DOCTOR OF PHILOSOPHY

The perception of emotion and identity in non-speech vocalisations

Pye, Annie

Award date:
2015

Awarding institution:
Bangor University

The perception of emotion and identity in non-speech vocalisations

Annie Pye

This thesis is submitted in partial fulfilment for the degree of Doctor of Philosophy

September 2015

School of Psychology, Bangor University

Declaration and Consent

Details of the Work

I hereby agree to deposit the following item in the digital repository maintained by Bangor University and/or in any other repository authorized for use by Bangor University.

Author Name:

……..

Title:

……..…… …….

Supervisor/Department:

.....

Funding body (if any):

.....

Qualification/Degree obtained:

…….

This item is a product of my own research endeavours and is covered by the agreement below in which the item is referred to as "the Work". It is identical in content to that deposited in the Library, subject to point 4 below.

Non-exclusive Rights

Rights granted to the digital repository through this agreement are entirely non-exclusive. I am free to publish the Work in its present version or future versions elsewhere.

I agree that Bangor University may electronically store, copy or translate the Work to any approved medium or format for the purpose of future preservation and accessibility. Bangor University is not under any obligation to reproduce or display the Work in the same formats or resolutions in which it was originally deposited.

Bangor University Digital Repository

I understand that work deposited in the digital repository will be accessible to a wide variety of people and institutions, including automated agents and search engines via the World Wide Web.

I understand that once the Work is deposited, the item and its metadata may be incorporated into public access catalogues or services, national databases of electronic theses and dissertations such as the British Library's EThOS or any service provided by the National Library of Wales.

I understand that the Work may be made available via the National Library of Wales Online Electronic Theses Service under the declared terms and conditions of use (http://www.llgc.org.uk/index.php?id=4676). I agree that as part of this service the National Library of Wales may electronically store, copy or convert the Work to any approved medium or format for the purpose of future preservation and accessibility. The National

Library of Wales is not under any obligation to reproduce or display the Work in the same formats or resolutions in which it was originally deposited.

Statement 1:

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree unless as agreed by the University for approved dual awards.

Signed …….. (candidate)

Date ……..

Statement 2:

This thesis is the result of my own investigations, except where otherwise stated. Where correction services have been used, the extent and nature of the correction is clearly marked in a footnote(s).

All other sources are acknowledged by footnotes and/or a bibliography.

Signed ……. (candidate)

Date …….

Statement 3:

I hereby give consent for my thesis, if accepted, to be available for photocopying, for inter-library loan and for electronic repositories, and for the title and summary to be made available to outside organisations.

Signed ……. (candidate)

Date …….

Acknowledgements

Thanks to everyone who made this possible and helped me along the way. You know who you are.

Contents

Abstract

The voice contains a wealth of information relevant for successful and meaningful social interactions. Aside from speech, the vocal signal also contains paralinguistic information such as the emotional state and identity of the speaker. The three empirical chapters reported in this thesis research the perceptual processing of paralinguistic vocal cues. The first set of studies uses unimodal adaptation to explore the mental representation of emotion in the voice. Using a series of different adaptor stimuli -human emotional vocalisations, emotive dog calls and affective instrumental bursts- it was found that aftereffects in human vocal emotion perception were largest following adaptation to human vocalisations. There was still an aftereffect present following adaptation to dog calls, however it was smaller in magnitude than that of the human vocalisation aftereffect and potentially as a result of the acoustic similarities between adaptor and test stimuli. Taken together, these studies suggest that the mental representation of emotion in the voice is not species specific but is specific to vocalisations as opposed to all affective auditory stimuli. The second empirical chapter examines the supramodal relationship between identity and emotion in face-voice adaptation. It was found that emotional faces have the ability to produce aftereffects in vocal emotion perception, irrespective of the identity of the adaptor and test stimuli being congruent. However, this effect was found to be dependent upon the adapting stimuli being dynamic as opposed to static in nature. The final experimental chapter looks at the mechanisms underlying the perception of vocal identity. A voice matching test was developed and standardised, finding large individual differences in voice matching ability. Furthermore, in an identity adaptation experiment, absolute difference in aftereffect size demonstrated a trend towards significance when correlated with voice matching ability, suggesting that there is a relationship between perceptual abilities and the degree of plasticity observed in response adaptation.

Chapter One: An introduction to voice perception

1.0 The Vocal Apparatus

Unlike other primates, humans have developed the ability to communicate through complex speech sounds. In order to understand the ways in which such sounds are perceived, it is necessary to review the ways in which these sounds are produced. During respiration, air passes out of the lungs and up to the larynx, the first of several structures implicated in speech production (Hardcastle, 1976). The larynx consists of several cartilages and connecting joints that serve to modify air-passage in order to generate an acoustic signal. The thyroid and cricoid cartilages are joined by the cricothyroid joint, the movement of which has important implications for the frequency of vibration of the vocal cords. In addition to this, there is a pair of arytenoid cartilages which are directly attached to the vocal cords, meaning that any movement of these cartilages has a direct effect on the tension of the vocal cords (Pittman, 1994). Shimmer is a measurable quality of the voice, that is determined by the temporal changes in the vibration of the vocal folds and it is reflective of vocal roughness or harshness (Farrús, Hernando & Ejarque, 2007). Similarly, another parameter reflective of roughness or harshness is jitter, which arises from the variability in the duration of vibration of the vocal folds. Another of the cartilages of the larynx is the epiglottis, the primary purpose of which is to prevent food from entering the airways when swallowing. In relation to the voice, the epiglottis has been shown to be implicated in the production of pharyngeal consonants, the vowel /a/ and whispered voice (Laufer, 1981).

The vocal cords stretch across the larynx and are attached to its cartilage. When a sound is voiced, the vocal cords vibrate as a result of air passage, muscular and elastic movement through the larynx (Titze, 2008). The vocal cords open and close during vocalisation, converting the air passage into an acoustic waveform. The tension of the vocal cords is related to the pitch of the voice, with increased tension reflected in higher pitched voicing. However, the pitch of the voice is also affected by the glottis, the opening through

which air passes at variable rates in both respiration and articulation. The more air that passes through the glottis in a given moment, the higher the pitch of the note articulated (Titze, 1989). The size of an individual's glottis determines to some degree their voice type with large glottis's being associated with deeper voices than smaller glottis's which tend to sound more shrill. The acoustic measure of pitch is fundamental frequency and is calculated in hertz.

The larynx opens up into the pharynx, the area at the back of the throat which stretches from the larynx to the base of the tongue (Hardcastle, 1976). The oropharynx is situated at the opening of the oral cavity and the nasopharynx at the opening of the nasal cavity. The velum acts to separate the nasal and oral cavities. The musculature of the pharynx causes changes in the shape, volume and length of the pharyngeal cavity which directly impacts upon voice quality.

Once air has passed through the glottis, it is modified in the vocal tract by means of a complex filtration process (Winsel, 1984). Bandpass filters allow certain frequencies to pass unchanged but restrict frequencies that are outside of a given range. These filters are known as formants, and are resonances at certain frequencies, the lowest of which is referred to as the first formant (F1), followed by the next highest resonant known as F2 and so on (Moore, 2003). These formant frequencies are independent of the pitch of the voice and change as a result of the shape of the vocal tract.

The source/ filter theory of speech production posits that there are two dissociable aspects of the speech signal: the source relates to the fundamental frequency or pitch of the voice, which is determined by the rate of oscillation of the vocal cords (Fant, 1960). The filter aspects of the speech signal are created through the filtering of the acoustic signal as it passes through the vocal tract by means of bandpass filters.

The wide range of movements that the tongue is able to achieve enables the diverse number of sounds used in complex speech (Hardcastle, 1976). For example, palatal sounds are created when the tongue contacts the palate of the mouth and guttural sounds are made from the back of the mouth. Another acoustic property of vocalisations is their high degree of harmonicity relative to other auditory sounds. A harmonic is an aspect of the auditory signal, the frequency of which is a numeric multiple of the fundamental frequency. Figure 1.0 shows the spectrographic representation of both a low pitched and a high pitched voice saying three different non-words. These visual representations of sound waves, demonstrates the spectrum of frequencies in the sound over time (Moore, 2003). The horizontal bands depict the harmonics present in the sound wave. Harmonics-to-noise ratio is a measure of the degree of hoarseness in the voice and is derived from the degree of noise present in the harmonic structure: the more noise present, the hoarser the voice sounds (Yumoto & Gould, 1982). Given these acoustic differences in the properties of the vocal signal as well as the social and environmental significance of the voice, researchers have begun to ask whether the mechanisms involved in the perception of the voice are specialised.

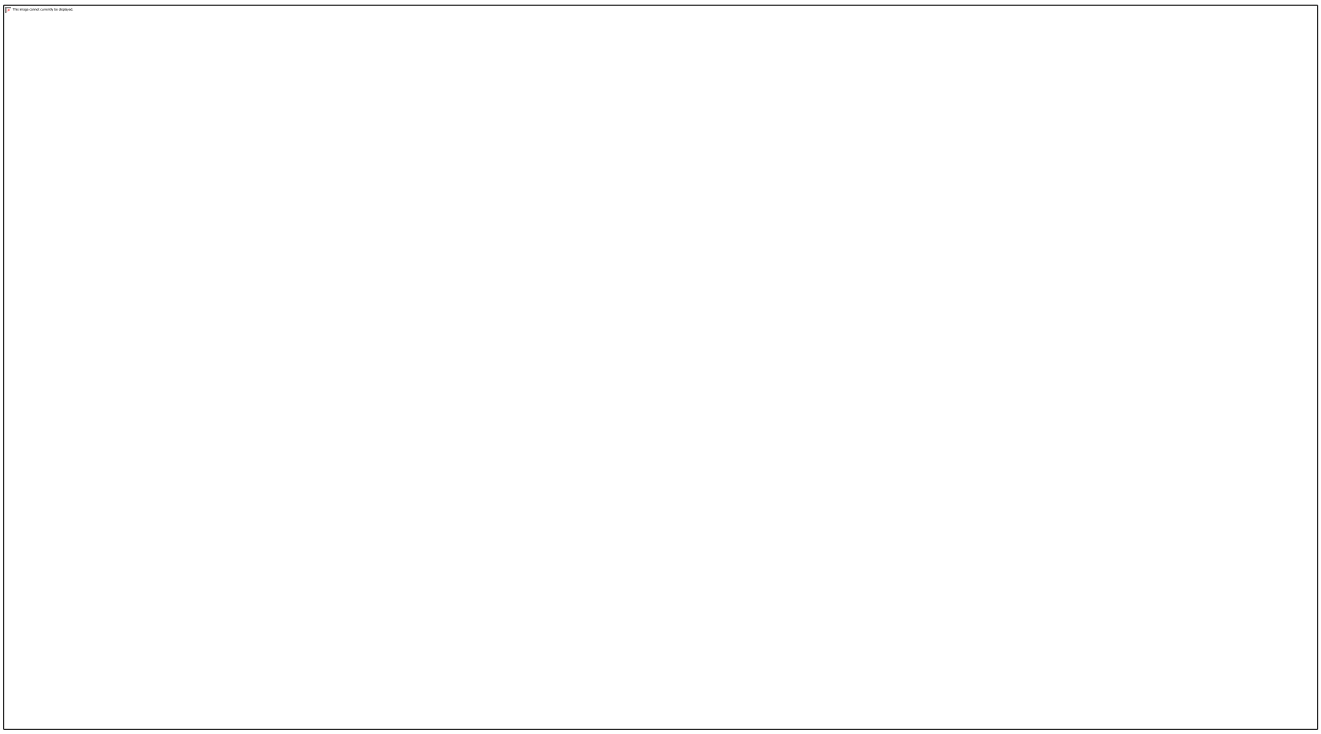


Figure 1.0 Spectrograms showing both low and high pitched voices saying three non-words.

1.1 The significance of the voice as an auditory stimulus

Previous research has demonstrated that even before birth, the foetus responds to the voice of the mother, demonstrated by an increase in heart rate, to a larger extent than they do to the voice of a stranger (Kisilevsky et al. 2003). Supporting this finding, it was found that auditory cortex was larger bilaterally for premature babies who had been exposed to maternal noise, including recordings of the mother's voice, relative to premature infants exposed to the natural auditory environment of the hospital (Webb, Heller, Benson & Lahav, 2015). These results demonstrate the significance of the voice in the childhood environment. From a young age, babies are able to recognise the voices of individuals who are familiar to them irrespective of their inability to comprehend or use speech.

Findings from neuroimaging suggest that there are indeed areas of the auditory cortex, located superiorly in the superior temporal sulcus, that are selectively sensitive to the processing of paralinguistic vocal features comparative to other auditory objects and environmental noise (Belin, Zatorre, Lafaille, Ahad & Pike. 2000; Lewis et al., 2009).

Furthermore, areas of the prefrontal cortex have shown greater activation in response to the voice than when compared with nonhuman sounds (Fecteau, Armony, Joanette & Belin, 2004). Supporting this, a voice specific event-related potential (ERP) was demonstrated approximately 320ms post stimulus onset which was absent in response to music stimuli. The authors concluded that this was reflective of increased attention to the voice due to it having greater salience than other auditory stimuli (Levy, Granot & Bentan, 2001; 2003). Similarly, other ERP results have demonstrated differences in amplitude between voice and non-voice stimuli as early as 164ms post stimulus (Charest et al, 2009). Research employing repetitive transcranial magnetic stimulation has demonstrated impaired ability to discriminate between voice and non-voice sounds when applied to the right temporal voice area (Bestelmeyer, Belin & Grosbras, 2011). These findings suggest the existence of mechanisms that are more responsive to the human voice than to other auditory sounds.

1.2 A model of voice perception

The voice provides us with a wealth of social information regarding, amongst other things, the gender, identity and affective state of the speaker even in the absence of speech. Primate studies allow us to make inferences about the function of vocalisations that do not incorporate speech as well as providing a degree of insight into the evolution of speech (Fitch, 2000). Prior to the evolution of speech, our ancestors were endowed with the ability to extract cues from vocalisations that provided valuable survival information, for example a cry as a means of indicating the presence of danger and alerting other individuals to it. The fact that such abilities were present before the use of speech suggests that the analysis of speech content may occur in a separate stream to that of the analysis of affective or identity information, and provides a good rationale for studying these paralinguistic features in isolation from the speech signal.

In a multistep model of voice perception Belin, Fecteau and Bédard (2004) proposed that voice perception follows a similar framework to that suggested by Bruce and Young in the domain of face perception. This model proposes that following low-level analysis, speech, affect and identity are processed by dissociable pathways. Furthermore, at each stage of the perceptual analysis, the voice and face are theorised to interact with one another (see Figure 1.1). This model conceptualises the voice as an 'auditory face' with vocal expressions such as sighs, cries and laughs to be equivalent to facial smiles, and frowns. Similarly, the age, gender and identity of an individual can be determined from both the face and the voice. Despite this model being relatively long-standing and widely cited in the literature, only a few studies have empirically tested its assumptions.

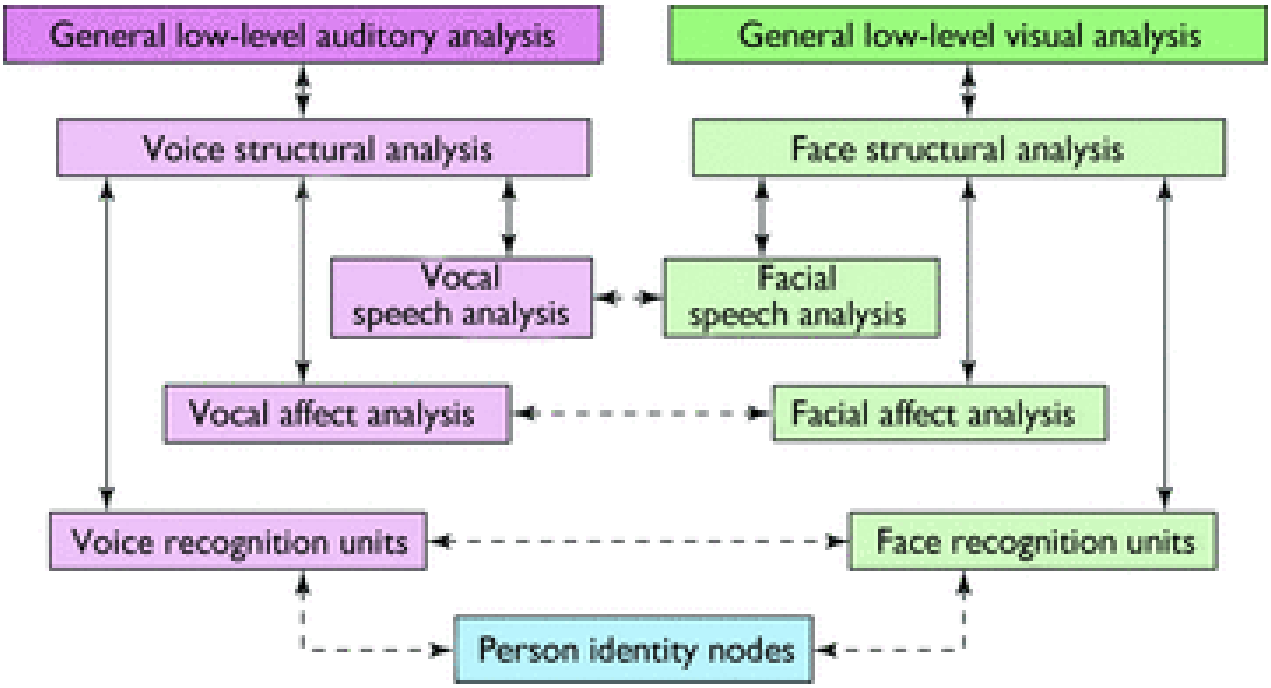


Figure 1.1 A model of voice perception taken from Belin et al. (2004), adapted from the model of face perception by Bruce and Young (1986).

1.3 The structure of the auditory pathways in relation to the 'auditory face' model of voice perception.

1.3.1 General low-level auditory analysis.

The first stage of this model is concerned with the low-level auditory analysis of sound. When a sound wave reaches the ear, there is a series of transformations that take place in order for us to translate that wave into a meaningful auditory object. Upon entering the ear canal, sound waves cause vibrations of the ear drum which then pass through the bones of the middle ear which are known as ossicles. The movement of the final bone causes fluid in the cochlea to move. This fluid then passes over hairs cells in the inner ear, transducing the vibrations into nerve impulses which are sent onto the auditory nerve. The action potentials created travel down the auditory nerve which joins up with the vestibular nerve forming the vestibulocochlear nerve, which transports the sound information through the structures of the midbrain. Low-level features of the auditory signal are then preliminarily analysed by sub-cortical stations including the cochlea nucleus and superior olivary complex which is theorised to analyse temporal, frequency and intensity aspects of the sound signal (Moore, 2003). From here, the sound information is carried to the inferior colliculi via the lateral lemniscus in the brainstem. The inferior colliculi are theorised to relay information, via the medial geniculate body, concerning the location of a sound source to the thalamus and primary auditory cortex. This low-level auditory analysis applies to the analysis of all sounds, whereas the rest of the stages of the model are specific to vocal sounds.

1.3.2 Voice structural analysis.

The structural acoustic representation is somewhat more abstract in nature than the low-level auditory code as it is this that enables us to extract information about various aspects of voice recognition. This is the stage at which the vocal speech, affective vocal information and information regarding the identity of the speaker are extracted from the

auditory code and passed on to their respective analysis streams, which are theorised to be independent of one another. Belin et al. (2004) theorise that this aspect of voice perception is likely to occur in the middle regions of the superior temporal sulcus (STS), in areas located in close proximity to the primary auditory cortex.

1.3.3. Vocal speech analysis.

A vast body of work has focussed upon the ways in which humans perceive and understand the speech signal. Reviewed in depth elsewhere (Hickok & Poeppel, 2000; Scott & Johnsrude, 2003), speech perception will not be the focus of the present thesis, which will instead focus upon paralinguistic vocal features: vocal affect analysis and vocal identity recognition.

1.3.4. Vocal affect analysis.

In situations where the linguistic content of speech is inaudible, it is still possible to gauge reasonably accurately the affective state of the individual talking through both the prosodic features of speech and the non-linguistic expression of emotion in sighs, cries, gasps and laughs. Several acoustic parameters, such as intensity, frequency and rate, provide us with information concerning the affective state of an individual. Integration of these cues is theorised to occur in the auditory 'what' pathway, which projects from the auditory cortices to bilateral superior temporal gyri and anterior superior temporal sulci (Schirmer and Kotz, 2006). Following this, higher-level cognitive judgements are made, with the right inferior frontal gyrus being implicated in the labelling of perceived expression (e.g. Wildgruber et al. 2005) and the right orbitofrontal cortex being implicated in the associated reward value of a given emotion (Rolls, 2000). In this respect, the auditory system is theorised to be hierarchical in the manner in which it processes information such that low-level analysis includes physical features of the acoustic code such as frequency, intensity and temporal changes and higher-level features require more top-down processing relying on semantic

decision-making (Sharp, Scott & Wise, 2004) and prior learning to assist in the accurate perception of acoustic cues (Giraud et al., 2004).

1.3.5. Voice recognition units.

It is at this stage of the acoustic analysis that a voice will be classed as familiar or unfamiliar. Patient studies, demonstrating a double dissociation between speech comprehension and vocal recognition, provide support for the existence of voice recognition units (VRUs) due to observed deficits in vocal identity perception, referred to as phonagnosia (Van Lancker & Canter, 1982). Furthermore, the deficits have been shown to be dissociable for familiar and unfamiliar voices, suggesting differential processing mechanisms for recognition of familiar voices and discrimination between unfamiliar voices (Van Lancker, Cummings, Kreiman & Dobkin, 1988). Neuroimaging studies have shown that vocal identity is associated with activation of the anterior temporal pole and middle superior temporal cortex, irrespective of the degree of familiarity of the speaker (Andics, McQueen, Petersson, Gál, Rudas, & Vidnyánszky, 2010) however areas of the inferior frontal gyrus demonstrated activation only to familiar voices (Latinus, Crabbe & Belin, 2009). However, there is very little literature on the nature of phonagnosia, potentially due to the lack of standardised tests with which to make a diagnosis. The development of such a test, measuring aspects of speaker identity perception would enable increased diagnosis of phonagnosic individuals and shed more light upon the nature of both the deficits present in such cases and the ways in which vocal identity information is processed in normal populations.

1.3.6. Person identity nodes.

Person identity nodes (PINs) refer to individualised nodes of information which contain all referential information for any given individual that we are familiar with. So that when you recognise a face as belonging to person x, you then access the related person identity node, giving you access to other information you might know about this individual,

such as their date of birth or their likes and dislikes. These representations are modality independent and provide access to biographical content relating to the speaker as well as the phonological codes required for the production of the individual's name (Bruce & Young, 1986). In attempt to decipher the nature of these PINs, Gainotti, Ferracciolo and Marra (2010) compared two patients who had suffered atrophy of the anterior temporal lobes of the left and right hemisphere, respectively. The results supported the notion that familiarity judgements are made prior to the supramodal level of representation, in the modality specific streams, where stored representations are compared with the incoming sensory code (Bruce & Young, 1986). In addition to this, it was found that right anterior temporal lobe damage resulted in deficits in facial recognition whereas left anterior temporal lobe damage was associated more with less severe deficits affecting both naming and recognition. Their data also indicated that PINs are not simply a gateway to a single semantic system, as suggested by some theories (e.g. Burton, Bruce & Johnston, 1990), due to the differential recall demonstrated by patients with damage to the temporal pole of both the right and left hemisphere, respectively. If there was a single semantic system for which the PINs provided access, it would be expected that recall of semantic content would be the same for both groups of patients (Gainotti et al. 2010). The contrasting view as to the nature of the semantic representation of PINs was originally proposed by Bruce and Young (1986) and suggests that semantic information is held within the nodes.

1.4 Evaluation of the independent processing of vocal identity, affect and speech.

Deficits specific to speech perception, known as Wernicke's aphasia, have been well documented, supporting the independence of this stream of processing from those that concern the analysis of identity and affect in the voice. Individuals with these deficits generally demonstrate preservation of paralinguistic vocal content (Feyereisen & Seron,

1982), showing significantly more correct and responsive judgements in relation to sentences that were of an emotional content (Boller, Cole, Vrtunski, Patterson & Kim, 1979). Furthermore, these patients also reacted significantly more when the sentences were read by an individual rather than played off a tape recording of the same individual, suggesting that mode of presentation plays an important role in the comprehension of paralinguistic information by aphasic patients.

Additional evidence for the independence of speech from other nonverbal forms of communication comes from primate studies which demonstrate communicative behaviours that are analogous to nonverbal uses of the human voice. For example, several species of monkey have been shown to communicate gender, emotional state and social group relations through their vocalisations (Steklis, 1984). These findings suggest that language evolved in a separate stream to the pre-existing, more primitive nonverbal forms of communication that have shown to be present in our ancestors. This theory is supported by social research on facial expressions which demonstrate consensus with regard to emotion perception seemingly universally (e.g. Ekman et al., 1987). However, more recent research has suggested that facial expressions are not universally recognised, insinuating that this view is possibly over-simplified (Jack, Garrod, Yu, Caldara & Schyns, 2012).

Evidence also exists supporting the independence of vocal identity and vocal emotion. Patients have demonstrated deficits in vocal identity processing whilst their abilities of vocal affect perception remain intact (Garrido et al. 2009; Hailstone, Crutch, Vesterdaard, Patterson & Warren, 2010; Neuner & Schweinberger, 2000). In addition to this, neuroimaging techniques also provide support for a dissociation in the processing of vocal identity and emotion information. An ERP experiment that required participants to make judgements with regard to the congruency of two vocal stimuli on aspects of emotion and identity demonstrated an earlier component for emotion matching trials, occurring at around 200ms,

comparative to identity matching trials for which the activation started at around 300ms. This research provides support for the independence of vocal emotion and identity processing (Sprekelmeyer, Kutar, Urbach, Altenmüller & Münte, 2009) and suggests that fixed aspects of the voice such as identity are processed differently from changeable aspects of the voice such as emotion. Similarly, a study using transcranial magnetic stimulation, suppressing activation in the sensorimotor cortex, demonstrated impairments in the processing of affective auditory content but not in speaker identification, providing further support for the dissociation in these two streams of processing (Banissy et al., 2010). Moreover, using positron emission tomography (PET), it was demonstrated that activation in response to judgements involving identity of the vocal stimulus were different from those involving the emotion portrayed in the vocal stimulus, suggesting that different anatomical pathways are involved in the analysis of these types of information (Imaizumi et al., 1997).

1.5 Acoustics of emotion perception

In the component process model of affective states, emotion is conceptualised as a process, determined by a series of stimulus evaluation checks (see Scherer, 1984a; Scherer, 1984b). Through this process of evaluation, it is assumed that emotional states are continuously re-appraised in relation to each novel stimulus evaluation check that is conducted. It is thought that this model can explain emotional vocalisations in that each stimulus evaluation check, alongside changes in various subsystems impacting upon the somatic nervous system, will produce a distinctive and measurable change in the musculature of the vocal tract.

The responses of the autonomic and somatic nervous systems in relation to different stimulus evaluations result in physiological changes in vocal expression (Scherer, 1986). The autonomic nervous system (ANS) is broken into two branches: the sympathetic and

parasympathetic nervous systems, with the first being associated with immediate responses to stressful situations (such as the fight or flight response) and the latter controlling more energy conserving processes aimed at regulation of automatic functions, making this the slower of the two branches of the ANS. Both aspects of the ANS are likely to impact upon vocal expression of emotion in different magnitudes for different emotions (Scherer, 1986). On the other hand, the somatic nervous system (SNS) is involved in the motor responses of the individual, responsible for aspects of posture, action and locomotion. Emotional arousal is associated with increased motor activity and tension of the respiratory, laryngeal and supralaryngeal musculature, which in turn impacts vocal production (Malmo, 1975).

Previous research has suggested that, through establishing links between theories of emotion perception and the neuromuscular parameters involved in the vocal expression of the various emotions, the knowledge surrounding emotional vocal communication would be greatly improved (Scherer, 1979). One of the acoustic parameters which provides a strong indication of the arousal state associated with a given emotion is that of fundamental frequency (F0). It is the case that emotions with a high, wide ranging and variable F0 often are associated with high arousal emotional states. Other acoustic cues indicative of such high arousal states include loudness and a faster tempo compared to low arousal emotional states whose acoustic profiles demonstrate a low F0 with a narrow range and a small variability. However, it seems that listeners are much more skilled in perception of emotion than can be expected simply through the differentiating profiles of these acoustic parameters.

This apparently paradoxical situation prompted research into acoustic cues that, although harder to obtain parameters for, might provide a more representational picture of the ways in which individuals are able to decode the various emotional states portrayed by the voice with such high accuracy. An example of one such parameter is that of voice quality, or timbre, an acoustic measurement of the pattern of energy distribution across a waveform.

There are several different acoustic profiles related to the various different emotions that are expressed vocally. Table 1 (taken from Scherer, Johnstone & Klasmeyer, 2003) provides a brief overview of the acoustic parameters associated with some of the basic emotional states.

Several studies have attempted to determine the acoustic cues of the emotional stimulus that are useful in decoding the emotional state of the individual. One method of investigating this is through correlation of acoustic parameters of the emotion being expressed with the listeners' perceptual judgements of the emotion. Using multiple regression to regress the acoustic characteristics of the expressions against the perceivers' judgements of emotion, Banse and Scherer (1996) found that between nine and ten acoustic parameters largely explained the variance in the judgements. These included aspects of F0, duration of voicing and mean energy in the expressed emotion vocal samples. Another method of investigating the acoustic cues that contribute to accurate perception of vocal emotion is that of filtering aspects of the acoustic signal. In one such study, using unintelligible speech samples, it was found that F0 and voice quality systematically varied with the strength of the listeners' judgements regarding the perceived emotion, suggesting that these are two important cues in the decoding of a speaker's emotional state (Scherer, Ladd & Silverman, 1984). Using re-synthesis techniques which manipulated neutral utterances on various acoustic parameters including, F0, intensity, duration and contour variability, it was demonstrated that emotional judgements were most influenced by changes in F0 range and high speech intensity (Ladd, Silverman, Tolkmitt, Bergmann & Scherer, 1985). Low F0 range was associated with sadness and large range in F0, and high speech intensity were associated with states of high arousal, negative emotions. Taken together, these studies demonstrate the valuable contribution of research in acoustics to the field of emotion expression and perception (for a review see Scherer, 2003).

Table 1.

Acoustic properties of vocal emotions relative to neutral. Where '<' is less than, '>' is greater than '=' is equivalent to neutal, '<='is less than or equal to and '>=' is greater than or equal to. Taken from Scherer, Johnstone and Klasmeyer, 2003

Acoustic parameters	Arousal/ Stress	Happiness/ elation	Anger/ rage	Sadness	Fear/ panic	Boredom
Speech Rate and Fluency						
Number of syllables per second	>	>=	<>	<	>	<
Syllable duration	<	<=	<>	>	<	>
Duration of accented vowels	>=	>=	>	>=	<	>=
Number and duration of pauses	<	<	<	>	<>	>
Relative duration of voiced segments			>		<>	
Relative duration of unvoiced segments			<		<>	
Voice source-F0 and Prosody						
F0 mean[3]	>	>	>	<	>	<=
F0 5th percentile[3]	>	>	=	<=	>	<=
F0 deviation[3]	>	>	>	<	>	<
F0 range[3]	>	>	>	<	<>	<=
Frequency of accented syllables	>	>=	>	<		
Gradient and F0 rising and falling[3,6]	>	>	>	<	<>	<=
F0 final fall: range and gradient[3,4,7]	>	>	>	<	<>	<=
Voice Source- Vocal Effort and Type of Phonation						
Intensity (dB) mean[5]	>	>=	>	<=		<=
Intensity (dB) deviation[5]	>	>	>	<		<
Gradient of intensity rising and falling[2]	>	>=	>	<		<=
Relative spectral energy in higher bands[1]	> <	> <	> <	< >	<> <>	<= >
Spectral slope[1]		=	=	>	>	=
Laryngealisation		>=	>=		>	=
Jitter[3]		>=	>=		>	=
Shimmer[3]		>	>	<	<	<=
Harmonics/Noise ratio[2,3]						
Articulation-Speed and Precision	?	=	>	<	<=	<=
Formants-precision of location	<		<	>		>=
Formant bandwidth						

Notes:

1. Depends on phoneme contributions, articulation, precision or tension of the vocal tract
2. Depends on prosodic features like accent realisation, rhythm etc.
3. Depends on speaker specific factors like age, gender, health etc.
4. Depends on sentence mode
5. Depends on microphone distance and amplification
6. For accented segments
7. For final portion of sentences

1.6 Identity perception

Several aspects of the speech signal can assist in identifying an individual. Low-level properties of the voice such as F0, intensity and amplitude as well as higher-level aspects of speech such as context, dialect and syntactical framing all contribute to recognition of speakers (Doddington, 1985). Studies using sine-wave replications of speech have provided some insight with regards to the acoustic features of the voice which contribute to successful perception of identity. Fellowes, Remez and Rubin (1997) found that listeners could perceive the identity of a speaker even when sine-wave speech blocked aspects of voice quality from the speech signal. These listeners seemingly relied upon the phonetic properties of the speech signal to differentiate between speakers. However, in relation to paralinguistic voice identification, F0 and formant frequencies have been demonstrated to play an important role (e.g. Bricker & Pruzansky, 1976).

Compton (1963) demonstrated an inverse relationship between speaker misidentification and ranked formant frequencies for familiar speakers: the closer the voices are in terms of formant frequency, the more likely it is that a speaker will be misidentified. Similarly, other research has also demonstrated that F0 and formant frequencies provided the greatest contribution to listeners distinguishing between pairs of voices (Matsumoto, Hiki, Sone & Nimura, 1973). However, Singh and Murray (1978) found that similarity judgements for voice pairs were largely explained in terms of the gender of the voice. The second most important contributor to similarity judgements in female voices was the duration of the speech sample whereas in males it was the fundamental frequency, suggesting that listeners use different cues to guide judgements of vocal identity based upon the gender of the speaker.

Some data suggest that identity in the human voice is perceived in a norm-based manner whereby voices are encoded in relation to an ‘average’ voice (Baumann & Belin, 2010). Within this framework, vocal identity perception is encoded in a ‘voice space’ where

the average voice is depicted at the centre of voice space whereas more distinctive voices are positioned around the perimeters, and are more easily recognised due to their increased distance from the average relative to less distinctive voices (Latinus & Belin, 2011a). Several lines of research provide support for this account (Latinus & Belin, 2011a, 2011b; Latinus, McAleer, Bestelmeyer & Belin, 2013; Zäske et al. 2010)**.** Behavioural data has demonstrated significant aftereffects in response to adaptation to anti-voices: when exposed to a repetitive anti-X-voice stimulus, participants classified subsequently presented identity ambiguous stimuli as being more similar to the voice X (Latinus & Belin, 2011a), suggesting that adaptation had resulted in a shift of the average voice within voice space. In addition to this, it was demonstrated that adaptation to a different identity as opposed to anti-identity produced smaller aftereffects despite physical and perceptual differences being held constant, suggesting that these results are not solely explicable in terms of simple repulsion from any adapting stimulus. Instead, these results are in line with a prototype based encoding of identity in the voice.

In addition to this, neuroimaging data support the idea of a norm-based coding strategy in the perception of vocal identity (see e.g Andics, McQueen & Petersson, 2013). Gender specific stimuli that were of a greater distance form a norm-based prototype were shown to elicit greater activation in the temporal voice area than were stimuli that were closer to the average voice. Moreover, stimuli that were a greater distance were perceived to be more distinctive than were stimuli that were close to the prototypical voice (Latinus, Mcaleer, Bestelmeyer & Belin, 2013). In another study, participants were trained to categorise ambiguous morphs taken from continua between two previously unknown identities. As well as demonstrating different neural activations for the processing of acoustic and identity information, supporting the ‘auditory face’ model of voice perception (Belin et al. 2004), it was also found that voices were coded in relation to an average voice (Andics et al. 2010).

Baumann and Belin (2010) attempted to reduce the dimensions of this multi-dimensional voice space to as few acoustic cues as possible when making judgements regarding the identity of an individual. In a study designed to ascertain any differences in the representation of both male and female voices respectively, participants were required to rate pairs of voices on similarity. Voices used were the vowels ‘a’, ‘i’ and ‘u’ due to the relative stability of the acoustic properties over time (Moore, 2003). Perceptual similarity ratings were then correlated with a number of acoustic properties of the sound wave such as, fundamental and formant frequencies as well as shimmer, jitter and dispersion. F0 provided the primary contribution for speaker differentiation in both genders. However, for females, the second axis was most highly correlated with F1 whereas for males, the second axis was most related to the dispersion between formants 4 and 5. It was concluded that a two dimensional space was an adequate representation of speaker similarity, with the dimensions roughly corresponding to the source and filter aspects of the speech signal. The source-filter theory of speech production posits that sounds have both an origin and a filter which shapes the sound waves, in this instance relating to formant and fundamental frequencies respectively. Taken together, these studies provide substantial support for the ability to decode identity in speakers from very few acoustic parameters, with fundamental and formant frequencies being important in this process.

1.7 What is adaptation?

The perceptual networks of the brain are constantly re-calibrating in relation to their environment, in other words, they exhibit a large degree of plasticity in the processing of sensory input (Clifford & Rhodes, 2005). Adaptation is a tool used in researching mechanisms that underlie these perceptual processes. Adaptation refers to the reduction in response of specific neural populations following sustained repetition of a stimulus and has

proven to be a useful tool in isolating neural populations that are tuned to the processing of specific stimulus attributes (Grill-Spector et al. 1999). This reduction subsequently causes a bias in perception in the opposing direction to that of the adapting stimulus. It is thought that these mechanisms serve to optimise perceptual processing. The efficient coding hypothesis postulates that within a processing system where capacity is limited, the most efficient neural strategy is employed in the evaluation of a given input (Barlow, 1961).

Adaptation could be reflective of many things: a reduction in neural firing in response to repeated presentation of a stimulus, changes in the neural tuning in response to stimulus repetitions or potentially a reduction in processing time for stimuli that are repeated (Grill-Spector, 2006). However, it is well established that adaptation effects are robust and apparent both behaviourally and at the neural level in terms of the responses observed to multiple presentations of an adapting stimulus.

It is thought that perceptual aftereffects occur as a result of prolonged stimulation of the neurons that selectively respond to properties of the adapting stimulus, causing changes in the response characteristics of these neurons (Barlow & Hill, 1963). In a single-unit study which examined the effects of adaptation and sensitivity of neurons in the monkey inferotemporal cortex, it was found that adaptation was maximal for repetitions of the same stimulus as opposed to related stimuli, which showed a lesser degree of neuronal adaptation (Sawamura et al. 2005). Similarly, research has demonstrated that a reduction in BOLD response is observed in fMRI paradigms not only to identical stimulus repetitions but also to related stimulus repetitions, suggesting that populations of neurons which exhibit this response are not sensitive to the differential properties of the adapting stimulus (Grill-Spector et al., 1999).

Aftereffects are greatest for the channels that are most in tune with those of the adapting stimulus and become progressively weaker for channels that represent progressively

more abstract categories (Storrs & Arnold, 2012). Data have suggested that neurons which are most strongly activated in response to the first stimulus presentation will become more adapted than neurons which are less in tune with the properties of the adapting stimulus (Li, Miller & Desimone, 1993). This makes adaptation a useful tool in establishing neural populations that selectively encode various stimulus attributes.

There are currently two accounts of the way in which aftereffects are generated: through bottom-up processes and top-down processes. Bottom-up processes could generate aftereffects through neuronal sensitisation from repeat firing, resulting in a diminished response to previously presented stimuli (Javadi & Wee, 2010). This diminished response results in the responses to novel stimuli being relatively more dominant, thus generating the observed aftereffects. On the other hand, top-down accounts stem from theories of hierarchical predictive coding. These theories posit that at each stage of the processing hierarchy, bottom-up representations are recalibrated in light of recent experience, by means of top-down input, resulting in a prediction error. Adaptation aftereffects in this instance are theorised to represent a greater degree of prediction error as a result of a larger discrepancy between bottom-up and top-down processes.

1.8 Voice morphing in the evaluation of theories of vocal emotion and identity perception

A useful means of investigating the ways in which we perceive different aspects of the voice is through the use of morphed stimuli, varying along a continuum in incremental stages. Such techniques have long been utilised in research in face perception but have more recently become available in the domain of voice perception. Using software that is specialised for use in speech synthesis, manipulation and analysis, it is possible to create vocal morphed continua from two different end point stimuli. This allows investigation of the

perceptual differences between sounds whilst controlling for the acoustic properties of the stimulus. These techniques have been previously employed in both vocal affect and identity research. An example of a vocal morph between two emotions (anger and fear) expressed by voicing the vowel sound /a/ is depicted in figure 1.3. The figure shows 7 morph steps ranging from 5-95% in increments of 15%. The spectrograms and wave forms demonstrate the differences in frequency and amplitude across the continuum.

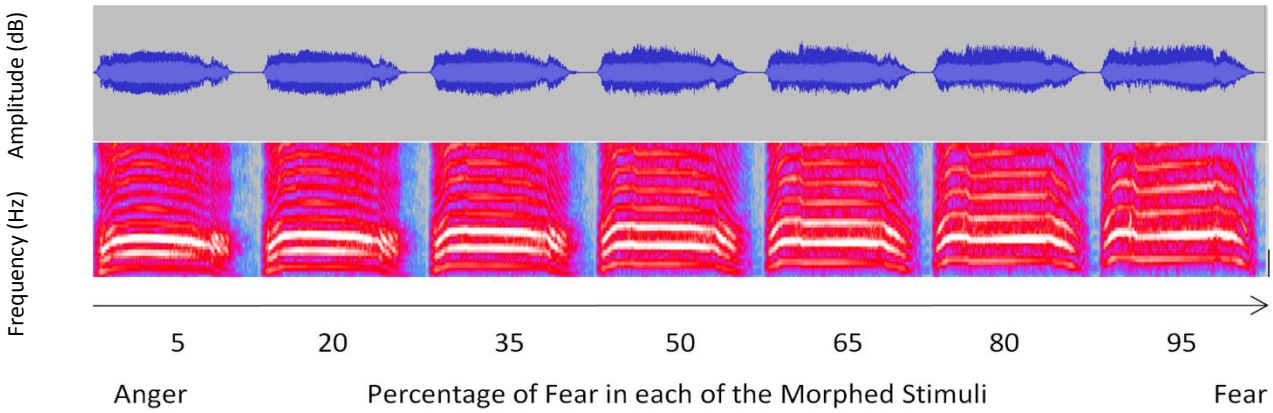


Figure 1.3 Waveforms and spectrograms for seven different morphed stimuli on an anger-fear vocal continuum.

Previous research has demonstrated that vocal emotions are perceived categorically (Laukka, 2005). Using two emotional vocal expressions, six different morph continua were created, ranging from one emotional expression to the other, whereby stimuli were of equal physical distance from one another. Using these stimuli, participants completed both a categorisation and a discrimination task. During the categorisation, stimuli were presented in isolation and were categorised as belonging to one or other of the endpoint emotions. Results resembled a sigmoid curve whereby morphs were clearly perceived as belonging to either one category or the other with the exception of the most ambiguous test morphs, which were categorised randomly. Participants were then required to perform a discrimination task in which three morphs were presented in succession, the third of which was identical to either

the first or the second and participants were required to respond indicating which of the two morphs matched. It was found that discrimination accuracy was greater for stimuli which crossed the category boundary between two emotions. These findings provide support for the categorical perception of emotion as opposed to theories that suggest perceptions vary according to dimensions such as valence and arousal. If this were the case, results from the categorisation task would have resembled a continuous perceptual shift between emotions whereas these results suggest that stimuli are readily perceived in terms of the end point values with the exception of the central morph which represents the category boundary between the two end points of the continuum.

Again using morphed stimuli, similar techniques have been employed in researching vocal identity perception (Latinus & Belin, 2011a). Using a vowel spoken by two different speakers as endpoint stimuli, morphed continua were created. Participants were then required to categorise the morphed stimuli as belonging to one of the two identities both before and after familiarisation with the speakers. Similarly to the perception of emotion, identity was perceived in a categorical manner, demonstrating the typical sigmoid curve. The gradient of the curve became steeper when participants were familiar with the identities that they were categorising, suggesting that learning assists in reinforcing the discrete categories of the individual identities. Taken together these results provide support for the categorical perception of emotion and identity in the voice, potentially as a function of category learning as demonstrated by the lessened categorical perception of unfamiliar versus familiar identities.

1.9 Aims of the thesis

Chapter 2: What does the mental representation of vocal emotion encompass?

Previous research has demonstrated that emotion is perceived in a categorical manner in the voice (Laukka, 2005). However, it is not clear at what stage of the processing hierarchy in the auditory pathways such decision making takes place. Through the use of behavioural paradigms, the thesis examines the likelihood of these judgements occurring as a result of a higher level evaluative process. In addition to this, the extent of generalisation of the neural representation of vocal emotion will be examined: is the representation species specific and does the representation encompass other emotive auditory sounds?

Chapter 3: Is there a representation of emotion that is supramodal in nature and if so, what is the relationship between the processing of emotion and identity at this level of the processing hierarchy?

Typically, research has examined the relationship between identity and emotion unimodally. The model of Belin et al. (2004) does not imply that there is a supramodal relationship between the separate aspects of processing. That is to say, there should be no evidence of identity information in the face influencing the perception of emotion information in the voice and vice versa. As of yet, no study has attempted to verify these claims and the thesis will attempt to do so through a series of behavioural experiments.

Chapter 4: How able are we at discrimination of speaker identity in the voice?

As of yet, no standardised means of gauging individual voice matching ability has been created. It is anticipated that such a test would provide valuable contribution to the understanding of individual differences in vocal identity perception. It will also potentially assist in detecting individuals with phonagnosia, allowing a more in depth analysis of the deficits that are associated with this condition.

Chapter 4: Can individual ability at voice perception influence the magnitude of the perceptual shifts observed in response to behavioural paradigms?

The study of individual differences in relation to perceptual abilities is a growing field. Whereas some individuals perform relatively poorly on tasks involving the perception of noisy or heavily accented speech signals, others perform very well (Gilbert, Tamati & Pisoni, 2013). Some of the factors underlying individual differences in speech perception have been investigated, with results suggesting that individuals who score highly on speech perception abilities demonstrate differences in aspects of working memory and executive function comparative to individuals who score poorly in such tests (Tamati, Gilbert & Pisoni, 2013). However, similar research in relation to paralinguistic aspects of voice perception remains to be carried out. The thesis aims to determine whether performance on a vocal identity matching task correlates with observed differences in the magnitude of perceptual aftereffects following vocal identity adaptation.

Chapter 2: Unimodal adaptation and the perception of emotion in the voice

As discussed in the preceding chapter, adaptation paradigms provide a valuable framework through which to investigate small biases in evaluative judgments and have been previously employed in researching the plasticity of the categorical boundaries of sensory perception. Perceptual processing in the auditory modality is theorised to be hierarchical in nature, meaning that there are both low and high levels of processing involved in the evaluation of any given stimuli. Low-level stimulus features refer to the acoustic properties of the stimulus, whereas higher-level features refer to the more abstract level representations (Nahum, Nelken & Ahissar, 2008). Low-level encoding of auditory information includes detailed representation of finely tuned acoustic features of the auditory stimulus, such as frequency and temporal aspects of the sound wave. In higher-level representations, the categories are somewhat broader and more abstract and entail a degree of semantic or evaluative input. Adaptation aftereffects have been shown at several different levels of processing including low-level stimulus properties such as light, motion, shape and colour (e.g. Wright, 1934) as well as for higher-level properties.

An example of a higher level aftereffect in the visual modality has been demonstrated in relation to facial perception. It was shown that prolonged exposure to an image of a face can generate significant aftereffects on an individual's perception of subsequent faces (Webster & MacLin, 1999) such that adaptation to a face showing distorted features resulted in a perceptual bias causing the test face to be interpreted as being distorted in the opposite direction to that of the adaptor stimulus. Similar aftereffects have been found in other aspects of facial perception such as gender (Webster, Kaping, Mizokami, & Duhamel, 2004; Yang, Shen, Chen, & Fang, 2011), age (Jordan, Johnson, & Fallah, 2008; O'Neil & Webster, 2011), identity (Hills, Elward, & Lewis, 2010; Leopold, O'Toole, Vetter, & Blanz, 2001), ethnicity (Webster, et al., 2004) and affect (Fox & Barton, 2007).

In addition to within-category aftereffects, cross-category aftereffects have also been demonstrated for visual stimuli. Cross-category aftereffects refer to perceptual biases that occur in the perception of stimuli taken from a different category to that of the adapting stimulus. Such aftereffects are of interest as they allow us to understand the nature of the neural representation of a given attribute, as well as providing a more compelling argument in relation to the observed aftereffect being higher-level in its nature. For example, it was found that adapting to bodies produced both gender and identity aftereffects in the subsequent perception of faces (Ghuman, McDaniel & Martin, 2010), suggesting that neuronal populations that respond to bodies are a part of the network involved in the encoding of facial gender and identity. Despite this, no cross category gender aftereffect was observed for adaptation to hands (Kovacs, Zimmer, Banko, Harza, Antal & Vidnyanszky, 2006) suggesting that these stimuli in isolation are not sufficient to activate the network associated with the encoding of facial gender.

Additionally, some studies have reported cross-category object effects for gender adaptation, i.e. gender specific objects such as high heeled shoes elicit perceptual aftereffects in faces, (Javadi & Wee, 2012) where others have failed to elicit such effects (Ghuman et al. 2010). It is possible that these discrepancies are as a result of differences in the paradigms used. Whereas Ghuman et al. showed a single adapting image for a period of 5 seconds, Javadi and Wee showed a series of gender related objects prior to the presentation of the gender ambiguous face, which then had to be gender classified. Participants were required to remember the adapting objects presented as following the facial gender categorisation, an object was presented which participants had to classify as old or novel. This paradigm required subjects to pay a greater degree of attention to the adapting stimuli, and it has previously been shown that attention mediates the magnitude of the aftereffect (Rhodes et al. 2011).

Fox and Barton (2007) investigated the neural underpinnings of perception of emotion in the face using a series of different adaptors. It was found that facial adaptors of the same and different identity to the test stimulus produced significant aftereffects in the perception of ambiguous images of expressive faces drawn from an anger-fear continuum. Furthermore, following adaptation to images of dogs in either fearful or angry poses, there was a trend towards a significant aftereffect in the perception of facial expression. However, verbal stimuli consisting of written adjective emotion words failed to elicit any significant aftereffects in the perception of emotion ambiguous facial morphs. It was demonstrated that greater aftereffects were observed for stimuli with the largest degree of overlap between adaptor and test. This suggests that a larger population of neurons was targeted when adaptors were similar to test stimuli, as opposed to the less corresponding stimuli for which the aftereffects were smaller in magnitude, potentially suggestive of a broader representation being targeted. Taken together, these findings suggest the existence of a more general visual-semantic, not entirely exclusive to human faces in the representation of facial emotion.

Despite the voice offering similar social cues to identity and affective state of an individual as is offered by the face, much less is understood about the representation and neural underpinnings involved in the perception of auditory information in comparison to visual. However, studies still report within-category auditory aftereffects in vocal perception following adaptation to gender (Schweinberger et al., 2008), age (Zäske & Schweinberger, 2011) and identity (Zäske, Schweinberger & Kawahara, 2010). Furthermore, adapting to vocally emotive auditory stimuli has been shown to elicit aftereffects in the perception of ambiguous voice morphs taken from both anger-fear (Bestelmeyer, Rouger, DeBruine & Belin, 2010) and anger-happy (Skuk & Schweinberger, 2013) continua. Results demonstrating that adaptation to caricatured voices still produced significant, although not more prominent, aftereffects in vocal emotion perception suggest that these findings cannot

solely be explained by low-level acoustic properties of the adapting stimulus, and imply the existence of a higher-level representation of vocal emotion (Bestelmeyer et al. 2010). Further supporting the notion that vocal aftereffects represent a higher-level adaptation, sinusoidal tones matched in frequency to that of male and female vocal adaptor stimuli failed to elicit any significant aftereffects in the perception of gender ambiguous vocal morphs (Schweinberger et al., 2008), suggesting that the adaptation aftereffects observed cannot be explained simply by adaptation to pitch.

The present series of adaptation experiments intends to explore which auditory stimuli have the capability of producing perceptual aftereffects in the human voice. This will be tested using three different adaptor-voice combinations: voice-voice, dog call-voice and instrumental expressions-voice. It is predicted that the experiment with vocal adaptors will produce the largest aftereffects such that following prolonged exposure to a voice expressing fear, ambiguous test morphs taken from an anger-fear continuum will be perceived as more angry and vice versa. In order to minimise the likelihood of these aftereffects being generated through acoustic properties of the stimuli, the adaptor and test voices will always be taken from two different individuals. Previous research has demonstrated that the magnitude of the aftereffect is dependent upon the perceived likeness between the adaptor and test stimulus (Hills, Edwards & Lewis, 2010). Given this, it is reasonable to assume that dog calls might still produce an aftereffect in the vocal test morphs but one that is smaller in magnitude than when the voice is used as the adapting stimulus. Due to the lesser correspondence between the instrumental bursts and the vocal test morphs, it is possible that there will be an absence of an aftereffect for these stimuli. It is anticipated that the results will engender current understanding of what is encompassed within the mental representation of vocal emotion.

Experiment 1[1]

Method

Participants. Nineteen participants (10 males, M_{age}=20.2years, *SD*=4.3) were recruited through the student participant panel at Bangor University. Participants were compensated for their time with course credits. All participants were required to provide full informed consent prior to participation and received debriefing upon completion. The study received full ethical approval from Bangor University's School of Psychology ethics committee prior to any data collection.

Stimuli. Adaptor stimuli were vocal recordings expressing anger and fear using the vowel sound /ah/. Audio was recorded using Cool Edit Pro 2.0, a Sennheiser microphone (MKH-40 P48) and Yamaha mixer (MG124c) and was recorded on a PC with a high-specification audio card (M-audio delta 1010). Stimuli were validated, intermixed with other vocal affective expressions, using nine different participants (2 male, *Mage* =23, *SD*=6.32) using a seven alternative forced-choice decision. Only vocal stimuli recognised at an accuracy of 80% or over were used in the experiment. Audio stimuli were edited using Cool Edit Pro 2.0, normalised in energy (root mean square) and presented in stereo via Beyerdynamic headphones at an intensity of 75dB SPL(C). Continua between anger and fear were created in seven morph steps at intervals of 5/95%, 20/80%, 35/65%, 50/50%, 65/35%, 80/20% and 95/5% for each of the identities. In order to establish that these continua were indeed perceived as categorical, 10 participants were asked to classify all stimuli using a two-alternate forced choice judgement task. Similar S-curve functions were demonstrated for these emotional continua as were found by Laukka (2005; see Figure 2.0). The 50% morph

[1] Experiment featured in Pye, A., & Bestelmeyer, P. E. (2015). Evidence for a supra-modal representation of emotion from cross-modal adaptation. *Cognition*, *134*, 245-251.

was determined to be the most perceptually ambiguous morph and therefore, was the focus of the subsequent data analyses.

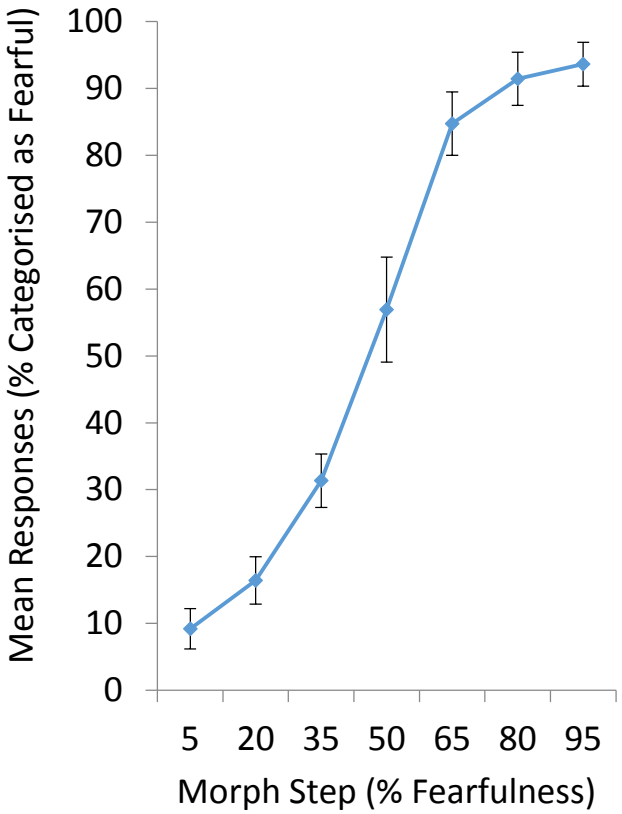


Figure 2.0. A graph to show the mean responses for the categorisation of the morphed vocal stimuli. Error bars represent standard error of the mean (S.E.M.).

Tandem-STRAIGHT (Banno et al., 2007; Kawahara et al., 2008) and Psychtoolbox (Brainard, 1997; Pelli, 1997; Kleiner, Brainard & Pelli, 2007) were used for stimulus manipulation and presentation respectively. Both programmes were run using MatlabR2012b (Mathworks, Inc.).

In order to reduce the likelihood that the aftereffects were driven by acoustic properties of the stimulus, the identity of the test and adaptor stimuli were always different. Thus, 4 adapting identities were used (2 female and 2 male) and 4 morphed continua were

tested. Ten participants were adapted to voice 1 and tested on voice 2 for one male and one female and the other ten participants were adapted to voice 2 and tested on voice 1.

Procedure. Each experiment consisted of 4 blocks: 2 identities (one of each gender) and 2 emotions. Participants had a minimum of a one minute break between genders and a five minute break between emotions. Each block consisted of an introduction to the identity corresponding to the voice used as the adaptor stimulus in the block, a pre-adaptation phase and an adaptation-test phase.

In order to introduce the actors, a short brief was presented containing the name of the identity, their profession, one activity they enjoy doing in their free time and audio clips portraying them as neutral, angry and fearful. Participants clicked through the briefs in their own time and were required to read aloud the information to ensure it was being attended to.

During the pre-adaptation phase, the vocal affective burst of the identity corresponding to the brief was presented sixty times consecutively in order to maximise adaptation. All the stimuli were presented unimodally in the auditory domain and portrayed either anger or fear, depending on the block. There was an inter-stimulus interval (ISI) of 200ms between adaptors and emotions and identities were counterbalanced across participants.

Participants were then required to complete the adaptation-test phase. Fifty six trials were completed per block, with each of the 7 morphs being rated 8 times. Trials consisted of 4 repeats of the adaptor voice, using the same ISI as in the pre-adaptation phase. Following this there was a 500ms interval, during which a plain grey screen [R:127 G:127 B:127] was displayed prior to the onset of the test stimulus. Whilst the test stimulus was played, a fixation cross was displayed on the centre of the screen which remained until the end of the trial. Participants were required to respond to the test morph indicating whether it sounded more fearful or more angry using the ‘z’ and ‘m’ keys respectively. A standard QWERTY

keyboard was used for response input. Trials were response terminated unless no response was given within 5-6 seconds (randomly jittered) in which case the next trial would commence automatically. In the case of a valid response, the next trial commenced following a 500ms inter trial interval (ITI).

A total of 112 trials were completed per adapting emotion (2 blocks x 56 trials). Ratings were averaged as a function of morph step and collapsed across identity for each of the adapting emotions.

Results

All trials in which no response was given were omitted from the analysis. The averaged results for each level of morph step for both anger and fear adaptation conditions are presented in Figure 2.1 (A). Average aftereffect size and individual aftereffect size are shown in Figure 2.1 (B). A main effect of adapting emotion was observed at the 50% morph level ($F(1,18)=13.388$; $p=0.002$; $\eta_p^2=0.427$). Adaptation to angry voices caused ambiguous voices to be perceived as more fearful than when individuals were adapted to fearful voices. Acoustic analyses for adaptor and morph stimuli appear in Appendices A and B.

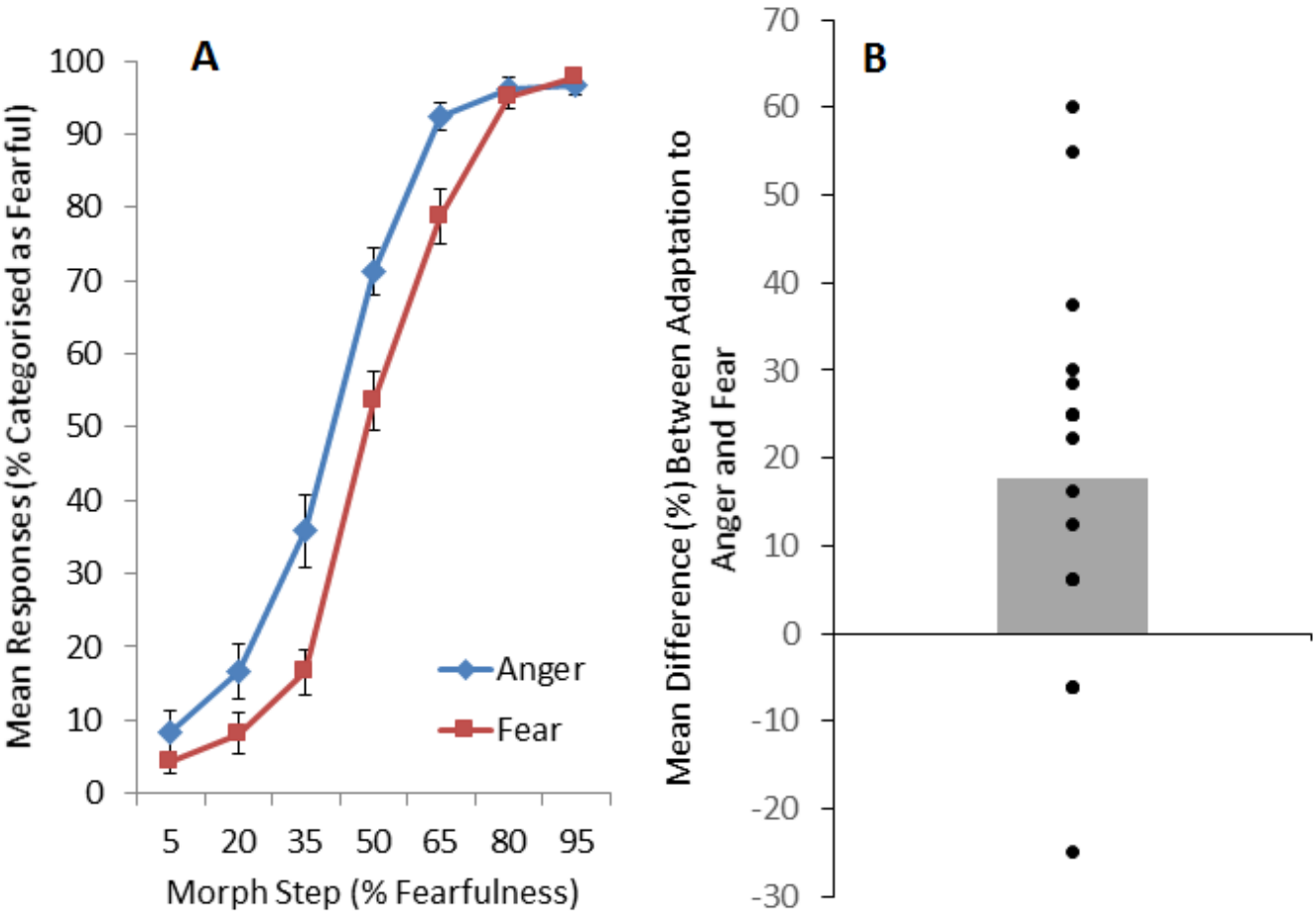


Figure 2.1. **A**: Average responses across participants as a function of morph step. Error bars represent S.E.M. **B**: The bar represents the average magnitude of the aftereffect (adaptation to fear–adaptation to anger). Data points represent individual participants' aftereffects. Data points can be representative of more than one participant.

Discussion

As predicted, and demonstrated in previous research (Bestelmeyer et al. 2010; Skuk & Schweinberger, 2013), aftereffects were observed following unimodal adaptation to affective vocal stimuli. It was shown that adaptation to fearful vocal stimuli caused subsequently presented test morphs taken from an anger-fear continuum to be perceived as more angry relative to when adapted to angry vocal stimuli. Previous research has attempted to determine the stage of the processing hierarchy that the adaptation effect is targeting (see e.g. Bestelmeyer et al. 2010). In relation to the aftereffect observed following adaptation to vocal

emotion, it has previously been demonstrated that caricatured voices fail to elicit an aftereffect that is larger in magnitude than when adapted to normal voices. This result suggests that simple acoustic properties cannot be responsible for the aftereffect, otherwise a greater magnitude of aftereffect would have been observed in response to the exaggerated voices. The current study used adaptor and test voices that were from different individuals, in attempt to minimise the likelihood that aftereffects occurred as a result of adaptation to acoustic properties of the adapting stimulus. Taken together, these results are suggestive of a higher level representation of vocal emotion being tapped as opposed to purely low-level sensory adaptation. It is still possible that there is some aspect of the effect that is driven by acoustic similarities in adaptor and test stimuli as regardless of identity, most individuals will exhibit higher pitched voices for fearful expressions relative to anger expressions.

In an attempt to uncover what is encompassed in this representation of vocal affect and what factors have the ability to manipulate it, further adaptation studies were run using different adaptor stimuli, namely dog noises and instrumental bursts. If an aftereffect is observed in perception of the human voice following adaptation to dog calls, it can be suggested that the representation of vocal affect is perhaps not species specific. It is predicted that if an aftereffect is present to dog calls, it will be smaller in magnitude than that of the unimodal voice experiment on the basis that a smaller population of neurons would be targeted than with the voice-voice adaptation experiment, symbolic of a broader representation of vocal emotion.

Experiment 2

Method

Participants. Ten participants (5 males, M_{age}=27.1years, *SD*=8.02) were recruited through advertisement via email and were paid £6 for participation.

Stimuli. Adaptor stimuli consisted of noises from a Labrador dog taken from a sound database on the internet (www.freesound.org). There was one adaptor stimulus for each of the two emotions: anger and fear. For the anger adaptation conditions, the adaptor was a growling vocalisation whereas for the fearful adaptation conditions, the dog was yelping. Test stimuli were the same as in the unimodal voice study. Therefore, each of the two adaptors was paired with both a male and a female voice continuum, resulting in four blocks in total (anger adaptation with female test, anger adaptation with male test, fear adaptation with female test, fear adaptation with male test). Stimuli were edited using Cool Edit Pro 2.0, normalised in energy (root mean square method) and presented to participants through Beyerdynamic headphones.

Procedure. The current study omitted the brief introduction and pre-adaptation phase so that only the adaptation test phase was completed by participants. Blocks and trial structures used the same parameters as in the voice adaptation experiment.

Results

As before, all trials in which no response was made were removed prior to the analysis. Results were averaged across participants as a function of morph step and are pictured in Figure 2.2 (A). Both the average and individual aftereffect sizes are shown in Figure 2.2 (B). A repeated measures ANOVA with two levels of adapting emotion (anger and fear) demonstrated a significant difference at the most ambiguous level of voice morph ($F(1,9)=6.44$; $p=0.032$; $\eta_p^2=0.417$). The result demonstrated the same effect as in the voice adaptation experiment whereby adaptation to angry stimuli caused the voice to be perceived as more fearful whereas adaptation to fearful stimuli resulted in the voice being perceived as more angry. Acoustic analyses for adaptor and test stimuli appear in Appendices C and D.

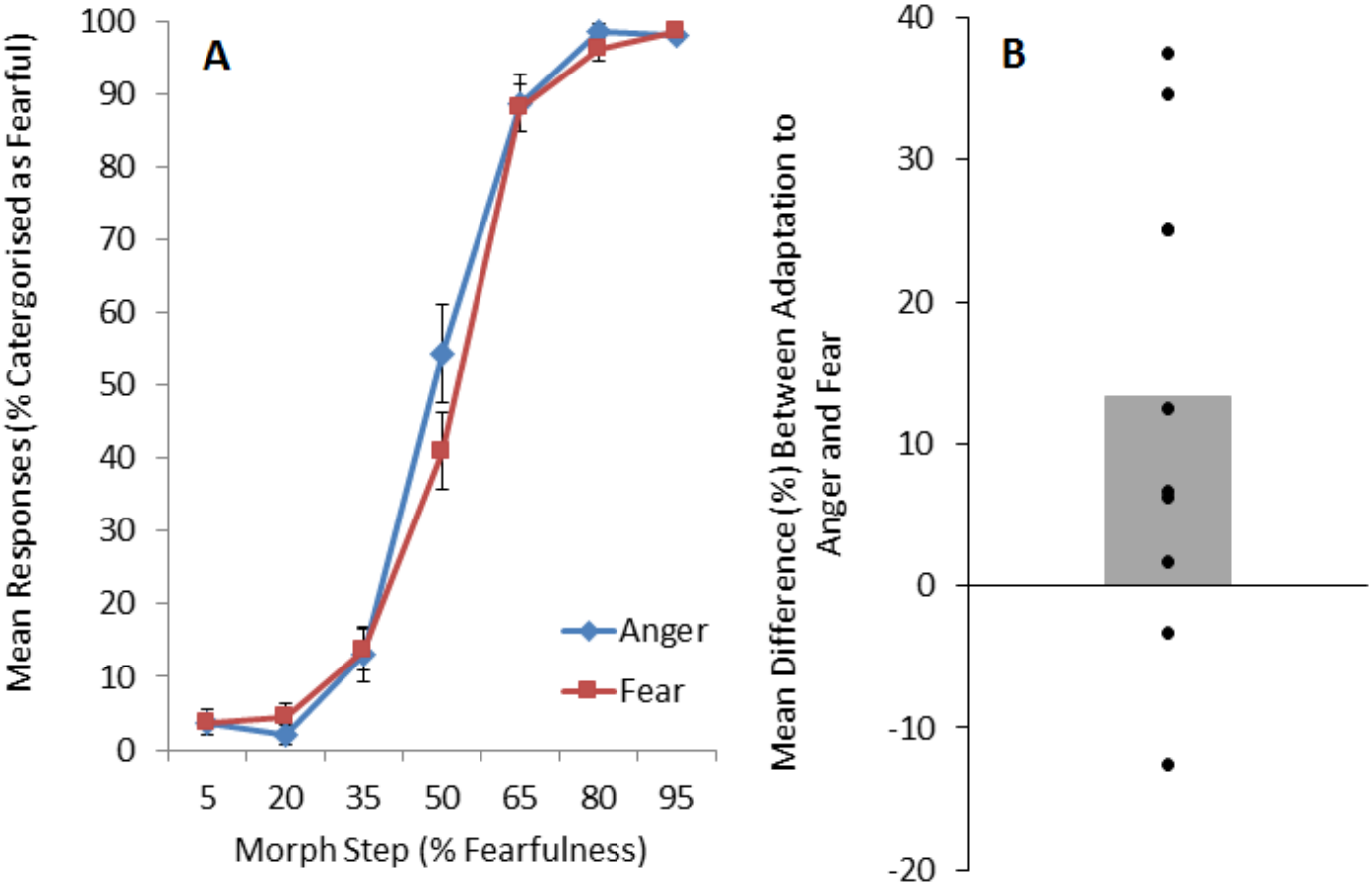


Figure 2.2. **A**: Average responses across participants as a function of morph step. Error bars represent S.E.M. **B**: The bar represents the average magnitude of the aftereffect (adaptation to fear–adaptation to anger). Data points represent individual participants' aftereffects. Data points can be representative of more than one participant.

Discussion

The current study investigated whether adaptation to affective dog calls could elicit significant auditory aftereffects in the perception of emotion in the human voice. Adaptation to noises of the dog growling, resulted in ambiguous voices taken from an anger-fear continuum as being perceived as more fearful relative to when adapted to noises of the dog yelping, which resulted in the ambiguous voices being perceived as more angry. Fox and Barton (2007) found that adapting to visual images of dogs in affective poses, produced a trend towards a significant aftereffect in the perception of emotion in the human face. However, their study used full images of drawings of dogs as opposed to just the face, with a

lesser degree of visual similarity comparative to the acoustic similarity between the dog and human vocalisations in the current study. It is possible that the effect observed here still relates to the acoustic features of the stimulus as the angry stimuli in both of the species vocalisations have a relatively low fundamental frequency. Similarly, in both instances, the fearful stimuli both have a somewhat higher fundamental frequency (see Appendices C and D, Figure A).

Preliminarily, the findings of this experiment suggest the existence of a broader semantic representation of auditory emotion that does not solely encompass that of the human voice. However, further research is required in order to determine that this effect is not solely as a result of acoustic properties of the stimuli. Visual scanning of the graphs suggests that the aftereffects in response to dog calls are smaller in magnitude than those observed in the voice adaptation experiment. In line with research on hierarchical processing, which suggests that higher level neurons have an increased degree of specificity relative to lower-level neurons (Xu, Dayan, Lipkin, & Qian, 2008), it could be proposed that a smaller population of neurons is targeted by the adapting stimulus in the current experiment.

There were some methodological differences between this experiment and the voice adaptor experiment in that the current experiment did not have a brief introduction or a pre-adaptation period. This is likely to have reduced the aftereffect in the present study compared with the first. However, it is likely that even if these phases had been included, the effect observed would still not be as large in magnitude as that of the voice experiment due to the acoustic correspondence between the adaptor and test stimulus being greater for the voice than the animal experiment (Hills et al., 2010; see Appendices A, B, C and D). The third experiment will use musical affective bursts as adaptor stimuli. It is possible that a contrastive aftereffect will not be present due to there being a larger degree of acoustic

difference between the adaptor and test stimuli. Similarly, if an aftereffect is observed, it is expected to be smaller in magnitude than in both of the preceding experiments.

Experiment 3

Method

Participants. Nineteen participants (5 males, M_{age}=20.42years, *SD*=6.7) were recruited using the same methods as in the voice adaptation study and were tested by undergraduate project students.

Stimuli. Adaptor stimuli consisted of musical emotional bursts (MEB) taken from the MEB database (Paquette, Peretz & Belin, 2013). Stimuli consisted of short bursts of musical improvisation portraying basic emotions. Four violin excerpts were used: 2 expressing the emotion of fear and 2 expressing happiness. All stimuli had previously been validated and had been recognised at an accuracy of 90% or more (see Paquette et al., 2013 for stimulus validation details).

Test stimuli were taken from the Montreal Affective Voices database (Belin, Fillion-Bilodeau, & Gosselin, 2008) and consisted of affective vocalisations of the vowel sounds /ah/. Two identities, one male and one female, were used and vocal morphs created on fear-pleasure continua using the same software and in the same incremental stages as in Experiment 1. Stimuli were matched on length, normalised in energy (root mean square) and presented at 75dB SPL (C) through Beyerdynamic headphones.

Procedure. Demographic information was collected; including details of any formal music training participants had received. Participants completed four blocks, one for each of the adaptor stimuli. Within each block, each voice morph was rated four times and morph gender was blocked. This resulted in 56 trials per block (7 morphs x 2 genders x 4 repetitions). Blocks were counterbalanced across participants. A one minute break was given

between blocks of the same adapting emotion and a 5 minute break was given between blocks of different adapting emotions. Stimulus presentation and response procedure followed the same parameters as in the two preceding experiments. Again, the brief introduction and pre-adaptation phases were omitted from the present study so participants only completed the adaptation test phase.

Results

Results were averaged across participants as a function of morph step and adapting emotion (See Figure 2.3 (A)). Both the average and individual aftereffect sizes are shown in Figure 2.3 (B). As with the previous two experiments, a main effect of adapting emotion was observed at the 50% morph level ($F(1,18)=5.04$; $p=0.038$; $\eta_p^2=0.219$). However, instead of showing an adaptation bias, a facilitation effect was observed: when adapted to the fearful musical bursts participants were more likely to rate the ambiguous vocal morphs as fearful whilst adaptation to pleasurable musical bursts increased the likelihood of the ambiguous test morphs being perceived as pleasurable. Acoustic analyses for adaptor and test stimuli appear in Appendix E.

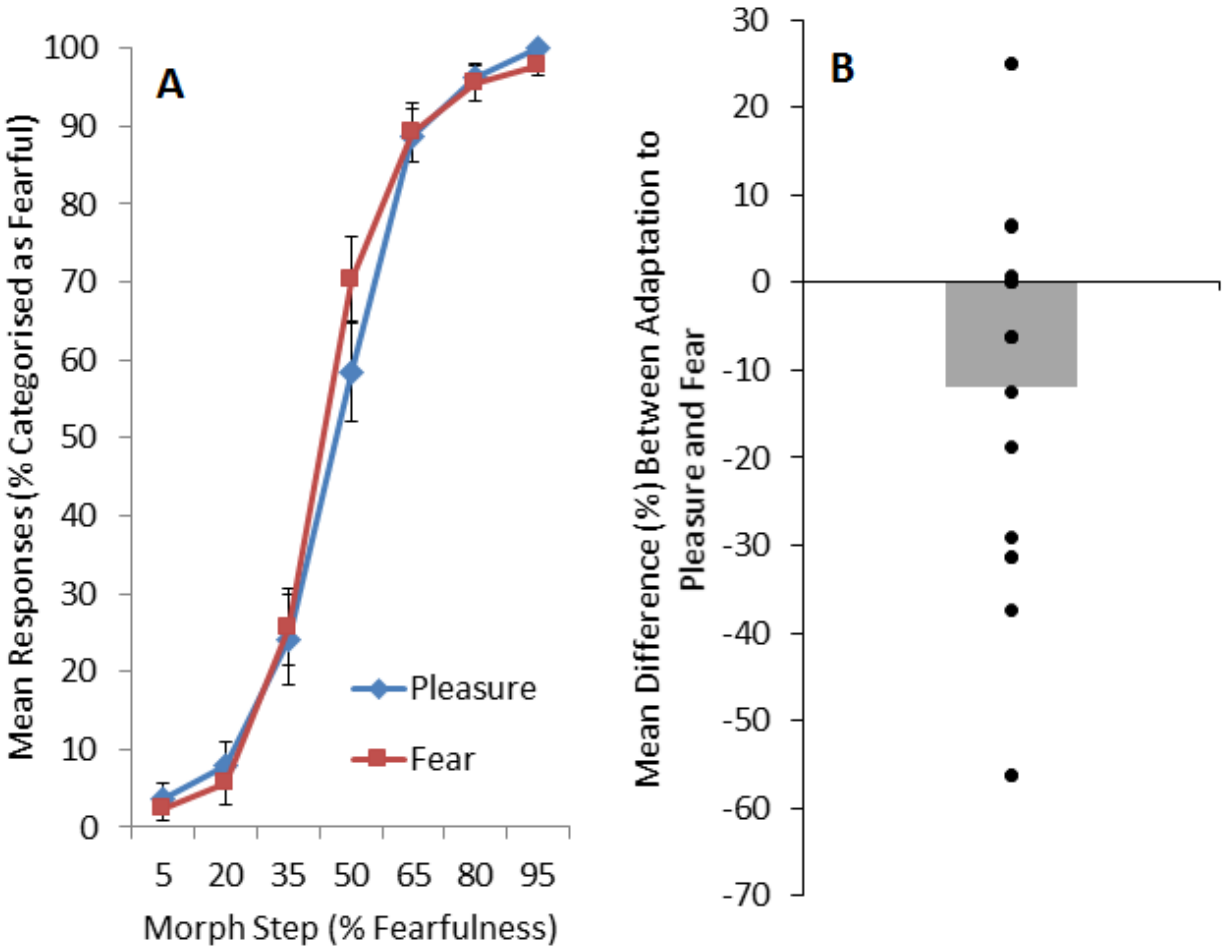


Figure 2.3. **A**: Average responses across participants as a function of morph step. Error bars represent S.E.M. **B**: The bar represents the average magnitude of the aftereffect (adaptation to fear–adaptation to pleasure). Data points represent individual participants' aftereffects. Data points can be representative of more than one participant.

Reaction Time analysis

In attempt to further explore the unexpected facilitation effect at the 50% morph, reaction time analyses were run. Reaction times were collapsed across participants and morph steps for each of the adapting emotions. If indeed the findings are representative of a true facilitation effect, we would expect a reaction time shift in relation to the opposite adapting emotion. However, the two reaction time profiles are very similar, regardless of adapting emotion (see Figure 2.4). This was confirmed using paired samples t-tests at each of the

morph levels in order to compare reaction time. None of these tests reached significance indicating that the reaction times did not significantly differ as a function of adapting emotion.

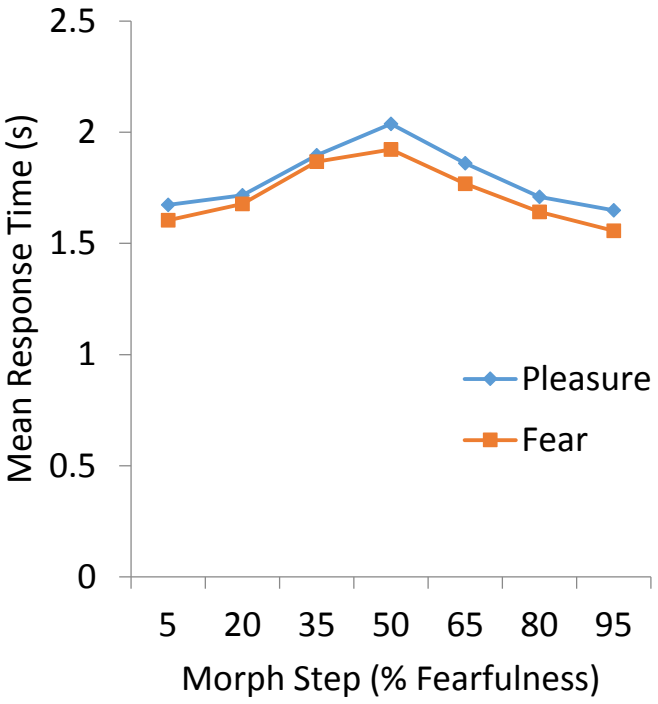


Figure 2.4. Average response times across participants to target stimuli as a function of morph step.

Discussion

The current study investigated aftereffects in the perception of ambiguous vocalisations taken from a fear-pleasure continuum following adaptation to instrumental expressions. The results showed that following adaptation to instrumental expressions of pleasure, ambiguous vocalisations were perceived as expressing more pleasure relative to when instrumental expressions of fear were used as adaptor stimuli. Although it has previously been suggested that the opposite effects elicited from priming and adaptation can be explained in terms of the timing parameters of the presentation of the adaptor/prime and test stimuli (Hills, Elward & Lewis, 2010), this fails to account for the differences observed between the present study and Experiments 1 and 2.

In a study where musical intervals were used as adaptor stimuli, it was demonstrated that adaptation to a major third (two notes played simultaneously, with a distance of four semitones between them), caused a perceptual shift of ambiguous intervals at the category boundary towards that of a minor third (two notes played simultaneously, with a distance of three semitones between them; Zatorre & Halpern, 1979). However, when attempting to determine whether the results were attributable to the intervals themselves or the absolute frequencies of the intervals, it was found that the adaptation effect only remained for adaptor-test stimuli of the same absolute frequency. That is, intervals from a different frequency range to that of the test continuum failed to elicit any perceptual aftereffects, suggesting that the higher level decision stage of perceptual processing is not being reached by these adaptor stimuli. It could be that the lack of adaptation effect in the current study could be explained by the differences in frequency between the adaptor and test stimuli, as was found by Zatorre and Halpern (1979). However, rather than no effect being present at the category boundary, we have a shift towards the adapting stimulus, suggesting that higher level judgements are being tapped by the paradigm, but in the opposite direction to that predicted and previously elicited by the same paradigm.

This result is in line with data from a cross-modal experiment looking at affective priming between musical emotion and facial emotion (Logeswaran & Bhattacharya, 2009). It was found that happy and sad musical excerpts of 15 seconds in length were able to enhance judgements of emotion in happy, sad and neutral facial stimuli. The effect was largest for neutral stimuli. However, the reaction time analysis for the present study suggests that this is not a true facilitation effect due to the highly similar reaction time profiles for both adapting emotions. Replications of this study in the lab do not show the facilitation effect at the 50% morph. Therefore, this result should be treated with caution.

As can be seen by the bar graphs in each of the results sections, this study has a considerably greater degree of response variability at the 50% morph than do the other two studies. It would be beneficial to re-run the study with a more stable set of morphs which produced less variable perceptual ratings in order to determine whether or not the current results are meaningful. Furthermore, the current study used different emotions than the other two experiments, potentially affecting the direction of the results by means of the acoustic differences between adapting emotional stimuli. It would therefore be advantageous to repeat the current study using the same adapting emotions as in the two preceding experiments.

Discussion of Acoustic Analyses

Acoustic analyses were run to determine the mean F0 and F1, harmonic to noise ratio and shimmer for each of the adapting stimuli and morph continua for each of the three experiments reported here (see Appendices A to E). The graphs are mapped with the adapting stimuli as 0 and 100% and the 7 test morphs ranging from 5-95% mapped in between the adaptors. As discussed in chapter 1, there are several acoustic differences between emotional expressions. The first thing to note in these graphic representations is the higher degree of correspondence between the adaptor and test stimuli in terms of all acoustic parameters for the unimodal voice experiment (see Appendices A and B, Figures A-D). Despite the fact that adaptor and test stimuli were taken from two different continua of the same gender, the acoustic parameters remain similar.

However, in the animal-call adaptation study, there are a number of acoustic parameters that vary between adaptor and test stimuli. For female continua, the mean F0 and F1 are lower in the anger adaptor stimuli than the high percentage anger morphs used in the continua (see Appendix C, Figures A & B) . Also, the mean shimmer is greater for anger in the adaptor stimuli than in the high percentage anger female test morphs (see Appendix C, Figure C). The mean harmonics to noise ratio is lower in female adaptors than in female test

stimuli (see Appendix C, Figure D). For males, there is a decreased F0 for anger adaptors relative to the morphs with a high percentage of anger (see Appendix D, Figure A). For fearful stimuli, there is an increase in both F0 and F1 for adaptors compared to male morph stimuli with a large percentage of fear (see Appendix D, Figures A & B). There is an increased shimmer for both anger and fear adaptors relative to their closest corresponding male morphed stimuli (see Appendix D, Figure C). Finally, there is a decreased harmonic to noise ratio for both anger and fear adaptor stimuli relative to their closest corresponding male morphs (see Appendix D, Figure D).

Similarly, there are several acoustic differences between adaptor and test stimuli in the affective instrumental bursts experiment. There is an increased F0 and harmonic to noise ratio in pleasure adaptor stimuli relative to the male morphs which contain the highest percentage of pleasure (see Appendix E, Figures A and D). The first formant is increased in pleasure adaptors relative to both male and female pleasurable morphs. However, the F1 for fearful adaptors is increased relative to male fearful morphs and decreased relative to female fearful morphs (see Appendix E, Figure B). Fearful adaptors had higher shimmer values than both male and female fearful morph stimuli (see Appendix E, Figure C). The HNR values were decreased in fearful adaptors relative to both male and female fearful morphs however, the pleasure adaptors had an increased HNR relative to male pleasurable morphed stimuli (see Appendix E, Figure D).

Taken together, these acoustic profiles can assist in explaining the different aftereffects observed in the present studies. The higher degree of acoustic correspondence between adaptor and test stimuli in the voice experiment probably contributes to the substantial aftereffects observed, suggesting that adaptation to low-level stimulus features might well be contributing to the effect.

General discussion

The present series of experiments explored the effects of different adaptor stimuli on the aftereffects observed in perception of emotion in the human voice. The studies report significant aftereffects following adaptation to both voices and animal noises whereby adapting to angry stimuli resulted in the ambiguous vocal morphs being perceived as more fearful relative to when adapted to fearful stimuli, which resulted in the morphs being perceived as more angry. Contrasting this, the third experiment in which participants were adapted to short affective instrumental bursts, demonstrated the opposite aftereffects whereby adaptation to snippets of fearful instrumental bursts resulted in the ambiguous voice morphs as being perceived as more fearful relative to adaptation to pleasurable instrumental bursts, which resulted in the ambiguous vocal morphs as being perceived as more pleasurable.

Whilst the findings from experiments 1 and 2 were expected, and in line with previous findings and theoretical predictions (e.g. Bestelmeyer et al. 2010; Skuk & Schweinberger, 2013), the results of the final experiment in which instrumental bursts were used proves more puzzling. This experiment should be interpreted with caution and in order to establish whether or not this effect is indeed genuine, it would be beneficial to re-run the study, potentially using the same emotions as in the voice and animal adaptation studies. Replications of the musical affective burst adaptation in the lab, using the same emotions as the present study fail to elicit the same facilitation effect at the 50% morph.

The differences observed in the direction of the aftereffect could be due to the nature of the adapting stimuli. Whereas the voice-voice and dog call-voice experiments use stimuli that are animate in nature, the musical affective bursts are not. This difference could reflect the increased significance of animate relative to inanimate stimuli in our acoustic environments. As discussed in chapter one, research suggests that we are ‘hard-wired’ to respond to the human voice. It could be that this is the case for vocal stimuli in general and

not just that of the same species. If indeed we do process vocal stimuli differentially to that of inanimate stimuli, this could go some way towards accounting for the differences in the perceptual aftereffects reported here.

There are several limitations to the present series of experiments which should be acknowledged. Firstly, it cannot be ruled out that the aftereffects observed in the current studies are not as a result of the acoustic properties of the stimuli. Further research should try and establish whether or not the representation of vocal emotion extends to different species by using animal expressions that differ more in terms of acoustics from the human vocalisations. Secondly, the inclusion of the introduction and pre-adaptation phases in the initial experiment make it difficult to compare the magnitude of aftereffects across studies. Such methodological differences should be addressed in order to establish a greater understanding of factors contributing to the magnitude of an aftereffect.

It is possible that the neural representation of vocal emotion encompasses some of the same neurons involved in the encoding of emotion more generally. Future research should explore the neural underpinnings of these cross-category emotion aftereffects and try to further establish the relationship between instrumental bursts and affective vocal stimuli. If indeed the neural representation of emotion is somewhat broader than a representation of vocal emotion, it is possible that recalibration of the auditory pathways has occurred through adaptation resulting in a large degree of prediction error between top-down and bottom-up processes. Given this, it can be suggested the degree of prediction error was somewhat larger for the voice-voice experiment than for the animal-voice experiment as the aftereffect observed was greater in magnitude. This effect is potentially reflective of the larger neuronal population targeted by the adaptation as a result of the higher degree of correspondence between adaptor and test stimuli, given that aftereffects are thought to be generated as a result

of prolonged stimulation altering the response properties of the neurons that are tuned to the features of the adapting stimulus (Barlow & Hill, 1963).

In order to establish more convincingly that this aftereffect is indeed higher-level in nature, it would be beneficial to run cross-modal adaptation experiments. If an aftereffect was observed in the voice following adaptation to the face, it could be concluded that higher-level representations are targeted by the adaptation, rather than low-level acoustic features of the adapting stimulus. However, the results from the current experiment provide a preliminary suggestion of the existence of a more general auditory semantic of emotion representation that is not specific to human voice.

Chapter three: Cross-modal adaptation and the perception of emotion in the voice

Introduction

Emotion and Identity

There has been a degree of controversy in the facial processing literature surrounding whether or not the subsystems for processing speech, affect and identity are indeed independent of one another as several theoretical models imply (Belin et al., 2004, 2011; Bruce and Young, 1986; Haxby, Hoffman & Gobbini, 2000; Le Gal and Bruce, 2002; Young, 1998 but see Young and Bruce, 2011 for a revised account). These models suggest that the processing of one functionally distinct processing dimension (e.g. emotional expression) is not influenced by a system with a functionally different purpose (e.g. identity). Indeed, primate studies have previously demonstrated that distinct neuron populations code facial identity and expression respectively (Hasselmo, Rolls & Bayliss, 1989).

Through the study of patients who have sustained brain injury resulting in a loss of function, it is possible to identify functionally specific brain systems, giving some insight into their structural organisation. An example of this in the domain of person perception is that of prosopagnosia, which results in the selective impairment of facial identity perception. Whilst still not tested for and not as easily recognised as prosopagnosia, phonagnosia is the impairment of vocal identity perception. In a report concerning two patients, one with a diagnosis of associative phonagnosia and another with a modality independent deficit in person perception, it was demonstrated that familiarity and identification of individuals from the voice was impaired relative to controls, but the ability to identify vocal emotion remained intact (Hailstone, Crutch, Vestergaard, Patterson, & Warren, 2010). These findings, along with similar results in the case of developmental phonagnosics (Garrido et al. 2009), provide support for the dissociation of the processing of emotion and identity in the voice. Similarly, there are clinical cases reporting double dissociations in facial identity and emotion perception (see e.g. Campbell, Landis & Regard, 1986; Etcoff, 1984; Kurucz & Feldmar,

1979; Young, Newcombe, de Haan, Small & Hay, 1993). However, the distinction is not always as straightforward as this, with many cases of prosopagnosics demonstrating deficits in both identity and affect perception in the face (see Calder & Young, 2005). Furthermore, in some cases these deficits have been shown to extend across sensory modalities with people demonstrating difficulties with affect perception in both visual and auditory domains (see e.g. Sprengelmeyer et al. 1999).

Neuroimaging studies also paint a mixed picture. In the 'distributed model' of face perception, Haxby et al. (2000) provide an anatomical framework for various aspects of face perception. This model posits that changeable aspects of the face, such as emotion, are likely to be associated with activation in the superior temporal sulcus (STS) whereas constant aspects of facial perception, such as identity are more likely to be associated with activation in the fusiform cortex. However, functional imaging research has shown fusiform cortex activity to be greater in response to emotional as opposed to neutral expressions, suggesting that this area is also involved in aspects of facial expression (see e.g. Winston**,** Vuilleumier, & Dolan, 2003). However, using an adaptation paradigm, in conjunction with functional magnetic resonance imaging (fMRI), it was demonstrated that facial identity was associated with activation in the fusiform and posterior STS whereas facial emotion perception was associated with activation in the anterior STS (Winston, Henson, Fine-Goulden & Dolan, 2004). These results provide support for distinctive systems involved in the processing of changeable and constant aspects of facial perception respectively.

Behavioural studies have also been used to address the question of the organisation of the mechanisms underlying invariant and changeable aspects of faces. These studies, often exploring identity and affect perception in the face, report mixed findings. Typically these studies employ Garner interference or matching paradigms, whereby stimuli are varied across the dimensions of both identity and emotion. Participants are required to classify stimuli for

both identity and emotion, and accuracy rates and reactions times are then studied with regard to the irrelevant stimulus dimension. Where some studies suggest that identity and emotion are processed independently (Campbell, Brooks, deHaan, & Roberts, 1996; Young, McWeeny, Hay, & Ellis, 1986), other studies in the face literature have contested this notion on the grounds of a growing body of evidence that fails to support such a distinction (Baudouin, Martin, Tiberghein, Verlut & Frank, 2001; Ganel & Goshen-Gottstein, 2004; Schweinberger, Burton & Kelly, 1999; Schweinberger & Soukup, 1998). Such results have caused speculation as to the degree of independence of these two aspects of facial perception (see Calder & Young, 2005).

Fox and Barton (2007) used a series of adaptation experiments to investigate the aftereffects on affect perception produced by different stimuli. They found that expression aftereffects were observed when static facial images of the same and different identities and genders were used for adaptor and test stimuli. However, a larger aftereffect was observed for same identity adaptation than different identity or different gender, both of which were comparable in the magnitude of the aftereffects observed. The results suggest that there are two neural representations of facial affect: a lower level identity dependent representation and a higher level identity independent, general representation of expression, a distinction since supported by other adaptation studies (Campbell & Burke, 2009; Ellamil, Susskind & Anderson, 2008; Pell & Richards, 2013; Skinner & Benton, 2012; Vida & Mondloch, 2009). Similarly, studies of unimodal vocal emotion adaptation in which the adaptor and test stimuli are of different identities (Bestelmeyer et al. 2010; Skuk & Schweinberger, 2013), suggest independence of systems processing these aspects of the voice.

The following series of experiments are designed to investigate the relationship between identity and expression at the supramodal level of perceptual processing. This will be investigated using a series of cross-modal adaptation paradigms manipulating the

congruency of the identity of the adaptor and test stimuli. Through utilising and combining aspects of designs that have previously demonstrated cross-modal adaptation effects, it is anticipated that perceptual aftereffects in vocal emotion perception will be present in response to adaptation to facially expressed emotion. Furthermore, based on unimodal data which has demonstrated aftereffects in emotion perception despite the adaptor and test stimuli being from different speakers, it is predicted that the congruency in identity between adaptor and test stimuli will not alter the perceptual aftereffects observed.

Multi-modal person perception

As discussed in the preceding chapter, a large body of research has investigated low-level and high-level aftereffects unimodally, both within and across categories. However, we cannot be certain that the aftereffects observed in the previous studies reported here are truly representative of adaptation at a higher-level evaluative stage of the processing hierarchy due to the commonalities in acoustic features of the adapting stimuli. More recently, studies have embarked upon testing cross-modal adaptation as a means of probing higher level, supramodal representation within the perceptual streams. Such aftereffects would demonstrate that mental representation of aspects of face and voice such as emotion, identity and gender are shared across modalities. However, the current evidence for cross-modal adaptation aftereffects has yielded mixed results.

One of the earliest studies to report significant cross-modal adaptation aftereffects was in the domain of identity perception (Hills et al, 2010). The study used several different adaptor stimuli in order to determine differences in magnitude of aftereffects between adaptors. Faces (both the same and different images), verbal name stimuli, voices, semantic information and individuals associated to the target were all used as adapting stimuli. It was found that significant perceptual aftereffects were observed following adaptation to all of the different adaptor stimuli. Furthermore, imagining an identity produced significant perceptual

aftereffects in target faces. These studies provide important information regarding the shared representations between aspects of identity association, suggesting that the mental representation of identity encompasses semantic and person related information and aspects of it are independent of the sensory modality in which information is presented.

In a study investigating cross-modal gender adaptation, it was found that adapting to sex typical and sex atypical stimuli produced significant unimodal and cross-modal face-voice and voice-face perceptual aftereffects (Little, Feinberg, DeBruine & Jones, 2013). It was found that unimodally, adapting to masculinised female voices caused masculinised test stimuli to be perceived as more normal than when adapted to feminised voices. However, this effect was not found for male voices following adaptation to female voices, suggesting that our mental representations are to some degree gender specific. Similarly, images of masculinised female faces produced perceptual aftereffects in masculinised female voices, whereby they were perceived as more normal following adaptation. This effect was again absent for male voices following adaptation to female faces. Adapting to masculinised female voices also resulted in masculinised female faces as being perceived as more normal. However, adapting to masculinised female voices had no effect upon the perception of masculinised male faces. Taken together, these results support the notion of a supramodal representation of gender that is separable for male and female stimuli.

The results reported by Little et al. (2013) are at odds with data from a study by Schweinberger et al. (2008) in which silent videos were used to adapt to the speaker gender and test morphs were taken from a male-female morphed vocal continuum. No significant aftereffect was observed possibly as a result of the methodological differences between the two studies. Schweinberger et al. (2008) used 4 repetitions of silent videos as gender specific adaptors followed 500ms later by a gender ambiguous voice morph. Little et al. (2013) used a block design in which participants were required to judge which was the more normal

sounding of a pair of voices/ faces that were masculinised or feminised both before and after a 60 second adaptation block. Little et al. (2013) suggest that the use of visual and auditory stimuli in such quick succession may promote integration across the senses and eradicate any aftereffects that are present.

Likewise, cross-modal adaptation in emotion perception studies report mixed findings, such that adapting to sentences spoken in an emotional tone did not distort perception of emotional faces (Fox & Barton, 2007). Auditory adaptors were emotionally spoken German sentences that were strung together into a five second adapting stimulus that contained both male and female voices. Test images were static photographs of emotional faces morphed between two end-point emotions. The authors concluded that the absence of effect observed in their cross-modal emotion adaptation was potentially as a result of the decreased salience of the auditory stimuli in comparison to that of the visual. It is also possible that the use of static as opposed to dynamic stimuli contributed to the lack of effect observed. The importance of temporal and spatial correspondence between face-voice stimuli has been noted by previous research (King & Palmer, 1985; Stein & Wallace, 1996).

Skuk and Schweinberger (2013) investigated perceptual aftereffects in emotion ambiguous vocal morphs following unimodal, bimodal and cross-modal adaptation to facial and vocal emotion. Adaptor stimuli consisted of consonant-vowel-consonant-vowel angry, happy and neutral utterances and their corresponding facial expressions which were presented as silent articulating videos in the cross-modal condition. In line with previous results (Bestelmeyer et al. 2010), unimodal adaptation produced significant contrastive aftereffects in ambiguous test morphs. Similarly, significant contrastive aftereffects were also demonstrated in the bi-modal presentation of adaptor stimuli. However, in the cross-modal adaptation condition, significant aftereffects were only found to be present for male listeners and not females. The authors suggest that this gender difference can be explained by research

which shows that men are more efficient at regulating emotions in certain conditions than are women (Whittle, Yücel, Yapp & Allen, 2011).

In another experiment which employed neuroimaging techniques, the behavioural data explored the impact of both unimodal and cross-modal emotion adaptation in face-face, voice-voice, face-voice and voice-face instances (Watson, Latinus, Noguchi, Garrod, Crabbe & Belin, 2014). Significant adaptation effects were found to be present for both of the unimodal conditions. However, the only significant cross-modal emotion aftereffect was in the direction of voice-face. The face-voice cross-modal condition failed to reach significance. This study did not use a standard adaptation paradigm as stimuli were presented bi-modally and adaptation effects were explored by analysing the speed of categorisations of stimuli with differing degrees of disparity between the face and voices morphs.

The current series of experiments aims to clarify the nature of the crossmodal aftereffect in relation to vocal emotion perception, using both static and dynamic facial images in order to determine if this is indeed one of the factors underlying the discrepancies in previous findings. It is hoped that by using aspects of both cross-modal paradigms, including an adaptation block as well as a test phase with top-up adaptation, the chances of finding a significant cross-modal aftereffect will be maximised.

Experiment 4- Static cross-modal face-voice adaptation [2]

Method

Participants. Nineteen undergraduate psychology students participated in the study. Participants consisted of 14 females and 5 males (M_{age}=18.84years, SD=1.06) and were recruited through the student participation panel and were compensated for their time with course credits. All participants were required to have normal or corrected normal vision and

[2] Experiment featured in Pye, A., & Bestelmeyer, P. E. (2015). Evidence for a supra-modal representation of emotion from cross-modal adaptation. *Cognition*, *134*, 245-251.

normal hearing, were required to be Caucasian due to known effects of race on perception (see e.g. Bestelmeyer et al., 2010) and were required to have no psychiatric or neurological problems due to the nature of the task. All studies were reviewed and approved by the Bangor University ethics board prior to commencement of data collection.

Stimuli. Adaptor stimuli were images of affective faces taken from The Radboud Faces Database (Langner et al. 2010). Two male and two female identities were used, each displaying the emotions of anger and fear, totalling 8 images. The background colour of the images was altered to grey [R:127; G:127; B:127] in keeping with the colour of the presentation screen and images were cropped to 18.24 cm in width and 28.14 cm in height.

The vocal test stimuli consisted of morphed recordings of vocal sounds containing varying degrees of emotion along a fear-anger continuum with the un-morphed stimuli being taken from the Montreal Affective Voices (Belin, Fillion-Bilodeau & Gosselin, 2008). Four different identities were used (2 male and 2 female). For each identity, 7 morphed stimuli were created differing in anger to fear ratios: 95/5%, 80/20%, 65/35%, 50/50%, 35/65%, 20/80% and 5/95%.

Procedure. Participants completed adaptation blocks for 4 individuals (2 males and 2 females). Identity was blocked but counterbalanced within gender across participants. During each of the 4 blocks, each of the voice morphs was categorised seven times, resulting in 42 trials per block (7 morphs x 6 ratings). Trials consisted of presentation of the adaptor stimuli for 6.5 seconds followed by an inter stimulus interval of 0.3 seconds, after which one of the voice test stimuli was played. The participant was then required to categorise the voice as either angry or fearful using the ‘z’ and ‘m’ keys on a standard QWERTY keyboard. Trials were response terminated unless no response was given, in which case the successive trial began 5-6 seconds (randomly jittered with a uniform distribution) following the presentation

of the test stimuli. In the case of a valid response, there was a 500ms inter-trial interval prior to the onset of the next adaptor stimuli. Adapting emotion was blocked. Participants had a minimum of a one minute break between same emotion blocks (based on Zäske et al.'s (2010) finding on durations of cross-modal aftereffects) and a five minute break between blocks of different emotions.

Results and discussion

Data were averaged across male and female adaptation conditions as a function of morph step and a paired sample t-test was used to compare responses between adaptation conditions at the most ambiguous level of morph (50%). Results revealed no significant differences ($t(18) = -1.084$; $p = .293$) in response to the most ambiguous vocal stimuli following adaptation to angry faces when compared with adaptation to fearful faces (see Figure 3.0 A). There was a large amount of individual variation with regard to the magnitude of the aftereffects observed as shown in Figure 3.0 B.

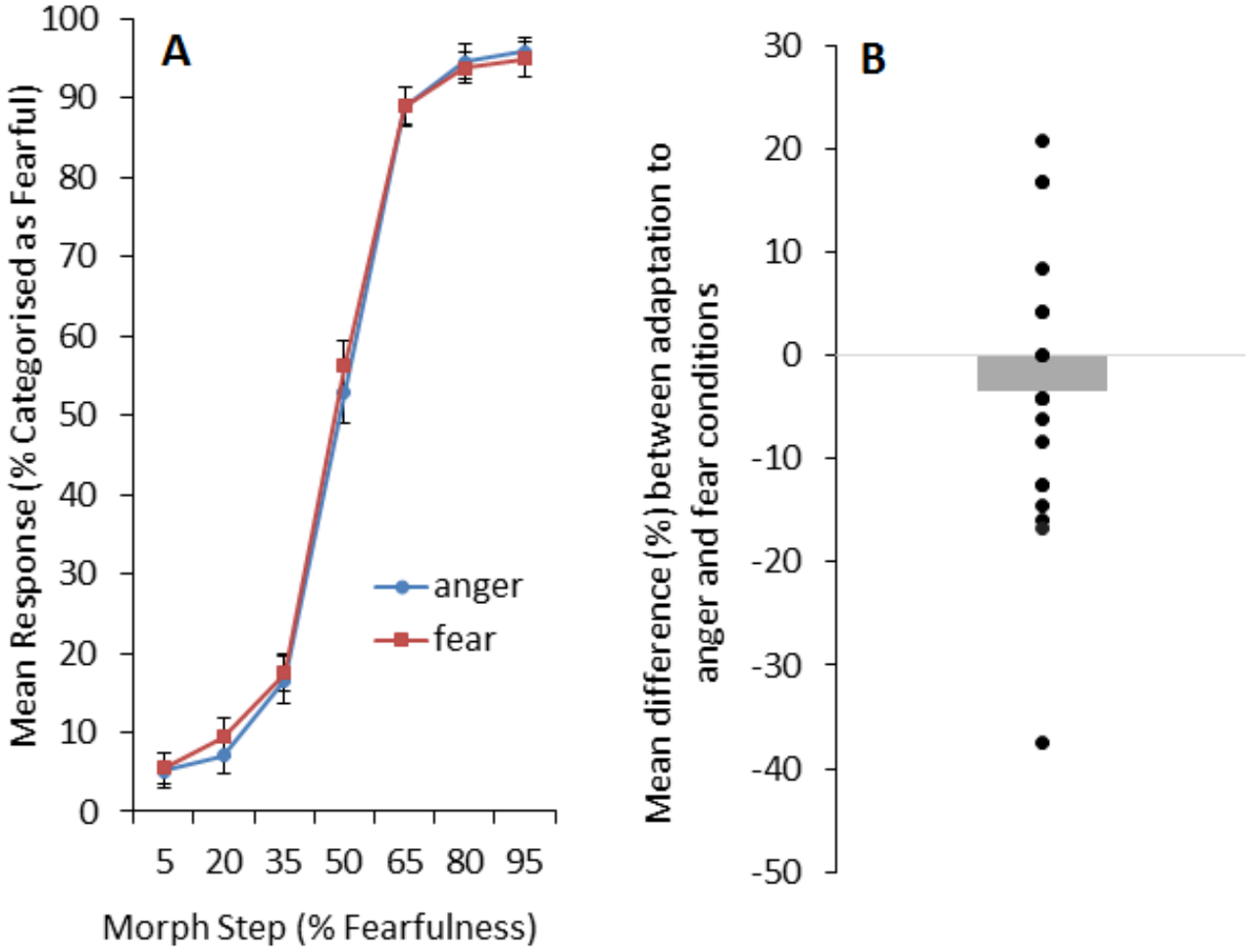


Figure 3.0. **A**: Average responses for both adaptation conditions as a function of morph step (data published in Pye & Bestelmeyer, 2015). Error bars show the standard error of the mean. **B:** The bar represents the average percentage difference between classifications of the most ambiguous vocal morphs following adaptation to angry and fearful faces. Data points represent the difference between the anger and fear adaptation conditions for each participant. One data point can be representative of more than one individual.

The absence of an aftereffect observed in the current study is potentially as a result of the images being static as opposed to dynamic in nature. The use of dynamic facial videos could enhance the degree of binding between adaptor and test stimuli, thus making any cross-modal aftereffect more apparent. Indeed, Calvert, Brammer and Iversen (1998) note the

influence of stimulus commonality on the integration across sensory inputs whereby increased commonality promotes cross-modal binding.

Experiment 5- Dynamic cross-modal face-voice adaptation[3]

Method

Participants. Twenty-five individuals participated in Experiment five (6 male, mean age= 23.4, SD=6.10). Participants were recruited and compensated for their time using the same methods as in the static cross-modal experiment.

Stimuli. Individuals were recruited from the area via poster advertisement for the making of the video stimuli. Twelve individuals were recorded expressing seven emotional affective bursts (anger, fear, surprise, disgust, happiness, sadness, neutral) both facially and vocally. Participants were instructed to express each interjection using the vowel "ah". Videos were filmed using a Canon EOS 5D Mark-II camera with a Canon EF 24-105mm f/4L IS USM lens. Videos were shot in full high-definition at a size of 1920 x 1080 pixels. The camera also recorded low-quality audio, which was used to synchronise the high-quality audio recordings during editing.

Videos were edited so that they started on the last neutral frame available prior to facial movement and ended on the first neutral frame available at the end of the expression. Videos were presented sized at 105 x 140mm and were positioned centrally in the frame, with a grey background (R:127; G:127; B:127). Stimuli were edited using Adobe Premiere CS6 software. High-quality audio was recorded using Cool Edit Pro 2.0, a Sennheiser microphone (MKH-40 P48) and Yamaha mixer (MG124c) and was written onto a PC with a high-specification audio card (M-audio delta 1010). Following recordings, the high quality audio tracks were overlaid on to the video footage.

[3] Experiment featured in Pye, A., & Bestelmeyer, P. E. (2015). Evidence for a supra-modal representation of emotion from cross-modal adaptation. *Cognition*, *134*, 245-251.

Prior to the experiment stimuli were validated in separate, randomised blocks for audio only, video only and combined audio-visual videos. Nine different participants (2 male, mean age=23, SD=6.32), who were naïve to the identities of the individuals in the videos, validated the stimuli using a seven alternative forced-choice task. The stimuli used for all studies reported were recognised at an accuracy of over 80% both uni- and bimodally.

Audio stimuli were edited using Cool Edit Pro 2.0, normalised in energy (root mean square) and presented in stereo via Beyerdynamic (DT 770 PRO 80 Ohm) headphones at an intensity of 75dB SPL(C). Continua between anger and fear were created in seven morph steps at intervals of 5% fear/95% anger, 20/80%, 35/65%, 50/50%, 65/35%, 80/20% and 95/5% for each of the test identities using Tandem-STRAIGHT (Banno et al., 2007; Kawahara et al., 2008). Psychtoolbox-3 (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997) was used for stimulus presentation. Toolboxes were run using MatlabR2012b (Mathworks, Inc.).

To ensure that the newly generated vocal morphs were perceived categorically, 10 of the participants from Experiment four were required to complete a two-alternative forced choice judgement as to which emotion each morph portrayed. The results replicated the sigmoid-shaped curves obtained in previous research (Bestelmeyer, et al., 2010b; Laukka, 2005) with the 50% morph being perceived as most ambiguous.

Procedure. There were four blocks in the experiment: male anger adaptor, female anger adaptor, male fear adaptor and female fear adaptor. Each block consisted of 3 stages: an introduction phase, an adaptation phase and a test with top-up adaptation phase (see Figure 3.1).

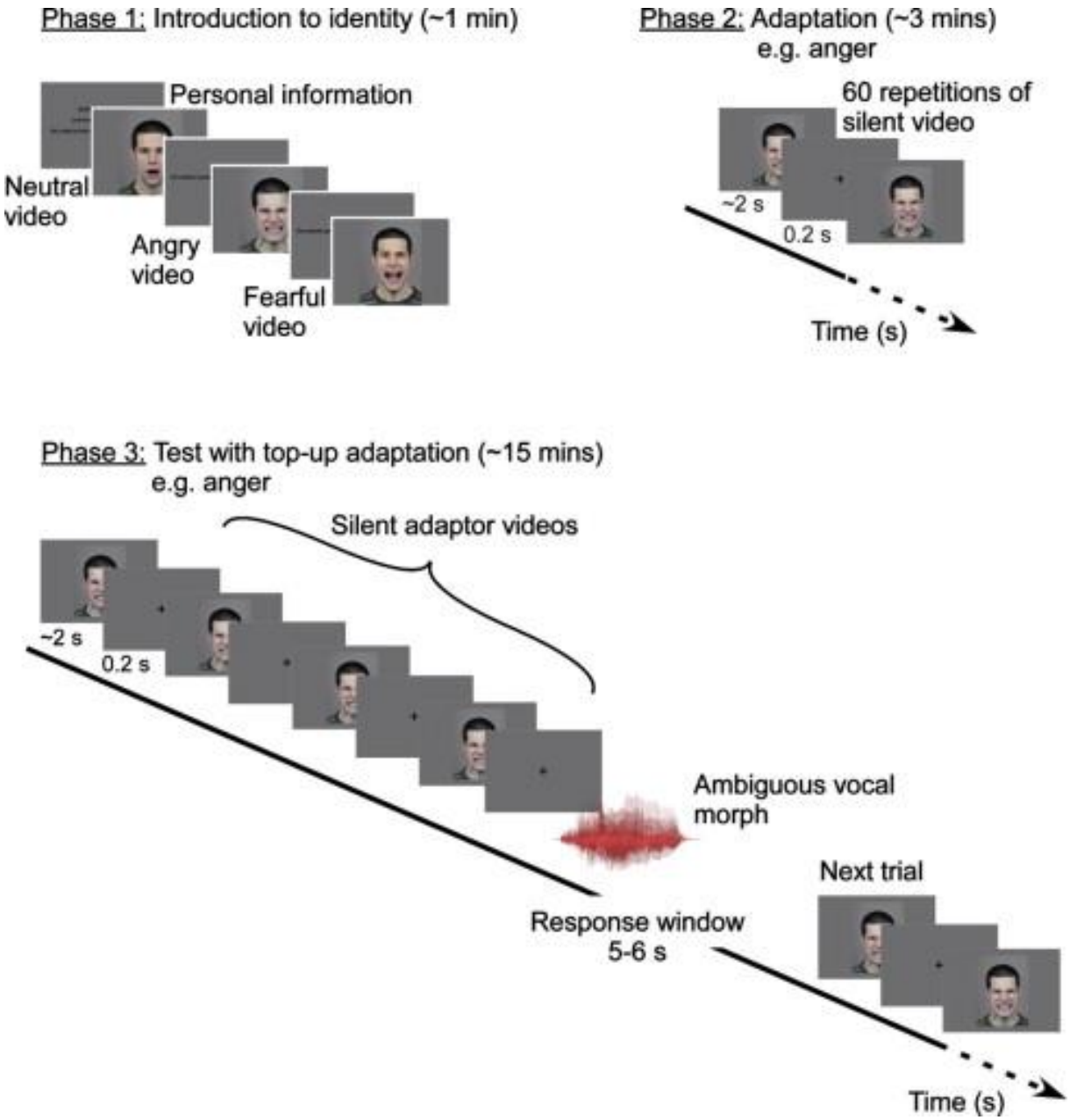


Figure 3.1. A schematic diagram of the three phases employed in the dynamic cross-modal experiments (taken from Pye & Bestelmeyer, 2015).

The introduction phase aimed to familiarise participants with the two identities (one male, one female) used in the study in the hope of increasing face-voice binding. Two short briefs were presented, each containing the name of the identity, their profession, one activity they enjoy doing in their free time and bimodal video clips portraying them as neutral, angry

and fearful. Participants clicked through the briefs in their own time and were required to read aloud the information to ensure it was being attended to.

The second phase was the adaptation phase. During this, the video of the identity corresponding to the brief was presented sixty times consecutively. Videos were presented without audio and portrayed either anger or fear. There was an inter-stimulus interval (ISI) of 200ms between adaptors. Participants were asked to observe the video carefully for as long as it was on the screen. The phase was included with the aim of maximising any cross-modal adaptation effects that might occur.

Participants were then required to complete the test with top-up adaptation phase. Trials consisted of four repeats of the adaptor video, using the same ISI as in the adaptation phase. Following this there was a 500ms interval, during which a plain grey screen was displayed, prior to the onset of the test stimulus (morphed vocal stimuli) which was presented through headphones whilst a fixation cross was displayed on the screen. Trials were response terminated unless no response was given within 5-6 seconds (randomly jittered) in which case, the next trial commenced automatically. The ITI was 500ms and each of the seven morph steps was rated 8 times, totalling 56 trials per block. The same inter block timings were used as in the preceding experiment.

Results and discussion

Figure 3.2 (A) presents average responses, in terms of percentage of fearful classifications, across participants for each morph step when adapted to facial expressions of anger and fear, respectively. A main effect of adapting emotion on the 50% morph was confirmed using a paired samples t-test ($t(24)=2.824$; $p=0.009$; $d=0.56$), comparing adaptation to anger with that of fear. Following adaptation to angry faces, ambiguous voices were perceived as more fearful relative to when adapted to fearful faces. Figure 3.2 (B)

shows the average aftereffect observed across participants as well as individual aftereffects. Again, a large amount of individual variability in aftereffect size was observed.

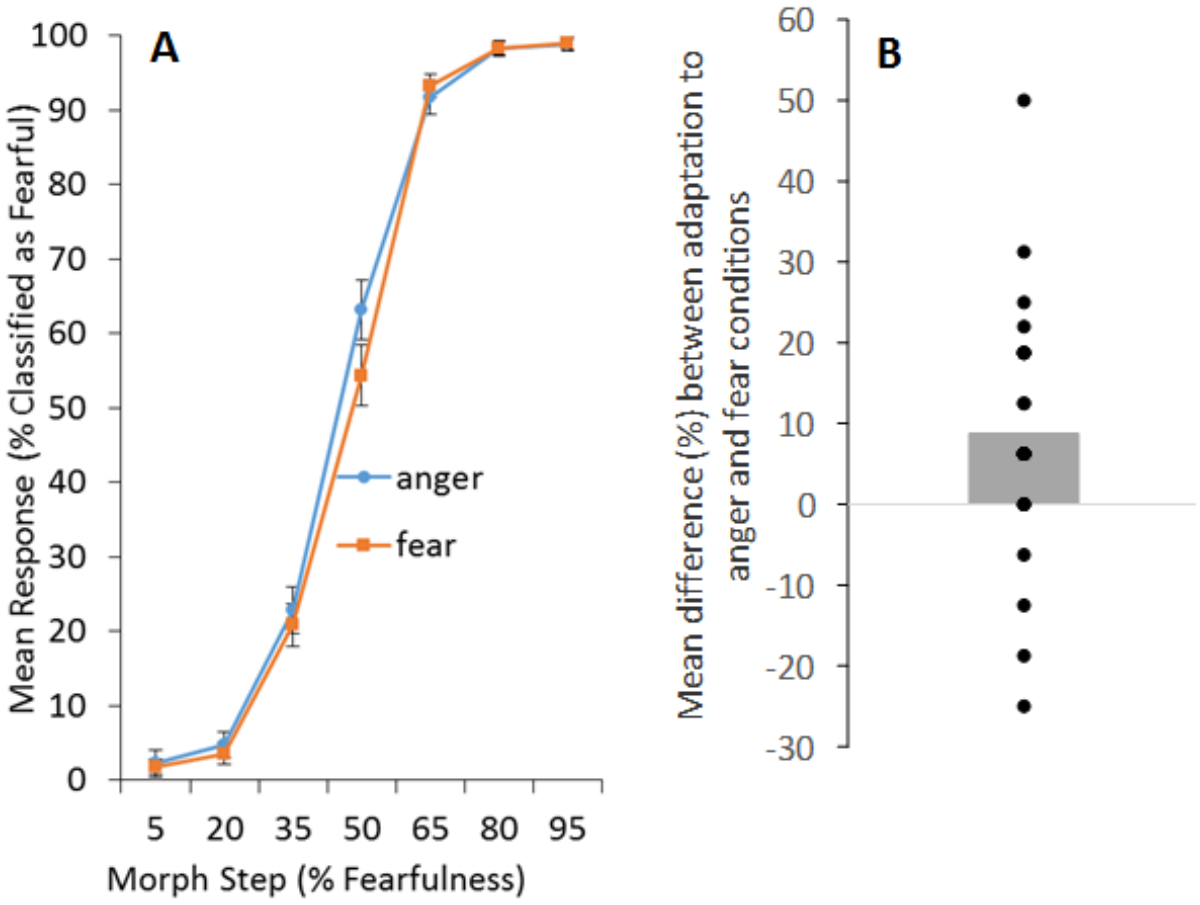


Figure 3.2. **A:** Average responses for both adaptation conditions as a function of morph step. Error bars show the standard error of the mean. **B:** The bar represents the average percentage difference between classifications of the most ambiguous vocal morphs following adaptation to angry and fearful faces. Data points represent the difference between the anger and fear adaptation conditions for each participant. One data point can be representative of more than one individual.

Here we demonstrate the ability of facial emotional expression to bias our subsequent perception of emotion in a different sensory domain, namely the voice. Similar cross-modal aftereffects were reported by Skuk and Schweinberger (2013). However, these were found to be contingent upon the gender of the participant, a factor that was not shown to affect results in the current study. The discrepancy in findings between the findings of Skuk and Schweinberger (2013), and the results of the present study are probably largely down to methodological differences between the two. Whereas Skuk and Schweinberger used adaptation within trials, we also incorporated an adaptation block into our design in order to maximise any perceptual aftereffects that may be observed. In order to further explore the relationship between emotion and identity at the supramodal level, a further experiment was run to determine whether the cross-modal effects were contingent upon the identity of the adaptor and the test stimulus being of the same identity.

Experiment 6- Dynamic cross-modal adaptation with different identities[4]

Method

Participants. Twenty-three participants (eight male, mean age=20.7 SD=4.09) completed the study, none of whom had participated in the two preceding studies. Participants were recruited through the student participant panel and awarded course credits for taking part.

Stimuli. For Experiment six, the two identities in Experiment five were used as the adaptor stimuli and voice morphs as well as three additional identity videos for each gender.

Procedure. As in Experiment five, participants completed four blocks, each with three phases. The introduction and adaptation phases were identical in both this and the preceding experiment. The only difference between the two studies was in the test with top-

[4] Experiment featured in Pye, A., & Bestelmeyer, P. E. G., (2015). Evidence for a supra-modal representation from cross-modal adaptation. *Cognition, 134*, 245-251.

up adaptation phase where four different identities were used for the videos in the 56 trials: the familiar identity from the first two phases plus three unfamiliar identities per gender. The same voice morphs were rated as in Experiment five, belonging to the two identities that the participant had been familiarised with. During a block each identity appeared on 14 trials and each level of morph step was rated twice per identity.

Results and discussion

In Experiment six participants knew that the voice morphs always belonged to the face of the introduced identity so that on some of the trials the adaptor-test pair was identity-congruent and on others it was incongruent. Responses were therefore averaged separately for congruent (Figure 3.3 A) and incongruent (Figure 3.3 C) adaptor-test identity trials as a function of morph step. A 2 x 2 repeated-measures ANOVA was run with two levels of adapting emotion (anger and fear) and two levels of adaptor-test identity (congruent and incongruent). Again a significant main effect of adapting emotion was observed at the 50% level ($F(1,22)=5.102$; $p=.034$; $\eta_p^2=0.188$) showing the same effect as in Experiment five. In addition, there was no significant effect of identity at the 50% morph level ($F(1, 22)=0.469$; $p=.501$; $\eta_p^2=0.021$) and no significant interaction between identity and emotion ($F(1, 22)=0.75$; $p=.786$; $\eta_p^2=0.003$), suggesting that identity does not mediate the cross-modal aftereffect in affect perception.

The variability in aftereffect size was greater for the non-adapted identity compared to the adapted identity (see bar charts, Figure 3.3 B and 3.3 D). This is due to the increased number of trials for the non-adapted identity: for every one trial of familiar adaptation, there were three of unfamiliar. So the total percentage correct responses for the familiarised identity was divided by eight (2 ratings per morph x 4 blocks), whereas the total percentage correct responses was divided by 24 for unfamiliar identities (2 ratings per morph x 3 identities x 4 blocks). This should not have an impact on the analysis as the assumption of

sphericity is only applicable to conditions with three or more levels. Additionally, the range of magnitudes of aftereffects is the same for both experiments.

The results from the current experiment replicate the cross-modal aftereffect in vocal emotion perception as was observed in Experiment five. In addition to this, the current results of no interaction between identity and emotion suggest that identity congruency does not mediate the cross-modal relationship in vocal emotion perception. This finding implies that the representations of emotion and identity are independent of one another at the supramodal level of the processing hierarchy. In order to determine if the magnitude of aftereffects was different for the unimodal vocal emotion adaptation experiment reported in chapter two to that of the cross-modal experiment reported here in experiment five, further analysis was run.

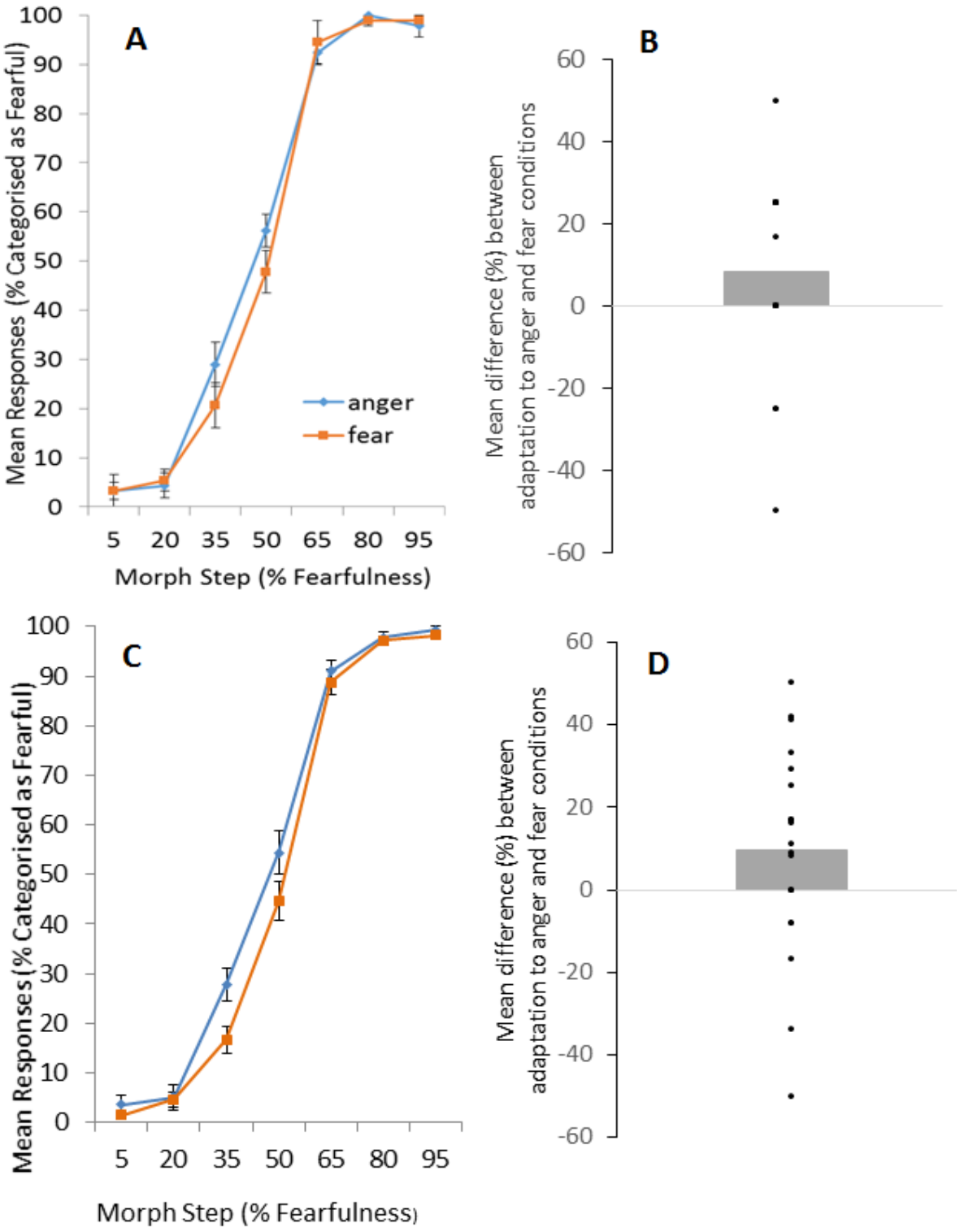


Figure 3.3 **A:** Average responses for both adaptation conditions as a function of morph step in the identity-congruent trials. Error bars show S.E.M. **B:** The bar represents the mean percentage categorisation across participants. Data points represent the difference between the anger and fear adaptation conditions for each participant. **C&D:** As in A & B but for identity-incongruent trials

Comparison of aftereffects in unimodal and cross-modal experiments

An independent samples t-test was run comparing the differences between adaptation conditions at the 50% morph level, in the unimodal voice experiment and the cross-modal blocked identity experiment. This was done in order to see if the aftereffect was significantly different between unimodal and cross-modal adaptation conditions. These two studies were chosen due to them having the closest proximity in aspects of methodology. No significant differences were observed in magnitude of aftereffect between unimodal and cross-modal adaptation conditions at the most ambiguous level of morph step ($t(42)=1.592$; $p=$(n.s.)).

General discussion

The current series of experiments investigated the factors necessary to elicit cross-modal aftereffects in the perception of vocal emotion. Static images of emotive faces failed to elicit any significant aftereffects whereas dynamic videos, bearing a higher degree of correspondence to the vocal test stimuli did indeed elicit contrastive aftereffects at the most ambiguous level of vocal morph. These effects were present irrespective of the congruence of identity between adaptor and test stimuli, suggesting that the supramodal representation of emotion is independent of identity. These results provide support for a high-level representation of emotion that is independent of sensory modality and low-level perceptual analysis.

Previous research has suggested sensory processing to be hierarchical with higher-level neurons theorised to have a higher degree of specificity than lower-level neuronal populations (Xu, Dayan, Lipkin, & Qian, 2008). Therefore it could be predicted that smaller effects should be observed from higher-level adaptation studies, reflective of the reduced number of neurons targeted by the adaptation effect. However, this was not found to be the case, with a t-test confirming no significant differences in the magnitude of aftereffect

following unimodal and cross-modal adaptation at the most ambiguous level of test morph. However, the adaptation effect is more extensive in terms of the morph levels at which effects are observed in the unimodal compared to the cross-modal condition, suggesting that there is an increased degree of shared representation between voice-voice adaptation compared with face-voice adaptation.

Experiment four demonstrated that static facial images were not powerful enough adaptors to elicit any cross-modal aftereffects. Similarly, Fox and Barton (2007) reported significant uni-modal but not cross-modal aftereffects for static facial expressions. In the cross-modal condition they used neutral sentences read with emotional prosody as adaptors and still images of facial expressions as test stimuli. In Experiments five and six, we have maximised adaptation by adapting participants prior to the test phase and again before each test stimulus with *dynamic* stimuli of the same event. Again, using emotional linguistic stimuli, Skuk and Schweinberger (2013) found cross-modal aftereffects but only for male participants. Maximising adaptation, the use of dynamic stimuli and non-linguistic expressions may be key in obtaining consistent cross-modal aftereffects behaviourally.

With regard to the supramodal relationship between affect and identity perception, our data tentatively support the view that these features are processed independently of one another. Although this finding cannot clarify the relationship between identity and expression at a unimodal level, it may suggest that supra-modal perception of identity and emotion are processed in parallel streams. However, it is possible that cross-modal effects may be so small that a modulation with identity is not detectable with behavioural experiments and therefore it could be interesting to explore these behavioural effects in combination with electrophysiological measures. In order to clarify the ways in which identity and emotion are processed supramodally, it would be useful to run comparative experiments reversing the manipulation keeping the identity of the adaptor stimuli constant but manipulating the facial

expression. Furthermore, it would also be of interest to ascertain whether voices as adaptor stimuli have a similar capacity to elicit cross-modal aftereffects in the face in order to establish whether or not emotional auditory information has the capacity to influence visual perceptual norms, as seems to be the case for the effect of emotional visual information on auditory perceptual norms.

There are several limitations to the present series of studies. Firstly, it is possible that the introductory briefs used to familiarise participants with the identities used were not substantial enough to ensure face-voice identity binding. It is possible that through the use of personally familiar stimuli, a mediating effect of identity on aftereffect size might be observed. This study was attempted in the lab, with stimuli recorded from lecturers in the psychology department. However the quality of the stimuli was poor and there were no instances in which both facial and vocal emotions were validated to be recognised at greater than 80% accuracy. Further research should attempt to discern whether personally familiar identity congruence mediates the magnitude of the aftereffect at the cross-modal level of adaptation.

Taken together, the current results support the existence of a representation of emotion that is independent of sensory modality. Our experiments show that high-level contrastive aftereffects of emotion are modality and perhaps also identity independent, providing support for Belin et al.'s (2004) 'auditory face' model. Future studies are necessary for the development and evaluation of multimodal person perception models and the neural correlates that underpin these processes.

Chapter Four: Individual differences and the perception of identity in the voice

Psychological testing is used to objectively measure individual differences in performance on a wide range of tasks. Such measures include matching and discrimination abilities, object recognition and naming as well as detection of motion and the experiencing of visual illusions. There are several factors that need to be taken into consideration when designing and constructing a new psychometric measure, the most notable of which include the concepts of reliability and validity. Reliability refers to the measure's ability to yield similar results upon multiple administrations whereas validity refers to the test's ability to measure what it is claiming to measure.

There are several types of reliability; inter-rater, test-retest, internal consistency and inter-method or parallel forms reliability. Inter-rater reliability refers to the degree of agreement between judges or raters on their assessments. Test-retest reliability refers to the test's ability to generate stable results upon repeat testing. Internal consistency is the degree to which items designed to measure the same construct correlate with one another within a given measure or scale. Finally, parallel forms reliability is demonstrated with similar performance across different versions of a test.

Similarly, there are three main types of validity: construct validity, content validity and criterion validity. Construct validity refers to how well a test relates to and measures the underlying construct that it is based upon and is established using both convergent and discriminant validation (Cronbach & Meehl, 1955). High convergent validity is demonstrated by strong correlations between measures of the same construct whereas high discriminate validity is shown by low correlations between the measure in question and constructs that are unrelated to it. Content validity refers to the ability of a measure to encompass all aspects of a given construct and requires a degree of subjective agreement upon the nature of the construct in question. Finally, criterion validity refers to the relationship between a given measure and real world criteria that should theoretically be related to performance on that

measure. Criterion validity consists of concurrent validity (the ability to differentiate between groups of individuals who should theoretically vary in performance on a measure) and predictive validity (the ability of a measure to predict performance on related tasks).

The preceding chapters have demonstrated large individual variability in the magnitude of aftereffects following both unimodal and cross-modal adaptation. One possible explanation for such variability is differences in the ability to perceive and discriminate paralinguistic vocal cues. One area in which perceptual ability has been widely studied is that of face matching, which refers to an individual's ability at matching images of the same face and discriminating between non-matching faces. There are currently several tests which are commonly employed in order to gauge the ability of individuals in the domain of face matching and discrimination. However, no such equivalent test exists for measuring voice matching ability. Such a test would assist in researching factors that contribute to and relate to voice matching ability.

What have we learned from face matching research?

Whilst it has been demonstrated that we are very good at identifying familiar faces, unfamiliar faces pose a greater problem. This has implications in several domains such as border control, sales of age restricted items and eye-witness accounts of a crime. For example, Kemp, Towell and Pike (1997) measured shop assistants' ability at matching a photograph to an individual. In some instances of mismatched identity trials, the photograph on the identification card was considered to be of minimal likeness to the shopper and these were correctly rejected only 67% of the time, suggesting that our ability to match and discriminate identity is far from optimal. Other applied research reports similarly low levels of performance accuracy when matching identity off CCTV footage as would often be required in the context of criminal identification (Davis & Valentine, 2009). In addition to

this, a recent study of passport officers demonstrates large variability in performance and high susceptibility to error irrespective of training received and the number of years spent in the job (White, Kemp, Jenkins, Matheson, & Burton, 2014). Such results demonstrate the need for screening measures to be implemented in both the recruitment of specialised personnel and in determining the credibility of witnesses with regard to their identification abilities. The implementation of standardised tests designed to gauge individual ability at such specialist tasks is one way this can be achieved.

Several tests currently exist that are designed to measure ability at unfamiliar face matching. However, some of these tests have been criticised on their validity, prompting revisions and modifications. The Warrington Recognition Memory for Faces test (RMF; Warrington, 1984) and the Benton Facial Recognition Test (BFRT; Benton, Silvan, Hamsher, Varney, & Spreen, 1983) are two tests that were once commonly used for this purpose. The RMF requires subjects to rate the pleasantness of 50 faces before testing memory for the images. This memory test is done by means of simultaneously presented pairs of faces, one of which the participant had previously encountered and the other a novel image with the participant being required to choose which they had seen before. Whereas this tests memory for unfamiliar faces, the BFRT tests an individual's ability to match images of unfamiliar faces. In the BFRT, an array of images is presented, comprising of one target photograph and an array of 6 other photographs. From this array, participants are required to identify 3 images which match the identity of the target photograph. However, an evaluation demonstrated that it is possible to achieve scores within the normal range in the absence of any facial features being visible (Duchaine & Weidenfeld, 2003). This finding suggests that participants with deficits in face matching can achieve normal scores through reliance upon features external to the facial structure such as hairline and clothing. In light of these shortcomings, the Cambridge Face Memory test (CFM) was devised (Duchaine &

Nakayama, 2006). Unlike the BFRT, stimuli are presented sequentially in order to avoid subjects using a feature-matching strategy when responding, again making it a test of memory as opposed to matching ability. The Glasgow Face Matching Test (GFMT; Burton, White & McNeill, 2010) however, is a task in which pairs of faces are presented simultaneously and the participant is required to say whether the images are of the same or of different individuals. The photographs presented in this test are from the same viewpoint but are taken using different cameras. Measures of visual short-term object memory, face recognition memory and matching familiar figures were run in conjunction with the GFMT, with the latter being found to correlate most highly with performance. These, and results from other studies insinuate that unfamiliar faces are processed on more of an object based level than are familiar faces (e.g. Megreya & Burton, 2006; Woodhead & Baddeley, 1981).

In attempt to establish baseline performance rates in different types of face matching tasks, Megreya and Burton (2008) ran a series of experiments. In the first of these, subjects were presented with a target followed by an array of ten test faces and were required a) to determine if the target was present and b) to identify the target amongst the test images. Average performance accuracy was demonstrated to be 70%. A second experiment followed the same procedure but the target and test faces were presented simultaneously, eliminating the involvement of memory. Despite this, performance was still only 68% accurate. In a final experiment, a same-different judgement task was employed and average performance accuracy reached 85%. Other studies have demonstrated similar levels of mean accuracy for such tasks (Bruce, Henderson, Newman & Burton, 2001; Henderson, Bruce & Burton, 2001)

Results in face matching tasks exhibit a wide range of performance, with some individuals performing well above average. For example, a range of ability (62-100%) was observed on the GFMT matching task with the CFM demonstrating similar results (60-100%). Several factors have been found to influence face matching ability. While some

results demonstrate that there is no relationship between accuracy and gender of the viewer (Burton et al., 2010; Duchaine & Nakayama, 2006), other studies have demonstrated an interaction between the gender of the stimulus and the gender of the viewer whereby accuracy levels are equivocal for viewers when the stimuli are male but women demonstrate higher accuracy when identifying female faces (Herlitz, & Lovén, 2013; Lewin & Herlitz, 2002; McKelvie, Standing, St. Jean, & Law, 1993; Rehnman & Herlitz, 2006). There is a large body of literature that suggests performance accuracy is increased when dealing with populations that are similar to our own. Research has documented several of these biases including own-gender (see e.g. Wright & Sladden, 2003), own-race (for a review see Meissner & Brigham, 2001) and own-accent biases (see e.g. Bestelmeyer, Belin & Ladd, 2014; Stevenage, Clarke & McNeill, 2012) amongst others. Age has also been investigated as a factor affecting performance. In the GFMT, no correlation between age and overall accuracy was observed but other studies have demonstrated an own-age bias in face recognition (Bartlett & Leslie, 1986; Fulton & Bartlett, 1991; for a review see Weise, Komes & Schweinberger, 2013).

Taken together, these studies researching face matching ability and factors which contribute to it have demonstrated large individual differences in accuracy and a tendency to be better at processing faces that are from similar categories to our own. Other studies have looked at the ability to match faces with voices and the factors which contribute to performance accuracy in such measures.

Cross-modal face-voice matching tasks

Previous research looking at matching of familiar faces and voices has demonstrated high accuracy of face identification despite the identity of the accompanying voice (Stevenage, Neil & Hamlin, 2014). However, faces interfered systematically with voice

identification whereby voices were identified with highest accuracy in conditions where the face and voice were of the same identity, accuracy was lower when the face was of a semantic relation to that of the voice and the lowest accuracy rates were seen in the "mismatched" condition where the face and voice had no relationship to one another. The authors' example was that the matched condition could be the face and voice of TV presenter Ant, the related condition would be the face of Ant but the voice of his co-presenter Dec and the mismatched condition would be the face of Ant paired with the voice of any individual who you would not normally associate with that face. Indeed these results support previous research that suggests the voice to be less easily recognised than the face, with equivalent performance only being achieved through the use of out of focus facial images (Damjanovic & Hanley, 2007; Hanley & Turner, 2000). Furthermore, research has looked at individuals' ability to retrieve semantic information relating to familiar voices and found performance to be far worse than the face equivalent task (Brédart, Barsics & Hanley, 2009; Hanley, Smith & Hadfield 1998). Despite this, matching identity across modality was shown to be equivalent for face-voice matching and voice-face matching (Kamachi, Hill, Lander & Vatikiotis-Basteson, 2003).

Taken together, previous findings in face perception and current tests that exist provide us with an insight into what a good voice matching test might look like and how it would be used to further our understanding of voice identification. As face matching ability has not demonstrated ceiling effects despite simplifying versions of the task, a preliminary measure of voice recognition accuracy might be better off focusing on the less complex measurement in order to gain an understanding of general ability. Doddington (1985) distinguishes speaker verification from speaker identification whereby verification refers to the ability to judge a stimulus as belonging to a certain identity. In this case, the two outcomes are that the stimulus is either classified as matching the identity or classified as a

mismatch in identity and judgements are made on the proximity of the stimulus to that of the identity being referenced. On the other hand, identification refers to tasks in which a stimulus is chosen from a number of options as belonging to a given identity. Therefore, the number of possible outcomes becomes dependent upon the number of reference speakers in the task and the strategy employed is that of discriminating against other reference stimuli in order to determine which of the given samples is the most proximal. Doddington suggests that the added difficulty present in identification tasks, relative to the verification, stems from the requirement of an increased understanding of the variability of aspects of the voice and speech features. Therefore, a preliminary measure of voice matching ability may benefit from using a verification task rather than an identification strategy due to the lessened complexity of such measures.

What factors contribute to voice matching ability?

One area in which such research has been carried out previously is that of earwitness testimony where memory for unfamiliar voices has been investigated. Studies have investigated the most reliable means of speaker discrimination, demonstrating that direct listening proved to be superior to both psychophysical scale ratings and analysis of physical waveform properties (Clarke & Becker, 1969). It has also been shown that there is large variability in people's ability to identify voices (Carbonell, Grignetti, Stevens, Williams & Wood, 1965; Stevens, Williams, Carbonell, & Woods, 1968; Williams, 1964). Several factors affecting voice recognition have been researched, reporting mixed findings. For example, some research has shown deterioration in accuracy with longer retention intervals between presentation and test (Clifford, Rathborn & Bull, 1981; Deffenbacher et al. 1989; McGehee, 1937; Yarmey, 1991a) whereas other research has demonstrated stability of voice recognition over time (Legge, Grosmann & Piper, 1984; Saslove & Yarmey, 1980; Yarmey, 1991b).

Mixed findings are also observed with regards to gender effects in vocal identity recognition. In terms of the gender of the target, McGehee (1937) reported that males were better at identifying female voices. However, Thompson (1985) found that recognition for male voices was better than that for females, irrespective of the gender of the witness.

Age has also been found to contribute to individual differences in earwitnessing ability. Research has demonstrated that witnesses over 40 are generally less accurate than younger adults. Furthermore, the ability to recognise vocal identity appears to develop over the course of adolescence, with peak performance being achieved during the later teenage years (Bull & Clifford, 1984). Mann, Diamond and Carey (1979) also investigated the developmental time course of unfamiliar voice recognition. It was found that ability increased between ages 6-10, with 10 year olds performing as well as adults. However, between ages 10-13, accuracy declined, only returning to the adult ability level at age 14. The length of the sound recordings has also been investigated as a factor affecting performance, demonstrating mixed results. Bricker and Pruzansky (1966) compared identification accuracy rates when presented with sentences, syllables and vowels of co-workers. A progressive decrease in accuracy was observed as the length of the speech sample was reduced, with mean performance being 98%, 84% and 56% respectively. Also investigating sound duration on accuracy, Compton (1963) determined that identification of a familiar speaker was possible in just 25ms, using a single vowel as the target stimulus. In addition to this, accuracy of identification of speakers was also shown to be greater for familiar rather than unfamiliar speakers (Bricker & Pruzansky, 1966). Naive listeners correctly identified 75% of voices when asked to decide whether they came from speaker A or speaker B, compared with 98% accuracy when listeners were familiar with the individuals speaking.

In a study using similar methodology to the current test, results demonstrated that overall mean accuracy on a voice matching test was 60%, which is considerably higher than

would be expected if performance was at chance level (33%). In attempt to determine whether vocal identification was subserved by language processing networks in the left hemisphere, or paralinguistic, nonverbal factors normally associated with right hemisphere networks, Doehring and Ross (1970) used a voice recognition task in which participants were required to match which one of three, nonsense consonant/vowel/consonant (CVC) syllables was spoken by the same individual as a target vowel. Two versions of the test were presented, one to the right ear and the other to the left. This was done in attempt to determine the presence of any right ear superiority but the results showed no difference in ear presentation. One factor this study failed to address is that of individual variability in performance and this will be the aim of the current project. Through the development and standardisation of a voice matching test, it would be easier to establish what factors contribute to voice matching ability and would help to clarify the discrepancies in current findings.

A test to determine voice-matching ability

As discussed in chapter one, there are several aspects of the speech signal which can assist in decoding the identity of a speaker. Using a principal component analysis, Baumann and Belin (2010) attempted to reduce the dimensions of voice space to as few as possible. Their conclusion was that males relied upon fundamental and first formant frequencies in decoding vocal identity whereas females used fundamental and fourth and fifth formants. However, recent research has demonstrated that fundamental and first formant frequency, (corresponding to the source and filter aspects of speech production), are sufficient for both genders to effectively make decisions concerning speaker identity (Latinus & Belin, 2011a; Latinus & Belin, 2012).

According to the auditory face model of voice perception (Belin, Fecteau & Béddard, 2004; Belin, Bestelmeyer, Latinus & Watson, 2011), aspects of vocal information, namely speech, emotion and identity information, are processed in interactive but functionally

dissociable pathways. Thus, the identity of a voice should be discernible in the absence of speech due to the fact that when we hear two individuals say the same sentence, we can distinguish the identity of the individual not based upon what is being said but upon the paralinguistic content of the speech signal. Indeed this has been demonstrated in studies which have used reversed speech samples as stimuli (Bricker & Pruzansky, 1966; Clarke, Becker & Nixon, 1966; Van Lancker & Kreiman, 1985; Wiliams, 1964).

The ability to match unfamiliar voices remains poorly explored and to date, no standardised voice matching test for unfamiliar identities exists. The aim of the current project is to provide a reliable and valid, standardised measure of gauging an individual's ability to match unfamiliar voices on identity. Such an instrument would be useful for a number of reasons. Deficits in auditory identification of individuals currently have no standardised measurement criteria. The development of a test which measured individual ability at voice matching would provide a valuable diagnostic tool for these populations which struggle with vocal identification. Such a test would provide a valuable means of exploring factors relating to individual differences in voice matching ability. Furthermore, the measurement of voice matching ability could assist in ascertaining the credibility of a witness in cases involving earwitness testimony.

As the most simple face matching tests demonstrate accuracies that are far from ceiling level in performance, the current test will be a simple same-different matching task, without any memory requirements. Pairs of voices will be presented to participants, requiring them to respond as to whether the voices come from the same speaker or from different speakers. The voice stimuli will be single CVC or vowel/consonant/vowel (VCV) syllables as opposed to speech, in line with the theory that identity and speech information are processed in dissociable pathways (Belin et al., 2004). Previous findings demonstrating that fundamental and formant frequencies are two substantial contributors to perceived speaker

similarity, enables mapping of voices in Euclidean space, providing us with the ability to manipulate trial difficulty (Baumann & Belin, 2010). By pairing voices that are close together in terms of first and formant frequencies, it is anticipated that it will be harder for individuals to discriminate between the identities when compared to trials in which first and fundamental frequency differ greatly.

Based on findings from the face matching literature, several predictions can be made. Firstly, it is expected that overall performance will potentially demonstrate worse overall accuracies than that of face matching literature due to the increased potency of the face relative to the voice, evidenced by findings in the cross-modal face-voice matching literature that suggests a greater degree of influence of the face on the voice than that of the voice on the face (Stevenage, Neil & Hamlin, 2014). Secondly, it can be predicted that there will be large individual differences in performance on the task based upon tests in the domain of face matching that have shown this to be the case (see e.g. Burton, White & McNeill, 2010).

Method

Participants. Three hundred and six participants (215 female; M_{age}=23.08; *SD*= 7.6 years) completed the voice test. Participants were recruited through university participant panels, email and poster advertisements and were either compensated £6 for their time or awarded course credits. All participants were required to give full informed consent prior to starting and were fully debriefed upon completion of the test. The study was approved by the Bangor University ethics committee preceding any data collection.

The Test. A voice database comprised of 330 recordings (148 males, 182 females) were edited and used as stimuli for the voice test. The database consisted of vocal recordings from both Glasgow and Bangor. So as not to let accent or minor differences in recording quality affect judgements of stimuli, voices recorded in Bangor were always paired with other voices from Bangor and likewise for Glaswegian voices.

For each of the speakers in the database, individual sound files of all CVC, VCV syllables and vowels were generated. The vowel sound ‘e’ was used to measure the fundamental frequency and first formant frequency (see Appendix F) of each of the voices in the database. ‘E’ was chosen as it provides a relatively stable sound wave (Moore, 2003). The voice files were analysed using PRAAT (Boersma and Weenink 2005) sound analysis software, where the most stable part of the sound wave was manually selected for each of the vowels. Following this, distances between each voice and every other voice were calculated using Pythagorean Theorem (as in Baumann & Belin, 2010; Bestelmeyer et al. 2012). The distances were then normalised between genders using the min max transform. Through normalising the distances, all of the distances were adjusted to be on the same scale of between 0 and 1. Voices were then paired according to three levels of difficulty: hard trials, consisting of stimuli that were closest together in terms of distance, medium difficulty trials which were an average distance from one another and easy trials which were the trials that were the greatest distance apart. For each gender, there were 30 hard trials, 30 middle difficulty trials and 12 easy trials from each of the two databases.

Stimuli consisted of VCV or CVC iterations of the words aba/ aga/ ada/ ibi/ igi/ idi/ ubu/ ugu/ udu/ had/ hid/ hod/ hud and hed. The vocal identity matching test consisted of 288 trials, half of which used same identity voice pairs and the other half used different identities. Of these trials, 96 were VCV-VCV pairs, 96 were VCV-CVC pairs and the final 96 were CVC-CVC pairs. Gender was blocked but the order the genders were presented in was randomised. The side of the screen to which the stimuli were assigned was randomised between subjects so that they were sometimes presented in an AB left to right format and other times in a BA left to right format. As well as this within trial randomisation, presentation was also randomised across trials but within gender.

Procedure. Following a brief overview of the task, an instruction screen was presented at the start of the experiment, which read ‘You will see two speaker icons on each trial. Click on each one (in any order) and decide whether the syllable (e.g. "had", "aba") was spoken by the same or two different speakers. You can re-play each voice but don't "overthink" your decision. We recorded speakers on multiple occasions so that a background hiss in one recording does not necessarily mean that the speakers are different! The experiment takes roughly 30 minutes. Take a break any time you like. Thank you for your help! Please press any key to continue.’

Once a key had been pressed, the test screen appeared. Voices were played and responses indicated by means of a mouse click, and a + sign was presented in the centre of the screen for 800ms following response input in order to indicate that the programme had moved on to the next trial. One experimenter was always present during testing.

Data analysis. A signal-detection analysis was performed on the data, in order to determine the response patterns of participants across trials. In a simple discrimination task, there are four possible outcomes to a same/different decision: correct rejections (where the individual correctly identifies the trial as containing two different speakers), false alarms (where the individual says that the voice samples are from a single speaker when they are actually from two different individuals), misses (where the individual says that the voice samples are from two different speakers when they are actually from the same individual) and hits (where the samples are from the same individual and the participant responds correctly). In a signal detection analysis, the proportion of false alarms is compared to the proportion of hits in order to determine the relative contribution of meaningful signal to that of noise in the response patterns. The d-prime statistic incorporates an adjustment for any response bias that might be present and when t-tested, using a one-sample t-test against 0, can determine whether a test is a significantly sensitive measure.

In addition to analyses of the entire dataset in terms of percentage accuracy, the data were also analysed in terms of the demographic information provided by participants concerning age, gender and nationality. Gender was disseminated in terms of both gender of the participant and gender of the voice in a given trial. Test-retest reliability analyses and split-half reliability calculations were also performed on the data.

Results

Performance

There was a large variation in accuracy ranging from 56-89%, with an average overall accuracy of 76% (*SD*=5.52). Figure 4.0 shows the cumulative percentage of response accuracy as well the frequency of responses. This graph enables easy comparison of individual scores against the population. As seen in the graph, the data for percentage accuracies are significantly negatively skewed (skewness -0.88 $p<.05$).

A 2 x 2 mixed ANOVA was run with 2 levels of voice gender and 2 levels of participant gender. Male voices were rated more accurately (78%) than females voices (74%; $F(1,304)=144.79$, $p<.001$, $\eta^2=0.32$) irrespective of participant gender which demonstrated no difference in performance (76% female participants, 75% male participants; $F(1,304)=2.5$, n.s.). Interestingly, there was a significant interaction between participant gender and voice gender ($F(1,304)=9.7$, $p=.002$, $\eta^2=0.031$), demonstrating that females were more accurate at identifying female voices than were males, however both genders performed equally well in the identification of male voices.

Participants performed more accurately on same identity trials (83%) than on different identity trials (70%). There was no significant correlation between age and overall accuracy but there was a significant correlation between age and time taken to complete the test ($r=.35$; $p<.001$), demonstrating that older individuals tended to take slightly longer than

younger people. On average, the test took 33 minutes to complete. There was no significant correlation between time taken and overall accuracy (r = .021, n.s.).

Two hundred and thirty one participants reported their nationality to be British, compared to 75 non-British participants. There was no significant difference in performance between British and non-British individuals ($t(304)=1.15$, n.s.), suggesting that nationality does not impact performance. However, non-British nationals took significantly longer ($M=35.1$ mins; $SD=8.79$) to complete the test than did British individuals ($M=32.7$ mins; $SD=6.28$; $t(304)=-2.591$, $p=.01$).

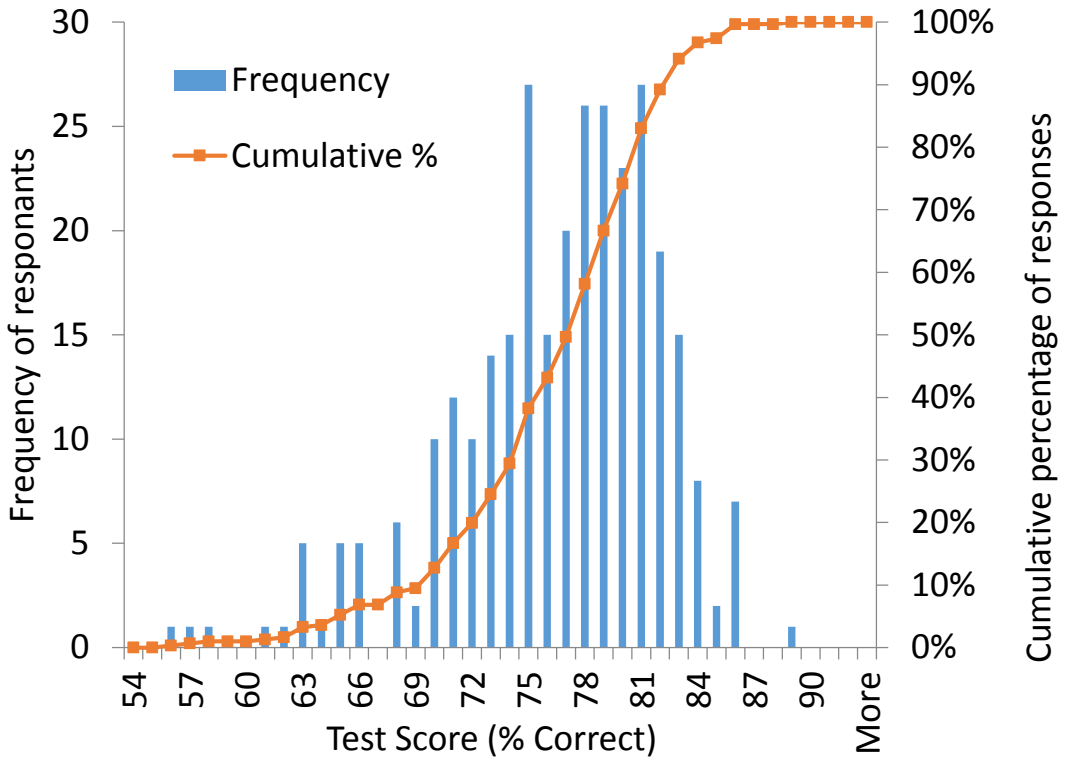


Figure 4.0. A graph to show the frequency of each of the percentage accuracies (blue), as well as the cumulative frequency at each response (red).

Signal Detection Analysis

The mean hit rate (i.e. the responses where the voices were the same and the participants responded that they were the same) for the voice test was 0.83, compared to 0.31 for false alarm trials (the voices are different but participants indicate that they are the same). There

was an average criterion bias of -1.0263, demonstrating an increased likelihood for participants to respond 'same'. However, the average d' across participants was 1.536 and when t-tested against 0, it was found that the voice test was a significantly sensitive measure ($t(305) = 73.702$; $p < .001$), whilst taking into account any bias.

Test-retest reliability

In order to determine whether the test was reliable, thirty participants (11 males, Mean age = 24.27, SD= 5.72) who had previously completed the test were asked to take the test again. Participants were again awarded with either course credits or £6 monetary compensation. A paired samples t-test was run comparing the mean accuracy for test one (M = 77.58%) with that of test 2 (M = 76.78%). It was shown that there was no significant difference between test scores ($t(29) = 1.115$; $p = .274$; see Figure 4.1 A). Furthermore, a bivariate correlation demonstrated that the first test scores were correlated highly significantly with the scores in the second test (r = .643; $p < .001$; see Figure 4.1 B). These results suggest that the test shows stable reliability.

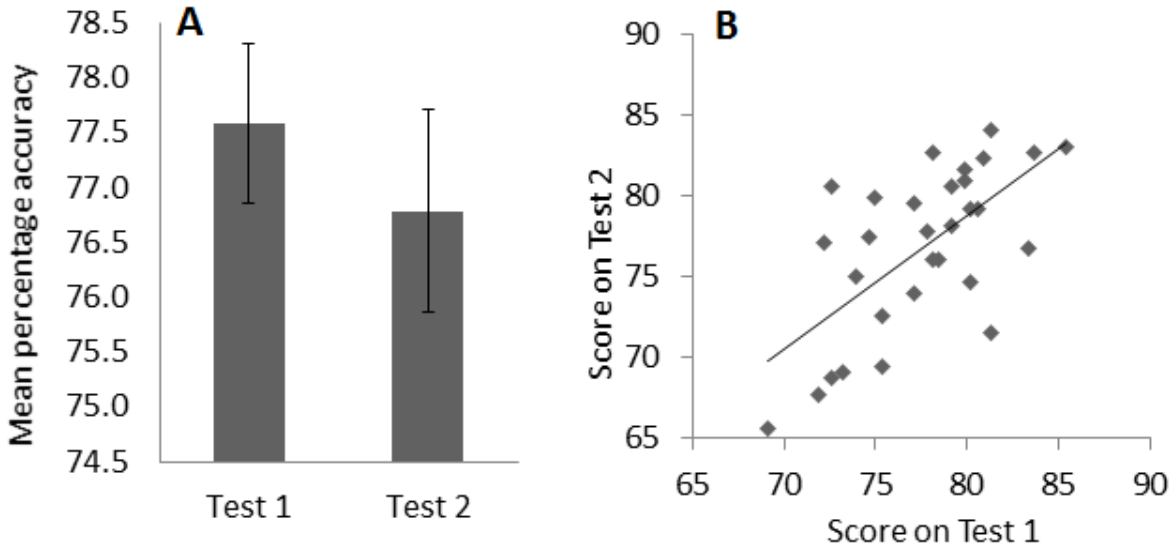


Figure 4. 1. **A:** Mean percentage accuracies for the original test and retest. **B:** A scatter plot to show the relationship between scores on Test 1 and Test 2.

The responses produced a Cronbach's alpha value of 0.849, suggesting good internal consistency. Furthermore, a split half reliability analysis was performed. Stimuli were split into odd an even numbers, enabling an even split of gender and trial difficulties. The correlation between halves was r =.757; p =<.001. In addition to this, the Spearman-Brown correction was 0.862, suggesting that the full length test has relatively strong internal reliability.

Discussion

The aim of the current work was to design, develop and standardise a new test for use in determining individual ability at unfamiliar voice matching. Whereas several tests exist which are designed to measure individual ability in face matching, no such equivalent exists in the auditory domain. A test was designed based on the conceptualisation of a voice space in which dimensions related to source and filter aspects of the vocal signal. Therefore, voices that lie closer together within this space are likely to be perceived as more similar and will subsequently be harder to discriminate between. Using a test with 288 trials, we tested 306 participants on their ability to discriminate between pairs of voices, determining whether they were from the same or different speakers. Average voice matching ability across the sample was shown to be 76%. However, there was a large degree of individual variability, with results ranging from 56-89%. This, along with the significant d' score suggest that this test provides a viable means of assessing individual accuracy at unfamiliar voice matching.

Further research should explore potential factors that affect voice matching ability. The GFMT was run in conjunction with tests for recognition memory, matching of familiar figures and visual short term memory tests. Face matching ability was demonstrated to be highly correlated with performance on face recognition memory tests and matching familiar figures. However, there was no significant correlation between performance accuracy on the

face-matching task and results from the visual short term memory tests. The researchers suggested that the large correlation between the GFMT and figure matching test was as a result of unfamiliar faces being processed more similarly to objects, in the absence of processes that promote similar levels of robust performance demonstrated on familiar face matching task. It would also be interesting to see if the same result is apparent in voice matching. Do we perceive and evaluate voices in a similar manner to auditory 'objects' or do we recruit the same processing strategies that we employ in familiar voice recognition? It would also be interesting to find out whether voice matching performance is related to the ability to accurately identify and remember previously presented voices. This would provide a more in depth knowledge of the interaction between perceptual processes and individual differences in these tasks.

Other studies have looked at several other aspects of processing and their relation to perceptual ability in the domain of face matching. For example, data looking at personality factors and face matching ability have found that females, but not males, demonstrated increased ability at face matching was related to low levels of anxiety and tension and increased emotional stability (Megreya & Bindemann, 2013). Similar gender differences were noted in a study looking at the effects of arousal on face recognition ability, which found that high arousal was associated with poorer recognition for females but not males (Brigham, Maass, Martinez & Whittenberger, 1983). More recent research has failed to find a relationship between extraversion and face matching ability, regardless of participant gender (Lander & Poyarekar, 2015). Other research has looked at facial recognition and its relationship to visual and verbal memory. Woodhead and Baddeley (1981) found that individuals who performed well at facial recognition tasks didn't necessarily perform well on verbal memory tasks, suggesting that visual perceptual skills are not intrinsically related to verbal memory skills.

Research on voice selective brain areas has demonstrated that three areas of the superior temporal sulcus are more sensitive, and indeed selective to voices as opposed to other auditory stimuli such as environmental sounds and human non-vocal sounds (Belin, Zatorre, Lafaill, Ahad & Pike, 2000). It was also found that filtering out low and high frequencies in the vocal signal caused a decrease in activation in these areas coupled with a decrease in subjects' ability to discriminate the sounds as being vocal as opposed to non-vocal and ability at gender identification. The voice test would be a good way to further investigate whether individual accuracy on voice matching is related to voice selective activation in areas of the STS and if so, which of the three areas of the temporal voice area demonstrate the highest degree of correspondence to voice matching ability.

A useful further development of the test would be to establish a shortened version, making it easier to use in conjunction with other measures. The GFMT produced a shortened version of the full length test comprising of the trials with the most errors and the results were highly correlated with the overall accuracies from the full length test. When the shortened version was piloted, mean overall accuracy was shown to be lower than that of the full length test, as a result of the choice of difficult trials. The shortened measure provides a quick and easy method of administering a test of face matching performance and a similar test could easily be devised from the findings of the current standardised voice test. In order to establish the trials for use of a shortened version of the same test, item analysis should be employed. This would enable the construction of a reliable sub-scale through identification of the trials with the most discriminative power. A shortened version of the test would enable the measure to be more easily administered in conjunction with other tests, furthering the tests' research potential.

In addition to the use of the test in research, there are also several real world instances in such a measure could be useful. As discussed in the introduction, administering a test

similar to the one reported here could assist in establishing the credibility of a witness in instances of ear-witness testimony. Similarly, in clinical settings, the diagnosis of phonagnosia often goes un-noticed due to a lack of standardised diagnostic tools designed to pick up on such specific deficits. The current test provides a good grounding from which to develop further tests to assist in these domains.

The results from the reliability analyses suggest that the voice test demonstrates relatively good test-retest and split half reliability. It is anticipated that further usage of the test will provide further scope with regards to the overall reliability of the test but preliminary results look promising. Similarly, the test appears to have a good degree of face validity as it does not incorporate verification techniques which would incorporate aspects of auditory memory. The use of identification procedures ensures a more direct measurement of voice matching ability and where face tests have been criticised on the ability of the participant to simply feature match across images of the individual, such a method is unachievable in the auditory domain due to the fact that stimuli are never presented concurrently. In addition to this, the use of short CVC and VCV phrases reduces reliance upon prosodic vocal features and the pairings of voices from recordings made in the same geographical location diminishes accent matching strategies as a means of differentiation.

The voice test was constructed on the basis that perceptual differences in voices are highly correlated to physical distances in Euclidean space between voices mapped on perceptual dimensions. Therefore, voices that are located close to one another when mapped on given perceptual dimensions are likely to be perceived as more similar than those which are of a greater distance to one another. Previous research has provided support for various acoustic dimensions being highly implicated in identifying individuals from the voice (Baumann & Belin, 2010). These perceptual dimensions provided the basis for the construction of the voice test and it would be interesting to see if these distances in voice

space are indeed related to performance across the sample. This should be established for both the current version of the test and a shortened version, should it be produced in order to confirm the construct reliability of the measure.

In conclusion, we have presented the first standardised measure of voice matching ability. Preliminary results suggest that it has promising test-retest and split-half reliability and is a sensitive means of gauging individual differences in voice matching. It is anticipated that through further development, in the means of a shortened version of the test, an easily administrable tool will be provided for use in both research and clinical settings.

Voice matching ability and vocal identity adaptation

Our preceding adaptation research has demonstrated a large degree of individual variability in adaptation aftereffect size. Previous research has attempted to determine some of the factors that contribute to the individual differences observed in both magnitude of aftereffect (Dennett, McKone, Edwards & Susilo, 2012; Rhodes, Jeffery, Taylor, Hayward & Ewing, 2014; Sussman, 1993) and differences in location of category boundaries (Webster, Kaping, Mizokami & Duhamel, 2004) across participants. For example, Webster and colleagues investigated some of the factors contributing to the large variability in category boundaries observed in relation to face adaptation. When adapted to facial gender, the boundary would shift towards the gender of the participant, suggesting that participants are more attuned to how a face diverges from those within the category to which they belong. Similar differences were observed between Caucasian and Japanese individuals in category boundary changes in response to race adaptation: there was increased sensitivity to within race differentiation. Furthermore, these differences were more pronounced for individuals who had only recently arrived in the United States relative to individuals who had been there for longer than one year suggesting that expertise with other-race faces will increase with

contact with a novel population. These results demonstrate a degree of environmental influence on the coding of various facial properties.

Valentine (1991) theorised that faces are coded in a multidimensional perceptual framework whereby faces are coded in relation to the average face. In this framework, typical faces lie towards the centre of this hypothetical space and distinctive faces lie towards the periphery. Aftereffects in face perception are often theorised to emerge as a result of a shift in the central point of 'face space'. It has previously been proposed that a similar coding framework is used in voice recognition ability (Yovel & Belin, 2013). Research in vocal identity adaptation has demonstrated aftereffects following adaptation to an anti-voice. An anti-voice is a caricature of the average voice in relation to a learned voice, located in the opposing direction of voice-space to that of the learned voice. Adapting to this anti-voice, causes a perceptual shift of the average away from the adapting stimulus, such that the average voice comes to be perceived as less average and the morphs between the average and the learned voice come to be perceived as being more average.

The idea that aftereffects in vocal identity perception arise as a result of a shift in the average voice-space, provide support for a norm-based coding strategy for vocal identity perception (Latinus & Belin, 2011b). Norm-based coding refers to a coding system whereby items are encoded in relation to an average. For example, a stereotypical female voice would be represented at the centre of a multidimensional voice space and all other female voices would be encoded in relation to this average, with more distinct voices being located towards the periphery of voice space. Neuroimaging research has provided support for the norm-based coding of vocal identity in that voices that were further in distance from the gender specific prototypical voice were perceived as more distinctive and were associated with an increased level of activity in the temporal voice area (Latinus, McAleer, Bestelmeyer, & Belin, 2013) an area previously found to be implicated in the acoustic representation of vocal

identity (Belin, Zatorre, Lafaille, Ahad & Pike, 2000). This relates somewhat to the format in which the voice test was generated in that perceived similarity of voices is based upon geometric distances across acoustic dimensions. From this it is possible to theorise that aftereffect sizes in vocal adaptation might have some relationship to voice recognition ability due to their common reliance upon norm-based coding strategies.

In a recent review article, the benefits of studying individual differences in relation to the comprehension of face processing are discussed (Yovel, Wilmer & Duchaine, 2014). One study employing this method of analysis found a positive correlation between face recognition ability, as measured by the CFMT, and magnitude of aftereffects when adapted to varying eye-heights in faces, whereby increased performance on the CFMT resulted in a larger aftereffect (Dennett, McKone, Edwards & Susilo, 2012). Furthermore, these correlations were found to be specific to faces and facial aftereffects as no significant correlations were observed between performance on the CFMT and adaptation to a T shaped manipulation with corresponding spatial properties to that of the eye-height manipulation used in the face adaptation paradigm. Additionally, no correlation was found between eye-height aftereffect size and the Cambridge Car Memory Test (CCMT), suggesting that the relationship is specific to face-matching ability rather than involving a more general object memory component.

Similarly, a recent study has demonstrated positive correlations between performance on the CFMT and aftereffect size following adaptation to face identity (Rhodes, Jeffery, Taylor, Hayward & Ewing, 2014). It was shown that significant correlations were evident between aftereffect size and the d' for own-race but were not significantly correlated with the d' for other-race faces. This result suggests that individuals who show strong adaptation aftereffects are more sensitive to identity differences in own-race as opposed to other-race faces. Additionally, this study again demonstrated that there was no correlation between

aftereffect size and performance on the CCMT. Taken together these results suggest that the better an individual is at facial memory, the larger the apparent aftereffect following face adaptation, irrespective of the individual's ability at object memory and the magnitude of aftereffects when adapted to spatial manipulations. The authors concluded that differences in adaptive coding of facial identity are related to the individual differences observed in face recognition ability. This suggestion is supported by research which demonstrates that adaptation effects are reduced in populations of individuals with face matching difficulties (see e.g. Pellicano, Rhodes & Calder, 2013).

A similar study in the auditory modality looked at differences in adaptation effects and discrimination abilities between children and adults (Sussman, 1993). Using an adaptation paradigm designed to gauge the effects of acoustic differences on the degree of adaptation observed in response to speech sounds, participants were required to complete adaptation conditions when adapted to the CV syllable "da" and tested on "da-ba" continuum and vice versa. In addition to this, participants complete auditory discrimination tasks, using the stimuli from the "ba-da" continuum and were required to indicate whether or not the syllable changed between trials. From this, a discrimination score (d-prime) was computed for each of the participants in the study. Children showed a lesser degree of adaptation than did adults. Furthermore, children performed more poorly on the discrimination task than did adults. As well as this effect being apparent between groups, it was also evident on the individual level whereby children who had the highest discriminability scores also demonstrated the greatest degree of adaptation.

Based on the literature the prediction is that there will be a positive relationship between an individual's ability at voice recognition and the magnitude of the observed aftereffect following vocal identity adaptation such that, the better an individual is at voice matching, the larger the aftereffect will be. The following experiment aims to establish

whether or not the individual differences observed in aftereffect size following adaptation are related to differences in perceptual abilities. In order to test these predictions, a unimodal identity adaptation task was run in conjunction with the voice test.

Method

Participants. Thirty three undergraduate psychology students from Bangor University took part in the study. One participant's data was removed from the analysis due to an accuracy of lower than 80% in the identity recognition pre-test. This left 32 participants in the final analysis (12 males, mean age = 19.68 years, SD = 0.97). Participants were recruited via the online participation panel and were awarded course credits in exchange for participation. Participants were required to be in the second year of their degree and familiar with the identity of two lecturers whose voices were used to generate the stimuli.

Stimuli. Voice recordings of 4 female psychology lecturers were made, saying the syllables used in voice test. Following this, two voices were chosen, which were not too accented to generate the test stimuli. Adaptor stimuli consisted of four different syllables from those used in the voice test, with the exception of both /aba/ and /hod/. These syllables were used for the test identity morphs and therefore did not feature as adaptors. Ambiguous identity vocal morphs were created in the same manner as preceding adaptation experiments, from 5-95% in increments of 15%.

Procedure. There were four different aspects to the study: completion of the voice test, an identity training phase, an identity test phase and an adaptation test phase. First participants completed the voice test, the parameters of which are outlined above. This took on average 35 minutes so participants were given a short break of up to 5 minutes before commencing the second part of the experiment.

In order to establish that the participants were indeed familiar with the identities used for the vocal stimuli, an identity training test was administered. This test consisted of iterations of any of the syllables used in the voice test by either of the two identities used in the adaptation phase. Participants were required to respond following the iteration with a classification judgement as to who they thought the word was spoken by. Auditory feedback was given on responses in terms of a buzzer for incorrect responses, and a ding for correct responses. Each of the iterations was rated once per identity resulting in 32 trials in total.

The identity test phase was structured identically to the previous phase with the exception that auditory feedback was not included and each of the iterations was rated twice per identity resulting in 64 trials in total. Only individuals who scored 80% or higher on this identity test were included in the final analysis.

The adaptation test phase was completed in four blocks: 2 x test syllables (hod and aba) and 2 x adaptor identities. Each trial consisted of four random syllables chosen from the following: /aga/ada/ibi/igi/idi/ubu/ugu/udu/had/hid/hud/hed, spoken by the adapting identity. No syllable was repeated within a trial to avoid low level adaptation to stimulus features. There was a 200ms inter-stimulus-interval between adaptor stimuli. Following this, there was 500ms interval prior to the onset of the ambiguous identity test morph, which participants were then required to categorise as being spoken by one of the two identities included, using the ‘z’ and ‘m’ keys on a standard QWERTY keyboard. Following the response, the succeeding trial would start after a 500ms inter-trial-interval. In the case of no valid response being made, the following trial would commence automatically after 5-6 seconds (randomly jittered). Each of the seven morphed test stimuli was rated four times per block, resulting in 56 trials per adapting identity. Stimuli were presented through Beyerdynamic DT770 80Ω headphones at 75 dB SPL(C).

Results

Data were averaged across participants according to the identity of the adaptor stimulus, and the morph step being rated. A paired samples t-test was used to contrast ratings of the most ambiguous voice morph following adaptation to identity A when compared with that of identity B. Adaptation to identity A resulted in ambiguous identity vocal morphs as being more likely to be categorised as belonging to identity B than was the case when adapted to identity B ($t(31)= 4.756$; $p <.001$), where more of the vocal morphs were rated as belonging to identity A (see Figure 4.2 A). These results replicate the findings of previous studies which have demonstrated contrastive aftereffects in vocal perception following identity adaptation (Zaeske, Schweinberger & Kawahara, 2010). Figure 4.2 B shows individual differences in magnitude of aftereffect following vocal identity adaptation.

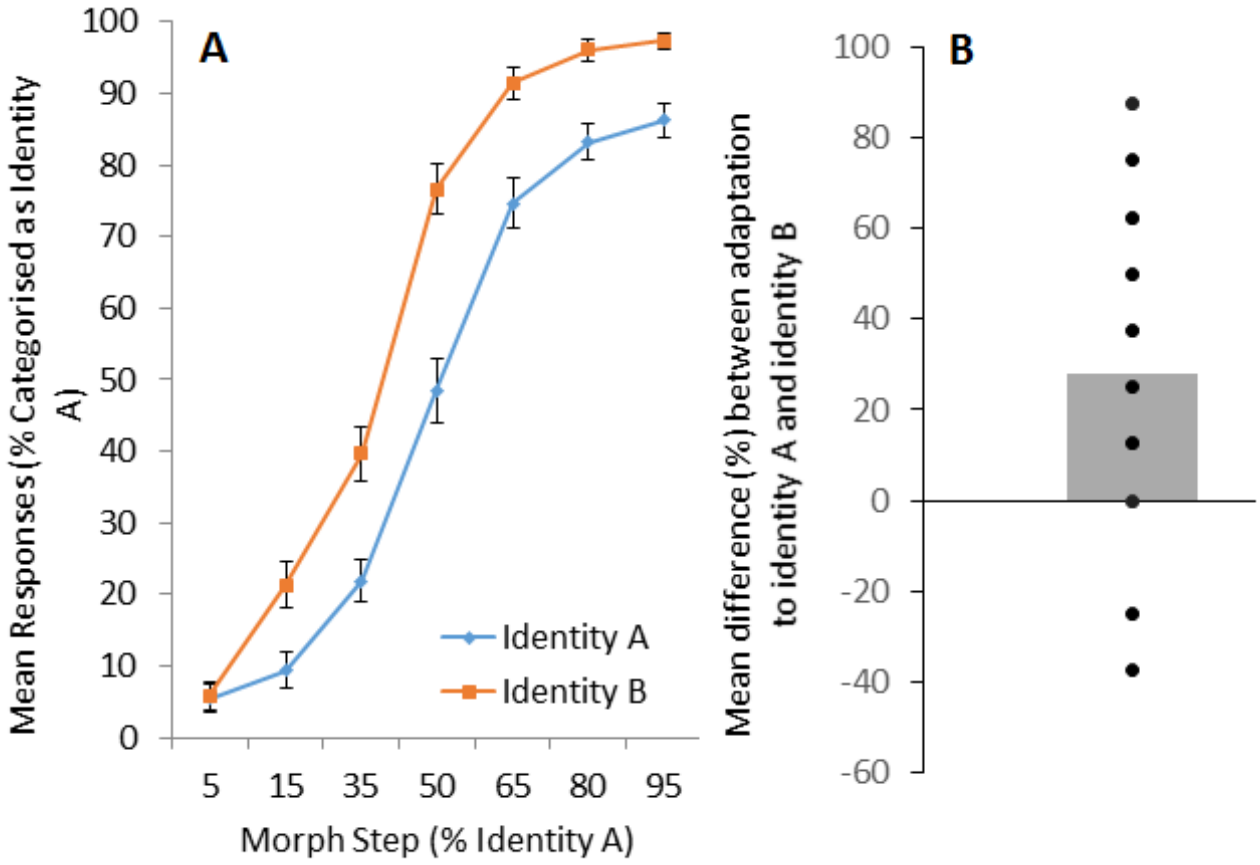


Figure 4.2. **A:** Average responses across participants as a function of morph step. Error bars represent S.E.M. **B:** The bar represents the average percentage difference between classifications of the most ambiguous vocal morphs following adaptation to Identity A and Identity B. Data points represent the difference between identity A and identity B adaptation conditions for each participant. One data point can be representative of more than one individual.

A difference score was computed at the 50% morph to obtain the aftereffect size for the most ambiguous vocal morph. A linear regression was then computed inputting voice test accuracy as the predictor variable and the absolute difference in aftereffect size as the outcome variable. Percentage accuracy on the unfamiliar voice matching test was a marginally significant predictor ($\beta = .329$; $p = .066$) indicating that the higher the accuracy on the voice test, the larger the magnitude of the observed aftereffect (see Figure 4.3). The

overall model fit was $R^2 = 0.108$, indicating that percentage accuracy on the voice matching test accounts for 10.8% of the variance in the absolute difference of the aftereffect size.

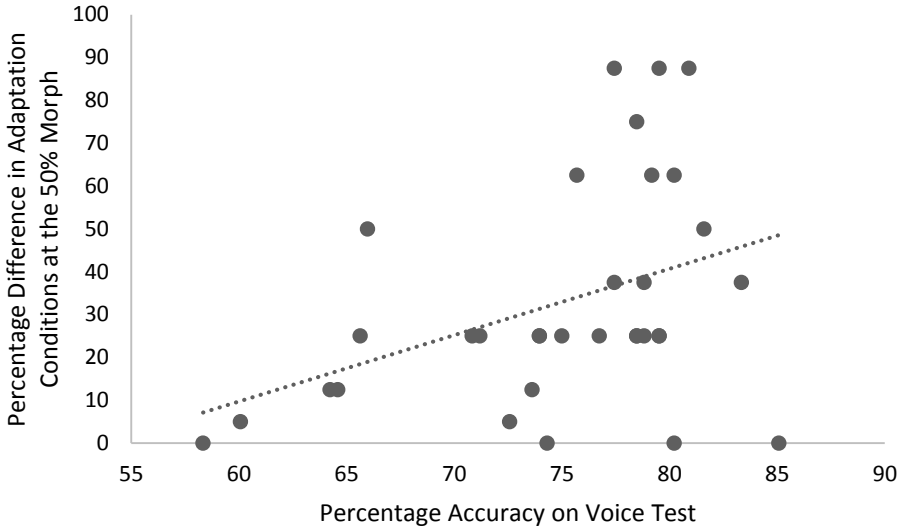


Figure 4.3. A scatter plot showing the relationship between percentage accuracy on the voice test and magnitude of aftereffect for the 50% morph. The dashed line represents the line of best fit.

Discussion

The current study aimed to establish whether or not there was any relationship between perceptual voice matching ability and the size of the aftereffects observed following unimodal adaptation to speaker identity. It was found that there was a significant difference in ratings of ambiguous identity morphs following adaptation to identity A when compared to that of identity B. When adapted to identity A, vocal morphs were more likely to be perceived as belonging to identity B relative to when adapted to identity B. It was also found that the absolute difference in aftereffect size marginally significantly predicted by the percentage accuracy on the voice test, which explained 10.8% of the variance, suggesting that individuals who are better at voice matching are more susceptible to greater shifts in the norm-based voice prototype within voice space following adaptation. The findings from the current research do show similarities to those found by Sussman (1993) who demonstrated

that individuals who performed better at phonetic discrimination in speech, also were the most susceptible to adaptation. In order to clarify whether or not this is a stable effect, it would be necessary to collect data from more participants. Therefore, the current results should be interpreted with caution and further research should attempt to confirm the apparent relationship between voice matching ability and magnitude of identity aftereffect size following voice identity adaptation.

In order to establish other factors that influence individual variability in aftereffect size, it would be useful to devise a test that measured familiar voice matching ability. Although the identity test given to participants here verified their ability to identify individuals used in the adaptation paradigm, a more general measure of familiar voice matching, potentially including famous voices so as to remove some of the bias regarding personal familiarity, would be useful to include in the model. It could be that ability of familiar vocal identity perception would be a better predictor of aftereffect size than would unfamiliar voice matching ability as face matching literature has demonstrated far greater accuracy for familiar compared to unfamiliar faces (Bruce, Henderson, Newman & Burton, 2001). Furthermore, strong correlations between unfamiliar and familiar face matching ability were observed only when familiar faces were inverted, suggesting that unfamiliar faces are potentially processed in a more feature based manner than familiar faces (Megreya & Burton, 2006). Unfamiliar face matching ability has been shown to correlate with visual short term memory, suggesting that performance on a voice matching task might be related to auditory short term memory, something which future research should investigate given the observed parallels in the coding of identity across modality.

Furthermore, it would be interesting to explore the relationship between same-race bias and aftereffect size in order to ascertain whether a similar relationship exists in voice perception as is apparent in face perception whereby significant correlations were observed

between own-race d' and face identity aftereffect size, however, there were no significant correlations between other-race d' scores and aftereffect size (Rhodes et al. 2014). Similarly, gender judgements caused the category boundary to shift towards the gender of the participant following facial gender adaptation (Webster et al. 2004). Exploring the existence of similar gender and race biases, in terms of accent differentiations, in voice perception would shed further light upon the nature of the prototypical voice and the degree to which this representation is susceptible to environmental shaping.

In extension of the present work, research should attempt to uncover the coding strategies underlying different aspects of vocal perception. Storrs and Arnold (2012) provide evidence to suggest that higher level face aftereffects can qualitatively differ in their underlying coding mechanisms. Whereas some aftereffects such as facial distortions appeared to be resultant of a norm-based opponent coding strategy, other factors such as facial gender, appear to adopt a multichannel coding approach. This research highlights the importance of investigating higher level aftereffects in terms of the diversity of the coding strategies which underlie them as it is probable that similar diversity exists in voice aftereffects as has become apparent in face aftereffects. Nevertheless, this study provides support for similar coding of identity across the auditory and visual modalities, a potentially useful framework considering the cross-modal integration strategies incorporated when decoding the identity of a given individual. In addition to this, the present findings suggest that a two dimensional representation of voice space with dimensions corresponding to F0 and F1 is an adequate framework, providing further validity to the voice test.

Conclusion

The aim of the current work was to design and validate an unfamiliar voice matching test and see if such a measure could be used to examine individual variability in aftereffect

size following identity adaptation. The voice matching test was standardised on a sample of 306 individuals and it is anticipated that further usage of the test will assist in generating a more representative standardised sample. The results from the standardisation revealed large individual differences in ability to match unfamiliar voices. In a separate experiment, the voice test was administered in conjunction with an auditory identity adaptation paradigm in order to determine whether unfamiliar voice matching ability had any effect on the size of the aftereffect observed in response to identity adaptation. A linear regression showed voice test performance to be a marginally significant predictor of absolute difference in aftereffect size when comparing adaptation to identity A to that of identity B.

Chapter 5: General discussion

Summary of findings

The thesis addressed three issues regarding the nature of perception of non-speech vocalisations. The first, was whether the mental representation of vocal emotion is specific to human vocal emotion or whether it encompasses emotional vocalisations from different species and other emotive auditory sounds, in this case, affective instrumental bursts. It was found that perceptual changes occurred following adaptation to both human and dog vocalisations however, musical affective bursts did not follow the same pattern of results. The findings of our unimodal adaptation studies suggest that the mental representation dealing with the perception of human vocal emotion, encompasses a representation for a more general, non-species specific representation of vocal emotion. These findings provide support for research that demonstrates selective activation in relation to vocal stimuli compared to other emotive auditory stimuli (Belin et al, 2000; Belin et al 2002) and may suggest that our brains are hard wired to attend to and process information contained within the vocal signal, further reinforcing its salience as an auditory stimulus. However, the significant aftereffects following adaptation to human voice and dog calls is possibly as a result of acoustic similarities between species, as the expression of fearful and angry calls in both human and dog vocalisations follow similar acoustic profiles. The unimodal voice experiment appeared to have aftereffects that were larger in magnitude than that of the dog call adaptation, supporting research that suggests that a higher degree of similarity between adaptor and test stimuli, results in larger aftereffects (e.g. Hills, Elward & Lewis, 2010). In addition to this, neuroimaging research has demonstrated greater activation to human voices in the STS than to animal vocalisations (Fecteau, Armony, Joanette & Belin, 2004).

Our second series of experiments aimed to determine whether or not the aftereffects observed in vocal emotion perception were indicative of a supramodal representation of emotion. Stable cross-modal perceptual aftereffects were achieved through the use of

dynamic adaptor stimuli of silent articulating facially expressed emotions and the use of a combination of block paradigms and top-up adaptation trials. The two cross-modal studies reflecting an aftereffect in vocal emotion perception following adaptation to facial expressions provide support for the existence of a population of neurons that respond generally to human portrayal of emotion, irrespective of the sensory modality in which it is perceived. The aftereffects observed in the cross-modal studies affect only a single morph whereas the unimodal studies appear to produce a more wide spread adaptation effect, suggesting that there might be a less sensitive response or a smaller number of neurons, tuned to the properties of the adapting stimulus in cross-modal paradigms. This finding provides support for a neural representation that is more abstract and higher-level than that observed in the unimodal studies.

In addition to this effect, we aimed to determine the nature of cross-modal relationships between identity and affect in voice perception. Perceptual aftereffects were robust to changes in the identity of the adapting stimulus, suggesting that at the supramodal level of processing, identity representations and affective information are processed in dissociable streams. In relation to the model of voice perception (Belin et al. 2004), the present results provide support for the multi-modal interactions between face and voice, as suggested by the dashed lines between the face and voice at each of the dissociable levels of processing. Furthermore, the model does not depict any relationship between any of the dissociable pathways at the supramodal level of evaluation, which is also supported by the results that suggest that adaptor and test stimulus identity congruency does not mediate the cross-modal aftereffect in vocal emotion perception. These results were found to be dependent upon the facial stimuli being dynamic as opposed to static in nature, further supporting research that suggests there are different perceptual mechanisms involved in the

perception of dynamic and static facial stimuli (Biele & Grabowska, 2006; Rubenstein, 2005).

More generally speaking, higher-level aftereffects involving perceptual attributes of a more abstract nature, suggest that rather than the sensory input being recalibrated, our cognitive, representational mechanisms are subject to a degree of plasticity. Webster (2011) discusses the difficulty involved in disentangling sensory and conceptual adjustments in the field of visual adaptation, but suggests that aftereffects to low-level colour adaptation are very similar in nature to those observed in high-level face adaptation. It is possible that the perception of auditory information works in a similar way, whereby it is adaptable at several different stages of the processing hierarchy as suggested by the results from the unimodal and cross-modal adaptation experiments reported here. It is possible that these 'higher-level' aftereffects represent a degree of temporary change in the location of the average voice in voice space, remapping subsequent stimuli in relation to this transient shift.

Research utilising adaptation has provided several important findings in relation to both low-level and higher-level aspects of processing, in both visual and auditory perception. Firstly, research suggests that several areas of our sensory perceptual streams show changes in response to adaptation, demonstrating that adaptability is potentially a central aspect of all perceptual processing. The similarities in the mechanisms that elicit these perceptual aftereffects have been shown to be similar in the visual adaptation literature in relation to a recalibration of reference being common to both colour and face adaptation (Webster & MacLeod, 2011). Similarly, in the realms of auditory perception, it is possible that both low-level (sensory processing areas and primary auditory cortex) and high-level aftereffects are dependent upon similar mechanisms, both reflecting the fine-tuned recalibration of our normal reference point in relation to the adapting stimulus. Thirdly, findings of higher-level auditory adaptation provides an ecologically relevant account of the ways in which the

auditory system processes environmental sounds. It is likely that our auditory system exhibits a high degree of perceptual plasticity across several stages of the processing hierarchy. This is evidenced not only by behavioural adaptation studies (e.g Schweinberger et al. 2008; Pye & Bestelmeyer, 2015) but also by studies that investigate individuals with various acquired hearing deficits, where the central auditory system exhibits profound changes in terms of both structure and function (for a review see Syka, 2002). Indeed this account complements a model whereby efficiency of coding is paramount to perception (Smith & Lewicki, 2006). Through constant readjustment of the normal representation in relation to the current mean of the auditory signal, our ability to discriminate significant signals from the environment is increased.

The final studies aimed to further uncover the nature of vocal identity perception. A voice matching test was created and standardised in order to observe individual differences in voice matching ability. A wide range of ability was observed across participants, suggesting that this could be a useful tool in both the detection of individuals with deficits of vocal identity perception and in the domain of earwitness testimony as it could potentially act as an index as to how credible a witnesses might be. The test could also be useful in determining any anatomical differences between individuals who perform well and individuals who perform poorly. As previously suggested, it would interesting to find out how perceptual ability at voice matching modulated activity in voice selective areas of the superior temporal sulcus, if indeed it does. The test will also provide a useful index with which to explore the ways in which individuals who perform well on this task differ in their performance on other perceptual tasks in relation to individuals who do not perform as well.

Lastly, a vocal identity adaptation experiment was run in conjunction with the voice test to see if performance on the voice test was related to the magnitude of identity adaptation observed. It was found that absolute aftereffect size correlated with performance accuracy on

the voice test, whereby increased accuracy on the test was associated with a larger magnitude of aftereffect in adaptation.

Unresolved issues and Methodological constraints

Some of the studies failed to elicit the predicted results. For example, in the first experimental chapter, the study using affective instrumental bursts as adaptor stimuli and emotion ambiguous vocal morphs as test stimuli demonstrated a bias towards the adapting stimulus as opposed to no bias or a bias opposing that of the adaptor stimulus. It is possible that these results reflect a type I error and should be interpreted with caution. A null effect would have been predicted in this instance due to the lessened degree of similarity between adaptor and test stimulus relative to that of the voice-voice or dog call-voice experiments, as well as the use of very similar paradigms in the three experiments. An aftereffect in the opposite direction to the preceding experiments was by no means expected and is somewhat hard to account for. However, it is possible that this opposing effect represents underlying differences in the acoustic properties of the adapting stimuli, as the reaction time data do not reflect a facilitation effect at the most ambiguous morph level.

As previously mentioned, it was anticipated that a further cross-modal adaptation study would be run using familiar identities to see if this had a mediating effect on the magnitude of aftereffects observed. This was thought to be a useful progression as it is possible that the brief introductions to the individuals used in the emotion-identity face-voice adaptation study reported here were not sufficient at familiarising participants with the identity of the individual. A stimulus database was collected of lecturers expressing emotional affective bursts bi-modally. However, when these were validated both unimodally (both audio and visual) as well as bi-modally, it was demonstrated that no singular affective bursts produced recognition accuracy of 80% or above in both sensory domains. Therefore, it was concluded that the stimuli were inadequate for use. It would be beneficial to collect

further stimuli in order to achieve appropriate levels of recognition accuracy amongst subjects such that the question of familiarity and the effect of this upon magnitude of aftereffect cross-modally can be further explored.

One question that arises from the use of adaptation paradigms is whether the results represent an instantaneous recalibration of the sensory system or whether they are actually more reflective of perceptual learning, which promotes discrimination ability (Yehezkel, Sagi, Sterkin, Belkin, & Polat, 2010). To ascertain the differences in the nature of aftereffects following adaptation and perceptual learning, research looking at the time scale over which the aftereffects develop and dissipate is useful in assisting in discerning whether results are reflective of a short lived sensory recalibration or a longer lasting learning effect. Research in the visual domain suggests that some adaptation studies do indeed demonstrate an improvement in learning, measured in terms of discrimination ability (e.g. McDermott, Malkoc, Mulligan & Webster, 2010). Moreover, research in perceptual learning has demonstrated that learning through the use of cueing in a Pavlovian conditioning sense, has the potential to bias our subsequent motion perception, demonstrating a profound perceptual effect of experience on current perceptions (Haijang, Saunders, Stone & Backus, 2005). In order to disentangle the contributions of adaptation and perceptual learning from the adaptation studies reported here, it would have been necessary to study the timescale of the development and dissipation of the aftereffect, including classification post adaptation tests within the design of the experiment. This is something that might prove useful in future research in this domain.

The studies reported here show that the use of morphed stimuli can assist in the understanding of voice perception in similar ways to those employed in the research of face perception. Developments of the software used to create the morphs would be beneficial to advancing this area of research. Currently, the process of morphing auditory stimuli requires

the user to provide manual identification of the fundamental frequency and the harmonics which require alignment. Stimuli are then re-synthesised, which can cause issues when the sound files are aperiodic in nature. This issue could be overcome through the development of software that automatised this process to some degree.

Directions for future research

As well as attempting to resolve the issues stated in the preceding section, future research should attempt to clarify the relationship, through replication of the unimodal affective bursts experiment, between adaptation to instrumental bursts and the effect of this upon vocal emotion perception. This would be useful as the results reported here were largely unexpected findings, which cannot be explained through reaction time data analysis. As the study was run by undergraduate students, it cannot be guaranteed that data collection followed a strict protocol that was uniform across participants. Additionally, further investigating other emotive auditory noises as adaptor stimuli, such as environmental sounds could provide a more in depth account of the effects of adaptor-test stimulus congruency on aftereffect size. The use of a larger variety of adaptor stimuli differing in acoustic properties would help to further explain the variance of the effect that is affected by the low-level aspects of the adapting stimulus.

The finding of supramodality should be explored from the opposite direction, using the paradigm utilised in the cross-modal studies reported here, in order to evaluate the contribution of the voice in relation to aftereffects in facial perception. This would be interesting in order to see if these are comparable in magnitude to the results reported here. Furthermore, the relationship between the theoretically dissociable aspects of both vocal and facial processing at the supramodal level should be further explored.

The results of the voice test and the identity adaptation study demonstrate the valuable contribution of the study of individual differences to aspects of perception research. Indeed,

all the adaptation studies reported in this thesis demonstrate large individual variability in aftereffect size and it is possible that the study of group averages causes a large degree of information to be lost from the present data. The current measure provides a means of disambiguating individuals with poor ability at voice matching from individuals who score highly on the test. This will provide a useful platform from which to explore the perceptual and anatomical differences between these groups of people. Furthermore, it can assist in the identification of individuals with deficits in vocal identity perception, potentially making it a useful clinical measure for use in research into phonagnosia.

Research in the face matching literature has attempted to establish some of the factors which relate to the large observed variability in face matching ability. Studies have looked at contributions of personality (e.g. Megreya & Bindermann, 2013), exposure to faces of the same and different races to our own (e.g. Sporer, 2001), general visual perceptual skills (Woodhead & Baddeley, 1981) and level of arousal (Brigham, Maass, Martinez & Whittenberger, 1983) amongst other things. Similar variables could be researched in the auditory domain in order to establish the degree to which some of these factors can explain the individual differences observed in voice matching ability.

Further research should investigate the differential processing mechanisms underlying perception of familiar and unfamiliar voices, in order to establish whether or not these aspects of perception are sub-served by the same processing mechanisms. Whilst it has previously been demonstrated that the learning of voices results in a steepening of the sigmoid curve in the categorisation of vocal morphs, demonstrating increased categorical perception, little is theorised about these apparent differences in perception in relation to the auditory face model of voice perception. Whilst the model focuses on aspects of voice recognition in the relation to the perception of familiar voices, few assumptions are made about how we discriminate between unfamiliar speakers.

There are indeed documented case studies demonstrating dissociative deficits in relation to the processing of familiar and unfamiliar voices (Van Lancker, Cummings, Kreiman & Dobkin, 1988; Van Lancker & Kreiman, 1987), suggesting that these two types of stimuli are potentially processed by dissociable streams, something which is not currently acknowledged by the model of voice perception of Belin et al. (2004). However, our results of a partially significant relationship between unfamiliar voice matching ability and absolute difference in aftereffect size following adaptation to a familiar speaker, suggest that there is a relationship between the underlying mechanisms involved in the evaluation of both familiar and unfamiliar voices. Previous research investigating the categorisation of voice morphs has demonstrated that as learning takes place, the S-shape curve representing the percentage categorisation of each of the stages of the morphed continuum, develops a steeper gradient. This increased gradient is likely to be as a result of category learning. It is possible that this perceptual change is indicative of the generation of person identity nodes (PIN) and previous research might endeavour to establish whether temporal differences in categorical learning are related to the degree of ability at unfamiliar voice matching. This would help to clarify the nature of the learning process involved in making an unfamiliar voice familiar.

Neuroimaging research has observed different activation in response to familiar and unfamiliar voice perception. Areas of the superior temporal pole (TP) in the right hemisphere have been shown to respond to acoustic differences in unfamiliar voices, whereas as voices were learned, the activation in this area decreased (Latinus et al. 2009). This had led to the suggestion that the superior right hemisphere TP is involved in acoustic representations of unfamiliar voices, potentially indicative of a neural correlate for the PIN (Belin, Bestelmeyer, Latinus & Watson, 2011).

Concluding remarks

Taken together, the studies reported in this thesis advance current knowledge of the perceptual mechanisms underlying paralinguistic aspects of the voice. The findings suggest that perceptual plasticity is an important aspect of auditory processing at both lower and higher-level stages of the processing hierarchy. The voice test provides a valuable means of further exploring individual differences in relation to the ways in which we process identity in the human voice. The trend towards a significant relationship between voice matching ability and absolute aftereffect size in an identity adaptation paradigm support the idea that perceptual ability is related to perceptual plasticity. The results from the emotion adaptation studies provide support for the auditory face model of voice perception in that identity and emotion were found to be independent of one another at the supramodal level of the processing hierarchy. Furthermore, the unimodal studies support the notion that neural mechanisms exist that are fine-tuned to processing features of the human voice, independently of any language component.

References

Andics, A., McQueen, J. M., & Petersson, K. M. (2013). Mean-based neural coding of voices. *Neuroimage*, *79*, 351-360.

Andics, A., McQueen, J. M., Petersson, K. M., Gál, V., Rudas, G., & Vidnyánszky, Z. (2010). Neural mechanisms for voice recognition. *Neuroimage*, *52*, 1528-1540.

Banissy, M. J., Sauter, D. A., Ward, J., Warren, J. E., Walsh, V., & Scott, S. K. (2010). Suppressing sensorimotor activity modulates the disctimination of auditory emotions but not speaker identity. *The Journal of Neuroscience*, *30*(41), 13552-13557.

Banno, H., Hata, H., Morise, M., Takahashi, T., Irino, T., & Kawahara, H. (2007). Implementation of realtime STRAIGHT speech manipulation system: Report on its first implementation. *Acoustical science and technology*, *28*(3), 140-146.

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, *70*(3), 614.

Barlow, H. B. (1961). The coding of sensory messages. *Current problems in animal behaviour*, 331-360.

Barlow, H. B., & Hill, R. M. (1963). Evidence for a physiological explanation of the waterfall phenomenon and figural after-effects.

Bartlett, J. C., & Leslie, J. E. (1986). Aging and memory for faces versus single views of faces. *Memory & Cognition*, *14*(5), 371-381.

Baudouin, J. Y., Martin, F., Tiberghien, G., Verlut, I., & Franck, N. (2002). Selective attention to facial emotion and identity in schizophrenia. *Neuropsychologia*, *40*(5), 503-511.

Baumann, O., & Belin, P. (2010). Perceptual scaling of voice identity: common dimensions for different vowels and speakers. *Psychological Research PRPF*,*74*(1), 110-120.

Belin, P., Bestelmeyer, P. E. G., Latinus, M., & Watson, R. (2011). Understanding Voice Perception. *British Journal of Psychology, 102,* 711-725. doi: 10.1111/j.2044-8295.2011.02041.x

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

Belin, P., Fillion-Bilodeau, S., & Gosselin, F. (2008). The Montreal Affective Voices: a validated set of nonverbal affect bursts for research on auditory affective processing. *Behavior research methods*, *40*(2), 531-539.

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

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

Benton, A. L., Sivan, A. B., Hamsher, K. De S., Varney, N. R., & Spreen, O. (1983). *Contributions to neuropsychological assessment: A clinical manual*. Oxford University Press.

Bestelmeyer, P. E. G., Belin, P., & Grosbras, M. H. (2011) Right temporal TMS impairs voice detection. *Current Biology*, *21*(20), R838-R839.

Bestelmeyer, P. E. G., Maurage, P. Rouger, J., Latinus, M., & Belin, P. (2014). Adaptation to vocal expressions reveals mulitstep perception of auditory emotion.*The Journal of Neuroscience*, *34*(24), 8098-8105.

Bestelmeyer, P. E., Belin, P., & Ladd, D. R. (2014). A Neural Marker for Social Bias Toward In-group Accents. *Cerebral Cortex*, bhu282.

Bestelmeyer, P. E., Rouger, J., DeBruine, L. M., & Belin, P. (2010). Auditory adaptation in vocal affect perception. *Cognition*, *117*(2), 217-223.

Biele, C., & Grabowska, A. (2006). Sex differences in perception of emotion intensity in dynamic and static facial expressions. *Experimental Brain Research*, *171*(1), 1-6.

Boersma, P., & Weenink, D. (2010). {P} raat: doing phonetics by computer.

Boller, F., Cole, M., Vrtunski, P. B., Patterson, M., & Kim, Y. (1979). Paralinguistic aspects of auditory comprehension in aphasia. *Brain and language*, *7*(2), 164-174.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial vision*, *10*, 433-436.

Brédart, S., Barsics, C., & Hanley, R. (2009). Recalling semantic information about personally known faces and voices. *European Journal of Cognitive Psychology*, *21*(7), 1013-1021.

Bricker, P. D., & Pruzansky, S. (1966). Effects of stimulus content and duration on talker identification. *The Journal of the Acoustical Society of America*, *40*(6), 1441-1449.

Brigham, J. C., Maass, A., Martinez, D., & Whittenberger, G. (1983). The effect of arousal on facial recognition. *Basic and Applied Social Psychology*, *4*(3), 279-293.

Bruce, V., & Young, A. W. (1986). Understanding face recognition, *British Journal of Psychology*, *77*, 305-327.

Bruce, V., Henderson, Z., Newman, C., & Burton, A. M. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied*, *7*(3), 207.

Bull, R., & Clifford, B. R. (1984). Earwitness voice recognition accuracy. *Eyewitness testimony: Psychological perspectives*, 92-123.

Burton, A. M., Bruce, V., & Johnston, R. A. (1990). Understanding face recognition with an interactive activation model. *British Journal of Psychology*, 81, 361-381.

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow face matching test. *Behavior Research Methods*, *42*(1), 286-291.

Calder, A. J., & Young, A. W. (2005). Understanding the recognition of facial identity and facial expression. *Nature Reviews Neuroscience*, *6*(8), 641-651.

Calvert, G. A., Brammer, M. J., & Iversen, S. D. (1998). Crossmodal identification. *Trends in cognitive sciences*, *2*(7), 247-253.

Campbell, J., & Burke, D. (2009). Evidence that identity-dependent and identity-independent neural populations are recruited in the perception of five basic emotional facial expressions. *Vision research*, *49*(12), 1532-1540.

Campbell, R., Brooks, E., deHaan, E., & Roberts, T. (1996). Dissociating face processing skills: Decisions about lip-read speech, expression and identity. *Quarterly Journal of Experimental Psychology: Section A*, *49*(2), 295-314.

Campbell, R., Landis, T., & Regard, M. (1986). Face recognition and lipreading. *Brain*, *109*(3), 509-521.

Carbonell, J. R., Grignetti, M.C., Stevens, K. N., Williams, C.E., & Woods, B. (1965). Speaker authentication techniques. Final Report Contract No. DA-28-043-AMC-00116 (E) with U. S. Army Electronics Laboratories, Fort Monmouth, New Jersey. Bolt, Beranek & Newman, Inc., Cambridge Mass

Charest, I., Pernet, C. R., Rousselet, G. A., Quiñones, I., Latinus, M., Fillion-Bilodeau, S., ... & Belin, P. (2009). Electrophysiological evidence for an early processing of human voices. *Bmc Neuroscience*, *10*(1), 127.

Clarke, F. R., & Becker, R. W. (1969). Comparison of techniques for discriminating among talkers. *Journal of Speech, Language, and Hearing Research*, *12*(4), 747-761.

Clarke, F. R., Becker, R. W., & Nixon, J. C. (1966). Characteristics that determine speaker recognition (Report ESD-TR-66-638). Hanscom Field, MA: Electronic Systems Division, Air Force Systems Command.

Clifford, B. R., Rathborn, H., & Bull, R. (1981). The effects of delay on voice recognition accuracy. *Law and Human Behavior*, *5*(2-3), 201.

Clifford, C. W., & Rhodes, G. (2005). *Fitting the mind to the world: Adaptation and after-effects in high-level vision* (Vol. 2). Oxford University Press.

Compton, A. J. (1963). Effects of filtering and vocal duration upon the identification of speakers, aurally. *The Journal of the Acoustical Society of America*, *35*(11), 1748-1752.

Cronbach, L. J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302.

Damjanovic, L., & Hanley, J. R. (2007). Recalling episodic and semantic information about famous faces and voices. *Memory & cognition*, *35*(6), 1205-1210.

Davis, J. P. & Valentine, T. (2009). CCTV on trial: Matching video images with the defendant in the dock. *Applied Cognitive Psychology, 23*(4), 482-505.

Deffenbacher, K. A., Cross, J. F., Handkins, R. E., Chance, J. E., Goldstein, A. G., Hammersley, R., & Read, J. D. (1989). Relevance of voice identification research to criteria for evaluating reliability of an identification. *The Journal of psychology*, *123*(2), 109-119.

Dennett, H.W., McKone, E., Edwards, M., & Susilo, T. (2012). Face aftereffects predict individual differences in face recognition ability. *Psychological Science, 23*(11), 1279-1287.

Doddington, G. R. (1985). Speaker recognition—Identifying people by their voices. *Proceedings of the IEEE, 73*(11), 1651-1664.

Doehring, D. G., & Ross, R. W. (1972). Voice recognition by matching to sample. *Journal of psycholinguistic research*, *1*(3), 233-242.

Duchaine, B. C., & Weidenfeld, A. (2003). An evaluation of two commonly used tests of unfamiliar face recognition. *Neuropsychologia, 41*(6), 713-720.

Duchaine, B., & Nakayama, K. (2006). The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia,44*(4), 576-585.

Ekman, P., Friesen, W. V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., ... & Tzavaras, A. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology*, *53*(4), 712.

Ellamil, M., Susskind, J. M., & Anderson, A. K. (2008). Examinations of identity invariance in facial expression adaptation. *Cognitive, Affective, & Behavioral Neuroscience*, *8*(3), 273-281.

Etcoff, N. L. (1984). Perceptual and conceptual organization of facial emotions: Hemispheric differences. *Brain and cognition*, *3*(4), 385-412.

Fant, G. (1960). Acoustic Theory of Speech Production, Mouton.

Farrús, M., Hernando, J., & Ejarque, P. (2007). Jitter and Shimmer Measurements for Speaker Recognition, Eurospeech, Antwerp, Belgium.

Fecteau, S., Armony, J. L., Joanette, Y., & Belin, P. (2004). Is voice processing species-specific in human auditory cortex? An fMRI study. *Neuroimage, 23*(3). 840-848. doi:10.1016/j.neuroimage.2004.09.019

Fellowes, J. M., Remez, R. E., & Rubin, P. E. (1997). Perceiving the sex and identity of a talker without natural vocal timbre. *Perception & Psychophysics*, *59*(6), 839-849.

Feyereisen, P., & Seron, X. (1982). Nonverbal communication and aphasia: A review: I. Comprehension. *Brain and language*, *16*(2), 191-212.

Fitch, W. T. (2000). The evolution of speech: A comparative review. *Trends in Cognitive Sciences, 4*(7), 258-267.

Fox, C. J., & Barton, J. J. S. (2007). What is adapted in face adaptation? The neural representations of expression in the human visual system. *Brain Research, 1127*(1), 80-89. doi: 10.1016/j.brainres.2006.09.104

Fulton, A., & Bartlett, J. C. (1991). Young and old faces in young and old heads: the factor of age in face recognition. *Psychology and aging*, *6*(4), 623.

Gainotti, G., Ferraccioli, M., Marra, C. (2010). The relation between person identity nodes, familiarity judgment and biographical information. Evidence from two patients with right and left anterior temporal atrophy. *Brain Research*, 1307, 11, 103-114, doi.org/10.1016/j.brainres.2009.10.009.

Ganel, T., & Goshen-Gottstein, Y. (2004). Effects of familiarity on the perceptual integrality of the identity and expression of faces: the parallel-route hypothesis revisited. *Journal of Experimental Psychology: Human Perception and Performance, 30*(3), 583.

Garrido, L., Eisner, F., McGettigan, C., Stewart, L., Sauter, D., Hanley, J. R., ... & Duchaine, B. (2009). Developmental phonagnosia: a selective deficit of vocal identity recognition. *Neuropsychologia*, *47*(1), 123-131.

Ghuman, A.S., McDaniel, J. R., & Martin, A. (2010). Face adaptation without a face. *Current Biology*, 20, 32-36. doi: 10.1016/j.cub.2009.10.077

Gilbert, J. L., Tamati, T. N., Pisoni, D. B. (2013). Development, reliability and validity of PRESTO: a new high-variability sentence recognition test. *Journal of the American Academy of Audiology, 24*(1), 26-36.

Giraud, A. L., Kell, C., Thierfelder, C., Sterzer, P., Russ, M. O., Preibisch, C., & Kleinschmidt, A. (2004). Contributions of sensory input, auditory search and verbal comprehension to cortical activity during speech processing. *Cerebral Cortex, 14*(3), 247-255.

Grill-Spector, K. (2006). Selectivity of adaptation in single units: implications for FMRI experiments. *Neuron, 49*(2), 170-171.

Grill-Spector, K., Kushnir, T., Edelman, S., Avidan, G., Itzchak, Y., & Malach, R. (1999). Differential processing of objects under various viewing conditions in the human lateral occipital complex. *Neuron, 24*(1), 187-203.

Grill-Spector, K., Kushnir, T.,Edelman, S., Avidan, G., Itzchak, Y., & Malach, R. (1999). Differential processing of objects under various viewing conditions in the human lateral occipital complex. *Neuron, 24*, 187-203.

Haijiang, Q., Saunders, J. A., Stone, R. W., & Backus, B. T. (2006). Demonstration of cue recruitment: Change in visual appearance by means of Pavlovian conditioning. *Proceedings of the National Academy of Sciences of the United States of America, 103*(2), 483-488.

Hailstone, J. C., Crutch, S. J., Vestergaard, M. D., Patterson, R. D., & Warren, J. D. (2010). Progressive associative phonagnosia: a neuropsychological analysis. *Neuropsychologia*, *48*(4), 1104-1114.

Hanley, J. R., & Turner, J. M. (2000). Why are familiar-only experiences more frequent for voices than for faces? *The Quarterly Journal of Experimental Psychology: Section A, 53*(4), 1105-1116.

Hanley, J. R., Smith, S. T., & Hadfield, J. (1998). I recognise you but I can't place you: An investigation of familiar-only experiences during tests of voice and face recognition. *The Quarterly Journal of Experimental Psychology: Section A, 51*(1), 179-195.

Hardcastle, W. J. (1976). *Physiology of speech production.* London: Academic Press.

Hargreaves, D. J., & Colman, A. M. (1981). The dimensions of aesthetic reactions to music. *Psychology of Music,* 9, 15-220.

Hasselmo, M. E., Rolls, E. T., & Baylis, G. C. (1989). The role of expression and identity in the face-selective responses of neurons in the temporal visual cortex of the monkey. *Behavioural brain research, 32*(3), 203-218.

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in cognitive sciences*, *4*(6), 223-233.

Henderson, Z., Bruce, V., & Burton, A. M. (2001). Matching the faces of robbers captured on video. *Applied Cognitive Psychology, 15*(4), 445-464.

Herlitz, A., & Lovén, J. (2013). Sex differences and the own-gender bias in face recognition: A meta-analytic review. *Visual Cognition*, *21*(9-10), 1306-1336.

Hickok, G., & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences, 4*(1), 131-138, doi:10.1016/S1364-6613(00)01463-7

Hills, P. J., Elward, R. L., & Lewis, M. B. (2010). Cross-modal face identity aftereffects and their relation to priming. *Journal of Experimental Psychology: Human Perception and Performance, 36*(4), 876.

Imaizumi, S., Mori, K., Kiritani, S., Kawashima, R., Sugiura, M., Fukuda, H., ... & Nakamura, K. (1997). Vocal identification of speaker and emotion activates differerent brain regions. *Neuroreport*, *8*(12), 2809-2812.

Jack, R. E., Garrod, O. G., Yu, H., Caldara, R., & Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, *109*(19), 7241-7244.

Javadi, A. H., & Wee, N. (2012). Cross-category adaptation: Objects produce gender adaptation in the perception of faces. *PLoS ONE, 7*(9): e46079.

Jordan, H., Johnson, M., & Fallah, M. (2008). Dual perceptual adaptation in human faces: Gender and age. *Journal of Vision*, *8*(6), 1140. doi: 10.1167/8.6.1140

Kamachi, M., Hill, H., Lander, K., & Vatikiotis-Bateson, E. (2003). Putting the face to the voice: Matching identity across modality. *Current Biology, 13,* 1709-1714.

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., & Banno, H. (2008, March). TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on* (pp. 3933-3936). IEEE.

Kemp, R., Towell, N., & Pike, G. (1997). When seeing should not be believing: Photographs, credit cards and fraud. *Applied Cognitive Psychology*, *11*(3), 211-222.

King, A. J., & Palmer, A. R. (1985). Integration of visual and auditory information in bimodal neurones in the guinea-pig superior colliculus. *Experimental Brain Research*, *60*(3), 492-500.

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

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception, 36,* 14-14.

Kovacs. G., Zimmer, M., Banko, E., Harza, U., Antal, A., & Vidnyanszky, Z. (2006). Electrophysiological correlates of visual adaptation to faces and body parts in humans. *Cerebral Cortex,15*(5), 742-753.

Kurucz, J., & Feldmar, G. (1979). Prosopo-affective agnosia as a symptom of cerebral organic disease. *Journal of the American Geriatrics Society*, *27*(5), 225-230.

Ladd, D. R., Silverman, K. E., Tolkmitt, F., Bergmann, G., & Scherer, K. R. (1985). Evidence for the independent function of intonation contour type, voice quality, and

F0 range in signaling speaker affect. *The Journal of the Acoustical Society of America, 78*(2), 435-444.

Lander, K., & Poyarekar, S. (2015). Famous face recognition, face matching, and extraversion. *The Quarterly Journal of Experimental Psychology*, (ahead-of-print), 1-8.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H., Hawk, S. T., & van Knippenberg, A. (2010). Presentation and validation of the Radboud Faces Database. *Cognition and Emotion*, *24*(8), 1377-1388.

Latinus, M., & Belin, P. (2011a). Anti-voice adaptation suggests prototype-based coding of voice identity. *Frontiers in Psychology, 2*, 175. doi:10.3389/fpsyg.2011.00175

Latinus, M., & Belin, P. (2011b). Human voice perception. *Current Biology, 21*(4), R143-R145.

Latinus, M., & Belin, P. (2012). Perceptual auditory aftereffects on voice identity using brief vowel stimuli. *PloS one, 7*(7), e41384.

Latinus, M., Crabbe, F., & Belin, P. (2009). fMRI investigations of voice identity perception. *Neuroimage*, *47*(1), 156

Latinus, M., McAleer, P., Bestelemeyer, P. E. G., & Belin, P. (2013). Norm-based coding of voice identity in human auditory cortex. *Current Biology*, *23*(12), 1075-1080. Doi: 10.1016/j.cub.2013.04.055

Latinus, M., McAleer, P., Bestelmeyer, P. E. G., & Belin, P. (2013). Norm-based coding of voice identity in human auditory cortex. Current Biology, 23(12), 1075-1080.

Laufer, A., & Condax, I. D. (1981). The function of the epiglottis in speech.*Language and Speech*, *24*(1), 39-62.

Laukka, P. (2005). Categorical perception of vocal emotion expressions. *Emotion,* 5, 277-295.

Le Gal, P. M., & Bruce, V. (2002). Evaluating the independence of sex and expression in judgments of faces. *Perception & Psychophysics, 64*(2), 230-243.

Legge, G. E., Grosmann, C., & Pieper, C. M. (1984). Learning unfamiliar voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition,10*(2), 298.

Leopold, D. A., O'Toole, A. J., Vetter, T., & Blanz, V. (2001). Prototype-referenced shape encoding revealed by high-level aftereffects. *Nature neuroscience, 4*(1), 89-94.

Levy, D. A., Granot, R., & Bentin, S. (2001). Processing specificity for human voice stimuli: electrophysiological evidence. *Cognitive Neuroscience and Neuropsychology, 12*(12), 2653-2657.

Levy, D. A., Granot, R., & Bentin, S. (2003). Neural sensitivity to human voices: ERP evidence of task and attentional influences. *Psychophysiology, 40*, 291-305.

Lewin, C., & Herlitz, A. (2002). Sex differences in face recognition—Women's faces make the difference. *Brain and cognition*, *50*(1), 121-128.

Lewis, J. W., Talkington, W. J., Walker, N. A., Spirou, G. A., Jajosky, A., Frum, C., & Brefczynski-Lewis, J. A. (2009). Human cortical organization for processing vocalizations indicates representation of harmonic structure as a signal attribute. *The Journal of Neuroscience*, *29*(7), 2283-2296.

Li, L., Miller, E. K., & Desimone, R. (1993). The representation of stimulus familiarity in anterior inferior temporal cortex. *Journal of neurophysiology*, *69*(6), 1918-1929.

Little, A. C., Feinberg, D. R., DeBruine, L. M., & Jones, B. C. (2013). Adaptation to Faces and Voices Unimodal, Cross-Modal, and Sex-Specific Effects. *Psychological science*, *24*(11), 2297-2305.

Logeswaran, N., & Bhattacharya, J. (2009). Crossmodal transfer of emotion by music. *Neuroscience letters*, *455*(2), 129-133.

Malmo, R. B, (1975). On emotions, needs, and our archaic brain. New York: Holt, Rinehart, & Winston.

Mann, V. A., Diamond, R., & Carey, S. (1979). Development of voice recognition: Parallels with face recognition. *Journal of experimental child psychology*, *27*(1), 153-165.

Matsumoto, H., Hiki, S., Sone, T., & Nimura, T. (1973). Multidimensional representation of personal quality of vowels and its acoustical correlates. *Audio and Electroacoustics, IEEE Transactions on*, *21*(5), 428-436.

McDermott, K. C., Malkoc, G., Mulligan, J. B., & Webster, M. A. (2010). Adaptation and visual salience. *Journal of vision*, *10*(13), 17.

McGehee, F. (1937). The reliability of the identification of the human voice. *The Journal of General Psychology*, *17*(2), 249-271.

McKelvie, S. J., Standing, L., Jean, D. S., & Law, J. (1993). Gender differences in recognition memory for faces and cars: Evidence for the interest hypothesis. *Bulletin of the Psychonomic Society*, *31*(5), 447-448.

Megreya, A. M., & Bindemann, M. (2013). Individual differences in personality and face identification. *Journal of Cognitive Psychology*, *25*(1), 30-37.

Megreya, A. M., & Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory & Cognition*, *34*(4), 865-876.

Megreya, A. M., & Burton, A. M. (2008). Matching faces to photographs: poor performance in eyewitness memory (without the memory). *Journal of Experimental Psychology: Applied*, *14*(4), 364.

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. Psychology, Public Policy & Law, 7(1), 3-35.

Moore, B. C. J. (2003). An introduction to the psychology of hearing (6th ed.). Bingley: Emerald Group Publishing Limited.

Nahum, M., Nelken, I., & Ahissar, M. (2008). Low-level information and high-level perception: the case of speech in noise. *PLoS Biol*, *6*(5), e126.

Neuner, F., & Schweinberger, S. R. (2000). Neuropsychological impairments in the recognition of faces, voices, and personal names. *Brain and cognition*,*44*(3), 342-366.

O'Neil, S. F., & Webster, M. A. (2011). Adaptation and the perception of facial age. *Visual Cognition*, *19*(4), 534-550. doi: 10.1080/13506285.2011.561262

Paquette, S., Peretz, I., & Belin, P. (2013). The "Musical Emotional Bursts": a validated set of musical affect bursts to investigate auditory affective processing. *Frontiers in psychology*, *4*.

Pell, P. J., & Richards, A. (2013). Overlapping facial expression representations are identity-dependent. *Vision research*, *79*, 1-7.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial vision*, *10*(4), 437-442.

Pellicano, E., Rhodes, G., & Calder, A. J. (2013). Reduced gaze aftereffects are related to difficulties in categorizing gaze direction in children with autism. *Neuropsychologica*, *51*, 1504-1509.

Pittman, J. (1994). Voice in social interaction: An interdisciplinary approach. California: Sage Publications.

Pye, A., & Bestelmeyer, P. E. (2015). Evidence for a supra-modal representation of emotion from cross-modal adaptation. *Cognition*, *134*, 245-251.

Rehnman, J., & Herlitz, A. (2006). Higher face recognition ability in girls: Magnified by own-sex and own-ethnicity bias. *Memory*, *14*(3), 289-296.

Rhodes, G., Jeffery, L., Evangelista, E., Ewing, L., Peters, M., & Taylor, L. (2011). Enhanced attention amplifies face adaptation. *Vision research, 51*(16), 1811-1819.

Rhodes, G., Jeffery, L., Taylor, L., Hayward, W. G., & Ewing, L. (2014). Individual differences in adaptive coding of face identity are linked to individual differences in face recognition ability. *Journal of Experimental Psychology: Human Perception and Performance, 40*(3), 897.

Rolls, E. T. (2000). The orbitofrontal cortex and reward. *Cerebral Cortex, 10*(3), 284-294.

Rubenstein, A. J. (2005). Variation in perceived attractiveness differences between dynamic and static faces. *Psychological Science, 16*(10), 759-762.

Saslove, H., & Yarmey, A. D. (1980). Long-term auditory memory: Speaker identification. *Journal of Applied Psychology, 65*(1), 111.

Sawamura, H., Georgieva, S., Vogels, R., Vanduffel, W., & Orban, G. A. (2005). Using functional magnetic resonance imaging to assess adaptation and size invariance of shape processing by humans and monkeys. *The Journal of neuroscience, 25*(17), 4294-4306.

Scherer, K. R. (1979). Nonlinguistic vocal indicators of emotion and psychopathology. In C. E. Izard (Ed.), Emotions in personality and psychopathology (pp. 493-529). New York: Plenum Press

Scherer, K. R. (1984a). Emotion as a multicomponent process: A model and some cross-cultural data. In P. Shaver (Ed.), Review of personality and social psychology: Vol. 5. Emotions, relationships and health (pp. 37-63). Beverly Hills, CA: Sage.

Scherer, K. R. (1984b). On the nature and function of emotion: A component process approach. In K. R. Scherer & P. Ekman (Eds.), Approaches to emotion (pp. 293-317). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Scherer, K. R. (1986). Vocal affect expression: a review and a model for future research. *Psychological bulletin*, *99*(2), 143.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech communication*, *40*(1), 227-256.

Scherer, K. R., & Oshinky, J. S. (1977). Cue utilization in emotion attribution from auditory stimuli. *Motivation and Emotion, 1*, 331-346.

Scherer, K. R., Johnstone, T., & Klasmeyer, G. (2003). Vocal expression of emotion. *Handbook of affective sciences*, 433-456.

Scherer, K. R., Ladd, D. R., & Silverman, K. E. (1984). Vocal cues to speaker affect: Testing two models. *The Journal of the Acoustical Society of America*,*76*(5), 1346-1356.

Schirmer, A., & Kotz, S. A. (2006). Beyond the right hemisphere: Brain mechanisms mediating vocal emotion processing. *Trends in Cognitive Sciences, 10*, 24-30.

Schweinberger, S. R., & Soukup, G. R. (1998). Asymmetric relationships among perceptions of facial identity, emotion, and facial speech. *Journal of Experimental Psychology: Human perception and performance*, *24*(6), 1748.

Schweinberger, S. R., Burton, A. M., & Kelly, S. W. (1999). Asymmetric dependencies in perceiving identity and emotion: Experiments with morphed faces. *Perception & Psychophysics*, *61*(6), 1102-1115.

Schweinberger, S. R., Casper, C., Hauthal, N., Kaufmann, J. M., Kawahara, H., Kloth, N., . . . Zaske, R. (2008). Auditory adaptation in voice perception. *Current Biology, 18*(9), 684-688.

Scott, S. K., & Johnsrude, I. S. (2003). The neuroanatomical and functional organization of speech perception. *Trends in Neuroscience, 26*, 100-107.

Sharp, D. J., Scott, S. K., & Wise, R. J. (2004). Retrieving meaning after temporal lobe infarction: the role of the basal language area. *Annals of neurology*, *56*(6), 836-846.

Singh, S., & Murry, T. (1978). Multidimensional classification of normal voice qualities. *The Journal of the Acoustical Society of America, 64*(1), 81-87.

Skinner, A. L., & Benton, C. P. (2012). The expressions of strangers: Our identity-independent representation of facial expression. *Journal of vision,12*(2), 12.

Skuk, V. G., & Schweinberger, S. R. (2013). Adaptation aftereffects in vocal emotion perception elicited by expressive faces and voices.

Smith, E. C., & Lewicki, M. S. (2006). Efficient auditory coding. *Nature,439*(7079), 978-982.

Sporer, S. L. (2001). Recognizing faces of other ethnic groups: An integration of theories. *Psychology, Public Policy, and Law, 7*(1), 36.

Spreckelmeyer, K. N., Kutas, M., Urbach, T., Altenmüller, E., & Münte, T. F. (2009). Neural processing of vocal emotion and identity. *Brain and Cognition, 69*(1), 121-126.

Sprengelmeyer, R., Young, A. W., Schroeder, U., Grossenbacher, P. G., Federlein, J., Buttner, T., & Przuntek, H. (1999). Knowing no fear. *Proceedings of the Royal Society of London B: Biological Sciences*, *266*(1437), 2451-2456.

Stein, B. E., & Wallace, M. T. (1996). Comparisons of cross-modality integration in midbrain and cortex. *Progress in brain research*, *112*, 289-299.

Steklis, H. D. (1984). Primate communication, comparative neurology and the origin of language re-examined. *Journal of Human Evolution, 14*, 157-173.

Stevenage, S. V., Clarke, G., & McNeill, A. (2012). The “other-accent” effect in voice recognition. *Journal of Cognitive Psychology, 24*(6), 647-653.

Stevenage, S. V., Neil, G. J., & Hamlin, I. (2014). When the face fits: Recognition of celebrities from matching and mismatching faces and voices.*Memory*, *22*(3), 284-294.

Stevens, K. N., Williams, C. E., Carbonell, J. R., & Woods, B. (1968). Speaker authentication and identification: a comparison of spectrographic and auditory presentations of speech material. *The Journal of the Acoustical Society of America*, *44*(6), 1596-1607.

Storrs, K. R., & Arnold, D. H. (2012). Not all face aftereffects are equal. *Vision Research,* 64, 7-16.

Storrs, K. R., & Arnold, D. H. (2012). Not all face aftereffects are equal. *Vision research*, *64*, 7-16.

Sussman, J. E. (1993). Auditory Processing in Children's Speech PerceptionResults of Selective Adaptation and Discrimination Tasks. *Journal of Speech, Language, and Hearing Research*, *36*(2), 380-395.

Syka, J. (2002). Plastic changes in the central auditory system after hearing loss, restoration of function, and during learning. *Physiological Reviews*, *82*(3), 601-636.

Tamati, T. N., Gilbert, J. L., & Pisoni, D. B. (2013). Some factors underlying individual differences in speech recognition on PRESTO: A first report. *Journal of the American Academy of Audiology*, *24*(7), 616-634.

Thompson, C. P. (1985). Voice identification: Speaker identifiability and a correction of the record regarding sex effects. *Human Learning: Journal of Practical Research & Applications*.

Titze, I. R. (1989). On the relation between subglottal pressure and fundamental frequency in phonation. *Acoustical Society of America,* 85, 901-906.

Titze, I. R. (2008). The human instrument. *Scientific American*, *298*(1), 94-101.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, *43*(2), 161-204, doi: 10.1080/14640749108400966

Van Lancker, D. R., & Canter, G. J. (1982). Impairment of voice and face recognition in patients with hemispheric damage. *Brain and Cognition, 1*, 185-195.

Van Lancker, D. R., Cummings, J. L., Kreiman, J., & Dobkin, B. H. (1988). Phonoagnosia: A dissociation between familiar and unfamiliar voices. *Cortex, 24*, 195-209.

Van Lancker, D., & Kreiman, J. (1987). Voice discrimination and recognition are separate abilities. *Neuropsychologia*, *25*(5), 829-834.

Van Lancker, D., Kreiman, J., & Emmorey, K. (1985). Familiar voice recognition: Patterns and parameters. Part I: Recognition of backward voices.*Journal of phonetics*, *13*(1), 19-38.

Vida, M. D., & Mondloch, C. J. (2009). Children's representations of facial expression and identity: Identity-contingent expression aftereffects. *Journal of experimental child psychology*, *104*(3), 326-345.

Warrington, E. K. (1984) Recognition Memory Test. Windosr, England; NFER-Nelson.

Watson, R., Latinus, M., Noguchi, T., Garrod, O., Crabbe, F., & Belin, P. (2014). Crossmodal adaptation in right posterior temporal sulcus during face-voice emotional integration. *The Journal of Neuroscience*, *34*(20), 6813-6821. doi: 10.1523/JNEUROSCI.4478-13.2014

Watson, R., Latinus, M., Noguchi, T., Garrod, O., Crabbe, F., & Belin, P. (2014). Crossmodal adaptation in right posterior superior temporal sulcus during face–voice emotional integration. *The Journal of Neuroscience*, *34*(20), 6813-6821.

Webb, A. R., Heller, H. T., Benson, C. B., & Lahav, A. Mother's voice and heartbeat sounds elicit auditory plasticity in the human brain before full gestation. *Proceedings of the National Academy of Sciences of the United States of America, 112*(10), 3152-3157. doi: 10.1073/pnas.1414924112

Webster, M. A. (2011). Adaptation and visual coding. *Journal of vision*, *11*(5), 3.

Webster, M. A., & MacLeod, D. I. A. (2011). Visual adaptation and face perception. *Philosophical Transactions of the Royal Society,* 366, 1702-1725. doi: 10.1098/rstb.2010.0360

Webster, M. A., & MacLin, O. H. (1999). Figural aftereffects in the perception of faces. *Psychonomic Bulletin & Review, 6*(4), 647-653.

Webster, M. A., Kaping, D., Mizokami, Y., & Duhamel, P. (2004). Adaptation to natural facial categories. *Nature, 428*(6982), 557-561.

White, D., Kemp, R. I., Jenkins, R., Matherson, M. & Burton, M. (2014). Passport officers errors in face matching. *PLoS ONE 9*(8): e103510 doi:10.137/journal.pone.0103510

Whittle, S., Yücel, M., Yap, M. B., & Allen, N. B. (2011). Sex differences in the neural correlates of emotion: evidence from neuroimaging. *Biological psychology*, *87*(3), 319-333.

Wiese, H., Komes, J., & Schweinberger, S. R. (2013). Ageing faces in ageing minds: A review on the own-age bias in face recognition. *Visual Cognition*,*21*(9-10), 1337-1363.

Wildgruber, D., Riecker, A., Hertrich, I., Erb, M., Grodd, W., Ethofer, T., & Ackermann, H. (2005). Identification of emotional intonation evaluated by fMRI. *Neuroimage, 24*, 1233-1241.

Williams, C. E. (1964). *The effects of selected factors on the aural identification of speakers. Sect 3*. Rept ESD-TDR-65-153, electronics systems division, air force systems command.

Winsel, R. (1966). *The anatomy of voice: An illustrated manual of vocal training*. Hudson House.

Winston, J. S., Henson, R. N. A., Fine-Goulden, M. R., & Dolan, R. J. (2004). fMRI-adaptation reveals dissociable neural representations of identity and expression in face perception. *Journal of neurophysiology*, *92*(3), 1830-1839.

Winston, J. S., Vuilleumier, P., & Dolan, R. J. (2003). Effects of low-spatial frequency components of fearful faces on fusiform cortex activity. *Current Biology*, *13*(20), 1824-1829.

Woodhead, M. M., & Baddeley, A. D. (1981). Individual differences and memory for faces, pictures, and words. *Memory & Cognition*, *9*(4), 368-370.

Wright, D. B., & Sladden, B. (2003). An own gender bias and the importance of hair in face recognition. *Acta psychologica*, *114*(1), 101-114.

Wright, W. D. (1934). The measurement and analysis of colour adaptation phenomena. *Proceedings of the Royal Society of London. Series B, Containing Papers of a Biological Character*, 49-87.

Xu, H., Dayan, P., Lipkin, R. M., & Qian, N. (2008). Adaptation across the cortical hierarchy: Low-level curve adaptation affects high-level facial-expression judgments. *Journal of Neuroscience*, *28*(13), 3374-3383. doi: 10.1523/jneurosci.0182-08.2008

Yang, H., Shen, J., Chen, J., & Fang, F. (2011). Face adaptation improves gender discrimination. *Vision Research*, *51*(1), 105-110. doi: 10.1016/j.visres.2010.10.006

Yarmey, A. D. (1991a). Descriptions of distinctive and non-distinctive voices over time. *Journal of the Forensic Science Society*, *31*(4), 421-428.

Yarmey, A. D. (1991b). Voice Identification Over the Telephone1. *Journal of Applied Social Psychology*, *21*(22), 1868-1876.

Yehezkel, O., Sagi, D., Sterkin, A., Belkin, M., & Polat, U. (2010). Learning to adapt: Dynamics of readaptation to geometrical distortions. *Vision research*,*50*(16), 1550-1558.

Young, A. (1998). *Face and mind*. Oxford University Press.

Young, A. W., & Bruce, V. (2011). Understanding person perception. *British Journal of Psychology*, *102*(4), 959-974.

Young, A. W., McWeeny, K. H., Hay, D. C., & Ellis, A. W. (1986). Matching familiar and unfamiliar faces on identity and expression. *Psychological research*, *48*(2), 63-68.

Young, A. W., Newcombe, F., Haan, E. H. D., Small, M., & Hay, D. C. (1993). Face perception after brain injury. *Brain*, *116*(4), 941-959.

Yovel, G., & Belin, P. (2013). A unified coding strategy for processing faces and voices. *Trends in Cognitive Sciences*, *17*(6), 263–271. doi:10.1016/j.tics.2013.04.004

Yovel, G., Wilmer, J. B., & Duchaine, B (2014). What can individual differences reveal about face processing? *Frontiers in Human Neuroscience,* 8, 562.

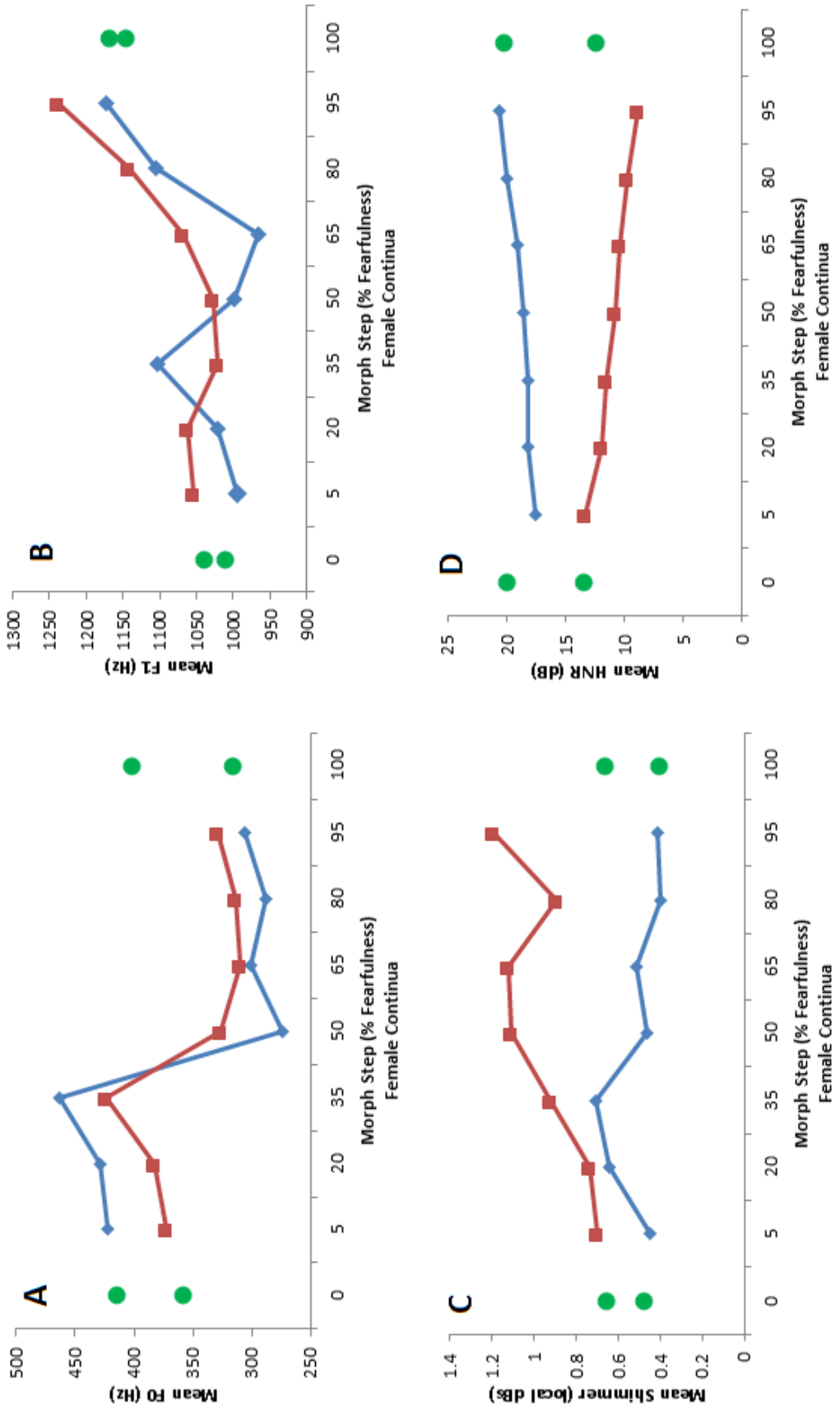
Yumoto, E., & Gould, W. J. (1982). Harmonics-to-noise ratio as an index of the degree of hoarsness. *Journal of the Acoustical Society of America, 71*(6), 1544-1550.

Zäske, R., & Schweinberger, S. R. (2011). You are only as old as you sound: Auditory aftereffects in vocal age perception. *Hearing Research, 282*(1-2), 283-288. doi: 10.1016/j.heares.2011.06.008

Zäske, R., Schweinberger, S. R., & Kawahara, H. (2010). Voice aftereffects of adaptation to speak identity. *Hearing Research*, 268, 38-45.

Zatorre, R. J., & Halpern, A. R. (1979). Identification, discrimination and selective adaptation of simultaneous musical intervals. *Perception & Psychophysics, 26*(5), 384-395.

Appendix A- Acoustic analyses for female continua unimodal voice experiment.



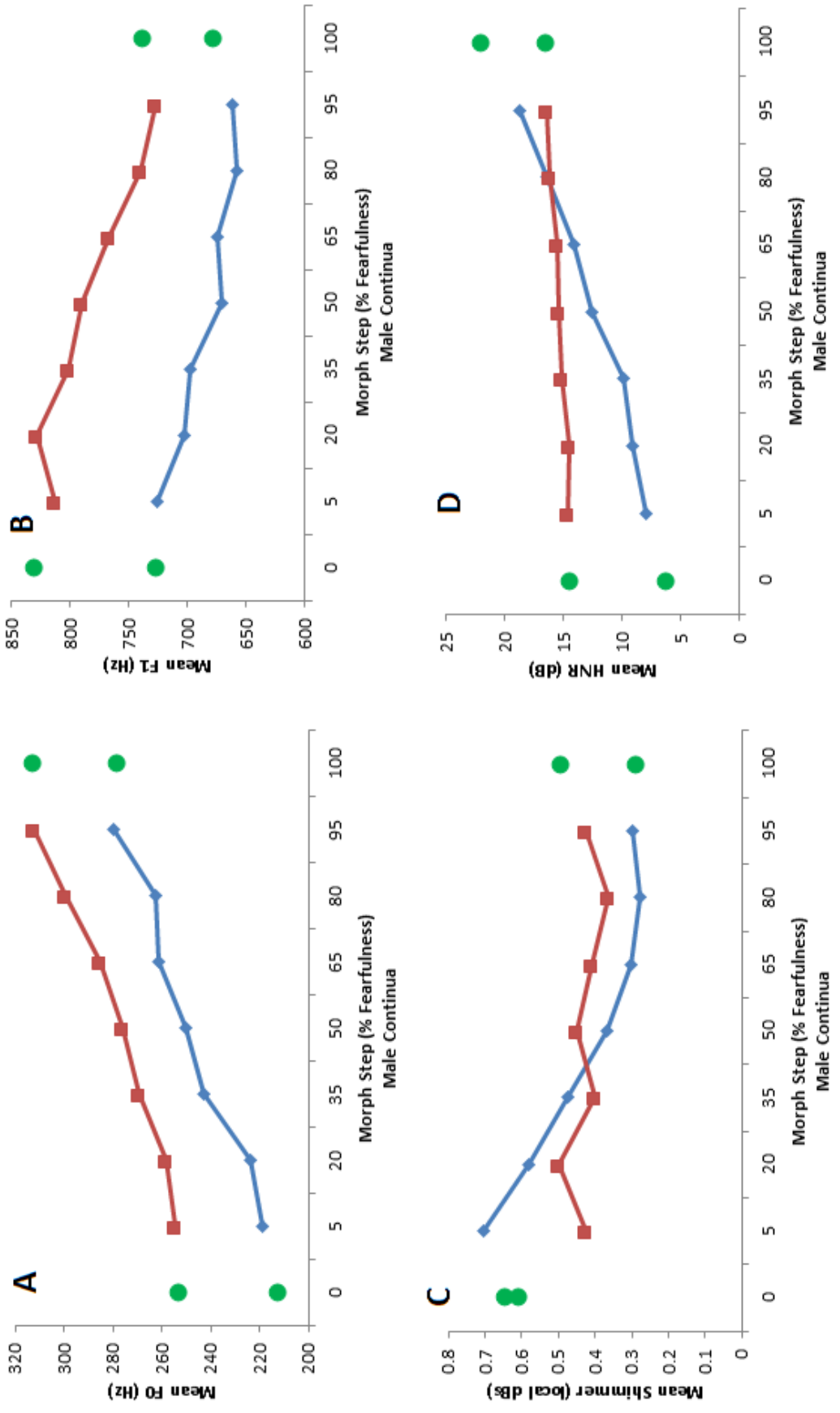
A: Mean F0 for adaptor and test stimuli for the two female vocal continua.

B: Mean F1 for adaptor and test stimuli for the two female vocal continua.

C: Mean Shimmer for adaptor and test stimuli for the two female vocal continua.

D: Mean HNR for adaptor and test stimuli for the two female vocal continua.

Appendix B- Acoustic analyses for male continua in unimodal voice experiment



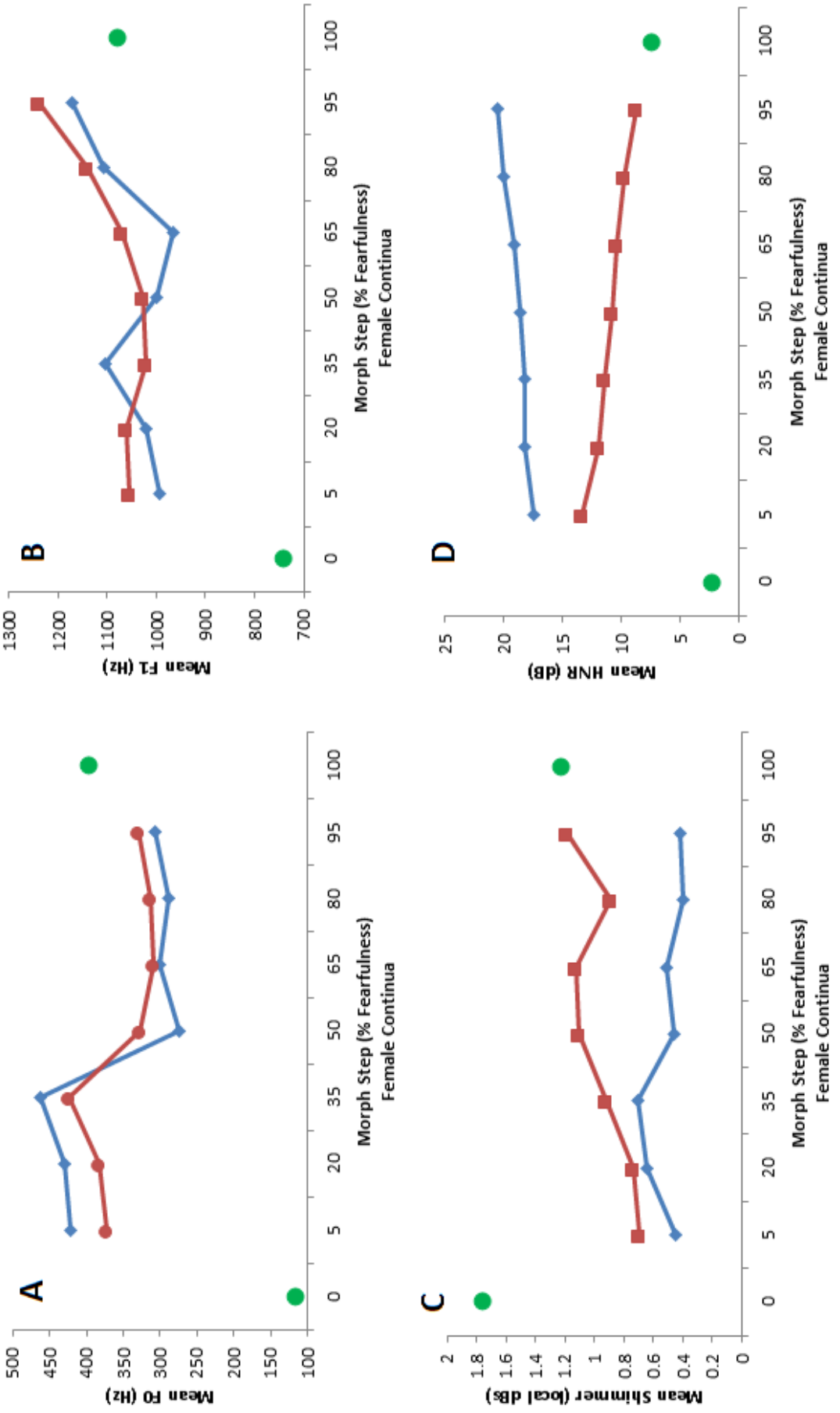
A: Mean F0 for adaptor and test stimuli for the two male vocal continua.

B: Mean F1 for adaptor and test stimuli for the two male vocal continua.

C: Mean Shimmer for adaptor and test stimuli for the two male vocal continua.

D: Mean HNR for adaptor and test stimuli for the two male vocal continua.

Appendix C- Acoustic analyses for female continua in unimodal dog call adaptation experiment



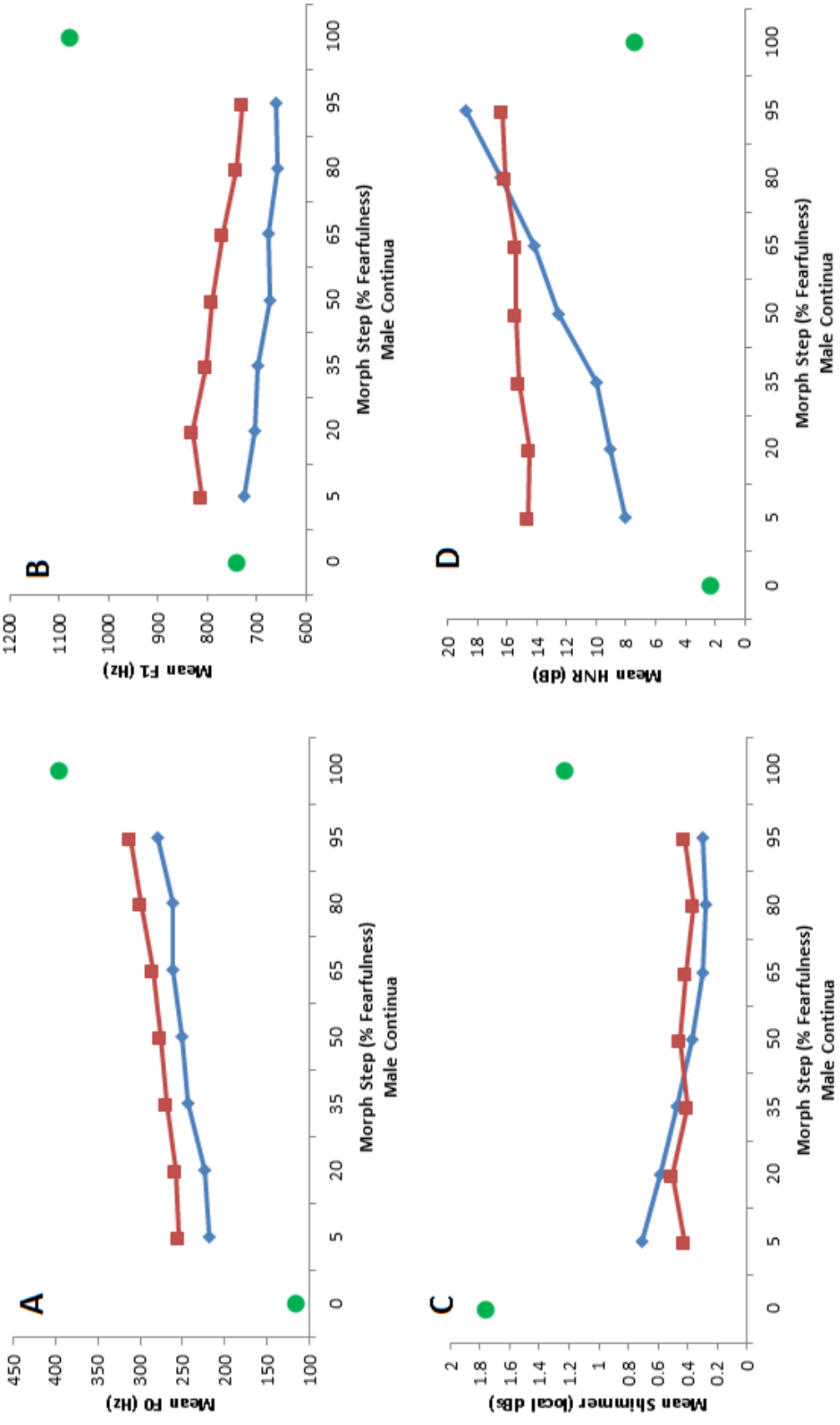
A: Mean F0 for adaptor and test stimuli for the two female vocal continua.

B: Mean F1 for adaptor and test stimuli for the two female vocal continua.

C: Mean Shimmer for adaptor and test stimuli for the two female vocal continua.

D: Mean HNR for adaptor and test stimuli for the two female vocal continua.

Appendix D- Acoustic analyses for male continua in unimodal dog call adaptation experiment



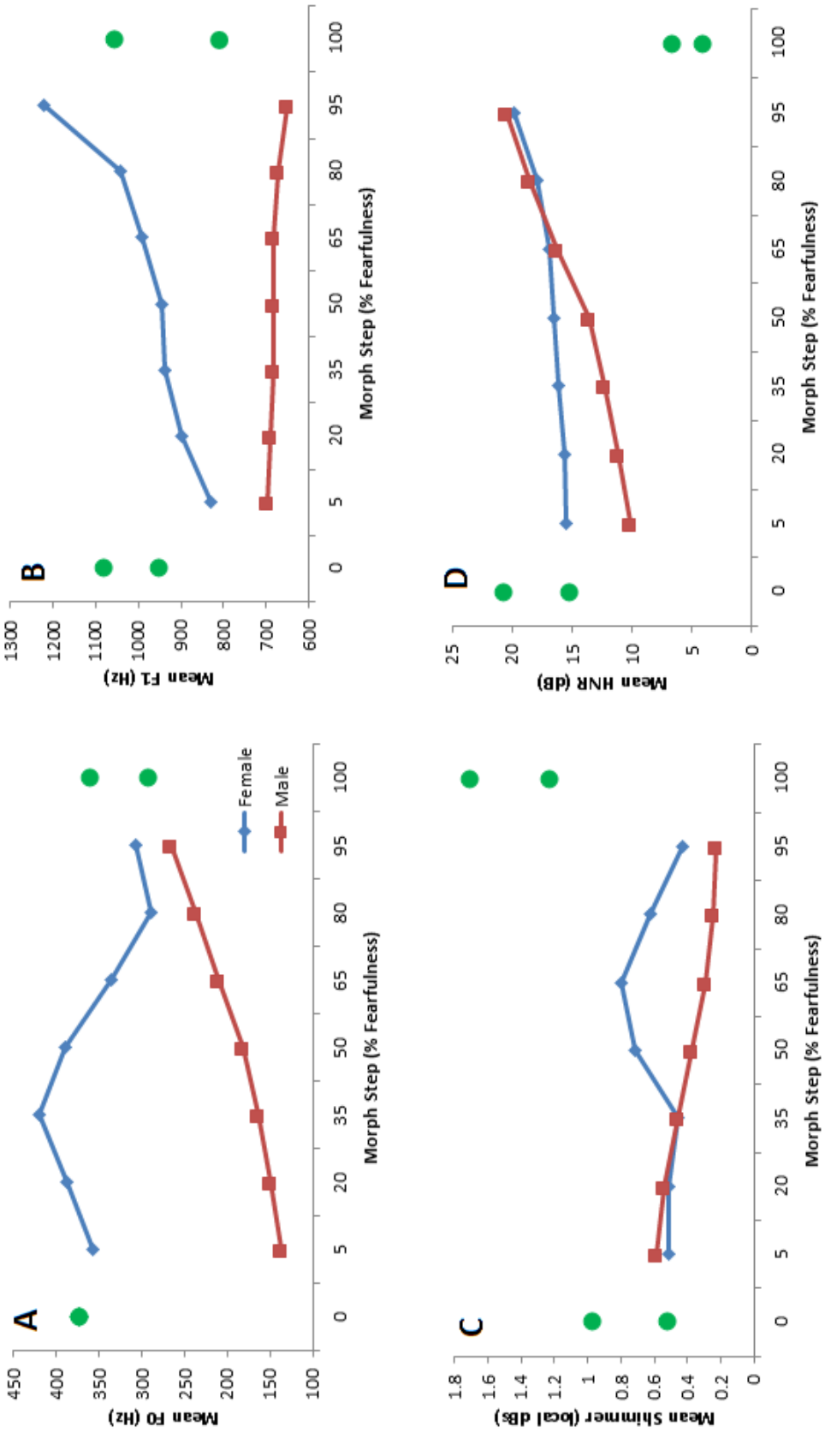
A: Mean F0 for adaptor and test stimuli for the two male vocal continua.

B: Mean F1 for adaptor and test stimuli for the two male vocal continua.

C: Mean Shimmer for adaptor and test stimuli for the two male vocal continua.

D: Mean HNR for adaptor and test stimuli for the two male vocal continua.

Appendix E- Acoustic analyses for male and female continua and all adapting stimuli for unimodal musical expressive bursts adaptation experiment



A: Mean F0 for adaptor and test stimuli

B: Mean F1 for adaptor and test stimuli

C: Mean Shimmer for adaptor and test stimuli

D: Mean HNR for adaptor and test stimuli

Appendix F- fundamental and first formants for all voices in database from which voice test stimuli were extracted

Females recorded in Bangor

		F0	F1
F100_long_e	e	210	385
F101_long_e	e	229	361
F102_long_e	e	235	283
F103_long_e	e	205	401
F104_long_e	e	227	425
F105_long_e	e	218	347
F106_long_e	e	214	387
F107_long_e	e	204	389
F108_long_e	e	231	334
F10_e	e	221	420
F11_e	e	168	318
F12_e	e	185	329
F13_e	e	205	396
F14_long_e	e	196	380
F15_e	e	159	315
F16_long_e	e	242	385
F17_e	e	272	281
F18_e	e	177	346
F19_long_e	e	219	338
F1_e	e	231	399
F20_long_e	e	147	300
F21_e	e	200	384
F22_e	e	246	434
F23_e	e	210	417
F24_long_e	e	205	404
F25_long_e	e	234	318
F26_e	e	199	396
F27_e	e	192	383
F28_e	e	230	404
F29_e	e	232	323
F2_long_e	e	243	436
F30_e	e	224	329
F31_e	e	232	394
F32_e	e	183	360
F33_e	e	201	363
F34_e	e	185	341
F35_e	e	213	334
F36_e	e	176	330
F37_long_e	e	224	336
F38_e	e	178	335
F39_e	e	197	337

F3_e	e	212	361
F40_e	e	199	389
F41_e	e	223	320
F42_e	e	228	329
F43_e	e	237	435
F44_long_e	e	216	402
F45_long_e	e	259	336
F46_e	e	218	409
F47_long_e	e	197	365
F48_long_e	e	233	379
F49_long_e	e	268	465
F4_e	e	228	334
F50_long_e	e	214	375
F51_long_e	e	201	336
F52_long_e	e	180	251
F53_long_e	e	193	283
F54_long_e	e	266	307
F55_long_e	e	202	380
F56_long_e	e	195	368
F57_long_e	e	200	362
F58_long_e	e	234	338
F59_long_e	e	185	314
F5_e	e	230	379
F60_long_e	e	230	420
F61_long_e	e	209	330
F62_long_e	e	206	409
F63_long_e	e	232	340
F64_long_e	e	227	353
F65_long_e	e	202	358
F66_long_e	e	208	358
F67_long_e	e	209	415
F68_long_e	e	248	432
F69_long_e	e	216	291
F6_e	e	252	281
F70_long_e	e	272	292
F71_long_e	e	289	295
F72_long_e	e	184	328
F73_long_e	e	236	383
F74_long_e	e	217	380
F75_long_e	e	213	432
F76_long_e	e	185	349
F77_long_e	e	204	373
F78_long_e	e	183	355
F79_long_e	e	175	337
F7_e	e	220	425
F80_long_e	e	224	276

F81_long_e	e	223	338
F82_long_e	e	218	258
F83_long_e	e	249	463
F84_e	e	223	312
F85_e	e	210	409
F86_e	e	200	397
F87_e	e	227	400
F88_e	e	204	375
F89_e	e	192	381
F8_e	e	212	409
F90_e	e	208	338
F91_e	e	179	349
F92_e	e	194	397
F93_e	e	205	410
F94_e	e	211	414
F95_e	e	205	395
F96_e	e	231	258
F97_e	e	232	291
F98_e	e	214	380
F99_long_e	e	235	375
F9_e	e	204	345

Females recorded in Glasgow

		F0	F1
F109_e	e	203	284
F110_e	e	194	349
F111_e	e	222	267
F112_e	e	222	328
F113_e	e	210	355
F114_e	e	232	260
F115_e	e	182	356
F116_e	e	192	380
F117_e	e	245	264
F118_e	e	229	281
F119_e	e	202	371
F120_e	e	245	278
F121_e	e	172	341
F122_e	e	260	441
F123_e	e	201	409
F124_e	e	198	375
F125_e	e	229	447
F126_e	e	238	416
F127_e	e	230	279
F128_e	e	231	413
F129_e	e	223	396

F130_e	e	235	410
F131_e	e	189	286
F132_e	e	199	321
F133_e	e	189	582
F134_e	e	222	324
F135_e	e	227	277
F136_e	e	175	391
F137_e	e	208	297
F138_e	e	206	282
F139_e	e	206	380
F140_e	e	200	359
F141_e	e	231	336
F142_e	e	220	376
F143_e	e	199	603
F144_e	e	204	370
F145_e	e	190	293
F146_e	e	203	379
F147_e	e	230	446
F148_e	e	201	382
F149_e	e	218	406
F150_e	e	191	270
F151_e	e	148	306
F152_e	e	219	315
F153_e	e	149	276
F154_e	e	256	365
F155_e	e	236	393
F156_e	e	264	452
F157_e	e	190	376
F158_e	e	194	381
F159_e	e	222	414
F160_e	e	238	305
F161_e	e	232	340
F162_e	e	243	386
F163_e	e	187	358
F164_e	e	206	292
F165_e	e	193	340
F166_e	e	242	255
F167_e	e	192	381
F168_e	e	237	253
F169_e	e	222	373
F170_e	e	219	354
F171_e	e	248	477
F172_e	e	217	421
F173_e	e	207	403
F174_e	e	218	418
F175_e	e	316	327

F176_e	e	187	283
F177_e	e	216	391
F178_e	e	193	351
F179_e	e	215	406
F180_e	e	190	288
F181_e	e	221	382
F182_e	e	227	446

Males recorded in Bangor

		F0	F1
M14_e	e	80	253
M15_e	e	116	282
M16_e	e	135	312
M17_e	e	152	299
M18_e	e	83	293
M19_long_e	e	126	261
M1_e	e	115	305
M20_e	e	131	310
M21_e	e	147	281
M22_e	e	102	275
M23_e	e	110	292
M24_e	e	102	296
M25_long_e	e	94	312
M26_e	e	122	349
M27_e	e	97	278
M28_e	e	100	272
M29_long_e	e	118	331
M2_e	e	114	314
M30_long_e	e	110	248
M31_long_e	e	118	236
M32_long_e	e	116	250
M33_long_e	e	114	233
M34_long_e	e	106	285
M35_long_e	e	136	278
M36_long_e	e	150	180
M37_long_e	e	103	300
M38_long_e	e	113	278
M39_long_e	e	109	260
M3_e	e	117	259
M40_long_e	e	119	274
M41_long_e	e	111	257
M42_long_e	e	124	307
M43_long_e	e	103	273
M44_long_e	e	124	247
M45_long_e	e	119	294

M46_long_e	e	109	312
M47_long_e	e	125	276
M48_long_e	e	97	284
M49_long_e	e	140	336
M4_e	e	118	242
M50_long_e	e	86	326
M51_long_e	e	94	335
M52_long_e	e	97	309
M53_long_e	e	111	266
M54_e	e	131	339
M55_e	e	103	237
M56_e	e	104	284
M57_long_e	e	141	282
M5_e	e	108	310
M6_e	e	110	316
M7_e	e	126	266
M8_e	e	103	288
M9_e	e	118	256
M118_e	e	112	247
M119_e	e	119	341
M120_e	e	120	245
M121_e	e	117	254
M122_e	e	151	287
M123_e	e	133	269
M124_e	e	101	265

Males recorded in Glasgow

		F0	F1
M101_e	e	126	274
M102_e	e	136	280
M103_e	e	124	238
M104_e	e	115	287
M105_e	e	101	276
M106_e	e	83	318
M107_e	e	109	314
M108_e	e	166	308
M109_e	e	150	299
M110_e	e	107	290
M111_e	e	119	253
M112_e	e	100	303
M113_e	e	102	265
M114_e	e	87	260
M115_e	e	141	283
M116_e	e	112	311
M58_e	e	118	290

M59_e	e	95	230
M60_e	e	122	269
M61_e	e	130	285
M62_e	e	116	299
M63_e	e	101	288
M64_e	e	118	293
M65_e	e	87	292
M66_e	e	101	254
M67_e	e	133	279
M68_e	e	116	295
M69_e	e	106	255
M70_e	e	138	274
M71_a	e	112	483
M72_e	e	104	330
M73_e	e	133	271
M74_e	e	95	224
M75_e	e	112	286
M76_e	e	101	299
M77_e	e	129	266
M78_e	e	127	266
M79_e	e	110	289
M80_e	e	109	277
M81_e	e	113	324
M82_e	e	127	285
M83_e	e	111	259
M84_e	e	129	286
M85_e	e	120	274
M86_e	e	135	258
M87_e	e	118	285
M88_e	e	153	311
M89_e	e	145	295
M90_e	e	118	266
M91_e	e	130	260
M92_e	e	161	281
M93_e	e	124	290
M94_e	e	101	301
M95_e	e	126	322
M96_e	e	113	297
M97_e	e	93	248
M98_e	e	129	281
M99_e	e	131	279
M100_e	e	116	270