Individual differences in voice adaptability are specifically linked to voice perception skill

Bestelmeyer, Patricia; Mühl, Constanze

Cognition

DOI:
10.1016/j.cognition.2021.104582

Published: 01/05/2021

Peer reviewed version

Individual differences in voice adaptability are specifically linked to voice perception skill

Patricia E.G. Bestelmeyer[1*] & Constanze Mühl[1]

[1]School of Psychology, Bangor University, Bangor, Gwynedd, UK

[*]Correspondence:

Dr Patricia E.G. Bestelmeyer

School of Psychology, Brigantia

Bangor University, Gwynedd, LL57 2AS, UK

E-mail: p.bestelmeyer@bangor.ac.uk

Phone: +44 (0)1248 383488

Abstract

There are remarkable individual differences in the ability to recognise individuals by the sound of their voice. Theoretically, this ability is thought to depend on the coding accuracy of voices in a low-dimensional “voice-space”. Here we were interested in how adaptive coding of voice identity relates to this variability in skill. In two adaptation experiments we explored first whether the aftereffect size to two familiar vocal identities can predict voice perception ability and second, whether this effect stems from general auditory skill (e.g. discrimination ability for tuning and tempo). Experiment 1 demonstrated that contrastive aftereffect sizes for voice identity predicted voice perception ability. In Experiment 2, we replicated this finding and further established that this effect is unrelated to general auditory abilities or general adaptability of listeners. Our results highlight the important functional role of adaptive coding in voice expertise and suggest that human voice perception is a highly specialised and distinct auditory ability.

Keywords: individual differences; adaptation; aftereffects; voice; identity; robust regression

1. Introduction

Successful communication relies on our ability to discriminate minute variations in vocal sounds. This skill allows us to differentiate and recognise the identity of speakers despite their overall similar acoustic structure (Hanley, Smith, & Hadfield, 1998). The fact that many species have developed the ability to recognise an individual's identity based on the sound of a conspecific's voice underlines the biological importance of this skill. Despite its significance, recent research has shown a wide inter-individual variability in the ability to recognise speakers from the sound of their voice alone (Mühl, Sheil, Jarutyte, & Bestelmeyer, 2018) and to remember voices (Aglieri et al., 2017). Even when linguistic information is carried by voices, a rich source of additional speaker-specific information (e.g. accent, speech rate), participants' identification rates vary greatly (Lavan, Burston, & Garrido, 2019; Roswandowitz et al., 2014; Shilowich & Biederman, 2016; Van Lancker, Kreiman, & Cummings, 1989).

These individual differences in skill may be underpinned by differences in the quality of norm-based coding of voice identity in a high dimensional voice-space (Baumann & Belin, 2010). This influential account suggests that voices are coded relative to an average, or prototypical, voice which is located at the centre of this acoustic space. Accordingly, and in line with Valentine's (1991) face-space idea, voices may be encoded along acoustic dimensions, which facilitate discrimination between voices (e.g. fundamental frequency, formant frequencies and the "smoothness" of voices (harmonics-to-noise ratio); Baumann & Belin, 2010; Latinus, McAleer, Bestelmeyer, & Belin, 2013). This framework predicts that voices that are further away from the prototypical (or mean) voice are perceived as more distinctive and are therefore easier to discriminate than voices that are closer to the central voice. This idea has received support from neuroimaging studies whereby voices that were morphed to be further away from the mean voice elicited greater activation in voice-sensitive cortices compared to voices that were closer to the mean (Bestelmeyer, Latinus, Bruckert, Crabbe, & Belin, 2012; Latinus et al., 2013).

Behavioural adaptation paradigms are a simple tool to explore the perceptual representation of certain stimulus attributes and probe how these may be coded in the brain. These attributes can range from simple features such as motion or colour to higher-level, abstract features such as the identity of a face (Burton, Jenkins, & Schweinberger, 2011; Leopold, O'Toole, Vetter, & Blanz, 2001; Little, Hancock, DeBruine, & Jones, 2012) or voice (Latinus & Belin, 2011, 2012; Zäske, Schweinberger, & Kawahara, 2010). Adaptation is ubiquitous in perception and refers to a process during which continued stimulation results in a biased perception towards opposite features of the adapting stimulus (Grill-Spector et al., 1999). Research using adaptation has revealed neural populations tuned to respond to specific stimulus attributes by isolating and subsequently distorting the perception of these attributes (Grill-Spector et al., 1999; Winston, Henson, Fine-Goulden, & Dolan, 2004). In the field of voice identity, researchers employ morphing to show that repeated exposure to a particular voice identity results in contrastive aftereffects in voice identity perception. For example, Latinus and Belin (2011) created a voice space by averaging the vowels of sixteen male speakers. Vowels of three familiar speakers (A, B and C) were independently morphed with the average of sixteen speakers to create three continua plus their respective “anti-identities” (i.e. extrapolations on the trajectory passing through the mean voice thus yielding the acoustic opposite of the familiar voices). Following adaptation to an anti-voice (e.g. anti-A), listeners were more likely to perceive the identity-ambiguous mean voice as belonging to familiar Speaker A compared to Speakers B or C. In other words, adaptation to one of the anti-identities provided the neutral mean voice with a “new” identity, each time in line with the particular identity trajectory in voice space. These contrastive aftereffects are consistent with the notion that voice identities are represented or encoded as deviations from a prototypical voice.

It is still debated what function adaptation serves in perception. A range of explanations have been proposed, most of which involve coding efficiency (Clifford et al., 2007; Wainwright, 1999; Wark, Lundstrom, & Fairhall, 2007; Webster, 2011). For example, adaptation may

continuously recalibrate perceptual norms to maintain a match between coding and environment (Webster, 2011). It may also enhance coding by reducing the sensitivity to continued stimulation, which in turn enhances sensitivity to change (Webster, 2011). There is recent evidence in the face literature to suggest that adaptive coding of faces may contribute to our ability to discriminate and recognise faces (Dennett, McKone, Edwards, & Susilo, 2012; Engfors, Jeffery, Gignac, & Palermo, 2017; Rhodes, Jeffery, Taylor, Hayward, & Ewing, 2014; Rhodes et al., 2015; see also Palermo et al. (2018) on adaptive coding of facial expression; see Table 1 for summary of previous findings in the face literature). These studies have shown that face identity aftereffects positively correlated with a face memory test but not a non-face, object memory test. These findings support a functional role for adaptive coding in face memory ability.

Table 1.

Summary of previous findings in the face literature reporting the relationship between face aftereffects and recognition ability.

Reference	*N*	Face aftereffect type	Face test	Pearson's r between aftereffect size and recognition ability
Dennet et al. (2012)	78	eye-height	CFMT	r = .23, p = .04
Engfors et al. (2016)	175	identity	CFMT	r = .45, p < .001
Palermo et al. (2018)	88	expression	emotion labelling	r = .38, p < .001
Rhodes et al. (2014)	129	identity	CFMT	r = .17., p = .049
Rhodes et al. (2015)	156	identity	CFMT	r = .19, p = .02

Here we were interested in whether individual differences in adaptive coding of voice identity can be linked to voice discrimination skill. We measured participants' ability to recognise speakers from the sound of their voice using the Bangor voice matching test (Mühl et al., 2018) and assessed whether the degree of adaptability to voice identity can be used as a predictor of performance on this test. Given the findings in the face literature, we predicted a positive relationship between the size of the voice identity aftereffect and voice perception skill. The second experiment had two aims, which addressed to what extent the relationship between voice identity aftereffect and voice perception skill is specifically due to voice-level coding rather than more generic sound coding. First, we investigated whether a relationship between voice perception ability and size of the voice identity aftereffect could be due to differences in general auditory abilities. We therefore administered a short auditory skills test that assessed the ability to match rhythms, melodies and other auditory features (Zentner & Strauss, 2017). Second, we created a non-voice adaptation task to serve as a control condition to establish whether larger aftereffects to any sound category simply come with better sound discrimination. If individual differences in voice adaptability are specifically linked to voice perception skill, we should find no relationship between a general auditory ability test and the size of the voice identity aftereffect. Supporting this notion, we did not expect to find a correlation between aftereffect sizes of the identity and control tasks.

2. Methods

2.1 Participants

Eighty-seven undergraduate students (73 females; mean age = 20.5; S.D. = 5.74) contributed data to the analysis of Experiment 1. Ninety-one different undergraduate students (82 females; mean age = 20.9; S.D. = 4.33) contributed data to the analysis of Experiment 2. Datasets from three participants were excluded from analysis of Experiment 1 and two from

Experiment 2 due to poor recognition performance of at least one speaker (correct identification < 65% on a familiarity task described in the Procedure section). All participants were English native speakers and reported normal hearing. Volunteers took part in return for course credit. The ethics committee of the School of Psychology at Bangor University approved the experimental protocols.

No previous literature exists on linking voice aftereffects to recognition skill and we did not have access to appropriate pilot data. We therefore based this sample size on our power calculation on the correlation sizes reported in the face literature that link aftereffect sizes to face memory or face recognition. Pearson's correlations range between r = .173 and r = .450 (Dennet et al., 2012; Engfors et al., 2017; Palermo et al., 2018; Rhodes et al., 2014, 2015). For a medium effect of r = .3, alpha level of .05 and power of .8 we calculated that we require a sample size of 84 participants (G*Power; Faul, Erdfelder, Lang & Buchner, 2007).

2.2 Stimuli

To elicit identity-specific aftereffects, we chose stimuli of personally familiar speakers. We therefore recorded eight psychology lecturers in a sound attenuated booth using a Rode NT1-A microphone and Audacity software (16-bit, 44.1 kHz sampling rate, mono). The recording protocol included the numbers 1-10 and 16 short non-sense syllables (aba, aga, ada, ubu, ugu, udu, ibi, igi, idi, had, hod, hed, hud, hood, hid, hide). We selected the sound recordings of two female lecturers who were most familiar to undergraduate students and who did not speak with a significant regional accent. Recordings were edited in Cool Edit Pro 2.0 and normalised in energy (root mean square; RMS) before morphing. Voice continua between the voice recordings of two lecturers were created for the syllable /aga/ using Tandem-STRAIGHT (Kawahara et al., 2008) in MatlabR2013a (The Mathworks, Inc). Voice continua consisted of seven morph steps that corresponded to 5%/95%, 20%/80%, 35%/65%, 50%/50%, 65%/35%, 80%/20%, 95%/5% of SpeakerA/SpeakerB. Tandem-STRAIGHT

performs an instantaneous pitch-adaptive spectral smoothing of each stimulus for separation of contributions to the voice signal arising from the glottal source (including f0) versus supralaryngeal filtering (distribution of spectral peaks, including the first formant frequency, F1). Voice stimuli were decomposed by Tandem-STRAIGHT into five parameters: fundamental frequency (f0; the perceived pitch of the voice), frequency, duration, spectrotemporal density and aperiodicity. Each parameter can be manipulated independently. For each recording of /aga/ we manually identified one time landmark with three frequency landmarks (corresponding to the first three formants) at the onset of phonation and the same number of landmarks at the offset of phonation. Morphed stimuli were then generated by re-synthesis based on the interpolation (linear for time; logarithmic for F0, frequency, and amplitude) of these time-frequency landmark templates (see also Schweinberger, Kawahara, Simpson, Skuk and Zäske (2014) for a discussion of the voice morphing technique). The remaining 15 syllables were used in their original form as adaptors. We used the same voice stimuli in Experiments 1 and 2.

For the control adaptation task in Experiment 2 we first generated two tones, middle C (C4; 262 Hz) and middle C-sharp (C4#; 277 Hz), with their respective second and third harmonics in Matlab. These two tones were then combined to form the maximally consonant (unison of C4/C4 or C4#/C4#) and dissonant (minor second; C4/C4# or C4#/C4) adaptor chords. To create the continua we generated seven additional tones with frequencies between C4 and C4# (and equivalent timbre). These seven steps were chosen to mimic the same sound mixture we used for the identity stimuli (i.e. from 5% to 95%: 262.75, 265, 267.25, 269.5, 271.75, 274 and 276.25 Hz). We then created one continuum by mixing down each one of these seven tones with the C4 (i.e. a continuum in seven steps from unison of C4 to minor second) and a second continuum by mixing down each of the seven tones with the C4#. This procedure ensured that the level of consonance was not confounded with the level of perceived pitch (i.e. the maximally dissonant sound was not also always the higher pitched

sound). This procedure resulted in two continua of seven morphs varying in levels of consonance.

Stimuli were root mean square (RMS) normalised. All tasks were implemented in Psychtoolbox-3 (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007) for Matlab. Sounds were presented via headphones (Beyerdynamic DT770 Pro; 250 Ohm) at 75dB SPL(C).

2.3 Procedure

Before the voice adaptation experiment, participants completed the Bangor Voice Matching Test (BVMT; Mühl et al., 2018). This is a standardised test for the assessment of voice perception skill and takes 10 minutes to complete. It requires a same/different identity judgment after hearing two voice samples. Following the BVMT, participants listened passively to a recording of the two lecturers whose voices were used in the adaptation task. The recordings consisted of each speaker counting from 1 to 10. Following that, in order to ascertain sufficient familiarity with the speakers, participants completed a familiarity task. In this three-alternative forced choice task, participants listened to a randomised sequence of three of their lecturers (Speaker A, B, and C) articulate 16 different non-sense syllables each. After each syllable, participants decided who the speaker was by pressing one of three keys. Auditory feedback for each decision was provided to indicate the correctness of the response. At the end of this familiarity test, participants' accuracy score was presented on the screen. Only participants who scored a minimum of 80% correct overall were asked to carry on with the experiment.

The voice adaptation experiment consisted of three blocks (1 categorisation; 2 adaptation; see also Pye (2015) for a pilot study). We always administered the categorisation task first to get a baseline rating of all seven morph steps along the Speaker A to B continuum. The categorisation block consisted of 56 randomised trials (8 instances of each of the 7 morph

steps). Participants categorised each sound as belonging to either Speaker A or B by means of a button press. Participants then completed two blocks of adaptation (one block adapting to Speaker A, and one block adapting to Speaker B). We counterbalanced the order of adaptation blocks across participants. Each adaptation block comprised 56 trials. Each trial consisted of four different adaptor syllables (randomly selected from a pool of 15 nonsense syllables) followed by 1s of silence. Participants then heard one of the seven morph steps, which they were asked to categorise as belonging to either Speaker A or B. Each morph step appeared 8 times in total per block. Once participants had completed their first adaptation block, they took a short break (~5 minutes) before starting the second adaptation block. This break was included to ensure the adaptation effect from the previous block would not carry over into the second block of adaptation. A reminder of which keys to press for each speaker remained on the screen for the duration of the blocks.

We ran Experiment 2 across two counterbalanced sessions scheduled on two different days. One session was identical to Experiment 1, the other session consisted of a musical ability test (Mini-PROMS; Zentner & Strauss, 2017) followed by a non-voice control task (adaptation to consonance or dissonance). The control task was prefaced by two examples of consonant (pleasant) and dissonant (unpleasant) chords. Participants then completed the categorisation task, which consisted of 56 randomised trials. Both control adaptation blocks followed the structure of the voice identity task. Following four adaptors that were either consonant or dissonant, depending on the block, participants were presented with one of the morphed sounds. Each experimental session took approximately one hour to complete.

2.4 Statistical Analyses

We used bootstrapping to assess statistical significance of the difference between the baseline and each adaptation condition using Matlab2015b with Statistics Toolbox (The MathWorks, Inc.). Each bootstrap sample was derived by randomly sampling from our

participants with replacement. Thus, a participant and associated data could be selected more than once or not at all according to conventional bootstrap methodology (Wilcox, 2012). Data from both adaptation conditions were selected for each sampled participant because of the within-subjects design of the experiments. For each participant, data were then averaged as a function of the seven morph steps and a psychophysical curve (based on the hyperbolic tangent function) was fitted for each condition. We then computed the difference between adaptation conditions. We repeated this process 9,999 times, which led to a distribution of 10,000 bootstrapped estimates of the fit to the psychophysical curves as well as differences between curves for the two conditions. Lastly, we calculated the 95% confidence intervals ($CI_{95\%}$) for the fitted curve (Figure 1A) and the differences between two conditions (Figure 1B). A difference between conditions is deemed significant if the mean of the differences and its $CI_{95\%}$ excludes 0 (e.g. Cumming, 2012). The point of subjective equality (PSE), i.e. the centre of symmetry of the psychophysical function, was also computed for the baseline condition. The PSE is illustrated with a star on the average baseline curve and consistently overlaps with the 50% morph, i.e. the mathematically most ambiguous morph.

We also computed the aftereffect size by calculating the difference in mean response between the 50% morph in the adaptation conditions. We used this measure and the z-scored BVMT data in a robust regression to assess the relationship between the two variables. While robust regression does not produce an R^2 value (a measure obtained with least squares regression), we opted for a robust regression analysis because it performs better than parametric methods when the data is not normally distributed and when outliers may be present by minimising (i.e. differently weighting) their impact (Wilcox, 2012). While the data from both tasks were normally distributed in Experiment 2, the Henze-Zirkler Multivariate Normality Test (Pernet, Wilcox, & Rousselet, 2013) was significant for Experiment 1 ($p = .041$) suggesting that the data does not have a normal distribution. We used Matlab's *robustfit* function and its default parameters, which employs a bisquare weighting algorithm, to achieve this down-weighting of possible outliers. Following this, we calculated percentile bootstrap confidence intervals around the slopes. We did this by sampling subjects with replacement,

keeping their corresponding z-scored voice test score and aftereffect size. We then performed regressions between each measure of the voice test and aftereffect size. We conducted these steps 600 times and each time we saved the resulting slope. Then, we sorted the bootstrapped slopes, and used the 2.5 and 97.5 percentiles to build the boundaries of 95% bootstrap confidence intervals to assess significance. We used the same statistical analysis for the data obtained in Experiment 2.

For direct comparison to results reported in the face literature, we report Pearson's r, without removal of outliers identified by our robust statistics so as not to inflate Type 1 error rates.

3. Results

3.1 Experiment 1

Figure 1A shows the fitted functions for the baseline condition and each of the two adaptation conditions for Experiment 1, bootstrap 95% confidence interval ($CI_{95\%}$) and error bars (SEM). Figure 1B illustrates the $CI_{95\%}$ of the differences between baseline and adaptation conditions and illustrates a significant identity aftereffect which was largest at the 50% morph, the mathematically and perceptually most ambiguous morph. Figure 1C is a scatterplot which shows a positive relationship between the size of the identity aftereffect and the BVMT score. The robust regression showed that the aftereffect size was a significant predictor of BVMT performance ($p = .03$; $CI_{95\%} = [.075, 1.650]$). Pearson's correlation was significant ($r = .24$; $p = .02$; $CI_{95\%} = [.032 .430]$).

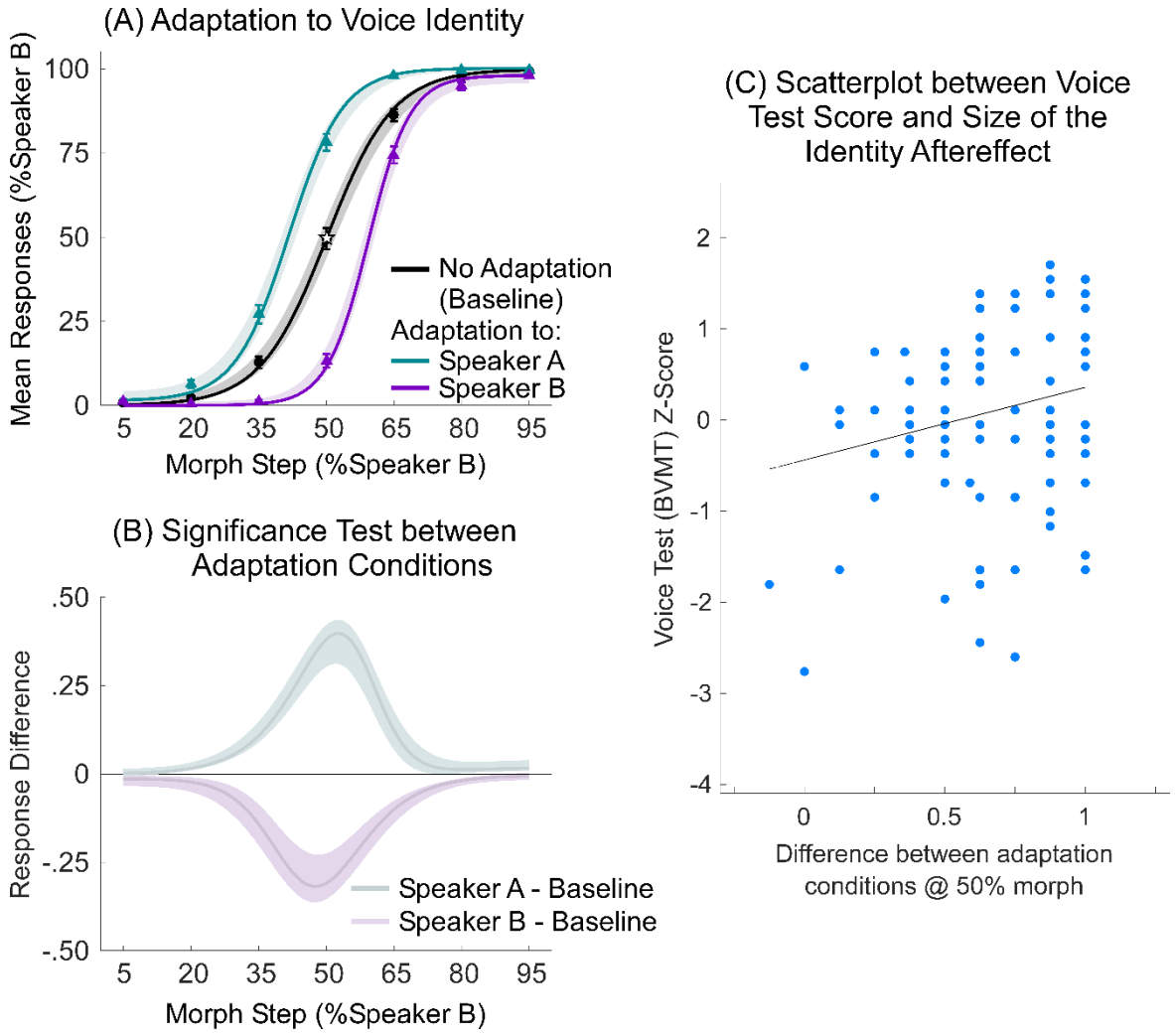


Fig. 1. Results of Experiment 1. (A) Average psychophysical functions for the baseline condition (black) and the two adaptation conditions (turquoise and purple). Error bars represent standard errors of the mean (SEM) and shaded areas illustrate the bootstrapped $CI_{95\%}$ of the fitted curve to participants' mean responses. (B) Mean difference and $CI_{95\%}$ (shaded area) of the differences between baseline and each adaptation condition. Significant differences between conditions occurred along the most ambiguous part of the continuum (i.e. where the $CI_{95\%}$ area does not overlap with the y = 0 line). (C) Scatterplot of the identity aftereffect size against the z-scored voice test (BVMT) results (with robust regression line).

3.2 Experiment 2

Figure 2 illustrates the results of Experiment 2. The voice identity task is a replication of Experiment 1 and the results are illustrated in Figure 2A-C. The results of the non-voice control task are illustrated in Figure 2D-F. Figure 2A shows the fitted functions for the baseline condition of the voice identity task and each of the two adaptation conditions, $CI_{95\%}$ and error bars (SEM). Figure 2B illustrates the $CI_{95\%}$ of the differences between the baseline and adaptation conditions for the voice task and shows a significant identity aftereffect which was largest at the 50% morph. Figure 2C is a scatterplot, which shows a positive relationship between the size of the identity aftereffect and the BVMT score. The robust regression showed that the identity aftereffect size was a significant predictor of BVMT performance ($p = .01$; $CI_{95\%} = [.107, 1.340]$). Results of the identity adaptation task of Experiment 2 directly replicate those of Experiment 1. The robust regression between the identity aftereffect size and music test score is illustrated in supplementary Figure 1A and was not significant ($p = .92$; $CI_{95\%} = [-.715 .827]$).

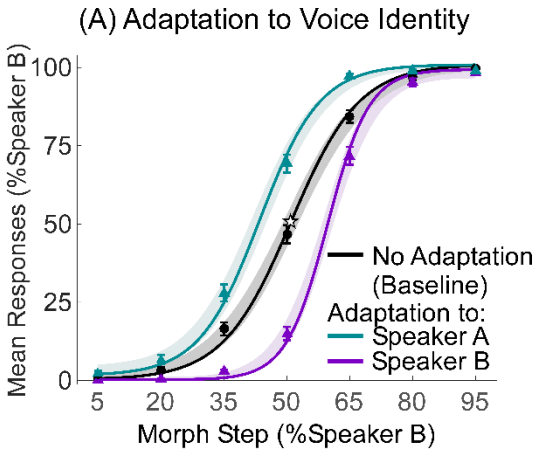
Figure 2D displays the fitted functions for the baseline condition of the consonance task and each of the two adaptation conditions, $CI_{95\%}$ and error bars (SEM). Figure 2E illustrates the $CI_{95\%}$ of the differences between the baseline and adaptation conditions for this control task and demonstrates a significant aftereffect which was generally largest at the 50% morph. Figure 2F is a scatterplot which shows a negative relationship between the size of this aftereffect and the PROMS score. The robust regression showed that the consonance aftereffect size was not a predictor of performance ($p = .17$; $CI_{95\%} = [-1.340 .201]$) on the music test (PROMS).

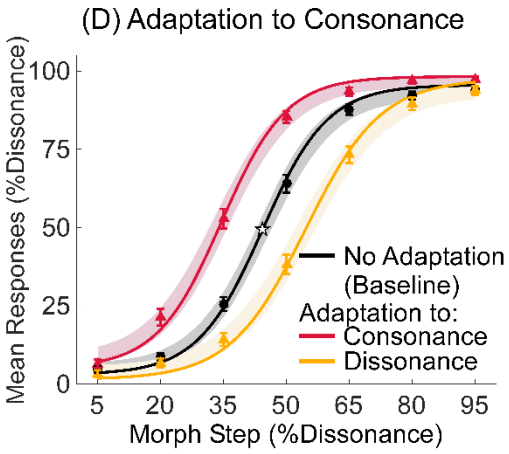
3.2.1 Pearson's correlations for comparison to the face literature

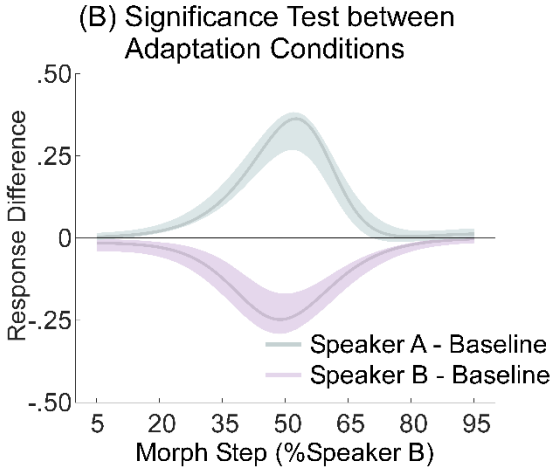
Significance of the Pearson's correlations were in line with the robust regressions. There was a significant positive correlation between voice test and identity aftereffect (r = .26; p = .01; $CI_{95\%}$ = [.052 .439]). In contrast, we found no correlation between the music test and identity aftereffect (r = .01; p = .92; $CI_{95\%}$ = [-.196 .216]). Importantly, the difference between these two correlations was significant (z-test = 1.82; p = .03; one-tailed; see Lenhard & Lenhard, 2014 for comparing two correlations from dependent samples).

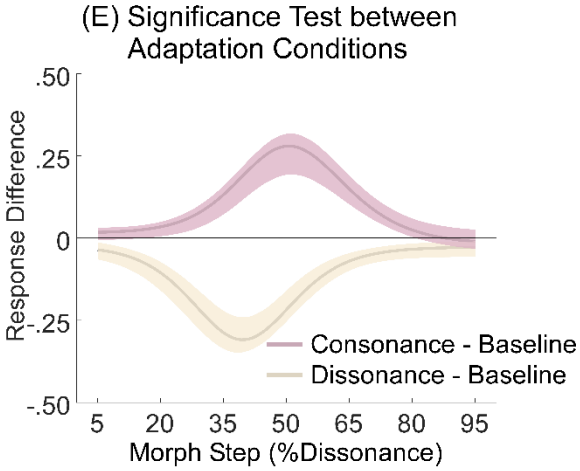
The correlation between music test and consonance aftereffect was not significant (r = -.15; p = .16; $CI_{95\%}$ = [-.345 .058]). The correlation between voice test and identity aftereffect size was significantly larger than the correlation between music test and consonance aftereffect size (z-test = 2.77; p = .003; one-tailed; Lenhard & Lenhard, 2014).

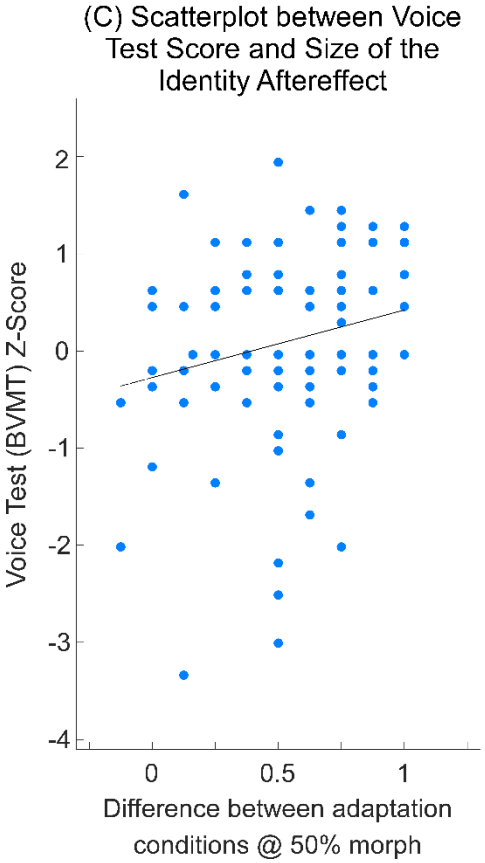
The aftereffect sizes obtained in our two adaptation tasks did not correlate (r = -.06; p = .56; $CI_{95\%}$ = [-.264 .146]) suggesting that adaptability is specific to the sound object. While skills required to successfully complete the BVMT likely overlap to some extent with the skills required to complete the PROMS, the positive correlation did not reach significance in the current sample (Pearson's r = .14; p = .17; $CI_{95\%}$ = [-.063 .340]).


(A) Adaptation to Voice Identity
Mean Responses (%Speaker B)
100
75
50
25
0
5
20
35
50
65
80
95
Morph Step (%Speaker B)
No Adaptation
(Baseline)
Adaptation to:
Speaker A
Speaker B


(D) Adaptation to Consonance
Mean Responses (%Dissonance)
100
75
50
25
0
5
20
35
50
65
80
95
Morph Step (%Dissonance)
No Adaptation
(Baseline)
Adaptation to:
Consonance
Dissonance


(B) Significance Test between
Adaptation Conditions
Response Difference
.50
.25
0
-.25
-.50
5
20
35
50
65
80
95
Morph Step (%Speaker B)
Speaker A - Baseline
Speaker B - Baseline


(E) Significance Test between
Adaptation Conditions
Response Difference
.50
.25
0
-.25
-.50
5
20
35
50
65
80
95
Morph Step (%Dissonance)
Consonance - Baseline
Dissonance - Baseline


(C) Scatterplot between Voice
Test Score and Size of the
Identity Aftereffect
Voice Test (BVMT) Z-Score
2
1
0
-1
-2
-3
-4
0
0.5
1
Difference between adaptation
conditions @ 50% morph

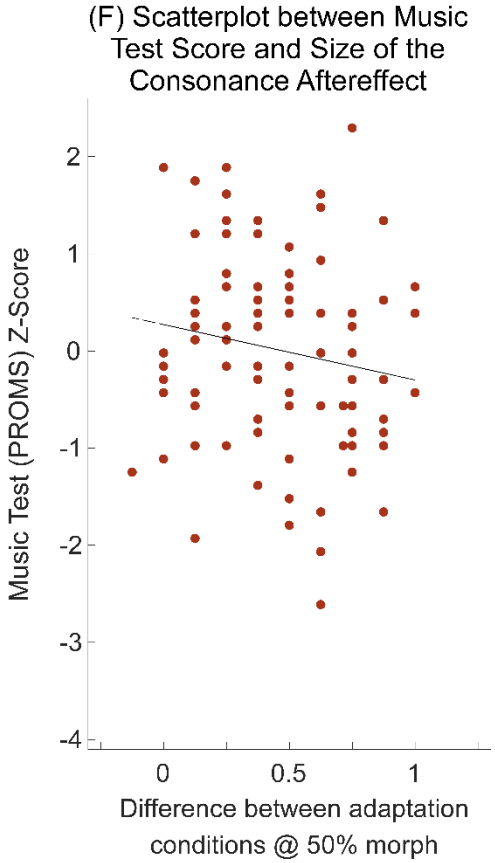

(F) Scatterplot between Music
Test Score and Size of the
Consonance Aftereffect
Music Test (PROMS) Z-Score
2
1
0
-1
-2
-3
-4
0
0.5
1
Difference between adaptation
conditions @ 50% morph

Fig. 2. Results of both tasks in Experiment 2. (A) Average psychophysical functions for the baseline condition (black) and the two identity adaptation conditions (turquoise and purple). Error bars represent SEM and shaded areas illustrate the bootstrapped $CI_{95\%}$ of the fitted curve to participants' mean responses. (B) Mean difference and $CI_{95\%}$ (shaded area) of the differences between baseline and each adaptation condition for the voice task. Significant differences between conditions occurred along the most ambiguous part of the continuum (i.e. where the $CI_{95\%}$ area does not overlap with the y = 0 line). (C) Scatterplot of the identity aftereffect size against the z-scored voice test (BVMT) results (with robust regression line). (D) Average psychophysical functions for the baseline condition (black) and the consonance and dissonance adaptation conditions (red and orange). Error bars represent SEM and shaded areas illustrate the bootstrapped $CI_{95\%}$ of the fitted curve to participants' mean responses. (E) Mean difference and $CI_{95\%}$ (shaded area) of the differences between baseline and each adaptation condition for this control task. Significant differences between conditions occurred along the most ambiguous part of the continuum (i.e. where the $CI_{95\%}$ area does not overlap with the y = 0 line). (F) Scatterplot of the consonance aftereffect size against the z-scored music test (PROMS) results (with robust regression line).

3.3 Reliability analyses of the tasks in Experiment 1 and 2

To assess internal consistency of the adaptation tasks we calculated the Spearman-Brown corrected split-half reliabilities. We computed the coefficient based on means from odd and even splits of trials for the most ambiguous test morph rather than randomly split halves because aftereffects increase with increasing exposure to the adaptor. The reliability of the task in Experiment 1 was acceptable at .72. In Experiment 2, split-half reliabilities were .77 for the identity aftereffect and .62 for the consonance aftereffect. Test-retest reliabilities of the BVMT and PROMS are reported to be high at r = .86 (Mühl et al., 2018) and r = .83 (Zentner & Strauss, 2017), respectively.

4. Discussion

We examined whether individual differences in the quality of adaptive coding of voice identity predict voice discrimination performance. In Experiment 1, voice identity aftereffects were positively linked with performance on the Bangor voice matching test, a standardised measure of voice discrimination ability. Experiment 2 directly replicated this effect. It also demonstrated that this link arises from voice-level processing rather than more basic auditory processing because voice identity aftereffect sizes did not predict performance on a more general auditory cognition test. In addition, there was no relationship between aftereffect sizes to identity and to consonance suggesting that it is not general adaptability that affects performance in voice identity perception.

In both experiments, we found robust contrastive aftereffects following adaptation to voice identity. This effect was strongest around the mathematically and perceptually most ambiguous morph. Our result is in line with previous findings where adaptation to a newly learned voice (Latinus & Belin, 2012) or to a personally familiar voice (Zäske et al., 2010) shifted the perception of ambiguous voices in the opposite direction. While low-level stimulus-dependent influences are difficult to remove completely, several studies have shown that contrastive aftereffects must reflect adaptation to high-level representations of the adapted feature (e.g. identity, emotion, gender). For example, Schweinberger et al. (2008) have shown that adapting to pure tones matched in fundamental frequency to a particular voice gender resulted in no gender aftereffects (see also Bestelmeyer, Rouger, DeBruine and Belin (2010), Pye and Bestelmeyer (2015), and Latinus and Belin (2012) on designs to reduce stimulus-driven adaptation effects of expression and identity). Our experimental design minimised adaptation to low-level features such as fundamental frequency, prosody and speech rate by using different syllables as adaptors and test morphs. Taken together, these findings underline

that adaptation is a universal mechanism in the perception of complex, paralinguistic information.

Our results support the idea that there are individual differences in the quality of voice-space coding and that these contribute to individual differences in voice perception ability. Aftereffects are ubiquitous in perception but the links between adaptive coding and performance seem to be selective to the adapted feature. In other words, voice identity aftereffects did not predict performance on general (i.e. not object-specific) auditory perception skills. This supports the notion of voices as a special auditory object that may be particularly important to us due to its relevance in social interactions.

In both experiments we found that the size of the aftereffects to familiar voice identities significantly predicts performance on a voice matching test consisting of unfamiliar voices. In order to explore voice identity aftereffects, it was necessary to provide voices that were familiar to our participants and use these in an identification task. The BVMT, on the other hand, requires a same/different decision between two unfamiliar voice samples. Previous research has suggested that voice recognition and voice discrimination are dissociable abilities that might depend on different underlying mechanisms and that can be selectively impaired (e.g. van Lancker & Kreiman, 1987; Schelinski, Roswandowitz, & von Kriegstein, 2017). However, with an increased interest in voice perception deficits, a simultaneous impairment of both mechanisms has also been reported (Roswandowitz et al., 2014), though these reports are rarer. Likewise, it has been suggested that the way we process unfamiliar versus familiar voices is different. Kreiman and Sidtis (2013) propose that unfamiliar voice perception relies mostly on the analysis of vocal features whereas familiar voices are perceived as a whole through Gestalt perception. A relationship between tasks that rely on recognition of familiar voices or discrimination of unfamiliar voices, as in our present experiments, might therefore seem surprising.

Previous research supports the notion that voice perception is a distinct auditory ability, and that human voices are a particularly salient auditory stimulus for us from an early age on, even when the voice is unfamiliar (Grossmann, Oberecker, Koch, & Friederici, 2010). As such, it suggests that there are mechanisms predominantly involved in processing human vocalisations over other auditory input (e.g. Belin et al., 2000). This notion could explain our findings of a robust positive relationship between performance on a voice discrimination test, and the size of a voice identity aftereffect, as both involve such voice-specific mechanisms. Furthermore, while Kreiman and Sidtis (2013) propose that two different kinds of analysis (feature vs. Gestalt) underlie unfamiliar and familiar voice perception, respectively, they also pose that they are not mutually exclusive. Feature analysis can still be part of the analysis of familiar voices, just as Gestalt perception can for unfamiliar voices. Our findings in the present experiments support the idea that there is, at least to some degree, an overlap between processes that support the perception of unfamiliar as well as familiar voices. Additionally, both accuracies in familiar and unfamiliar voice perception are affected to a similar degree when the nature of the vocalisation changes, e.g. from spoken word to laughter (Lavan, Scott, & McGettigan, 2016). A recent study into possible cross-modality between superior face recognisers and superior voice recognisers included three voice perception tasks measuring unfamiliar voice discrimination, memory for unfamiliar voices, and recognition of famous voices (Jenkins et al., 2020). In this study, significant albeit small correlations were found between all three voice tests. This finding suggests common processes underlying familiar and unfamiliar voice perception.

Our pattern of results and their effect sizes are in line with results from the face literature. Here, several studies have shown a positive link between adaptive coding of facial identity or facial expression, for example, with recognition performance specific to the adapted attributes (Dennett et al., 2012; Engfors et al., 2017; Palermo et al., 2018; Rhodes et al., 2014; Rhodes et al., 2015). These parallels highlight again the similarities in coding mechanisms of voice and face representations. However, they also highlight that additional factors must account for

the variance in facial and vocal identity perception skill. These additional factors are largely unexplored in voice research. In face research, on the other hand, face perception ability has been shown to be a highly specific skill. Studies with mono- and dizygotic twins have revealed that face perception skills are hereditary, and not linked to more general cognitive abilities like intelligence, memory, or global attention (Wilmer et al., 2010; Zhu et al., 2010). Additionally, there is evidence that face perception ability is largely independent from environmental factors, yet still follows a trajectory of long maturation and later decline across the lifespan (see Wilmer et al., 2017 for review). Judging from the existing literature on voice perception deficits, it seems likely that voice perception ability also does not depend on general intelligence. The individuals reported to have developmental phonagnosia by Roswandowitz and colleagues as well as the first reported case (see Garrido et al., 2009) all completed studies at university level. Research on individual differences in voice perception finds a wide range of ability levels in samples that can be assumed to have a fairly high average IQ throughout (samples consisting of young adults in higher education, as reported in Mühl et al., 2017 and Shilowich & Biederman, 2016), though a systematic exploration using standardised intelligence tests is absent.

While there is converging evidence for a functional role of adaptive coding mechanisms in identity recognition expertise, the reason for this positive link is still speculative. One possible reason is that adaptation calibrates coding mechanisms to the faces and voices we regularly perceive, and the discrimination performance may therefore vary with the proficiency of this calibration process. Adaptation to a given population of faces enhances identification performance of faces from that population suggesting that calibration to match the population prototype helps to maximise sensitivity to change (Rhodes, Watson, Jeffery, & Clifford, 2010; see also Dennett et al. (2012) for an alternative explanation).

There was no relationship between the general auditory abilities test and the voice adaptation or control adaptation tasks. Additionally, we found no relationship between the

aftereffect sizes in our two different tasks. Taken together, these findings may suggest that the aftereffects resulting from the adaptation to consonance or dissonance likely also involve adaptation in the auditory periphery (e.g. Bidelman & Krishnan, 2009), rather than reflecting higher-level coding mechanisms necessary for voice-specific processing.

5. Conclusion

Our results provide support for the notion that individual differences in the quality of voice-space coding exist and that these are linked to a functional role of adaptation in voice identity perception. Our findings also suggest that voices are special auditory objects that are distinct from other auditory stimuli and highlight the similarity in perceptual coding strategies for face and voice identity.

Acknowledgments

Constanze Mühl was funded by a Ph.D. studentship sponsored by the School of Psychology at Bangor University.

Supplementary material

Raw data to this article can be found online at http://dx.doi.org/10.17632/8h9prhx5xp.2.

References

Aglieri, V., Watson, R., Pernet, C., Latinus, M., Garrido, L., & Belin, P. (2017). The Glasgow Voice Memory Test: Assessing the ability to memorize and recognize unfamiliar voices. *Behavior Research Methods, 49*(1), 97-110.

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

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

Bestelmeyer, P. E. G., Latinus, M., Bruckert, L., Crabbe, F., & Belin, P. (2012). Implicitly perceived vocal attractiveness modulates prefrontal cortex activity. *Cerebral Cortex, 22*, 1263-1270.

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

Bidelman, G. M., & Krishnan, A. (2009). Neural Correlates of Consonance, Dissonance, and the Hierarchy of Musical Pitch in the Human Brainstem. *Journal of Neuroscience, 29*(42), 13165-13171.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10*, 433-436.

Burton, A. M., Jenkins, R., & Schweinberger, S. R. (2011). Mental representations of familiar faces. *British Journal of Psychology, 102*, 943-958.

Clifford, C. W. G., Webster, M. A., Stanley, G. B., Stocker, A. A., Kohn, A., Sharpee, T. O., et al. (2007). Visual adaptation: Neural, psychological and computational aspects. *Vision Research, 47*(25), 3125-3131.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.

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.

Engfors, L. M., Jeffery, L., Gignac, G. E., & Palermo, R. (2017). Individual differences in adaptive norm-based coding and holistic coding are associated yet each contributes uniquely to unfamiliar face recognition ability. *Journal of Experimental Psychology: Human Perception and Performance, 43*(2), 281–293.

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

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

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.

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

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. *Quarterly Journal of Experimental Psychology Section a-Human Experimental Psychology, 51*(1), 179-195.

Jenkins, R. E., Tsermentseli, S., Monks, C. P., Robertson, D. J., Stevenage, S. V., Symons, A. E., & Davis, J. P. (2020). Are super-face-recognisers also super-voice-recognisers? Evidence from cross-modal identification tasks. PsyArXiv. https://psyarxiv.com/7xdp3/

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., & Banno, H. (2008). *Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation.* Paper presented at the ICASSP.

Kleiner, M., Brainard, D., & Pelli, D. (2007). *What's new in Psychtoolbox-3?* Paper presented at the ECVP.

Kreiman, J., & Sidtis, D. (2013). Foundations of voice studies: An interdisciplinary approach to voice production and perception. Malden, MA: Wiley-Blackwell.

Latinus, M., & Belin, P. (2011). Anti-voice adaptation suggests prototype-based coding of voice identity. *Frontiers in psychology, 2*, 175-175.

Latinus, M., & Belin, P. (2012). Perceptual Auditory Aftereffects on Voice Identity Using Brief Vowel Stimuli. *Plos One, 7*(7).

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.

Lavan, N., Burston, L. F. K., & Garrido, L. (2019). How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices. *British Journal of Psychology, 110*(3), 576-593.

Lavan, N., Scott, S. K., & McGettigan, C. (2016). Impaired generalization of speaker identity in the perception of familiar and unfamiliar voices. *Journal of Experimental Psychology: General, 145*(12), 1604-1614.

Lenhard, W. & Lenhard, A. (2014). Hypothesis Tests for Comparing Correlations. available: https://www.psychometrica.de/correlation.html. Bibergau (Germany): Psychometrica.

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

Little, A. C., Hancock, P. J. B., DeBruine, L. M., & Jones, B. C. (2012). Adaptation to antifaces and the perception of correct famous identity in an average face. *Frontiers in Psychology, 3*.

Mühl, C., & Bestelmeyer, P. E. G. (2017). Assessing susceptibility to distraction along the vocal processing hierarchy. *Quarterly Journal of Experimental Psychology, 72*(7), 1657-1666.

Mühl, C., Sheil, O., Jarutyte, L., & Bestelmeyer, P. E. G. (2018). The Bangor Voice Matching Test: A standardized test for the assessment of voice perception ability. *Behavior Research Methods, 50*(6), 2184-2192.

Palermo, R., Jeffery, L., Lewandowsky, J., Fiorentini, C., Irons, J. L., Dawel, A., et al. (2018). Adaptive Face Coding Contributes to Individual Differences in Facial Expression Recognition Independently of Affective Factors. *Journal of Experimental Psychology-Human Perception and Performance, 44*(4), 503-517.

Pernet, C. R., Wilcox, R., & Rousselet, G. A. (2013). Robust correlation analyses: false positive and power validation using a new open source matlab toolbox. *Frontiers in Psychology, 3*.

Pye, A. (2015). *The perception of emotion and identity in nonspeech vocalisations.* Unpublished PhD, Bangor University, Bangor.

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

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), 897-903.

Rhodes, G., Pond, S., Burton, N., Kloth, N., Jeffery, L., Bell, J., et al. (2015). How distinct is the coding of face identity and expression? Evidence for some common dimensions in face space. *Cognition, 142*, 123-137.

Rhodes, G., Watson, T. L., Jeffery, L., & Clifford, C. W. G. (2010). Perceptual adaptation helps us identify faces. *Vision Research, 50*(10), 963-968.

Roswandowitz, C., Mathias, S. R., Hintz, F., Kreitewolf, J., Schelinski, S., & von Kriegstein, K. (2014). Two Cases of Selective Developmental Voice-Recognition Impairments. *Current Biology, 24*(19), 2348-2353.

Schelinski, S., Roswandowitz, C., & von Kriegstein, K. (2017). Voice identity processing in autism spectrum disorder. *Autism Research, 10*, 155-168.

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

Schweinberger, S. R., Kawahara, H., Simpson, A. P., Skuk, V. G., & Zäske, R. (2014). Speaker perception. *Wiley Interdisciplinary Reviews-Cognitive Science, 5*(1), 15-25.

Shilowich, B. E., & Biederman, I. (2016). An estimate of the prevalence of developmental phonagnosia. *Brain and Language, 159*, 84-91.

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.

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

Van Lancker, D. R., Kreiman, J., & Cummings, J. (1989). VOICE PERCEPTION DEFICITS NEUROANATOMICAL CORRELATES OF PHONAGNOSIA. *Journal of Clinical and Experimental Neuropsychology, 11*(5), 665-674.

Wainwright, M. J. (1999). Visual adaptation as optimal information transmission. *Vision Research, 39*(23), 3960-3974.

Wark, B., Lundstrom, B. N., & Fairhall, A. (2007). Sensory adaptation. *Current Opinion in Neurobiology, 17*(4), 423-429.

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

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Williams, M., Loken, E., Nakayama, K., & Duchaine, B. (2010). Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of Sciences, 107*(11), 5238-5241.

Wilmer, J. B. (2017). Individual Differences in Face Recognition: A Decade of Discovery. *Current Directions in Psychological Science, 26*(3), 225-230.

Wilcox, R. R. (2012). *Introduction to Robust Estimation and Hypothesis Testing.* (3rd ed.). Oxford: Academic Press.

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.

Zäske, R., Schweinberger, S. R., & Kawahara, H. (2010). Voice aftereffects of adaptation to speaker identity. *Hearing Research, 268*(1-2), 38-45.

Zentner, M., & Strauss, H. (2017). Assessing musical ability quickly and objectively: development and validation of the Short-PROMS and the Mini-PROMS. *Annals of the New York Academy of Sciences, 1400*(1), 33-45.

Zhu, Q., Song, Y., Hu, S., Li, X., Tian, M., Zhen, Z., Dong, Q., Kanwisher, N, & Liu, J. (2010). Heritability of the specific cognitive ability of face perception. *Current Biology, 20*(2), 137-142.