. Soc. Am. 96 (3), 1263–1282.
Watkins, A.J., Makin, S.J., 1996a. Effects of spectral contrast on perceptual compensation for spectral-envelope distortion. J. Acoust. Soc. Am. 99 (6), 3749–3757.
Watkins, A.J., Makin, S.J., 1996b. Some effects of filtered contexts on the perception of vowels and fricatives. J. Acoust. Soc. Am. 99 (1), 588–594.
Weary, D.M., 1989. Categorical perception of bird song: how do great tits (Parus major) perceive temporal variation in their song? J. Comp. Psychol. 103 (4), 320–325.
Weary, D.M., Krebs, J.R., 1992. Great tits classify songs by individual voice characteristics. Anim. Behav. 43, 283–287.
Weisman, R., Njegovan, M., Ito, S., 1994. Frequency ratio discrimination by zebra finches (*Taeniopygia guttata*) and humans (Homo sapiens). J. Comp. Psychol. 108 (4), 363–372.
Werker, J., Tees, R., 1984. Cross-language speech perception: evidence for perceptual reorganization during the first year of life. Infant Behav. Dev. 7, 49–63.
White, S.A., Fisher, S.E., Geschwind, D.H., Scharff, C., Holy, T.E., 2006. Singing mice, songbirds, and more: models for FOXP2 Function and dysfunction in human speech and language. J. Neurosci. 26 (41), 10376–10379.
Williams, H., 1989. Multiple representations and auditory-motor interactions in the avian song system. Ann. N. Y. Acad. Sci. 563, 148–164.
Williams, H., Cynx, J., Nottebohm, F., 1989. Timbre control in zebra finch (*Taeniopygia guttata*) song syllables. J. Comp. Psychol. 103 (4), 366–380.
Wohlgemuth, M.J., Sober, S.J., Brainard, M.S., 2010. Linked control of syllable sequence and phonology in birdsong. J. Neurosci. 30 (39), 12936–12949.
Wong, P., Nusbaum, H., Small, S., 2004. Neural bases of talker normalization. J. Cogn. Neurosci. 16, 1–13.
Wyttenbach, R.A., May, M.L., Hoy, R.R., 1996. Categorical perception of sound frequency by crickets. Science 273 (5281), 1542–1544.
Yip, M.J., 2006. The search for phonology in other species. Trends Cognit. Sci. 10 (10).