Matching novel face and voice identity using static and dynamic facial images

Harriet M. J. Smith

Thesis submitted in partial fulfilment of the requirements of Nottingham Trent University for the degree of Doctor of Philosophy

May 2016

ABSTRACT

Research suggests that both static and dynamic faces share identity information with voices. However, face-voice matching studies offer contradictory results. Accurate face-voice matching is consistently above chance when facial stimuli are dynamic, but not when facial stimuli are static. This thesis aims to account for previous inconsistencies, comparing accuracy across a variety of two-alternative forced-choice (2AFC) procedures to isolate the features that support accuracy. In addition, the thesis provides a clearer and more complete picture of face-voice matching ability than that available in the existing literature. Same-different procedures are used to address original research questions relating to response bias and the delay between face and voice presentation.

The overall findings indicate that faces and voices offer concordant source identity information. When faces and voices are presented close together in time, matching accuracy is consistently above chance level using both dynamic and static facial stimuli. Previous contradictory findings across studies can be accounted for by procedural differences and the characteristics of specific stimulus sets. Multilevel modelling analyses show that some people look and sound more similar than others. The results also indicate that when there is only a short (~1 second) interval between faces and voices, people exhibit a bias to assume that they belong to the same person.

The findings presented in this thesis have theoretical and applied relevance. They highlight the value of considering person perception from a multimodal point of view, and are consistent with evidence for the existence of early perceptual integrative mechanisms between face and voice processing pathways. The results also offer insights into how people successfully navigate complex social situations featuring a number of novel speakers.

ACKNOWLEDGEMENTS

For making my PhD experience so rewarding, stimulating, and interesting I cannot possibly thank my SUPER supervision team enough. I owe a great deal to Dr. Andrew Dunn for helping me out of a difficult time, sweeping me up in his infectious enthusiasm for faces and voices, and constantly challenging me to be the best that I could be. He has taught me some extremely valuable lessons, not all of which have been work-related, but all of which I am sure I will remember forever. (I cannot promise I will always include definite articles though.) I have also benefitted hugely from the help of 'serious statistician', Professor Thom Baguley, who has patiently dedicated many hours to helping me get to grips with multilevel modelling. Countless times, Dr. Paula Stacey's magical ability to boil down my complicated and confused ideas into more logical and simple ones has lent coherence to both my writing and the contents of my head. I feel extremely lucky to have had the guidance and support of this fantastic team, not only because of the influence they have had on my PhD, but also because of the way that they have shaped the knowledge, opinions, and approaches I will carry forward into my academic career.

I have met some amazing people at NTU; people who have provided moral support along this journey, and made it so much more enjoyable by helping me fuel up with alcohol, sugar and caffeine. On a related note, I must also thank technicians Ben, Roy and Josh for resurrecting my laptop after I accidentally deposited an entire cup of coffee on it.

None of what I have achieved over the past few years would have been possible without the love and support of my parents, Janet and Jollyon. It would be impossible for me to explain how grateful I am for everything they have done. Their belief in me has never faltered, and their patience seems never-ending. This thesis is dedicated to them.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. CHAPTER 1: OVERVIEW OF THE THESIS

This thesis investigates whether people look like they sound, and sound like they look. It reports experiments testing whether participants can match face and voice stimuli for identity. An overview of the contents of each chapter is provided below.

1.1 Chapter 2: Literature review

Chapter 2 places the thesis within the wider context of existing literature. The chapter begins by reviewing evidence for integrated face and voice processing, which provides an important theoretical foundation for the studies presented in this thesis. The chapter then considers the extent to which faces and voices share redundant information, focusing on dynamic faces and voices in light of speech perception research, and static faces and voices in light of the evolutionary psychology literature. The findings provide a rationale for directly comparing dynamic to static face-voice matching. Despite the two strands of research independently informing hypotheses that both types of matching should be possible, the subsequent review of the relevant literature shows that whilst dynamic face-voice matching is consistently above chance level, static face-voice matching is more variable. There are a number of procedural confounds between studies that might help to explain differences in performance. In discussing the existing face-voice matching research in detail, a number of gaps in knowledge emerge. On the basis of these, the following research questions are formulated:

- Research question 1: Do voices share redundant information with dynamic as well as static faces?
- Research question 2: Is it possible to match voices and static faces, or is accurate face-voice matching contingent on encoding information about visual articulatory patterns?

- Research question 3: Do procedural differences account for inconsistencies in the previous literature regarding static face-voice matching?
- Research question 4: Are there matching performance asymmetries according to the order of stimulus presentation?
- Research question 5: How do response biases operate in face-voice matching?

1.2 Chapter 3: Face and voice stimuli: Methodological and statistical issues

Chapter 3 explains the methodological and statistical challenges of appropriately dealing with inter and intra stimulus variation in faces and voices. The chapter shows that in order to generalise from stimuli, it is necessary to use a stimulus sample that features as many individuals as possible, and always more than one. More importantly, appropriate statistical analyses should be employed. This chapter therefore provides a rationale for the use of multilevel modelling. Whilst conventional statistical analyses such as ANOVA aggregate over stimuli or items, only approaches such as multilevel modelling simultaneously account for participant and stimulus variability.

1.3 Chapter 4: Testing the back-up signal hypothesis: Do faces and voices offer redundant information?

Chapter 4 reports an experiment addressing whether voices share redundant information with dynamic as well as static faces (Research question 1). The evolutionary psychology literature suggests that together, faces and voices provide multimodal signals for dimensions of fitness and quality. In Experiment 1 we tested whether this information is complementary or redundant. Participants rated faces and voices on scales for masculinity/femininity, age, health, height and weight. The results show that independent ratings of the same person’s face and voice are strongly related, regardless of whether the face is static or dynamic. Faces and voices therefore appear to offer redundant rather than complementary information. This evidence is used to inform the hypothesis that both static and dynamic face-voice matching should be possible.

1.4 Chapter 5: Matching novel face and voice identity using two-alternative forced-choice procedures

Chapter 5 reports three experiments investigating static face-voice matching and dynamic face-voice matching performance using two-alternative forced-choice (2AFC) procedures. In each experiment, participants had to decide which face-voice combination was made up of a single identity. Experiments 2a, 2b and 2c addressed whether static face-voice matching is possible (Research question 2). The experiments employed different versions of 2AFC tasks in order to establish whether contradictions across previous studies might be accounted for by procedural differences (Research question 3). Experiments 2a and 2b also included a manipulation of stimulus presentation order to investigate the possibility that performance differs according to whether the face is seen before the voice is heard, or vice-versa (Research question 4). Taken together, the results suggest that above chance static face-voice matching is possible, although it is sensitive to the experimental procedure employed. In addition, inconsistencies in previous research might depend on the specific stimulus set used; multilevel modelling reveals that some people look and sound more similar than others.

1.5 Chapter 6: Position bias in two-alternative forced-choice procedures

In Experiments 2a and 2b, participants were more accurate when the same identity stimulus appeared in position 1, compared to position 2, of a sequential 2AFC task. The two experiments presented in Chapter 6 attempted to account for this position effect by testing for the existence of a response bias (Research question 5). In Experiments 3a and 3b, the same identity stimulus was never present at test. The overall pattern of matching responses was consistent with the conclusion that sequential 2AFC face-voice matching tasks are inherently biased; people are more likely to select the stimulus appearing in position 1. This may reflect a tendency to integrate a face and voice presented close together in time as belonging to the same person.

1.6 Chapter 7: Matching novel face and voice identity using same-different procedures

As an alternative to the biased 2AFC task, Chapter 7 tests static face-voice matching and dynamic face-voice matching (Research question 2) using a same-different procedure. This procedure includes both signal and noise trials, so facilitates investigation of how both sensitivity and response biases operate in face-voice matching (Research question 5), and whether biases differ according to stimulus order, i.e. whether the face is presented before the voice, or the voice is presented before the face (Research question 4). In Experiments 4a and 4b, participants saw a face and heard a voice. They had to decide whether the face and voice belonged to the same person. On signal trials, the correct response was *same identity*, whereas on noise trials the correct response was *different identity*. The results replicate the results presented in Chapter 4, showing that both static and dynamic face-voice matching is possible. In both experiments participants exhibited a bias to respond that the face and voice in each trial belonged to the same person. This bias was stronger when the face was presented before the voice. This finding is discussed in light of voices providing weaker identity cues than faces; voices perhaps tend to be subsumed by the identity of preceding faces.

1.7 Chapter 8: The effect of increasing the inter-stimulus interval on face-voice matching performance

The experiments presented in Chapter 8 extend the existing literature, as well as the new findings presented in previous chapters, by investigating how face-voice matching performance operates when faces and voices are separated by an inter-stimulus interval of 5 seconds (Experiments 5a and 5c) or 10 seconds (Experiments 5b and 5d). In order to investigate the effect on accuracy (Research question 2) and response bias (Research question 5), these experiments used a same-different procedure, and manipulated the order of stimulus presentation (Research question 4). The results show that as the inter-stimulus interval increases, people are less likely to be able to accurately match face and voice identity. Accurate matching appears to depend on being able to compare high quality visual and auditory perceptual representations for identity information. The bias to respond *same identity* also weakens as the interval increases, suggesting that the bias observed in previous experiments is related to temporal contiguity. Integrating the face and voice into a single multimodal signal appears to be more challenging when they become temporally separated.

1.8 Chapter 9: Summary and general discussion

Chapter 9 summarises the main findings of the 12 experiments comprising this thesis, highlighting how they constitute an original contribution to the literature. The results are discussed in reference to the five research questions outlined above. The chapter offers some recommendations for future research, and comments on the applied relevance of the findings.

2 CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

Both faces and voices are highly salient social stimuli, thought to signal important and related information in order to facilitate social communication (Belin, Bestelmeyer, Latinus & Watson, 2011; Belin, Fecteau & Bedard, 2004; Campanella & Belin, 2007; Schweinberger, Kawahara, Simpson, Skuk & Zäske, 2014; Stevenage & Neil, 2014). During social interactions, faces and voices tend to be perceived simultaneously. However, the extent to which faces and voices offer concordant source identity information is relatively under-researched (Wells, Baguley, Sergeant & Dunn, 2013; Wells, Dunn, Sergeant & Davies, 2009). Testing whether novel faces and voices can be accurately matched provides a measure of the extent to which they offer redundant information. However, the literature has not resolved uncertainty regarding the extent to which accurate face-voice matching is contingent on encoding visual articulatory patterns and linking these to the sound of a voice (Kamachi, Hill, Lander & Vatikiotis-Bateson, 2003; Lachs & Pisoni, 2004a). Some studies show that there is sufficient redundant information available in static faces and voices to facilitate novel face-voice matching (Krauss, Freyberg & Morsella, 2002; Mavica & Barenholtz, 2013).

This literature review discusses pertinent research investigating the concordance of source identity information offered by faces and voices. The chapter will review recent theories of face-voice processing before outlining the role of audiovisual information in speech perception, a research area highlighting redundancies between voices and dynamic articulating faces. The review will then turn to the evolutionary psychology literature, which has considered whether static faces and voices communicate similar information about dimensions of fitness and quality. In light of these two distinct strands of research, current literature regarding static and dynamic face-voice matching will be addressed to build hypotheses for the experiments featured in future chapters. The hypotheses will also be informed by the relevant methodological literature. The following discussion of the literature therefore serves as a framework on which the research questions in this thesis have been formulated.

2.2 Models of face and voice perception: Independent or integrated processes?

It is necessary to address how face-voice matching ability is accommodated within existing models of person perception. The following section explains how the cognitive architecture supports face and voice processing, and considers whether cognitive models conceive of the two pathways as being either independent or integrated.

The existence of highly selective cortical face and voice regions underlines their central and basic importance in supporting everyday social functioning (Yovel & Belin, 2013). Functional magnetic resonance imaging (fMRI) studies have identified a number of cortical areas responding selectively to faces, such as the fusiform face area (FFA), occipital face area (OFA), and the right posterior superior temporal sulcus (STS) (e.g. Chao, Martin & Haxby, 1999; Grill-Spector, Knouf & Kanwisher, 2004; Hoffman & Haxby, 2000; Pitcher, Walsh, Yovel & Duchaine, 2007). Similarly, fMRI evidence indicates the existence of voice-specific regions, or temporal voice areas (TVAs). These are located in the superior temporal gyrus (STG) (e.g. Ahrens, Hasan, Giordano & Belin, 2014; Belin, Zatorre & Ahad, 2002; Charest, Pernet, Latinus, Crabbe & Belin, 2013; Ethofer et al., 2013; von Kriegstein, Eger, Kleinschmidt & Giraud, 2003).

The majority of the literature investigating paralinguistic aspects of face and voice perception has traditionally regarded face and voice processing as occurring relatively independently of each other (Belin et al., 2004, 2011). The well-known computational Interactive Activation and Competition (IAC) model (Burton, Bruce & Johnston, 1990) predicts that following the structural encoding of face and voice representations, identity information about familiar people converges from different modalities at the post-perceptual stage of person identity nodes (PINs) (Ellis, Jones & Mosdell, 1997). However, an increasing body of recent behavioural evidence suggests that face and voice processing are not totally independent until this late stage (Joassin, Pesenti, Maurage, Verreckt, Bruyer & Campanella, 2011; Schweinberger, Herholz & Stief, 1997; Sheffert & Olson, 2004; Zäske, Schweinberger & Kawahara, 2010). The literature now tends to consider person perception from a more multimodal perspective (Schweinberger et al., 2014). This supports Belin et al.'s (2004) adaptation of Bruce and Young's (1986) model of face perception, which includes the addition of a voice processing pathway.

The auditory face model (Belin et al., 2004) proposes that face and voice processing occur in parallel integrated pathways to facilitate the efficient exploitation of redundant information (Belin, Bestelmeyer, Latinus & Watson, 2011). Organisation of the functional architecture of voice perception is similar to Bruce and Young (1986)'s conception of face processing. After being processed for structural analysis, both faces and voices are processed for information about speech, emotion, and identity. However, according to this model, the three parallel visual and auditory pathways also interact with each other (Belin et al., 2004, 2011; Campanella & Belin, 2007; Stevenage & Neil, 2014). The auditory face model is supported by brain imaging evidence. As well as the existence of cortical areas selective to faces and voices, a number of brain areas have been identified as possible loci for supramodal, multimodal person perception. These include the amygdala, STS and superior colliculus (see Belin et al., 2011). Further support for the model is offered by imaging studies indicating crosstalk and functional connections between selective face and voice areas (Blank, Anwander & von Kriegstein, 2011; von Kriegstein, Kleinschmidt, Sterzer & Giraud, 2005).

Achieving accuracy in a face-voice matching task would rely on the successful extraction of redundant multimodal information. Accurate matching of voices to both dynamic articulating faces, as well as static faces, sits well with Belin et al.'s (2004) model. Whilst viewing dynamic articulating faces might facilitate the exploitation of redundant speech information, viewing static faces should be sufficient for the extraction of redundant identity information. The auditory face model therefore hypothetically supports a role for both static and dynamic information in explaining accurate novel face-voice matching. It is rather more difficult to reconcile the possibility of accurate static face-voice matching with Burton et al.'s (1990) model. There is little provision for the existence of redundant multimodal identity information in a model requiring familiarisation before a modality-free representation (i.e. PIN) can be activated.

2.3 Audiovisual speech perception

The majority of studies investigating audiovisual face-voice processing have focused on speech perception (Yovel & Belin, 2013). This research is relevant to understanding the nature of redundancies between voices and dynamic faces.

The results of audiovisual speech perception research highlight the existence of links between auditory and visual modalities. As originally shown by Sumby and Pollack (1954), the visual perception of a speaker's face improves speech intelligibility in noisy conditions (Benoit, Mohamadi & Kandel, 1994; MacLeod & Summerfield, 1987; Summerfield, 1987; see Rosenblum, 2005 for a review). McGurk and MacDonald's seminal research demonstrates the existence of automatic perceptual links between voices and dynamic faces during speech perception (MacDonald & McGurk, 1978; McGurk & MacDonald, 1976). When phonetic information from a speaker's face (e.g. [ga]) and voice (e.g. [ba]) is discrepant, the information is fused, and perceived as something that did not occur in either modality (e.g. [da] or [tha]). This is known as the McGurk effect. Similar results have been observed under a variety of conditions (see Massaro, 1998).

More recent evidence for the integration of face and voice information in speech perception comes from brain imaging studies. The same brain areas respond to both visual and auditory speech cues (e.g. Besle, Fischer, Bidet-Caudelet, Lecaignard, Bertrand & Giard, 2008; Miller & D'Esposito, 2005; Paulesu et al., 2003; Skipper, Van Wassenhove, Nusbaum & Small, 2007). For example, silent lip-reading activates cortical areas, which were previously believed to respond selectively to the sound of a voice (Calvert et al., 1997). Findings such as these may help to explain why the benefit of increased familiarity with a speaker in one modality (auditory or visual) transfers to the other modality. Familiarity with a person's voice improves people's ability to lip-read that person's silent speech, and vice-versa (Rosenblum, Miller & Sanchez, 2007; Sanchez, Dias & Rosenblum, 2013). Some researchers therefore argue that speech perception is better understood as an amodal process, which is blind to the specific modality input because auditory and visual information are functionally inseparable (e.g. Rosenblum, 2005; Rosenblum, 2008). Other researchers argue for independent face and voice processing and late-integration (e.g. Bernstein, Auer & Takayanagi, 2004; Braida, 1991; Massaro, 1987, 1998).

An argument against the latter position is that speech has closely related auditory and visual characteristics (Lachs & Pisoni, 2004a). Idiosyncratic speaking styles dictate both what voices sound like, and how faces move (Dohen, Loevenbruck, Cathiard & Schwartz, 2004; Lander, Hill, Kamachi, & Vatikiotis-Bateson, 2007; Yehia, Rubin, & Vatikiotis-Bateson, 1998). An illustration of the voice production process provides a key to understanding how and why visual and auditory speech are so closely related. According to the source-filter model (Fant, 1960), voices are produced by vibrations in the vocal chords, which are situated in the larynx. These vibrations modulate airflow from the respiratory system so that acoustic energy can be filtered through the vocal tract. Articulators in the vocal tract, such as the tongue, teeth, and soft palate, work in combination with the vocal folds to produce the necessarily wide and intricate array of sounds involved in human speech (Fitch, 2000; Jenkins, 1998; Titze, 1994a). The idiosyncratic neuromuscular movement of the internal vocal apparatus is redundantly reflected in both the face and the voice (Yehia, Kuratate & Vatikiotis-Bateson, 2000). The cheeks and lips constitute the outer surface of the vocal tract. The production of speech involves not only the movement of these features, but also the jaw (Vatikiotis-Bateson, Munhall, Hirayama, Lee & Terzopoulos, 1996). For example, Yehia et al. (1998) showed that tongue movement is closely related to jaw movement during speech. A number of visual and auditory speech correlates have also been observed. For example, the fundamental frequency of the vocal fold vibration is related to head position and orientation (Yehia, Kuratate & Vatikiotis-Bateson, 2002).

There is strong evidence for perceptual links between voices and dynamic articulating faces. The evidence reviewed above suggests that, in line with predictions made by Belin et al.'s (2004) auditory face model, redundant information is offered by the visual and auditory modalities during speech production.

2.4 Multimodal signals in faces and voices: Back-up signals or multiple messages?

Faces and voices transmit far more identity-specific information than speech alone (Campanella & Belin, 2007; Yovel & Belin, 2013). Both convey information about a number of other dimensions, including gender, personality and emotion (e.g. Belin et al., 2004; de Gelder & Vroomen, 2000; Dolan, Morris & de Gelder, 2001; Massaro & Egan, 1996; Mavica & Barenholtz, 2013; Warner & Sugarman, 1986). This section focuses on the evolutionary psychology literature in order to explore the possibility that static faces and voices offer redundant information about a number of relatively stable dimensions, which might help to indicate common source identity. The evolutionary literature, which has primarily dealt with ratings of static rather than dynamic faces and voices, suggests that both types of stimulus offer reliable and related information about mate value and fitness (Collins & Missing, 2003; Feinberg, 2008; Feinberg et al., 2005; Fraccaro, Feinberg, DeBruine, Little, Watkins & Jones, 2010; Pisanski, Mishra & Rendall, 2012).

Together, faces and voices convey multimodal signals. Such signals are common in animals, and occur when information about an underlying trait is communicated by more than one modality. As most research has focused on face and voice ratings independently of each other (Wells et al., 2009; Wells et al., 2013), relatively little is known about multimodal signals in humans. Multimodal signals are either back-up signals (Johnstone, 1997), or multiple messages (Møller & Pomiankowski, 1993), and are likely to have adaptive value in terms of mate choice. Back-up signals are redundant in meaning: they offer similar information, and elicit the same response, thereby helping to reduce inaccurate trait assessments (Møller & Pomiankowski, 1993).

An example of a back-up signal in the animal world can be observed in male wolf spiders (*schizocosa ocreata*), which combine seismic and visual aspects in courtship displays (Uetz & Roberts, 2002). These signals provoke the same response from female wolf spiders when presented in isolation as they do when presented together (Uetz, Roberts & Taylor, 2009). Multiple messages on the other hand offer complementary information and prompt different responses (see Partan & Marler, 1999). Taken together, multimodal information can offer a fuller assessment of mate quality (Candolin, 2003). For example, in zebra finches (*Taeniopygia guttata*) song rate and beak colour are condition-dependent sexual signals (Zann, 1996). However, when Birkhead, Fletcher and Pellatt (1998) manipulated diet quality in the laboratory, these two signals reacted at different rates to a poor seed-only diet compared to a supplemented diet. They concluded that whilst song rate indicates present condition, beak colour is indicative of longer-term condition (Candolin, 2003).

From an evolutionary perspective, faces and voices provide valuable clues about fitness. For example, in terms of attractiveness they appear to constitute reliable and concordant signals of genetic quality (e.g. Abend, Pflüger, Koppensteiner, Coquerelle & Grammer, 2015; Collins & Missing 2003; Feinberg, 2008; Feinberg et al., 2005; Fraccaro, Feinberg, DeBruine, Little, Watkins & Jones, 2010; Saxton, Caryl & Roberts, 2006; Thornhill & Gangestad, 1999; Thornhill & Grammer, 1999; Wheatley et al., 2014; Zahavi & Zahavi 1997; see also Puts, Jones & DeBruine, 2012 for a review). A number of studies have found that people who have faces that rate highly for attractiveness also tend to have voices that rate highly for attractiveness (e.g. Collins & Missing, 2003; Saxton et al., 2006, but see Oguchi & Kikuchi, 1997; Rezlescu, Penton, Walsh, Tsujimura, Scott & Banissey, 2015; Wells et al., 2013). Making similar judgements about a person regardless of whether you see their face or hear their voice might help to indicate common source identity. With the exception of the attractiveness literature, previous research has rarely compared judgements made from faces and voices, focusing instead on judgements informed by a single modality (e.g. Neiman & Applegate, 1990; Penton-Voak & Chen, 2004; Perrett et al., 1998; Pisanski et al., 2012). There are a number of reasons why we may expect concordance between face-voice ratings in terms of masculinity and femininity, health, age, height, and weight. Some of these reasons are addressed below.

2.4.1 Masculinity/femininity

Levels of reproductive hormones are likely to inform perceptions of both facial and vocal femininity and masculinity. For example, testosterone increases the size and thickness of vocal folds (Beckford, Rood & Schaid, 1985), resulting in lower fundamental frequency

(Fant, 1960), which influences perceptions of masculinity (Pisanski et al., 2012). In addition, high levels of testosterone are associated with characteristics of facial masculinity (Penton-Voak & Chen, 2004; Perrett et al., 1998), such as larger jaws, chins and noses (Miller & Todd, 1998). In women, oestrogen slows down vocal fold development, and is associated with higher vocal pitch (Abitbol, Abitbol & Abitbol, 1999; O'Connor, Re & Feinberg, 2011). Oestrogen levels are also related to markers of facial femininity (Thornhill & Grammer, 1999) such as larger lips, smaller lower faces, and fat deposits on the upper cheeks (Perrett et al., 1998).

2.4.2 Health

We might also expect ratings of health made from faces and voices to be similar. According to the handicap principle (Zahavi & Zahavi, 1997), masculine males and feminine females are perceived as high quality. This is because they are able to bear the immunocompetence handicap associated with high levels of reproductive hormones, imposed because metabolising hormones draws resources away from other bodily functions (Folstad & Karter, 1992). Previous research suggests that cues relating to higher levels of reproductive hormones are reliable indicators of fitness and quality (Folstad & Karter, 1992; Thornhill & Gangestad, 2006; Zahavi & Zahavi, 1997). Indeed, some studies suggest that measures of sexual dimorphism are linked to health ratings and actual health in both men and women (Ellison, 1999; Gray, Berlin, Law Smith et al., 2006; McKinlay & Longcope, 1991; Rhodes, Chan, Zebrowitz & Simmons, 2003). For example, the self-reported incidence and duration of respiratory disease is negatively associated with measures of sexual dimorphism (Thornhill & Gangestad, 2006). Medical health in males has been linked to ratings of facial masculinity (Rhodes, Chan, Zebrowitz & Simmons, 2003) and actual testosterone levels (Gray et al., 1991). In women, higher levels of reproductive hormones reflect reproductive health, such as the increased chance of successful conception (Ellison, 1999).

2.4.3 Age

Faces and voices index information about biological age, a cue that is relevant to reproductive fitness in both males and females (Thornhill & Gangestad, 1999). Numerous visual markers act as indicators of older age, such as decreased skin elasticity, wrinkles, discolouration and reduced clarity in skin tone (Burt & Perrett, 1995). In terms of voices, older people speak with a slower speech rate (Linville, 1996), and age-related hormonal changes affect pitch. For example, female voice pitch lowers after the menopause, whereas older male voices become higher-pitched with increasing age (Linville, 1996). People can estimate a speaker's age from their voice relatively accurately (to within about 10 years) (Braun, 1996; Neiman & Applegate, 1990; Ptacek & Sander, 1966; Smith & Baguley, 2014).

2.4.4 Height and weight

Body size is a further indicator of quality (Collins & Missing, 2003; Thornhill & Gangestad, 1999). However, although people tend to agree about height and weight judgements made from a voice (Collins, 2000), this does not indicate that they are necessarily accurate (Bruckert, Liénard, Lacroix, Kreutzer & Leboucher, 2006; Collins, 2000; van Dommelen & Moxness, 1995). Despite the apparent inaccuracy of height judgements made from voices, people judge height from faces with relative accuracy (Schneider, Hecht, Stevanov & Carbon, 2013), using cues such as facial elongation. People with longer faces are judged as being taller (Re et al., 2013). Judgements from faces are also accurate for weight estimates (Coetzee, Chen, Perrett & Stephen, 2010). Lass and Colt (1980) compared visual and auditory height and weight ratings. The results indicated significant differences between weight ratings from female faces and voices, suggesting that for some characteristics, faces and voices may not offer concordant information. Recent research has not addressed the extent of concordance between body size information offered by faces and voices. Although

Krauss et al. (2002) asked participants to rate the age, height and weight of speakers from faces and voices, they only measured accuracy, rather than the relationship between the two sets of ratings.

2.4.5 The influence of pitch variability on face-voice concordance

Pitch is the most perceptually salient characteristic of the human voice (Banse & Scherer, 1996). The research reviewed above suggests that voice pitch is likely to play an important role in explaining the positive relationship between ratings of static faces and voices. It likely plays a role in informing ratings of masculinity/femininity, health, age, height and weight. However, although voice pitch is a physiologically determined sexually dimorphic characteristic (Abitbol et al., 1999; Dabbs & Mallinger 1999; Hollien, 1960), it is not fixed (Titze, 1994b), and is influenced by muscular settings (Abercrombie, 1967).

Cultural differences in voice production dictate different average voice pitches across countries. Japanese women speak with relatively higher pitched voices than women from Western cultures in order to transmit cultural ideals such as modesty and politeness, traditionally associated with femininity (Loveday, 1981; van Bezooijen, 1995). Evidence also shows that both males and females modulate their pitch according to social situations (e.g. Gregory, 1996; Hughes, Farley & Rhodes, 2010; Falk, 2005; Farley, Hughes & LaFayette, 2013; Leongómeza et al., 2014). When competing against a man they perceive to be physically 'weaker' than themselves, men modulate their voice pitch downwards. If they perceive their competitor to be more dominant, men modulate their voice pitch upwards (Puts, Gaulin & Verdolini, 2006). Through intra-sexual pressure, male voice pitch has developed as a signal for aggression, dominance and position within a social hierarchy (Hodges-Simeon, Gurven, Puts & Gaulin, 2014; Puts, Apicella & Cárdenas, 2012; Puts et al., 2006; Puts, Hodges, Cárdenas & Gaulin, 2007). Women also modulate according to social context, speaking with a higher-pitched voice, for example, if they find a man attractive (Fraccaro et al., 2011). The notable variability of voice pitch may therefore reduce the extent to which faces and voices offer redundant information overall, perhaps making it more difficult to accurately attribute common source identity.

2.4.6 Information concordance in faces and voices: The story so far

Faces and voices are highly complex social stimuli. Both offer a wealth of socially relevant information about emotion, personality and the content of speech (Yovel & Belin, 2013), which may not all necessarily be concordant (Campanella & Belin, 2007; Rezlescu et al., 2015). Added to this, the relatively fluid nature of pitch means that the relationship between ratings of faces and voices is unlikely to be perfect. Nevertheless, taken together, the research outlined above suggests that static faces and voices offer at least some redundant information. However, apart from literature addressing the relationship between facial and vocal attractiveness, the extent to which faces and voices communicate similar or overlapping information has seldom been tested.

2.5 Static vs. dynamic facial information

2.5.1 Ratings

Aside from the importance of dynamic facial images in providing information about articulatory movement (see section 2.3), in comparison to static faces, dynamic faces also offer extra information about emotion (Chiller-Glaus, Schwaninger, Hofer, Kleiner & Knappmeyer, 2011) and 3-D facial shape (O'Toole, Roark & Abdi, 2002). If dynamic faces communicate additional information compared to static faces, facial stimulus type may also affect the extent to which face and voice information is concordant. This could in turn influence the accuracy of face-voice matching. In one of the only studies to address this question, Lander (2008) found that male face and voice attractiveness were only related when faces were dynamic.

Recent evolutionary psychology literature investigating mate value and attractiveness has compared ratings of static and dynamic faces. Most studies have used static facial stimuli (photos) (e.g. Coetzee et al., 2010; Main, DeBruine, Little & Jones, 2010; Scott & Penton-Voak, 2011), but in everyday social situations we encounter people in motion. There has been a recent move amongst face researchers to use dynamic facial stimuli (videos) in order to improve ecological validity (Gangestad & Scheyd, 2005; Kościński, 2013; Penton-Voak & Chang, 2008; Roberts et al., 2009b). Some studies have found that facial stimulus type (static or dynamic) influences attractiveness judgements, although the overall results are somewhat mixed. Rubenstein's (2005) results indicate that attractiveness judgements of females made from static and dynamic images are not strongly or significantly correlated. Other studies have observed significant correlations between static and dynamic images of female faces but not male faces (Lander, 2008; Penton-Voak & Chang, 2008), whilst Roberts et al. (2009a) detected significant correlations using male images. In reviewing previous studies, and investigating methodological differences between them, Roberts et al. (2009b) found that correlations between ratings from static and dynamic facial stimuli were stronger when rated by the same participants, likely because of carryover effects.

If judgements from faces are immediate (<100ms), automatic, and robust (Rhodes et al., 2011; Willis & Todorov, 2006), the extra information from time-varying dynamic cues should not be particularly influential. In other words, because we reach judgments so soon after the initial presentation of a face, the judgments based on static and dynamic faces are likely to be similar. However, as patterns of facial movement vary according to sex (Morrison, Gralewski, Campbell & Penton-Voak, 2007), viewing dynamic images might be more likely to lead to the revision of initial judgments; it is conceivable that masculinity/femininity ratings will be more extreme when viewing dynamic faces. In a recent study, Kościński (2013) found strong correlations between attractiveness ratings of static and dynamic faces. In this experiment, ratings of femininity were also taken, but influenced attractiveness ratings similarly in both conditions.

Research comparing face ratings has exclusively concentrated on attractiveness, rather than considering how static or dynamic faces might influence assessments of age, height or weight etc. Overall, the existing literature reflects diverging evidence regarding the influence of dynamic information. The evidence is not sufficiently conclusive to inform a strong prediction about the influence of facial stimulus type on face-voice information concordance.

2.5.2 Matching novel face and voice identity

Drawing on hypotheses from both the audiovisual speech perception literature and the evolutionary psychology literature, a number of recent studies have used face-voice matching as a measure of crossmodal redundancy. Taking the existing literature together as a whole, it is unclear whether accurate face-voice matching depends on encoding dynamic visual information about articulatory patterns, or whether sufficient redundant information is available in static faces.

2.5.2.1 Dynamic visual information facilitates face-voice matching

Audiovisual speech perception research demonstrates that participants can match sequentially presented dynamic images of articulating faces to the voice of the same speaker. In Lachs and Pisoni's (2004a) experiment, people accurately matched the visual component of the word 'cat' to the auditory component above-chance level. These results have been replicated using repetitions of full sentences (Rosenblum, 2002).

Identity matching is not totally dependent on overlapping linguistic content. Using the same procedure as Lachs and Pisoni (2004a), Kamachi et al. (2003) ran experiments in which the face and voice in each trial said similar sentences, as well as separate sets of trials in which they said either identical or very different sentences. Although performance was marginally better when there was some linguistic overlap (Kamachi et al., 2003), voices and dynamic faces can still be matched at above chance level when the voice says a completely different sentence to the face (Kamachi et al., 2003; Lander et al., 2007). Indeed, overlap in terms of the manner in which a sentence is spoken appears to be more important than the content (Lander et al., 2007).

Both Lachs and Pisoni (2004a) and Kamachi et al. (2003) ran separate matching experiments using static faces and voices to test the hypothesis that crossmodal source identity information is contingent on encoding dynamic visual articulatory patterns. In both studies static face-voice matching performance was at chance level (Kamachi et al., 2003; Lachs & Pisoni, 2004a). The apparent importance of time-varying articulatory information is underlined by the fact that participants can match faces and voices using movement information alone. Studies isolating articulatory movement using a point-light technique observe accurate matching of auditory utterances to dynamic displays (Lachs & Pisoni, 2004b; Rosenblum, Smith, Nichols, Hale & Lee, 2006).

2.5.2.2 Static visual information facilitates face-voice matching

Some research challenges the conclusion that dynamic visual information is crucial to crossmodal matching. Krauss et al. (2002) showed that people could match a voice to a static image with above chance accuracy. Participants heard a recording of a voice saying a 7-syllable sentence. After 1 second, they were presented with two simultaneously presented full-length static photographs (a target of the same identity and a distractor of a different identity), and asked to decide which photograph featured the speaker. However, whilst the studies observing chance level matching performance with static faces and voices used stimuli of the same sex and a similar age and ethnicity in each trial (e.g., Kamachi et al., 2003; Lachs & Pisoni, 2004a), Krauss et al.'s (2002) stimuli were from a wider age range (20-60 years). The stimuli were also full-length images rather than images of faces, which may have provided additional cues to inform accurate matching. However, Mavica and Barenholtz (2013) replicated Krauss et al.'s (2002) results using static headshots of age-matched stimuli. Face-voice matching was above chance in both of the experiments they report. These results offer evidence, supported by the evolutionary psychology literature (see section 2.4), that source identity information available in static faces corresponds to information offered by voices.

2.6 Procedural issues relating to novel face-voice matching performance

Procedural differences between studies may account for some of the apparently contradictory results outlined above. Audiovisual speech perception studies (e.g., Kamachi et al., 2003; Lachs & Pisoni, 2004a, 2004b; Lander et al., 2007) have tended to use a crossmodal matching task (Lachs, 1999). This is a sequential 2-alternative forced-choice (2AFC) procedure. In the visual to auditory (V-A) condition, a face is shown then two voices are presented at test, one after the other. In the auditory to visual (A-V) condition, this procedure is reversed: participants hear a voice, and then see two sequentially presented faces at test. One of the alternatives is therefore always the same identity as the target, while the other is a different identity distractor. The participant must decide which of the two alternatives matches the identity of the other-modality stimulus. Studies that have used this procedure have generally emphasised the importance of dynamic articulatory information in facilitating face-voice matching; above chance face-voice matching is typically found for dynamic but not static faces (Kamachi et al., 2003; Lachs & Pisoni, 2004a; Lachs & Pisoni, 2004b; Lander et al., 2007). In contrast, the majority of experiments observing above chance levels of matching accuracy using static facial stimuli have not used this exact procedure, making it unwise to compare results directly. For instance, Krauss et al. (2002) presented a voice followed by two simultaneously presented full-length images. Mavica and Barenholtz's (2013) stimuli (one voice and two test faces) were presented simultaneously in Experiment 1. However, it is important to note that Mavica and Barenholtz's (2013) second experiment replicated above chance level matching with static facial stimuli using the A-V condition of the standard crossmodal matching task (Lachs, 1999). Although the V-A condition was not included, this result hints that even if procedural differences across studies hold some explanatory value, additional factors may also affect performance and help to explain existing contradictions. Nevertheless, the impact of procedural differences on face-voice matching accuracy deserves further attention.

2.6.1 Relative vs. absolute judgements

The eyewitness literature provides evidence to suggest that the simultaneous or sequential presentation of test options may affect matching accuracy by prompting the adoption of different response strategies. Witnesses in real forensic situations might be asked to identify the suspect from either a sequential lineup, in which they see each face one after the other, or a simultaneous lineup, in which all of the faces are presented at the same time. Deciding which procedure is diagnostically superior has been the subject of fierce debate (e.g. Carlson, Gronlund, & Clark, 2008; Ebbesen & Flowe, 2002; Flowe, Smith, Karoğlu, Onwuegbusi & Rai, 2015; Gronlund, 2005; Lindsay, Mansour, Beaudry, Leach & Bertrand, 2009; Meissner, Tredoux, Parker, MacLin, 2005; McQuiston-Surrett, Malpass, Tredoux, 2006; Wells, Steblay, & Dysart, 2012). A simultaneous procedure is believed to encourage witnesses to compare members of a lineup to each other, in order to decide which member best matches their memory for the perpetrator. This strategy is referred to as a relative judgement (Lindsay et al., 1991; Lindsay & Wells, 1985; Wells, 1984), and likely supports accuracy when the perpetrator is present by making it easier to select the best option. In contrast, sequential lineups, in which only one face is visible at a time, are thought to encourage absolute judgements (Lindsay & Wells, 1985). Owing to the difficulty of making comparisons, each lineup member is therefore compared predominantly to the memory of the perpetrator (Wells, Small, Penrod, Malpass, Fulero & Brimacombe, 1998). If the witness correctly identifies the perpetrator, this is a *hit*. If they select an innocent member of the lineup, this is a *false alarm*. Many studies have observed different patterns of accuracy according to lineup procedure, with many showing higher hit rates for simultaneous lineups (Clark, Howell & Davey, 2008; Ebbesen & Flowe, 2002; Steblay et al., 2001, Steblay, Dysart & Wells, 2011). An even more robust finding is that sequential lineups reduce the false alarm rate (Ebbesen & Flowe, 2002; Kneller, Memon & Stevenage, 2001; Steblay et al., 2001), suggesting that sequential procedures simply make participants less likely to make a positive identification. This is a desirable outcome on target absent lineups, but not on target present lineups. However, recent studies have employed Receiver Operating Characteristic analysis to assess more appropriately the diagnostic accuracy of sequential and simultaneous lineups (Gronlund, Wixted & Mickes, 2014; Mickes, Flowe & Wixted, 2012). The results of these experiments generally indicate that the simultaneous procedure is in fact superior in terms of supporting memory sensitivity (i.e. the hit rate).

Lineup procedures differ from forced-choice procedures in that it is always possible to select none of the options, thereby rejecting the whole lineup. A 2AFC task, as employed in the face-voice matching literature, is therefore not exactly comparable. False alarm rates cannot be calculated because participants are forced to make a positive identification, selecting either the stimulus in position 1 or 2. However, in terms of hit rates, which reflect sensitivity, participants might still respond differently when the test options are presented sequentially compared to when they are presented simultaneously. In line with predictions from the eyewitness literature, relative judgements could facilitate higher rates of face-voice matching accuracy. This hypothesis might go some way to explaining why static face-voice matching seems more likely to be above chance level when a simultaneous 2AFC procedure is adopted (Krauss et al., 2002; Mavica & Barenholtz, 2013).

2.6.2 Memory for static and dynamic faces

Research has suggested that memory for dynamic facial images is better than for static facial images (e.g. Knappmeyer, Thornton & Bülthoff, 2003; Lander & Chuang, 2005). Matching procedures that impose a higher memory load may particularly undermine static face-voice matching accuracy, by making it harder for participants to hold the face in working memory for long enough to compare it with the voice for source identity information. This would be most relevant to procedures like the crossmodal matching task, in which the stimuli are presented sequentially rather than simultaneously.

In a review, O'Toole et al. (2002) provide two explanations for the increased memorability of dynamic faces. According to the 'representation enhancement hypothesis', dynamic images facilitate the perception of the 3-D facial structure. Structural information has been shown to be particularly important in facilitating face recognition, and knowledge of 3-D structure is thought to underlie the accuracy of familiar face recognition (Burton, Jenkins & Schweinberger, 2011). Unfamiliar face recognition on the other hand relies more on 2-D pictorial codes, which provide less information (Hancock, Bruce & Burton, 2000). According to the alternative explanation put forward by O'Toole et al. (2002), the 'supplemental information hypothesis', motion offers additional signature information about the given person. However, overall, the benefit of motion is more robust for familiar face recognition than unfamiliar recognition (Christie & Bruce, 1998; O'Toole et al., 2002; Pike, Kemp, Towell & Phillips, 1997).

In line with the argument that dynamic facial images are more memorable, differential processing of static and dynamic facial images is supported by brain imaging evidence. Specific areas of the brain such as the superior temporal sulcus face area (STS-FA) are sensitive to dynamic facial images (Allison, Puce & McCarthy, 2000; Gobbini et al., 2011), whilst the occipital face area (OFA) and fusiform face area (FFA) are sensitive to static facial images (McCarthy, Puce, Gore & Allison, 1997; Yovel & Kanwisher, 2004). There is also evidence of functional dissociations between brain areas responsive to static and dynamic images, further strengthening the argument that these two types of input are processed separately (Pitcher, Dilks, Saxe, Triantafyllou & Kanwisher, 2011; Polosecki, Moeller, Schwers, Romanski, Tsao & Freiwald, 2013).

Based on the literature reviewed above, a possible effect of facial stimulus type on face-voice matching should not be disregarded without further testing, particularly when the stimuli are presented sequentially and therefore must be held in working memory. In an attempt to rule out memory explanations for the results of Experiment 1, which detected above-chance static face-voice matching using a simultaneous 2AFC procedure, Mavica and Barenholtz (2013) used sequential presentation in Experiment 2, running the A-V condition of the crossmodal matching procedure (Lachs, 1999). Although they did not include an V-A condition, the results replicated the finding that static faces could be matched to voices significantly above chance level. This does not entirely rule out an explanation for discrepancies across studies based on memory effects because in both experiments Mavica and Barenholtz (2013) did not include a dynamic face-voice matching condition for comparison.

2.6.3 Order of presentation in cross-modal matching tasks: Auditory-visual or visual-auditory

Other aspects of face-voice matching performance also warrant additional attention. As mentioned above, studies employing the standard crossmodal matching task (Lachs, 1999) have manipulated stimulus presentation order (Kamachi et al., 2003; Lachs & Pisoni, 2004a, 2004b; Lander et al., 2007). Participants either see a face then decide between two voices (V-A), or they hear a voice then decide between two faces (A-V).

The manipulation of order is motivated by face-voice asymmetries identified in audiovisual speech perception research, as well as the assumption that accurate face-voice matching is contingent on encoding visual articulatory patterns in dynamic faces. In terms of speech, voices are more informative than faces; it is easier to perceive what is being said from hearing a voice than it is from lip-reading (see Massaro, 1987). Lachs and Pisoni (2004a) hypothesised that as memory for the details of auditory speech is likely to be superior to memory for visual speech, it would be easier to compare two auditory stimuli (as in the V-A condition) than it would be to compare two (dynamic) visual stimuli (as in the A-V condition). Previous studies have not observed a difference between V-A and A-V conditions when the auditory stimuli consist of normal forwards speech and the visual stimuli are dynamic articulating faces (Kamachi et al., 2003; Lachs & Pisoni, 2004a; Lander et al., 2007). It is not clear from the existing literature whether a performance asymmetry according to the order of stimulus presentation might operate in static face-voice matching, because studies detecting above-chance accuracy have not included both A-V and V-A conditions (Krauss et al., 2002; Mavica & Barenholtz, 2013).

One rationale for manipulating order in face-voice matching, regardless of whether the faces are static or dynamic, relates to sensory memory. Echoic memory for sounds lasts longer than iconic memory for images, as shown by robust modality effects (see Crowder & Morton, 1969; Penney, 1989). Therefore, if the voice stimulus is presented first, the representation might persist for longer, making it easier to compare to the subsequently presented visual stimuli for source identity information (Lachs & Pisoni, 2004a). This is relevant to both static and dynamic face-voice matching.

The recognition literature provides a further rationale for investigating order effects. Whilst voices are more central to speech comprehension, faces offer more reliable identity information (see Stevenage & Neil, 2014 for a review). Studies have consistently observed asymmetries between faces and voices in terms of the rates of recognition accuracy, which have been attributed to differential link strength in the two perception pathways (e.g. Damjanovic & Hanley 2007; Hanley & Turner 2000; Stevenage, Hugill & Lewis, 2012), and more weakly encoded mental representations of voices (Stevenage, Howland & Tippelt, 2011; Stevenage, Neil, Barlow, Dyson, Eaton-Brown & Parsons, 2013). Therefore, it might be expected that when matching voices and static faces, a performance advantage would be afforded in the V-A (compared to the A-V) condition because richer and more clearly encoded information is presented first, thereby facilitating a comparison with the auditory information.

Based on the existing literature it would clearly be premature to disregard the order of stimulus presentation as a factor in face-voice matching. Although face and voice processing is believed to take place in parallel and integrated pathways (Belin et al., 2004), this does not mean that face and voice stimuli are processed identically (see Stevenage & Neil, 2014). Indeed, the evidence reviewed above supports the hypothesis that the order of stimulus presentation may potentially play a role in influencing matching accuracy.

2.6.4 Face-voice matching: 2AFC vs. same-different procedures

All previous studies investigating face-voice matching have used a 2AFC procedure (Kamachi et al., 2003; Lachs & Pisoni, 2004a, 2004b; Lander et al., 2007; Mavica & Barenholtz, 2013). This is one way of experimentally measuring decisions made in conditions of uncertainty, but other procedures are also appropriate. According to signal detection theory (Green & Swets, 1966), two aspects of performance are important when analysing decisions. The first, sensitivity, equates to hit rates, and is concerned with the ease of detecting a signal. It is the ability to correctly respond positively to the signal when it is present. The second aspect of performance is specificity, which equates to the true negative rate, or the ability to correctly identify when the signal is absent. This measure is concerned with criterion placement, or response bias, during the decision making process. These definitions of sensitivity and specificity are used throughout this thesis. They should not be confused with the rather more common view, in which sensitivity reflects accuracy in both correctly identifying and correctly rejecting the presence of a signal. Similarly, measures of response bias traditionally use a balance between hit rate and false alarm rate (1-true negative rate). (For further explanation, please see section 7.2.2.)

Forced-choice tasks only measure sensitivity. Participants are forced to make a binary decision between two test options. According to the standard difference model (see Thurstone, 1927a, 1927b), which underlies signal detection theory (Dyjas, Bausenhart & Ulrich, 2012; García-Pérez & Alcalá-Quintana, 2011), decisions are based entirely on the comparison of the two test options, allowing the participant to select the option that represents the best fit. The 2AFC task therefore assumes that there is no response bias, which means that responses should be distributed evenly across alternatives if both alternatives are equally viable (Green & Swets, 1966; Macmillan & Creelman, 2005; Wickens, 2001). This assumption may also be due in part to the statistical complexity of modelling a possible response bias in a 2AFC task using a signal detection theory approach (DeCarlo, 2012; Green

& Swets, 1966; Luce, 1963). However, evidence for the unbiased nature of the 2AFC procedure appears to be rather questionable, thereby refuting assumptions underlying the standard difference model. Although the data are typically pooled across positions for analysis (García-Pérez & Alcalá-Quintana, 2011), as has been the case in all previous face-voice matching studies (Kamachi et al., 2003; Lachs & Pisoni, 2004a, 2004b; Lander et al., 2007; Mavica & Barenholtz, 2013), levels of observed accuracy in 2AFC tasks appear to differ frequently according to position (Dyjas, Bausenhart & Ulrich, 2012; García-Pérez & Alcalá-Quintana, 2010, 2011; Rammsayer & Ulrich, 2012; Ulrich & Vorberg, 2009; Yeshurun, Carrasco & Maloney, 2008). Having re-analysed 17 published 2AFC experiments testing different kinds of visual sensitivity, Yeshurun et al. (2008) present strong evidence for position biases, making the compelling argument that if the standard difference model is refuted, attempting to use data from 2AFC tasks to represent meaningful measures of sensitivity is problematic. This is because the decisional processes operating during the task and contributing to position effects are wholly unclear. Yeshurun et al. (2008) conclude with a recommendation that 2AFC tasks should be used with caution, if at all.

An alternative procedure, the same-different task, is also commonly used to measure decision making under conditions of uncertainty (Green & Swets, 1966). Same-different tasks measure both sensitivity (hit rate) and specificity (true negative rate), because participants can either respond positively or negatively. In these tasks, two stimuli are presented for participants to respond to. There are signal trials, in which the correct answer is to respond positively, and noise trials in which the correct answer is to respond negatively (Stanislaw & Todorov, 1999). A same-different task, in which participants respond *same identity* if they think the face and voice belong to the same person, and *different identity* if they do not, would be appropriate for face-voice matching.

Response bias may be an important aspect of face-voice matching performance. If people have a tendency to accept a face and voice as belonging to the same identity, this would not be modelled using the 2AFC paradigm. In light of the overall pattern of false alarms identified in the eyewitness literature (e.g. Steblay et al., 2001; Ebbesen & Flowe, 2002; Kneller et al., 2001) (see section 2.6.1), response bias is perhaps even more likely to differ than detection sensitivity according to whether test options are presented simultaneously or sequentially. Although order of presentation effects have not been observed by previous face-voice matching studies using a 2AFC task (Kamachi et al., 2003; Lachs & Pisoni, 2004a, 2004b; Lander et al., 2007) (see section 2.6.3), it is possible that an order effect may operate in terms of response bias.

As face-voice matching has only ever been tested using 2AFC procedures, it is important to test the validity of previous findings using alternative experimental procedures. In areas of research such as recognition memory and vision, levels of accuracy are frequently reported to differ according to whether 2AFC or same-different tasks are employed (Azzopardi & Cowey, 1998; Balsdon & Azzopardi, 2015; Jang, Wixted & Huber, 2009), most probably because participants are forced to rely on different strategies in order to complete these tasks. 2AFC tasks force participants to make a positive decision: the answer is either option 1 or option 2. In a same-different task they can respond according to their response criterion: *different identity* if they have adopted a conservative response criterion and *same identity* if their response criterion is more liberal.

Owing to the reliance on 2AFC procedures, previous face-voice matching studies have never investigated how response bias operates in face-voice matching. Investigation of this aspect of performance using a same-different procedure may offer an important insight both into how the faces and voices of novel people are processed.

2.7 Summary

Although faces and voices constitute distinct types of sensory stimuli, their processing exhibits many parallels. Extensive brain imaging and behavioural evidence supports the auditory face model (Belin et al., 2004), which highlights the importance of crossmodal redundancies in aiding social communication (Belin et al., 2011; Campanella & Belin, 2007).

Whilst audiovisual speech perception research has shown that redundant information is offered by voices and dynamic faces (e.g. Dohen et al., 2004; Lander et al., 2007; Munhall & Vatikiotis-Bateson, 1998; Yehia et al., 1998; Yehia et al., 2000), evolutionary psychology research suggests that redundant information is also available in voices and static faces (e.g., Abend et al., 2015; Collins & Missing 2003; Thornhill & Gangestad 1999; Thornhill & Grammer 1999; Feinberg et al., 2005; Feinberg, 2008; Saxton et al., 2006; Wheatley et al., 2014; Zahavi & Zahavi 1997). Both areas of research independently offer compelling evidence for voices and dynamic faces, as well as voices and static faces, sharing redundancies. It seems plausible that common source identity information is crossmodally available in voices and faces, regardless of whether facial stimuli are static or dynamic.

The face-voice matching literature illustrates a rather more confusing picture of crossmodal redundancy. Although voices and dynamic faces are consistently matched above chance level, static face-voice matching is more variable (Krauss et al., 2002; Lachs & Pisoni, 2004a; Kamachi et al., 2003; Lander et al., 2007; Mavica & Barenholtz, 2013). It is therefore unclear whether encoding dynamic visual articulatory speech patterns is crucial to accurate face-voice matching.

There are a number of possible explanations for the apparent contradictions across face-voice matching studies, none of which have been thoroughly investigated or resolved by previous research. For example, procedural differences might help to explain differing patterns of results. An alternative explanation is based on the hypothesis that poorer memory for static compared to dynamic facial images (Knappmeyer et al., 2003; Lander & Chuang, 2005; O'Toole et al., 2002; Pike et al., 1997) affects performance.

The results of existing studies leave a number of important questions unanswered, and do not fully reveal how face-voice matching performance operates. By relying exclusively on 2AFC procedures, researchers have unwittingly neglected to address possible response biases in face-voice matching, which may constitute a key aspect of performance. This literature review has highlighted some important gaps in knowledge, which the subsequent experiments seek to fill.

2.8 Aims

2.8.1 Research questions

The specific research questions to be addressed throughout this thesis include:

- Research question 1: Do voices share redundant information with dynamic as well as static faces?
- Research question 2: Is it possible to match voices and static faces, or is accurate face-voice matching contingent on encoding information about visual articulatory patterns?
- Research question 3: Do procedural differences account for inconsistencies in the previous literature regarding static face-voice matching?
- Research question 4: Are there matching performance asymmetries according to the order of stimulus presentation?
- Research question 5: How do response biases operate in face-voice matching?

3. CHAPTER 3: FACE AND VOICE STIMULI: METHODOLOGICAL AND STATISTICAL ISSUES

3.1 Introduction

There is a large amount of inter- and intra-stimulus variation associated with both faces and voices. Not only do people look and sound different across images and utterances, but they also look and sound different from each other (Burton, 2013; Belin, Zatorre & Ahad, 2002; Schweinberger et al., 2014; Stevenage & Neil, 2014; Valentine, Lewis & Hills, 2015). Investigating face and voice perception and modelling the resulting data poses a number of challenges that must be met in order for the findings to be generalisable (Clark, 1973; Judd, Westfall & Kenny, 2012; Wells et al., 2013; Wells & Windschitl, 1999).

Wells and Windschitl's (1999) widely cited paper on stimulus sampling warns against basing conclusions on functional sample sizes of *N*=1 when stimuli within a category differ from each other (see also Brunswick, 1947; Kenny, 1985). Their paper argues that failing to use an adequate sample of stimuli threatens external validity, limiting generalisability and construct validity by potentially confounding a single stimulus (e.g. one face) with a whole category (e.g. all faces) (Wells & Windschitl, 1999). However Wells and Windschitl (1999) acknowledge that including an appropriate sample of stimuli only addresses one aspect of the challenge associated with modelling variability. In order to maximise generalisability it is also necessary to employ statistical analyses that avoid aggregating over either stimuli or participants, because aggregating involves ignoring a source of variability that is relevant to the interpretation of the results (Clark, 1973; Judd et al. 2012; Wells et al., 2013).

This chapter is split into three main sections. The first section (3.2) details why stimulus sampling is important when investigating face and voice perception. The second section (3.3) outlines the advantages of using multilevel modelling over traditional ANOVA, and the third section (3.4) describes the stimuli used throughout this thesis.

3.2 Stimulus variability

3.2.1 Variability in faces

Stimulus sampling is particularly relevant to experiments featuring facial stimuli. Faces vary from each other on a number of different dimensions, such as height, width, feature size, skin texture, age and attractiveness (Jenkins, White, Van Mountford & Burton, 2011; Valentine et al., 2015). The face-space model (Valentine, 1991) explains how variability might affect face processing. According to this model, representations of faces are located at different spatial positions within a multidimensional space. The organising principle of face representations is their similarity to a central, prototypical face. Faces that resemble each other are arranged close together, whilst a larger distance separates those with less in common. Representations of distinctive faces therefore lie towards the edge of the face-space, while representations of typical faces are clustered around the mid-point. Owing to the likelihood that distinctive faces will have more empty space surrounding them than typical faces, the model predicts that distinctive faces will be encoded with less error (Valentine, 1991). If this is the case, distinctive faces should be easier to recognise. Indeed, this has been consistently found to be the case (Bartlett, Hurry & Thorley, 1984; Light, Kayra-Stuart & Hollander, 1979; Vokey & Read, 1992).

Burton (2013) emphasises the importance of taking inter-stimulus variability into account when investigating face perception. He argues that the frequent failure to do so is an important factor in explaining the slow progression of research in face recognition. The common practice of aggregating over stimuli in conventional statistical analyses (see Wells et al., 2013) averages over a huge amount of variability within the face space, thereby ignoring the way that face processing is affected by inter-stimulus variability.

3.2.2 Variability in voices

Voices differ from each other in terms of fundamental frequency, speech rate, nasality, accent and age etc. (Handkins & Cross, 1985; Mathias & von Kriegstein, 2014; Mullennix & Pisoni, 1990). Evidence from fMRI studies is consistent with the conclusion that, in a similar way to faces, representations of voices are also located within a multidimensional space organised with reference to a prototypical voice (Latinus, McAleer, Bestelmeyer & Belin, 2013). In line with this voice-space model, the literature on distinctiveness supports the hypothesis that stimulus variation is an important factor in voice perception (Schweinberger et al., 2014; Stevenage & Neil, 2014). For example, Barsics and Brédart (2012) observed a distinctiveness advantage for voices in terms of the retrieval of semantic information. Research into voice recognition also suggests that performance varies not only across participants, but also across stimuli (e.g. Mullennix, Ross, Kuykendall, Conard & Barb, 2011; Orchard & Yarmey, 1995; Van Lancker, Kreiman & Emmorey, 1985). As is the case for faces, averaging over voice variability is likely to minimise stimulus effects and reduce generalisability (Stevenage & Neil, 2014).

3.2.3 Implications for face-voice matching

Stimulus level variability in faces and voices may affect face-voice matching performance. It is likely that matching decisions are highly dependent on specific stimuli pairings; perhaps some people look and sound more similar than others. Previous studies have used varying numbers of face-voice pairs when testing crossmodal matching, which may help to account for the apparent contradictions outlined in the literature review (see section 2.6). For example, whilst Lachs and Pisoni (2004a) used 8 face-voice pairs, Kamachi et al. (2003) used 40. Matching performance appears to vary according to specific stimulus pairings. Mavica and Barenholtz (2013) reported that matching accuracy varied between 35% and 70% for the 64 models whose faces and voices featured in their study. Although some previous face-voice matching studies include by-stimulus analyses (Kamachi et al., 2003; Mavica & Barenholtz, 2013), simultaneously accounting for the variance associated with stimuli and participants is a problem that can only be appropriately dealt with by a statistical model that incorporates both sources of variability, such as a multilevel model (Baguley, 2012; Judd et al., 2012).

3.3 Multilevel modelling

3.3.1 Problems associated with conventional statistical analyses

In cases when different participants encounter a number of stimuli over trials, the data is best understood as being organised into a hierarchy because the observations from each participant are not independent (see Baayen Davidson & Bates, 2008 for a discussion of nested and cross-classified random effects; Nezlek, 2008). The stimuli at level 1, and the participants at level 2, both constitute a sample, and variance is associated with each of them (Baayen et al., 2008). Both sources of sampling error must be taken into account in order to avoid the ecological fallacy. This fallacy arises when it is falsely assumed that patterns observed for participant means also hold for data at a lower level of analysis such as individual trials (level 1) repeated within participants (level 2) (e.g., see Robinson, 1950; Wells et al., 2013). Performing a traditional regression on individual observations for this kind of data would violate the assumption of independence. However, commonly used alternatives such as least squares dummy-codes do not appropriately account for sampling error (Hoffman & Rovine, 2007; Nezlek, 2001; Nezlek, 2008).

The solution of aggregating over one level of analysis, as is the procedure when performing a by-participants (most common) or by-stimulus ANOVA, is equally problematic (Judd et al., 2012) because it only accommodates one fixed effect at a time. It is important to distinguish between fixed effects, which are constant across participants, and random effects, which vary (Kreft, Kreft & De Leeuw, 1998). In an example in which participants encounter a number of different faces and voices, the participants, face stimuli, and voice stimuli should all be treated as random effects (Judd et al., 2012). The majority of papers investigating face and voice perception have tended to rely on conventional analyses (for exceptions see Morrison et al., 2007; Wells et al. 2013), which involve treating the stimuli as a fixed effect (Clark, 1973). Multilevel modelling represents a more desirable method of dealing with the variability associated with facial and vocal stimuli.

3.3.2 The advantages of multilevel modelling

Multilevel modelling is a recently developed statistical method, which addresses the problems of conventional analyses outlined above (Baayen, 2008; Baayen et al., 2008; Judd et al., 2012; Wright & London, 2009). Although some researchers may be hesitant to adopt this seemingly complex statistical innovation (Quené & Van den Bergh, 2004), the method is likely to be increasingly adopted for hierarchical data in future psychological research (Wright & London, 2009).

One of the main advantages of multilevel modelling is that it can simultaneously take into account the variability associated with individual performance and different stimuli. In multilevel modelling, variability is allowed at multiple levels, thereby explaining the different sources of variance.

By avoiding aggregating data (see Wells et al., 2013), and separating the sampling error from the treatment effect, multilevel modelling successfully reduces the risk of Type 1 error (Baguley, 2012; Clark, 1973; Judd et al., 2012). Unless the ignored source of variability is negligible, multilevel modelling is always more conservative than separate by-stimulus or by-participant analyses. The outcome of analyses performed using traditional ANOVA compared to those using multilevel modelling can vary, seriously affecting the resulting conclusions (Westfall, Kenny & Judd, 2014). An example using data from this thesis (Experiment 2a) is presented in Appendix A.

Accounting for variability appropriately is particularly important when investigating face-voice matching. The crucial issue in much of the previous literature is whether static face-voice matching is above or below chance level, the level of accuracy that reflects guessing (Kamachi et al., 2003; Lachs & Pisoni, 2004a; Lander et al., 2007; Krauss et al., 2002; Mavica & Barenholtz, 2013). In 2AFC tasks, which have been used in all previous face-voice matching studies, chance level is 50%. Measuring whether performance is truly above chance can be achieved by observing whether the 95% confidence intervals overlap with 50%. The confidence intervals should always be calculated in a way that incorporates both sources of variability to avoid Type 1 error (i.e. incorrectly concluding that performance is above chance level).

As shown above, the challenges of investigating face and voice perception are therefore two-fold. In line with the recommendations of Wells and Windschitl (1999), an adequate sample of stimuli should be used. Additionally, in order to generalise beyond the sample of faces and voices used in experiments, the resulting data should be analysed in a way that simultaneously takes into account both the variability associated with the stimuli and the participants. Multilevel modelling provides a way of achieving this.

3.4 Stimuli used in the thesis

The following section describes the stimuli featured in each experiment of this thesis. Stimulus faces and voices were taken from the GRID audiovisual sentence corpus (Cooke, Barker, Cunningham & Shao, 2006), a multi-talker corpus featuring head and shoulder videos of British adult speakers saying 1,000, 6-word sentences each in an emotionally neutral manner. All the videos are recorded against a plain blue background. Each sentence follows the same 6-word structure: 1) command (*set/lay/bin/place*), 2) colour (*red/blue/green/white*), 3) preposition (*at/by/with/in*), 4) letter (*A-Z*, although *W* was excluded because it has more than one syllable), 5) digit (*1-9*), 6) adverb (*now/soon/please/again*), for example, "*Place blue at J 9 now*". Although there is overlap in terms of the words articulated across and within speakers, none of the exact sentences in the corpus are ever repeated.

The corpus features 34 speakers. In total 18 speakers were selected: 9 male and 9 female. Ideally all 34 would have been useable, as this would have increased the size of the stimulus sample. However in order to facilitate comparisons with previous face-voice matching studies (Kamachi et al., 2003; Krauss et al., 2002; Lachs & Pisoni, 2004; Lander et al., 2007) it was necessary to compromise stimulus sample size in favour of matching the stimuli for ethnicity (white British), accent (English) and age (18-30). Of the selected stimuli, 2 of the males and 2 of the females wore glasses.

Each set of experiments in this thesis features static faces, dynamic (muted) faces, and voices from the GRID corpus. The method of selecting and editing these files is explained below. Images and transcripts for each of the 18 stimulus people are presented in Appendix B.

3.4.1 The stimulus set

Three videos (.mpegs) were selected at random from numbered files using an online research randomiser (Urbaniak & Plous, 2013). All of the videos for each stimulus person were recorded during the same session.

3.4.1.1 Static faces

One of the three videos was used to create static pictures of faces. Pictures were extracted using the snapshot function on Windows Movie Maker (2012), and presented in .png format. In keeping with Schweinberger, Robertson and Kaufmann (2007), the static picture for each talker was the first frame of the video. Some of the stimuli were opening their mouth to prepare their first word, but none were in the process of articulating. The image measured 368 x 288 pixels. An example static face is shown in Figure 3.1.

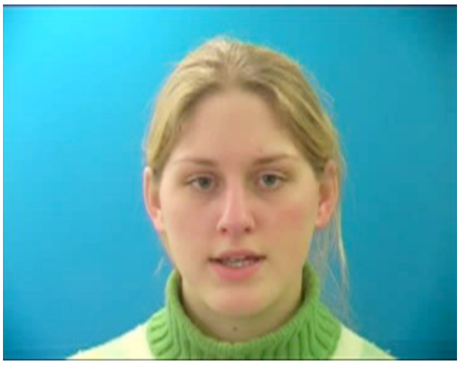


Figure 3.1: *Example static facial stimulus*

3.4.1.2 Dynamic faces

Another of the three video files was used to construct the dynamic stimuli. The file was muted using Windows Movie Maker, and converted back into .mpeg format using Mobile Media Converter (v1.7.7). The video measured 368 x 288 pixels, and played at a rate of 25 progressive frames per second.

3.4.1.3 Voice recordings

Voices played from the last of the three video files (.mpeg), and featured audio quality of 256 kbits per second.

3.5 Conclusion

The first challenge of modelling face and voice data is to include an adequate sample of stimuli. All the experiments reported in this thesis featured a sample of faces and voices belonging to 18 different people. In addition, face-voice matching trials were constructed so that stimuli from one modality (e.g. faces) did not always occur with the same distractor stimuli from the other modality (e.g. voices). In order to address the second challenge, which is to maximise the chances of being able to generalise from both stimuli and participants, multilevel modelling was used for all appropriate analyses.

4. CHAPTER 4: TESTING THE BACK-UP SIGNAL HYPOTHESIS: DO FACES AND VOICES OFFER REDUNDANT INFORMATION?

4.1 Introduction

Faces and voices are informative about dimensions of fitness and quality (Belin et al., 2004; Collins & Missing, 2003; Feinberg, 2008; Feinberg et al., 2005; Fraccaro et al., 2010; Pisanski et al., 2012; Yovel & Belin, 2013). Aside from some research investigating attractiveness cues (e.g. Abend et al., 2015; Collins & Missing, 2003; Oguchi & Kikuchi, 1997; Saxton et al., 2006, 2009; Wells et al., 2013), little is known about how multimodal signals for other dimensions of fitness operate in humans. Motivated by findings from the attractiveness literature, this experiment tests whether faces and voices elicit concordant judgements about masculinity/femininity, age, health, height and weight.

Combined information from faces and voices might provide overlapping information (a back-up signal) (Johnstone, 1997) or complementary information (a multiple message) (Møller & Pomiankowski, 1993). It is possible to distinguish between multiple messages and back-up signals by empirically testing the effect of multimodal signals on a recipient (Partan & Marler, 2005). If a multimodal signal present in human faces and voices is a back-up signal for a certain dimension, ratings on this dimension should correlate, whereas uncorrelated ratings would reflect the presence of multiple messages (Wells et al., 2009; Wells et al., 2013).

Previous face-voice matching studies, despite ostensibly dealing with face-voice redundancy, have not directly addressed the extent to which faces and voices offer redundant information. For example, Krauss et al. (2002) asked participants to rate the age, height and weight of speakers. One group judged voice recordings, another judged the speakers' full-length photographs. Ratings were compared against the speakers' actual age, height and weight. Their results indicated that although the participants were slightly more accurate when rating photographs, they were almost as accurate when rating voices. Krauss et al. (2002) only focused on how accurate the face and voice ratings were, rather than how concordant they were. Similarly, Mavica and Barenholtz's (2013) participants rated photographs of faces and recordings of voices for age, height and weight, as well as dimensions relating to socioeconomic status and personality. However, the focus of their analysis was whether the average difference score for each dimension predicted matching. They did not report any details about face-voice concordance on the different scales, or give an indication of how closely related the face and voice ratings were.

In the present study, ratings of masculinity/femininity, age, health height and weight were recorded, then correlated, from independent ratings of faces and voices.[1]

4.1.1 Aim

In order to build hypotheses regarding the accuracy of both static and dynamic face-voice matching, Experiment 1 aimed to establish whether faces and voices communicate similar information (back-up signals) or different but complementary information (multiple messages) about people. Participants judged faces and voices separately, estimating age (in years), and completing Likert-style rating scales for femininity/masculinity, health, height and weight. In light of the contradictory findings regarding the extent to which attractiveness judgements made from static and dynamic facial stimuli are related (see section 2.5.1), the study also tested whether the relationship between face and voice ratings differs according to facial stimulus type. As the previous literature suggests that both faces and voices honestly

[1] The data from Experiment 1 have been published (Smith, Dunn, Baguley & Stacey, 2016a**)** (see Appendix D)

signal quality, it was expected that judgements made independently from faces and voices should be similar.

4.2 Method

4.2.1 Design

This experiment employed a mixed design. The between subjects factor was facial stimulus type (static or moving), and the within subjects factor was modality (visual or auditory). The dependent variables were age estimates (in years) and ratings on scales for femininity/masculinity, health, height and weight.

4.2.2 Participants

The participants (*N* = 48) were recruited from the Nottingham Trent University Psychology Division's Research Participation Scheme and by convenience sampling. There were 12 male and 36 female participants (age range = 18 to 28 years, *M* = 20.54, *SD* = 2.59). All participants reported having normal or corrected vision and hearing. Student participants received research credits in line with course requirements. The university's BLSS (Business, Law and Social Science) College Research Ethics Committee granted ethical approval for this, and subsequent experiments (ref: 2013/37).

4.2.3 Apparatus and materials

The stimuli were presented on an Acer Aspire laptop (2.5GHz processor, screen size 15.6 inches, resolution 1366 x 768 pixels, Dolby Advanced Audio), with brightness set to the maximum level. The laptop was placed approximately 8.5cm away from the edge of the desk at which participants sat. The experiment ran on Psychopy v1.77.01 (Peirce 2009), an open-source software package designed for running experiments in Python. To reduce background noise, participants listened to the recordings binaurally through Apple EarPods, which have a frequency range of 5Hz to 21,000Hz. This exceeds the range of human hearing (Feinberg et al. 2005). Voices were played at a comfortable listening volume (30% of the maximum volume). Two versions of the experiment were constructed: one using static faces and voices, and one using dynamic faces and voices. In both versions, all 18 faces and voices were presented. All of the stimuli were presented for 2 seconds each.

4.2.4 Procedure

The participants were randomly allocated to either the static face or the dynamic face version of the experiment using an online research randomiser (Urbaniak & Plous, 2013). They read the information sheet, completed the consent form, and provided demographic information. Testing took place in a quiet cubicle. Participants completed two counterbalanced blocks of testing. As illustrated in Figure 4.1, in one block participants viewed faces (visual (V) condition), in the other they heard voices (auditory (A) condition). Participants were not told that the voices and faces featured in the experiment belonged to the same people. Each block consisted of a practice trial, followed by 18 randomly ordered experimental trials. After each face or voice, participants estimated the age of the stimulus person in years and completed 7-point Likert-style rating scales in the following order: femininity/masculinity (1 – *very feminine*, 7 – *very masculine*), health (1 – *very unhealthy*, 7 – *very healthy*), height (1 – *very short*, 7 – *very tall*) and weight (1 – *very underweight*, 7 – *very overweight*). The participants responded by pressing number keys on the laptop keyboard.

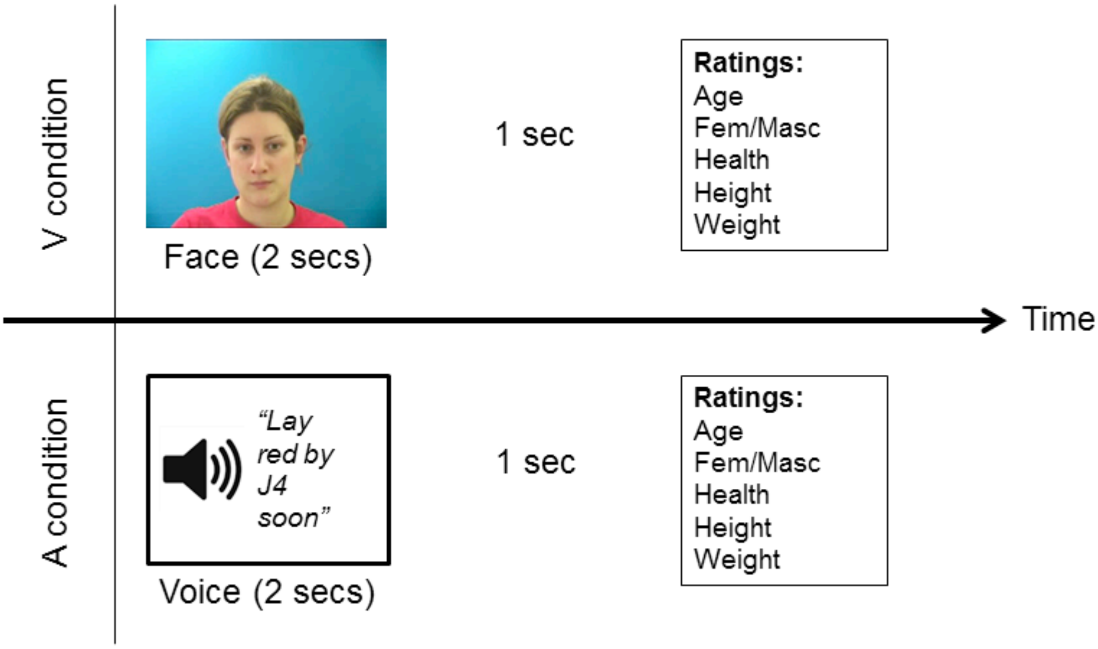


Figure 4.1: *An illustration of the procedure used in Experiment 1*

4.3 Results

The mean estimated age, and ratings for femininity/masculinity, health, height, and weight, are recorded in Appendix B for each stimulus person's dynamic face, static face, and voice. Datasets and executable R code for each experiment reported in this thesis can be accessed via the Google Drive link provided in Appendix C.

4.3.1 Absolute difference between face and voice ratings

The absolute difference between face and voice ratings was measured by comparing each rating participants had given to a face and voice belonging to the same person. Following this, the mean absolute difference (MAD) for each stimulus person on each rating scale (age, masculinity/femininity, health, height and weight) was calculated. Descriptive statistics (Table 4.1) indicated that typical ratings for faces and voices fall within a similar range.

Table 4.1

Mean absolute difference (MAD) and 95% confidence intervals for the MAD between face and voice ratings by stimulus type condition

	Static facial stimuli				Dynamic facial stimuli			
			95% CI				95% CI	
Rating scale	*M*	*SD*	LB	UB	*M*	*SD*	LB	UB
Age	3.91	1.51	3.27	4.55	3.62	1.58	2.95	4.29
Masculinity/Femininity	1.05	.35	.90	1.19	1.00	.36	.85	1.15
Health	1.24	.34	1.10	1.39	1.12	.27	1.00	1.23
Height	1.10	.29	.98	1.23	1.04	.36	.89	1.19
Weight	.92	.25	.81	1.02	1.00	.27	.88	1.11

On all scales apart from age, face and voice ratings only differ on average by about 1 point (14%) on a 7-point rating scale, and MADs were similar across static and dynamic facial stimuli. The difference between face and voice ratings in terms of age appears larger than that of the other rating scales. However, rather than being rated on a 7-point scale, age estimates were given in years. This prevents a neat comparison between the rating scales.

4.3.2 Correlation between face and voice ratings

The results in Table 4.1 show that face and voice ratings tend to be close together in terms of the range they fall into. A logical next step is to quantify the extent to which voice and face ratings co-vary in the same individual. For this purpose, a simple correlation coefficient between voice and face ratings would either ignore the dependency within participants or rely only on aggregate data (mean ratings for each participant). Use of multilevel models means that both participant and stimuli variation can be accounted for when correlating voice ratings with face ratings for estimated age, and ratings for femininity/masculinity, health, height and weight (see section 3.3). These correlations are scaled in the same way as Pearson's correlation. For each variable, an intercept-only model was fitted with the rating as an outcome, using the lme4 package in R (Bates, Maechler, Bolker & Walker, 2014). A crucial aspect of each model was to estimate separate variance for face and voice ratings as well as the correlation between face and voice ratings across both stimuli and participants. The correlation between face and voice ratings within participants is, for present purposes, a nuisance term (merely indicating that participants who give high ratings to voices also tend to give high ratings to faces) and is not reported here. The correlations reported in Table 4.2 are those within stimuli and demonstrate that, for a given item, voice and face ratings are positively correlated.

Table 4.2

Within stimulus correlations between face and voice ratings

	Correlation coefficient				
Condition	Age	Masc/Fem	Health	Height	Weight
Static facial stimuli	.60	.97	.70	.83	.40
Dynamic facial stimuli	.32	.92	.91	.86	.17
All facial stimuli	.46	.95	.77	.84	.28

Table 4.2 provides evidence that mean face and voice ratings for the same identity appear to be positively related for all rating types. Correlations between face and voice ratings on scales for masculinity/femininity, health, and height were particularly high, regardless of whether the facial stimuli were static or dynamic. Correlations between the mean face and voice ratings for age and weight were moderate when facial stimuli were static, with some suggestion that the correlations were diminished for dynamic stimuli. However correlations did not vary according to facial stimulus type in direction, or by more than .3 on any scale. The difference between the static and dynamic correlations was tested by fitting models with separate variance terms for each stimulus type. Comparing a model that includes separate variance and covariance terms for static and dynamic stimuli with one

that does not, did not improve the model fit for any of the ratings (*p* >.14). This complements the results shown in Table 4.2, suggesting that the extent to which faces and voices offer similar information is not greatly influenced by whether the facial stimuli are static or dynamic.

4.4 Discussion

Experiment 1 investigated the extent to which novel faces and voices offer concordant information about dimensions of fitness and quality. The results indicate that not only do face and voice ratings fall within a similar range, but that independent ratings of an individual's face and voice are positively correlated. These results complement other studies showing that faces and voices offer related information about mate value (Collins & Missing, 2003; Feinberg, 2008; Feinberg et al., 2005; Fraccaro et al., 2010).

The strongest correlations between face and voice ratings occurred on scales for masculinity/femininity, health, and height. The striking relatedness of face and voice ratings observed on these dimensions is underlined by the fact that results were obtained using multilevel modelling. This method is more conservative than traditional methods of statistical analysis (Baguley, 2012), and avoids the ecological fallacy (Robinson, 1950; Wells et al., 2013) (see section 3.3.1).

It is necessary to acknowledge that rating scales were always completed in the same order. The first scale was always masculinity/femininity, and the possibility that there was some carryover when completing subsequent scales for health, height and weight cannot be dismissed. However, this is unlikely to have influenced the results in a way that undermines the overall conclusion that faces and voices provide related information about mate value. If the results could exclusively be explained by carryover, much stronger and more consistent relationships across the scales might have been anticipated, particularly as the strongest relationship was observed on the first scale. As it was, the strength of correlations varied across the rating scales in a way that corresponds with the previous literature. This is evident when considering the results relating to body size.

Despite the previous literature indicating a tendency for unimodal voice ratings of body size to be less accurate than unimodal face ratings (Bruckert et al., 2006; Coetzee et al., 2010; Collins, 2000; Re et al., 2013; van Dommelen & Moxness, 1995), Experiment 1 showed that regardless of accuracy, body size judgements made from faces and voices fall within a similar range. However, correlations were strong for height, but only weak-moderate for weight. This corresponds with the pattern of findings reported by Lass and Colt (1980), who observed significant differences between weight ratings from male faces and voices, but not between height ratings.

The stimuli were from a narrow demographic (see section 3.4), meaning that they are unlikely to have varied very much from each other. In Appendix B it is clear that the participants did not make use of the full range (1-7) of each rating scale, and all ratings fell between values of 2 and 5. Although this might help to explain why the ratings for faces and voices tended to fall within such a small range, it does not explain the correlation results. These indicate that the average face and voice ratings were ordered extremely similarly, particularly in terms of masculinity/femininity, health and height. So for example, regardless of the range of rating values used, the results appear to reflect the fact that if someone looks taller than another person, they also tend to sound taller than that person. Similarly, on the basis of these results it is difficult to argue that the results are due to people guessing and merely attributing a mid value. If responses were truly arbitrary it is almost unfeasible that the results would be echoed across face and voice ratings in the way that the correlations show; many of the relationships were very strong.

Frequently contradictory findings regarding attractiveness ratings of static and dynamic facial stimuli have been reported in the literature (see Roberts et al., 2009b). This is one of the only experiments to consider how facial stimulus type affects face-voice rating concordance. Whilst Lander (2008) found that judgements of male face and voice attractiveness were related only when faces were dynamic, here in Experiment 1, a lack of difference between static and dynamic face-voice correlations (on the dimensions tested) appears to be robust. No difference was observed on any of the five rating scales, so these signals appear to be stable across dynamic and static faces. It therefore seems justifiable to use this set of results to inform hypotheses regarding the relationship between static and dynamic face-voice matching.

4.4.1 Using the ratings results to inform hypotheses about face-voice matching

Interpretation of the present set of results is not intended to propose that if static face-voice matching is possible it is wholly attributable to dimensions relating to fitness and quality. Faces and voices convey a wide spectrum of socially relevant information (Belin et al., 2004, 2011; Campanella & Belin, 2007). Nevertheless, the results constitute sufficient evidence to counter the hypothesis, based on audiovisual speech perception research, that dynamic face-voice matching is possible, but static face-voice matching is not (Kamachi et al., 2003; Lachs & Pisoni, 2004a). If face and voice ratings are so closely related, on any dimension, static face-voice matching should be hypothetically possible.

The results of Experiment 1 indicate that ratings made from faces and voices of the same identity, presented in isolation, offer redundant signals (Johnstone, 1997) on a number of dimensions. Information about masculinity/femininity, height and health is particularly similar across faces and voices. The extent to which faces and voices offer concordant information is not affected by whether the face is static or dynamic. These results support the suggestion that it is possible to match novel voices and static faces (Krauss et al., 2002; Mavica & Barenholtz, 2013) as well as voices and dynamic faces (Kamachi et al., 2003; Lachs & Pisoni, 2004a, 2004b; Lander et al., 2007).

5. CHAPTER 5: MATCHING NOVEL FACE AND VOICE IDENTITY USING TWO-ALTERNATIVE FORCED-CHOICE PROCEDURES

5.1 Introduction

Experiment 1 showed that observers perceive a number of dimensions extremely similarly in faces and voices. The relationship between face and voice ratings on these dimensions did not vary according to whether facial stimuli were static or dynamic. This chapter explores the prediction that crossmodal source-identity information is shared both by voices and static faces as well as voices and dynamic faces, testing face-voice matching performance across three different 2AFC procedures.

Overall, the hypothesis that static face-voice matching is possible receives rather inconclusive support from the existing literature. Although voices are consistently matched to dynamic articulating faces significantly above chance level, static face-voice matching performance varies across studies (Kamachi et al., 2003; Krauss et al., 2002; Lachs & Pisoni, 2004a, 2004b, Lander et al., 2007; Mavica & Barenholtz, 2013). In line with predictions informed by audiovisual speech perception research, evidence of chance level static face-voice matching has been taken to suggest that accurate matching depends on being able to encode visual articulatory movement (Kamachi et al., 2003; Lachs & Pisoni, 2004a). As faces and voices offer such a wide range of socially relevant information (e.g. Belin et al., 2004, 2011; Campanella & Belin, 2007; de Gelder & Vroomen, 2000; Dolan et al., 2001; Massaro & Egan, 1996), chance level static face-voice matching may reflect a lack of redundancy on dimensions aside from those tested in Experiment 1.

Some studies have observed above-chance static face-voice matching (Krauss et al., 2002; Mavica & Barenholtz, 2013). One possible explanation for the apparent contradictions hinges on procedural differences across studies. Whilst studies observing chance level face-voice matching using static facial stimuli have employed a standard crossmodal matching task (Kamachi et al., 2003; Lachs & Pisoni, 2004a), studies reporting above-chance level performance have used procedures (e.g. Krauss et al., 2002; Mavica & Barenholtz, 2013, Experiment 1) that might have encouraged different response strategies which better support matching accuracy (Lindsay et al., 1991; Lindsay & Wells, 1985; Wells, 1984). Although it is important to acknowledge that Mavica and Barenholtz (2013, Experiment 2) did observe above-chance level static face-voice matching using the A-V condition of the crossmodal matching procedure (Lachs, 1999), they omitted the V-A condition. On the basis of these results it would be premature to conclude that procedural differences do not influence performance accuracy.

A further explanation for the contradictory results is offered by the face recognition literature. Some research suggests that memory is better for dynamic compared to static faces (e.g. Knappmeyer et al., 2003; Lander & Chuang, 2005). If memory load is higher using sequential procedures such as the crossmodal matching task (Lachs, 1999), this might disproportionately affect static compared to dynamic face-voice matching accuracy. A position effect might occur in a sequential 2AFC task, whereby accuracy is higher if the same identity other-modality stimulus appears in position 1 rather than position 2. Previous face-voice matching studies have not included analyses of responses by position, so the impact of this factor is unknown. However, position effects in 2AFC tasks are well documented in the psychological literature, and so require attention in this context (García-Pérez & Alcalá-Quintana, 2010, 2011; Yeshurun et al., 2008).

In order to thoroughly investigate static and dynamic face-voice matching, key manipulations must be appropriately scrutinized. Some previous studies have not attended to the impact of stimulus presentation order (Visual-Auditory or Auditory-Visual) on matching accuracy (Mavica & Barenholtz, 2013). However, asymmetries in terms of sensory memory for faces and voices (Crowder & Morton, 1969; Penney, 1985) might affect responses. In addition, as faces provide more reliable identity information, and voices are processed more for speech information (see Stevenage & Neil, 2014), this could influence accuracy according to the order of stimulus presentation. Performance may depend on whether speech or visual identity information is most important in facilitating matching.

Addressing the impact of facial stimulus type is crucial. If visual articulatory movement is so critical to matching accuracy, there may be a significant difference in accuracy according to whether the faces are static or dynamic. Although audio-visual speech perception researchers have tested face and voice matching using both static and dynamic facial stimuli, they have not statistically compared the data (Kamachi et al., 2003; Lachs & Pisoni 2004a). Neither of the studies identifying above chance level static face-voice matching have included trials using dynamic articulating faces (Kamachi et al., 2003; Krauss et al., 2002). Failure to include both static and dynamic face conditions prevents direct comparison of crossmodal matching explanations based on static facial information (e.g. Krauss et al., 2002; Mavica & Barenholtz, 2013) to those focusing on dynamic facial information (Kamachi et al., 2003; Lachs & Pisoni, 2004a, 2004b; Lander et al., 2007; Rosenblum et al., 2006).

5.1.1 Aim

In the face of contradictory results, this chapter aims to clarify whether static face-voice matching is possible using stimuli of the same age, sex and ethnicity, comparing matching accuracy across three different 2AFC procedures.

In an attempt to tease apart the relative contribution of static or dynamic face information in facilitating crossmodal matching, performance using static and dynamic faces was compared in both Experiments 2a and 2b. Performance on A-V and V-A trials was also compared in these two experiments. In case better memory for dynamic facial stimuli affects matching accuracy, memory load was varied across studies. In Experiment 2a, all of the stimuli were presented sequentially, so memory load was higher, whereas in Experiment 2b, face-voice combinations were presented simultaneously. In a further test of whether static face-voice matching is sensitive to procedural differences, Experiment 2c adopts the procedure of Krauss et al. (2002), in which alternatives in a 2AFC task are presented simultaneously. In an attempt to clarify how memory load and procedure affects performance, all three experiments included a manipulation of position to test whether accuracy is higher when the same identity stimulus appears in position 1 rather than position 2.[2]

5.2 Experiment 2a: Sequential face-voice presentation in a 2AFC matching task

Experiment 2a used a standard crossmodal matching task (Lachs, 1999) to compare static and dynamic face-voice matching. In most experiments in which this procedure has been used, the results have shown only dynamic face-voice matching to be at above chance level (Kamachi et al., 2003; Lachs & Pisoni, 2004a; Lander et al., 2007; cf. Mavica & Barenholtz, 2013, Experiment 2). The balance of evidence therefore predicted static face-voice matching to be at chance level using this particular procedure.

5.2.1 Methods

5.2.1.1 Design

Experiment 2a employed a 2 x 2 x 2 mixed factorial design. The between subjects factor was facial stimulus type (static or dynamic), and the within subjects factors were order

[2] The data from Experiments 2a, 2b and 2c have been published (Smith, Dunn, Baguley & Stacey, 2016b) (see Appendix E)

(visual to auditory (V-A) or auditory to visual (A-V)), and position (position 1 or position 2). The dependent variable was matching accuracy.

5.2.1.2 Participants

The participants (*N* = 82) were recruited from the Nottingham Trent University Psychology Division's Research Participation Scheme and by convenience sampling. There were 26 male and 56 female participants (age range = 18 to 66 years, *M* = 23.70, *SD* = 8.56). All reported having normal or corrected vision and hearing. None of the participants had taken part in Experiment 1. In line with course requirements, student participants received research credits.

5.2.1.3 Apparatus and materials

The apparatus used in Experiment 2a was exactly the same as in Experiment 1. Four versions of the experiment were created so that trials could be constructed using different combinations of stimuli in order to maximise stimulus sampling (Wells & Windschitl, 1999). Each version comprised 12 trials in total, and each trial featured 3 stimuli. In the V-A condition, a face (stimulus 1) was followed by two sequentially presented voices (stimuli 2 and 3): a target (a same identity, other modality stimulus) and a distractor (a different identity, other modality stimulus). In the A-V condition, a voice (stimulus 1) was followed by sequentially presented target and distractor faces (stimuli 2 and 3). Across versions, whether someone's face/voice appeared as stimulus 1, 2 or 3, and whether it was used in a V-A or A-V trial, was randomly varied. The position of the target stimulus at test (position 1 or position 2) was also randomly and equally varied. In each experimental version, all 18 faces and voices appeared. All of the stimuli were presented for 2 seconds each. None of the faces or voices appeared more than once in any version. Each of the 4 versions was used for the between subjects manipulation of facial stimulus (static or dynamic). In total there were 8 versions of the experiment.

5.2.1.4 Procedure

The participants were randomly allocated to one of the 8 versions of the experiment using an online research randomiser (Urbaniak & Plous, 2013). In the dynamic facial stimulus condition the participants were accurately informed that the face and the voice were saying different sentences to prevent the use of speech-reading (Kamachi et al., 2003).

The participants completed 2 counterbalanced experimental blocks. There was a practice trial, followed by 6 randomly ordered experimental trials. As illustrated in Figure 5.1, in one block of trials participants saw a face first. After a 1 second gap they heard the first voice. The text 'Voice 1' was visible in the middle of the screen while the recording was playing. After another 1 second gap they heard the second voice, with the text 'Voice 2' visible in the middle of the screen. In the other block of trials, participants heard a voice first, and then saw 2 faces, presented one after the other. Gaps of 1 second were inserted between all stimuli, and the text 'Face 1' or 'Face 2' appeared below each picture. At test, using number keys on the laptop keyboard, the participants were asked to select either *1* or *2*, as the face/voice that was the same identity as the first stimulus.

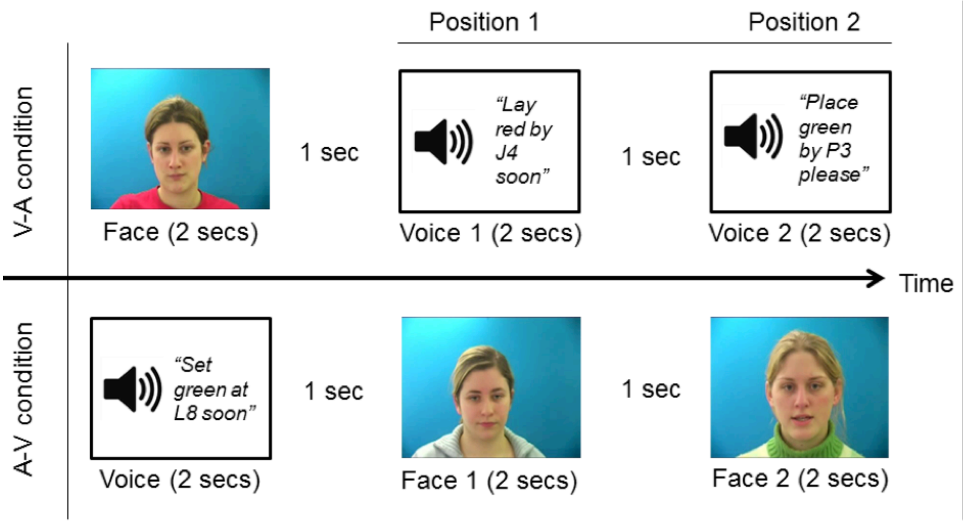


Figure 5.1: *An illustration of the procedure used in Experiment 2a*

5.2.2 Results

All of the data were analysed using multilevel models in order that both the participants and the stimuli could be treated as random effects. The random effects were fully crossed; every participant encountered all 36 stimuli (18 faces, 18 voices) in each version of the experiment. Matching accuracy was analysed using multilevel logistic regression with the lme4 version 1.06 package in R (Bates et al., 2014). Four nested models were compared, all fitted using restricted maximum likelihood, and with accuracy (0 or 1) as the dependent variable. The first model included a single intercept; the second included the main effects of each factor (order, position and facial stimulus type). The third added the two-way interactions, and the final model included the three-way interaction. This method of analysis allowed us to test for individual effects in a similar way to traditional ANOVA. However, as *F* tests derived multilevel models tend not to be accurate, profile likelihood ratio tests provided by lme4 are reported instead. These are more robust, and are obtained by dropping each effect in turn from the appropriate model (e.g., testing the three-way interaction by dropping it from the model including all effects, and testing the two-way interactions by dropping each effect in turn from the two-way model).

Table 5.1 shows the profile likelihood chi-square statistic (G^2) and *p* value associated with dropping each effect. Table 5.1 also reports the coefficients and standard errors (on a log odds scale) for each effect in the full three-way interaction model. Variability for the first stimulus in each trial (the voice in the A-V condition, and the face in the V-A condition) was modelled separately from the foil stimulus. The random effect for the first stimulus captures the variability of both faces and voices because corresponding faces and voices are highly correlated. For foils, it was more appropriate to model separate random effects for faces and voices because the corresponding voice or face was never present. In the three-way model, the estimate of *SD* of the first stimulus random effect was 0.535, for the voice foils it was 0.634, and for face foils it was 0.484. The estimated *SD* for the participant effect was less than 0.0001. A similar pattern held for the null model. Thus, although individual differences were negligible in this instance, a conventional by-participants analysis that did not incorporate variance associated with the stimuli could have been extremely misleading.

Table 5.1

Parameter estimates (b) and profile likelihood tests for the 2x2x2 factorial analysis, Experiment 2a: Sequential face-voice presentation in a 2AFC matching task

Source	*df*	*b*	*SE*	G^2	*p*
Intercept	1	0.444	0.315	.	.
Position	1	0.062	0.374	5.92	.015
Order	1	0.333	0.371	0.68	.410
Facial stimulus type	1	0.676	0.277	3.42	.064
Position x Order	1	0.870	0.516	0.35	.553
Position x Facial stimulus type	1	0.625	0.390	0.02	.884
Order x Facial stimulus type	1	0.775	0.382	0.59	.441
Position x Order x Facial stimulus type	1	1.159	0.549	4.34	.037

The main effect of position was significant (p = .015), along with the 3-way interaction between position, order and facial stimulus type (p = .037). Figure 5.2 aids interpretation of the effects and interaction, showing means and 95% confidence intervals for matching accuracy in each condition of the factorial design. The confidence intervals were obtained by simulating the posterior distributions of cell means in R (arm package, version 1.6) (Gelman & Su, 2013).

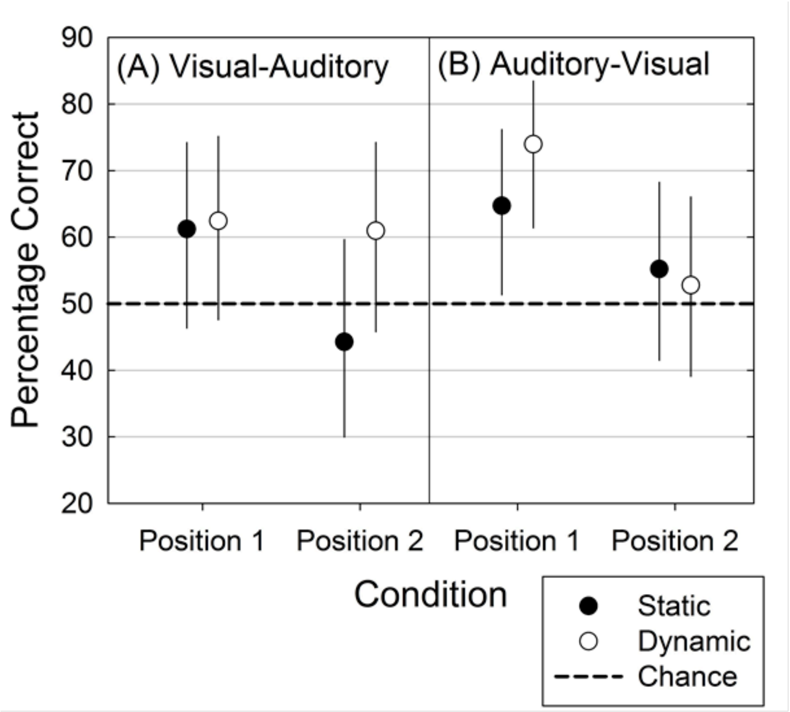


Figure 5.2: *Face-voice matching accuracy on V-A (panel A) and A-V (panel B) trials for sequentially presented faces and voices in a 2AFC matching task. Error bars show 95% CI for the condition means*

Overall matching performance was significantly above chance (50%) level, M = 59.7%, 95% CI [50.8, 68.0]. However, confidence intervals for percentage accuracy in the static, M = 57.6%, 95% CI [47.5, 67.1], and dynamic, M = 63.7%, 95% CI [53.8, 72.5], conditions show that only performance on dynamic facial stimulus trials was significantly above chance level. Figure 5.2 shows the main effect of position, with accuracy levels consistently higher when the same identity stimulus was presented in position 1 compared to when it was presented in position 2. The results from the V-A condition are shown in panel A, while results from the A-V condition appear in panel B. Using visual analysis to guide an interpretation, it appears that the basis of the three-way interaction relates to performance when the same identity other-modality stimulus appears in position 2 in the V-A condition. In the V-A condition there is no position effect in the dynamic facial stimulus condition. However, as with any factorial design testing multiple effects it would be imprudent to over-interpret a single non-predicted interaction that is only just statistically significant (p = .037).

5.2.3 Discussion

Using the standard crossmodal matching task (Lachs, 1999) employed in audiovisual speech perception research, Experiment 1 observed above chance dynamic face-voice matching, but chance level static face-voice matching. Although there was no significant difference between static and dynamic face-voice matching accuracy, and static face-voice matching was close to being above chance level, this pattern of results appears to support the conclusion that source identity information shared by dynamic articulating faces and voices explains accurate face-voice matching. The results are consistent with two previous studies (Kamachi et al., 2003; Lachs & Pisoni, 2004a), but in conflict with Mavica and Barenholtz (2013, Experiment 2), who observed above chance level static face-voice matching using this procedure.

The presence of a position effect in Experiment 2a additionally suggests that memory load might be hindering performance, especially in the static facial stimulus condition. Face-voice matching was more accurate when the same identity face and voice were presented in relatively closer temporal proximity (position 1), than when the same identity other-modality stimulus was further away (position 2). In line with research suggesting that memory is better for dynamic than static faces (Christie & Bruce, 1998; Knappmeyer et al., 2003), the position effect appears not to manifest in the dynamic facial stimulus, V-A condition. Although this interpretation must not be overstated, based as it is on visual analysis, it is important to consider a possible explanation for the three-way interaction. In the V-A condition, the face (stimulus 1) needs to be held in memory whilst being compared to voice 1 and voice 2. If dynamic faces are more durable in memory than static faces, their representation might endure better across both voices, meaning that a position effect does not occur. As voices are less durable than faces (Stevenage et al., 2011, 2012, 2013), comparisons across two faces might be particularly difficult in the A-V condition. In keeping with this explanation, the bias in the A-V condition occurs regardless of whether the subsequent faces are static or dynamic.

5.3 Experiment 2b: Simultaneous face-voice presentation in a 2AFC matching task

In order to clarify the effect of procedural differences across previous studies, Experiment 2b used a modified presentation procedure from Experiment 2a. Experiment 2b presented 2 different face-voice combinations. This time the face and voice in each combination were presented simultaneously, instead of sequentially. It was hypothesised that relieving the memory load should make it easier to identify incongruent face-voice combinations (Lander et al., 2007). Therefore, matching accuracy should be higher when faces and voices are presented simultaneously, and above chance for static face-voice matching.

5.3.1 Method.

The methods for Experiment 2b were identical to Experiment 2a, with exceptions outlined below.

5.3.1.1 Participants

There were 7 male and 33 female adult participants (*N*=40) with an age range of 18 to 33 years (*M* = 21.38, *SD*= 3.57). None of the participants had taken part in previous experiments.

5.3.1.2 Procedure

The procedure used in Experiment 2b is illustrated in Figure 5.3. Participants in the V-A condition saw a face accompanied by a recording of a voice. The text ‘Voice 1’ was visible underneath the face. After a 1 second gap they saw the same face accompanied by a different voice. The text ‘Voice 2’ appeared beneath the face. In the A-V condition, participants heard a voice accompanied by a face, followed by a 1 second intervening gap, after which they heard the same voice accompanied by a different face. The text ‘Face 1’ and ‘Face 2’ appeared below the first and second combination respectively. Participants had to decide which combination featured same identity stimuli by pressing *1* on the laptop keyboard for face/voice 1, or *2* for face/voice 2.

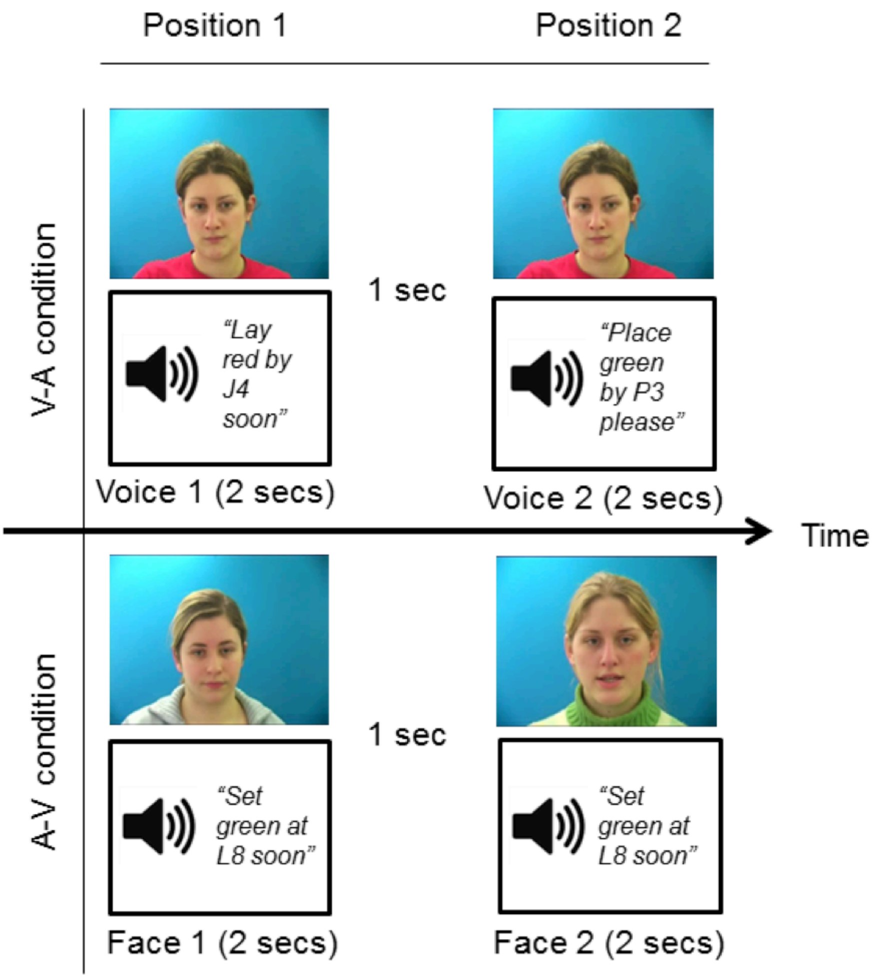


Figure 5.3: *An illustration of the procedure used in Experiment 2b*

5.3.2 Results

Matching accuracy was analysed using the same method as Experiment 2a. Table 5.2 shows the profile likelihood chi-square statistic (G^2) and p value associated with dropping each effect in turn from the appropriate model. Coefficients and standard errors (on a log odds scale) for each effect in the full three-way interaction model are also reported in Table 5.2. A similar pattern of *SD*s was observed for the random effects, with more variability at the stimulus level than the participant level. In the three-way model, the estimate of *SD* of the first stimulus random effect was 0.778, for the voice foils it was 0.324, and for the face foils it was 0.103. The estimated *SD* for the participant effect was 0.007.

Table 5.2

Parameter estimates (b) and profile likelihood tests for the 2x2x2 factorial analysis, Experiment 2b: Simultaneous face-voice presentation in a 2AFC matching task

Source	*df*	*b*	*SE*	G^2	*p*
Intercept	1	0.266	0.365	.	.
Position	1	0.550	0.462	17.40	<.001
Order	1	0.755	0.431	<0.01	.952
Facial stimulus type	1	0.314	0.391	0.37	.545
Position x Order	1	1.402	0.653	1.95	.162
Position x Facial stimulus type	1	0.140	0.568	1.09	.295
Order x Facial stimulus type	1	0.771	0.549	0.37	.544
Position x Order x Facial stimulus type	1	1.121	0.804	1.90	.169

Only the main effect of position was significant (*p* < .001). Figure 5.4 aids interpretation of this main effect, showing the means and 95% confidence intervals for accuracy in each of the 8 conditions, obtained using the arm package (version 1.6) (Gelman & Su, 2013).

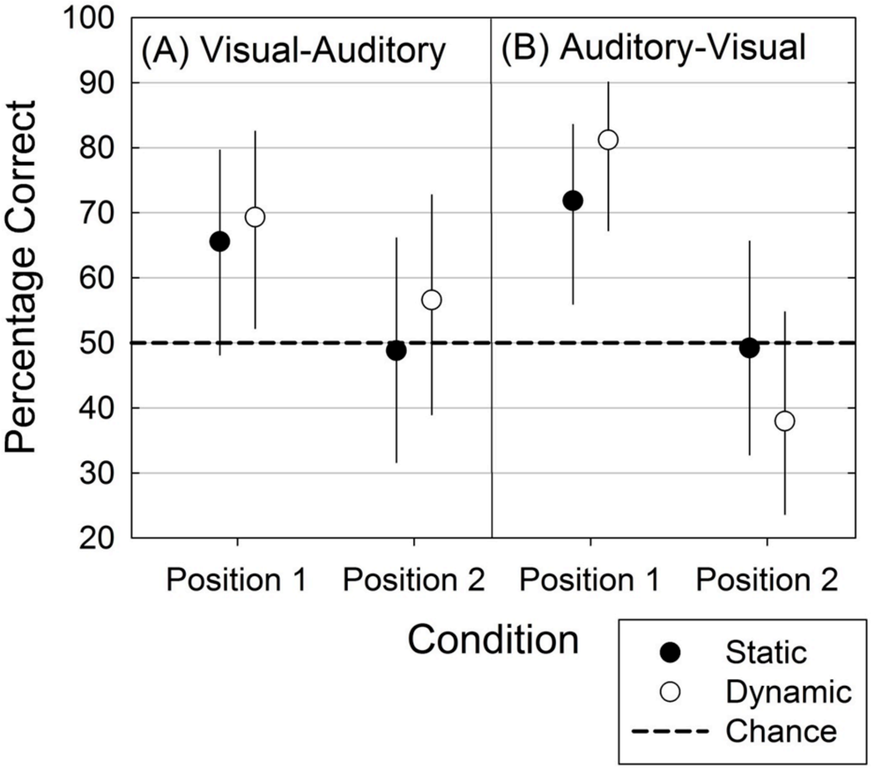


Figure 5.4: *Face-voice matching accuracy on V-A (panel A) and A-V (panel B) trials for simultaneously presented faces and voices in a 2AFC matching task. Error bars show 95% CI for the condition means*

As in Experiment 2a, overall matching performance was significantly above chance (50%) level, *M* = 60.9%, 95% CI [50.4, 70.5]. Overall, the dynamic facial stimulus trials were significantly above chance, *M* = 62.5%, 95% CI [50.1, 73.6], but static facial stimulus trials were not, *M* = 59.8%, 95% CI [47.2, 71.2]. As is clear from Figure 5.4, the main effect of position exhibits the same pattern as Experiment 2a, with accuracy levels consistently higher when the same identity face-voice combination is presented in position 1. There was, however, no three-way interaction.

5.3.3 Discussion

Overall the pattern of results observed in Experiment 2b is largely similar to that observed in Experiment 2a, when the stimuli were presented sequentially. The participants in

Experiment 2b exhibited a bias towards selecting the first face-voice combination they encountered. As the position effect was observed in both experiments, this may be less attributable to memory load, and more related to the nature of the 2AFC-task: when alternatives are presented sequentially, the first alternative is disproportionately favoured. Indeed, this explanation corresponds well with other studies, which have found widespread similar evidence of temporal position biases using 2AFC procedures (García-Pérez & Alcalá-Quintana, 2010, 2011; Yeshurun et al., 2008). However, it would be premature to rule out explanations based on memory at this stage. In contrast to the results presented in Experiment 2a, there was no three-way interaction; the position effect also occurred in the dynamic facial stimulus V-A condition. The interaction in Experiment 2a was explained in terms of the differential durability of dynamic faces, static faces, and voices. It was suggested that comparisons would be particularly difficult in conditions when less durable stimuli must be held in memory for a longer time. It is possible that the durability of a face-voice combination is only as strong as its weakest element, i.e. the voice (Stevenage et al., 2011, 2012, 2013). If this is the case, a uniform position effect would be expected across conditions when sequentially presented alternatives consist of face-voice combinations. This is what the results of Experiment 2b show.

5.4 Experiment 2c: Simultaneously presented alternatives in a 2AFC matching task

The results from Experiment 2b showed that simultaneously presenting faces and voices did not improve static face-voice matching. This was contrary to what was expected; it seems that the pattern of results from Experiment 2a were not attributable to increased memory load impairing the comparison of the first stimulus to the same identity other-modality stimulus in position 2. Experiment 2c was designed to test whether chance level static face-voice matching could be attributable to the sequential presentation of alternatives in a 2AFC task. Evidence from the forensic eyewitness literature suggests that simultaneously presenting faces in a lineup array prompts a different pattern of results in comparison to when faces are presented sequentially (Clark et al., 2008; Ebbesen & Flowe, 2002; Steblay et al., 2011). This is possibly because of differential use of relative and absolute judgements (Kneller et al., 2001). Relative judgements (Wells, 1984) are employed when choosing the best option from simultaneously presented alternatives, whereas sequential presentation of alternatives encourages absolute judgements because of the difficulty of making comparisons (Kneller et al., 2001; Wells et al., 1998).

Some previous experiments finding above chance face-voice matching accuracy with static stimuli have used a procedure in which test alternatives are presented simultaneously, and can therefore be more easily compared (Krauss et al., 2002; Mavica & Barenholtz, 2013 Experiment 1). Experiment 2c tested whether static face-voice matching is above chance level when the alternatives in a 2AFC task are presented simultaneously. Because of the nature of this procedure, and the difficulty of presenting voices simultaneously at test, Experiment 2c only included an A-V condition. The main aim of Experiment 2c was to account for static face-voice matching, replicating the procedure of Krauss et al. (2002) as closely as possible. Considering the null effect of facial stimulus type in Experiments 2a and 2b, this experiment does not include a dynamic face condition. Taking into account the results of Krauss et al. (2002), in conjunction with the observation that faces and voices offer redundant information on a number of dimensions (Experiment 1), it seemed likely that this particular procedure would elicit above chance static face-voice matching. The manifestation of a position effect was not anticipated when the 2 face alternatives were presented simultaneously.

5.4.1 Method

The methods for Experiment 2c were identical to Experiment 2a and 2b, with exceptions outlined below.

5.4.1.1 Design

Experiment 2c employed a within subjects design, with one factor: position (left = position 1 or right = position 2). The dependent variable was matching accuracy.

5.4.1.2 Participants

There were 8 male and 22 female adult participants (*N* = 30) with an age range of 18 to 44 years (*M* = 20.70, *SD* = 5.20). None had taken part in either Experiment 2a or Experiment 2b.

5.4.1.3 Apparatus and materials

In the absence of a between subjects manipulation, only 4 versions of Experiment 2c were constructed, all of which featured different combinations of stimuli. Each version featured 1 block of 18 trials, in which a voice was followed by the presentation of 2 faces. The same-identity face was always present at test, with its position (left = position 1 or right = position 2) randomly and equally varied. Each voice was only heard once in each version. Each of the stimulus faces appeared twice, but only once as the same identity stimulus. This was in keeping with the procedure of Krauss et al. (2002), who also re-used visual stimuli as foils.

5.4.1.4 Procedure

The participants were randomly allocated to one of the four versions of the experiment using an online research randomiser (Urbaniak & Plous, 2013). The procedure for Experiment 2c is illustrated in Figure 5.5. The participants heard a voice for 2 seconds. After a 1 second gap they saw 2 images of faces presented side by side. The text 'Face 1' was visible underneath the face on the left, and the text 'Face 2' appeared underneath the face on the right. This screen was visible for 2 seconds. Participants were then instructed to decide which face matched the voice they had heard, indicating their answer by pressing *1* on the laptop keyboard for 'Face 1', and *2* for 'Face 2'.

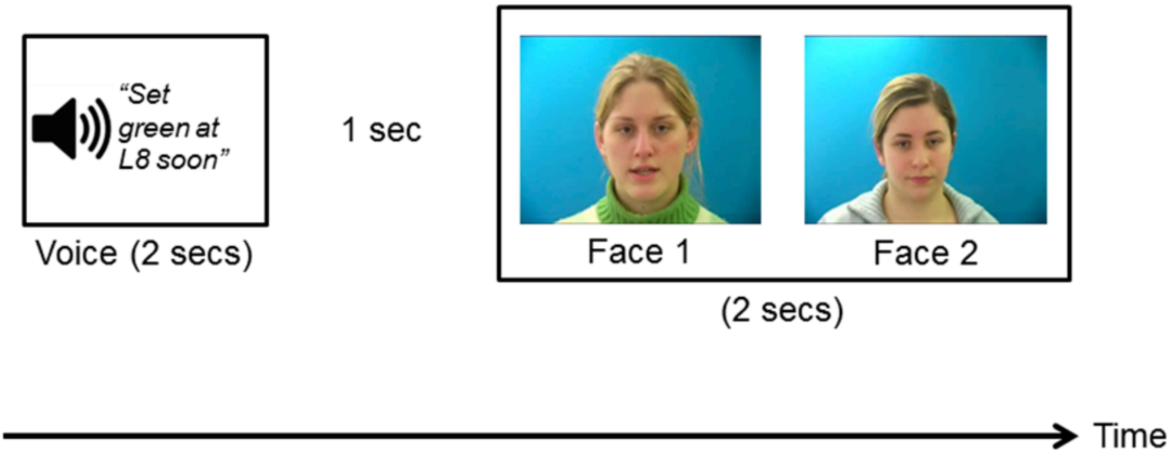


Figure 5.5: *An illustration of the procedure used in Experiment 2c*

5.4.2 Results

Face-voice matching accuracy was analysed using the same method as Experiment 2a and 2b. Since there is only one within subjects factor, only the profile likelihood chi-square statistic (G^2) and *p* value associated with dropping the main effect from the null model is reported. The coefficients and standard error (on a log odds scale) for the effect of position in the main effect model are reported in Table 5.3. In the main effect model, the estimate of *SD* of the voice random effect was 0.487, and for the face foil it was 0.0002. The estimated *SD* for the participant effect was less than 0.0001.

Table 5.3

Parameter estimates (b) and profile likelihood tests for the analysis, Experiment 2c: Simultaneously presented alternatives in a 2AFC matching task

Source	*df*	*b*	*SE*	*G^2*	*p*
Intercept	1	0.446	0.147	.	.
Spatial position	1	0.199	0.203	0.98	.329

The main effect of position was non-significant (*p* = .329). Overall matching accuracy with simultaneously presented static facial stimuli was above chance level (50%), *M* = 61.0%, 95% CI [54.1, 67.6].

5.4.3 Discussion

The results of Experiment 2c indicate that when test alternatives are presented simultaneously, static face-voice matching is above chance level. In keeping with previous findings (Mavica & Barenholtz, 2013), this confirms that static face-voice matching is possible. The results also replicate the findings of Krauss et al. (2002) using headshots rather than full-length images. Considered alongside the results presented in Experiments 2a and 2b, it would appear that static face-voice matching performance is sensitive to procedure, thus offering one possible explanation for contradictions between previous studies.

Experiments 2a and 2b showed that there is a temporal position bias when test options are presented sequentially. However, Experiment 2c suggests that there is no corresponding spatial position bias; when the test options are presented simultaneously, the position bias is negligible.

5.5 General Discussion

In an attempt to resolve discrepancies across previous face-voice matching studies, this chapter tested whether crossmodal source identity information is exclusively dependent on encoding visual articulatory patterns, or whether static faces and voices offer sufficient concordant information to facilitate above chance performance. Taken together, the results of Chapter 5 are consistent with the conclusion that whilst articulatory movement might be important in facilitating face-voice matching (Experiments 2a and 2b), it is also possible to match static faces and voices when a 2AFC procedure facilitates comparisons between alternatives (Experiment 2c). Therefore, it seems that procedural differences between previous studies offer a possible explanation for discrepant results in the literature. Furthermore, as shown by the variance associated with stimuli in the multilevel modelling analysis, people vary in the extent to which they look and sound similar. This offers a complementary explanation for contradictions in previous studies, because previous results may be highly dependent on the particular stimuli used.

5.5.1 Static vs. dynamic face-voice matching

Experiments 2a and 2b presented test alternatives in the 2AFC task sequentially. The results replicate those of audiovisual speech perception studies showing that although dynamic faces and voices can be matched significantly above chance level, static faces and voices cannot (Kamachi et al., 2003; Lachs & Pisoni, 2004a). As shown by the results of Experiment 2c, and in keeping with the alternative hypothesis that static faces and voices offer concordant source identity information (Feinberg et al., 2005; Krauss et al., 2002; Mavica & Barenholtz, 2013; Saxton et al., 2006), performance was significantly above chance when the alternatives were presented simultaneously. The overall results are therefore not consistent with the conclusion that dynamic articulatory movement is exclusively responsible for explaining cross-modal matching (e.g., Kamachi et al., 2003; Lachs & Pisoni, 2004a), although they do not rule out the audiovisual speech perception argument that visual articulatory movement shares source identity information with voices (Kamachi et al., 2003; Lachs & Pisoni, 2004a, 2004b; Rosenblum et al., 2006).

The absence of any statistical difference between static and dynamic face-voice matching in Experiment 2a and 2b warns against overstating the importance of visual articulatory movement in accounting for crossmodal matching accuracy. That said, the absence of an effect of facial stimulus type is not necessarily at odds with the results of studies detecting accurate face-voice matching when movement is isolated using point-light displays, and static information is unavailable (Lachs & Pisoni, 2004b; Rosenblum et al., 2006). Dynamic point-light displays could offer sufficient information to enable accurate face-voice matching independently of the structural information available in static images.

5.5.2 Procedural differences

On both static and dynamic facial stimulus trials, there was a uniform position effect in Experiment 2b when the memory load was reduced. Our findings are more consistent with the conclusion that the position effect is attributable to the nature of the 2AFC task (García-Pérez & Alcalá-Quintana, 2010, 2011; Yeshurun et al., 2008) when the two test alternatives are presented sequentially. This undermines the suggestion that 2AFC procedures are unbiased (Green & Swets, 1966; Macmillan & Creelman, 2005; Wickens, 2002), hinting that alternative procedures, such as a same-different task, might be more appropriate for investigating face-voice matching. The observed position bias appears to be temporal rather than spatial. However, presenting test alternatives simultaneously in a 2AFC task is not ideal. Investigating order of presentation effects using this procedure is problematic because of the undesirability of presenting two voices at the same time for comparison. As 2AFC position effects have not been addressed in previous face-voice matching research, this topic would benefit from further investigation, and is the subject of the next chapter.

In comparing the results of Experiments 2a and 2b to Experiment 2c, it appears that static face-voice matching is sensitive to the procedure employed. The similarity of results across Experiments 2a (sequential face-voice presentation) and 2b (simultaneous face-voice presentation) suggest that contradictions between previous studies are not attributable to superior performance when faces and voices are presented simultaneously. This may be because the more critical comparison to make in facilitating matching accuracy is between alternatives, rather than between the face and the voice. When the two alternatives are presented simultaneously, as in Experiment 2c, the key comparison, a relative judgement (Wells, 1984), is easier to make.

At this point it should be noted that in previous face-voice matching experiments using a crossmodal matching procedure, a standard inter-stimulus interval of 500ms has been used (e.g. Lachs & Pisoni, 2004a, 2004b; Mavica & Barenholtz, 2013), which is half as long as the interval featured in the experiments reported here. With 1-second intervals in Experiment 2a we observed chance level static face-voice matching when the stimuli were presented sequentially. Using 500ms intervals, Mavica and Barenholtz (2013, Experiment 2) observed above-chance level matching accuracy. It is necessary to consider the possible impact of this methodological dissimilarity. It could be argued that a longer interval might increase the load on auditory and visual sensory memory, making the task more difficult. The results we report support the argument that sensory memory pressures do not account for the chance level static facial stimulus results in Experiment 2a. Experiment 2b, in which faces and voices were presented simultaneously, was designed to alleviate memory load. The results were very similar to the results of Experiment 2a; static face-voice matching was still at chance level.

5.5.3 Variability associated with stimuli

An explanation based on procedural differences does not accommodate all the results in the previous literature. Mavica and Barenholtz (2013) observed above chance static face-voice matching using sequential presentation of alternatives in the A-V condition of the standard crossmodal matching task (Lachs, 1999). Alongside procedural differences, this set of three experiments also highlights the importance of stimulus variability in providing an additional but complementary explanation for contradictions between previous studies. Other studies have used varying numbers of face-voice pairs when testing crossmodal matching. For example, whilst Lachs and Pisoni (2004a) used 8 pairs, Kamachi et al. (2003) used 40 pairs. The results of the multilevel modelling analyses described in Experiments 2a, 2b and 2c reveal that some people look and sound more similar than others; relatively high levels of variance associated with stimuli were observed for the 18 face-voice pairs used here, and in all three experiments the overall variance associated with stimuli was far greater than that associated with the participants. Consistent with this, Mavica and Barenholtz (2013) report that for their stimuli, levels of matching accuracy varied widely, between 35% and 70%, across 64 face-voice pairs. Overall, Mavica and Bareholtz's (2013) stimuli pairings of voices and static faces may have been easier to match than the pairings featured in Experiment 2a, 2b and 2c, or those featured in previous studies (Kamachi et al., 2003; Lachs & Pisoni, 2004a).

5.5.4 No effect of order in 2AFC tasks

In line with other studies (Kamachi et al., 2003, forwards and backwards conditions; Lachs and Pisoni, 2004a; Lander et al., 2007), neither Experiment 2a nor 2b showed an effect of the order of stimulus presentation. Therefore, certainly in terms of detection sensitivity, as measured by accuracy in the 2AFC procedure, face-voice matching appears to be unaffected by differential sensory memory for faces and voices (Crowder & Morton, 1969; Penney, 1985), the greater contribution of identity information by faces, or the greater contribution of speech information by voices (see Stevenage & Neil, 2014).

5.5.5. Conclusion

The results presented in this chapter are consistent with the results of Experiment 1, suggesting that source identity is shared by dynamic articulating faces and voices, as well as static faces and voices. The findings help to resolve previous uncertainty about whether static face-voice matching is possible, presenting two complementary explanations for apparent contradictions. The data suggest that static face-voice matching is more likely to be above chance level when alternatives in a 2AFC task are presented simultaneously. In addition, the variance associated with stimuli indicates that some people look and sound more similar than others, an issue which has not been properly accounted for by analyses undertaken in previous research, but which helps to explain why static face-voice matching performance across previous studies might be inconsistent.

The overall results of this chapter therefore support the conclusion that dynamic visual information about articulatory patterns facilitates matching accuracy (Kamachi et al., 2003; Lachs & Pisoni, 2004a, 2004b; Lander et al., 2007; Rosenblum et al., 2006), but that this alone cannot explain the existence of shared source identity information with voices. Crossmodal source identity information appears to be available in both static and dynamic faces.

6. CHAPTER 6: POSITION BIAS IN TWO-ALTERNATIVE FORCED-CHOICE PROCEDURES

6.1 Introduction

In Experiments 2a and 2b, a temporal position bias was observed. The participants were more accurate when the same identity stimulus appeared in position 1 of a 2AFC task. This bias was observed when the memory load was higher and all the stimuli were presented sequentially (Experiment 2a), as well as when the memory load was lower and face-voice combinations were presented simultaneously (Experiment 2b). As all previous face-voice matching studies have employed a 2AFC procedure, with the majority presenting test alternatives sequentially (Kamachi et al., 2003; Lachs & Pisoni 2004a, 2004b; Lander et al., 2007; Mavica & Barenholtz, 2013, Experiment 2, but see Krauss et al., 2002; Mavica & Barenholtz, 2013, Experiment 1), identifying a temporal position effect is an important finding. Its presence casts doubt on this procedure offering a wholly unbiased method of investigating face-voice matching, and supports previous findings of 2AFC position biases in other areas of research (García-Pérez & Alcalá-Quintana, 2010, 2011; Yeshurun et al., 2008). For example, Yeshurun et al. (2008) re-analysed the data from 17 studies measuring visual sensitivity. The data had originally been collapsed across positions in these studies, but the re-analyses revealed clear position biases. In some cases the alternative in position 1 was differentially favoured, in other cases it was the alternative in position 2 (see section 2.6.4).

There are a number of possible explanations for temporal position biases in 2AFC tasks. When participants are uncertain they may not assign guesses equally to the alternatives presented in position 1 and 2 (García-Pérez & Alcalá-Quintana, 2010, 2011; Jäkel & Wichmann, 2006). This could be due to something as simple as key '1' being in a more comfortable position to press than key '2'. Alternatively, the position bias may manifest because sensitivity differs across positions 1 and 2, so for example, it might be easier to be accurate when the correct alternative is in position 1 (Yeshurun et al., 2008). One way to distinguish between these two explanations is to test how responses are distributed when the same identity stimulus is not present at test. If the latter explanation is supported, there should be no position bias in 2AFC face-voice matching tasks.

6.1.1 Aim

To demonstrate the pattern of responses in 2AFC face-voice matching tasks, it is unnecessary to include the same identity stimulus. In Experiments 3a and 3b, the same identity target stimulus was not present at test. Removing the signal emanating from the target by including two (different identity) distractor stimuli allows for a clearer test of the position bias hypothesis. Rather than testing whether participants could discriminate between the target and distractor, the aim was to measure whether there was a bias to select the first or second sequentially presented test alternative in a 2AFC task. If the procedure is unbiased, alternatives in position 1 and 2 should be selected equally as often as each other. Therefore, in this set of two experiments, instead of the dependent variable being matching accuracy, it was the percentage of responses selecting the first test alternative (the stimulus in position 1) as being the same identity target. As the experiments presented in Chapter 5 were the first ever to analyse face-voice matching data for a position effect, the experiments in Chapter 6 were also undertaken in part to test whether the effect would be replicated.

6.2 Experiment 3a: Position bias and the 2AFC matching task: Sequential face-voice presentation

Experiment 3a used a cross-modal matching procedure (Lachs, 1999) to compare position biases in static and dynamic face-voice matching. In light of the results from Experiment 2a, we expected that a position bias would operate, with the alternative in position 1 being selected as the same identity target more often than the alternative in position 2.

6.2.1 Methods

The methods for Experiment 3a were identical to Experiment 2a, apart from the following exceptions:

6.2.1.1 Design

Experiment 3a employed a 2 x 2 mixed factorial design. The between subjects factor was facial stimulus type (static or dynamic), and the within subjects factor was order (visual to auditory (V-A) or auditory to visual (A-V)). The dependent variable was a *position 1* response (i.e. selecting the stimulus in position 1 as the same identity target).

6.2.1.2 Participants

There were 12 male and 28 female participants (*N*=40), ranging from 18 to 35 years (*M* = 21.98, *SD* = 4.40). They were recruited by convenience sampling and from the Nottingham Trent University Psychology Division's Research Participation Scheme. In accordance with this scheme, students received research credits in return for their participation. All of the participants reported having normal or corrected vision and hearing, and none had taken part in any previous experiments.

6.2.1.3 Apparatus and materials

For each of the 4 versions of the experiment, the stimulus set was re-randomised using an online research randomiser (Urbaniak & Plous, 2013) to construct trials consisting of different stimuli combinations to Experiment 2a. In the V-A condition, a face (stimulus 1) was followed by 2 sequentially presented voices (stimuli 2 and 3): both of them were a different identity to the voice. In the A-V condition, a voice (stimulus 1) was followed by 2 sequentially presented different identity faces (stimuli 2 and 3). As in Experiment 2a, each of the 4 versions was used for the between subjects manipulation of facial stimulus type (static or dynamic), so in total there were 8 versions of the experiment: 4 featuring static facial stimuli and 4 featuring dynamic facial stimuli.

6.2.1.4 Procedure

The participants received identical instructions to the participants in Experiment 2a. They were not informed that trials consisted entirely of distractor stimuli and that the same identity target would never be present at test.

6.2.2 Results

Matching performance was analysed using the same method as Experiment 2a. However, because there were only two factors (facial stimulus type and order), three nested models were compared, all fitted using restricted maximum likelihood, and with response: position 1 (0 or 1) as the dependent variable. The first model included a single intercept; the second included the main effects of each factor (order, facial stimulus type), and the third added the two-way interaction.

Table 6.1 shows the profile likelihood chi-square statistic (G^2) and p value associated with dropping each effect in turn from the appropriate model. Coefficients and standard errors (on a log odds scale) for each effect in the full two-way interaction model are also reported in Table 6.1. In Experiment 2a, variability for the first stimulus in each trial (the voice in the A-V condition, and the face in the V-A condition) was modelled separately from the foil stimulus because same identity faces and voices were highly correlated. However, in Experiment 3a, each trial featured stimuli from 3 different identities, so random effects for each stimulus (1, 2 and 3) were all modelled separately. In the two-way model, the estimate of *SD* of the first stimulus random effect was 0.458, for the second stimulus it was 0.357, and for the third stimulus it was 0.303. The estimated *SD* for the participant effect was less than 0.001. A similar pattern held for the null model.

Table 6.1

Parameter estimates (b) and profile likelihood tests for the 2x2 factorial analysis, Experiment 3a: Position bias and the 2AFC matching task: Sequential face-voice presentation

Source	*df*	*b*	*SE*	G^2	*p*
Intercept	1	0.125	0.257	.	.
Order	1	0.250	0.291	0.18	.669
Facial stimulus type	1	0.262	0.271	0.26	.609
Order x Facial stimulus type	1	0.323	0.382	0.69	.407

There were no main effects and no interactions (*p* > .407).

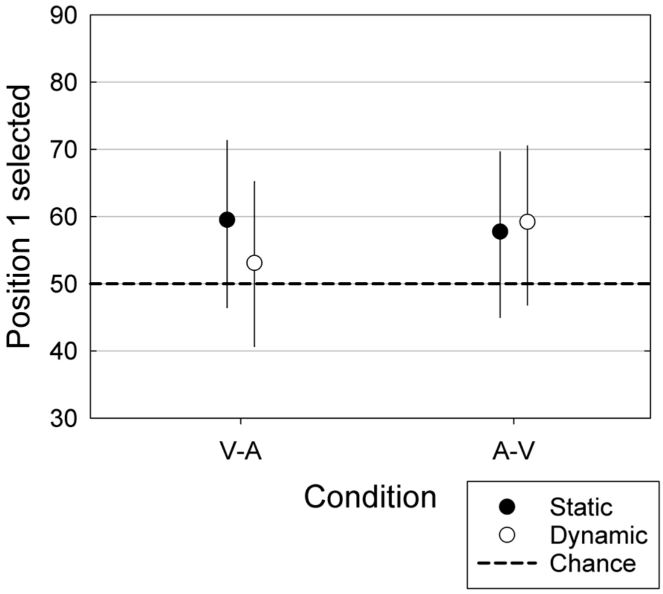


Figure 6.1: *Position 1 responses on V-A (panel A) and A-V (panel B) trials for sequentially presented faces and voices in a 2AFC matching task. Error bars show 95% CI for the condition means*

Figure 6.1 shows the means and 95% confidence intervals for the percentage of ‘position 1’ responses in each condition. Overall, the stimulus in position 1 was not selected significantly above chance level, *M* = 57.38%, 95% CI [48.12, 66.23].

6.2.3 Discussion

Although the stimulus in position 1 was not selected significantly above chance level, descriptively speaking the mean response favours position 1 in each of the 4 conditions. The descriptive statistics correspond with the pattern of results observed in Experiment 2a, suggesting that regardless of whether the same identity stimulus is present at test, on balance the stimulus in position 1 is slightly more likely to be selected. Also in keeping with the results of Experiment 2a, there was no effect of facial stimulus type.

6.3 Experiment 3b: Position bias and the 2AFC matching task: Simultaneous face-voice presentation

Experiment 3b compared position biases in static and dynamic face-voice matching using the same procedure as Experiment 2b, in which face-voice combinations were presented simultaneously. This experiment tested whether a position bias operates when the same identity target stimulus is absent at test. Based on the results of Experiment 2b, we anticipated that the alternative in position 1 would be selected more often than the alternative in position 2. However, owing to the results of Experiment 3a, it was unclear whether the imbalance would reach significance.

6.3.1 Methods

The methods for Experiment 3b were identical to Experiment 3a, apart from the following exceptions:

6.3.1.1 Participants

There were 6 male and 34 female participants (*N*=40), with an age range of 18 to 48 years (*M* = 21.98, *SD* = 6.94). None of the participants had taken part in previous experiments.

6.3.1.2 Apparatus and materials

Each of the 8 versions of the experiment was identical to Experiment 2b, apart from the fact that the re-randomised stimuli featured in Experiment 3a were used to construct trials featuring stimuli from 3 different identities.

6.3.2 Results

Matching performance was analysed using the same method as Experiment 3a. Table 6.2 shows the profile likelihood chi-square statistic (G^2) and *p* value associated with dropping each effect in turn from the appropriate model. Coefficients and standard errors (on a log odds scale) for each effect in the full two-way interaction model are also reported in Table 6.2. In the two-way model, the estimate of *SD* of the first stimulus random effect was 0.001, for the second stimulus it was 0.303, and for the third stimulus it was 0.303. The estimated *SD* for the participant effect was less than 0.001. A similar pattern was observed in the null model.

Table 6.2

Parameter estimates (b) and profile likelihood tests for the 2x2 factorial analysis, Experiment 3b: Position bias and the 2AFC matching procedure: Simultaneous face-voice presentation

Source	*df*	*b*	*SE*	G^2	*p*
Intercept	1	0.250	0.219	.	.
Order	1	0.105	0.274	0.18	.674
Facial stimulus type	1	0.035	0.262	0.08	.779
Order x Facial stimulus type	1	0.035	0.370	0.01	.926

There were no main effects and no interactions (*p* > .674). Figure 6.2 shows the means and 95% confidence intervals for the percentage of responses selecting the stimulus in position 1 in each condition. Overall, the stimulus in position 1 was not selected significantly above chance level, *M* = 55.56%, 95% CI [48.66, 62.28].

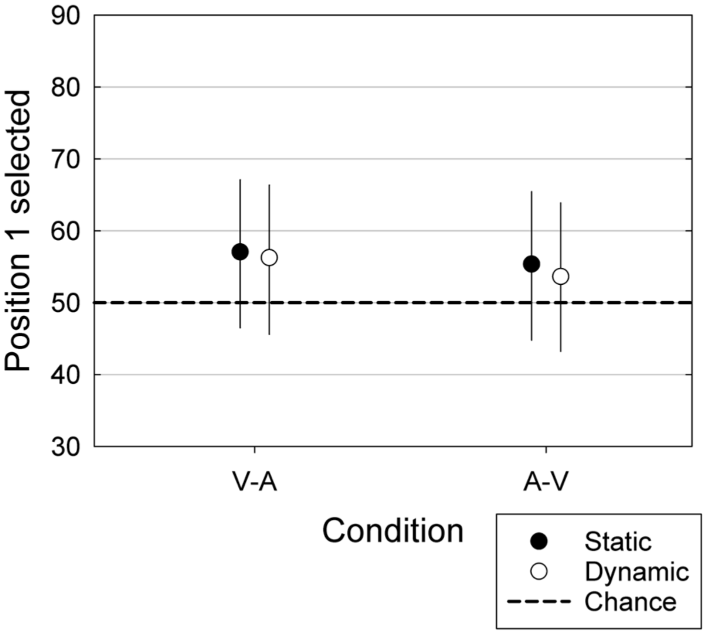


Figure 6.2: ‘*Position 1’ responses on V-A (panel A) and A-V (panel B) trials for sequentially presented faces and voices in a 2AFC matching task. Error bars show 95% CI for the condition means*

6.3.3 Discussion

Experiment 3b shows the same pattern of results as Experiment 3a, with the mean response favouring position 1 in each of the 4 conditions. Despite the face-voice combination in position 1 not being selected significantly above chance level, the descriptive statistics correspond with the results of Experiment 2b, which pointed to the existence of a bias to respond *position 1* across all conditions. As in Experiments 2a and 2b, the same pattern of responses is observed when face-voice combinations are presented sequentially (Experiment 3a) or simultaneously (Experiment 3b). This suggests that any bias should not be attributed to higher memory load during sequential presentation. In Experiments 2a and 3a the participants had to hold the first stimulus in mind across the two alternatives, whereas when the stimuli are presented simultaneously they do not have to refer to a stored mental representation in order to consider a face-voice combination. Once again, the distribution of *position 1* and *position 2* responses is consistent across static and dynamic facial stimulus trials.

6.4 General Discussion

In line with hypotheses based on the findings reported in Experiments 2a and 2b, the distribution of *position 1* and *position 2* responses in every condition across Experiments 3a and 3b indicates that a temporal position bias operates in 2AFC face-voice matching tasks when the target is not present. This finding is not consistent with the position bias being attributable to differing sensitivity across positions (Yeshurun et al., 2008). Rather, the effect appears to reflect decision bias under uncertainty (García-Pérez & Alcalá-Quintana, 2010). As expected, this bias does not vary according to whether the facial stimuli are static or dynamic.

The position bias detected in both Experiment 3a and 3b must be interpreted with care. The magnitude of this bias should not be overstated, as it is not statistically above chance level. That said, across two experiments, the mean response in all 8 conditions was consistently above 50%. Although the mean response selecting the option in position 1 was always less than 60% (10% above chance), it is useful to refer back to the results presented in Figures 5.2 and 5.4. The strength of the position biases across Experiments 2a, 2b, 3a and 3b are comparable. Indeed, a reduction in accuracy in Experiments 2a and 2b of less than 10% when the target was in position 1 (with an accompanying 10% increase in accuracy when the target was in position 2) would, in the majority of conditions, be sufficient to flatten out the pattern of results and make the position bias disappear. Evidence from Experiments 2a and 2b therefore suggests that a bias of similar strength to that observed in Experiments 3a and 3b translates into a significant difference in accuracy when the target appears in position 1 compared to position 2. Therefore, it is sensible to interpret the results presented in this chapter as reflecting a small, albeit non-significant, temporal position bias. This offers further corroborating evidence that 2AFC procedures may not be altogether appropriate for investigating face-voice matching.

The position bias might favour the alternative in position 1 because faces and voices are most commonly encountered close together in time during social interactions. It makes intuitive sense that faces and voices encountered together would belong to the same person. This could be expressed as a bias for people to accept a face and voice presented in relative temporal proximity (Experiment 2a and 3a), or the first face-voice combination they encounter (Experiments 2b and 3b), as belonging to the same identity. However, the basis for this position bias is unclear (Yeshurun et al., 2008). As biases for selecting the alternative in position 1 have also been identified in a wide range of unrelated 2AFC tasks, the pattern of results reported here may merely be attributable to the nature of the 2AFC procedure, rather than being specific to face-voice matching (Dyjas et al., 2012).

In order to more appropriately test whether people exhibit a bias to accept a face and voice as belonging to the same person, it is necessary to employ a same-different procedure. Unlike 2AFC tasks, same-different procedures are designed to measure both detection sensitivity and response bias. This procedure is therefore more appropriate for investigating response bias in face-voice matching.

6.4.1 Conclusion

Taken together, the results of Experiments 2a, 2b, 3a and 3b support the need for caution when employing 2AFC procedures (Yeshurun et al., 2008), showing that this warning generalises to face-voice matching. The following chapter investigates face-voice matching using a same-different task. Use of this methodology provides an opportunity to explore whether there is converging evidence for static face-voice matching, and to examine possible response biases in more detail.

7. CHAPTER 7: MATCHING NOVEL FACE AND VOICE IDENTITY USING SAME-DIFFERENT PROCEDURES

7.1 Introduction

The results of Chapter 5 offer compelling evidence that a temporal interval bias operates in 2AFC face-voice matching procedures. At test, the alternative presented in position 1 is differentially favoured over the alternative presented in position 2. The 2AFC procedure does not represent an unbiased way of testing face-voice matching, contrary to assumptions based on the previous literature (Green & Swets, 1966; Macmillan & Creelman, 2005; Thurstone 1927a, 1927b; Wickens, 2002). Chapter 6 showed that the position effect likely reflects decision bias under uncertainty (García-Pérez & Alcalá-Quintana, 2010). It is possible that people display a bias to accept a face and voice as belonging to the same person. In a 2AFC task in which alternatives are presented sequentially, this might manifest as an increased tendency to select the test alternative presented in position 1 because of its temporal proximity to the other-modality stimulus. One way to test the existence of a response bias is to use a same-different procedure, in which the participants are shown two stimuli and have to decide whether they are the same or different. This has not been previously undertaken; other face-voice matching studies (Kamachi et al., 2003; Krauss et al. 2002; Lachs & Pisoni, 2004a, 2004b; Lander et al., 2007; Mavica & Barenholtz, 2013) have all employed versions of the 2AFC matching procedure.

Although dynamic face-voice matching was consistently above chance in all three 2AFC experiments reported in Chapter 5, static face-voice matching was only above chance when test alternatives were presented simultaneously (Experiment 2c). The results were interpreted as providing evidence that accurate static face-voice matching is possible, although it is sensitive to the type of experimental procedure employed. So far static face-voice matching has only been above chance level in one experiment. It is necessary to run further tests of static face-voice matching in order to strengthen the conclusion that this finding is statistically robust.

7.1.1 Aim

In Chapter 7, static and dynamic face-voice matching was explored using a sequential (Experiment 4a) and simultaneous (Experiment 4b) same-different procedure. The intention was to provide a further test of whether static faces and voices offer concordant information. Using a same-different matching procedure for the first time, the experiments presented in this chapter test how response biases operate in face-voice matching.

7.2 Experiment 4a: Sequential face-voice presentation in a same-different matching task

As static faces and voices provide concordant information (Experiment 1), and static face-voice matching has been shown to be possible when the temporal element of comparing alternatives is removed (Experiment 2c), overall static face-voice matching accuracy was expected to be above chance level when tested using a sequential same-different procedure. Based on evidence from 2AFC tasks (Experiments 2a, 2b, 3a and 3b), which suggests that people accept the first face-voice identity combination they are presented with, employing a same-different task was expected to reveal the presence of a response bias, reflecting an overall tendency to accept a face and voice as belonging to the same person.[3]

7.2.1 Methods

7.2.1.1 Design

[3] The data from Experiment 4a have been published (Smith, Dunn, Baguley & Stacey, 2016a) (see Appendix D)

Experiment 4a employed a 2 x 2 x 2 mixed factorial design. In the matching accuracy analysis, the between subjects factor was facial stimulus type (static or dynamic). The within subjects factors were identity (same or different) and order (visual to auditory (V-A) or auditory to visual (A-V)). The dependent variable was matching accuracy.

The matching response analysis employed the same 2 x 2 x 2 mixed factorial design, but the dependent variable was a *same identity* response.

7.2.1.2. Participants

There were 40 male and 40 female adult participants (*N*=80) with an age range of 18 to 66 years (*M* = 25.44, *SD* = 8.36). Participants were recruited by convenience sampling and from the Nottingham Trent University Psychology Division's Research Participation Scheme. In accordance with this scheme, students received research credits in exchange for participation. All participants reported having normal or corrected vision and hearing. None had taken part in previous experiments.

7.2.1.3 Apparatus and materials

Experiment 4a used identical apparatus to that of previous experiments. Four different versions of the experiment were created so that same identity and different identity face-voice combinations could be constructed using different stimulus people. For each of the versions, stimuli were randomly selected to be used either for one of the 8 same identity or 8 different identity trials. None of the 18 faces or voices in the stimulus set appeared more than once in each version of the experiment. The stimuli that remained following randomisation were used for the practice trials. The combination of stimuli in each of the 4 versions was repeated for static and dynamic conditions, making a total of 8 versions.

7.2.1.4 Procedure

The participants were randomly allocated to one of the 8 versions of the experiment using an online research randomiser (Urbaniak & Plous, 2013). In the dynamic facial stimulus condition the participants were accurately informed that the face in the muted video and the voice in the recording were not saying the same thing. This was to prevent them using speech-reading to match the face and voice (Kamachi et al., 2003).

The participants completed 2 counterbalanced experimental blocks, each consisting of a practice trial, followed by 8 randomly ordered experimental trials. As illustrated in Figure 7.1, in one block the participants saw the face first (V-A), and in the other they heard the voice first (A-V). In each trial, there was a 1 second gap between presentation of the face and voice stimuli. When the face was visible the text 'Face' appeared below the face, and while the voice recording was being played the text 'Voice' was visible in the middle of the screen. At test, participants pressed *1* on the laptop keyboard if they thought the face and voice were from the same identity, and *0* if they thought they were from different identities.

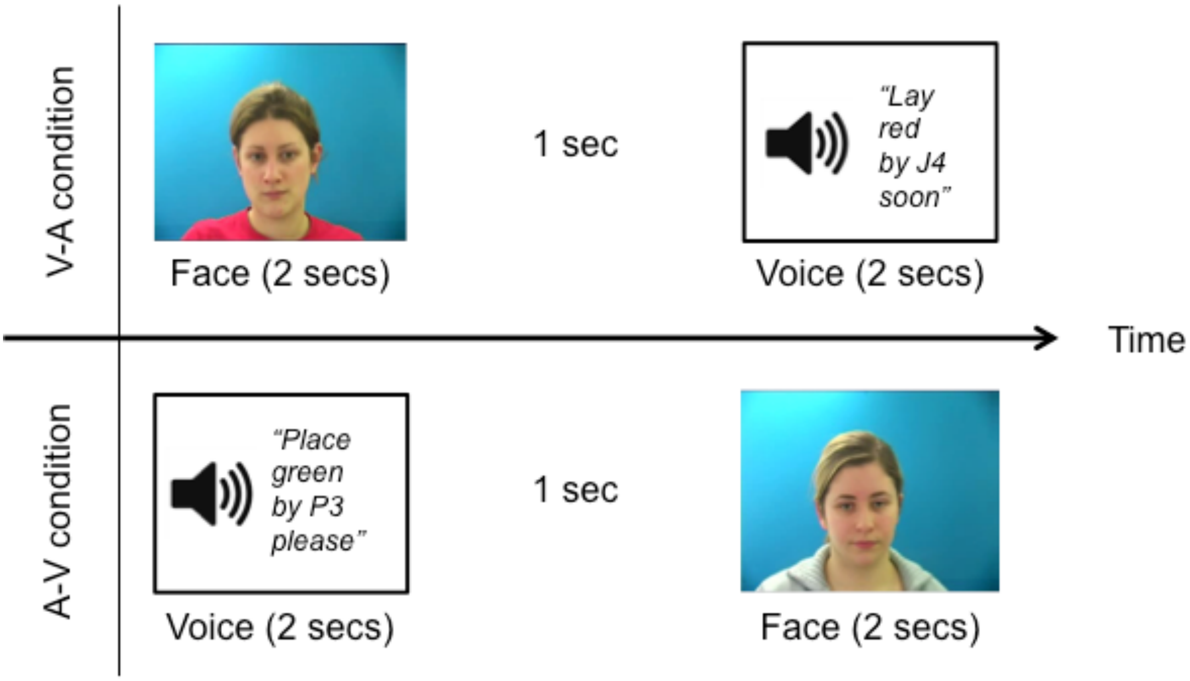


Figure 7.1: *An illustration of the procedure used in Experiment 4a*

7.2.2 Results

The traditional approach to signal detection involves partitioning same-different data into hits, false alarms, misses and correct rejections. For each participant, an aggregate measure of accuracy would be calculated, and statistics performed on these values. The problems associated with performing analyses on aggregate data are summarised in Chapter 3, and are particularly salient here because of the high level of variability associated with face and voice stimuli (Burton, 2013; Mathias & von Kriegstein, 2014; Mullenix & Pisoni, 1990; Valentine et al., 2015). The multilevel modelling analyses of previous experiments (Experiment 2a, 2b, 2c, 3a and 3b) show that stimulus variability is an important factor in face-voice matching. Therefore, the traditional approach to signal detection is not appropriate for the current set of data (Wright, Horry & Skagerberg, 2009).

Our analysis of the matching accuracy data is undertaken using multilevel modelling. It uses the hit rate as a measure of sensitivity and the true negative rate as a measure of specificity, rather than adopting the more common definitions of these terms (see section 2.6.4). Observed accuracy across same identity and different identity trials is compared against chance level performance (50%) in order to separate the signal from the noise. To measure response bias, *same identity* responses across all trials are compared against chance level.

7.2.2.1 Matching accuracy

As in previous chapters, matching performance was analysed using multilevel logistic regression (lme4 v. 1.06, Bates et al., 2014). Four nested models with matching accuracy (0 or 1) as the dependent variable were compared. All models were fitted using restricted maximum likelihood. The first model included a single intercept, and was later used to obtain confidence intervals for the overall accuracy. The second model also included the main effects of each factor (identity, order and stimulus type). The third model added all two-way interactions and the final model added the three-way interaction.

Table 7.1 shows the profile likelihood chi-square statistic (G^2) and *p* value associated with dropping each effect from the appropriate model. Table 7.1 also reports the coefficients and standard errors (on a log odds scale) for each effect in the full three-way interaction model. In the three-way model the estimate of *SD* of the face random effect was 0.353 while for voice it was 0.207. The estimated *SD* for the participant effect was less than 0.0001. A similar pattern held for the null model.

Table 7.1

Parameter estimates (b) and profile likelihood tests for the 2x2x2 factorial analysis of accuracy in Experiment 4a: Sequential face-voice presentation in a same-different matching task

Source	*df*	*b*	*SE*	G^2	*p*
Intercept	1	-0.445	0.196	.	.
Identity	1	1.382	0.254	57.84	< .001
Order	1	0.509	0.241	2.28	.131
Facial stimulus type	1	0.133	0.231	0.13	.717
Identity x Order	1	0.601	0.358	4.20	.040
Identity x Facial stimulus type	1	0.165	0.339	0.32	.572
Order x Facial stimulus type	1	0.052	0.324	0.01	.916
Identity x Order x Facial stimulus type	1	0.058	0.474	0.01	.903

Only the main effect of identity ($p < .001$) and the two-way interaction of identity and order ($p = .040$) were statistically significant. To aid interpretation of these effects, the means and confidence intervals were calculated for the percentage accuracy of the 8 conditions in the factorial design. These confidence intervals were obtained through simulations of the posterior distributions of the cell means using arm package version 1.6 in R (Gelman & Su, 2013). The means and the associated 95% confidence intervals are shown in Figure 7.2.

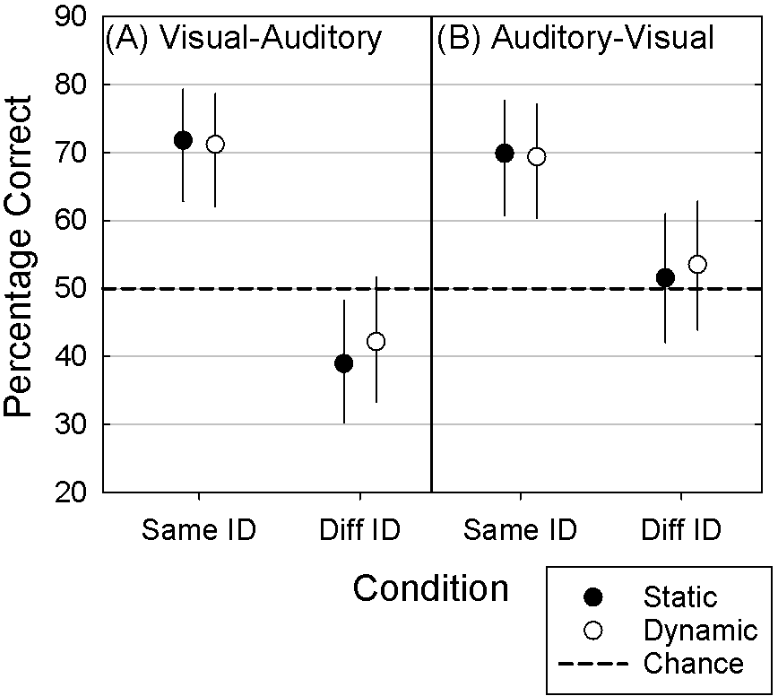


Figure 7.2: *Face-voice matching accuracy on V-A (panel A) and A-V (panel B) trials for sequentially presented faces and voices using a same-different matching task. Error bars show 95% CI for the condition means*

From Figure 7.2 it is clear that overall matching accuracy was significantly above chance (50%) level, *M* = 59.7 %, 95% CI [51.9, 66.9]. Static face-voice matching was above chance, *M* = 59.19%, 95% CI [50.94, 66.84], as was dynamic face-voice matching, *M* = 60.12%, 95% CI [51.97, 67.74]. Whilst performance on A-V trials was also above chance level, *M* = 62.78%, 95% CI [54.89, 70.03], performance on V-A trials was not, *M* = 56.45%, 95% CI [48.47, 64.11]. Figure 7.2 reveals the main effect of identity, with the hit rate (same identity trials) consistently higher than the true negative rate (different identity trials), and the former but not the latter consistently above chance. It also reveals the basis of the identity by order interaction. The results from the V-A trials are shown in panel A. The results from the

A-V trials are shown in panel B. Using visual analysis to guide an interpretation, it appears that the hit rate did not differ across conditions, but the true negative rate was higher in the A-V condition.

7.2.2.2 Matching response

Overall, faces and voices were accepted as belonging to the same person above chance level, *M* = 61.75%, 95% CI [54.99, 68.06]. This was the case in both the V-A condition, *M* = 64.53%, 95% CI [57.52, 71.08], and the A-V condition, *M* = 58.66,% 95% CI [51.38, 65.62].

7.2.3 Discussion

The results of the matching accuracy analysis replicate the findings of Experiment 2c, offering further evidence that static faces and voices offer sufficient concordant information that they can be matched above chance level. As in previous experiments, there was no significant difference between static and dynamic face-voice matching performance. Also in keeping with previous results (Experiments 2a and 2b), the hit rate (sensitivity) does not differ according to the order of stimulus presentation.

On different identity trials, the participants performed at chance level (A-V trials), or below chance level (V-A trials), and were significantly less accurate than on same identity trials; the hit rate was higher than the true negative rate. This points to the existence of a response bias. Using a same-different procedure, Experiment 6a shows for the first time how response biases manifest in face-voice matching. Analyses detected the existence of a bias to accept a face and voice as belonging to the same person, with overall *same identity* responses occurring above chance level. This reflects liberal response criterion placement.

Although *same identity* responses occurred above-chance level in both order conditions, the interaction between identity and order reveals that the bias to accept a face and voice as belonging to the same person was more pronounced in the V-A condition, when the face was presented before the voice. Therefore, the results suggest that the response bias differs according to stimulus presentation order in face-voice matching tasks. As faces are stronger cues to identity than voices (Damjanovic & Hanley 2007; Hanley & Turner 2000; Stevenage et al., 2011, 2012 2013; Stevenage & Neil, 2014; Stevenage, Neil & Hamlin, 2014b), it is possible that the voice in the V-A condition is swept up with the identity of the face, thereby increasing the likelihood that it will be accepted as belonging to the same identity as the face. Consistent with this explanation, the asymmetry observed in Experiment 4a corresponds with audiovisual integration studies, in which tolerance for stimulus offset is greater when the voice occurs after the face (Munhall, Gribble, Sacco & Ward, 1996; Robertson & Schweinberger, 2010; Van Wassenhove, Grant & Poeppel, 2007).

7.3 Experiment 4b: Simultaneous face-voice presentation in a same-different matching task

Although accuracy rates in 2AFC tasks operate similarly regardless of whether face-voice options are presented sequentially (Experiment 2a) or simultaneously (Experiment 2b), in order to test how response bias operates, Experiment 4b investigated face-voice matching using a simultaneous same-different procedure. Faces and voices were presented at the same time. Studies investigating audiovisual integration have shown that events are more likely to be perceived as emanating from the same source when they are presented simultaneously (Howard & Templeton, 1966). This effect is also observed for synchronous visual and auditory speech; integration occurs at lags of up to 300ms (Munhall et al., 1996; Robertson & Schweinberger, 2010; Van Wassenhove et al., 2007).

7.3.1 Methods

Experiment 4b used the same methods as Experiment 4a. Any exceptions are outlined below.

7.3.1.1 Design

Experiment 4b employed a 2 x 2 mixed factorial design. For the matching accuracy analysis, the between subjects factor was facial stimulus type (static or dynamic), and the within subjects factor was identity (same or different). The dependent variable was matching accuracy.

The matching response analysis employed the same 2 x 2 mixed factorial design, but the dependent variable was a *same identity* response.

7.3.1.2 Participants

There were 12 male and 36 female participants ($N = 48$), with an age range of 18 – 44 years ($M = 21.94$, $SD = 5.54$).

7.3.1.3 Procedure

The procedure used in Experiment 4b is illustrated in Figure 7.3. Participants saw a face and heard a recording of a voice presented at the same time. The face-voice combination was presented for 2 seconds. Participants pressed *1* on the laptop keyboard if they thought the face and voice belonged to the same identity, and *0* if they thought they were from different identities.

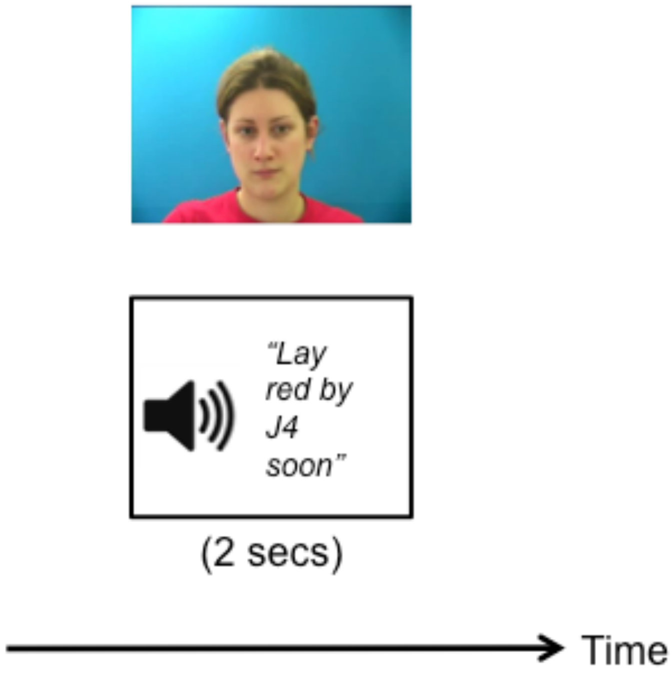


Figure 7.3: *An illustration of the procedure used in Experiment 4b*

7.3.2 Results

7.3.2.1 Matching accuracy

Matching accuracy was analysed using the same method as Experiment 4a. As there were only 2 factors, 3 nested models were compared, with accuracy (0 or 1) as the dependent variable. The first model included a single intercept, the second model included the main effects of each factor (identity and facial stimulus type), while the third model added the two-way interactions.

The profile likelihood chi-square statistic (G^2) and *p* value associated with dropping each effect from the appropriate model are shown in Table 7.2. Coefficients and standard errors (on a log odds scale) for each effect in the full two-way interaction model are also reported in this table. In the two-way model the estimate of *SD* of the face random effect was 0.399 while for voice it was 0.245. The estimated *SD* for the participant effect was less than 0.001. A similar pattern held for the null model.

Table 7.2

Parameter estimates (b) and profile likelihood tests for the 2x2 factorial analysis, Experiment 4b: Simultaneous face-voice presentation in a same-different matching task

Source	*df*	*b*	*SE*	G^2	*p*
Intercept	1	0.146	0.189	.	.
Identity	1	1.220	0.234	26.62	<.001
Facial stimuli type	1	0.195	0.207	0.27	.601
Identity x Facial stimulus type	1	0.599	0.307	3.75	.053

Only the main effect of identity was significant (*p* < .001). Figure 7.4 shows the means and 95% confidence intervals for percentage accuracy in each condition of the factorial design.

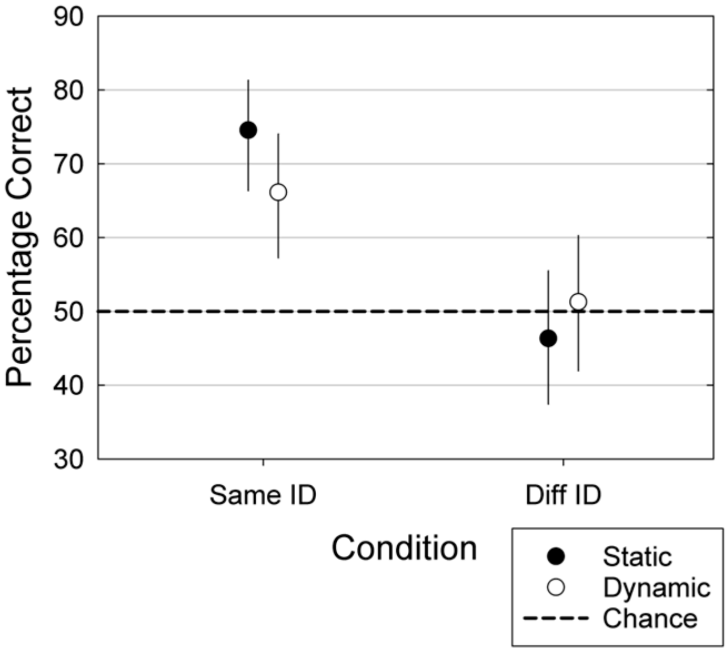


Figure 7.4: *Face-voice matching accuracy for simultaneously presented faces and voices using a same-different matching task. Error bars show 95% CI for the condition means*

Static, *M* = 61.33%, 95% CI [51.96, 69.83], and dynamic, *M* = 59.39%, 95% CI [50.13, 68.04] face-voice matching were both significantly above chance level. Figure 7.4

reveals the main effect of identity. The hit rate (same identity trials) was consistently higher than the true negative rate (different identity trials).

7.3.2.2 Matching response

Overall, faces and voices were attributed to the same identity above chance level, *M* = 61.20%, 95% CI [52.63, 69.22].

7.3.3 Discussion

The results of the matching accuracy analysis correspond to the results of Experiment 4a. Voices and static faces, as well as voices and dynamic faces, were accurately matched above chance level. The results indicate that simultaneously presenting face and voice stimuli does not make incongruent matches more obvious, as the true negative rate remained at chance level. In Experiment 4b, the pattern of accuracy according to identity was similar to that observed in Experiment 4a. Even when faces and voices are presented simultaneously, hit rates (same identity trials) are significantly higher than the true negative rate (different identity trials). As in Experiment 4a, there was an overall bias to assign faces and voices to the same identity.

7.4 General discussion

Experiments 4a and 4b used a same-different procedure, replicating the results of Experiment 2c to show that static face-voice matching is possible, both when faces and voices are presented sequentially and when they are presented simultaneously. The same-different procedure showed that people demonstrate a bias to assign faces and voices to the same identity in face-voice matching tasks. This bias is more pronounced when the face is presented before the voice (V-A condition). These are the first experiments to ever analyse response bias in face-voice matching.

7.4.1 Matching accuracy

Overall, the results of Experiments 4a and 4b detected accuracy levels significantly above chance level. The findings are consistent with previous findings (Krauss et al., 2002; Mavica & Barenholtz, 2013), and the conclusion of Chapter 5. People can use redundant information to match voices and dynamic faces, as well as voices and static faces, for identity. As in previous experiments, there was no difference between static and dynamic facial stimulus trials, further weakening the conclusion of audiovisual speech perception studies that accurate face-voice matching is wholly dependent on encoding visual articulatory movement (Kamachi et al., 2003; Lachs & Pisoni, 2004a).

Both the overall pattern of results and the observed accuracy rates were very similar across Experiments 4a and 4b. The results therefore replicate those of Experiments 2a and 2b, showing that accuracy does not differ according to whether the face-voice combination is presented sequentially or simultaneously. Limiting memory load by presenting faces and voices simultaneously does not appear to affect the hit rate on same identity face-voice matching trials, nor does simultaneous presentation increase the true negative rate on different identity trials by making incongruent identities more obvious.

Consistent with the results of the multilevel modelling analyses presented in Chapter 5, the pattern of variance associated with participants and stimuli in Experiments 4a and 4b further highlights that people vary in the extent to which they look and sound similar. These results offer further support to the conclusion that accurate face-voice matching is likely to be highly dependent on the particular stimulus set used.

7.4.2 Matching response

In both Experiments 4a and 4b there was a main effect of identity. The hit rate was higher than the true negative rate, showing that it is easier to accurately accept a face-voice match than to reject a mismatch. Whilst hit rates were consistently above chance level, the true negative rate indicated that participants were guessing on different identity trials. Owing to the exclusive adoption of 2AFC procedures, previous face-voice matching studies have not measured whether people have a bias to accept a face and voice presented in relative temporal proximity as sharing a common source identity (Kamachi et al., 2003; Krauss et al., 2002; Lachs & Pisoni, 2004a, 2004b; Lander et al., 2007; Mavica & Barenholtz, 2013). Experiments 4a and 4b showed for the first time that when matching novel face and voice identity, people do exhibit such a bias. This likely helps to explain the effect of identity present in both experiments; such a bias would particularly undermine accuracy on different identity trials when responding *same* is an incorrect response. In neither Experiment 4a nor Experiment 4b did the response bias differ according to facial stimulus type. Specificity was the same, regardless of whether the faces were static or dynamic.

7.4.2.1 Sequential vs. simultaneous face-voice presentation

In terms of overall *same identity* responses, the means and 95% CIs were very similar in both Experiment 4a and 4b. Despite previous audiovisual integration literature suggesting that common source attributions are more likely when stimuli are presented synchronously (Howard & Templeton, 1966) or with a small temporal offset (Munhall et al., 1996; Robertson & Schweinberger, 2010; Van Wassenhove et al., 2007), there was no evidence of a stronger bias to respond *same identity* when the face and voice were presented simultaneously in Experiment 4b. The lack of difference between simultaneous and sequential presentation in terms of response bias may be explained by the fact that the dynamic faces and voices in this study were not saying the same sentence. On the basis of results showing that voice recognition is compromised by the presentation of timesynchronised articulating faces of a different identity, but not static faces of a different identity, Schweinberger et al. (2007) argued that integration does not occur for static faces and voices. Taken together with the results of other audiovisual integration literature, this would suggest that speech synchrony, even if slightly offset, is a key component in explaining integration. In a situation when the face and voice say different sentences, there is no speech synchrony. This should not be taken as a dismissal of explanations based on integration, as the task adopted by Schweinberger et al. (2007) involves an indirect measure of whether people integrate the identity of faces and voices. Overall it seems that a general bias to accept a face and voice as belonging to the same person, as observed in this set of two experiments, regardless of whether the face is dynamic or static, may provide at the very least a useful foundation for audiovisual integration, thereby helping to facilitate social communication.

7.4.2.2 Order of stimulus presentation

In keeping with the results of Chapter 5, as well as previous face-voice matching studies using 2AFC paradigms (Kamachi et al. 2003; Lachs & Pisoni 2004a, 2004b; Lander et al., 2007), Experiment 4a found no difference between V-A and A-V performance in terms of sensitivity. However, the interaction between identity and order observed in Experiment 4a showed that response bias varies according to stimulus order. Specificity was higher in the A-V condition, showing that participants exhibited a more liberal response criterion in trials when the face was presented before the voice (V-A condition). A performance asymmetry according to stimulus order is consistent with previous literature highlighting differences in the way these two stimuli are processed. There is no reason to assume that performance on V-A and A-V face-voice matching trials should be identical. Given that voices carry more speech information than faces (Lachs & Pisoni, 2004a), and faces carry more reliable identity information than voices (e.g. Damjanovic & Hanley 2007; Hanley & Turner 2000; Stevenage et al., 2011, 2012 2013; Stevenage & Neil, 2014; Stevenage et al., 2014b), this may influence criterion placement when a decision is being made about common source identity. If the face is presented first, the voice may be automatically encompassed by the identity of the preceding face, and processed primarily for speech information, rather than being interrogated for identity information.

The nature of the response bias asymmetry observed here is consistent with patterns of results from audiovisual integration studies. Face-voice integration occurs from an auditory lead of up to around 100ms, and an auditory lag of around 300ms (Munhall et al., 1996; Robertson & Schweinberger, 2010; Van Wassenhove et al., 2007). The results presented in this chapter hint at the existence of parallel biases in face-voice matching and audiovisual face-voice integration, such that there is a greater tendency to accept a face and voice as belonging to the same person when there is an auditory lag (V-A condition) compared to when there is a visual lag as in the A-V condition. This supports the argument that the general bias to accept a face and voice as belonging to the same person is useful in supporting audiovisual integration.

7.4.3 Conclusion

The set of results presented in this chapter build on those presented in Chapter 5. Taken together, the results justify adopting the working conclusion that static face-voice matching is possible. By modelling response bias, the adoption of same-different procedures has detected the existence of a general bias to accept a face and voice as sharing common source identity, as well as more liberal response criterion placement on V-A trials.

8. CHAPTER 8: THE EFFECT OF INCREASING THE INTER-STIMULUS INTERVAL ON FACE-VOICE MATCHING PERFORMANCE

8.1 Introduction

The procedures adopted in all the experiments reported so far in this thesis, as well as in the previous literature (Kamachi et al., 2003; Krauss et al., 2002; Lachs, 1999; Lachs & Pisoni, 2004a, 2004b; Lander et al., 2007; Mavica & Barenholtz, 2013), have presented faces and voices close together in time, with a maximum 1 second gap between each stimulus in matching tasks. In previous chapters, the results show that both dynamic faces and voices, as well as static faces and voices, offer concordant information (Chapter 4), that it is possible to accurately match (static and dynamic) faces and voices for identity (Chapters 5 and 7), and that participants exhibit a bias to respond that novel faces and voices belong to the same identity (Chapter 6 and 7). Chapter 8 addresses whether a similar pattern of results holds when faces and voices are temporally offset to a greater extent (> 1 second).

Chapter 5, 6 and 7 suggest that reducing the memory load by using a simultaneous rather than sequential procedure does not affect the overall pattern of responses (Chapter 6), the hit rate (Chapter 5 and 7), or the true negative rate (Chapter 7). This could be taken to indicate that reducing the memory load does not influence face-voice matching performance. However, a maximum 1 second interval between sequentially presented faces and voices means that the results do not reflect how matching performance may operate in everyday social situations, when faces and voices might be offset by greater time intervals.

Precise representations of both visual and auditory information degrade quickly, so it is possible that increasing the inter-stimulus interval beyond 1 second will affect overall face-voice matching accuracy. Iconic memory, the brief storage of highly detailed visual information, typically lasts for a few hundred milliseconds (Coltheart, 1980; Neisser, 1967; Sperling, 1960). It may last longer though; recent evidence has been put forward for the existence of an intermediate, high capacity visual store persisting for up to 4 seconds with the help of afterimages (Sligte, Scholte & Lamme, 2008, 2009). Visual information is then transferred to the limited capacity visual short-term memory (VSTM) system where it is stored temporarily for anything up to 30 seconds (Blake, Cepeda & Hiris, 1997; Magnussen, Idås & Myhre, 1998; Pasternak & Greenlee, 2005). The time-course of the degradation of auditory stimuli is slightly different from that of visual stimuli. Echoic memory, the auditory equivalent of iconic memory, persists for longer (Crowder & Morton, 1969; Penney, 1985), up to a period of about 5 seconds (Glanzer & Cunitz, 1966; Lu, Williamson & Kaufman, 1992; Treisman, 1964; Wickelgren, 1969). Auditory information then follows the same sequence of storage as visual information, passing into the limited capacity auditory short-term memory (ASTM) store (Baddeley, 2007).

The short inter-stimulus intervals employed in previous studies (Kamachi et al., 2003; Krauss et al., 2002; Lachs & Pisoni, 2004a; Lander et al., 2007; Mavica & Barenholtz, 2013) is likely within the limits of both iconic and echoic memory, meaning that high quality representations of faces and voices can be compared to each other for source-identity information. The more precise mental representations of faces and voices are, the more accurate we might expect face-voice matching to be. A short inter-stimulus interval may facilitate comparisons between the stimuli, thereby supporting sensitivity.

If the bias to respond *same identity* (Chapter 7) is temporally dependent, then increasing the inter-stimulus interval may also affect response bias. Certainly the nature of the bias observed in Chapters 5 and 6, which showed a temporal position effect in 2AFC tasks, does suggest that this might be the case. Participants exhibited a bias to accept a face and voice presented in relative temporal proximity as sharing a common identity. The research relating to the beneficial effects of temporal contiguity in facilitating associations between events and stimuli support the hypothesis that attributions of common identity will be more likely when faces and voices are presented within a brief time-frame. For example, when two events are presented close together in time, attributions of causality are inferred; a 2 second window appears to be the crucial time period (Reed, 1992; Shanks, Pearson & Dickinson, 1989), although the exact length of time is likely to depend on expectations associated with the specific stimuli (Buehner & May, 2003). The educational psychology literature has repeatedly demonstrated significant learning gains when temporal contiguity between information is increased (for a meta-analysis, see Ginns, 2006).

Temporal contiguity is also relevant to face and voice processing (Stevenage et al., 2014b). Audiovisual speech perception research suggests that face-voice speech integration occurs during a short temporal window (Munhall et al., 1996; Robertson & Schweinberger, 2010; Van Wassenhove et al., 2007). There might be a corresponding temporal window of opportunity during which people exhibit a bias to attribute a novel face and voice to the same identity.

No differences between static and dynamic facial stimulus trials have been observed in previous experiments presented in this thesis. To date, no experiments have tested how short-term memory for static and dynamic faces might influence both sensitivity and specificity in face-voice matching. As there is evidence for better memory for dynamic compared to static facial stimuli (e.g. Knappmeyer et al., 2003; Lander & Chuang, 2005), it is possible that dynamic face-voice matching accuracy will persist better than static face-voice matching over longer inter-stimulus intervals.

8.1.1 Aim

This chapter considers how face-voice matching performance is affected by increasing the inter-stimulus interval to 5 seconds (Experiment 5a and 5c) and 10 seconds (Experiment 5b and 5d). Previous results indicated that response biases are an important element of novel face-voice matching performance (Experiments 2a, 2b, 3a, 3b, 4a and 4b). Same-different tasks are the most appropriate procedure to employ here, because they measure this aspect of performance. Thus, a sequential same-different procedure is adopted in all experiments reported in Chapter 8.

8.2 Experiment 5a: Face-voice matching using a sequential same-different task: 5 second inter-stimulus interval

Experiment 5a tested static and dynamic face-voice matching using the same procedure as Experiment 4a, but with an inter-stimulus interval of 5 seconds. Previous studies have tested face-voice matching using short (<1 second) inter-stimulus intervals (Kamachi et al., 2003; Krauss et al., 2002; Lachs, 1999; Lachs & Pisoni, 2004a; Lander et al., 2007; Mavica & Barenholtz, 2013). An interval of 5 seconds is likely to be the absolute temporal limit of high-capacity sensory storage, the point at which auditory and visual information could reasonably be expected to have transferred to the lower capacity short-term memory store (Glanzer & Cunitz, 1966; Lu, et al., 1992; Sligte et al., 2008, 2009; Treisman, 1964; Wickelgren, 1969). If accurate face-voice matching relies on the ability to compare highly detailed mental representations of faces and voices, performance may be at chance level when there is an inter-stimulus interval of 5 seconds. If the bias to respond *same identity* only operates when faces and voices are presented within a short temporal window, it is possible that overall *same identity* responses will also be at chance level.

8.2.1 Methods

The methods were the same as those used in Experiment 4a. Any exceptions are outlined below.

8.2.1.1 Participants

There were 48 participants (46 females and 2 males), with an age range of 18 to 35 years (*M* = 19.73, *SD* = 3.39). They were recruited by convenience sampling and from the Nottingham Trent University Psychology Division's Research Participation Scheme. Students received research credits in return for their participation. All of the participants reported having normal or corrected vision and hearing, and none of them had taken part in previous experiments.

8.2.1.2 Apparatus and Materials

In all previous experiments, the participants listened to voices binaurally through Apple EarPods. In Experiment 5a, and in all future experiments, voices were presented binaurally through Sennheiser (HD 205) headphones. The decision to change was due to apparatus availability, and is unlikely to affect comparisons across experiments because both the Apple EarPods and Sennheiser (HD205) headphones have a frequency response exceeding the range of human hearing. The main advantage of the Sennheiser (HD205) headphones relates to the superior suppression of external and ambient noise. This is particularly important for the present experiment, which features a silent 5-second interval between the face and voice.

8.2.1.3 Procedure

The procedure is illustrated in Figure 8.1.

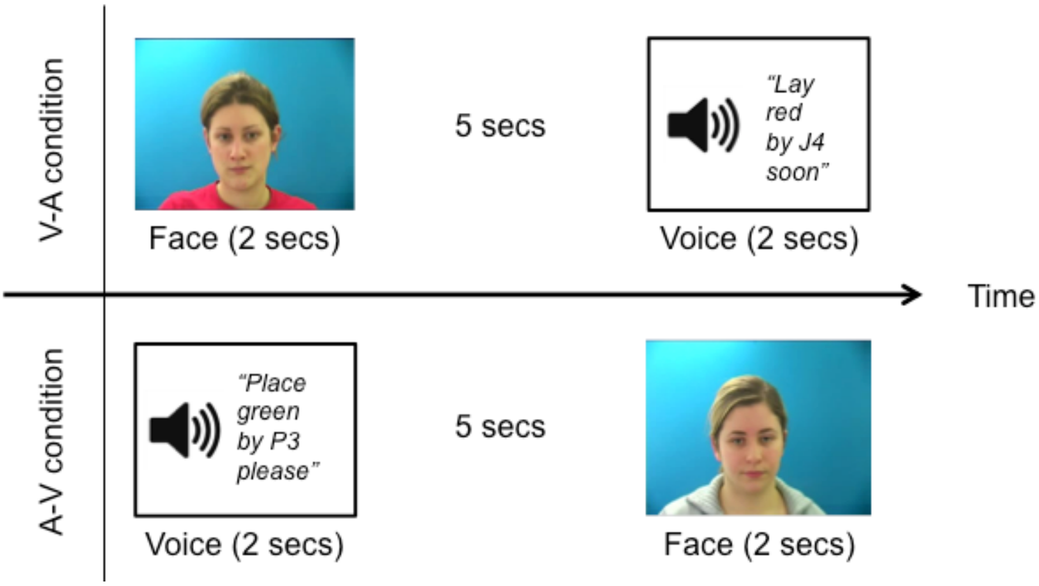


Figure 8.1: *An illustration of the procedure used in Experiment 5a*

8.2.2 Results

8.2.2.1 Matching accuracy

Matching accuracy was analysed using the same methods as Experiment 4a. Table 8.1 shows the profile likelihood chi-square statistic (G^2) and p value associated with dropping each effect from the appropriate model. Table 8.1 also reports the coefficients and standard errors (on a log odds scale) for each effect in the full three-way interaction model. In the three-way model the estimate of *SD* of the face random effect was 0.453 while for voice stimulus it was 0.161. The estimated *SD* for the participant effect was less than 0.001. A similar pattern held for the null model.

Table 8.1

Parameter estimates (b) and profile likelihood tests for the 2x2x2 factorial analysis, Experiment 5a: Face-voice matching using a sequential same-different task: 5 second inter-stimulus interval

Source	*df*	*b*	*SE*	G^2	*p*
Intercept	1	0.301	0.247	.	.
Identity	1	1.284	0.331	16.48	<.001
Order	1	0.472	0.310	0.69	.406
Facial stimulus type	1	0.048	0.298	3.32	.069
Identity x Order	1	1.116	0.461	13.00	<.001
Identity x Facial stimulus type	1	0.145	0.443	<0.01	.979
Order x Facial stimulus type	1	0.459	0.427	1.15	.284
Identity x Order x Facial stimulus type	1	0.268	0.620	0.19	.665

There was a main effect of identity (*p* < .001). The interaction between identity and order was significant (*p* < .001). The cell means and 95% confidence intervals for matching accuracy are shown in Figure 8.2.

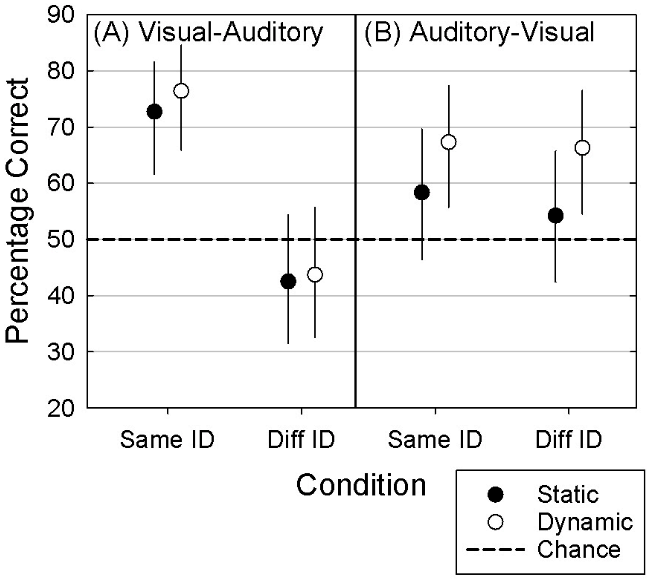


Figure 8.2: *Face-voice matching accuracy on V-A (panel A) and A-V (panel B) trials with a 5-second inter-stimulus interval. Error bars show 95% CI for the condition means*

Overall accuracy was above chance level, *M* = 61.42%, 95% CI [53.02, 69.25]. Performance was above chance level for dynamic facial stimulus trials, *M* = 64.70%, 95% CI [55.57, 72.89], but not for static facial stimulus trials, *M* = 58.11%, 95% CI [48.79, 66.85]. In terms of stimulus order, although matching accuracy was above chance level for V-A trials, *M* = 63.58%, 95% CI [54.29, 71.84], it was not for A-V trials, *M* = 59.15%, 95% CI [49.83, 67.85]. The hit rate (same identity trials) was consistently above chance level, *M* = 69.00%, 95% CI [61.74, 75.35], but the true negative rate (different identity trials) was not, *M* = 51.79%, 95% CI [44.02, 59.52]. As illustrated in Figure 8.2, the main effect of identity reveals that the hit rate was reliably higher than the true negative rate. Based on visual inspection, it seems that the interaction between identity and order reflects the true negative rate being higher in the A-V condition (panel B) than in the V-A condition (panel A).

8.2.2.2 Matching response

Overall, *same identity* responses were not made significantly above chance level, *M* = 59.10%, 95% CI [48.85, 68.62]. Faces and voices were attributed to the same identity above chance level in V-A trials, *M* = 62.98%, 95% CI [52.08, 72.79], but not in A-V trials, *M* = 55.18%, 95% CI [44.09, 65.98].

8.2.3 Discussion

The results of the matching accuracy analysis show that when faces and voices are separated by an inter-stimulus interval of 5 seconds, overall it is still possible to match the two for identity. However, matching accuracy on A-V trials, as well as static facial stimulus trials, was at chance level. Performance in both of these conditions was above chance level in Experiment 4a when the inter-stimulus interval only lasted for 1 second.

Overall, there was not a bias to accept a face and voice as belonging to the same person when the stimuli were separated by 5 seconds. *Same identity* matching responses were not made above chance level. This finding supports the hypothesis that biases in face-voice matching are influenced by the degree of temporal contiguity (Buehner & May, 2003; Ginns, 2006; Reed, 1992; Shanks et al., 1989), because when the inter-stimulus interval was shorter (1 second), participants did make *same identity* responses above chance level (Experiment 4a).

Experiment 5a showed the same pattern of results as Experiment 4a, with a main effect of identity and 2-way interaction between order and identity. The basis of this interaction is that whilst sensitivity did not differ across conditions, the true negative rate (specificity) was lower in the V-A condition. Both Experiments 4a and 5a therefore highlight the existence of a stronger bias to respond *same identity* when the face is presented before the voice. The bias observed in Experiment 4a was explained in terms of strong identity cues associated with faces sweeping up the subsequent voice and making participants more likely to respond *same identity* in the V-A condition. Experiment 5a shows that the bias endures over a 5 second inter-stimulus interval, further highlighting the potency of facial identity cues in comparison to those associated with voices. This interpretation is supported by the results of the matching response analysis. There was a significant bias to respond *same identity* in the V-A condition, but not in the A-V condition.

8.3 Experiment 5b: Face-voice matching using a sequential same-different task: 10 second inter-stimulus interval

Experiment 5b investigated face-voice matching performance with a longer inter-stimulus interval. When there is a 10 second inter-stimulus interval, the first stimulus should be well beyond the range of echoic and iconic memory by the time the second stimulus is presented (Coltheart, 1980; Glanzer & Cunitz, 1966; Lu et al., 1992; Neisser, 1967; Sligte et al., 2008, 2009; Sperling, 1960; Treisman, 1964; Wickelgren, 1969). Guided by our interpretation of the results of Experiment 5a, we expected overall accuracy to have deteriorated to chance level, and for there to be no bias to accept a face and voice as belonging to the same person.

8.3.1 Methods

Apart from the following exceptions, the methods were identical to Experiment 5a.

8.3.1.1 Participants

There were 48 participants (43 females and 5 males), with an age range of 18 to 54 years (*M* = 23.90, *SD* = 8.52).

8.3.1.2 Procedure

The inter-stimulus interval was 10 seconds.

8.3.2 Results

8.3.2.1 Matching accuracy

The matching accuracy data were analysed using identical methods to Experiment 5a. Table 8.2 shows the profile likelihood chi-square statistic (G^2) and *p* value associated with dropping each effect from the appropriate model, as well as the coefficients and standard errors (on a log odds scale) for each effect in the full three-way interaction model. In the three-way model the estimate of *SD* of the face random effect was 0.599 while for voice stimulus it was 0.526. The estimated *SD* for the participant effect was 0.176. A similar pattern held for the null model.

Table 8.2

Parameter estimates (b) and profile likelihood tests for the 2x2x2 factorial analysis, Experiment 5b: Face-voice matching using a sequential same-different task: 10 second inter-stimulus interval

Source	*df*	*b*	*SE*	G^2	*p*
Intercept	1	0.485	0.287	.	.
Identity	1	1.132	0.337	7.81	.005
Order	1	0.505	0.332	1.52	.217
Facial stimulus type	1	0.463	0.312	6.23	.013
Identity x Order	1	1.013	0.474	4.71	.030
Identity x Facial stimulus type	1	0.511	0.437	0.85	.357
Order x Facial stimulus type	1	0.208	0.436	1.95	.162
Identity x Order x Facial stimulus type	1	0.454	0.619	0.54	.464

There was a main effect of identity (*p* = .005), and facial stimulus type (*p* = .013). There was also a significant interaction between identity and order (*p* = .030). The cell means and 95% confidence intervals for matching accuracy are shown in Figure 8.3.

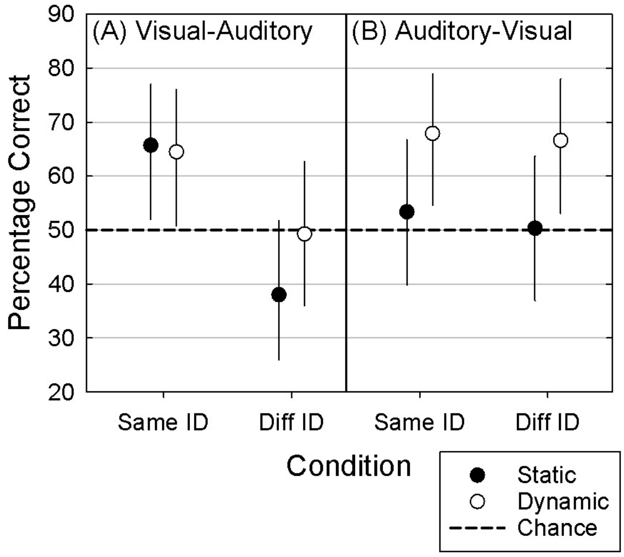


Figure 8.3: *Face-voice matching accuracy on V-A (panel A) and A-V (panel B) trials with a 10-second inter-stimulus interval. Error bars show 95% CI for the condition means*

Overall matching accuracy was at chance level, *M* = 57.58%, 95% CI [48.17, 66.42]. This was also the case for static facial stimulus trials, *M* = 52.54%, 95% CI [42.46, 62.29], but not dynamic, *M* = 62.72%, 95% CI [52.85, 71.63]. Performance was above chance level when voices were presented before faces (A-V), *M* = 60.99%, 95% CI [51.04, 70.18], but not when faces were presented before voices (V-A), *M* = 54.30%, 95% CI [44.25, 64.02]. Whilst the hit rate (same identity trials) was above chance level, *M* = 63.49%, 95% CI [53.48, 72.53], the true negative rate (different identity trials) was not, *M* = 50.71%, 95% CI [40.23, 61.16]. As illustrated in Figure 8.3, the main effect of identity reveals that the hit rate was higher than the true negative rate. The basis of the main effect of facial stimulus type is that dynamic face-voice matching performance is more accurate than static face-voice matching. According to visual inspection of Figure 8.3, the interaction between identity and order shows that the true negative rate was higher in the A-V condition (panel B) than in the V-A condition (panel A).

8.3.2.2 Matching response

Overall, faces and voices were not attributed to the same identity significantly above chance level, *M* = 56.24%, 95% CI [45.75, 66.14]. *Same identity* responses were not made above chance level on either V-A, *M* = 60.15%, 95% CI [49.18, 70.49], or A-V trials, *M* = 52.28%, 95% CI [41.03, 63.24].

8.3.3 Discussion

When the inter-stimulus interval was extended to 10 seconds, overall face-voice matching accuracy was at chance level. Taken together with the results from Experiments 4a (1 second inter-stimulus interval), and Experiment 5a (5 second inter-stimulus interval), this supports the hypothesis that accurate matching is not possible when the inter-stimulus interval is extended beyond a certain duration.

Unexpectedly, overall matching accuracy was above chance level in the A-V condition. Performance in this condition was not above chance level in Experiment 5a. As there is no theoretical explanation for this, it seems prudent not to overt-interpret the result at this stage. From Figure 8.3, it is clear that only performance on same identity, dynamic, A-V face-voice matching is accounting for the overall above-chance result in this condition. In Experiment 5a, performance was only above chance on dynamic facial stimulus trials, yet there was no difference between static and dynamic face-voice matching. In Experiment 5b there was an advantage afforded by dynamic over static facial stimulus trials reflected by the main effect of facial stimulus type.

As in Experiment 5a, there was a main effect of identity, and a significant interaction between identity and order. Figure 8.3 indicates that the basis of this interaction is a lower true negative rate in the V-A condition. However, as shown by the matching response analysis, when there is a 10 second inter-stimulus interval, this interaction does not translate into a significant bias to respond that a face and voice belong to the same person in the V-A condition. Consistent with predictions based on the results of Experiment 5a, overall, participants did not exhibit a bias to respond *same identity*. Therefore this experiment, along with the results of Experiment 5a, provides evidence that the bias weakens when faces and voices are temporally separated to a greater extent.

8.4 Experiment 5c: Face-voice matching using a sequential same-different task: Reorienting attention in the 5 second inter-stimulus interval

In Experiments 5a and 5b there was a trend towards less accurate performance and weakening of the bias to respond *same identity* as the inter-stimulus interval was extended to 5 and 10 seconds. This interpretation is based on the observation that overall performance and the bias to respond *same identity* in the A-V condition were above chance level when the interstimulus interval was 5s, but at chance level when it was 10s. It seems reasonable to argue that accurate face-voice matching performance therefore depends on being able to match high quality perceptual representations of faces and voices which are temporarily stored in echoic and iconic memory. However, an alternative explanation for the results is that participants were simply not paying attention at the onset of the second stimulus, making them both less accurate, and less likely to assume that faces and voices belong to the same identity. In order to test whether attention lapses account for the results, we adapted the procedure to maximise the chances that participants were attending to the task when the second stimulus was presented. Experiment 5c employed a 5-second inter-stimulus interval (the same duration as the interval in Experiment 5a).

8.4.1 Methods

Apart from the following exceptions, the methods were identical to Experiment 5a.

8.4.1.1 Participants

There were 48 participants (36 females and 12 males), with an age range of 18 to 46 years (*M* = 21.46, *SD* = 5.39).

8.4.1.2 Procedure

To increase the likelihood that participants focused their attention to the matching task and were not distracted at the onset of the second stimulus, a central cross-hair ('+') was visible on the screen for the duration of the inter-stimulus interval. It disappeared when the second stimulus was presented. One second before the onset of the second stimulus a short beep (250ms) played. Participants were informed that the beep signalled the impending presentation of the second stimulus.

8.4.2 Results

8.4.2.1 Matching accuracy

The data were analysed using identical methods to Experiment 5a. Table 8.3 shows the profile likelihood chi-square statistic (G^2) and *p* value associated with dropping each effect from the appropriate model, as well as the coefficients and standard errors (on a log odds scale) for each effect in the full three-way interaction model. In the three-way model the estimate of *SD* of the face random effect was 0.564 while for voice it was 0.552. The estimated *SD* for the participant effect was 0.268. A similar pattern held for the null model.

Table 8.3

Parameter estimates (b) and profile likelihood tests for the 2x2x2 factorial analysis, Experiment 5c: Face-voice matching using a sequential same-different task: Reorienting attention in the 5 second inter-stimulus interval

Source	*df*	*b*	*SE*	G^2	*p*
Intercept	1	0.515	0.307	.	.
Identity	1	1.410	0.347	17.12	<.001
Order	1	0.862	0.341	0.24	.623
Facial stimulus type	1	0.486	0.318	0.03	.867
Identity x Order	1	1.085	0.491	5.09	.024
Identity x Facial stimulus type	1	0.407	0.454	0.29	.589
Order x Facial stimulus type	1	0.725	0.435	2.53	.112
Identity x Order x Facial stimulus type	1	0.449	0.627	0.49	.483

The main effect of identity was significant ($p < .001$), along with the 2-way interaction between identity and order ($p = .024$). Figure 8.4, which shows the cell means and 95% confidence intervals for matching accuracy, aids interpretation of the main effect and interaction.

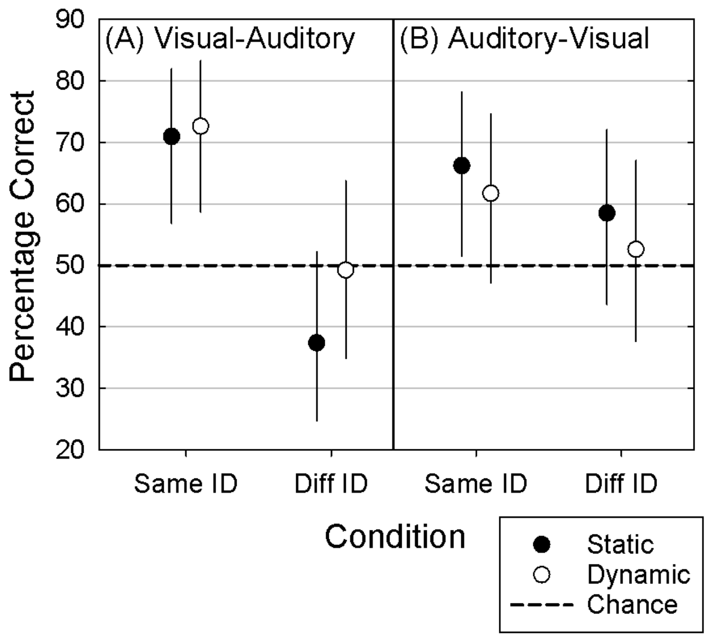


Figure 8.4: *Face-voice matching accuracy on V-A (panel A) and A-V (panel B) trials when attention is reoriented during the 5-second inter-stimulus interval. Error bars show 95% CI for the condition means*

Overall matching accuracy was at chance level, *M* = 59.80%, 95% CI [48.83, 69.85]. This was the case for both static, *M* = 59.44%, 95% CI [47.90, 70.28], and dynamic face-voice matching, *M* = 60.18%, 95% CI [48.48, 70.85], as well as both order conditions: A-V, *M* = 57.59%, 95% CI [45.70, 68.55], and V-A, *M* = 61.86%, 95% CI [49.98, 72.21]. Whilst the hit rate was above chance level, *M* = 68.29%, 95% CI [58.01, 77.09], the true negative rate was not, *M* = 49.54%, 95% CI [38.41, 60.48]. According to visual inspection, the basis of the 2-way interaction is that the true negative rate is lower in the V-A condition (shown in panel A) compared to the A-V condition (shown in panel B).

8.4.2.2 Matching response

Overall, faces and voices in each trial were not positively matched for identity significantly above chance level, *M* = 59.18%, 95% CI [48.93, 68.61]. Faces and voices were positively matched above chance level in the V-A condition, *M* = 62.92%, 95% CI [51.87, 72.87], but not in the A-V condition, *M* = 55.25%, 95% CI [44.02, 65.96].

8.4.3 Discussion

The matching accuracy results are inconsistent with the interpretation that participant inattention explains the results of Experiment 5a. With the inclusion of a central cross-hair, designed to maintain the participants' attention, and a short beep to reorient possible lapsed attention, overall matching accuracy was at chance level. In fact, as overall performance was above chance level in Experiment 5a, it seems that the introduction of the fixation and beep could have had the opposite effect from that which was intended. They may have distracted the participants by orienting attention away from the face (V-A condition) or voice (A-V condition) they were attempting to hold in memory, thereby making comparison with the second stimulus more difficult. Performance on dynamic as well as static facial stimulus trials was at chance level. Overall accuracy was above chance level without these procedural additions in Experiment 5a, so it appears that the cross-hair and beep may in fact have disrupted accurate performance.

Even with the addition of the fixation and beep, the matching response analysis indicated that the strength of the bias to respond *same identity* declines over a 5 second inter-stimulus interval. The overall pattern of the matching responses is identical to Experiment 5a, adding to evidence of a stronger bias to respond *same identity* in the V-A condition than in the A-V condition. Therefore, it would seem that identity cues associated with faces encompass subsequent voices, even if the voice is presented after a short (5s) inter-stimulus interval containing distracting visual and auditory events.

8.5 Experiment 5d: Face-voice matching using a sequential same-different task: Reorienting attention in the 10 second inter-stimulus interval

Experiment 5d provided a further test of the possibility that the fixation and beep prior to the presentation of the second stimulus have a disruptive influence on matching performance. Experiment 5d included a cross-hair and beep in the 10 second inter-stimulus interval. Based on the results of Experiment 5c, we did not expect overall performance to be above chance level.

8.5.1 Methods

Apart from the following exceptions, the methods were identical to those used in Experiment 5c.

8.5.1.1 Participants

There were 8 male and 38 female participants (*N*=46), with an age range of 18 to 29 years, *M* = 19.96, *SD* = 2.26.

8.5.1.2 Procedure

The duration of the inter-stimulus interval was 10 seconds.

8.5.2 Results

8.5.2.1 Matching accuracy

Table 8.4 shows the profile likelihood chi-square statistic (G^2) and *p* value associated with dropping each effect from the appropriate model, as well as the coefficients and standard errors (on a log odds scale) for each effect in the full three-way interaction model. In the three-way model the estimate of *SD* of the face random effect was 0.538 while for voice stimulus it was 0.284. The estimated *SD* for the participant effect was less than 0.001. A similar pattern held for the null model.

Table 8.4

Parameter estimates (b) and profile likelihood tests for the 2x2x2 factorial analysis, Experiment 5d: Face-voice matching using a sequential same-different task: Reorienting attention in the 10 second inter-stimulus interval

Source	*df*	*b*	*SE*	G^2	*p*
Intercept	1	0.453	0.270	.	.
Identity	1	1.077	0.330	12.68	<.001
Order	1	0.630	0.327	2.90	.089
Facial stimulus type	1	0.238	0.307	1.01	.316
Identity x Order	1	0.845	0.470	2.57	.109
Identity x Facial stimulus type	1	0.334	0.435	0.05	.826
Order x Facial stimulus type	1	0.094	0.435	0.31	.581
Identity x Order x Facial stimulus type	1	0.538	0.619	0.74	.390

The main effect of identity was significant (*p* < .001). There were no other main effects and no interactions (*p* > .089). Figure 8.5 shows the cell means and 95% confidence intervals for matching accuracy.

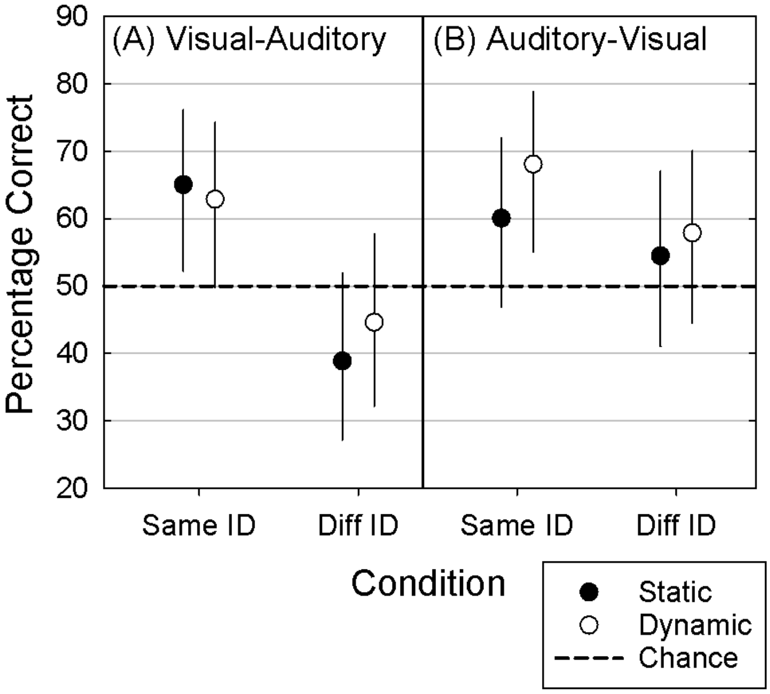


Figure 8.5: *Face-voice matching accuracy on V-A (panel A) and A-V (panel B) trials when attention is reoriented during the 10-second inter-stimulus interval. Error bars show 95% CI for the condition means*

Overall accuracy was at chance level, *M* = 56.98%, 95% CI [47.37, 65.97]. Both static, *M* = 55.12%, 95% CI [45.03, 64.87] and dynamic, *M* = 58.86%, 95% CI [48.77, 68.14] matching accuracy were at chance level overall, as was performance on the V-A order condition, *M* = 52.93%, 95% CI [43.42, 62.23]. In the A-V order condition, performance was above chance level, *M* = 60.92%, 95% CI [51.29, 69.70]. The hit rate (same identity trials) was above chance level, *M* = 64.28%, 95% CI [54.87, 72.61], but the true negative rate (different identity trials) was not, *M* = 48.59%, 95% CI [38.84, 58.41].

8.5.2.2 Matching response

Overall, faces and voices in each trial were not positively matched for identity significantly above chance level, *M* = 56.73%, 95% CI [46.53, 66.23]. This was the case in both the V-A condition, *M* = 60.18%, 95% CI [49.03, 70.23], and the A-V condition, *M* = 53.23%, 95% CI [42.23, 64.01].

8.5.3 Discussion

Overall matching accuracy was at chance level. Consistent with the results of Experiment 5c, when a cross-hair and short beep were included in the inter-stimulus interval, dynamic face-voice matching performance was no more accurate than static face-voice matching. These results are consistent with the fixation and beep actually disrupting performance. In Experiment 5b, which included a 10 second inter-stimulus interval but no fixation or beep, dynamic face-voice matching was above chance level. In Experiment 5d, performance on the trials featuring static facial stimuli, as well as those featuring dynamic facial stimuli, was at chance level. However, overall performance on A-V trials was above chance level. As in Experiment 5b, same identity dynamic accuracy accounts for this result; the other 3 conditions were at chance level (see panel B, Figure 8.5). The apparent performance advantage afforded in this condition may be explained by the fact that voices and dynamic faces share both speech and identity information, whereas voices and static faces share only identity information (Lachs & Pisoni, 2004a; Lander et al., 2007). As speech information is necessarily time-varying, it makes sense that identifying crossmodal redundancy across short intervals is possible. It also makes sense that performance is only above chance in the A-V condition because voices provide more reliable speech information than faces (Stevenage & Neil, 2014). Overlapping speech information may therefore provide a fallback in situations when overlapping identity information is less easy to access, such as when the face and voice are temporally separated.

The overall pattern of matching responses in Experiment 5d is the same as Experiment 5b. With a 10-second inter-stimulus interval, there was no overall bias to respond *same identity*, and the bias in both the A-V and V-A condition was at chance level.

8.6 General Discussion

The results presented in this chapter indicate how face-voice matching performance varies according to the time course of stimulus presentation. Taken together, all four experiments reveal that overall performance accuracy deteriorates when the inter-stimulus interval is extended beyond a few seconds. Although there is some evidence that dynamic face-voice matching ability fares better than static face-voice matching, the results of Experiment 5c and 5d show just how easily performance can be disrupted. Whilst participants exhibit a bias to respond *same identity* when faces and voices are presented close together in time (Experiments 4a and 4b), the results of Experiments 5a, 5b, 5c and 5d clearly show that the bias to respond *same identity* depends on the degree of temporal contiguity.

8.6.1 Matching accuracy

The results of all four experiments show that accurate static face-voice matching is not possible when faces and voices are presented beyond a short time interval (1 second). Performance was at chance level when the inter-stimulus interval was 5 seconds long (Experiments 5a and 5c), and when it was 10 seconds long (Experiments 5b and 5d). These results are consistent with the interpretation that above-chance matching accuracy depends on being able to compare high-quality perceptual representations of faces and voices temporarily stored in echoic and iconic memory. The literature suggests that these representations are likely to have significantly decayed after 5 seconds (Coltheart, 1980; Glanzer & Cunitz, 1966; Lu et al., 1992; Neisser, 1967; Sligte et al., 2008, 2009; Sperling, 1960; Treisman, 1964; Wickelgren, 1969).

The results of Experiment 5a and 5b show that dynamic face-voice matching accuracy is above chance level, even when the interval is 10 seconds. Figure 8.3 illustrates that dynamic face-voice matching is most accurate when the voice is presented before the face (A-V condition). In this condition, the voice must be held in working memory for longer than the face, so the results cannot be explained by superior memory for dynamic over static faces (e.g. Knappmeyer et al., 2003; Lander & Chuang, 2005). In the A-V condition, dynamic faces might operate as a memory cue for voices. We have shown in previous experiments (Experiments 4a and 4b) that dynamic faces and voices do not share more diagnostic identity information than static faces and voices. However, an articulating face and a voice do have a wider range of information in common than static faces and voices, in that both are involved in speech production (Lachs & Pisoni, 2004a; Lander et al., 2007). Redundant information relating to articulatory patterns perhaps becomes particularly useful in matching tasks when it is difficult to access overlapping identity information. Such a situation might occur when the interstimulus interval is increased beyond a certain duration. Therefore, in a face-voice matching task with a longer inter-stimulus interval (>1 second), it is feasible that a dynamic articulating face re-activates, or maintains, the perceptual representation of the preceding voice more effectively than a static face does because articulatory movement provides a memory cue for speech information conveyed by the voice, thereby facilitating matching accuracy. It makes sense to explain A-V performance in Experiments 5b and 5d in the context of this interpretation, because above chance matching accuracy is accounted for by performance in the same identity dynamic facial stimulus condition. This explanation is based purely on the observation that this is the only condition in which performance is significantly above chance level. There was however no 3-way interaction to indicate that performance in this condition differed significantly from the others.

The overall pattern of declining matching accuracy as the interval increases is not explained by an attention lapse prior to the onset of the second stimulus. In fact, in Experiment 5c and 5d the beep and central cross-hair appear to disrupt accurate matching performance. One explanation for this is that the cross-hair and beep introduce interference, which undermines the quality of the perceptual representations temporarily residing in limited memory stores (Baddeley, 2007). As a result, the participants' ability to compare these representations, and to make accurate matching decisions, is likely to be impaired.

Although the overall results in Experiments 5a and 5c were descriptively similar, the pattern of variance in the multilevel modelling analyses are informative. Whilst in Experiment 5a the *SD* of the participant random effect was minimal, in Experiment 5c it was larger; the participants who saw a cross-hair and heard a beep in the 5 second interval responded less uniformly to the stimuli in each trial. The increased variance may be attributable to individual differences in memory. At 5 seconds, detailed representations, less resistant to disruption by a cross-hair or beep, may persist in some people's memory, but not in others' (Todd & Marois, 2005; Vogel & Machizawa, 2004). Alternatively, the level of disruption might be mediated by the extent to which the participants were paying attention to the task in the first place. Those who were paying close attention might have been able to hold a detailed representation in mind and therefore have been less distracted than those who were not paying close attention. In Experiment 5d, which also included a beep and fixation, the participant variance was minimal. This may be because the 10 second interval had pushed the first stimulus well out of the range of echoic and iconic memory. Therefore, detailed representations had likely decayed for all participants, regardless of the amount of attention they were paying to the task.

The results of Experiment 5a and 5b suggest that overlapping speech information shared by dynamic faces and voices might provide a fallback when overlapping identity information is more difficult to access. However, it seems that even dynamic information is relatively transient; the central cross-hair and beep introduced in Experiments 5c and 5d were particularly disruptive to dynamic face-voice matching accuracy. Despite above chance matching performance in the dynamic facial stimulus condition in Experiment 5a and 5b, performance was at chance level when the beep and cross-hair were included in the interval. In a social setting, faces and voices are not usually encountered in silent situations, so interference is more likely when the interval is longer. The results of Experiments 5c and 5d are therefore consistent with the conclusion that it is easier to accurately attribute common source identity to faces and voices when the two stimuli are presented in close temporal proximity.

The matching accuracy results should be considered in terms of social functioning. During social interactions involving a number of individuals, faces and voices belonging to the same people are usually encountered at the same time. It makes sense that the ability to attribute common identity only occurs when faces and voices are presented within a short time frame. Being able to accurately link faces and voices that are temporally offset to a greater extent would incur an unnecessary cost in terms of cognitive load.

8.6.2 Matching response

The bias to respond *same identity* appears to depend on faces and voices being presented close together in time. Although an overall bias was observed in Experiments 4a and 4b when the inter-stimulus interval was 1 second, it does not persist when 5 or 10-second intervals separate faces and voices (Experiments 5a, 5b, 5c and 5d). This sits well with predictions informed by temporal contiguity research (Buehner & May, 2003; Ginns, 2006; Reed, 1992; Shanks et al., 1989).

Taken together with the results of Experiment 4a, the results in this chapter add to evidence of a stronger response bias in the V-A condition than in the A-V condition. In Experiment 5a and 5c, which both featured a 5 second inter-stimulus interval, specificity was higher when the voice was presented before the face (A-V condition). Consistent with this, matching response analyses show that whilst the overall response bias to accept faces and voices in each trial as belonging to the same identity does not persist overall at 5 seconds (Experiment 5a and 5c) in the A-V condition, it does persist in the V-A condition. However, participants were not more likely to respond *same identity* in either condition when there was a 10 second inter-stimulus interval (Experiments 5b and 5d).

Beyond a short time-frame, the overall lack of a bias to respond *same identity* is unsurprising. In speech perception, audiovisual integration only occurs when articulating faces and voices are presented close together in time (Munhall et al., 1996; Robertson & Schweinberger, 2010; Van Wassenhove et al., 2007). The results of these four experiments therefore fit with the results of Experiment 4a and 4b, offering additional support for the argument that the bias to attribute common identity to faces and voices provides a useful foundation for successful audiovisual speech integration, thereby helping to facilitate social communication.

8.6.3 Conclusion

Taken together, this set of four experiments show that performance deteriorates as the interval between a face and voice increases, and that accuracy is disrupted by intervening interference. Furthermore, when offset by between 5 and 10 seconds, people no longer exhibit a bias to attribute common identity to faces and voices. Face-voice matching performance is clearly dependent on the time-course of stimulus presentation.

9. CHAPTER 9: SUMMARY AND GENERAL DISCUSSION

9.1 Introduction

This Chapter summarises and discusses the findings of the 12 experiments presented in previous chapters. It suggests some future directions for face-voice matching research and considers the applied relevance of the findings.

9.2 Summary and main conclusions

This thesis investigated whether people look and sound similar, using face-voice matching as a measure of whether faces and voices index redundant identity information. The overall picture of face-voice matching ability offered by previous studies is contradictory and incomplete. This thesis has attempted to resolve contradictions, as well as extending the existing literature. The following section briefly summarises the main findings and conclusions.

In support of the hypothesis that it should be possible to match voices to static faces, Experiment 1 showed that both static faces and voices, as well as dynamic faces and voices, offer strikingly concordant information about a number of dimensions. The relationship between face and voice ratings of masculinity, femininity, height and health were particularly strong.

It was not clear from the previous literature whether accurate face-voice matching relies on the ability to encode visual articulatory movement present in dynamic faces. Experiment 2a, 2b and 2c tested face-voice matching across different 2AFC procedures in order to compare performance using static and dynamic facial stimuli. The results showed that dynamic face-voice matching was consistently above chance level. Static face-voice matching is also possible, but it is sensitive to the exact experimental procedure employed.

Performance was only above chance level when the test options were presented simultaneously (Experiment 2c), likely because this procedure facilitates direct comparisons between alternatives. In highlighting that some people look and sound more similar than others, the multilevel modelling analyses of Experiments 2a, 2b and 2c offered an additional explanation for previous contradictions. The results of face-voice matching studies are likely to depend on the exact stimuli used.

In Experiments 2a and 2b, test alternatives in 2AFC tasks were presented sequentially. There was a temporal position bias, whereby matching accuracy was higher when the same identity alternative appeared first. Experiments 3a and 3b tested matching performance when the same identity stimulus was absent at test. Descriptively speaking, the participants were consistently more likely to select the first of the two alternatives. Together, the findings presented in these 5 experiments cast doubt on the suitability of using 2AFC procedures to investigate face-voice matching, and highlight the need to investigate the role of bias in more depth using alternative methodologies.

Experiments 4a and 4b adopted a same-different procedure, an arguably more appropriate method of testing face-voice matching, which facilitates the investigation of response biases. The results offered corroborating evidence for accurate static face-voice matching, as well as showing that participants exhibit a bias to respond that faces and voices belong to the same person. This bias was strongest when the face was presented before the voice.

The remaining experiments addressed the effect of increasing the inter-stimulus interval on face-voice matching performance. Experiments 5a and 5b used a same-different procedure, inserting a 5 second (Experiment 5a) and 10 second (Experiment 5b) interval between the stimuli in order to push the first stimulus out of the range, or at least to the very limits, of sensory memory. Accurate face-voice matching appears to depend on being able to compare temporarily stored, high-quality, representations of faces and voices. The results were not due to attention lapses occurring during the inter-stimulus interval (Experiments 5c and 5d). Matching response analyses showed that beyond 5-10 seconds, there is no bias to attribute common identity to faces and voices (Experiments 5b and 5d).

In considering the results as a whole, three main conclusions can be drawn. Each conclusion constitutes an original contribution to the literature:

- Faces and voices offer common source identity information. Accordingly, when presented close together in time, novel faces and voices can be accurately matched for identity above chance level.
- Above-chance matching is not contingent on encoding information about visual articulatory movement. There is no difference in matching accuracy when comparing trials using static and dynamic facial stimuli. Contradictions across previous literature can be explained by methodological differences.
- People exhibit a bias to attribute common identity to faces and voices when they are temporally proximal.

9.3 Research questions

Five research questions were outlined in the Literature Review (see section 2.8.1). The following section deals with each of these in turn, drawing together evidence from different chapters to help facilitate a detailed consideration of the overall results. It expands on the main conclusions referred to above in order to explain how the findings extend existing knowledge.

9.3.1 Research question 1: Do voices share redundant information with dynamic as well as static faces?

Previous research investigating multimodal signals in faces and voices has concentrated almost exclusively on attractiveness (e.g. Abend et al., 2015; Thornhill & Gangestad, 1999; Thornhill & Grammer 1999; Feinberg et al., 2005; Feinberg, 2008; Oguchi & Kikuchi, 1997; Wells et al., 2013; Wheatley et al., 2014; Zahavi & Zahavi, 1997). Studies have found that a face and voice belonging to the same person tend to be rated similarly on scales for attractiveness (Collins & Missing, 2003; Saxton et al., 2006). A minority of studies have ventured beyond this research question, considering whether faces and voices offer accurate information about dimensions such as body size. It has previously been shown that there are significant differences between estimates made from faces and voices (Lass & Colt, 1980), and that participants are slightly more accurate when rating body size from photographs than voice recordings (Krauss et al., 2002). More recently, Rezlescu et al. (2015) addressed the relationship between trait ratings, finding that the contribution of facial and vocal information to final judgments varies according to the trait being communicated. Overall though, existing knowledge of whether multimodal signals in humans constitute back-up signals (Johnstone, 1997) or multiple-messages (Møller & Pomiankowski, 1993) is limited.

Experiment 1 tested whether faces and voices constitute back-up signals (Johnstone, 1997) for a number of dimensions, and whether the extent of concordance varies by facial stimulus type. As well as extending existing knowledge regarding the nature of multimodal signals in humans, this experiment aimed to provide evidence on which to build hypotheses regarding face-voice matching accuracy using both static and dynamic facial images. Experiment 1 showed for the first time, that faces and voices offer concordant information about a number of dimensions relevant to fitness and quality, particularly in terms of femininity, masculinity, height and health. Furthermore, face-voice concordance was not affected by whether the faces were static or dynamic. In showing that recipients respond so similarly to the visual and auditory aspects of these multimodal signals (Partan & Marler, 2005), our results suggest that faces and voices constitute back-up signals (Johnstone, 1997) rather than multiple messages (Møller & Pomiankowski, 1993).

At this stage, it is important to comment on the stimulus set used throughout this thesis. As explained in Chapter 3, faces and voices in the GRID audiovisual sentence corpus (Cooke et al., 2006) were emotionally neutral throughout the duration of each 2 second video. The faces were only ever visible from one angle, and the lighting did not change over the course of the video. The voices said nonsense sentences in a monotone fashion. The only noteworthy thing that differed between static pictures and dynamic videos was that the mouth was moving. This may have limited the extent to which dynamism associated with visual articulation offered additional information on any of the dimensions tested in Experiment 1. However, in terms of stimulus testing, the results of Experiment 1 are crucial to the subsequent experiments. As one of the main aims was to establish whether both static and dynamic face-voice matching is possible (see section 9.3.2) it was necessary to first establish whether people draw similar inferences from both sets of stimulus faces. These results were therefore important in helping to build hypotheses for later chapters, which used the same set of stimuli. Importantly, the broad characteristics of the stimuli described above are also true of previous face-voice matching experiments (Kamachi et al., 2003; Lachs & Pisoni, 2004a; Lander et al., 2007; Mavica & Barenholtz, 2013), which is helpful in facilitating comparisons across studies.

The findings of Experiment 1 are consistent with the hypotheses derived from Belin et al.’s (2004) auditory face model, which predicts that voice and face perception occur in integrated and parallel pathways dedicated to processing speech, emotion, and identity information. Based on this model, it seems likely that whilst redundant speech information is available in articulating (dynamic) faces and voices, viewing a static face should be sufficient to extract identity information shared with that person's voice. The results from Experiment 1 support the hypothesis that static face-voice matching is possible.

9.3.2 Research question 2: Is it possible to match voices and static faces, or is accurate face-voice matching contingent on encoding information about visual articulatory patterns?

Based on experiments observing chance level static face-voice matching performance, audiovisual speech perception researchers have concluded that encoding auditory and visual information about idiosyncratic speaking style is crucial to accurate face-voice matching (Kamachi et al., 2003; Lachs & Pisoni, 2004a; Lander et al., 2007). This conclusion is challenged by other studies showing that static face-voice matching is possible (Krauss et al., 2002; Mavica & Barenholtz, 2013). Notably however, no previous studies have directly compared face-voice matching using static facial stimuli to matching using dynamic facial stimuli. Regardless of whether static face-voice matching is above chance level, if there is no significant difference between the two, this would undermine the conclusion that face-voice matching depends entirely on the availability of information about articulatory patterns.

Chapters 5, 7 and 8 addressed this gap in the literature, comparing static and dynamic face-voice matching accuracy using different experimental procedures, including sequential and simultaneous 2AFC tasks (Experiments 2a, 2b and 2c), as well as same-different tasks (Experiments 4a, 4b, 5a, 5b, 5c and 5d). The stimulus set featured in each experiment (Cooke et al., 2006, see Chapter 3) is particularly appropriate to establishing whether visual articulatory movement is crucial to accurate face-voice matching. The emotionally neutral sentences coupled with the uniform head position make it easier to isolate, as far as possible, whether visual articulatory movement explains accurate face-voice matching.

There was no significant effect of facial stimulus type (static or dynamic) in any of the experiments employing a 1 second inter-stimulus interval (Experiment 2a, 2b, 4a, 4b). Purely on the basis of this null effect, it seems logical to conclude that the additional information provided by visual articulatory movement fails to explain face-voice matching ability, thereby undermining the arguments of previous audiovisual speech perception studies (e.g. Kamachi et al., 2003; Lachs & Pisoni, 2004a; Lander et al., 2007). This interpretation is supported by the complementary finding that static face-voice matching is above-chance level using certain experimental procedures (Experiments 2c, 4a and 4b). However, the results presented in Chapter 8 provide an important qualification. Whilst accurate dynamic face-voice matching is possible over longer inter-stimulus intervals (5-10 seconds), accurate static face-voice matching is not (Experiments 5a and 5b). Whilst this finding does not undermine the conclusion that static face-voice matching is possible, the significant effect of facial stimulus type when the inter-stimulus interval was 10 seconds (Experiment 5b) shows that access to common source identity information in static faces and voices is relatively transient. Static faces and voices share identity information, whereas dynamic faces and voices share both identity information and speech information. As speech unfolds over time and is commonly punctuated with pauses, it is possible that common source identity available in dynamic faces is relatively more tolerant to the temporal separation of faces and voices. This issue is addressed in further detail in section 9.3.3.

The ability to accurately match voices and faces for identity (regardless of whether the face is static or dynamic) is likely to have an important function. Above chance level matching may help people to navigate complex social interactions, which frequently feature a number of novel speakers. It is common to hear a voice whilst not looking in the direction of the speaker. Being able to accept or reject a face match quickly may aid social communication by facilitating attention shifts. As faces and voices are both important in speech perception (Benoit et al., 1994; MacLeod & Summerfield, 1987; Rosenblum, 2005; Summerfield, 1987; Sumby & Pollack, 1954), underlying awareness of redundant identity information might also facilitate coherence. This may go some way to explaining the apparent dynamic (articulating) facial stimulus advantage in some conditions.

9.3.3 Research question 3: Do procedural differences account for inconsistencies in the previous literature regarding static face-voice matching?

In Experiment 1, it was shown that static faces and voices, as well as dynamic faces and voices, offer strikingly concordant information about a number of dimensions. However, whilst some studies have observed above-chance static face-voice matching (Krauss et al., 2002; Mavica & Barenholtz, 2013), others have only observed accurate dynamic face-voice matching (Kamachi et al., 2003; Lachs & Pisoni, 2004a; Lander et al., 2007). Previous face-voice matching studies have used a variety of different procedural versions of the 2AFC task. It was hypothesised that these differences could help to account for the incompatible results across studies.

Chapter 4 reported tests of face-voice matching using different versions of 2AFC tasks in order to establish which procedural elements support accurate static face-voice matching. In this chapter, three different procedures were compared: sequential face-voice presentation (Experiment 2a), simultaneous face-voice presentation (Experiment 2b), and simultaneously presented alternatives (Experiment 2c). Undertaking this comparison facilitated the isolation of specific procedural characteristics. The findings offered two explanations for previous inconsistencies. Static face-voice matching was only above chance level in Experiment 2c, highlighting that performance is sensitive to the type of experimental procedure employed. Matching accuracy is more likely to be above chance level when the procedure enables participants to compare simultaneously presented alternatives at test.

However, it was clear from the existing literature that an explanation based on procedural differences constituted only part of the story. Mavica and Barenholtz (2013, Experiment 2) used the same procedure as in Experiment 2a, but observed above chance level static face-voice matching. The multilevel modelling analyses offered an additional explanation for apparently inconsistent static face-voice matching performance. In all three experiments (2a, 2b and 2c), there was a high level of variability associated with the face and voice stimuli, far greater than the variability at the participant level. A similar pattern occurred in the experiments reported in other chapters (Experiments 3a, 3b, 4a, 4b, 5a, 5b and 5d). As some people evidently look and sound more similar than others, the specific stimuli used in face-voice matching studies are likely to affect the overall results. In line with the literature reviewed in Chapter 3, this finding supports calls for the use of appropriate statistical techniques that simultaneously account for sources of variability associated with stimuli and participants, as well as emphasising the importance of using sufficiently large samples of stimuli (see section 9.6 for further discussion of these issues).

It is necessary to consider in more detail why sequentially presenting test alternatives in a 2AFC task might compromise static more than dynamic face-voice matching accuracy. Chapter 8 presented four experiments investigating the effect of increasing the inter-stimulus interval on matching performance. Considering these results alongside the results of Experiments 2a, 2b and 2c may help to explain the apparent matching accuracy advantage using dynamic facial stimuli. As highlighted above, dynamic face-voice matching accuracy was consistently above chance level in all experiments when the inter-stimulus interval was 1 second (Experiments 2a, 2b, 2c, 4a and 4b), regardless of whether the alternatives in the 2AFC task were presented sequentially (Experiments 2a and 2b) or simultaneously

(Experiment 2c). It also remained above chance level when the inter-stimulus interval was extended to 5 seconds (Experiment 5a) and 10 seconds (Experiment 5b). Static face-voice matching was at chance level in all experiments when alternatives in the 2AFC task were presented sequentially (Experiments 2a and 2b) as well as those featuring longer (>1 second) inter-stimulus intervals (Experiments 5a, 5b, 5c and 5d).

Experiments 2a and 2b have a notable feature in common with Experiments 5a, 5b, 5c and 5d. The test element of the trials featured in these experiments is longer in duration than the experiments in which alternatives are presented simultaneously (Experiment 2c), or the same-different tasks with a maximum 1second inter-stimulus interval (Experiment 4a and 4b). Therefore, when facial stimuli are dynamic, matching accuracy appears to be more robust to procedural differences that temporally extend the test element of the face-voice matching task. A parsimonious explanation is that when the cognitive load is higher, the additional time-varying speech information contained in dynamic faces better supports matching decisions. Establishing why this might be the case is more challenging. It may be related to articulatory movement providing participants with additional, and perhaps more memorable information. This might create an extra layer of facial-vocal overlap that can be capitalised on when matching decisions are more difficult.

The data do not allow a distinction to be made between the exact differences in information shared by static faces and voices compared to dynamic faces and voices, although there is evidently sufficient commonality in static faces and voices to support accurate matching (see section 9.3.2). It would appear from previous literature that the information driving matching decisions in each of the facial stimulus conditions is not identical. When characteristics of dynamic faces are isolated from the characteristics of static faces using point-light displays, accurate face-voice matching is still possible (Lachs & Pisoni, 2004b; Rosenblum et al., 2006). Owing to shared information about speaking style, voices and dynamic faces may act as memory cues for each other when the task is more difficult. Alternatively, the crossmodal representation (including, but not limited to, identity information (Belin et al., 2004)), which is created following exposure to faces and voices, might be less susceptible to disruption when the face is dynamic because it also includes bimodal speaking style information. If, as has been suggested here, accurate face-voice matching provides a foundation for audiovisual speech integration, accessing common source identity information might be particularly necessary during conversations, when faces are dynamic.

9.3.4 Research question 4: Are there matching performance asymmetries according to the order of stimulus presentation?

Face-voice matching studies concerned with audiovisual speech perception have compared accuracy in V-A conditions (a face followed by 2 voices) to accuracy in A-V conditions (a voice followed by 2 faces) using 2AFC standard crossmodal matching tasks (Kamachi et al., 2003; Lachs, 1999; Lachs & Pisoni, 2004a, 2004b; Lander et al., 2007). As speech perception primarily involves voices (Massaro & Simpson, 2014), Lachs and Pisoni (2004a) suggested that it might be easier to compare information from 2 voices (V-A condition) than from 2 faces (A-V condition) when making matching decisions. However, previous studies have not detected differences in terms of accuracy. As the manipulation of stimulus presentation order is explicitly motivated by hypotheses formulated on the basis of speech perception research (Lachs, 1999), it is unsurprising that this manipulation has not been adopted in matching studies concerned exclusively with static faces (Krauss et al., 2002; Mavica & Barenholtz, 2013). Nevertheless, the wider literature hints that the manipulation of order is important, because of differences between face and voice processing. Face-voice matching tasks require participants to make identity decisions, but faces are more reliable indicators of identity than voices (e.g. Damjanovic & Hanley 2007; Hanley & Turner 2000; Stevenage et al., 2011, 2012, 2013, 2014b; Stevenage & Neil, 2014).

Studies observing above chance level matching using static facial stimuli have used a variety of 2AFC procedures (Kamachi et al., 2003; Lachs & Pisoni, 2004a, 2004b; Lander et al., 2007; Mavica & Barenholtz, 2013). The results in Chapter 5 illustrate that face-voice matching is sensitive to the type of procedure employed (Experiments 2a, 2b, and 2c). This finding warns against assuming that the order results from crossmodal matching tasks (Kamachi et al., 2003; Lachs, 1999; Lachs & Pisoni, 2004a, 2004b; Lander et al., 2007) generalise to other types of tasks. Order effects have not been investigated using alternative matching procedures.

In each matching experiment in which faces and voices, or face-voice combinations were presented sequentially (Experiment 2a, 2b, 3a, 3b, 4a, 5a, 5b, 5c and 5d), a manipulation of order was included. Accuracy on a 2AFC task did not differ across order conditions, either when the faces were dynamic or static (Experiments 2a, 2b). This replicated the results of audiovisual speech perception studies (Kamachi et al., 2003; Lachs, 1999; Lachs & Pisoni, 2004a, 2004b; Lander et al., 2007). No previous face-voice matching studies have investigated order effects in same-different tasks. Employing this procedure in Chapter 7, the results showed no difference in terms of sensitivity between the A-V and V-A conditions (Experiment 4a). The clear lack of difference in terms of hit rates across all of these experiments indicates that despite faces offering more reliable identity information (Damjanovic & Hanley 2007; Hanley & Turner 2000; Stevenage et al., 2011, 2012, 2013, 2014b; Stevenage & Neil, 2014), being presented before the voice(s) does not increase sensitivity to identity matches. Based on the results from Experiment 1, this may be because people make such similar judgements about people from their faces and voices.

The results from Chapter 8 are interpreted as providing evidence that face-voice matching depends on making identity decisions based on comparing high quality visual and auditory representations. Following a 1 second interval, it is likely that both visual and auditory representations of faces and voices are still in a high capacity immediate memory store (Glanzer & Cunitz, 1966; Lu et al., 1992; Sligte et al., 2008; 2009; Treisman, 1964; Wickelgren, 1969). Therefore, in terms of perceptual quality, the order of stimulus presentation should not matter. This is what we found in Experiments 2a, 2b and 4a.

Observed modality effects in sensory memory, whereby auditory representations persist longer than visual representations (Crowder & Morton, 1969; Penney, 1985), predict that the order manipulation might have been more likely to affect performance in experiments including longer (>1 second) inter-stimulus intervals (Chapter 8: Experiment 5a, 5b, 5c and 5d). The voice in the A-V condition would perhaps have been of a higher perceptual quality than the face in the V-A condition, thereby boosting accuracy. However, hit rates did not differ across order conditions in any of the experiments in Chapter 8. The specific duration of the inter-stimulus intervals might explain why this was the case. By 5 seconds, it is likely that both visual and auditory representations have already passed, or are at least in the process of passing, to the short term memory stores (Glanzer & Cunitz, 1966; Lu et al., 1992; Sligte et al., 2008; 2009; Treisman, 1964; Wickelgren, 1969). In this case, modality effects in sensory memory would no longer be relevant. Future research might investigate the time-course of accurate face-voice matching in more detail, employing 2, 3 and 4 second inter-stimulus intervals to further test order effects.

In employing same-different procedures (Experiments 4a, 5a and 5b), we were able to test not only sensitivity, as with 2AFC tasks (Experiments 2a and 2b), but also whether order affected specificity and response bias. Discussion of this set of results is more appropriately addressed in the following section (9.3.5).

9.3.5 Research question 5: How do response biases operate in face-voice matching?

Despite all previous face-voice matching studies using variations of a 2AFC procedure (Kamachi et al., 2003; Krauss et al., 2002; Lachs & Pisoni, 2004a, 2004b; Lander et al., 2007; Mavica & Barenholtz, 2013), none have tested whether presenting the same identity stimulus in position 1 or 2 affects accuracy. Some literature suggests that 2AFC procedures might be inherently biased (Dyjas et al., 2012; Garcia Perez et al., 2010, 2011; Rammsayer & Ulrich, 2012; Ulrich & Vorberg, 2009; Yeshurun et al., 2008). When using this procedure to disentangle contradictions between previous studies, the data were analysed for position effects (Experiments 2a, 2b and 2c), and a clear temporal position bias was observed. When the two alternatives in a 2AFC task were presented sequentially, matching accuracy was higher if the correct (same identity) alternative appeared in position 1, compared to when it appeared in position 2 (Experiments 2a and 2b). When the two alternatives were presented simultaneously, there was no main effect of position, reflecting the absence of a spatial position bias (Experiment 2c).

The presence of a temporal position effect highlights the biased nature of 2AFC procedures for testing face-voice matching. Having identified the bias (Experiments 2a and 2b), it was necessary to account for it in order to clarify the decision processes informing performance. One possible explanation for the effect is that sensitivity differs according to position (Yeshurun et al., 2008). In order to rule this explanation out, it was necessary to test whether the bias still operated when the same identity stimulus was not present (Experiments 3a and 3b). This bias did not reach significance, but in all conditions *position 1* responses were numerically above 50%. The distribution of *position 1* and *position 2* responses was similar in Experiment 3a and 3b (target not present) to that in Experiment 2a and 2b (target present). Therefore, the results were more consistent with an explanation based on participants exhibiting decision bias under uncertainty than an explanation based on differing sensitivity across conditions (García-Pérez & Alcalá-Quintana, 2010).

The results of Experiments 3a and 3b did not clarify why the bias manifested in this particular pattern, with the alternative presented in position 1 consistently being favoured. It was clear from considering Experiments 2a and 2b together that it was not because of pressures on sensory memory. The results of Experiment 2a, in which all of the stimuli were presented sequentially, appear to be consistent with the interpretation that quickly degrading representations make it easier to compare the first stimulus to the alternative in position 1 than the alternative in position 2. However in Experiment 2b, faces-voice combinations were presented simultaneously. The position bias persisted even when the memory load was reduced. On the basis of these results, it seems more likely that the observed position bias operated because of a general tendency to attribute common identity to faces and voices, perhaps influenced by the fact that faces and voices belonging to the same person most commonly occur close together in time during social interactions. This question has not been addressed in the previous literature.

A same-different procedure was adopted in Experiments 4a, 4b, 5a, 5b, 5c and 5d. In line with expectations based on the results of the 2AFC Experiments (Experiments 2a, 2b,2c, 3a and 3b), there was evidence of an overall bias to respond that faces and voices share the same identity. *Same identity* responses were made above chance level in experiments with a 1 second inter-stimulus interval (Experiments 4a and 4b), undermining accuracy on different identity (noise) trials, but supporting accuracy on same identity (signal) trials. Accordingly, in Experiments 4a and 4b there was a main effect of identity. Participants were significantly more accurate on same identity trials. There was also a main effect of identity when the inter-stimulus interval was 5 seconds (Experiment 5a and 5c) and 10 seconds (Experiment 5b and 5d), but this did not equate to an overall bias to assume that novel faces and voices belong to the same identity. The results of Experiments 5a, 5b, 5c and 5d show that the bias operates according to temporal contiguity, which helps to explain the position bias pattern observed in Experiments 2a and 2b. The participants were more likely to accept the first combination presented to them in time (position 1), thereby ruling out the second alternative (position 2).

It is useful to consider the consistent main effect of identity (Experiment 4a, 4b, 5a, 5b, 5c and 5d) in the context of the person recognition literature. A pattern of responses reflecting asymmetric performance on noise and signal trials, typically with higher accuracy on signal than noise trials, is common in both unfamiliar face recognition (Bruce, Burton, & Dench, 1994; Hancock, Burton & Bruce, 1996; Lewis & Johnston, 1997; Vokey & Read, 1992), and unfamiliar face matching (Bruce et al., 1999; Megreya & Burton, 2006, 2007). Similarly, voice recognition studies frequently observe a particularly high rate of false positives on voice lineups (Kerstholt et al., 2004; Yarmey & Matthys, 1992). This is not consistent with the robust mirror effect observed in recognition studies using non-human stimuli such as high and low frequency words (see Glanzer & Adams, 1985, 1990; Glanzer, Adams, Iverson & Kim, 1993). The mirror effect refers to a situation when recognition performance on signal trials mirrors performance on noise trials. In this case the hit rate and true negative rate would be very similar. The reason for the unrelated nature of these two aspects of performance using human stimuli is not altogether clear (Megreya & Burton, 2007), but it is plausible that the explanation relates to social functioning. That is to say, the cost of an incorrect positive response to human stimuli may be greater than the cost of an incorrect negative response. Therefore, in the case of recognition it might be more important to be able to recognise someone you have previously encountered than to know that you have never seen them before. Perhaps in the case of face-voice matching there is some adaptive value in being able to identify congruence over incongruence. For example, identifying congruence would be crucial in helping to quickly direct attention from an unfamiliar person’s voice to their face.

Although the order of stimulus presentation does not affect sensitivity (see section 9.3.4), it does appear to affect the bias to respond *same identity*. In Experiment 4a, the accuracy analysis showed a significant interaction between identity and order. Specificity was better in the A-V condition than in the V-A condition. Although the matching response analyses revealed a significant bias in both order conditions, the interaction reflected a stronger bias in the V-A condition. The same pattern of results was observed in Experiments 5a and 5c, showing that as well as being stronger, the response bias in the V-A condition also withstands a longer inter-stimulus interval (5 seconds). After 5 seconds the bias in the A-V condition had disappeared (Experiments 5a and 5c). Section 9.3.4 explains that modality effects are unlikely to account for observed face-voice matching performance in the reported experiments. The order effect according to bias is more likely attributable to the strength of identity information associated with faces and voices (Damjanovic & Hanley 2007; Hanley & Turner 2000; Stevenage et al., 2011, 2012, 2013, 2014b; Stevenage & Neil, 2014). The bias is stronger when the face is presented before the voice. As faces provide more reliable cues to identity than voices, it is feasible that voices tend to be subsumed by the identity of preceding faces. During conversations it is possible to view a face continuously, but voices are only audible when the interlocutor is speaking. It makes sense to rely on the face to a greater extent as a cue to identity, automatically accepting a voice as belonging to the same person.

9.4 Considering face-voice matching performance within an overarching framework

Face-voice matching performance exhibits a number of characteristics that are evident when considering the 12 experiments in this thesis together as a whole. The following section attempts to account for these characteristics within a single framework.

Face and voice processing are integrated processes. Because of this, people can access crossmodal identity information present in both static faces and voices, and dynamic faces and voices. Although dynamic faces and voices also share information about idiosyncratic speaking style, this information is not more informative during matching tasks than the identity information shared with static faces. If it were we could expect a significant difference between static and dynamic conditions when faces and voices were separated by 1 second. No such effect was observed (Experiments 2a, 2b, 4a, 4b). However, because speech information is time-varying, this information may be more tolerant to temporal offsets and pauses; the information can be accessed as a fallback when redundant identity information is less easy to access. This likely explains why dynamic face-voice matching is above chance level when the test element of matching tasks are extended beyond a few seconds, but static face-voice matching is not (Experiments 2a, 2b, 5a, 5b). However, the value of this information should not be overstated because even this additional dynamic information is susceptible to disruption, as shown by the fact that performance was at chance level in the dynamic condition of Experiments 5c and 5d, which included a fixation and beep.

Maintaining access to redundant information across faces and voices in complex social settings featuring a number of different speakers, as well as numerous other visual and auditory events, would likely impose a huge cognitive load without offering appreciable benefits. The transient nature of overlapping face-voice identity information therefore makes sense, and may explain why people exhibit a bias to attribute a face and voice to the same identity. As faces and voices of the same person often occur close together in space and time, this bias provides a useful cognitive shortcut. It could be viewed as an additional guarantee that faces and voices belonging to the same person will be correctly attributed to the same identity as a way of organizing the social environment in a meaningful and useful way. This would explain the nature of the temporal position bias observed in Experiments 2a and 2b, as well as the overall bias to respond *same* in Experiments 4a and 4b.

The fact that the bias operates more strongly in the V-A condition than the A-V condition probably reflects matching decisions being driven by identity information rather than information about idiosyncratic speaking style. This explanation links to the conclusion that overlapping speech information is only additionally helpful when this identity information is unavailable. Identity signals emanating from faces are stronger and more reliable than those emanating from voices (Stevenage et al., 2011, 2012, 2013). It is logical that identity signals from faces have a further reach and are therefore more likely to encompass a voice than a voice would be to encompass a face. In keeping with this explanation, when speech information is relied upon to inform accurate matching decisions, as in Experiment 5b where there was a significant difference between static and dynamic face-voice matching, no bias was observed in either the A-V or the V-A condition.

Overall, the results are consistent with the conclusion that the process of face-voice integration begins at an early perceptual stage. It is facilitated by the presence of redundant information. Capitalising on these redundancies is possible even when people have not been exposed to a person's face and voice co-occurring in real life. This ability is not perfect, but provides a useful foundation for full integration at later perceptual stages, as explained in the following section.

9.5 Putting face-voice matching in the context of the wider literature

The early face-processing literature suggested that face-voice integration occurs purely at the post-perceptual PIN stage (Burton et al., 1990; Ellis et al., 1997). According to the IAC model, the PIN contains multimodal signature information about people (e.g. their facial appearance, the sound of their voice, the style of their handwriting), and is strengthened during the process of familiarisation (Burton et al., 1990). The findings presented here are not wholly consistent with the interpretation that integration only occurs at this stage of face and voice processing. Throughout this thesis it has been shown that unfamiliar faces and voices belonging to the same person offer redundant information, and that unfamiliar face-voice matching is possible. These results reflect that the processing of facial and vocal identity information is not totally independent, and is not contingent on familiarisation. The results are therefore more consistent with, and extend, the recent literature highlighting the existence of early perceptual integrative mechanisms between face and voice processing pathways (e.g. Belin et al., 2004). The observation of accurate face-voice matching may help to clarify, or at least to formulate hypotheses about, the construction of multisensory person representations. Even after 2 seconds exposure to novel faces and voices, people can make accurate identity matches (Experiments 2a, 2b, 2c, 4a, 4b). Awareness of redundant face-voice information following such limited exposure, and in the absence of familiarity, may facilitate the building of stable multisensory representations.

As referred to in Chapters 7 and 8, it seems likely that face-voice redundancies provide an important foundation for the successful integration of visual and auditory speech information. The bias to respond *same* exhibits the same asymmetrical pattern as that observed in studies investigating audiovisual speech integration. Successful integration can tolerate auditory lags better than visual lags (Munhall et al., 1996; Robertson & Schweinberger, 2010; Van Wassenhove et al., 2007). Redundancies may additionally act as a foundation for the integration of identity and affect information. There is evidence that bimodally available emotion information plays an important role in social functioning. For example, categorisation of affect is faster when expressed in the face and voice, compared to when it is available in just one modality (e.g. Collignon et al., 2008; Kreifelts, Ethofer, Grodd, Erb, & Wildgruber, 2007). In terms of identity, a number of studies have demonstrated the existence of crossmodality priming (Ellis et al., 1997; Schweinberger et al., 1997; Stevenage et al., 2012; Stevenage, Hale, Morgan & Neil, 2014a). The ability to easily exploit redundancies may be important when building multimodal identity representations during the process of familiarisation, helping to support rapid identity decisions.

Overall, this thesis highlights the value and importance of considering person perception from a multimodal point of view. This is consistent with recent advances in the field and the current state of thinking (see Schweinberger et al., 2014). Person stimuli are best understood as coherent, multimodal wholes. Ignoring this risks attending to artificial constituent parts that do not adequately reflect how people are actually perceived in day-to-day life.

9.6 Limitations of the stimulus set

The corpus (Cooke et al., 2006) used in this thesis contained 34 stimulus individuals, but only 18 of these were matched for age (18-30) and ethnicity. Using a sample matched on these dimensions was critical in order to address gaps in the literature. Previous studies offered contradictory results, but all had tested face-voice matching using highly homogenous stimulus samples (Kamachi et al., 2003; Lachs & Pisoni, 2004a, 2004b; Lander et al., 2007; Mavica & Barenholtz, 2013). For the present purposes therefore, homogeneity was an advantage. However, in terms of maximising generalisabilty it would be desirable to test face-voice matching with a much wider range of individuals in the future.

Guaranteeing homogeneity involved compromising on stimulus sample size, and not including 16 of the people featured in the corpus. This is a limitation of the thesis. Having only 18 stimulus people in the sample meant that there were a small number of trials in each experiment: 12 trials in 2AFC experiments, and 18 in same-different experiments. Although it would have been possible to increase power by repeating trials, this would have introduced the possibility that participants could have learnt face-voice associations, and responded according to decisions taken in previous blocks.

Considering the high level of variability associated with stimuli, it would certainly be advantageous to have access to a larger set of stimuli for testing face-voice matching accuracy. Future face-voice matching studies should aim to use more stimuli than was unfortunately practicable here. Nevertheless, stimulus variability was minimised in each experiment by matching the stimuli in each trial for sex, ethnicity, and age. All of the people in this stimulus set were from similar educational backgrounds (Cooke et al., 2006), and none exhibited strong regional accents. Although Simmons et al. (2011) recommend sample sizes in excess of 20, many studies have used far smaller samples when investigating person perception (see Wells & Windshitl, 1999), as have other face-voice matching studies (e.g. Lachs & Pisoni, 2004a). Furthermore, multilevel modelling enabled us to generalise from stimuli as well as participants. Even in studies using a large sample of stimuli, generalisability is limited by the common practice of aggregating over stimuli (Clark, 1973; Wells et al., 2013; Judd et al., 2012). Ultimately, the question of adequate sample size of stimuli or participants in experimental designs such as those reported here is a question of statistical power (e.g., see Westfall, Kenny & Judd, 2014).

One further limitation of the stimulus set was the potential for content overlap across voice clips. Although each of the 1000 sentences spoken across and within speakers were unique, the sentences were made up of a relatively small pool of words. Therefore, it is possible that some of the sentences spoken by the dynamic face and voice could have featured the same words. However, as dynamic and visual articulations were extracted from separate videos, the words were articulated on different occasions. This means that the participants would not have been able to perform exact pattern matching. Although it is of course plausible that the participants could have attempted to use articulatory matching of specific words in order to inform their decisions, the overall results presented both in the previous literature and in this thesis do not support the conclusion that this extra information is likely to have been particularly helpful. The previous literature has shown that overlap in the style/manner in which a sentence is said is far more informative than overlapping content in terms of supporting face-voice matching accuracy (Kamachi et al., 2003; Lander et al., 2007). Had overlapping content been especially beneficial, significant differences between static and dynamic face-voice matching would have occurred consistently. As it was, there was only a significant difference in one single experiment (Experiment 5b).

9.7 Implications for future research

The results in this thesis offer a number of recommendations for future research. The most specific recommendation relates to the investigation of face-voice integration. People exhibit knowledge of face-voice identity concordance prior to familiarisation (Experiments 2a, 2b, 2c, 4a, 4b and 5a). This must be taken into account when designing integration experiments. The results of some behavioural studies demonstrating priming effects (e.g. Ellis et al., 1997; Stevenage et al., 2012 etc.), and interference effects (e.g. Schweinberger et al., 2007) could in fact be attributed to cognitive processes separate from the kind of perceptual binding which occurs during the process of familiarisation. Familiarisation involves a multimodal person representation being stored in memory (Burton et al., 1990). However, it may in fact be the case that participants can capitalise on face-voice redundancies even when they have had no prior exposure to a person. Testing participants using novel faces and voices should be included as a control condition to help establish whether observed effects are exclusively attributable to face-voice integration following familiarisation.

A more serious issue is that some studies have tested integration effects using supposed face-voice pairs that in fact belong to different people. Joassin et al. (2011) used faces from the Stirling Face Database, and (Belgian) voices recorded in the laboratory to measure brain activity with functional magnetic resonance imaging (fMRI) during unimodal and bimodal recognition. Faces and voices from the same person should be used when addressing face-voice integration in person perception; people may respond to them differently compared to faces and voices that do not share an identity. The results from studies failing to satisfy this criterion may have limited generalisability to everyday social contexts.

A further recommendation relates to methods of statistical analysis. Based on hypotheses informed by the face-space model (Valentine, 1991), plus the existence of inter-stimulus variability in terms of how faces look and voices sound (Burton, 2013; Stevenage & Neil, 2014; Valentine et al., 2015), the use of multilevel modelling was strongly recommended in Chapter 3 (see section 3.3.1). In all the experiments reported here, there was a high level of variability at the stimuli level, showing that people differ in the extent to which they look and sound similar. Notably, in the majority of experiments, inter-participant variability was limited. Participants tended to respond similarly to the same stimuli. This finding suggested one possible explanation for previous contradictions in the literature (Chapter 5). In this case, using traditional ANOVA would have led to a different set of results and conclusions, as demonstrated by the results presented in Appendix A. Therefore, it is important to underline previous recommendations (Clark, 1973; Judd et al., 2012; Kreft & De Leeuw, 1998) calling for studies using sets of variable stimuli (e.g. faces, voices, words etc.) to employ multilevel modelling as a matter of course.

The last recommendation for future research is the most general. 2AFC tasks have traditionally been a staple of psychological investigation into a wide variety of topics. Whilst it would be highly naïve to expect this procedure to never be used in the future, it is important to highlight some issues associated with its use. The reported results support calls for caution regarding 2AFC tasks (Garcia-Perez & Alcala-Quintana, 2010, 2011; Yeshurun et al., 2008). The procedure certainly does not appear to represent an unbiased way of testing performance, as has been previously suggested (Green & Swets, 1973; Macmillan & Creelman, 2005; Wickens, 2002). If 2AFC tasks are to be used in experiments, it is necessary to appropriately interrogate the results, considering whether there is a position effect, whether the position effect is temporal or spatial, and why the position effect manifests according to a particular pattern. Depending on the research question, having considered these issues, it may be necessary to account for the results by using alternative tasks (e.g. same-different) to disentangle decision processes driving performance.

9.8 Outstanding research questions and possible future directions

Face-voice matching has been addressed in very few studies. However, this is an issue with theoretical implications in terms of multimodal person perception. The topic also has applied relevance (Section 9.9). Clearly additional research is necessary in order to further clarify how face-voice matching operates. A number of outstanding research questions arise from the findings presented in this thesis, some of which are considered below.

If, as suggested in Chapter 8, accurate face-voice matching relies on comparing high-quality perceptual representations, further strengthening these perceptual representations by increasing temporal exposure to faces and voices might improve performance. As emphasised by previous research, although face and voice information is processed in parallel, this does not mean that the processes are identical (Belin et al., 2004). In all of the experiments in this thesis, the exposure time to faces and voices was equal (2 seconds). Comparing whether matching is more accurate if exposure time to the face or voice is increased would address the relative contribution of face and voice information to matching decisions.

In this thesis, primarily British participants were tested using exclusively British stimuli. In keeping with the own-ethnicity bias in face recognition, the ability to match faces and voices might have an important cultural underpinning relating to expertise and exposure (e.g. see Levin, 2000; Meissner & Brigham, 2001; Tanaka, 2001). For instance, British participants might have expertise in exploiting concordant information to make accurate face-voice matching decisions when the stimuli are British, but not when the stimuli are, for example, Japanese or African. Expertise might play a role in enabling accurate matching because of cultural differences in voice production. For example, in Japan women speak with a higher pitch than Western women in order to appear modest, polite and feminine (Loveday, 1981; van Bezooijen, 1995, 1996). This could make it difficult for British people to match Japanese faces and voices for identity.

Very recently, Stevenage, Hamlin and Ford (2015) considered what types of strategies people might be using to reach accurate face-voice matching decisions. They found that the strategies identified by participants did not predict performance. However, although overlapping cues might be present in faces and voices (Experiment 1) this does not mean that people necessarily utilise the most informative cues when making a matching decision. They might not even be conscious of the influences operating on their choices. Further research might investigate whether it is possible to prime participants to use the most informative cues, and therefore improve the accuracy of their matching decisions. For example, the results presented in Experiment 1 showed that the correlation between faces and voices on a scale of masculinity/femininity was .95. Participants could be instructed to try and base their identity decisions on whether the face and voice exhibited similar levels of perceived masculinity/femininity. Accuracy might be significantly higher than if they used a less informative cue, such as weight.

9.9 Applied relevance of the findings

Following a crime, witnesses might be required to identify a perpetrator at lineup from their face, and in some cases from their voice. Both types of evidence can be admitted to court, and often constitute pivotal evidence. Unfamiliar voice identification is particularly unreliable (Ormerod, 2001), and is significantly less accurate than face identification (Stevenage et al., 2011, 2012). Performance is frequently at chance level (Yarmey, 2007), which suggests that voice lineup decisions are based on guessing. This poses a particular problem if the witness never sees the perpetrator's face, but does hear his/her voice. The results of Chapter 4 suggest that certain information provided by a voice might still be useful in a forensic setting. People tend to agree with each other about judgements made from faces and voices, and also make similar judgements based on faces and voices (Experiment 1). Therefore, witnesses' perception of the perpetrator based on their voice (e.g. in terms of masculinity femininity, health or height), providing this perception was well retained in memory, is likely to correspond to the way that person looks. An earwitness' information might be useful in helping to narrow down a list of suspects, or perhaps images of suspects captured by CCTV.

The findings have further forensic implications in terms of mapping identity blueprints. The SuperIdentity (SID) Project (Guest, Miguel-Hurtado, Stevenage, Neil & Black, 2014) adopts a multimodal and multi-dimensional approach to the investigation of identity. The project tests how elements of identity, expressed across different contexts (e.g. face, voice and behaviour) link together to create an holistic biometric 'picture'. The generation of identity maps has obvious utility for security and intelligence services. The experiments presented in this thesis demonstrate clear links between face and voice identity as identified by humans. The SID Project has considered the differences between diagnostic identity decisions made by machines and humans (Stevenage, Walpole, Neil & Black, 2014c), although not in the context of face-voice matching. Covert recordings of voices might be used in court as evidence. It is possible to imagine a situation when a voice recording needs to be compared to a mugshot, or perhaps to the image of a deceased person. Future research might consider whether machines can isolate features of the voiceprint that predict visual structural features of the face, even whether they can categorise faces and voices according to identity. This is theoretically possible if hormonal profiles affect facial structure (Miller & Todd, 1998; Penton-Voak & Chen, 2004; Perrett et al., 1998; Thornhill & Grammer, 1999) as well as the physiology of the vocal apparatus (Abitbol et al., 1999; Beckford et al., 1985; Hollien, 1960; O'Connor et al., 2011). Human performance will likely rely on alternative strategies, and be influenced differently by bias (Dror & Charlton, 2006; Dror & Hampikian, 2011; Nakhaeizadeh, Dror & Morgan, 2014). Bias clearly affects face-voice matching performance (Experiments 2a, 2b, 3a, 3b, 4a, 4b), and as suggested above, encouraging people to attend to the most informative redundant cues when making a matching decision might optimise human accuracy. Therefore, a high level of diagnosticity could potentially be achieved by taking into account responses generated by both humans and machines.

The reported findings are also relevant to the entertainment industry. British and American films are commonly exported to other countries, where actors' voices are dubbed in the native language. Dubbing is now far more common than subtitling in many countries because it imposes less of a cognitive load, and also improves impact and the feeling of presence (Chaume, 2013; Wissmath, Weibel & Groner, 2009). Voice actors frequently provide dubbing voices for more than one British/American actor. Although of course a

British or American actor could not feasibly provide their own voice for dubbing in a different language, it would be advisable for dubbing companies to carefully check whether the different identity face-voice pairs were perceived as being a good match. Perhaps obvious incongruence might be distracting, or make the characters less convincing. Future research could consider whether very seemingly wrong face-voice pairs compromise the ability of the audience to follow the story line. Disjointed identities could be disorienting, or might significantly affect the audience's perceived enjoyment of the film. This could have financial implications for the film industry.

9.10 Conclusion

This thesis makes a number of original contributions to the existing literature, providing a clearer understanding of face-voice matching performance than it is possible to glean from previous studies. The results show that people look and sound similar. Faces and voices presented close together in time can be accurately matched for identity, although a bias operates to make it more likely that faces and voices will be attributed to the same identity. The overall findings offer some clues about how people might successfully navigate complex social situations. In addressing issues relating to experimental procedure, highlighting the shortcomings of 2AFC tasks, as well as the need to use multilevel modelling when analysing face and voice data, this thesis draws to attention a number of methodological issues of more general application to the investigation of face and voice processing. Considering the clear theoretical and applied relevance of face-voice matching, this is an important topic for the future.

References

Abend, P., Pflüger, L. S., Koppensteiner, M., Coquerelle, M., & Grammer, K. (2015). The sound of female shape: a redundant signal of vocal and facial attractiveness. *Evolution and Human Behavior*, *36*(3), 174-181. doi: 10.1016/j.evolhumbehav.2014.10.004

Abercrombie, D. (1967). Elements of general phonetics. Edinburgh: Edinburgh University Press

Abitbol, J., Abitbol, P., & Abitbol, B. (1999). Sex hormones and the female voice. *Journal of Voice*, *13*(3), 424-446. doi: 10.1016/S0892-1997(99)80048-4

Ahrens, M. M., Hasan, B. A. S., Giordano, B. L., & Belin, P. (2014). Gender differences in the temporal voice areas. *Frontiers in Neuroscience*, *8,* 1-8. doi: 10.3389/fnins.2014.00228

Allison, T., Puce, A., & McCarthy, G. (2000). Social perception from visual cues: role of the STS region. *Trends in Cognitive Sciences*, *4*(7), 267-278. doi: 10.1016/S1364-6613(00)01501-1

Azzopardi, P., & Cowey, A. (1998). Blindsight and visual awareness. *Consciousness and Cognition*, *7*(3), 292-311. doi: 10.1006/ccog.1998.0358

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390-412. doi: 10.1016/j.jml.2007.12.005

Baayen, R.H. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using*

R. Cambridge: Cambridge University Press

Baddeley, A. (2007). *Working memory, thought, and action*. Oxford: Oxford University Press

Baguley, T. (2012). *Serious stats: A guide to advanced statistics for the behavioral sciences.* Basingstoke: Palgrave

Balsdon, T., & Azzopardi, P. (2015). Absolute and relative blindsight. *Consciousness and Cognition*, *32*, 79-91. doi: 10.1016/j.concog.2014.09.010

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology, 70*(3), 614-636. doi: 10.1037/0022-3514.70.3.614

Barsics, C., & Brédart, S. (2012). Recalling semantic information about newly learned faces and voices. *Memory*, *20*(5), 527-534. doi: 10.1080/09658211.2012.683012

Bartlett, J. C., Hurry, S., & Thorley, W. (1984). Typicality and familiarity of faces. *Memory & Cognition*, *12*(3), 219-228. doi: 10.3758/BF03197669

Bates, D, Maechler, M., Bolker, B., & Walker, S. (2014) lme4: Linear mixed-effects models using Eigen and S4. R package version 1.0-6. Available at http://CRAN.R-project.org/package=lme4

Beckford, N. S., Rood, S. R., & Schaid, D. (1985). Androgen stimulation and laryngeal development. *Annals of Otology, Rhinology, and Laryngology, 94,* 634–640.

Belin, P., Bestelmeyer, P. E., Latinus, M., & Watson, R. (2011). Understanding voice perception. *British Journal of Psychology*, *102*(4), 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. doi: 10.1016/j.tics.2004.01.008

Belin, P., Zatorre, R. J., & Ahad, P. (2002). Human temporal-lobe response to vocal sounds. *Cognitive Brain Research*, *13*(1), 17-26. doi: 10.1016/S0926-6410(01)00084-2

Benoit, C., Mohamadi, T., & Kandel, S. (1994). Effects of phonetic context on audio-visual intelligibility of French. *Journal of Speech, Language, and Hearing Research*, *37*(5), 1195-1203. doi: 10.1044/jshr.3705.1195

Bernstein, L. E., Auer, E. T., & Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading. *Speech Communication*, *44*(1), 5-18. doi: 10.1016/j.specom.2004.10.011

Besle, J., Fischer, C., Bidet-Caulet, A., Lecaignard, F., Bertrand, O., & Giard, M. H. (2008). Visual activation and audiovisual interactions in the auditory cortex during speech perception: Intracranial recordings in humans. *The Journal of Neuroscience*, *28*(52), 14301-14310. doi: 10.1523/JNEUROSCI.2875-08.2008

Birkhead, T. R., Fletcher, F., & Pellatt, E. J. (1998). Sexual selection in the zebra finch Taeniopygia guttata: condition, sex traits and immune capacity. *Behavioral Ecology and Sociobiology*, *44*(3), 179-191.doi: 10.1007/s002650050530

Blake, R., Cepeda, N. J., & Hiris, E. (1997). Memory for visual motion. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(2), 353-369. doi: 10.1037/0096-1523.23.2.353

Blank, H., Anwander, A., & von Kriegstein, K. (2011). Direct structural connections between voice-and face-recognition areas. *The Journal of Neuroscience*, *31*(36), 12906-12915. doi: 10.1523/JNEUROSCI.2091-11.2011

Braida, L. D. (1991). Crossmodal integration in the identification of consonant segments. *The Quarterly Journal of Experimental Psychology, 43*(3), 647–677. doi: 10.1080/14640749108400991

Braun, A. (1996). Age estimation by different listener groups. *International Journal of Speech Language and the Law*, *3*(1), 65-73. doi: 10.1558/ijsll.v3i1.65

Bruce, V., Burton, M. A., & Dench, N. (1994). What's distinctive about a distinctive face? *The Quarterly Journal of Experimental Psychology*, *47*(1), 119-141. doi: 10.1080/14640749408401146

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, *77*(3), 305-327. doi: 10.1111/j.2044-8295.1986.tb02199.x

Bruce, V., & Young, A. W. (2012). *Face perception.* London: Psychology Press.

Bruckert, L., Liénard, J. S., Lacroix, A., Kreutzer, M., & Leboucher, G. (2006). Women use voice parameters to assess men's characteristics. *Proceedings of the Royal Society, B: Biological Sciences*, *273*(1582), 83-89. doi: 10.1098/rspb.2005.3265

Brunswik, E. (1947). *Systematic and representative design of psychological experiments*. Berkeley: University of California Press

Buehner, M. J., & May, J. (2003). Rethinking temporal contiguity and the judgement of causality: Effects of prior knowledge, experience, and reinforcement procedure. *The Quarterly Journal of Experimental Psychology Section A*, *56*(5), 865-890. doi:

10.1080/02724980244000675

Burt, D. M., & Perrett, D. I. (1995). Perception of age in adult Caucasian male faces: Computer graphic manipulation of shape and colour information. *Proceedings of the Royal Society, B: Biological Sciences*, *259*(1355), 137-143. doi: 10.1098/rspb.1995.0021

Burton, A. M. (2013). Why has research in face recognition progressed so slowly? The importance of variability. *The Quarterly Journal of Experimental Psychology*, *66*(8), 1467-1485. doi: 10.1080/17470218.2013.800125

Burton, A. M., Bruce, V., & Johnston, R. A. (1990). Understanding face recognition with an interactive activation model. *British Journal of Psychology*, *81*(3), 361-380. doi: 10.1111/j.2044-8295.1990.tb02367.x

Burton, A. M., Jenkins, R., & Schweinberger, S. R. (2011). Mental representations of familiar faces. *British Journal of Psychology*, *102*(4), 943-958. doi: 10.1111/j.2044-8295.2011.02039.x

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., ... & David, A. S. (1997). Activation of auditory cortex during silent lipreading. *Science*, *276*(5312), 593-596. doi: 10.1126/science.276.5312.593

Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences*, *11*(12), 535-543. doi: 10.1016/j.tics.2007.10.001

Candolin, U. (2003). The use of multiple cues in mate choice. *Biological Reviews*, *78*(4), 575-595. doi: 10.1017/S1464793103006158

Carlson, C. A., Gronlund, S. D., & Clark, S. E. (2008). Lineup composition, suspect position, and the sequential lineup advantage. *Journal of Experimental Psychology: Applied*, *14*(2), 118-128. doi: 10.1037/1076-898X.14.2.118

Chao, L. L., Martin, A., & Haxby, J. V. (1999). Are face ‐ responsive regions selective only for faces? *Neuroreport*, *10*(14), 2945-2950. doi: 10.1097/00001756-199909290-00013

Charest, I., Pernet, C., Latinus, M., Crabbe, F., & Belin, P. (2013). Cerebral processing of voice gender studied using a continuous carryover fMRI design. *Cerebral Cortex*, *23*(4), 958-966. doi: 10.1093/cercor/bhs090

Chaume, F. (2013). The turn of audiovisual translation: New audiences and new technologies. *Translation Spaces*, *2*(1), 105-123.

Chiller-Glaus S. D., Schwaninger A., Hofer F., Kleiner M., Knappmeyer B. (2011). Recognition of emotion in moving and static composite faces. *Swiss Journal of Psychology, 70*(4), 233–240. doi: 10.1024/1421-0185/a000061

Christie, F., & Bruce, V. (1998). The role of dynamic information in the recognition of unfamiliar faces. *Memory & Cognition, 26*(4), 780-790. doi: 10.3758/BF03211397

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*(4), 335-359. doi: 10.1016/S0022-5371(73)80014-3

Clark, S. E., Howell, R. T., & Davey, S. L. (2008). Regularities in eyewitness identification. *Law and Human Behavior*, *32*(3), 187-218. doi: 10.1007/s10979-006-9082-4

Coetzee, V., Chen, J., Perrett, D. I., & Stephen, I. D. (2010). Deciphering faces: Quantifiable visual cues to weight. *Perception*, *39*(1), 51-61. doi: 10.1068/p6560

Collignon, O., Girard, S., Gosselin, F., Roy, S., Saint-Amour, D., Lassonde, M., & Lepore, F. (2008). Audio-visual integration of emotion expression. *Brain Research*, *1242*, 126-135. doi: 10.1016/j.brainres.2008.04.023

Collins, S. A. (2000). Men's voices and women's choices. *Animal Behaviour*, *60*(6), 773-780. doi: 10.1006/anbe.2000.1523

Collins, S. A., & Missing, C. (2003). Vocal and visual attractiveness are related in women. *Animal Behaviour*, *65*(5), 997-1004. doi: 10.1006/anbe.2003.2123

Coltheart, M. (1980). Iconic memory and visible persistence. *Perception & Psychophysics*, *27*(3), 183-228. doi: 10.3758/BF03204258

Cooke, M., Barker, J., Cunningham, S., & Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America, 120*(5), 2421–2424. doi: 10.1121/1.2229005.

Crowder, R. G., & Morton, J. (1969). Precategorical acoustic storage (PAS). *Perception & Psychophysics*, *5*(6), 365-373. doi: 10.3758/BF03210660

Dabbs Jr, J. M., & Mallinger, A. (1999). High testosterone levels predict low voice pitch among men. *Personality and Individual Differences*, *27*(4), 801-804. doi: 10.1016/S0191-8869(98)00272-4

Damjanovic, L., & Hanley, J. R. (2007). Recalling episodic and semantic information about famous faces and voices. *Memory & Cognition, 35*(6), 1205-1210. doi: 10.3758/BF03193594

de Gelder, B., & Vroomen, J. (2000). The perception of emotions by ear and by eye.

Cognition & Emotion, *14*(3), 289-311. doi: 10.1080/026999300378824

DeCarlo, L. T. (2012). On a signal detection approach to m-alternative forced choice with bias, with maximum likelihood and Bayesian approaches to estimation. *Journal of Mathematical Psychology*, *56*(3), 196-207. doi: 10.1016/j.jmp.2012.02.004

Dohen, M., Lœvenbruck, H., Cathiard, M. A., & Schwartz, J. L. (2004). Visual perception of contrastive focus in reiterant French speech. *Speech Communication*, *44*(1), 155-172. doi: 10.1016/j.specom.2004.10.009

Dolan, R. J., Morris, J. S., & de Gelder, B. (2001). Crossmodal binding of fear in voice and face. *Proceedings of the National Academy of Sciences*, *98*(17), 10006-10010. doi: 10.1073/pnas.171288598

Dror, I. E., & Charlton, D. (2006). Why experts make errors. *Journal of Forensic Identification*, *56*(4), 600-616.

Dror, I. E., & Hampikian, G. (2011). Subjectivity and bias in forensic DNA mixture interpretation. *Science & Justice*, *51*(4), 204-208. doi: 10.1016/j.scijus.2011.08.004

Dyjas, O., Bausenhart, K. M., & Ulrich, R. (2012). Trial-by-trial updating of an internal reference in discrimination tasks: Evidence from effects of stimulus order and trial sequence. *Attention, Perception, & Psychophysics*, *74*(8), 1819-1841. doi: 10.3758/s13414-012-0362-4

Ebbesen, E. B., & Flowe, H. D. (2002). *Simultaneous v. sequential lineups: What do we really know?* Unpublished manuscript.

Ellis, H. D., Jones, D. M., & Mosdell, N. (1997). Intra-and inter-modal repetition priming of familiar faces and voices. *British Journal of Psychology, 88*(1), 143-156. doi: 10.1111/j.2044-8295.1997.tb02625.x

Ellison, P. T. (1999). Reproductive ecology and reproductive cancers. In C. Pater-Brick and C. Worthman (Eds.), *Hormones, health, and behavior: A socio-ecological and lifespan perspective* (pp. 184-209). Cambridge: Cambridge University Press.

Ethofer, T., Bretscher, J., Wiethoff, S., Bisch, J., Schlipf, S., Wildgruber, D., & Kreifelts, B. (2013). Functional responses and structural connections of cortical areas for processing faces and voices in the superior temporal sulcus. *Neuroimage*, *76*(1), 45-56. doi: 10.1016/j.neuroimage.2013.02.064

Falk, D. (2005). Prelinguistic evolution in early hominins: Whence motherese? *Behavioral and Brain Sciences, 27*(4), 491–503. doi: 10.1017/S0140525X04000111

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

Farley, S. D., Hughes, S. M., & LaFayette, J. N. (2013). People will know we are in love: Evidence of differences between vocal samples directed toward lovers and friends. *Journal of Nonverbal Behavior*, *37*(3), 123-138. doi: 10.1007/s10919-013-0151-3

Feinberg, D. R. (2008). Are human faces and voices ornaments signaling common underlying cues to mate value? *Evolutionary Anthropology: Issues, News, and Reviews*, *17*(2), 112-118. doi: 10.1002/evan.20166

Feinberg, D. R., Jones, B. C., DeBruine, L. M., Moore, F. R., Law Smith, M. J., Cornwell, R. E., ... & Perrett, D. I. (2005). The voice and face of woman: One ornament that signals quality? *Evolution and Human Behavior*, *26*(5), 398-408. doi: 10.1016/j.evolhumbehav.2005.04.001

Fitch, W. T. (2000). The evolution of speech: a comparative review. *Trends in Cognitive Sciences*, *4*(7), 258-267. doi: 10.1016/S1364-6613(00)01494-7

Flowe, H. D., Smith, H. M., Karoğlu, N., Onwuegbusi, T. O., & Rai, L. (2015). Configural and component processing in simultaneous and sequential lineup procedures. *Memory*, (ahead-of-print), 1-9. doi: 10.1080/09658211.2015.1004350

Folstad, I., & Karter, A. J. (1992). Parasites, bright males, and the immunocompetence handicap. *American Naturalist, 139*(3), 603-622. doi: 10.1086/285346.

Fraccaro, P. J., Feinberg, D. R., DeBruine, L. M., Little, A. C., Watkins, C. D., & Jones, B. C. (2010). Correlated male preferences for femininity in female faces and voices. *Evolutionary Psychology*, *8*(3), 447-461. doi: 10.1.1.174.4476

Fraccaro, P., Jones, B., Vukovic, J., Smith, F., Watkins, C., Feinberg, D., ... & Debruine, L. (2011). Experimental evidence that women speak in a higher voice pitch to men they find attractive. *Journal of Evolutionary Psychology*, *9*(1), 57-67. doi: 10.1556/JEP.9.2011.33.1

Gangestad, S. W., & Scheyd, G. J. (2005). The evolution of human physical attractiveness. *Annual Review of Anthropology*, *34*, 523-548. doi: 10.1146/annurev.anthro.33.070203.143733

García-Pérez, M. A., & Alcalá-Quintana, R. (2010). The difference model with guessing explains interval bias in two-alternative forced-choice detection procedures. *Journal of Sensory Studies, 25*(6), 876-898. doi: 10.1111/j.1745-459X.2010.00310.x

García-Pérez, M. A., & Alcalá-Quintana, R. (2011). Improving the estimation of psychometric functions in 2AFC discrimination tasks. *Frontiers in Psychology*, *2*, 96. doi: 10.3389%2Ffpsyg.2011.00096

Gelman, A. E., & Su, Y. S. (2013). arm: Data analysis using regression and multilevel/hierarchical models. R package version 1.6-05. Available at http://CRAN.R-project.org/package=arm

Ginns, P. (2006). Integrating information: A meta-analysis of the spatial contiguity and temporal contiguity effects. *Learning and Instruction*, *16*(6), 511-525. doi: 10.1016/j.learninstruc.2006.10.001

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition*, *13*(1), 8-20. doi: 10.3758/BF03198438

Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: data and theory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *16*(1), 5-16. doi: 10.1037//0278-7393.16.1.5

Glanzer, M., & Cunitz, A. R. (1966). Two storage mechanisms in free recall. *Journal of Verbal Learning and Verbal Behavior*, *5*(4), 351-360. doi: 10.1016/S0022-5371(66)80044-0

Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review, 100*(3), 546-567. doi: 10.1037/0033-295X.100.3.546

Gobbini, M. I., Gentili, C., Ricciardi, E., Bellucci, C., Salvini, P., Laschi, C., ... & Pietrini, P. (2011). Distinct neural systems involved in agency and animacy detection. *Journal of Cognitive Neuroscience*, *23*(8), 1911-1920. doi: 10.1162/jocn.2010.21574

Gray, A., Berlin, J. A., McKinlay, J. B., & Longcope, C. (1991). An examination of research design effects on the association of testosterone and male aging: results of a meta-analysis. *Journal of Clinical Epidemiology*, *44*(7), 671-684. doi: 10.1016/0895-4356(91)90028-8

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Gregory, S. W. (1996). A nonverbal signal in voices of interview partners effectively predicts communication accommodation and social status perceptions. *Journal of Personality and Social Psychology, 70*(6), 1231–1240. doi: 10.1037/0022-3514.70.6.1231

Grill-Spector, K., Knouf, N., & Kanwisher, N. (2004). The fusiform face area subserves face perception, not generic within-category identification. *Nature Neuroscience*, *7*(5), 555-562. doi: 10.1038/nn1224

Gronlund, S. D. (2005). Sequential lineup advantage: Contributions of distinctiveness and recollection. *Applied Cognitive Psychology*, *19*(1), 23-37. doi: 10.1002/acp.1047

Gronlund, S. D., Wixted, J. T., & Mickes, L. (2014). Evaluating eyewitness identification procedures using receiver operating characteristic analysis. *Current Directions in Psychological Science, 23*(1), 3-10. doi: 10.1177/0963721413498891

Guest, R., Miguel-Hurtado, O., Stevenage, S. V., Neil, G. J., & Black, S. (2014). Biometrics within the SuperIdentity project: A new approach to spanning multiple identity domains. In *2014 International Carnahan Conference on Security Technology (ICCST),* (pp. 1-6).

Hancock, P. J., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences*, *4*(9), 330-337. doi: 10.1016/S1364-6613(00)01519-9

Hancock, P. J., Burton, A. M., & Bruce, V. (1996). Face processing: Human perception and principal components analysis. *Memory & Cognition, 24*(1), 26-40. doi: 10.3758/BF03197270

Handkins, R. E., & Cross, J. F. (1985). Can a voice lineup be too fair? Paper presented at the annual meeting of the Midwestern Psychology Association, Chicago, IL.

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. doi: 10.1080/713755942

Hodges-Simeon, C. R., Gurven, M., Puts, D. A., & Gaulin, S. J. (2014). Vocal fundamental and formant frequencies are honest signals of threat potential in peripubertal males. *Behavioral Ecology*, *25*(4), 984-988. doi: 10.1093/beheco/aru081

Hoffman, E. A., & Haxby, J. V. (2000). Distinct representations of eye gaze and identity in the distributed human neural system for face perception. *Nature Neuroscience*, *3*(1), 80-84. doi: 10.1038/71152

Hoffman, L., & Rovine, M. J. (2007). Multilevel models for the experimental psychologist: Foundations and illustrative examples. *Behavior Research Methods*, *39*(1), 101-117. doi: 10.3758/BF03192848

Hollien, H. (1960). Some laryngeal correlates of vocal pitch. *Journal of Speech, Language, and Hearing Research, 3*(1), 52-58. doi: 10.1044/jshr.0301.52

Howard, I. P., & Templeton, W. B. (1966). *Human spatial orientation.* New York: Wiley

Hughes, S. M., Farley, S. D., & Rhodes, B. C. (2010). Vocal and physiological changes in response to the physical attractiveness of conversational partners. *Journal of Nonverbal Behavior, 34*(3), 155-167. doi: 10.1007/s10919-010-0087-9

Jäkel, F., & Wichmann, F. A. (2006). Spatial four-alternative forced-choice method is the preferred psychophysical method for naïve observers. *Journal of Vision, 6*(11), 1307-1322. doi: 10.1167/6.11.13

Jang, Y., Wixted, J. T., & Huber, D. E. (2009). Testing signal-detection models of yes/no and two-alternative forced-choice recognition memory. *Journal of Experimental Psychology: General*, *138*(2), 291-306. doi: 10.1037/a0015525

Jenkins, J. S. (1998). The voice of the castrato. *The Lancet*, *351*(9119), 1877-1880. doi: 10.1016/S0140-6736(97)10198-2

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition*, *121*(3), 313-323. doi: 10.1016/j.cognition.2011.08.001

Joassin, F., Pesenti, M., Maurage, P., Verreckt, E., Bruyer, R., & Campanella, S. (2011). Cross-modal interactions between human faces and voices involved in person recognition. *Cortex*, *47*(3), 367-376. doi: 10.1016/j.cortex.2010.03.003

Johnstone, R. A. (1997). The evolution of animal signals. In J.R. Krebs and N. B. Davies (Eds.), *Behavioural ecology: An evolutionary approach* (pp. 155-178). Oxford: Blackwell.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103*(1), 54-69. doi: 10.1037/a0028347

Kamachi, M., Hill, H., Lander, K., & Vatikiotis-Bateson, E. (2003). Putting the face to the voice: Matching identity across modality. *Current Biology, 13*(19), 1709-1714. doi: 10.1016/j.cub.2003.09.005

Kenny, D. A. (1985). Quantitative methods for social psychology. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (Vol. 1, pp. 487-508). New York: Random House.

Kerstholt, J. H., Jansen, N. J., Van Amelsvoort, A. G., & Broeders, A. P. A. (2004). Earwitnesses: effects of speech duration, retention interval and acoustic environment. *Applied Cognitive Psychology*, *18*(3), 327-336. doi: 10.1002/acp.974

Knappmeyer, B., Thornton, I. M., & Bülthoff, H. H. (2003). The use of facial motion and facial form during the processing of identity. *Vision Research*, *43*(18), 1921-1936. doi: 10.1016/S0042-6989(03)00236-0

Kneller, W., Memon, A., & Stevenage, S. (2001). Simultaneous and sequential lineups: Decision processes of accurate and inaccurate eyewitnesses. *Applied Cognitive Psychology*, *15*(6), 659-671. doi: 10.1002/acp.739

Kościński, K. (2013). Perception of facial attractiveness from static and dynamic stimuli. *Perception*, *42*(2), 163-175. doi: 10.1068/p7378

Krauss, R. M., Freyberg, R., & Morsella, E. (2002). Inferring speakers' physical attributes from their voices. *Journal of Experimental Social Psychology*, *38*(6), 618-625. doi: 10.1016/S0022-1031(02)00510-3

Kreft, I. G., Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: Sage.

Kreifelts, B., Ethofer, T., Grodd, W., Erb, M., & Wildgruber, D. (2007). Audiovisual integration of emotional signals in voice and face: an event-related fMRI study. *Neuroimage*, *37*(4), 1445-1456. doi: 10.1016/j.neuroimage.2007.06.020

Lachs, L. (1999). A voice is a face is a voice: Cross-modal source identification of indexical information in speech. In *Research on Spoken Language Processing, Progress Report No. 23* (pp. 241-258) Bloomington, Indiana: Speech Research Laboratory, Department of Psychology, Indiana University.

Lachs, L., & Pisoni, D. B. (2004a). Crossmodal source identification in speech perception. *Ecological Psychology*, *16*(3), 159-187. doi: 10.1207/s15326969eco1603_1

Lachs, L., & Pisoni, D. B. (2004b). Specification of cross-modal source information in isolated kinematic displays of speech. *The Journal of the Acoustical Society of America*, *116*(1), 507-518. doi: 10.1121/1.1757454

Lander, K. (2008). Relating visual and vocal attractiveness for moving and static faces. *Animal Behaviour, 75*(3), 817-822. doi: 10.1016/j.anbehav.2007.07.001

Lander, K., & Chuang, L. (2005). Why are moving faces easier to recognize? *Visual Cognition*, *12*(3), 429-442. doi: 10.1080/13506280444000382

Lander, K., Hill, H., Kamachi, M., & Vatikiotis-Bateson, E. (2007). It's not what you say but the way you say it: Matching faces and voices. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(4), 905-914. doi: 10.1037/0096-1523.33.4.905

Lass, N. J., & Colt, E. G. (1980). A comparative study of the effect of visual and auditory cues on speaker height and weight identification. *Journal of Phonetics*, *8*(3), 277-285.

Latinus, M., McAleer, P., Bestelmeyer, P. E., & 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

Law Smith, M. J., Perrett, D. I., Jones, B. C., Cornwell, R. E., Moore, F. R., Feinberg, D. R., ... & Hillier, S. G. (2006). Facial appearance is a cue to oestrogen levels in women. *Proceedings of the Royal Society B: Biological Sciences*, *273*(1583), 135-140. doi: 10.1098/rspb.2005.3296

Leongómez, J. D., Binter, J., Kubicová, L., Stolařová, P., Klapilová, K., Havlíček, J., & Roberts, S. C. (2014). Vocal modulation during courtship increases proceptivity even in naive listeners. *Evolution and Human Behavior*, *35*(6), 489-496. doi: 10.1016/j.evolhumbehav.2014.06.008

Levin, D. T. (2000). Race as a visual feature: using visual search and perceptual discrimination tasks to understand face categories and the cross-race recognition deficit. *Journal of Experimental Psychology: General*, *129*(4), 559-574. doi: 10.1037/0096-3445.129.4.559

Lewis, M. B., & Johnston, R. A. (1997). The Thatcher illusion as a test of configural disruption. *Perception*, *26*(2), 225-227. doi: 10.1068/p260225

Light, L. L., Kayra-Stuart, F., & Hollander, S. (1979). Recognition memory for typical and unusual faces. *Journal of Experimental Psychology*, *5*(3), 212-228. doi: 10.1037/0278-7393.5.3.212

Lindsay, R. C. L., & Wells, G. L. (1985). Improving eyewitness identifications from lineups: Simultaneous versus sequential lineup presentation. *Journal of Applied Psychology*, *70*(3), 556-564. doi: 10.1037/0021-9010.70.3.556

Lindsay, R. C. L., Lea, J. A., Nosworthy, G. J., Fulford, J. A., Hector, J., LeVan, V., & Seabrook, C. (1991). Biased lineups: Sequential presentation reduces the problem. *Journal of Applied Psychology*, *76*(6), 796-802. doi: 10.1037/0021-9010.76.6.796

Lindsay, R. C. L., Mansour, J. K., Beaudry, J. L., Leach, A. M., & Bertrand, M. I. (2009). Sequential lineup presentation: Patterns and policy. *Legal and Criminological Psychology*, *14*(1), 13-24. doi: 10.1348/135532508X382708

Linville, S. E. (1996). The sound of senescence. *Journal of Voice*, *10*(2), 190-200. doi: 10.1016/S0892-1997(96)80046-4

Loveday, L. (1981). Pitch, politeness and sexual role: An exploratory investigation into the pitch correlates of English and Japanese politeness formulae. *Language and Speech*, *24*(1), 71-89. doi: 10.1177/002383098102400105

Lu, Z. L., Williamson, S., & Kaufman, L. (1992). Behavioral lifetime of human auditory sensory memory predicted by physiological measures. *Science*, *258*(5088), 1668-1670. doi: 10.1126/science.1455246

Luce, R. D. (1963). Detection and recognition. In R.D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 103-189). New York: Wiley.

MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Perception & Psychophysics, 24*(3)*,* 253–257. doi: 10.3758/BF03206096

MacLeod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology*, *21*(2), 131-141. doi: 10.3109/03005368709077786

Macmillan, N. A., & Creelman, C. D. (2005). *Detection Theory: A User's Guide* (2nd ed.). New York: Lawrence Erlbaum Associates.

Magnussen, S., Idås, E., & Myhre, S. H. (1998). Representation of orientation and spatial frequency in perception and memory: a choice reaction-time analysis. *Journal of*

Experimental Psychology: Human Perception and Performance, *24*(3), 707-718. doi: 10.1037/0096-1523.24.3.707

Main, J. C., DeBruine, L. M., Little, A. C., & Jones, B. C. (2010). Interactions among the effects of head orientation, emotional expression, and physical attractiveness on face preferences. *Perception, 39*(1), 62-71. doi: 10.1068/p6503

Massaro, D. (1998). *Perceiving talking faces: From speech perception to behavioural principles.* Cambridge, MA: MIT Press.

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry.* Hillsdale, NJ: Erlbaum.

Massaro, D. W., & Egan, P. B. (1996). Perceiving affect from the voice and the face. *Psychonomic Bulletin & Review*, *3*(2), 215-221. doi: 10.3758/BF03212421.

Mathias, S. R., & von Kriegstein, K. (2014). How do we recognise who is speaking? *Frontiers in Bioscience, 6*, 92-109. doi: 10.2741/S417

Mavica, L. W., & Barenholtz, E. (2013). Matching voice and face identity from static images. *Journal of Experimental Psychology: Human Perception and Performance, 39*(2), 307-312. doi: 10.1037/a0030945

McCarthy, G., Puce, A., Gore, J. C., & Allison, T. (1997). Face-specific processing in the human fusiform gyrus. *Journal of Cognitive Neuroscience*, *9*(5), 605-610. doi: 10.1162/jocn.1997.9.5.605

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588)*,* 746–748. doi: 10.1038/264746a0

McQuiston-Surrett, D., Malpass, R. S., & Tredoux, C. G. (2006). Sequential vs. simultaneous lineups: A Review of methods, data, and theory. *Psychology, Public Policy and Law*, *12*(2), 137-419. doi: 10.1037/1076-8971.12.2.137

Megreya, A. M., & Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory & Cognition, 34*(4), 865-876. doi: 10.3758/BF03193433

Megreya, A. M., & Burton, A. M. (2007). Hits and false positives in face matching: A familiarity-based dissociation. *Perception & Psychophysics*, *69*(7), 1175-1184. 10.3758/BF03193954

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, and Law*, *7*(1), 3-35. doi: 10.1037/1076-8971.7.1.3

Meissner, C. A., Tredoux, C. G., Parker, J. F., & MacLin, O. H. (2005). Eyewitness decisions in simultaneous and sequential lineups: A dual-process signal detection theory analysis. *Memory & Cognition*, *33*(5), 783-792. doi: 10.3758/BF03193074

Mickes, L., Flowe, H. D., & Wixted, J. T. (2012). Receiver operating characteristic analysis of eyewitness memory: comparing the diagnostic accuracy of simultaneous versus sequential lineups. *Journal of Experimental Psychology. Applied, 18*(4), 361-376. doi: 10.1037/a0030609

Miller, B. T., & D'Esposito, M. (2005). Searching for “the top” in top-down control. *Neuron*, *48*(4), 535-538. doi: 10.1016/j.neuron.2005.11.002

Miller, G. F. & Todd, P. M. (1998). Mate choice turns cognitive. *Trends in Cognitive Sciences, 2*(5)*,* 190-198. doi: 10.1016/S1364-6613(98)01169-3

Møller, A. P., & Pomiankowski, A. (1993). Why have birds got multiple sexual ornaments? *Behavioral Ecology and Sociobiology*, *32*(3), 167-176. doi: 10.1007/BF00173774

Morrison, E. R., Gralewski, L., Campbell, N., & Penton-Voak, I. S. (2007). Facial movement varies by sex and is related to attractiveness. *Evolution and Human Behavior*, *28*(3), 186-192. doi: 10.1016/j.evolhumbehav.2007.01.001

Mullennix, J. W., & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics*, *47*(4), 379-390. doi: 10.3758/BF03210878

Mullennix, J. W., Ross, A., Smith, C., Kuykendall, K., Conard, J., & Barb, S. (2011). Typicality effects on memory for voice: Implications for earwitness testimony. *Applied Cognitive Psychology*, *25*(1), 29-34. doi: 10.1002/acp.1635

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics*, *58*(3), 351-362. doi: 10.3758/BF03206811

Nakhaeizadeh, S., Dror, I. E., & Morgan, R. M. (2014). Cognitive bias in forensic anthropology: Visual assessment of skeletal remains is susceptible to confirmation bias. *Science & Justice*, *54*(3), 208-214. doi: 10.1016/j.scijus.2013.11.003

Neiman, G. S., & Applegate, J. A. (1990). Accuracy of listener judgments of perceived age relative to chronological age in adults. *Folia Phoniatrica et Logopaedica*, *42*(6), 327-330. doi: 10.1159/000266090

Neisser, U. (1967). *Cognitive psychology*. Englewood Cliffs, N.J.:Prentice-Hall

Nezlek, J. B. (2001). Multilevel random coefficient analyses of event-and interval-contingent data in social and personality psychology research. *Personality and Social Psychology Bulletin*, *27*(7), 771-785. doi: 10.1177/0146167201277001

Nezlek, J. B. (2008). An introduction to multilevel modeling for social and personality psychology. *Social and Personality Psychology Compass*, *2*(2), 842-860. doi: 10.1111/j.1751-9004.2007.00059.x

O'Connor, J. J., Re, D. E., & Feinberg, D. R. (2011). Voice pitch influences perceptions of sexual infidelity. *Evolutionary Psychology*, *9*(1), 64-78. doi: 10.1177/147470491100900109

O'Toole, A. J., Roark, D., & Abdi, H. (2002). Recognizing moving faces: A psychological and neural synthesis. *Trends in Cognitive Science, 6*(6)*,* 261–266. doi: 10.1016/S1364-6613(02)01908-3

Oguchi, T., & Kikuchi, H. (1997). Voice and interpersonal attraction. *Japanese Psychological Research*, *39*(1), 56-61. doi: 10.1111/1468-5884.00037

Orchard, T., & A. Yarmey, A. D. (1995). The effects of whispers, voice sample duration and voice distinctiveness on criminal speaker identification. *Applied Cognitive Psychology,* 9(3)*,* 249-260. doi: 10.1002/acp.2350090306

Ormerod, D. (2001). Sounds familiar? Voice identification evidence. *Criminal Law Review*, 595-622.

Partan, S. R., & Marler, P. (2005). Issues in the classification of multimodal communication signals. *The American Naturalist*, *166*(2), 231-245. doi: 10.1086/431246

Partan, S., & Marler, P. (1999). Communication goes multimodal. *Science*, *283*(5406), 1272-1273. doi: 10.1126/science.283.5406.1272

Pasternak, T., & Greenlee, M. W. (2005). Working memory in primate sensory systems. *Nature Reviews Neuroscience*, *6*(2), 97-107. doi: 10.1038/nrn1603

Paulesu, E., Perani, D., Blasi, V., Silani, G., Borghese, N. A., De Giovanni, U., ... & Fazio, F. (2003). A functional-anatomical model for lipreading. *Journal of Neurophysiology*, *90*(3), 2005-2013. doi: 10.1152/jn.00926.2002

Peirce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics, 2*(10), 1-8. doi: 10.3389/neuro.11.010.2008

Penney, C. G. (1989). Modality effects and the structure of short-term verbal memory. *Memory & Cognition*, *17*(4), 398-422. doi: 10.3758/BF03202613

Penton-Voak, I. S., & Chen, J. Y. (2004). High salivary testosterone is linked to masculine male facial appearance in humans. *Evolution and Human Behavior*, *25*(4), 229-241. doi: 10.1016/j.evolhumbehav.2004.04.003

Penton-Voak, I., & Chang, H. (2008). Attractiveness judgements of individuals vary across emotional expression and movement conditions. *Journal of Evolutionary Psychology*, *6*(2), 89-100. doi: 10.1556/JEP.2008.1011

Perrett, D. I., Lee, K. J., Penton-Voak, I., Rowland, D., Yoshikawa, S., Burt, D. M., Henzi, S. P., Castles, D. L. & Akamatsu, S. (1998). Effects of sexual dimorphism on facial attractiveness. *Nature*, *394*(6696), 884-887. doi: 10.1038/29772

Pike, G. E., Kemp, R. I., Towell, N. A., & Phillips, K. C. (1997). Recognizing moving faces: The relative contribution of motion and perspective view information. *Visual Cognition*, *4*(4), 409-438. doi: 10.1080/713756769

Pisanski, K., Mishra, S., & Rendall, D. (2012). The evolved psychology of voice: Evaluating interrelationships in listeners' assessments of the size, masculinity, and attractiveness of unseen speakers. *Evolution and Human Behavior*, *33*(5), 509-519. doi: 10.1016/j.evolhumbehav.2012.01.004

Pitcher, D., Dilks, D. D., Saxe, R. R., Triantafyllou, C., & Kanwisher, N. (2011). Differential selectivity for dynamic versus static information in face-selective cortical regions. *Neuroimage*, *56*(4), 2356-2363. doi: 10.1016/j.neuroimage.2011.03.067

Pitcher, D., Walsh, V., Yovel, G., & Duchaine, B. (2007). TMS evidence for the involvement of the right occipital face area in early face processing. *Current Biology*, *17*(18), 1568-1573. doi: 10.1016/j.cub.2007.07.063

Polosecki, P., Moeller, S., Schweers, N., Romanski, L. M., Tsao, D. Y., & Freiwald, W. A. (2013). Faces in motion: selectivity of macaque and human face processing areas for dynamic stimuli. *The Journal of Neuroscience*, *33*(29), 11768-11773. doi: 10.1523/JNEUROSCI.5402-11.2013

Ptacek, P. H., & Sander, E. K. (1966). Age recognition from voice. *Journal of Speech & Hearing Research*. *9*(2), 273-277. doi: 10.1044/jshr.0902.273

Puts, D. A., Apicella, C. L., & Cárdenas, R. A. (2012). Masculine voices signal men's threat potential in forager and industrial societies. *Proceedings of the Royal Society of London B: Biological Sciences*, *279*(1728), 601-609. doi: 10.1098/rspb.2011.0829

Puts, D. A., Gaulin, S. J., & Verdolini, K. (2006). Dominance and the evolution of sexual dimorphism in human voice pitch. *Evolution and Human Behavior, 27*(4), 283-296. doi: 10.1016/j.evolhumbehav.2005.11.003

Puts, D. A., Hodges, C. R., Cárdenas, R. A., & Gaulin, S. J. (2007). Men's voices as dominance signals: vocal fundamental and formant frequencies influence dominance attributions among men. *Evolution and Human Behavior*, *28*(5), 340-344. doi: 10.1016/j.evolhumbehav.2007.05.002

Puts, D. A., Jones, B. C., & DeBruine, L. M. (2012). Sexual selection on human faces and voices. *Journal of Sex Research, 49*(2-3), 227-243. doi: 10.1080/00224499.2012.658924

Quené, H., & Van den Bergh, H. (2004). On multi-level modeling of data from repeated measures designs: A tutorial. *Speech Communication*, *43*(1), 103-121. doi: 10.1016/j.specom.2004.02.004

Rammsayer, T., & Ulrich, R. (2012). The greater temporal acuity in the reminder task than in the 2AFC task is independent of standard duration and sensory modality. *Canadian Journal of Experimental Psychology, 66*(1), 26-31. doi: 10.1037/a0025349

Re, D. E., Hunter, D. W., Coetzee, V., Tiddeman, B. P., Xiao, D., DeBruine, L. M., ... & Perrett, D. I. (2013). Looking like a leader–facial shape predicts perceived height and leadership ability. *PloS one*, *8*(12), e80957. doi: 10.1371/journal.pone.0080957

Reed, P. (1992). Effect of a signalled delay between an action and outcome on human judgement of causality. *Quarterly Journal of Experimental Psychology: Section B*, *44*(2), 81-100. doi: 10.1080/02724999208250604

Rezlescu, C., Penton, T., Walsh, V., Tsujimura, H., Scott, S. K., & Banissy, M. J. (2015).

Dominant Voices and Attractive Faces: The Contribution of Visual and Auditory Information to Integrated Person Impressions. *Journal of Nonverbal Behavior*, *39*(4), 355-370. Doi: 10.1007/s10919-015-0214-8

Rhodes, G., Chan, J., Zebrowitz, L. A., & Simmons, L. W. (2003). Does sexual dimorphism in human faces signal health? *Proceedings of the Royal Society of London B: Biological Sciences*, *270*(Suppl 1), S93-S95. doi: 10.1098/rsbl.2003.0023

Rhodes, G., Lie, H. C., Thevaraja, N., Taylor, L., Iredell, N., Curran, C., ... & Simmons, L. W. (2011). Facial attractiveness ratings from video-clips and static images tell the same story. *PloS one*, *6*(11), e26653. doi: 10.1371/journal.pone.0026653

Roberts, S. C., Little, A. C., Lyndon, A., Roberts, J., Havlicek, J. & Wright, R. L. (2009a). Manipulation of body odour alters men's self-confidence and judgements of their visual attractiveness by women. *International Journal of Cosmetic Science, 31*(1), 47-54. doi: 10.1111/j.1468-2494.2008.00477.x

Roberts, S. C., Saxton, T. K., Murray, A. K., Burriss, R. P., Rowland, H. M., & Little, A. C. (2009b). Static and dynamic facial images cue similar attractiveness judgements. *Ethology*, *115*(6), 588-595. doi: 10.1556/JEP.7.2009.1.4

Robertson, D. M., & Schweinberger, S. R. (2010). The role of audiovisual asynchrony in person recognition. *The Quarterly Journal of Experimental Psychology*, *63*(1), 23-30. doi: 10.1080/17470210903144376

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15*(3), 351–357. doi: 10.2307/2087176

Rosenblum, L. D. (2005). The primacy of multimodal speech perception. In D. Pisoni & R. Remez (Eds.), *Handbook of speech perception* (pp. 51-78). Malden, MA: Blackwell.

Rosenblum, L. D. (2008). Speech perception as a multimodal phenomenon. *Current Directions in Psychological Science*, *17*(6), 405-409. doi: 10.1111/j.1467-8721.2008.00615.x

Rosenblum, L. D., Miller, R. M., & Sanchez, K. (2007). Lip-read me now, hear me better later: Cross-modal transfer of talker-familiarity effects. *Psychological Science*, *18*(5), 392-396. doi: 10.1111/j.1467-9280.2007.01911.x

Rosenblum, L. D., Smith, N. M., Nichols, S. M., Hale, S., & Lee, J. (2006). Hearing a face: Cross-modal speaker matching using isolated visible speech. *Perception & Psychophysics*, *68*(1), 84-93. doi: 10.3758/BF03193658

Rosenblum, L.D. (2002). The perceptual basis for audiovisual speech integration. In J.H.L. Hansen and B. Pellom (Eds.), *International Conference on Spoken Language Processing* (pp. 1461–1464). Adelaide, Australia: Causal Productions Pty Ltd..

Rubenstein, A. J. (2005). Variation in perceived attractiveness: Differences between dynamic and static faces. *Psychological Science, 16*(10), 759-762. doi: 10.1111/j.1467-9280.2005.01610.x

Sanchez, K., Dias, J. W., & Rosenblum, L. D. (2013). Experience with a talker can transfer across modalities to facilitate lipreading. *Attention, Perception, & Psychophysics*, *75*(7), 1359-1365. doi: 10.3758/s13414-013-0534-x

Saxton, T. K., Caryl, P. G., & Roberts, C. S. (2006). Vocal and facial attractiveness judgments of children, adolescents and adults: The ontogeny of mate choice. *Ethology*, *112*(12), 1179-1185. doi: 10.1111/j.1439-0310.2006.01278.x

Schneider, T. M., Hecht, H., Stevanov, J., & Carbon, C. C. (2013). Cross-ethnic assessment of body weight and height on the basis of faces. *Personality and Individual Differences, 55*(4), 356-360. doi: 10.1016/j.paid.2013.03.022

Schweinberger, S. R., Herholz, A., & Stief, V. (1997). Auditory long term memory: Repetition priming of voice recognition. *The Quarterly Journal of Experimental Psychology*, *50A*(3), 498-517. doi: 10.1080/713755724

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. doi: 10.1002/wcs.1261

Schweinberger, S. R., Robertson, D., & Kaufmann, J. M. (2007). Hearing facial identities. *The Quarterly Journal of Experimental Psychology*, *60*(10), 1446-1456. doi: 10.1080/17470210601063589

Scott, I. M., & Penton-Voak, I. S. (2011). The validity of composite photographs for assessing masculinity preferences. *Perception, 40*(3), 323-331. doi: 10.1068/p6723

Shanks, D. R., Pearson, S. M., & Dickinson, A. (1989). Temporal contiguity and the judgement of causality by human subjects. *The Quarterly Journal of Experimental Psychology*, *41*(2), 139-159. doi: 10.1080/14640748908401189

Sheffert, S. M., & Olson, E. (2004). Audiovisual speech facilitates voice learning. *Perception & Psychophysics*, *66*(2), 352-362. doi: 10.3758/BF03194884

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11). 1359-1366. doi: 10.1177/0956797611417632

Skipper, J. I., van Wassenhove, V., Nusbaum, H. C., & Small, S. L. (2007). Hearing lips and seeing voices: How cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex*, *17*(10), 2387-2399. doi: 10.1093/cercor/bhl147

Sligte, I. G., Scholte, H. S., & Lamme, V. A. (2008). Are there multiple visual short-term memory stores? *PLOS one*, *3*(2), e1699. doi: 10.1371/journal.pone.0001699

Sligte, I. G., Scholte, H. S., & Lamme, V. A. (2009). V4 activity predicts the strength of visual short-term memory representations. *The Journal of Neuroscience*, *29*(23), 7432-7438. doi: 10.1523/JNEUROSCI.0784-09.2009

Smith, H. M. J. & Baguley, T. (2014). Unfamiliar voice identification: Effect of post-event information on accuracy and voice ratings. *Journal of European Psychology Students, 5*(1), 59-68. doi: 10.5334/jeps.bs

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied*, *74*(11), 1-29. doi: 10.1037/h0093759

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers, 31*(1), 137-149. doi: 10.3758/BF03207704

Steblay, N. K., Dysart, J. E., & Wells, G. L. (2011). Seventy-two tests of the sequential lineup superiority effect: A meta-analysis and policy discussion. *Psychology, Public Policy, and Law*, *17*(1), 99-139. doi: 10.1037/a0021650

Steblay, N., Dysart, J., Fulero, S., & Lindsay, R. C. (2001). Eyewitness accuracy rates in sequential and simultaneous lineup presentations: a meta-analytic comparison. *Law and Human Behavior*, *25*(5), 459-473. doi: 10.1023/A:1012888715007

Stevenage, S. V., & Neil, G. J. (2014). Hearing faces and seeing voices: The integration and interaction of face and voice processing. *Psychologica Belgica*, *54*(3), 266-281. doi: 10.5334/pb.ar

Stevenage, S. V., Hale, S., Morgan, Y., & Neil, G. J. (2014a). Recognition by association: Within- and cross- modality associative priming with faces and voices. *British Journal of Psychology*, *105*(1), 1-16. doi: 10.1111/bjop.12011

Stevenage, S. V., Hamlin, I., & Ford, R. (2015). *You look exactly as I had imagined: Matching faces and voices across line-up and matching trials.* Paper presented at Annual BPS Cognitive Section Conference, University of Kent.

Stevenage, S. V., Howland, A., & Tippelt, A. (2011). Interference in eyewitness and earwitness recognition. *Applied Cognitive Psychology*, *25*(1), 112-118. 10.1002/acp.1649

Stevenage, S. V., Hugill, A. R., & Lewis, H. G. (2012). Integrating voice recognition into models of person perception. *Journal of Cognitive Psychology, 24*(4), 409-419. doi: 10.1080/20445911.2011.642859

Stevenage, S. V., Neil, G. J., & Hamlin, I. (2014b). When the face fits: Recognition of celebrities from matching and mismatching faces and voices. *Memory, 22*(3), 284-294. doi: 10.1080/09658211.2013.781654

Stevenage, S. V., Neil, G. J., Barlow, J., Dyson, A., Eaton-Brown, C., & Parsons, B. (2013). The effect of distraction on face and voice recognition. *Psychological Research*, *77*(2), 167-175. doi: 10.1007/s00426-012-0450-z

Stevenage, S. V., Walpole, C., Neil, G. J., & Black, S. M. (2014c). Testing the reliability of hands and ears as biometrics: The importance of viewpoint. *Psychological Research*, *79*(6), 989-999. doi: 10.1007/s00426-014-0625-x

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, *26*(2), 212-215. doi: 10.1121/1.1907309

Summerfield, A. Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception, in R. Campbell and B. Dodd (Eds.), *Hearing by Eye: The Psychology of Lip-Reading* (pp. 3–51), London, U.K.: Lawrence Erlbaum

Tanaka, J. W. (2001). The entry point of face recognition: Evidence for face expertise. *Journal of Experimental Psychology: General*, *130*(3), 534-543. doi: 10.1037/0096-3445.130.3.534

Thornhill, R., & Gangestad, S. W. (1999). Facial attractiveness. *Trends in Cognitive Sciences*, *3*(12), 452-460. doi: 10.1016/S1364-6613(99)01403-5

Thornhill, R., & Gangestad, S. W. (2006). Facial sexual dimorphism, developmental stability, and susceptibility to disease in men and women. *Evolution and Human Behavior*, *27*(2), 131-144. doi: 10.1016/j.evolhumbehav.2005.06.001

Thornhill, R., & Grammer, K. (1999). The body and face of woman: One ornament that signals quality? *Evolution and Human Behavior*, *20*(2), 105-120. doi: 10.1016/S1090-5138(98)00044-0

Thurstone, L. L. (1927a). The method of paired comparisons for social values. *The Journal of Abnormal and Social Psychology, 21*(4), 384-400. doi: 10.1037/h0065439

Thurstone, L. L. (1927b). Psychophysical Analysis. *The American Journal of Psychology,* *38*(3), 368-389. doi: 10.2307/1415006

Titze, I. R. (1994a). *Principles of voice production.* Englewood Cliffs, NJ: Prentice Hall

Titze, I. R. (1994b). Toward standards in acoustic analysis of voice. *Journal of Voice*, *8*(1), 1-7. doi: 10.1016/S0892-1997(05)80313-3

Todd, J. J., & Marois, R. (2005). Posterior parietal cortex activity predicts individual differences in visual short-term memory capacity. *Cognitive, Affective, & Behavioral Neuroscience*, *5*(2), 144-155. doi: 10.3758/CABN.5.2.144

Treisman, A. (1964). Monitoring and storage of irrelevant messages in selective attention. *Journal of Verbal Learning and Verbal Behavior*, *3*(6), 449-459. doi: 10.1016/S0022-5371(64)80015-3

Uetz, G. W., & Roberts, J. A. (2002). Multisensory cues and multimodal communication in spiders: insights from video/audio playback studies. *Brain, Behavior and Evolution*, *59*(4), 222-230. doi: 10.1159/000064909

Uetz, G. W., Roberts, J. A., & Taylor, P. W. (2009). Multimodal communication and mate choice in wolf spiders: Female response to multimodal versus unimodal signals. *Animal Behaviour*, *78*(2), 299-305. doi: 10.1016/j.anbehav.2009.04.023

Ulrich, R., & Vorberg, D. (2009). Estimating the difference limen in 2AFC tasks: Pitfalls and improved estimators. *Attention, Perception, & Psychophysics, 71*(6), 1219-1227. doi: 10.3758/APP.71.6.1219

Urbaniak, G. C., & Plous, S. (2013). Research Randomizer (Version 4.0) [Computer software]. Available from http://www.randomizer.org/

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

Valentine, T., Lewis, M. B., & Hills, P. J. (2015). Face-space: A unifying concept in face recognition research. *The Quarterly Journal of Experimental Psychology*, (ahead-of-print), 1-24. doi: 10.1080/17470218.2014.990392

van Bezooijen, R. (1995). Sociocultural aspects of pitch differences between Japanese and Dutch women. *Language and Speech*, *38*(3), 253-265. doi: 10.1177/002383099503800303

van Dommelen, W. A., & Moxness, B. H. (1995). Acoustic parameters in speaker height and weight identification: sex-specific behaviour. *Language and Speech*, *38*(3), 267-287. doi: 10.1177/002383099503800304

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.

Van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*, *45*(3), 598-607. doi: 10.1016/j.neuropsychologia.2006.01.001

Vatikiotis-Bateson, E., Munhall, K. G., Hirayama, M., Lee, Y. V., & Terzopoulos, D. (1996). The dynamics of audiovisual behavior in speech. In D. Stork and M. Hennecke (Eds.), *Speechreading by humans and machines* (pp. 221-232). Berlin: Springer-Verlag

Vogel, E. K., & Machizawa, M. G. (2004). Neural activity predicts individual differences in visual working memory capacity. *Nature*, *428*(6984), 748-751. doi: 10.1038/nature02447

Vokey, J. R., & Read, J. D. (1992). Familiarity, memorability, and the effect of typicality on the recognition of faces. *Memory & Cognition*, *20*(3), 291-302. doi: 10.3758/BF03199666

von Kriegstein, K., Eger, E., Kleinschmidt, A., & Giraud, A. L. (2003). Modulation of neural responses to speech by directing attention to voices or verbal content. *Cognitive Brain Research*, *17*(1), 48-55. doi: 10.1016/S0926-6410(03)00079-X

von Kriegstein, K., Kleinschmidt, A., Sterzer, P., & Giraud, A. L. (2005). Interaction of face and voice areas during speaker recognition. *Journal of Cognitive Neuroscience*, *17*(3), 367-376. doi: 10.1162/0898929053279577

Warner, R. M., & Sugarman, D. B. (1986). Attributions of personality based on physical appearance, speech, and handwriting. *Journal of Personality and Social Psychology*, *50*(4), 792-799. doi: 10.1037/0022-3514.50.4.792

Wells, G. L. (1984). The Psychology of lineup identifications. *Journal of Applied Social Psychology*, *14*(2), 89-103. doi: 10.1111/j.1559-1816.1984.tb02223.x

Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin*, *25*(9), 1115-1125. doi: 10.1177/01461672992512005

Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human Behavior*, *22*(6), 603-647. doi: 10.1023/A:1025750605807

Wells, G. L., Steblay, N. K., & Dysart, J. E. (2012). Eyewitness Identification Reforms: Are Suggestiveness-Induced Hits and Guesses True Hits. *Perspectives on Psychological Science*, *7*(3), 264-271. doi: 10.1177/1745691612443368

Wells, T. J., Dunn, A. K., Sergeant, M. J., & Davies, M. N. (2009). Multiple signals in human mate selection: A review and framework for integrating facial and vocal signals. *Journal of Evolutionary Psychology, 7*(2), 111-139. doi: 10.1556/JEP.7.2009.2.2

Wells, T., Baguley, T., Sergeant, M., & Dunn, A. (2013). Perceptions of human attractiveness comprising face and voice cues. *Archives of Sexual Behavior, 42*(5), 805-811. doi: 10.1007/s10508-012-0054-0

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General, 143*(5), 2020-2045. doi: 10.1037/xge0000014

Wheatley, J. R., Apicella, C. A., Burriss, R. P., Cárdenas, R. A., Bailey, D. H., Welling, L. L., & Puts, D. A. (2014). Women's faces and voices are cues to reproductive potential in industrial and forager societies. *Evolution and Human Behavior*, *35*(4), 264-271. doi: 10.1016/j.evolhumbehav.2014.02.006

Wheatley, J. R., Apicella, C. A., Burriss, R. P., Cárdenas, R. A., Bailey, D. H., Welling, L. L., & Puts, D. A. (2014). Women's faces and voices are cues to reproductive potential in industrial and forager societies. *Evolution and Human Behavior*, *35*(4), 264-271. doi: 10.1016/j.evolhumbehav.2014.02.006

Wickelgren, W. A. (1969). Auditory or articulatory coding in verbal short-term memory. *Psychological Review, 76*(2), 232-235. doi: 10.1037/h0027397

Wickens, T. (2001). *Elementary Signal Detection Theory.* Oxford: Oxford University Press

Willis, J., & Todorov, A. (2006). First impressions making up your mind after a 100-ms exposure to a face. *Psychological Science*, *17*(7), 592-598. doi: 10.1111/j.1467-9280.2006.01750.x

Wissmath, B., Weibel, D., & Groner, R. (2009). Dubbing or subtitling? Effects on spatial presence, transportation, flow, and enjoyment. *Journal of Media Psychology, 21*, 114–125. doi: 10.1027/ 1864-1105.21.3.114

Wright, D. B., & London, K. (2009). Multilevel modelling: Beyond the basic applications. *British Journal of Mathematical and Statistical Psychology*, *62*(2), 439-456. doi: 10.1348/000711008X327632

Wright, D. B., Horry, R., & Skagerberg, E. M. (2009). Functions for traditional and multilevel approaches to signal detection theory. *Behavior Research Methods*, *41*(2), 257-267. doi: 10.3758/BRM.41.2.257

Yarmey, A. D. (2007). The psychology of speaker identification and earwitness memory. In R. C. L. Lindsay, D. F. Ross, J. D. Read & M. P. Toglia (Eds.), *The handbook of eyewitness psychology, vol II: Memory for people* (pp. 101-136). Mahwah, NJ: Lawrence Erlbaum Associates.

Yarmey, A. D., & Matthys, E. (1992). Voice identification of an abductor. *Applied Cognitive Psychology*, *6*(5), 367-377. doi: 10.1002/acp.2350060502

Yehia, H. C., Kuratate, T., & Vatikiotis-Bateson, E. (2002). Linking facial animation, head motion and speech acoustics. *Journal of Phonetics*, *30*(3), 555-568. doi: 10.1006/jpho.2002.0165

Yehia, H., Kuratate, T., & Vatikiotis-Bateson, E. (2000). Facial animation and head motion driven by speech acoustics. In *5th Seminar on Speech Production: Models and Data* (pp. 265-268). Kloster Seeon, Germany.

Yehia, H., Rubin, P., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication*, *26*(1), 23-43. doi: 10.1016/S0167-6393(98)00048-X

Yeshurun, Y., Carrasco, M., & Maloney, L. T. (2008). Bias and sensitivity in two-interval forced choice procedures: Tests of the difference model. *Vision Research*, *48*(17), 1837-1851. doi: 10.1016/j.visres.2008.05.008

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., & Kanwisher, N. (2004). Face perception: Domain specific, not process specific. *Neuron*, *44*(5), 889-898. doi: 10.1016/j.neuron.2004.11.018

Zahavi, A., & Zahavi, A. (1997). *The Handicap Principle.* Oxford: Oxford University Press.

Zann, R. A. (1996). *The zebra finch: A synthesis of field and laboratory studies* (Vol. 5). Oxford: Oxford University Press.

Zäske, R., Schweinberger, S. R., & Kawahara, H. (2010). Voice aftereffects of adaptation to speaker identity. *Hearing Research*, *268*(1), 38-45. doi: 10.1016/j.heares.2010.04.011

APPENDIX A: COMPARISON OF MULTILEVEL MODELLING AND TRADITIONAL ANOVA ANALYSES

Chapter 3 argued that appropriate analysis of face and voice data is crucial. Unlike conventional analyses, which tend to aggregate over stimuli, multilevel modelling takes into account the potentially huge amount of variability associated with both faces (Burton, 2013) and voices (Stevenage & Neil, 2014). The results of traditional analyses compared to multilevel modelling can have a significant impact on resulting conclusions (Quené & Van den Bergh, 2004). In order clearly illustrate the necessity of using multilevel modelling to investigate face-voice matching, Appendix A compares the analysis of data from Experiment 2a (as reported in section 4.2.2) to an analysis of the same data using traditional ANOVA.

In Experiment 2a, participants completed a standard 2AFC crossmodal matching task (Lachs, 1999) in which all stimuli were presented sequentially (see Figure 4.1).

A.1 Experiment 2a: Analysis using multilevel models

By way of a recap, the multilevel modelling analysis showed a main effect of position, as well as 3-way interaction between position, order and facial stimulus type (see Table 4.1). Overall matching accuracy was significantly above chance level. However, when data was broken down into the two facial stimulus type conditions, dynamic face-voice matching was above chance level, but static face-voice matching was not (see Figure 4.2).

A.2 Experiment 2a: Analysis by traditional ANOVA

A conventional analysis of these data would involve aggregating over stimuli, and running a mixed 2 x 2 x 2 ANOVA. The results of this analysis are presented in Table A.1.

Table A.1

F ratios for the mixed factorial ANOVA

Source	*F*(1,80)	*p*	η_p^2
Position	22.237	<.001	.218
Order	2.362	.128	.029
Facial stimulus type	2.079	.153	.025
Position x Order	0.930	.338	.011
Position x Facial stimulus type	0.047	.828	.001
Order x Facial Stimulus type	0.648	.423	.008
Position x Order x Facial stimulus type	3.477	.066	.042

This analysis shows that the main effect of position was significant. There were no other main effects and no interactions. Figure A.1 shows mean matching performance (with 95% CI error bars) in each condition calculated using ANOVA.

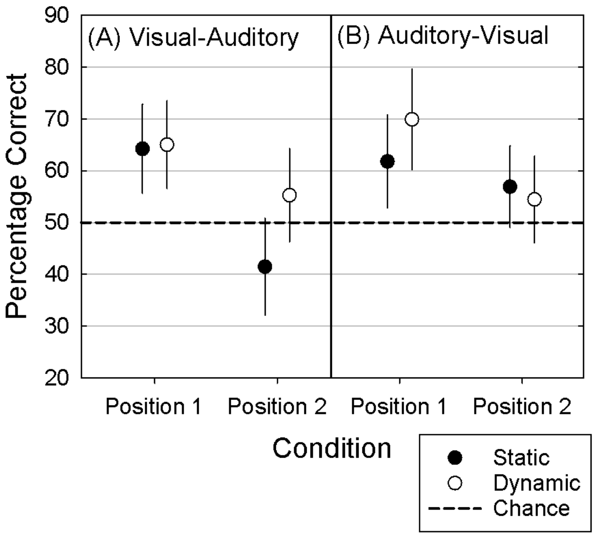


Figure A.1: *Re-analysis of Experiment 2a data using traditional ANOVA: Face-voice matching accuracy on V-A (panel A) and A-V (panel B) trials for sequentially presented faces and voices in a 2AFC task. Error bars show 95% CI for the condition means*

Performance was significantly above chance (50%) level for both static, *M* = 56.10%, 95% CI [51.58%, 60.62%], and dynamic, *M* = 61.18%, 95% CI [55.67%, 66.68%] conditions.

A.3 Comparison of resulting conclusions from both analyses

As previous studies had variously found static face-voice matching to be either at chance level (Lachs & Pisoni, 2004a; Kamachi et al., 2003) or above chance level (Krauss et al., 2002; Mavica & Barenholtz, 2013), the main aim of Experiment 2a was to make a preliminary attempt to resolve these contradictions, testing whether static face-voice matching was significantly above chance using a standard crossmodal matching procedure (Lachs, 1999).

Conclusions based on the overall pattern of main effects and interactions in each analysis were not markedly different. Both analyses detected the main effect of position, with higher levels of accuracy when the same identity stimulus appeared in position 1 compared to position 2. The multilevel analysis detected a three-way interaction between position order and facial stimulus type, but the ANOVA did not. However, this non-predicted interaction (which was only just significant) did not affect the general conclusion.

As noted in Chapter 5, the multilevel modelling analysis showed the variance associated with stimuli to be much higher than the variance associated with participants. The variance associated with stimuli is not accounted for in the ANOVA, resulting in the condition means being slightly different in the two analyses, and the 95% CIs being wider in the multilevel modelling analysis. The most notable difference between the two sets of results is clear when comparing Figures 5.2 and A.1. Whilst static face voice matching is above chance when data is analysed using ANOVA, it is at chance level when analysed using multilevel modelling. This difference is crucial to the conclusions reached in Chapter 5, and to explaining the previous contradictory sets of results in the literature.

In Chapter 5, above-chance static face-voice matching was found to be procedurally dependent. This was notable because previous studies have employed a number of different procedures. A further explanation for contradictions was that some people look and sound more similar than others. Traditional ANOVA analyses would not have detected evidence of either conclusion in the data.

APPENDIX B: STIMULI: IMAGES, RATINGS AND TRANSCRIPTS

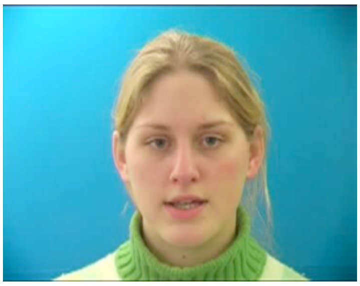
The stimulus set used throughout this thesis is described in Chapter 3 (see section 3.4.1) In the following pages, further details are provided about each of the stimuli. For each stimulus person (1-18), the following information is included:

- The static facial image
- The transcript for the voice recording
- The transcript of the muted sentence articulated by the dynamic face

In Experiment 1, participants rated the stimuli on scales for masculinity/femininity, health height, and weight. They also estimated the age of the person in years. Facial stimulus type was manipulated between subjects, so participants either rated static faces and voices, or dynamic faces and voices. Both sets of results are illustrated for each stimulus person (1-18).

The figures depict the mean rating on each of the scales. As indicated in the legend, face ratings are shown by a filled black dot, and voice ratings are shown by a white dot. The error bars show 95% CIs for the mean ratings.

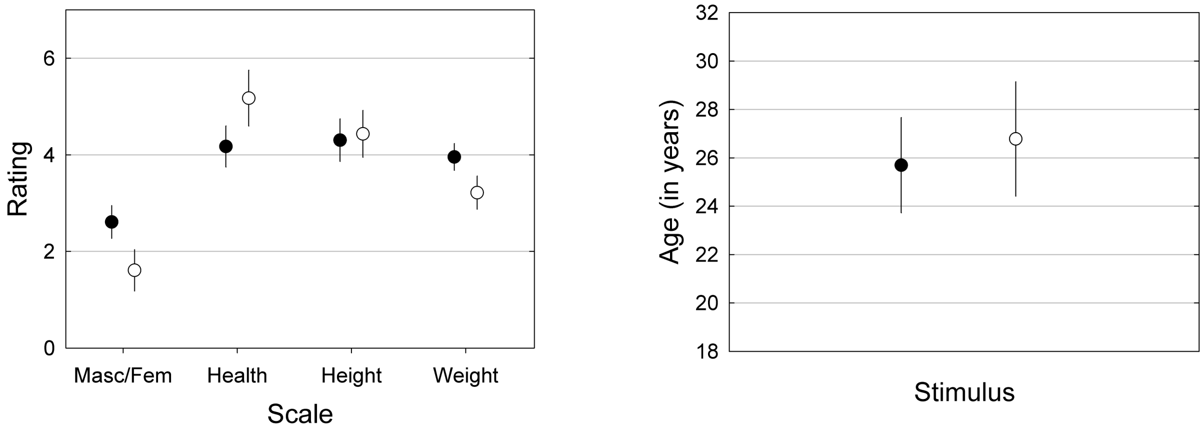
Person 1



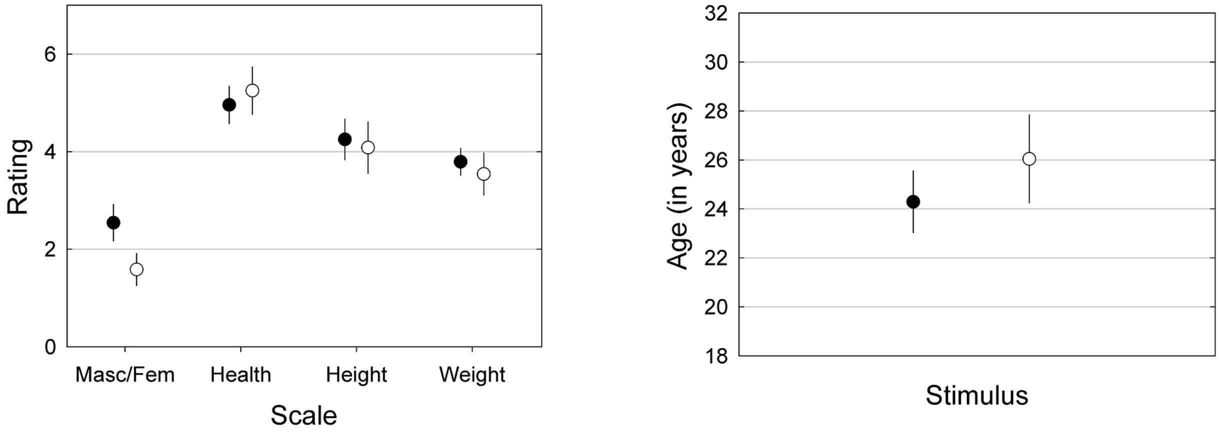
Voice: “set green at L 8 soon”

Dynamic face (muted): “set blue at K 2 again”

Mean ratings of voice and static face



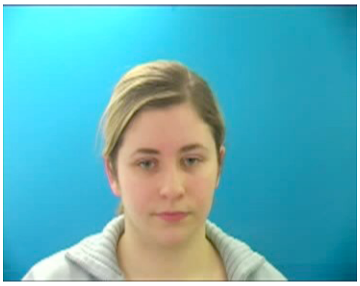
Mean ratings of voice and dynamic face



Legend

● Face
○ Voice

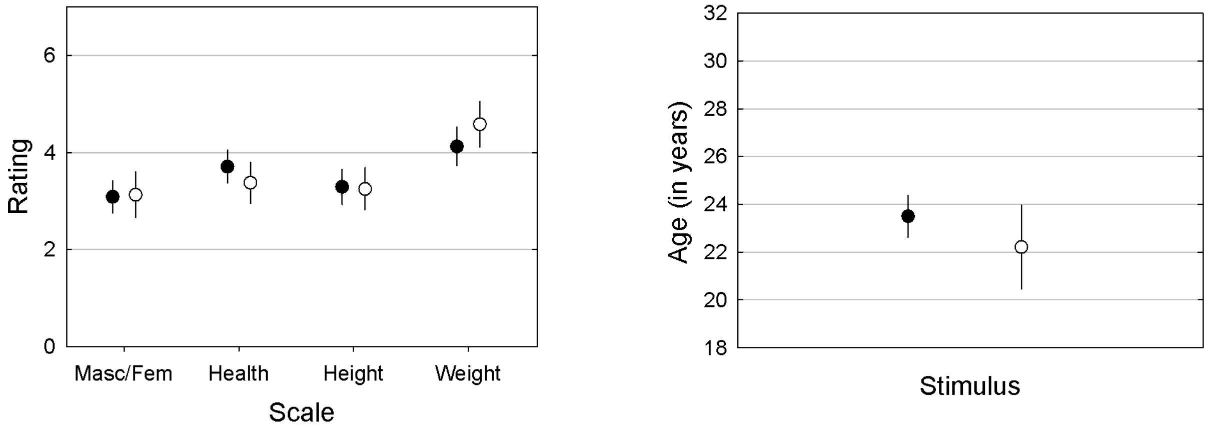
Person 2



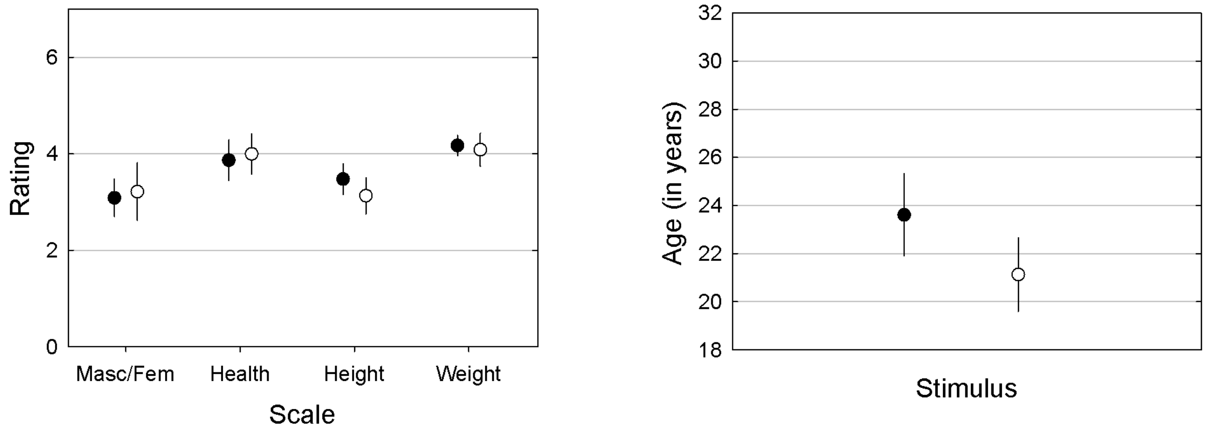
Voice: “place red by V 4 please”

Dynamic face (muted): “set green by B 9 again”

Mean ratings of voice and static face



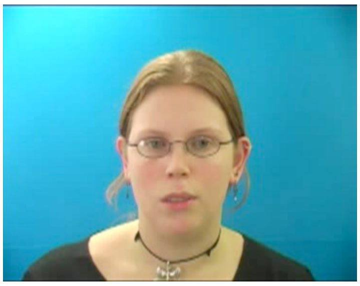
Mean ratings of voice and dynamic face



Legend

● Face
○ Voice

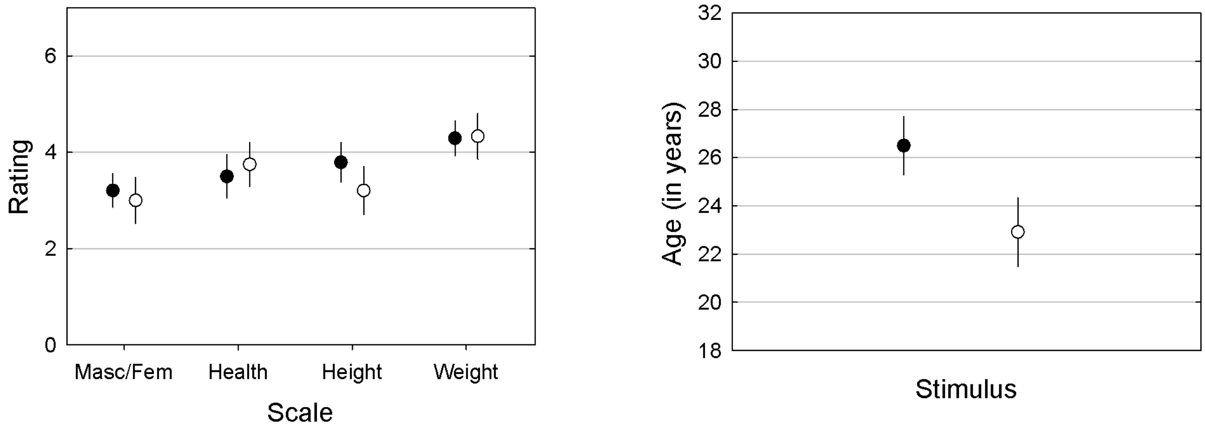
Person 3



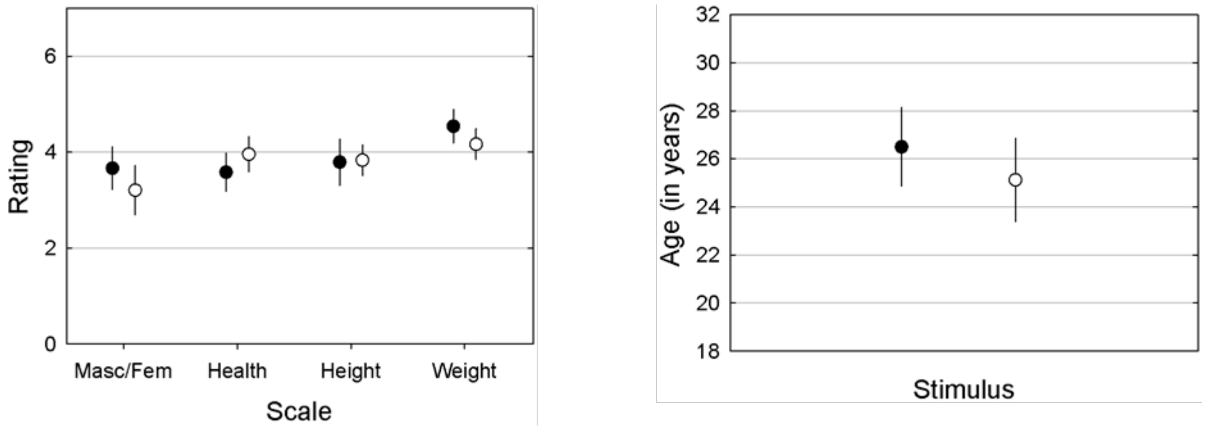
Voice: “set red by T 2 soon”

Dynamic face (muted): “bin blue at D 8 soon”

Mean ratings of voice and static face



Mean ratings of voice and dynamic face



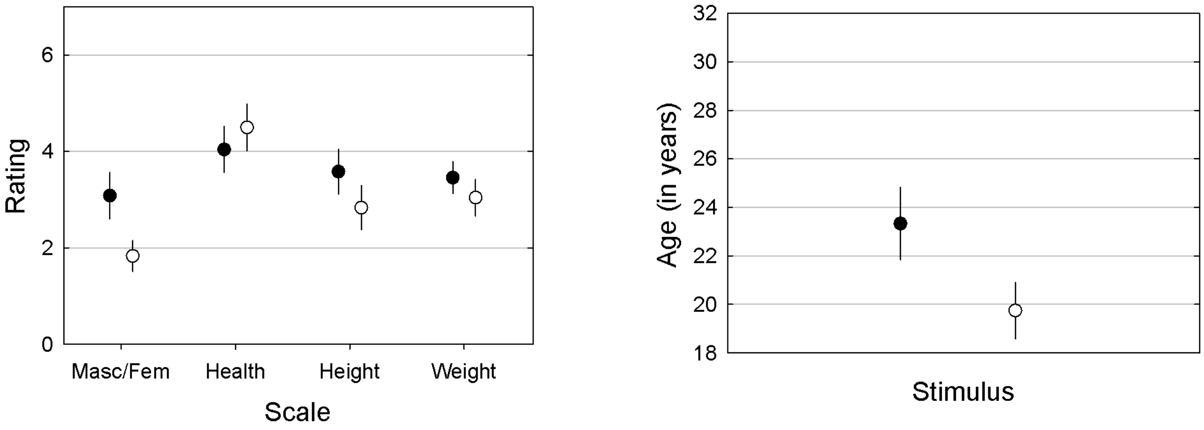
Legend

● Face
○ Voice

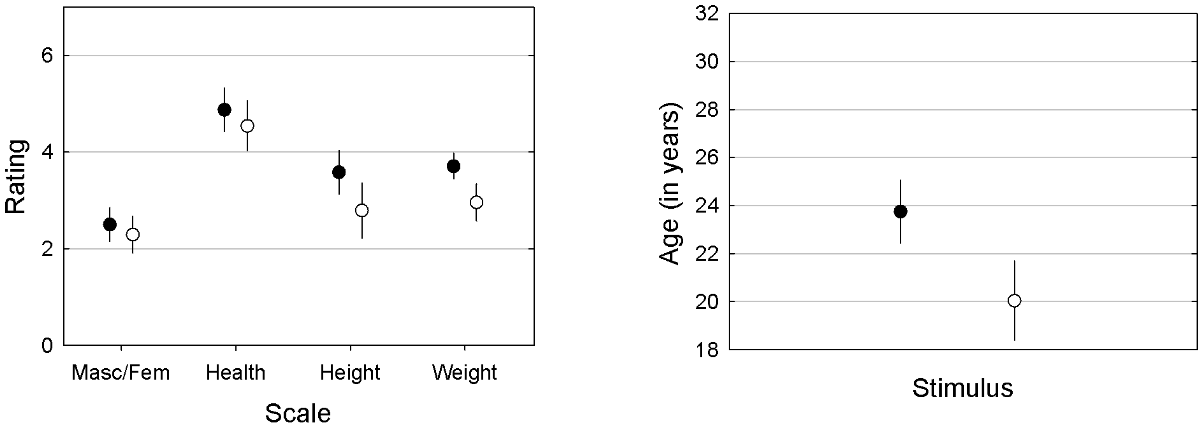
Person 4

Voice:	"lay red by J 4 soon"
Dynamic face (muted):	"set white at T 4 again"

Mean ratings of voice and static face



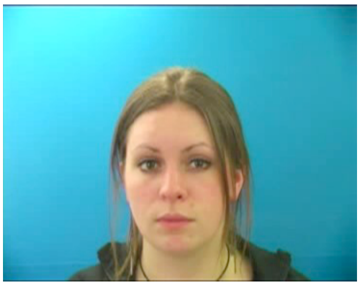
Mean ratings of voice and dynamic face



Legend

● Face
○ Voice

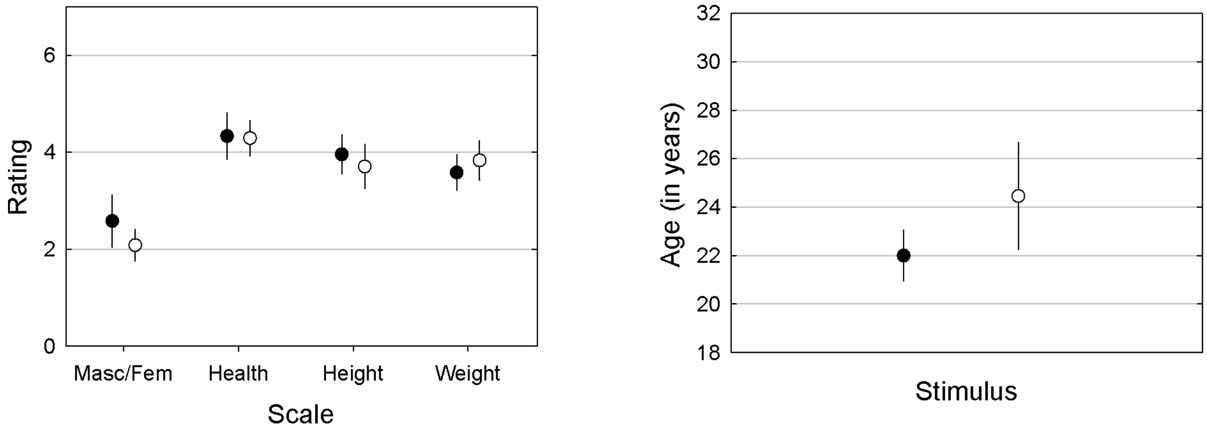
Person 5



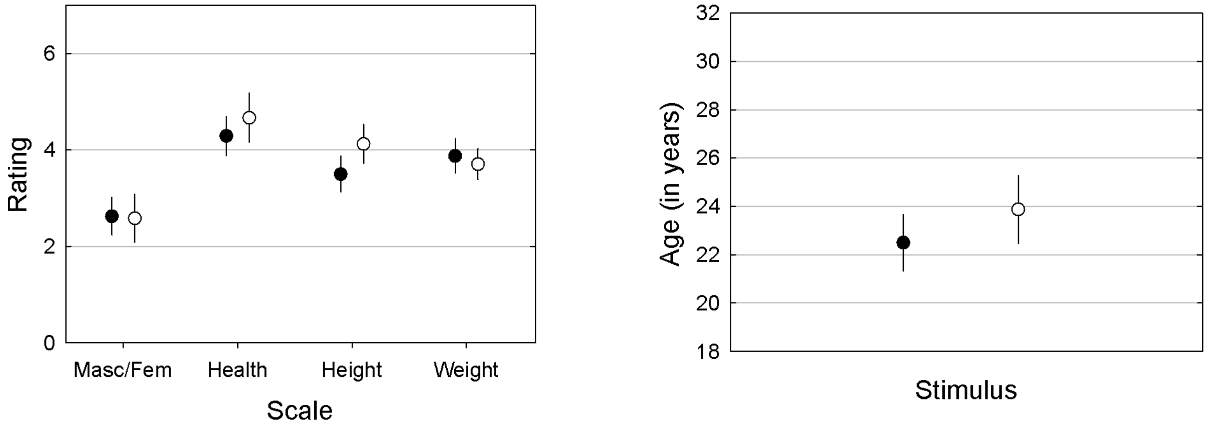
Voice: “place green by P 3 please”

Dynamic face (muted): “set white in M 1 now”

Mean ratings of voice and static face



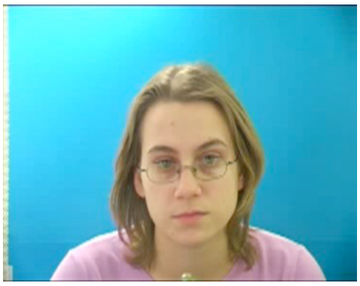
Mean ratings of voice and dynamic face



Legend

● Face
○ Voice

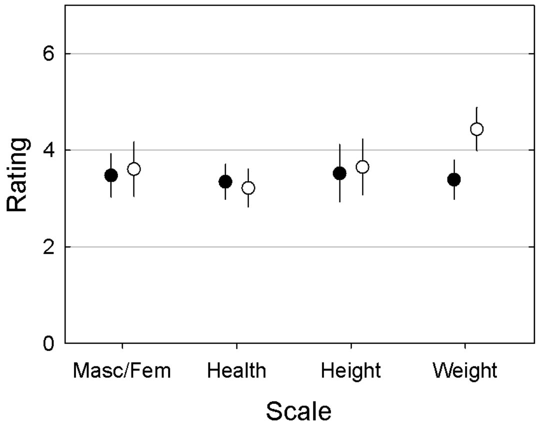
Person 6

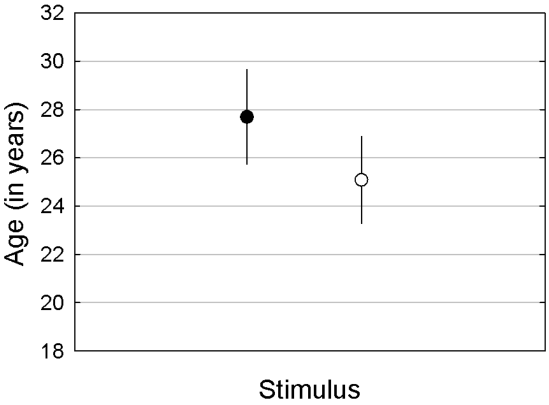


Voice: "bin white with G 8 please"

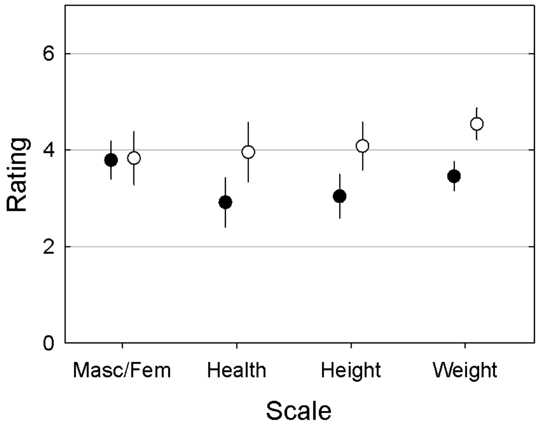
Dynamic face (muted): "bin blue by F 1 soon"

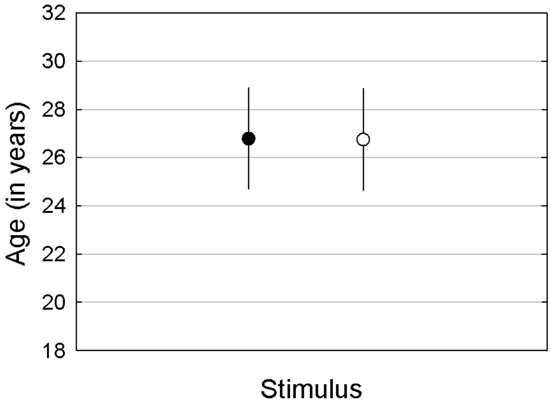
Mean ratings of voice and static face





Mean ratings of voice and dynamic face

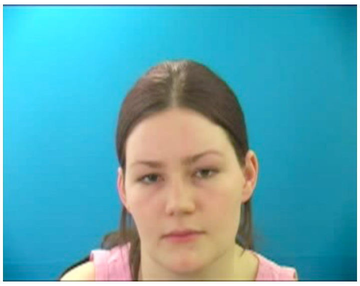




Legend

●	Face
○	Voice

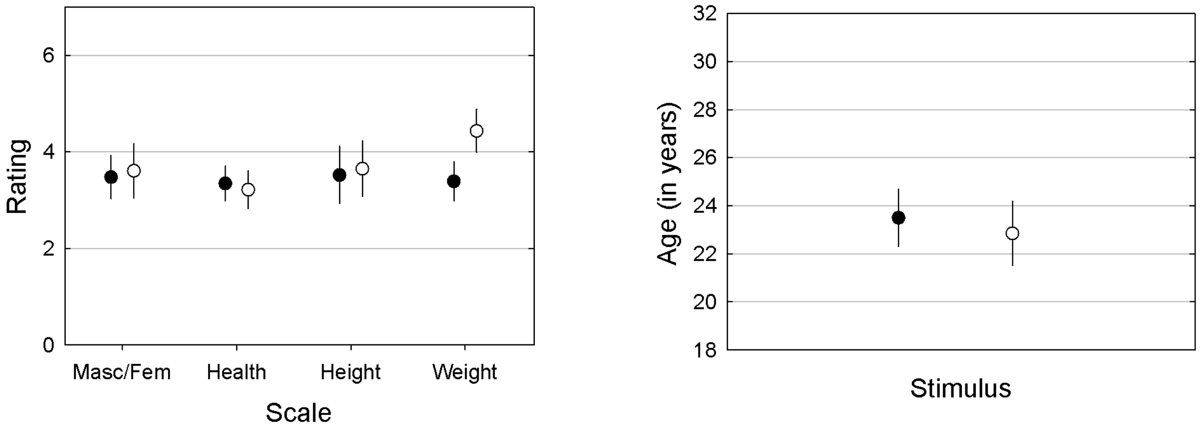
Person 7



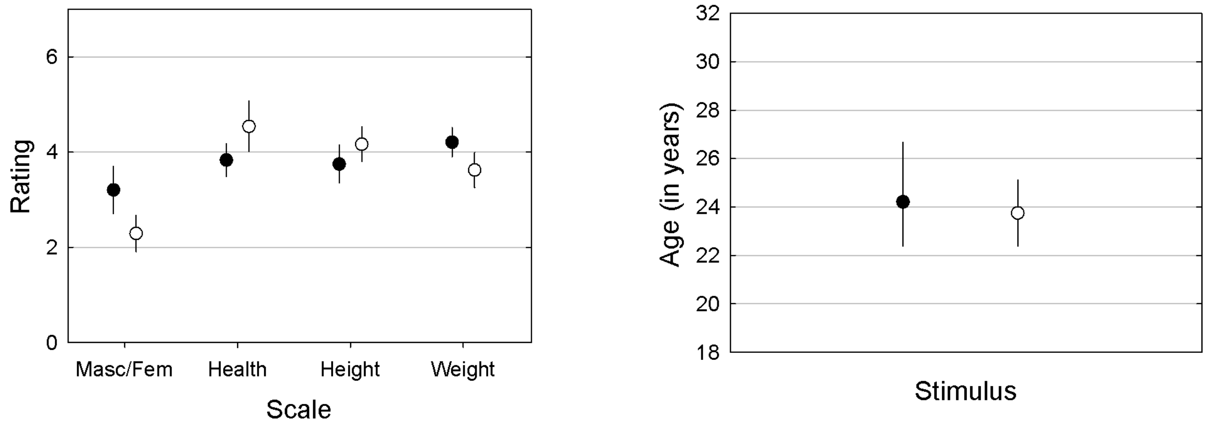
Voice: “lay white by D 3 please”

Dynamic face (muted): “set white by G 4 soon”

Mean ratings of voice and static face



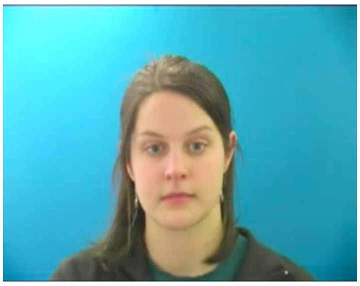
Mean ratings of voice and dynamic face



Legend

● Face
○ Voice

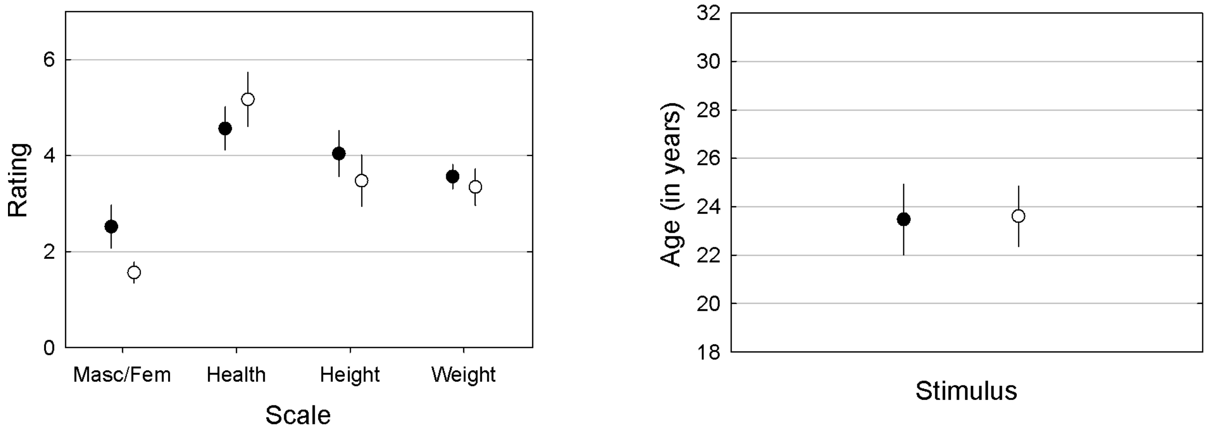
Person 8



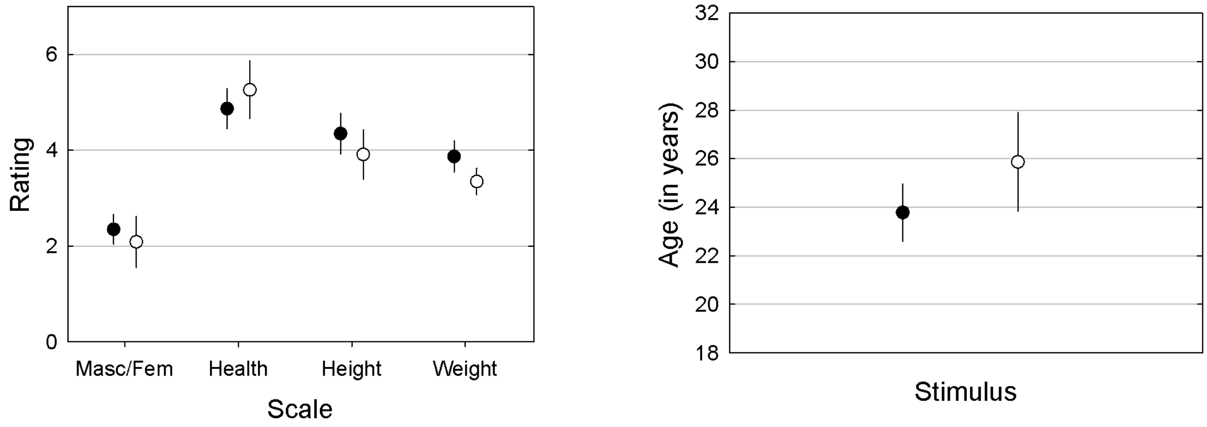
Voice: “bin blue with Q 6 now”

Dynamic face (muted): “bin white at Y 4 now”

Mean ratings of voice and static face



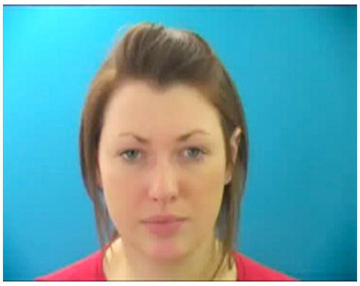
Mean ratings of voice and dynamic face



Legend

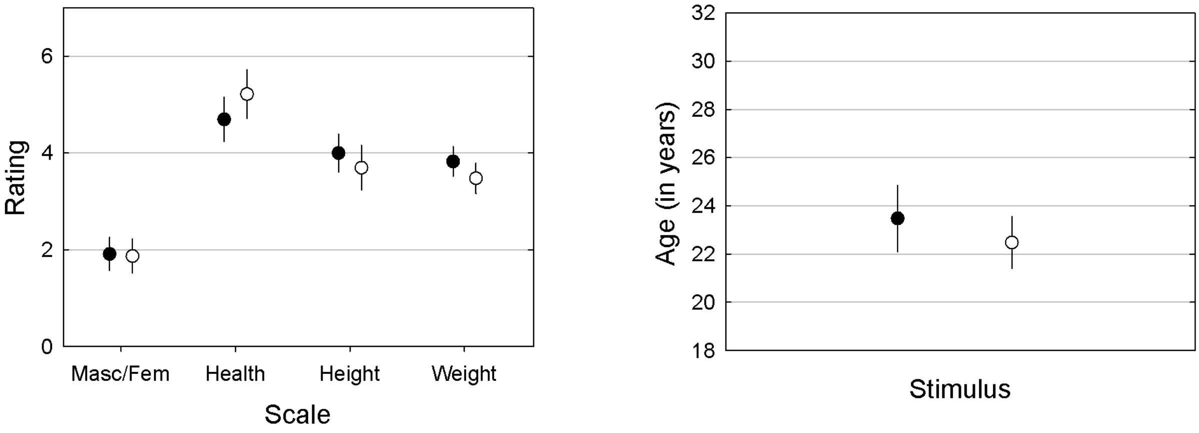
● Face
○ Voice

Person 9

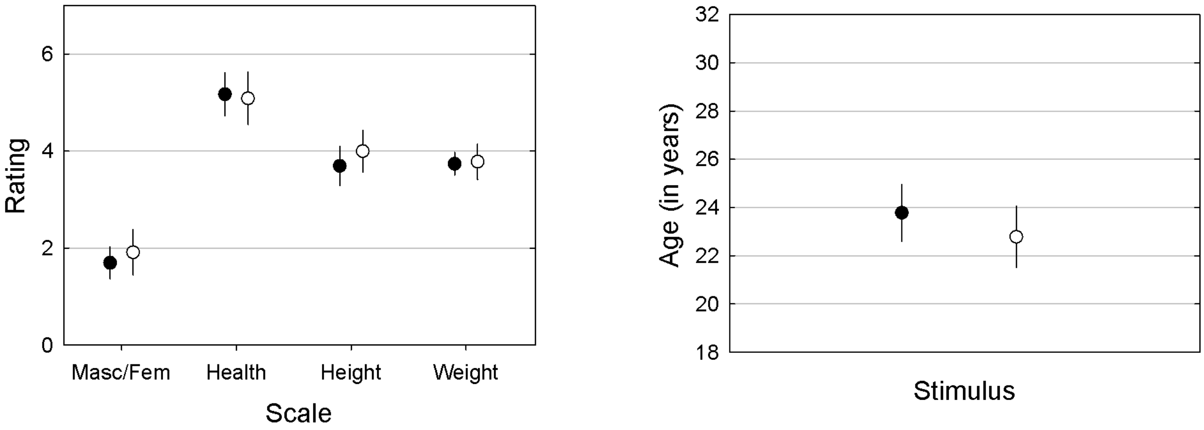


Voice:	“place white at U 2 please”
Dynamic face (muted):	“set blue by F 1 again”

Mean ratings of voice and static face



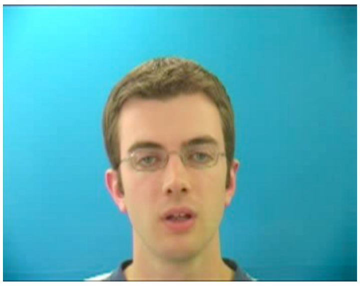
Mean ratings of voice and dynamic face



Legend

● Face
○ Voice

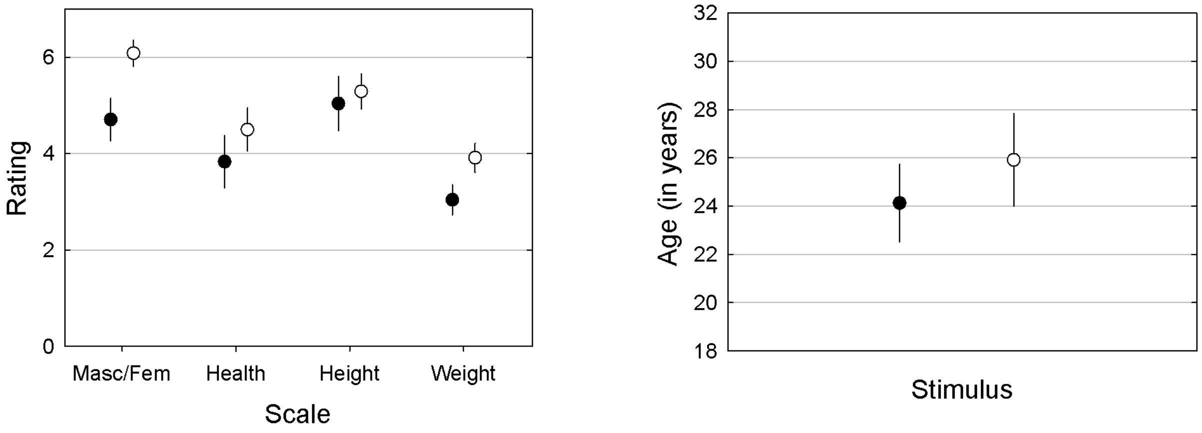
Person 10



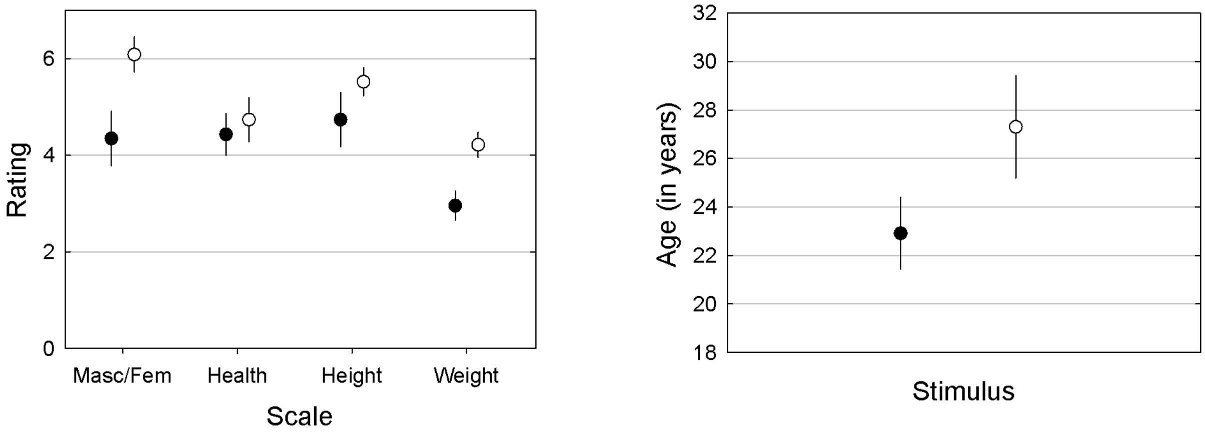
Voice: “bin green at G 8 please”

Dynamic face (muted): “bin red at M 0 please”

Mean ratings of voice and static face



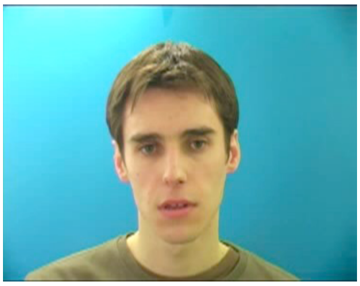
Mean ratings of voice and dynamic face



Legend

● Face
○ Voice

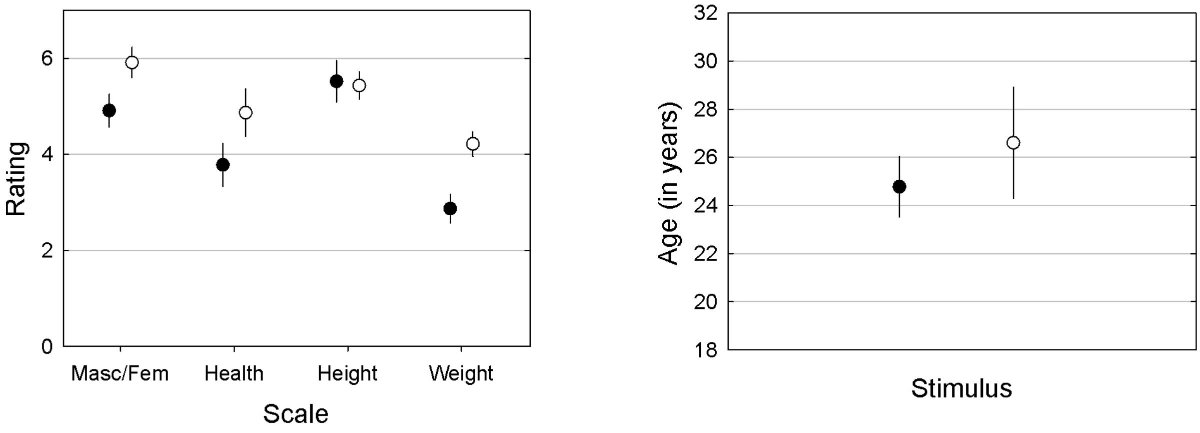
Person 11



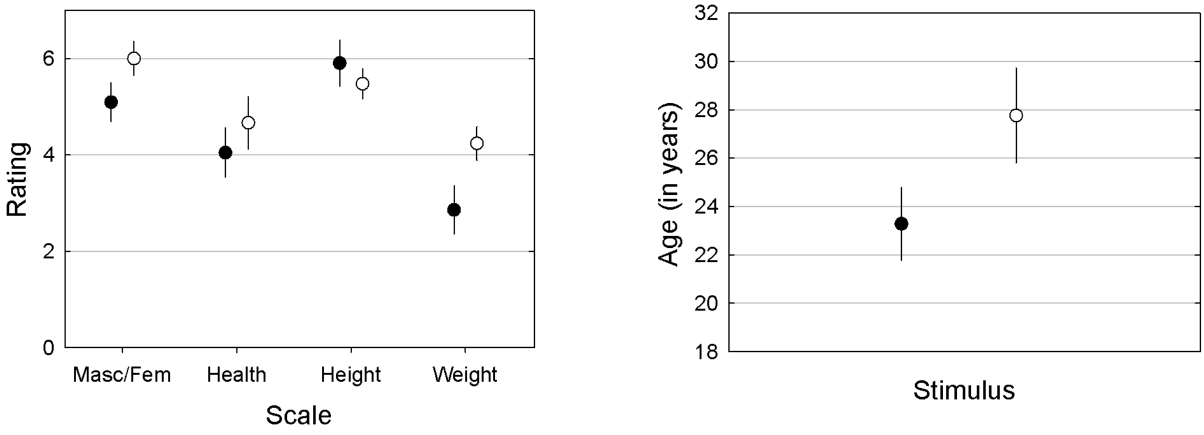
Voice: “set blue at Q 3 soon”

Dynamic face (muted): “set red at D 8 please”

Mean ratings of voice and static face



Mean ratings of voice and dynamic face



Legend

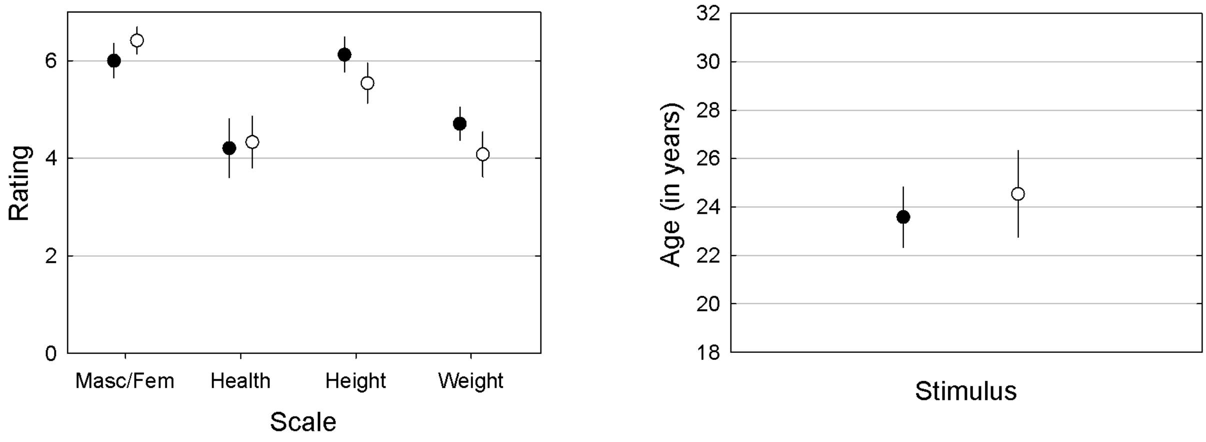
● Face
○ Voice

Person 12

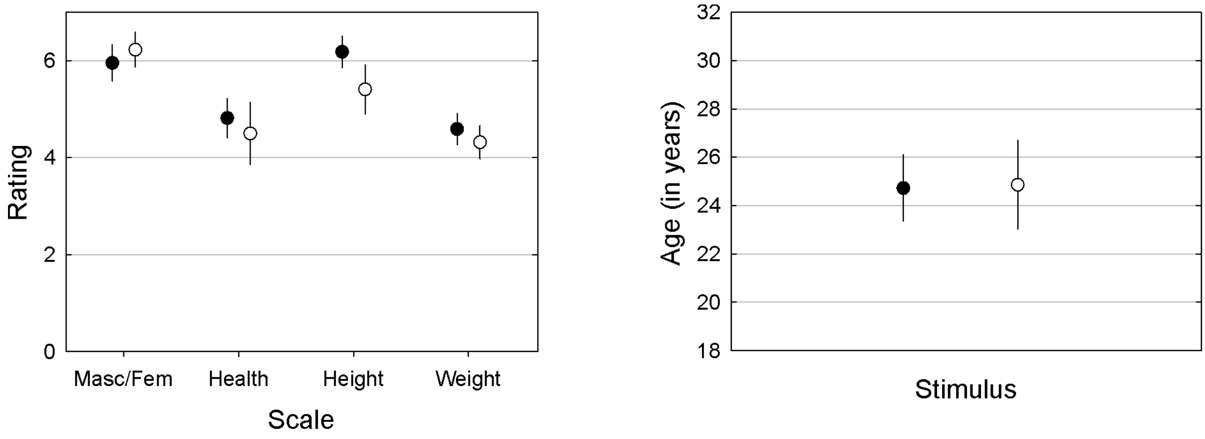
Voice: “lay blue by J 3 now”

Dynamic face (muted): “lay white with L 1 please”

Mean ratings of voice and static face



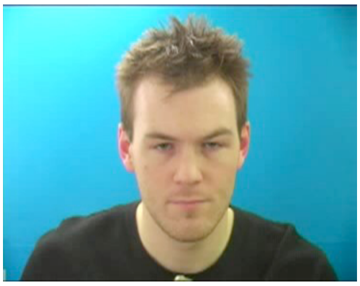
Mean ratings of voice and dynamic face



Legend

● Face
○ Voice

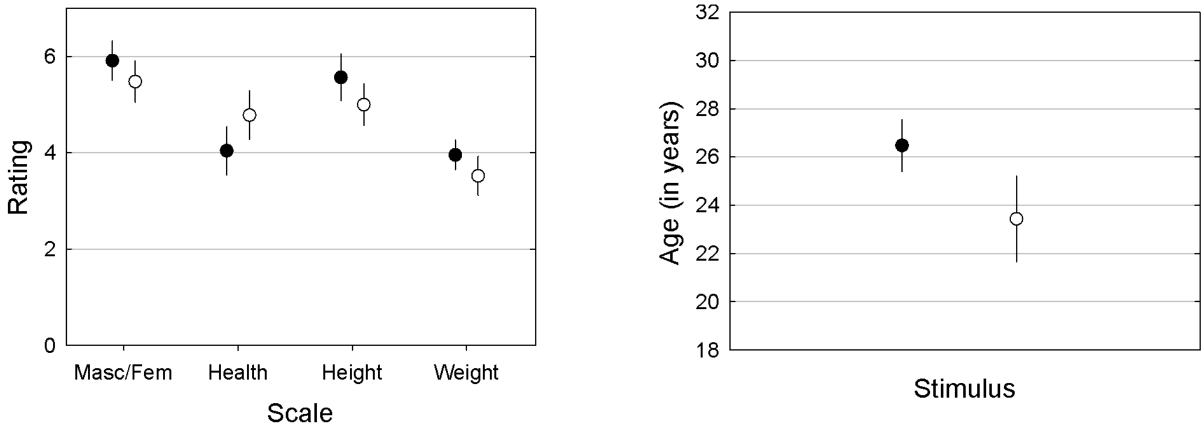
Person 13



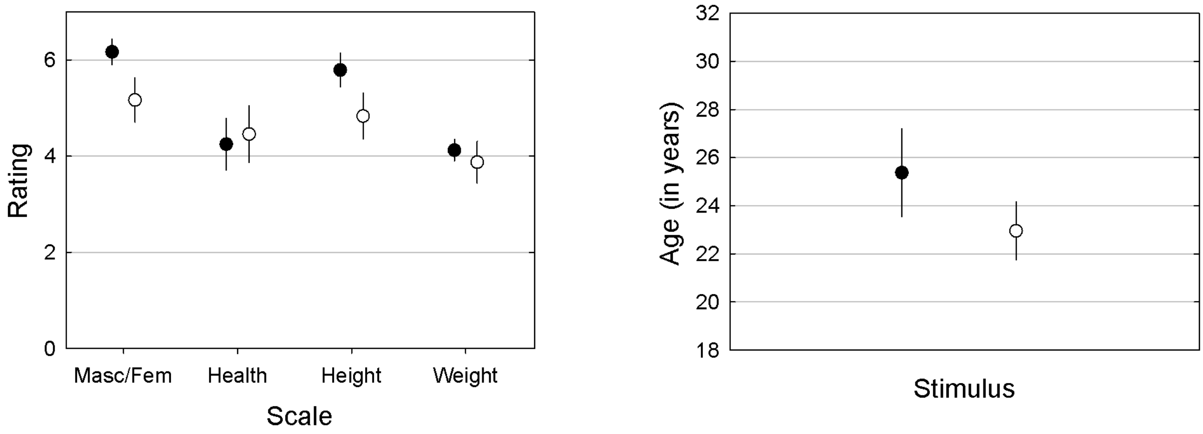
Voice: “place blue at U 3 again”

Dynamic face (muted): “place green in C 2 now”

Mean ratings of voice and static face



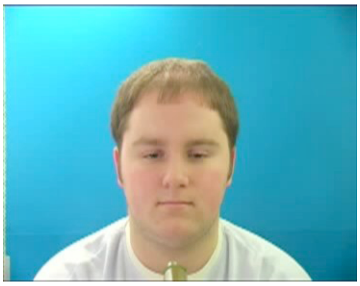
Mean ratings of voice and dynamic face



Legend

● Face
○ Voice

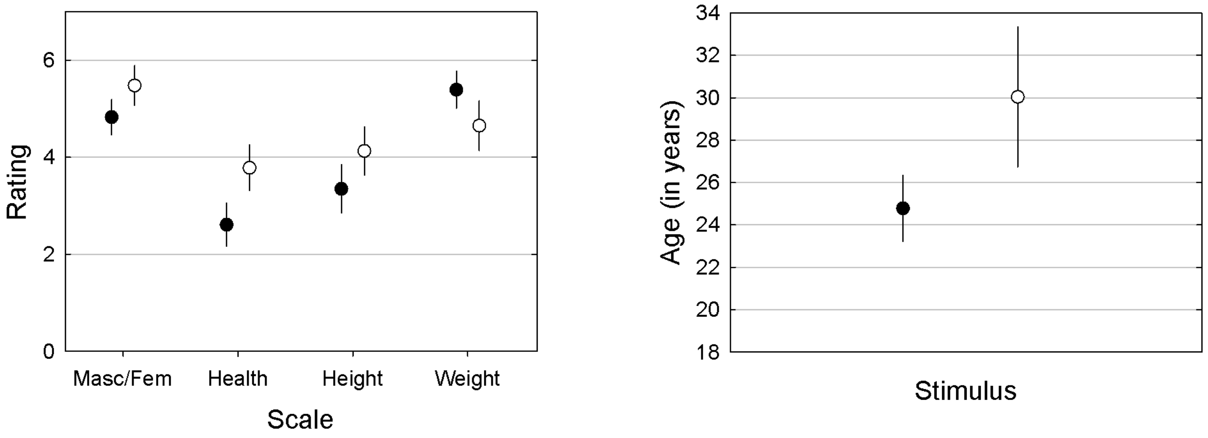
Person 14



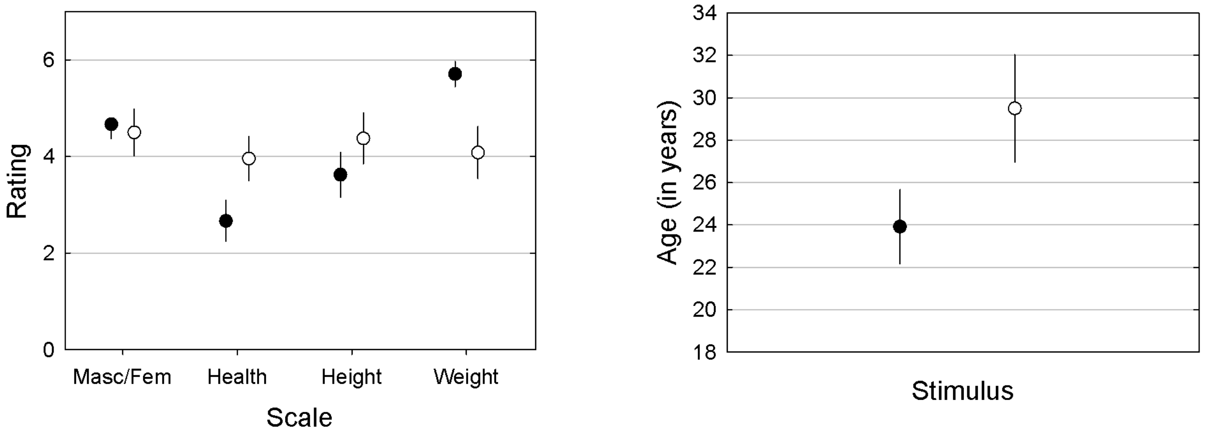
Voice: “set white at G 9 please”

Dynamic face (muted): “bin white in E 7 now”

Mean ratings of voice and static face



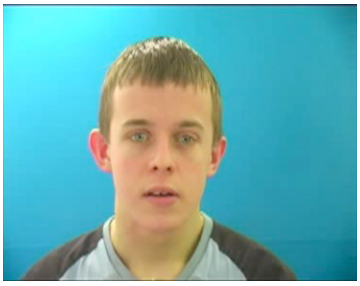
Mean ratings of voice and dynamic face



Legend

● Face
○ Voice

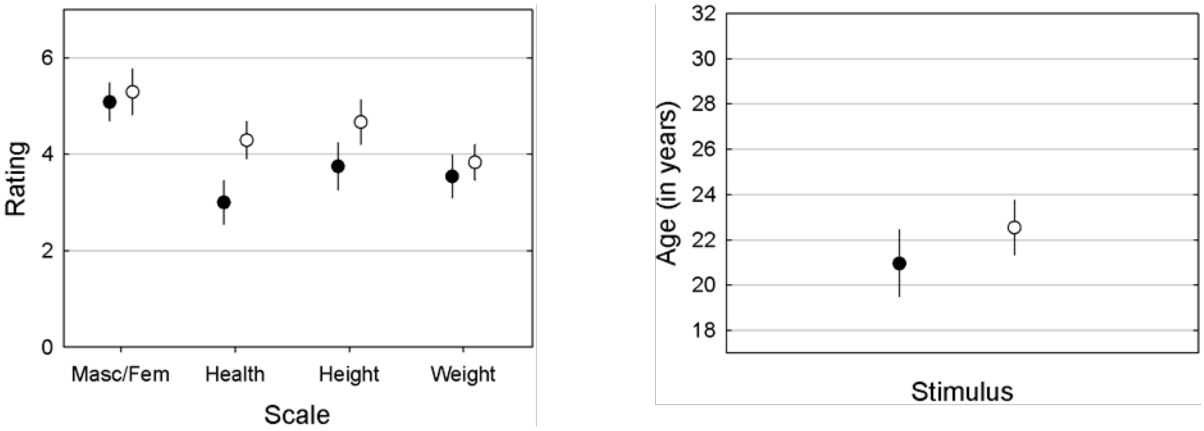
Person 15



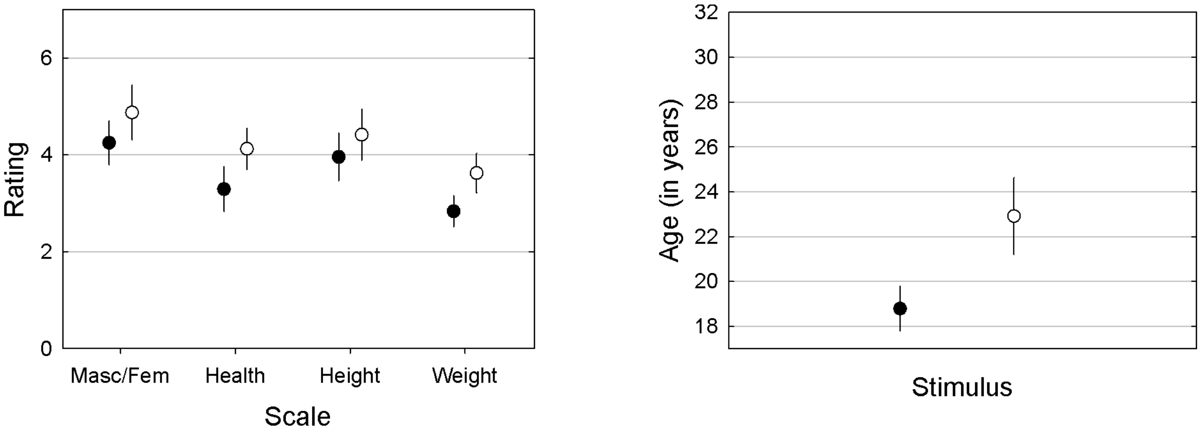
Voice: “set green at H 1 soon”

Dynamic face (muted): “place red by I 0 please”

Mean ratings of voice and static face



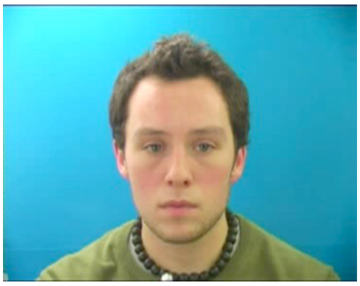
Mean ratings of voice and dynamic face



Legend

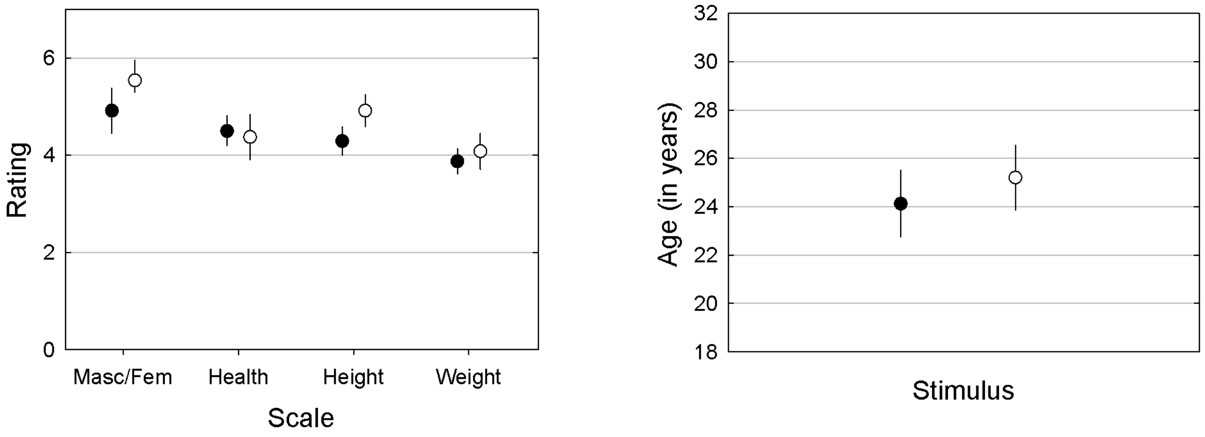
● Face
○ Voice

Person 16

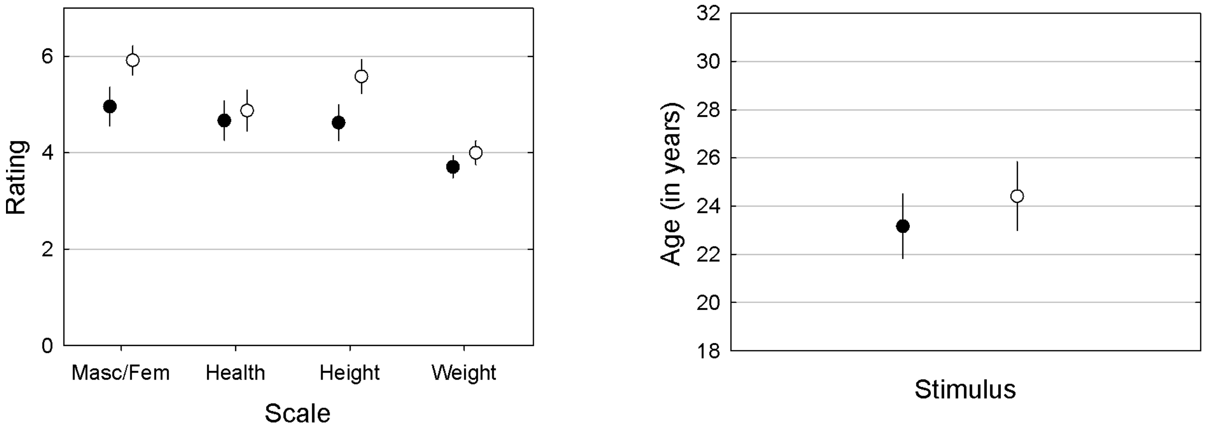


Voice:	“bin red at Q 8 now”
Dynamic face (muted):	“place green by P 2 now”

Mean ratings of voice and static face



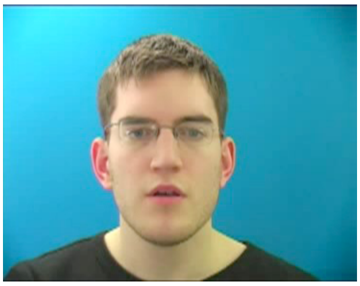
Mean ratings of voice and dynamic face



Legend

● Face
○ Voice

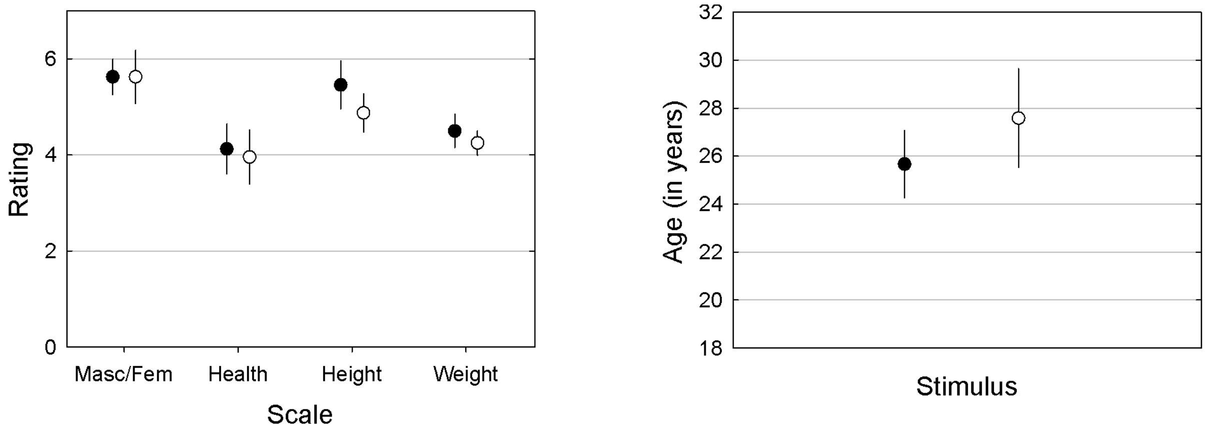
Person 17



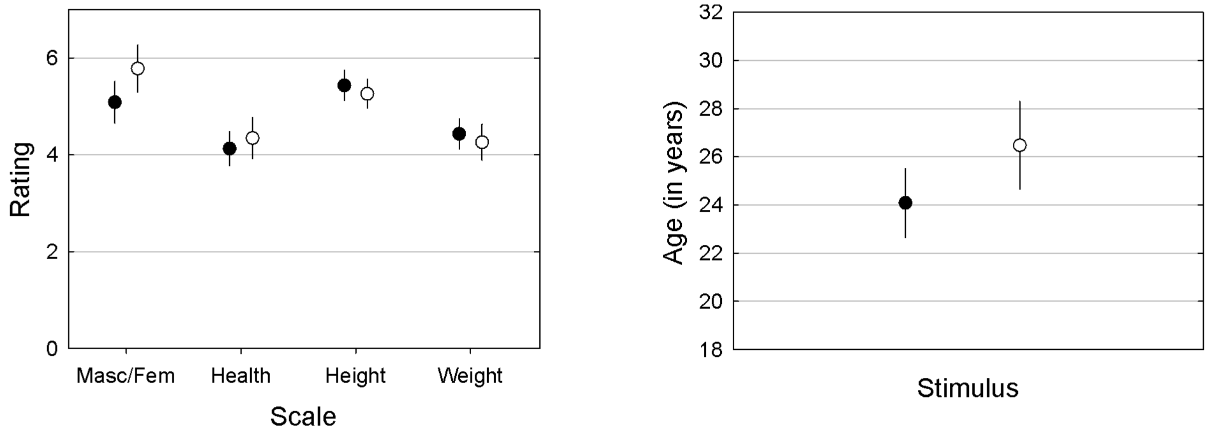
Voice: “lay green by D 2 now”

Dynamic face (muted): “place blue in F 4 now”

Mean ratings of voice and static face



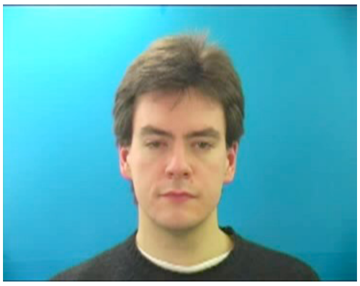
Mean ratings of voice and dynamic face



Legend

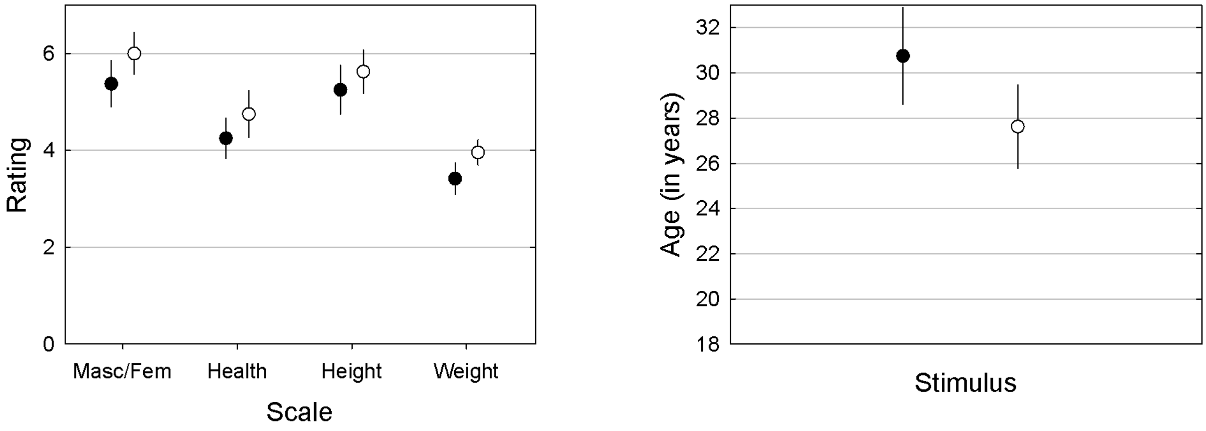
● Face
○ Voice

Person 18

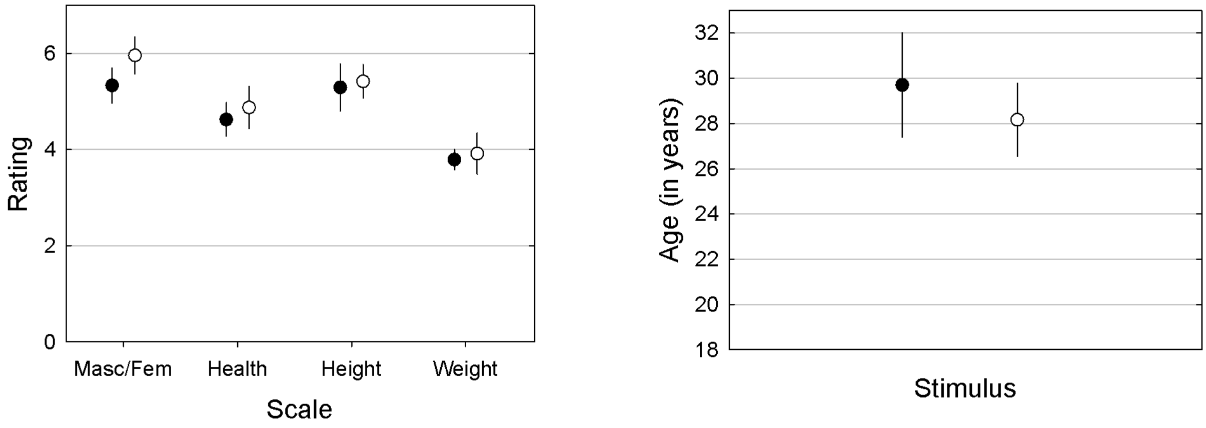


Voice:	“place blue in U 8 please”
Dynamic face (muted):	“place white at P 9 soon”

Mean ratings of voice and static face



Mean ratings of voice and dynamic face



Legend

● Face
○ Voice

APPENDIX C: DATA SETS AND R SCRIPT

Appendix C provides the data sets and executable R script for each experiment reported in this thesis. R is a free, open-source statistical package. In order to run the script, please first download the R package from: http://cran.r-project.org

The data sets and script can be accessed via the following link: https://goo.gl/pnT97T

The data sets and R code are in separate folders for each experiment. The following files can be accessed:

Chapter 4_Exp 1 →	Exp1.csv Exp1_R script.R
Chapter 5_Exp2a →	Exp 2a.csv Exp2a_R script.R
Chapter 5_Exp2b →	Exp2b.csv Exp2b_R script.R
Chapter 5_Exp2c →	Exp2c.csv Exp2c_R script.R
Chapter 6_Exp3a →	Exp3a.csv Exp3a_R script.R
Chapter 6_Exp3b →	Exp3b.csv Exp3b_R script.R

Chapter 7_Exp4a →	Exp4a_Accuracy.csv
	Exp4a_Accuracy_R script.R
	Exp4a_MatchngResponse.csv
	Exp4b_MatchingResponse_R script.R
Chapter 7_Exp4b →	Exp4b_Accuracy.csv
	Exp4b_Accuracy_R script.R
	Exp4b_MatchngResponse.csv
	Exp4b_MatchingResponse_R script.R
Chapter 8_Exp5a →	Exp5a_Accuracy.csv
	Exp5a_Accuracy_R script.R
	Exp5a_MatchngResponse.csv
	Exp5a_MatchingResponse_R script.R
Chapter 8_Exp5b →	Exp5b_Accuracy.csv
	Exp5b_Accuracy_R script.R
	Exp5b_MatchngResponse.csv
	Exp5b_MatchingResponse_R script.R
Chapter 8_Exp5c →	Exp5c_Accuracy.csv
	Exp5c_Accuracy_R script.R
	Exp5c_MatchngResponse.csv
	Exp5c_MatchingResponse_R script.R
Chapter 8_Exp5d →	Exp5d_Accuracy.csv
	Exp5d_Accuracy_R script.R
	Exp5d_MatchngResponse.csv
	Exp5d_MatchingResponse_R script.R

APPENDIX D: PUBLISHED ARTICLE: Smith, Dunn, Baguley & Stacey, (2016a)

Article

Evolutionary Psychology
January-March, 2016: 1–10
© The Author(s) 2016
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1474704916630317
evp.sagepub.com
SAGE

Concordant Cues in Faces and Voices: Testing the Backup Signal Hypothesis

Harriet M. J. Smith[1], Andrew K. Dunn[1], Thom Baguley[1], and Paula C. Stacey[1]

Abstract

Information from faces and voices combines to provide multimodal signals about a person. Faces and voices may offer redundant, overlapping (backup signals), or complementary information (multiple messages). This article reports two experiments which investigated the extent to which faces and voices deliver concordant information about dimensions of fitness and quality. In Experiment 1, participants rated faces and voices on scales for masculinity/femininity, age, health, height, and weight. The results showed that people make similar judgments from faces and voices, with particularly strong correlations for masculinity/femininity, health, and height. If, as these results suggest, faces and voices constitute backup signals for various dimensions, it is hypothetically possible that people would be able to accurately match novel faces and voices for identity. However, previous investigations into novel face–voice matching offer contradictory results. In Experiment 2, participants saw a face and heard a voice and were required to decide whether the face and voice belonged to the same person. Matching accuracy was significantly above chance level, suggesting that judgments made independently from faces and voices are sufficiently similar that people can match the two. Both sets of results were analyzed using multilevel modeling and are interpreted as being consistent with the backup signal hypothesis.

Keywords

face, voice, static, dynamic, backup signal

Date received: September 18, 2015; Accepted: November 14, 2015

Together, faces and voices convey multimodal signals. Such signals are common in animals and occur when information about an underlying trait is communicated by more than one modality. As most research has focused on face and voice ratings independently of each other (Wells, Baguley, Sergeant, & Dunn, 2013; Wells, Dunn, Sergeant, & Davies, 2009), relatively little is known about multimodal signals in humans. Multimodal signals are either backup signals (Johnstone, 1997), or multiple messages (Møller & Pomiankowski, 1993), and are likely to have adaptive value in terms of mate choice. Backup signals are redundant in meaning: they offer similar information and elicit the same response, thereby helping to reduce inaccurate trait assessments (Møller & Pomiankowski, 1993). It is therefore possible to distinguish between multiple messages and backup signals by empirically testing the effect of multimodal signals on a recipient (Partan & Marler, 1999). If a multimodal signal present in human faces and voices is a backup signal for a certain dimension, ratings on this dimension should correlate, whereas uncorrelated ratings would reflect the presence of multiple messages (Wells et al., 2013; Wells et al., 2009).

Multimodal Signals in Faces and Voices

Faces and voices are salient social stimuli, offering a multitude of identity and affective information (Belin, Fecteau, & Bedard, 2004). From an evolutionary perspective, faces and voices provide valuable clues about fitness. For example, in terms of attractiveness they appear to constitute reliable and concordant signals of genetic quality (e.g., Collins & Missing, 2003; Feinberg, 2008; Feinberg et al., 2005; Fraccaro et al., 2010; Saxton, Caryl, & Roberts, 2006; Thornhill & Gangestad, 1999; Thornhill & Grammer, 1999; Wheatley et al., 2014; Zahavi & Zahavi, 1997; see also Puts, Jones, & DeBruine, 2012 for a review), and a number of studies have found that people who have faces that rate highly for attractiveness also

[1] Psychology Division, Nottingham Trent University, Nottingham, UK

Corresponding Author:
Harriet M. J. Smith, Psychology Division, Nottingham Trent University, Burton Street, Nottingham, NG1 4BU, UK.
Email: harriet.smith2011@my.ntu.ac.uk

tend to have voices that rate highly for attractiveness (e.g., Collins & Missing, 2003; Saxton et al., 2006, but see Oguchi & Kikuchi, 1997; Wells et al., 2013).

With the exception of the attractiveness literature, previous research has rarely compared judgments made from faces and voices, focusing instead on judgments informed by a single modality (e.g., Neiman & Applegate, 1990; Penton-Voak & Chen, 2004; Perrett et al., 1998; Pisanski, Mishra, & Rendall, 2012). However, there are a number of reasons as to why we may expect concordance between face and voice ratings in terms of masculinity and femininity, health, age, height, and weight. Some of these reasons are detailed below.

Masculinity/femininity. Levels of reproductive hormone levels are likely to influence perceptions of both facial and vocal femininity and masculinity. For example, testosterone increases the size and thickness of vocal folds (Beckford, Rood, & Schaid, 1985), resulting in lower fundamental frequency (Fant, 1960), which influences perceptions of masculinity (Pisanski et al., 2012). In addition, high levels of testosterone are associated with characteristics of facial masculinity (Penton-Voak & Chen, 2004; Perrett et al., 1998), such as larger jaws, chins, and noses (Miller & Todd, 1998). In women, estrogen slows down vocal fold development and is associated with higher vocal pitch (Abitbol, Abitbol, & Abitbol, 1999; O'Connor, Re, & Feinberg, 2011). Estrogen levels are also related to markers of facial femininity (Thornhill & Grammer, 1999) such as larger lips, smaller lower faces, and fat deposits on the upper cheeks (Perrett et al., 1998).

Health. We might also expect ratings of health made from faces and voices to be similar. Previous research suggests that cues relating to higher levels of reproductive hormones are reliable indicators of fitness and quality (Folstad & Karter, 1992; Thornhill & Gangestad, 2006; Zahavi & Zahavi, 1997), and, indeed, some studies suggest that measures of sexual dimorphism are linked to health ratings and actual health in both men (Gray, Berlin, McKinlay, & Longcope, 1991; Rhodes, Chan, Zebrowitz, & Simmons, 2003) and women (Ellison, 1999; Law Smith et al., 2006).

Age. Faces and voices index information about biological age, a cue which is relevant to reproductive fitness in both males and females (Thornhill & Gangestad, 1999). Numerous visual markers act as indicators of older age, such as decreased elasticity in the skin, wrinkles, discoloration, and reduced clarity in skin tone (Burt & Perrett, 1995). In terms of voices, older people speak with a slower speech rate (Linville, 1996), and age-related hormonal changes affect pitch. For example, female voice pitch lowers after the menopause, whereas older male voices become higher pitched (Linville, 1996). People can estimate a speaker's age from their voice relatively accurately (to within about 10 years; Braun, 1996; Neiman & Applegate, 1990; Ptacek & Sander, 1966; Smith & Baguley, 2014).

Height and weight. Body size is a further indicator of quality (Collins & Missing, 2003; Thornhill & Gangestad, 1999). However, although people tend to agree about height and weight judgments made from a voice (Collins, 2000), this does not indicate that they are necessarily accurate (Bruckert, Liénard, Lacroix, Kreutzer, & Leboucher, 2006; Collins, 2000; van Dommelen & Moxness, 1995). Despite the apparent inaccuracy of height judgments made from voices, people judge height from faces with relative accuracy (Schneider, Hecht, Stevanov, & Carbon, 2013), using cues such as facial elongation. People with longer faces are judged as being taller (Re et al., 2013). Judgments from faces are also accurate for weight estimates (Coetzee, Chen, Perrett, & Stephen, 2010). Lass and Colt (1980) compared visual and auditory height and weight ratings. Results showed significant differences between weight ratings from female faces and voices, suggesting that for some characteristics, faces and voices may not offer concordant information. Recent research has not addressed the extent of concordance between body size information offered by faces and voices. Although Krauss, Freyberg, and Morsella (2002) asked participants to rate the age, height, and weight of speakers from faces and voices, they only tested whether the ratings were accurate, rather than whether there was a relationship between face and voice ratings.

Static and Dynamic Faces

The extent to which faces and voices offer concordant information might be affected by whether the face is static or dynamic. For example, Lander (2008) found that male face and voice attractiveness was only related when faces were dynamic. Studies investigating facial attractiveness and human mate preferences most frequently use static facial stimuli (photos). However, there has been a recent move to use dynamic facial stimuli (videos) in order to improve ecological validity (Gangestad & Scheyd, 2005; Penton-Voak & Chang, 2008; Roberts, Saxton et al., 2009b). Some studies have found that facial stimulus type (static or dynamic) influences attractiveness judgments, although the overall results are somewhat mixed (e.g., Lander, 2008; Penton-Voak & Chang, 2008; Roberts, Little, et al., 2009a; Rubenstein, 2005). In reviewing previous studies and investigating methodological differences between them, Roberts, Saxton et al. (2009b) reported that correlations between ratings from static and dynamic facial stimuli were stronger when rated by the same participants, likely because of carryover effects. As patterns of facial movement vary according to sex (Morrison, Gralewski, Campbell, & Penton-Voak, 2007), it is conceivable that masculinity/femininity ratings will be more extreme when viewing dynamic faces. In light of these findings, it is necessary to consider the influence of facial stimulus type when testing the concordance of face–voice judgments.

Face–voice matching provides a further test of the extent to which faces and voices offer redundant information. However, it is not clear from the literature whether accurate face–voice matching using static facial stimuli is possible. While Kamachi, Hill, Lander, and Vatikiotis-Bateson (2003) showed that participants could match dynamic muted faces saying different sentences to voices of the same identity, participants performed at chance level when the facial stimuli were static. Similar results were reported by Lachs and Pisoni (2004). However, Mavica and Barenholtz (2013) observed above chance level accuracy on trials featuring static faces, suggesting that above chance matching ability is not dependent on being able to encode visual articulatory patterns but rather on concordant information offered by faces and voices.

Aims

This article investigates the extent to which faces and voices offer concordant information, thereby providing a test of the backup signal hypothesis (Johnstone, 1997). Using both static and dynamic facial stimuli, we tested cross-modal concordance by asking participants to make judgments from faces and voices about perceived femininity/masculinity, health, age, height, and weight. In a further test of face–voice concordance, we investigated whether it is possible to accurately match novel static or dynamic faces and voices of the same identity. If faces and voices offer similar information, and it is possible to match the two, this would offer support for the backup signal hypothesis.

Experiment 1

Experiment 1 tested whether faces and voices offer concordant information about dimensions of fitness and quality, aiming to establish whether people make similar judgments about a novel person, regardless of whether they see their face or hear their voice. We expect that as the previous literature suggests that both faces and voices honestly signal quality, judgments made independently from faces and voices should be similar. In light of the contradictory findings regarding judgments made from static and dynamic facial stimuli, the study also tested whether the relationship between face and voice ratings differs according to facial stimulus type (static vs. dynamic).

Method

Design

This experiment employed a mixed design. The between-subject factor was facial stimulus type (static or dynamic), and the within-subject factor was modality (face or voice)

Participants

The participants ($n = 48$) were recruited from the Nottingham Trent University Psychology Division's Research Participation Scheme. There were 12 male and 36 female participants (age range = 18–28 years, $M = 20.54$, $SD = 2.59$). Participants gave informed consent and received a research credit in line with course requirements. The College Research Ethics Committee for Business, Law and Social Sciences granted ethical approval for the study (ref: 2013/37). All participants reported having normal to corrected hearing and vision.

Apparatus and Materials

Stimulus faces and voices were taken from the Grid audiovisual sentence corpus (Cooke, Barker, Cunningham, & Shao, 2006), a multi-talker corpus featuring head and shoulder videos of British adult speakers saying 1,000, six-word sentences each in an emotionally neutral manner recorded against a plain blue background. Each sentence follows the same six-word structure: (1) command, (2) color, (3) preposition, (4) letter, (5) digit, and (6) adverb, for example, "Place blue at J 9 now." None of the speakers in the corpus say the same sentence. A total of 18 speakers were selected from the corpus: 9 males and 9 females. Speakers were matched for ethnicity (White British), accent (English), and age (18–30).

The stimuli were presented on an Acer Aspire laptop (screen size 15.6 inches, resolution 1,366 × 768 pixels, Dolby Advanced Audio) placed approximately 8.5 cm away from the edge of the desk at which participants sat. The experiment was run using Psychopy v1.77.01 (Peirce, 2009), an open-source software package designed for running experiments in Python. Three videos (.mpegs) were selected at random from the GRID corpus for each speaker, using an online research randomizer (Urbaniak & Plous, 2013). The study used static faces, dynamic faces, and voices. One of the three videos was used to create static pictures of faces. Pictures were extracted using the snapshot function on Windows Movie Maker (2012) and presented in .png format. The static picture for each talker was the first frame of the video. Another of the three video files was used to construct the dynamic stimuli. The file was muted using Windows Movie Maker and converted back into .mpeg format. All facial stimuli measured 384 × 288 pixels and were presented in color for 2 s, with brightness settings at the maximum level. Voice recordings were also played for 2 s, from the third .mpeg file, but the face was not visible at presentation. To reduce the background noise, participants listened to the recordings binaurally through Apple earphones with a frequency range of 5–21,000 Hz. This exceeds the range of human hearing (Feinberg et al., 2005). Voices were played at a comfortable listening volume (30% of the maximum volume). Two versions of the experiment were constructed: one using static faces and voices and the other using dynamic faces and voices. In both versions, all 18 faces and voices appeared.

Procedure. Participants were randomly allocated to either the static face or the dynamic face version of the experiment. They read the information sheet, completed the consent form, and provided demographic information. Testing took place in a quiet cubicle. Participants completed two counterbalanced blocks of testing. In one block participants viewed faces, in the other they heard voices. Participants were not told that the voices and faces featured in the experiment belonged to the same people. Each block consisted of a practice trial followed by 18 randomly ordered experimental trials. After each face or voice, participants estimated the age of the stimulus person in years and completed the 7-point Likert-style rating scales in the following order: femininity/masculinity (1 = *very feminine*, 7 = *very masculine*), health (1 = *very unhealthy*, 7 = *very healthy*), height (1 = *very short*, 7 = *very tall*), and weight (1 = *very underweight*, 7 = *very overweight*).

Data Analysis and Multilevel Modeling

Data were analyzed using multilevel models, rather than performing conventional analyses on data averaged over either participants or stimuli (see Wells et al., 2013). This avoids the ecological fallacy which arises when it is falsely assumed that patterns observed for participant means also hold for data at a lower level of analysis such as individual trials repeated within participants (e.g., see Robinson, 1950; Wells et al., 2013). Multilevel modeling allows both participants and stimuli to be simultaneously treated as random effects, thereby maximizing generalizability (Clark, 1973; Judd, Westfall, & Kenny, 2012). When the random effects are fully crossed (i.e., when all participants experience all stimuli), conventional analyses (including separate by-items or by-subjects analyses) can lead to massive Type 1 error inflation (Baguley, 2012; Clark, 1973; Judd et al., 2012). The most appropriate analysis therefore takes into account both sources of variability. Unless the ignored source of variability is negligible, this is always more conservative than separate by-stimuli or by-participants analyses.

Results

We calculated the absolute difference between face and voice ratings by comparing each rating participants had given to a face and voice belonging to the same person. Then we calculated the mean absolute difference (MAD) for each stimuli person on each rating scale (age, masculinity/femininity, health, height, and weight). Descriptive statistics (Table 1) indicate that typical ratings for faces and voices fall within a similar range.

On all scales apart from age, face and voice ratings only differ on average by about 1 point (14%) on a 7-point rating scale, and MADs were similar across static and dynamic facial stimuli. The difference between face and voice ratings in terms of age appears larger than that of the other rating scales. However, rather than being rated on a 7-point scale, age estimates were given in years. This prevents a neat comparison between the rating scales.

The results in Table 1 show that face and voice ratings tend to be close together in terms of the range they fall into. A logical next step is to quantify the extent to which voice and face ratings covary in the same individual. For this purpose, a simple correlation coefficient between voice and face ratings would either ignore the dependency within participants or rely only on aggregate data (mean ratings for each participant). We therefore used multilevel models to account for both participant and stimuli variation when correlating voice ratings with face ratings for estimated age and ratings for femininity/masculinity, health, height, and weight. For each variable, we fitted an intercept-only model with the rating as an outcome, using the lme4 package in R (Bates, Maechler, Bolker, & Walker, 2014). A crucial part of each model was to estimate separate variance for face and voice ratings as well as the correlation between face and voice ratings across both stimuli and participants. The correlation between face and voice ratings within participants is, for present purposes, a nuisance term (merely indicating that participants who give high ratings to voices also tend to give high ratings to faces) and is not reported here. The correlations reported in Table 2 are those within stimuli and demonstrate that, for a given item, voice and face ratings are positively correlated.

Table 1. MAD and 95% Confidence Intervals for the MAD Between Face and Voice Ratings by Stimulus-Type Condition.

	Static Facial Stimuli				Dynamic Facial Stimuli			
			95% CI				95% CI	
Rating scale	*M*	*SD*	LB	UB	*M*	*SD*	LB	UB
Age	3.91	1.51	3.27	4.55	3.62	1.58	2.95	4.29
Masculinity/femininity	1.05	0.35	0.90	1.19	1.00	.36	0.85	1.15
Health	1.24	.34	1.10	1.39	1.12	0.27	1.00	1.23
Height	1.10	.29	0.98	1.23	1.04	0.36	0.89	1.19
Weight	0.92	0.25	0.81	1.02	1.00	0.27	0.88	1.11

Note. MAD = mean absolute difference.

Table 2. Within-Stimulus Correlations Between Face and Voice Ratings.

	Correlation coefficient				
Condition	Age	Masc/fem	Health	Height	Weight
Static facial stimuli	.60	.97	.70	.83	.40
Dynamic facial stimuli	.32	.92	.91	.86	.17
All facial stimuli	.46	.95	.77	.84	.28

Table 2 provides evidence that mean face and voice ratings for the same target appear to be positively related for all rating types. Correlations between face and voice ratings on scales for masculinity/femininity, health, and height were particularly high, regardless of whether the facial stimuli were static or dynamic. Correlations between mean face and voice ratings for age and weight were moderate when facial stimuli were static—with some suggestion that the correlations were diminished for dynamic stimuli. However, correlations did not vary according to facial stimulus type in direction or by more than .3 on any scale. The difference between the static and dynamic correlations was tested by fitting models with separate variance terms for each stimulus type. Comparing a model which includes separate variance and covariance terms for static and dynamic stimuli with one that does not did not improve the model fit for any of the ratings ($p > .14$). This complements the results shown in Table 1, suggesting that the extent to which faces and voices offer similar information is not greatly influenced by whether the facial stimuli is static or dynamic.

Discussion

Experiment 1 showed that observers glean concordant information about different dimensions of quality from faces and voices, particularly in terms of masculinity and femininity, health, and height. On each dimension, the relatedness of face and voice ratings is not affected by facial stimulus type, showing that the signals tested here are stable across static and dynamic faces. These results support the hypothesis that on various dimensions of quality, faces and voices constitute backup signals.

Experiment 2

Experiment 2 tested whether faces and voices offer sufficiently concordant information that people can match novel faces to voices. Previous studies have addressed this question, with conflicting results. Krauss et al. (2002) showed that people are relatively accurate at inferring physical information from a voice. After only hearing a voice excerpt, participants selected the speaker's full-length photograph from one of two possible options with above chance accuracy. Mavica and Barenholtz (2013) tested whether people could use information from a voice to distinguish between two static images of different faces. Accuracy was significantly above chance level, despite contradictory results presented in previous studies (Kamachi et al., 2003; Lachs & Pisoni, 2004) suggesting that successful matching of faces and voices depends on the ability to encode dynamic properties of speaking (muted) faces (Mavica & Barenholtz, 2013).

Previous face–voice matching studies (Kamachi et al., 2003; Krauss et.al., 2002; Mavica & Barenholtz, 2013) have used a two-alternative forced choice paradigm (2AFC), which unlike a same–different paradigm does not model whether people are also able to correctly reject a match when a face and voice are from different people. The 2AFC tasks therefore give no information about possible response biases. Experiment 2 uses a same–different paradigm to give a clearer picture of face–voice matching ability.

Experiment 2 addresses three main questions. First, whether it is possible to accurately match novel faces and voices of the same age (20–30), sex, and ethnicity (White British). Second, whether matching accuracy is affected by facial stimulus type (static or dynamic). Third, in line with cross-modal matching procedures (Kamachi et al., 2003; Lachs & Pisoni, 2004), we investigated whether people are more accurate at face–voice matching when visual information (a face) is presented first, compared to when auditory information (a voice) is presented first. If faces and voices primarily constitute backup signals, people should be able to match novel faces and voices above chance level.

Method

The methods for Experiment 2 were the same as for Experiment 1, with exceptions explained in the following subsections.

Design

This experiment employed a 2 × 2 × 2 mixed factorial design. The between-subject factor was facial stimulus type (static or dynamic). The within-subject factors were identity (same or different) and order (face first or voice first). The dependent variable was accuracy.

Participants

There were 40 male and 40 female adult participants ($n = 80$) with an age range of 18–66 years ($M = 25.44$, $SD = 8.36$).

Materials

Four different versions of the experiment were created so that matching and not-matching pairs of faces and voices could be constructed using different stimulus people. Stimuli were randomly selected to be used for either one of the eight same identity or eight different identity trials. None of the faces or voices appeared more than once in each version. On different identity trials, the face and voice were matched for age, gender, and ethnicity. The stimuli that remained were used for the practice trials. Each version was repeated for static and dynamic conditions. In total, there were eight versions.

Procedure

Participants were randomly allocated to one of the eight versions of the experiment. In the dynamic facial stimulus condition, participants were also correctly informed that the face in the muted video and the voice in the recording were not saying the same thing. This was to prevent them using speech reading to match the face and voice (Kamachi et al., 2003).

Participants completed two counterbalanced experimental blocks, each consisting of a practice trial followed by eight randomly ordered experimental trials. In one block, participants saw the face first, and in the other they heard the voice first. None of the stimuli appeared more than once in each version of the experiment. In each trial, there was a 1-s gap between presentation of the face and voice stimuli. At test, participants pressed "1" if they thought the face and voice were "matching" (same identity), and "0" if they thought it was "not matching" (different identity).

Results

Performance accuracy was analyzed using multilevel logistic regression with the lme4 version 1.06 package in R (Bates et al., 2014). Four nested models with accuracy (0 or 1) as the dependent variable were compared (and all models were fitted using restricted maximum likelihood). The first model included a single intercept (and was later used to obtain confidence intervals for the overall accuracy). The second model also included the main effects of each factor (identity, order, and stimulus type). The third model added all two-way interactions and the final model added the three-way interaction. Setting up the model in this way allows us to test for individual effects in a manner similar to that of a traditional analysis of variance. However, as F-tests-derived multilevel models are not, in general, accurate, we report the more robust profile likelihood ratio

Table 3. Parameter Estimates (*b*) and Profile Likelihood Tests for the 2 × 2 × 2 Factorial Analysis of Accuracy in Experiment 2.

Source	*df*	*b*	*SE*	G^2	*p*
Intercept	1	−0.445	0.196		
Identity	1	1.382	0.254	57.84	<.001
Order	1	0.509	0.241	2.28	.131
Facial stimulus type	1	0.133	0.231	0.13	.717
Identity × Order	1	0.601	0.358	4.20	.040
Identity × Facial Stimulus Type	1	0.165	0.339	0.32	.572
Order × Facial Stimulus Type	1	0.052	0.324	0.01	.916
Identity × Order × Facial Stimulus Type	1	0.058	0.474	0.01	.903

tests provided by lme4. These were obtained by dropping each effect in turn from the appropriate model (e.g., testing the three-way interaction by dropping it from the model including all effects, and testing the two-way interactions by dropping each effect in turn from the two-way model).

Table 3 shows the profile likelihood chi-square statistic (G^2) and *p*-value associated with dropping each effect. Table 3 also reports the coefficients and standard errors (on a log odds scale) for each effect in the full three-way interaction model. In the three-way model, the estimate of *SD* of the face random effect was 0.353, while for voice it was 0.207. The estimated *SD* for the participant effect was less than 0.0001. A similar pattern held for the null model. Thus, although individual differences were negligible in this instance, a conventional by-participants analysis that did not incorporate both voice and face variation could be extremely misleading.

Only the main effect of identity and the two-way interaction of identity and order were statistically significant. To aid interpretation of these effects, we obtained means and confidence intervals for the percentage accuracy of the eight conditions in the factorial design. These confidence intervals were obtained through simulations of the posterior distributions of the cell means using arm package version 1.6 in R (Gelman & Su, 2013). These means and the associated 95% confidence intervals are shown in Figure 1.

From Figure 1 it is clear that overall matching performance was significantly above chance (50%) level, $M = 59.7\%$, 95% CI [51.9, 66.9]. Static face–voice matching was above chance, $M = 59.19$, 95% CI [50.94, 66.84], as was dynamic face–voice matching, $M = 60.12$, 95% CI [51.97, 67.74]. Figure 1 also reveals the main effect of identity, with performance for same trials consistently higher than for different trials (and the former but not the latter consistently above chance). It also reveals the basis of the identity by order interaction. The results from the face first trials are shown in Panel A. The results from the voice first trials are shown in Panel B. Although same identity trials showed better performance than different trials for both face first and voice first trials, this advantage is greater in the face first conditions. Given that performance on the face first different trials is on average worse than chance (and significantly so for the static stimuli), this pattern suggests the operation of a response bias, such that participants exhibited a bias to accept faces and voices as belonging to the same identity when they saw the face before hearing the voice.

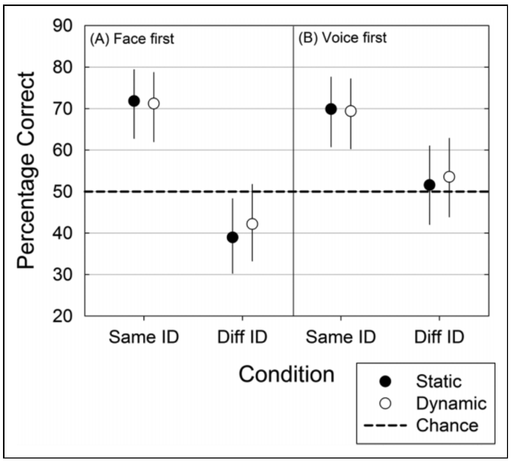


Figure 1. Face–voice matching accuracy on face first (Panel A) and voice first (Panel B) trials. Error bars show 95% CI for the condition means. CI = confidence interval.

Discussion

In Experiment 2, we observed that both dynamic faces and voices, and static faces and voices, can be matched for identity above chance level. These results are consistent with the hypotheses informed by the results of Experiment 1, which show that faces and voices offer a high level of concordant information on various dimensions. Face–voice matching performance does not differ according to facial stimulus type. Therefore, accuracy does not appear to depend on encoding visual information about speaking style but rather on redundant signals available in voices and static faces.

General Discussion

The results of Experiment 1 are consistent with the hypothesis that faces and voices offer redundant signals for various dimensions of quality. Mean face and voice ratings for the same target were positively related for all rating types. Correlations between face and voice ratings on scales for masculinity/femininity, health, and height were particularly strong, regardless of whether the facial stimuli were static or dynamic. The results of Experiment 2 show that the information signaled by faces and voices is so similar that people can match novel faces and voices of the same sex, ethnicity, and age-group at a level significantly above chance. Taken together, results suggest that faces and voices constitute backup signals, reinforcing the same information about quality (Johnstone, 1997) rather than complementary but different information (Møller & Pomiankowski, 1993).

Face and Voice Ratings

With the exception of the attractiveness literature, previous research has rarely compared judgments made from faces and voices, focusing instead on judgments informed by a single modality (e.g., Penton-Voak & Chen 2004; Perrett et al., 1998; Pisanski et al., 2012; Neiman & Applegate, 1990, and so on) or comparing face and voice ratings to actual measurements of physical characteristics (e.g., Krauss et al., 2002) rather than to each other. The results of Experiment 1 show that not only do face and voice ratings fall within a small range but independent ratings of an individual's face and voice are positively correlated. These results complement other studies, showing that faces and voices offer related information about fitness and mate value (Collins & Missing, 2003; Feinberg, 2008; Feinberg et al., 2005; Fraccaro et al., 2010).

The strongest correlations between face and voice ratings occurred on scales for masculinity/femininity, health, and height. Despite the previous literature suggesting that unimodal voice ratings of body size are less accurate than unimodal face ratings (Bruckert et al., 2006; Coetzee et al., 2010; Collins, 2000; Re et al., 2013; van Dommelen & Moxness, 1995), Experiment 1 showed that regardless of accuracy, the MAD between body size judgments made from faces and voices was small. However, correlations were strong for height but only weak-moderate for weight. This corresponds with Lass and Colt (1980) who found significant differences between weight ratings for female faces and voices.

Face and Voice Matching

Overall, face–voice matching accuracy in Experiment 2 was significantly above chance. This result is consistent with previous findings (Krauss et al., 2002; Mavica & Barenholtz, 2013) and shows that people can use redundant information to match faces and voices of the same identity. Furthermore, the use of multilevel modeling allows us to generalize these findings beyond the sample of faces and voices used, thereby overcoming a common limitation of previous studies.

Although overall matching accuracy is at 59.7%, there is still a substantial proportion of unexplained variance which could be due to the existence of discordant rather than concordant face–voice information. Beyond the characteristics tested in Experiment 1, faces and voices also convey a multitude of other information, including personality characteristics and emotion (Belin et al., 2004; Mavica & Barenholtz, 2013), some of which might be complementary. Nevertheless, the results from Experiment 2 suggest that on balance, faces and voices provide concordant information because overall performance is significantly above chance level. These results are consistent with the results presented in Experiment 1.

On different identity trials, participants performed at chance level (voice first trials), or below chance level (face first trials), and were significantly less accurate than on same identity trials. This indicates that participants were better at detecting a correct match than rejecting an incorrect one. In line with the argument presented above, based purely on the findings from Experiment 1, we might have expected that accurately rejecting mismatches would be possible because the ratings were so closely related. It seems that participants are using other information to inform their matching decisions on different identity trials. On the other hand, the pattern of results across same–different trials might be partially explained by the existence of a response bias.

While previous face–voice matching studies using 2AFC procedures have found no difference between face first and voice first performance (Kamachi et al., 2003; Lachs & Pisoni, 2004), our results using a same–different task suggest people exhibit a bias to respond that a face and voice belong to the same identity, particularly when the face is presented before the voice. A performance asymmetry, according to stimuli order, is consistent with the previous literature. For instance, studies have consistently found asymmetries between faces and voices in terms of rates of recognition accuracy, which have been attributed to differential link strength in the two perception pathways (e.g., Damjanovic & Hanley, 2007; Hanley & Turner, 2000; Stevenage, Hugill, & Lewis, 2012). Therefore, there is no reason to assume that face first and voice first matching performance should be identical. However, based on the finding that familiar faces prime familiar voices better than familiar voices prime familiar faces (Stevenage et al., 2012), we might have expected the asymmetry to operate the other way around. Nevertheless, it is feasible that voices give more information about faces than faces do about voices, and aside from conveying semantic information about the spoken message, the other important role of voices is to allow people to infer socially relevant visual information about the speaker, such as information about masculinity/femininity, body size, health, and age. This idea is in keeping with the finding that showing participants mismatched celebrity face–voice pairs disrupts voice recognition to a greater extent than it disrupts face recognition (Stevenage, Neil, & Hamlin, 2014). During social interactions, it is common to hear a voice while not looking in the direction of the speaker. Being able to accept or reject a face match quickly may aid social communication by facilitating attention shifts.

Static and Dynamic Faces

Informed by contradictory findings relating to the effect of static and dynamic facial stimuli on ratings of attractiveness (e.g., Lander, 2008; Roberts, Little, et al., 2009a; Rubenstein, 2005) and face–voice matching ability (Kamachi et al., 2003; Lachs & Pisoni, 2004; Mavica & Barenholtz, 2013), we tested whether facial stimulus type affected the extent of face–voice concordance. In both experiments, performance was unaffected by whether the facial stimuli were dynamic or static. This suggests that information on these dimensions is stable across dynamic and static faces. Novel face–voice matching ability is not due to encoding visual articulatory patterns (Mavica & Barenholtz, 2013) but to the availability of redundant information.

Stimulus Sample Size

The findings of the multilevel models we report emphasize the importance of stimulus sample size in estimating effects. These models provide the tools to generalize over both participants and stimuli, but obtaining large samples of stimuli is challenging. The corpus (Cooke et al., 2006) we used only contained 18 stimulus individuals matched for age, gender, and ethnicity. This reduced the set of stimuli available for study but also reduced extraneous variability. In addition, all of the people in this stimulus set were from similar educational backgrounds (Cooke et al., 2006), and none of them exhibited strong regional accents. As there is a high level of interstimulus variability in both faces (Valentine, Lewis, & Hills, 2015) and voices (Stevenage & Neil, 2014), we would encourage future face–voice matching studies to aim for larger samples of stimuli, having demonstrated that it is variation in faces and voices that is the limiting factor on statistical power in experiments such as these (as face and voice variation is consistently higher than participant variation). However, many published studies have used samples of stimuli far smaller than 18 when investigating person perception (see G. L. Wells & Windshitl, 1999), as have other face–voice matching studies (e.g., Lachs & Pisoni, 2004). Crucially, only by accounting for variability in stimuli is it reasonable to generalize from stimuli as well as participants. Even in studies using large sample of stimuli, generalizability is limited by the common practice of aggregating over stimuli (Clark, 1973; Judd et al., 2012; Wells et al., 2013). Ultimately, the adequate sample size of stimuli or participants in experimental designs such as those reported here is a question of statistical power (e.g., see Westfall, Kenny, & Judd, 2014).

Conclusion

Faces and voices of the same identity offer redundant signals about a number of dimensions associated with quality and fitness. Information about masculinity/femininity, height, and health is particularly similar across faces and voices. We have shown that the level of redundancy between faces and voices is sufficient that it is possible to accurately match them for identity. In summary, the results of Experiments 1 and 2 are more consistent with the backup signal hypothesis (Johnstone, 1997) than the multiple messages hypothesis (Møller & Pomiankowski, 1993). As multimodal signals for various indicators of quality, faces, and voices offer concordant rather than complementary information.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The first author was supported by a Ph.D. studentship from the Division of Psychology, Nottingham Trent University.

References

Abitbol, J., Abitbol, P., & Abitbol, B. (1999). Sex hormones and the female voice. *Journal of Voice*, *13*, 424–446. doi:10.1016/S0892-1997(99)80048-4

Baguley, T. (2012). Calculating and graphing within-subject confidence intervals for ANOVA. *Behavior Research Methods*, *44*, 158–175. doi:10.3758/s13428-011-0123-7

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.0-6. Retrieved September 12, 2014, from http://CRAN.R-project.org/package=lme4

Beckford, N. S., Rood, S. R., & Schaid, D. (1985). Androgen stimulation and laryngeal development. *Annals of Otology, Rhinology, and Laryngology*, *94*, 634–640.

Belin, P., Fecteau, S., & Bedard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences*, *8*, 129–135. doi:10.1016/j.tics.2004.01.008

Braun, A. (1996). Age estimation by different listener groups. *International Journal of Speech Language and the Law*, *3*, 65–73. doi:10.1558/ijsll.v3i1.65

Bruckert, L., Liénard, J. S., Lacroix, A., Kreutzer, M., & Leboucher, G. (2006). Women use voice parameters to assess men's characteristics. *Proceedings of the Royal Society, B: Biological Sciences*, *273*, 83–89. doi:10.1098/rspb.2005.3265

Burt, D. M., & Perrett, D. I. (1995). Perception of age in adult Caucasian male faces: Computer graphic manipulation of shape and colour information. *Proceedings of the Royal Society, B: Biological Sciences*, *259*, 137–143. doi:10.1098/rspb.1995.0021

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*, 335–359. doi:10.1016/S0022-5371(73)80014-3

Coetzee, V., Chen, J., Perrett, D. I., & Stephen, I. D. (2010). Deciphering faces: Quantifiable visual cues to weight. *Perception*, *39*, 51–61. doi:10.1068/p6560

Collins, S. A. (2000). Men's voices and women's choices. *Animal Behaviour*, *60*, 773–780. doi:10.1006/anbe.2000.1523

Collins, S. A., & Missing, C. (2003). Vocal and visual attractiveness are related in women. *Animal Behaviour*, *65*, 997–1004. doi:10.1006/anbe.2003.2123

Cooke, M., Barker, J., Cunningham, S., & Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, *120*, 2421–2424. doi:10.1121/1.2229005

Damjanovic, L., & Hanley, J. R. (2007). Recalling episodic and semantic information about famous faces and voices. *Memory & Cognition*, *35*, 1205–1210. doi:10.3758/BF03193594

Ellison, P. T. (1999). Reproductive ecology and reproductive cancers. In C. Pater-Brick & C. Worthman (Eds.), *Hormones, health, and behavior: A socio-ecological and lifespan perspective* (pp. 184–209). Cambridge, England: Cambridge University Press.

Fant, G. (1960). *The acoustic theory of speech production*. The Hague, the Netherlands: Mouton.

Feinberg, D. R. (2008). Are human faces and voices ornaments signaling common underlying cues to mate value? *Evolutionary*

Anthropology: Issues, News, and Reviews, *17*, 112–118. doi:10.1002/evan.20166

Feinberg, D. R., Jones, B. C., DeBruine, L. M., Moore, F. R., Law Smith, M. J., Cornwell, R. E., . . . Perrett, D. I. (2005). The voice and face of woman: One ornament that signals quality? *Evolution and Human Behavior*, *26*, 398–408. doi:10.1016/j.evolhumbehav.2005.04.001

Folstad, I., & Karter, A. J. (1992). Parasites, bright males, and the immunocompetence handicap. *American Naturalist*, *139*, 603–622. doi:10.1086/285346

Fraccaro, P. J., Feinberg, D. R., DeBruine, L. M., Little, A. C., Watkins, C. D., & Jones, B. C. (2010). Correlated male preferences for femininity in female faces and voices. *Evolutionary Psychology*, *8*, 447–461. doi:10.1177/147470491000800311

Gangestad, S. W., & Scheyd, G. J. (2005). The evolution of human physical attractiveness. *Annual Review of Anthropology*, *34*, 523–548. doi:10.1146/annurev.anthro.33.070203.143733

Gelman, A. E., & Su, Y. S. (2013). arm: Data analysis using regression and multilevel/hierarchical models. R package version 1.6-05. Retrieved September 12, 2014, from http://CRAN.R-project.org/package=arm

Gray, A., Berlin, J. A., McKinlay, J. B., & Longcope, C. (1991). An examination of research design effects on the association of testosterone and male aging: Results of a meta-analysis. *Journal of Clinical Epidemiology*, *44*, 671–684. doi:10.1016/0895-4356(91)90028-8

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*, 1105–1116. doi:10.1080/713755942

Johnstone, R. A. (1997). The evolution of animal signals. In J. R. Krebs & N. B. Davies (Eds.), *Behavioural ecology: An evolutionary approach* (pp. 155–178). Oxford, England: Blackwell.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*, 54. doi:10.1037/a0028347

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. doi:10.1016/j.cub.2003.09.005

Krauss, R. M., Freyberg, R., & Morsella, E. (2002). Inferring speakers' physical attributes from their voices. *Journal of Experimental Social Psychology*, *38*, 618–625. doi:10.1016/S0022-1031(02)00510-3

Lachs, L., & Pisoni, D. B. (2004). Crossmodal source identification in speech perception. *Ecological Psychology*, *16*, 159–187. doi:10.1207/s15326969eco1603_1

Lander, K. (2008). Relating visual and vocal attractiveness for moving and static faces. *Animal Behaviour*, *75*, 817–822. doi:10.1016/j.anbehav.2007.07.001

Lass, N. J., & Colt, E. G. (1980). A comparative study of the effect of visual and auditory cues on speaker height and weight identification. *Journal of Phonetics*, *8*, 277–285.

Law Smith, M. J., Perrett, D. I., Jones, B. C., Cornwell, R. E., Moore, F. R., Feinberg, D. R., . . . Hillier, S. G. (2006). Facial appearance is a cue to oestrogen levels in women. *Proceedings of the Royal Society B: Biological Sciences*, *273*, 135–140. doi:10.1098/rspb.2005.3296

Linville, S. E. (1996). The sound of senescence. *Journal of Voice*, *10*, 190–200. doi:10.1016/S0892-1997(96)80046-4

Mavica, L. W., & Barenholtz, E. (2013). Matching voice and face identity from static images. *Journal of Experimental Psychology: Human Perception and Performance*, *39*, 307–312. doi:10.1037/a0030945

Miller, G. F., & Todd, P. M. (1998). Mate choice turns cognitive. *Trends in Cognitive Sciences*, *2*, 190–198. doi:10.1016/S1364-6613(98)01169-3

Møller, A. P., & Pomiankowski, A. (1993). Why have birds got multiple sexual ornaments? *Behavioral Ecology and Sociobiology*, *32*, 167–176. doi:10.1007/BF00173774

Morrison, E. R., Gralewski, L., Campbell, N., & Penton-Voak, I. S. (2007). Facial movement varies by sex and is related to attractiveness. *Evolution and Human Behavior*, *28*, 186–192. doi:10.1016/j.evolhumbehav.2007.01.001

Neiman, G. S., & Applegate, J. A. (1990). Accuracy of listener judgments of perceived age relative to chronological age in adults. *Folia Phoniatrica et Logopaedica*, *42*, 327–330. doi:10.1159/000266090

O'Connor, J. J., Re, D. E., & Feinberg, D. R. (2011). Voice pitch influences perceptions of sexual infidelity. *Evolutionary Psychology*, *9*, 64–78. doi:10.1177/147470491100900109

Oguchi, T., & Kikuchi, H. (1997). Voice and interpersonal attraction. *Japanese Psychological Research*, *39*, 56–61. doi:10.1111/1468-5884.00037

Partan, S., & Marler, P. (1999). Communication goes multimodal. *Science*, *283*, 1272–1273. doi:10.1126/science.283.5406.1272

Peirce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics*, *2*, 1–8. doi:10.3389/neuro.11.010.2008

Penton-Voak, I., & Chang, H. (2008). Attractiveness judgements of individuals vary across emotional expression and movement conditions. *Journal of Evolutionary Psychology*, *6*, 89–100. doi:10.1556/JEP.2008.1011

Penton-Voak, I. S., & Chen, J. Y. (2004). High salivary testosterone is linked to masculine male facial appearance in humans. *Evolution and Human Behavior*, *25*, 229–241. doi:10.1016/j.evolhumbehav.2004.04.003

Perrett, D. I., Lee, K. J., Penton-Voak, I., Rowland, D., Yoshikawa, S., Burt, D. M., . . . Akamatsu, S. (1998). Effects of sexual dimorphism on facial attractiveness. *Nature*, *394*, 884–887. doi:10.1038/29772

Pisanski, K., Mishra, S., & Rendall, D. (2012). The evolved psychology of voice: Evaluating interrelationships in listeners' assessments of the size, masculinity, and attractiveness of unseen speakers. *Evolution and Human Behavior*, *33*, 509–519. doi:10.1016/j.evolhumbehav.2012.01.004

Ptacek, P. H., & Sander, E. K. (1966). Age recognition from voice. *Journal of Speech & Hearing Research*, *9*, 273–277. doi:10.1044/jshr.0902.273

Puts, D. A., Jones, B. C., & DeBruine, L. M. (2012). Sexual selection on human faces and voices. *Journal of Sex Research*, *49*, 227–243. doi:10.1080/00224499.2012.658924

Re, D. E., Hunter, D. W., Coetzee, V., Tiddeman, B. P., Xiao, D., DeBruine, L. M., . . . Perrett, D. I. (2013). Looking like a leader–facial shape predicts perceived height and leadership ability. *PloS one*, *8*, e80957. doi:10.1371/journal.pone.0080957

Rhodes, G., Chan, J., Zebrowitz, L. A., & Simmons, L. W. (2003). Does sexual dimorphism in human faces signal health? *Proceedings of the Royal Society of London B: Biological Sciences*, *270*, S93–S95. doi:10.1098/rsbl.2003.0023

Roberts, S. C., Little, A. C., Lyndon, A., Roberts, J., Havlicek, J., & Wright, R. L. (2009a). Manipulation of body odour alters men's self-confidence and judgements of their visual attractiveness by women. *International Journal of Cosmetic Science*, *31*, 47–54. doi:10.1111/j.1468-2494.2008.00477.x

Roberts, S. C., Saxton, T. K., Murray, A. K., Burriss, R. P., Rowland, H. M., & Little, A. C. (2009b). Static and dynamic facial images cue similar attractiveness judgements. *Ethology*, *115*, 588–595. doi:10.1556/JEP.7.2009.1.4.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, *15*, 351–357. doi:10.2307/2087176

Rubenstein, A. J. (2005). Variation in perceived attractiveness: Differences between dynamic and static faces. *Psychological Science*, *16*, 759–762. doi:10.1111/j.1467-9280.2005.01610.x

Saxton, T. K., Caryl, P. G., & Roberts, C. S. (2006). Vocal and facial attractiveness judgments of children, adolescents and adults: The ontogeny of mate choice. *Ethology*, *112*, 1179–1185. doi:10.1111/j.1439-0310.2006.01278.x

Schneider, T. M., Hecht, H., Stevanov, J., & Carbon, C. C. (2013). Cross-ethnic assessment of body weight and height on the basis of faces. *Personality and Individual Differences*, *55*, 356–360. doi:10.1016/j.paid.2013.03.022

Smith, H. M. J., & Baguley, T. (2014). Unfamiliar voice identification: Effect of post-event information on accuracy and voice ratings. *Journal of European Psychology Students*, *5*, 59–68. doi:10.5334/jeps.bs

Stevenage, S. V., Hugill, A. R., & Lewis, H. G. (2012). Integrating voice recognition into models of person perception. *Journal of Cognitive Psychology*, *24*, 409–419. doi:10.1080/20445911.2011.642859

Stevenage, S. V., & Neil, G. J. (2014). Hearing faces and seeing voices: The integration and interaction of face and voice processing. *Psychologica Belgica*, *54*, 266–281. doi:10.5334/pb.ar

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*, 284–294. doi:10.1080/09658211.2013.781654

Thornhill, R., & Gangestad, S. W. (1999). Facial attractiveness. *Trends in Cognitive Sciences*, *3*, 452–460. doi:10.1016/S1364-6613(99)01403-5

Thornhill, R., & Gangestad, S. W. (2006). Facial sexual dimorphism, developmental stability, and susceptibility to disease in men and women. *Evolution and Human Behavior*, *27*, 131–144. doi:10.1016/j.evolhumbehav.2005.06.001

Thornhill, R., & Grammer, K. (1999). The body and face of woman: One ornament that signals quality? *Evolution and Human Behavior*, *20*, 105–120. doi:10.1016/S1090-5138(98)00044-0

Urbaniak, G. C., & Plous, S. (2013). *Research randomizer* (Version 4.0) [Computer software]. Retrieved from http://www.randomizer.org/

Valentine, T., Lewis, M. B., & Hills, P. J. (2015). Face-space: A unifying concept in face recognition research. *The Quarterly Journal of Experimental Psychology*, 1–24. doi:10.1080/17470218.2014.990392

van Dommelen, W. A., & Moxness, B. H. (1995). Acoustic parameters in speaker height and weight identification: Sex-specific behaviour. *Language and Speech*, *38*, 267–287. doi:10.1177/002383099503800304

Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin*, *25*, 1115–1125. doi:10.1177/01461672992512005

Wells, T., Baguley, T. S., Sergeant, M. J. T., & Dunn, A. K. (2013). Perceptions of human attractiveness comprising face and voice cues. *Archives of Sexual Behavior*, *42*, 805–811. doi:10.1007/s10508-012-0054-0

Wells, T., Dunn, A. K., Sergeant, M. J. T., & Davies, M. N. O. (2009). Multiple signals in human mate selection: A review and framework for integrating facial and vocal signals. *Journal of Evolutionary Psychology*, *7*, 111–139. doi:10.1556/JEP.7.2009.2.2

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*, 2020–2045. doi:10.1037/xge0000014

Wheatley, J. R., Apicella, C. A., Burriss, R. P., Cárdenas, R. A., Bailey, D. H., Welling, L. L., & Puts, D. A. (2014). Women's faces and voices are cues to reproductive potential in industrial and forager societies. *Evolution and Human Behavior*, *35*, 264–271. doi:10.1016/j.evolhumbehav.2014.02.006

Zahavi, A., & Zahavi, A. (1997). *The handicap principle*. New York, NY: Oxford University Press.

APPENDIX E: PUBLISHED ARTICLE: Smith, Dunn, Baguley & Stacey, (2016b)

Atten Percept Psychophys (2016) 78:868–879
DOI 10.3758/s13414-015-1045-8

Matching novel face and voice identity using static and dynamic facial images

Harriet M. J. Smith[1,2] · Andrew K. Dunn[1] · Thom Baguley[1] · Paula C. Stacey[1]

Published online: 5 January 2016

Abstract Research investigating whether faces and voices share common source identity information has offered contradictory results. Accurate face–voice matching is consistently above chance when the facial stimuli are dynamic, but not when the facial stimuli are static. We tested whether procedural differences might help to account for the previous inconsistencies. In Experiment 1, participants completed a sequential two-alternative forced choice matching task. They either heard a voice and then saw two faces or saw a face and then heard two voices. Face–voice matching was above chance when the facial stimuli were dynamic and articulating, but not when they were static. In Experiment 2, we tested whether matching was more accurate when faces and voices were presented simultaneously. The participants saw two face–voice combinations, presented one after the other. They had to decide which combination was the same identity. As in Experiment 1, only dynamic face–voice matching was above chance. In Experiment 3, participants heard a voice and then saw two static faces presented simultaneously. With this procedure, static face–voice matching was above chance. The overall results, analyzed using multilevel modeling, showed that voices and dynamic articulating faces, as well as voices and static faces, share concordant source identity information. It seems, therefore, that above-chance static face–voice matching is sensitive to the experimental procedure employed. In addition, the inconsistencies in previous research might depend on the specific stimulus sets used; our multilevel modeling analyses show that some people look and sound more similar than others.

Keywords Static · Dynamic · Face · Voice · Crossmodal matching

Redundant information offered by faces and voices facilitates everyday social communication (Campanella & Belin, 2007). Testing whether novel (and therefore unfamiliar) faces and voices can be accurately matched provides a measure of the extent to which faces and voices offer redundant source identity information. Although some research has suggested that crossmodal matching of novel faces and voices is only possible when dynamic visual information about articulatory patterns is available (Kamachi, Hill, Lander, & Vatikiotis-Bateson, 2003; Lachs & Pisoni, 2004a), other research has suggested that it is possible to match static faces to voices because they offer concordant source identity information (Krauss, Freyberg, & Morsella, 2002; Mavica & Barenholtz, 2013; Smith, Dunn, Baguley, & Stacey, 2015). We tested whether differences between the experimental procedures across previous studies might account for these apparently inconsistent results.

A crucial role for dynamic visual articulatory patterns?

Idiosyncratic speaking styles dictate what voices sound like and how faces move (Lander, Hill, Kamachi, & Vatikiotis-Bateson, 2007; Yehia, Rubin, & Vatikiotis-Bateson, 1998). Audiovisual speech perception researchers have emphasized the existence of links between auditory and visual sensory

✉ Harriet M. J. Smith
harriet.smith2011@my.ntu.ac.uk

[1] Nottingham Trent University, Nottingham, UK

[2] Psychology Division, Nottingham Trent University, Burton Street, Nottingham NG1 4BU, UK

modalities (e.g., Kuhl & Meltzoff, 1984; MacDonald & McGurk, 1978; McGurk & MacDonald, 1976) and have demonstrated that participants can match sequentially presented dynamic images of articulating faces to speakers (Lachs & Pisoni, 2004a), even when the voice and face are producing different sentences (Kamachi et al., 2003; Lander et al., 2007). The conclusion that crossmodal source identity information is contingent on encoding dynamic visual articulatory patterns has been supported by studies finding that static face–voice matching performance is at chance level (Kamachi et al., 2003; Lachs & Pisoni, 2004a). The importance of time-varying articulatory information is underlined by the fact that participants can match faces and voices using movement information alone. Studies isolating articulatory movement using a point-light technique have produced accurate matching of utterances to dynamic displays (Lachs & Pisoni, 2004b; Rosenblum, Smith, Nichols, Hale, & Lee, 2006).

Other research challenges the conclusion that dynamic visual information is crucial to crossmodal matching. Krauss et al. (2002) showed that people could match a voice to one of two full-length static images of different people with above-chance accuracy. Whereas the studies observing chance-level matching performance using static faces and voices used stimuli of a similar age, gender, and ethnicity in each trial (e.g., Kamachi et al., 2003), Krauss et al.'s stimuli were from a wider age range (20–60 years). The stimuli were also full-length images rather than images of faces, which may have provided additional cues to inform accurate matching. However, Mavica and Barenholtz (2013) replicated Krauss et al.'s results using static headshots of age-matched stimuli, and face–voice matching was above chance in both of the experiments they reported. Similarly, Smith et al. (2015) also observed above-chance static face–voice matching. These three studies offer growing evidence that the source identity information available in static faces overlaps with the information offered by voices.

Concordant information in faces and voices

In light of research investigating the extent to which faces and voices offer similar information about personal characteristics, above-chance static face–voice matching makes intuitive sense. Studies testing the concordance between ratings of attractiveness from static faces and voices suggest that both validly signal genetic quality (Collins & Missing, 2003; Feinberg et al., 2005; Saxton, Caryl, & Roberts, 2006; T. Wells, Baguley, Sergeant, & Dunn, 2013). Hormone levels are reflected in both faces (Penton-Voak & Chen, 2004; Perrett et al., 1998; Thornhill & Grammer, 1999) and voices (Abitbol, Abitbol, & Abitbol, 1999; Beckford, Rood, & Schaid, 1985; O'Connor, Re, & Feinberg, 2011; Pisanski, Mishra, & Rendall, 2012). A man who sounds masculine should therefore also tend to look masculine, and similarly, feminine-sounding women should tend to look feminine. In a recent study, Smith et al. (2015) asked participants to complete a number of rating scales for faces and corresponding voices. Faces and voices were presented in two separate blocks. The results showed that independent judgments about femininity and masculinity made from faces and voices were strongly and positively correlated. Positive correlations were also found between face and voice ratings of age, health, height, and weight (Smith et al., 2015). Interestingly, the strength of correlations did not vary according to whether the faces were static or dynamic. These results suggest that static face–voice matching is possible (Krauss et al., 2002; Mavica & Barenholtz, 2013; Smith et al., 2015) because faces do not need to be dynamic in order to share concordant information with voices.

Procedural differences between studies

Procedural differences between studies may account for some of the apparently contradictory results outlined above. Audiovisual speech perception studies (e.g., Kamachi et al., 2003; Lachs & Pisoni, 2004a, b; Lander et al., 2007), have tended to use a "crossmodal matching task" (Lachs, 1999). This is a sequential two-alternative forced choice (2AFC) procedure. In the visual to auditory (V–A) condition, a face is shown and then two voices are presented at test, one after the other. In the auditory to visual (A–V) condition, this procedure is reversed: Participants hear a voice and then see two sequentially presented faces at test. At test, one of the alternatives is therefore always the same-identity target, whereas the other is a different-identity distractor. The participant must decide which of the two alternatives matches the identity of the other-modality stimulus. Studies that have used this procedure have generally emphasized the importance of dynamic articulatory information in facilitating face–voice matching; above-chance face–voice matching is typically found for dynamic but not for static faces (Kamachi et al., 2003; Lachs & Pisoni, 2004a, b; Lander et al., 2007). In contrast, the majority of experiments observing above-chance levels of matching accuracy using static facial stimuli have not used this exact procedure, making it unwise to directly compare the results. For instance, Krauss et al. (2002) presented a voice followed by two simultaneously presented full-length images. Smith et al. (2015) used a same–different procedure in which participants saw a face and heard a voice, and then had to decide whether or not the face and voice shared the same identity. Mavica and Barenholtz's (2013) stimuli (one voice and two test faces) were presented simultaneously in Experiment 1. However, it is important to note that Mavica and Barenholtz's second experiment replicated above-chance-level matching with static facial stimuli using the A–V condition of the standard crossmodal matching task (Lachs, 1999). Although the V–A condition was not included, this result hints that even if procedural differences across studies hold some explanatory value, additional factors may also affect performance and help to explain the existing contradictions. Nevertheless, the impact of procedural differences on face–voice matching accuracy deserves further attention.

A possible explanation for the differences in face–voice matching between static and dynamic stimuli is associated with memory demands. Some research has suggested that memory for dynamic facial images is better than that for static facial images (e.g., Christie & Bruce, 1998; Knappmeyer, Thornton, & Bülthoff, 2003; Lander & Chuang, 2005). In a review, O'Toole, Roark, and Abdi (2002) put forward two explanations for this increased memorability. According to the "representation enhancement hypothesis," dynamic images facilitate the perception of 3-D facial structure. In the "supplemental information hypothesis," motion is thought to provide additional signature information about the given person. Therefore, when stimuli are presented sequentially (as in a crossmodal matching task), poorer memory for static images could make it harder for participants to hold the face in working memory long enough to compare with the voice for source identity information. In an attempt to rule out memory explanations for the results of their first experiment, which detected above-chance static face–voice matching, Mavica and Barenholtz (2013) used sequential presentation in their Experiment 2. Their results did not entirely rule out an explanation for the discrepancies across studies based on memory effects. In neither experiment did Mavica and Barenholtz include a dynamic face–voice matching condition. If memory load affects performance, we might expect to find a position effect in a 2AFC task, whereby accuracy is higher if the correct other-modality stimulus appears in Position 1 rather than Position 2. Previous studies have not included analyses of responses by position, and thus the impact of this factor is unknown, although position effects for 2AFC tasks are well-documented in the literature (García-Pérez & Alcalá-Quintana, 2011; Yeshurun, Carrasco, & Maloney, 2008).

Failure to include both static and dynamic face conditions therefore prevents a direct comparison of crossmodal matching explanations based on static facial information (e.g., Krauss et al., 2002; Mavica & Barenholtz, 2013) with those focusing on dynamic facial information (e.g., Kamachi et al., 2003; Lachs & Pisoni, 2004a, b; Lander et al., 2007; Rosenblum et al., 2006). To date, only one study has directly compared matching performance using static and dynamic facial stimuli in the same experiment, and it found no difference in matching accuracy across conditions (Smith et al., 2015). Further clarification of these results using a crossmodal matching procedure will be necessary. However, as has been suggested by other results (Kamachi et al., 2003; Lachs & Pisoni, 2004a), it is feasible that participants tested using dynamic facial stimuli may significantly outperform those in static conditions because dynamic stimuli make both temporal and spatial information available to inform matching decisions.

Aims

In the face of these contradictory results, in the experiments presented here we aimed to clarify whether static face–voice matching is possible using stimuli of the same age, sex, and ethnicity. In an attempt to tease apart the relative contributions of static and dynamic face information in facilitating crossmodal matching, performance using static and dynamic faces was compared in both Experiments 1 and 2. In case better memory for dynamic facial stimuli affects matching accuracy, memory load was varied across the experiments: In Experiment 1, all stimuli were presented sequentially, so memory load was higher, whereas in Experiment 2, face–voice combinations were presented simultaneously. In a further test of whether static face–voice matching is sensitive to procedural differences, for Experiment 3 we adopted the procedure of Krauss et al. (2002), in which the alternatives in a 2AFC task are presented simultaneously. To clarify how memory load and task type affect the results, in all three experiments we also investigated whether accuracy is higher when the correct, matching other-modality stimulus appears in Position 1 rather than Position 2.

Experiment 1

In Experiment 1 we used a standard crossmodal matching task (Lachs, 1999) to compare static and dynamic face–voice matching. In most experiments in which this procedure has been used, the results have shown only dynamic face–voice matching to be above chance level (Kamachi et al. 2003; Lachs & Pisoni, 2004a; Lander et al., 2007; cf. Mavica & Barenholtz, 2013, Exp. 2). Informed by the balance of evidence, we expected static face–voice matching to be at chance level.

Method

Design Experiment 1 employed a 2 × 2 × 2 mixed factorial design. The between-subjects factor was Facial Stimulus Type (static or dynamic), and the within-subjects factors were Order (visual then auditory [V–A] or auditory then visual [A–V]) and Position (1 or 2). The dependent variable was matching accuracy.

Participants The participants (N = 82) were recruited from the Nottingham Trent University Psychology Division's

Research Participation Scheme by convenience sampling. A total of 26 male and 56 female participants took part (age range = 18 to 66 years, M = 23.70, SD = 8.56). All participants reported having normal or corrected vision and hearing. In line with course requirements, student participants received three research credits. Ethical approval for this and subsequent experiments was granted by the university's BLSS (Business, Law, and Social Science) College Research Ethics Committee.

Apparatus and materials The stimuli were taken from the GRID audiovisual sentence corpus (Cooke, Barker, Cunningham, & Shao, 2006). The corpus features head and shoulder videos of British adults recorded against a plain background saying six-word sentences in an emotionally neutral manner. Each sentence follows the same structure: (1) command, (2) color, (3) preposition, (4) letter, (5) digit, and (6) adverb—for example, *Place red at F2 please*. A total of 18 speakers were selected from the corpus: nine male and nine female. All of the speakers were between 18 and 30 years of age and were white British with an English accent.

The stimuli were presented on an Acer Aspire laptop (screen size = 15.6 in., resolution = 1,366 × 768 pixels, Dolby Advanced Audio), with brightness set to the maximum level. The experiment ran on PsychoPy version 1.77.01 (Peirce, 2009), an open-source software package for running experiments in Python. The study used the same static faces, dynamic faces, and voices as Smith et al. (2015). Three .mpeg-format videos were randomly selected from the GRID corpus for each of the 18 speakers. The videos were selected using an online research randomizer (Urbaniak & Plous, 2013). One of the three videos was used to create static pictures of faces (.png format). The static picture for each talker was the first frame of the video. Another of the three video files was used to construct the dynamic stimuli by muting the sound. Facial stimuli measured 384 × 288 pixels and were presented for 2 s, in color. Voice recordings were also played for 2 s. To reduce background noise, participants listened to the recordings binaurally through Apple EarPods at a comfortable listening volume (30 % of the maximum). Apple EarPods have a frequency range of 5 to 21000 Hz. This is wider than the normal range of human hearing (Feinberg et al. 2005.

Four versions of the experiment were created, so that trials could be constructed using different combinations of stimuli. Each version consisted of 12 trials in total, and each trial featured three stimuli. In the V–A condition, a face (Stimulus 1) was followed by two sequentially presented voices (Stimuli 2 and 3): a target and a distractor. In the A–V condition, a voice (Stimulus 1) was followed by sequentially presented target and distractor faces (Stimuli 2 and 3). Across versions, whether someone's face/voice appeared as Stimulus 1, 2, or 3, and whether it was used in a V–A or A–V trial, was randomly varied. The position of the same-identity other-modality stimulus at test (Position 1 or 2) was also randomly and equally varied. None of the faces or voices appeared more than once in each experimental version. Each of the four versions was used for the between-subjects manipulation of facial stimuli (static or dynamic), so in total there were eight versions of the experiment.

Procedure The participants were randomly allocated to one of the eight versions of the experiment using an online research randomizer (Urbaniak & Plous, 2013). In the dynamic facial stimulus condition, participants were accurately informed that the face and the voice were saying different sentences, to prevent the use of speech-reading (Kamachi et al. 2003.

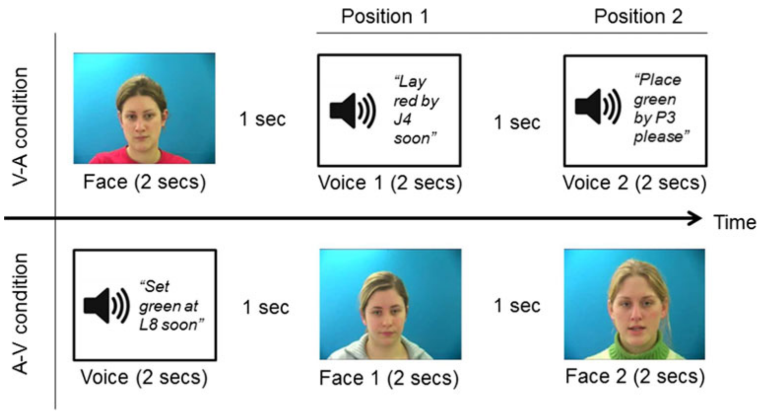
The participants completed two counterbalanced experimental blocks. The procedure is illustrated in Fig. 1. First, participants received a practice trial, followed by six randomly ordered trials. In one block of trials, participants saw a face first. After a 1-s gap, they heard the first voice. The text "Voice 1" was visible in the middle of the screen while the recording was playing. After another 1-s gap, they heard the second voice, with the text "Voice 2" visible in the middle of the screen. In the other block of trials, participants heard a voice first, and then saw two faces, presented one after the other. Gaps of 1 s were inserted between all stimuli, and the text "Face 1" or "Face 2" appeared below each picture. At test, participants were asked to select either "1" or "2" as the face/voice that had the same identity as the first stimulus.

Data analysis and multilevel modeling All data were analyzed using multilevel models so that both participants and stimuli could be treated as random effects. The random effects were fully crossed; every participant encountered all 36 stimuli (18 faces, 18 voices) in each version of the experiment. Multilevel modeling avoids aggregating data (see Smith et al. 2015; Wells et al. 2013) and inflating the risk of Type I error (Baguley, 2012; Clark, 1973; Judd, Westfall, & Kenny, 2012). Accordingly, multilevel modeling was the most appropriate analysis, because it takes into account the variability associated with individual performance and different stimuli. The variance associated with stimuli may be particularly important when investigating face–voice matching. Mavica and Barenholtz (2013) reported that matching performance varied between 35 % and 70 % for the 64 models whose faces and voices they used as stimuli. Disregarding this source of variance would risk the ecological fallacy (see Robinson, 1950), by falsely assuming that the observed patterns for participant means also occur at the level of individual trials.

Results

Matching accuracy was analyzed using multilevel logistic regression with the lme4, version 1.06, package in R (Bates, Maechler, Bolker, & Walker, 2014). This is the same method

Fig. 1 The procedure used in Experiment 1



of analysis used in Smith et al. (2015). Four nested models were compared, all fitted using restricted maximum likelihood, and with accuracy (0 or 1) as the dependent variable. The first model included a single intercept; the second included the main effects of each factor (Order, Position, and Facial Stimulus Type). The third added the two-way interactions, and the final model included the three-way interaction. This method of analysis allowed us to test for individual effects in a way similar to traditional analysis of variance (ANOVA). However, as *F* tests derived from multilevel models tend not to be accurate, we report the likelihood ratio tests provided by lme4. These are more robust and are obtained by dropping each effect in turn from the appropriate model (e.g., testing the three-way interaction by dropping it from the model including all effects, and testing the two-way interactions by dropping each effect in turn from the two-way model).

Table 1 shows the likelihood chi-square statistic (G^2) and *p* value associated with dropping each effect. Table 1 also reports the coefficients and standard errors (on a log odds scale) for each effect in the full three-way interaction model. Variability for the first stimulus in each trial (the voice in the A–V condition, and the face in the V–A condition) was modeled separately from the foil stimulus. The random effect for the first stimuli captures the variability of both faces and voices, because corresponding faces and voices are highly correlated. For foils we modeled separate random effects for faces and voices, because the corresponding voice or face was never present. In the three-way model, the estimated *SD* of the first-stimulus random effect was .535; for the voice foils it was .634; and for face foils it was .484. The estimated *SD* for the participant effect was less than .0001. A similar pattern held for the null model. Thus, although individual differences were negligible in this instance, a conventional by-participants analysis that did not simultaneously incorporate the variance associated with the stimuli could be extremely misleading.

The main effect of position was significant, along with the three-way interaction between position, order, and facial stimulus type. Figure 2 aids interpretation of the effects and interaction, showing means and 95 % confidence intervals for the percentage accuracies in each condition of the factorial design. The confidence intervals were obtained by simulating the posterior distributions of the cell means in R (arm package, version 1.6; Gelman & Su, 2013).

Overall, matching performance was significantly above the chance (50 %) level, *M* = 59.7 %, 95 % CI [50.8, 68.0]. However, the confidence intervals for percentage accuracy in the static (*M* = 57.6 %, 95 % CI [47.5, 67.1]) and dynamic (*M* = 63.7 %, 95 % CI [53.8, 72.5]) conditions show that only performance on dynamic facial stimulus trials was significantly above chance level. Figure 2 shows the main effect of position, with accuracy levels being consistently higher when the correct, matching other-modality stimulus was presented in Position 1 than when it was presented in Position 2. The results from the V–A condition are shown in panel A, whereas results from the A–V condition appear in panel B. The basis of

Table 1 Parameter estimates (*b*) and likelihood ratio tests for the 2 × 2 × 2 factorial analysis, Experiment 1: Sequential face–voice presentation

Source	*df*	*b*	*SE*	G^2	*p*
Intercept	1	0.444	0.315	–	–
Position	1	0.062	0.374	5.92	.015
Order	1	0.333	0.371	0.68	.410
Facial Stimulus Type	1	0.676	0.277	3.42	.064
Position × Order	1	0.870	0.516	0.35	.553
Position × Facial Stimulus Type	1	0.625	0.390	0.02	.884
Order × Facial Stimulus Type	1	0.775	0.382	0.59	.441
Position × Order × Facial Stimulus Type	1	1.159	0.549	4.34	.037

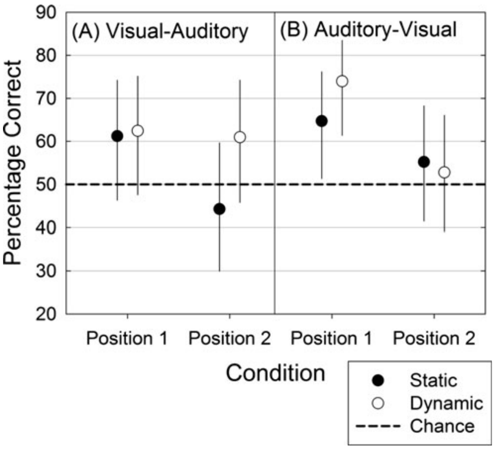


Fig. 2 Face–voice matching accuracy on visual–auditory (panel A) and auditory–visual (panel B) trials for sequentially presented faces and voices in a two-alternative forced choice task. Error bars show 95 % confidence intervals for the condition means

the three-way interaction appears to relate to performance when the matching other-modality stimulus appears in Position 2 in the V–A condition. In that condition there was no position effect in the dynamic facial stimulus condition. However, as with any factorial design testing multiple effects, it would be imprudent to overinterpret a single nonpredicted interaction that is only just statistically significant (p = .037).

Discussion

Using the standard crossmodal matching task (Lachs, 1999) employed in audiovisual speech perception research, in Experiment 1 we observed above-chance dynamic face–voice matching, but chance-level static face–voice matching. Although there was no significant difference between static and dynamic face–voice matching accuracy, and although static face–voice matching was close to being above chance level, this pattern of results appears to support the conclusion that the source identity information shared by dynamic articulating faces and voices explains accurate face–voice matching. The results are consistent with those of two previous studies (Kamachi et al. 2003; Lachs & Pisoni, 2004a), but are in conflict with Mavica and Barenholtz (2013, Exp. 2), who observed above-chance-level static face–voice matching using this procedure.

The presence of a position effect in Experiment 1 additionally suggests that memory load might be hindering performance, especially in the static facial stimulus condition. Matching was more accurate when the matching face and voice were presented close together in time (Position 1) than when the matching other-modality stimulus was further away, in Position 2. In line with research suggesting that memory is better for dynamic than for static faces (Christie & Bruce, 1998; Knappmeyer et al. 2003), the position effect did not manifest in the dynamic facial stimulus, V–A condition. This is the condition in which the face (Stimulus 1) would need to be held in memory for the longest time.

Experiment 2

In order to clarify the effect of procedural differences across previous studies, in Experiment 2 we used a modified version of the presentation procedure from Experiment 1. Experiment 2 presented two different face–voice combinations. This time, the face and voice in each combination were presented simultaneously, instead of sequentially. By reducing the memory load, we hypothesized that matching accuracy might be higher when faces and voices were presented simultaneously, and above chance for static face–voice matching.

Method

The methods for Experiment 2 were identical to those of Experiment 1, with the exceptions outlined below.

Participants Seven male and 33 female adult participants (N = 40) took part in the experiment, with an age range of 18 to 33 years (M = 21.38, SD = 3.57). None of the participants had taken part in Experiment 1.

Procedure The procedure used in Experiment 2 is illustrated in Fig. 3. Participants in the V–A condition saw a face accompanied by a recording of a voice. The text "Voice 1" was visible underneath the face. After a 1-s gap, they saw the same face accompanied by a different voice, and the text "Voice 2" appeared beneath the face. In the A–V condition, participants heard a voice accompanied by a face, then a 1-s intervening gap, before hearing the same voice accompanied by a different face. The text "Face 1" and "Face 2" appeared below the first and second combinations, respectively. Participants had to decide which combination was correct by pressing "1" for face–voice Combination 1, or "2" for face–voice Combination 2.

Results

Face–voice matching accuracy was analyzed using the same method as in Experiment 1. Table 2 shows the likelihood chi-square statistic (G^2) and p value associated with dropping each effect in turn from the appropriate model. The coefficients and standard error (on a log odds scale) for each effect in the full three-way interaction model are also reported in Table 2. We observed a similar pattern of SDs for the random effects. In the

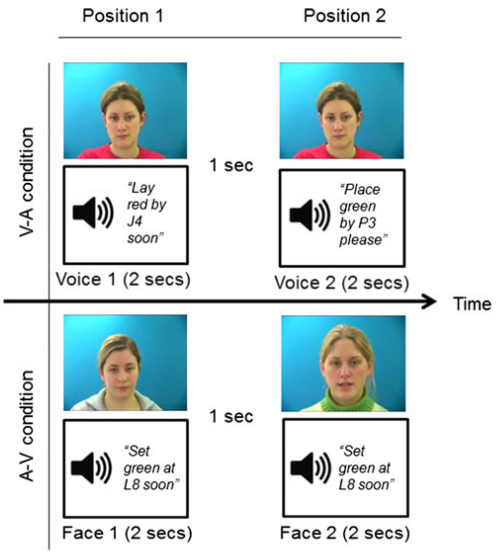


Fig. 3 Procedure used in Experiment 3

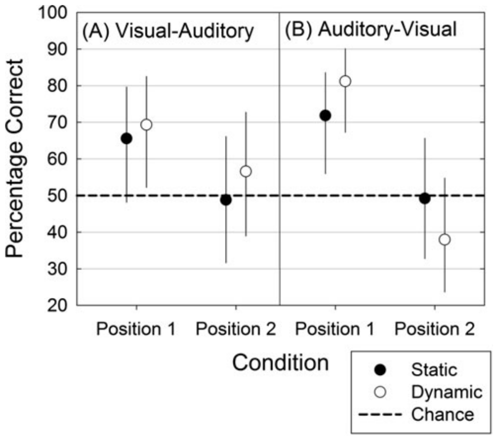


Fig. 4 Face–voice matching accuracy on visual–auditory (panel A) and auditory–visual (panel B) trials for simultaneously presented faces and voices in a two-alternative forced choice task. Error bars show 95 % confidence intervals for the condition means.

three-way model, the estimated *SD* of the first-stimulus random effect was .778; for the voice foils it was .324; and for the face foils it was .103. The estimated *SD* for the participant effect was .007.

Only the main effect of position was significant. Figure 4 aids interpretation of this main effect, showing the means and 95 % confidence intervals for accuracy in each of the eight conditions, obtained using the arm package (version 1.6; Gelman & Su, 2013).

As in Experiment 1, the overall matching performance was significantly above chance (50 %) level, *M* = 60.9 %, 95 % CI [50.4, 70.5]. Dynamic facial stimulus trials overall were significantly above chance (*M* = 62.5 %, 95 % CI [50.1, 73.6]), but static facial stimulus trials were not (*M* = 59.8 %, 95 % CI [47.2, 71.2]). As is clear from Fig. 4, the main effect of position exhibits the same pattern as in Experiment 1, with accuracy levels being consistently higher when the correct face–voice combination is presented in Position 1. There is, however, no three-way interaction.

Discussion

Overall, the pattern of results observed in Experiment 2 is largely similar to that observed in Experiment 1, when all of the stimuli were presented sequentially. The participants in Experiment 2 exhibited a bias toward selecting the first face–voice combination they encountered. As the position effect was observed in both experiments, this may be attributable to the nature of the 2AFC task: When alternatives are presented sequentially, the first alternative is disproportionately favored. Indeed, as we noted in the introduction, other studies have shown widespread evidence of position biases using 2AFC procedures (García-Pérez & Alcalá-Quintana, 2011; Yeshurun et al. 2008). No three-way interaction was detected in Experiment 2. Thus, although the position effect may vary in strength depending on stimulus type and order, the two

Table 2 Parameter estimates (*b*) and likelihood ratio tests for the 2 × 2 × 2 factorial analysis, Experiment 2: Simultaneous face–voice presentation

Source	*df*	*b*	*SE*	G^2	*p*
Intercept	1	0.266	0.365	–	–
Position	1	0.550	0.462	17.40	<.001
Order	1	0.755	0.431	<0.01	.952
Facial Stimulus Type	1	0.314	0.391	0.37	.545
Position × Order	1	1.402	0.653	1.95	.162
Position × Facial Stimulus Type	1	0.140	0.568	1.09	.295
Order × Facial Stimulus Type	1	0.771	0.549	0.37	.544
Position × Order × Facial Stimulus Type	1	1.121	0.804	1.90	.169

experiments presented here do not provide compelling evidence for this conclusion.

Experiment 3

The results from Experiment 2 showed that simultaneously presenting faces and voices does not improve static face–voice matching. This was contrary to what we expected; it seems that the pattern of results from Experiment 1 was not attributable to increased memory load impairing the comparison of the first stimulus to the matching other-modality stimulus in Position 2. In Experiment 3, we aimed to test whether chance-level static face–voice matching could be attributable to the sequential presentation of alternatives in a 2AFC task. Evidence from the forensic eyewitness literature suggests that simultaneously presenting faces in a lineup array produces a different pattern of results than when faces are presented sequentially (Clark, Howell, & Davey, 2008; Ebbesen & Flowe, 2002; Steblay, Dysart, & Wells, 2011). This possibly occurs because of the differential use of relative and absolute judgments (Kneller, Memon, & Stevenage, 2001). Relative judgments (G. L. Wells, 1984) are employed when choosing the best option from simultaneously presented alternatives, whereas the sequential presentation of alternatives encourages absolute judgments because of the difficulty of making comparisons (G. L. Wells et al. 1998).

Some previous experiments finding above-chance accuracy with static stimuli have used a procedure in which the test alternatives were presented simultaneously, and can therefore be compared more easily (Krauss et al., 2002; Mavica & Barenholtz, 2013, Exp. 1). Experiment 3 tested whether static face–voice matching is above chance level when the alternatives in a 2AFC task are presented simultaneously. Because of the nature of this procedure, and the difficulty of presenting voices simultaneously at test, Experiment 3 only included an A–V condition. Although we did not expect a spatial position effect to manifest when the two face alternatives were presented simultaneously, we were unsure (in face of the contradictory previous research) whether this procedure would elicit above-chance static face–voice matching.

Methods

Design For Experiment 3, we employed a within-subjects design, with one factor: Spatial Position (left = Position 1, or right = Position 2). The dependent variable was matching accuracy.

Participants Eight male and 22 female adult participants (N = 30) took part, with an age range of 18 to 44 years (M = 20.70, SD = 5.20). The participants were recruited in the same way as in Experiments 1 and 2, although none had taken part in previous experiments. All participants reported having normal or corrected vision and hearing.

Apparatus and materials The software and equipment used in Experiments 1 and 2 were also used in Experiment 3. The voice stimuli and static facial stimuli were also the same as in the previous experiments. In the absence of a between-subjects manipulation, only four versions of Experiment 3 were constructed, all of which featured different combinations of stimuli. Each version featured one block of 18 trials, in which a voice was followed by the presentation of two faces. The same-identity face was always present at test, with its spatial position (left = Position 1 or right = Position 2) being randomly and equally varied. Each voice was only heard once in each version. Each of the stimulus faces appeared twice, but only once as the correct, matching stimulus. This was in keeping with the procedure of Krauss et al. (2002), who also reused the visual stimuli as foils within blocks.

Procedure The participants were randomly allocated to one of the four experimental versions using an online research randomizer (Urbaniak & Plous, 2013). As is illustrated in Fig. 5, participants heard a voice for 2 s. After a 1-s gap, they saw two images of faces presented side by side. The text “Face 1” was visible underneath the face on the left, and the text “Face 2” appeared underneath the face on the right. This screen was visible for 2 s. Participants were then instructed to decide which face matched the voice they had heard, indicating their answer by pressing “1” for “Face 1” or “2” for “Face 2.”

Results

Face–voice matching accuracy was analyzed using the same method as in Experiments 1 and 2. Since there was only one within-subjects factor, we only report the likelihood chi-square statistic (G^2) and p value associated with dropping the main effect from the null model. The coefficients and standard error (on a log odds scale) for the effect of spatial position in the main effect model are reported in Table 3. In the main effect model, the estimated SD of the voice random effect was .487, and that for the face foil was .0002. The estimated SD for the participant effect was less than .0001.

The main effect of spatial position was nonsignificant, and the overall matching accuracy with simultaneously presented static facial stimuli was above chance level (50 %), M = 61.0 %, 95 % CI [54.1, 67.6].

Discussion

The results indicate that when test alternatives are presented simultaneously, static face–voice matching is above chance level. In keeping with the previous results (Mavica & Barenholtz, 2013; Smith et al., 2015), this confirms that static

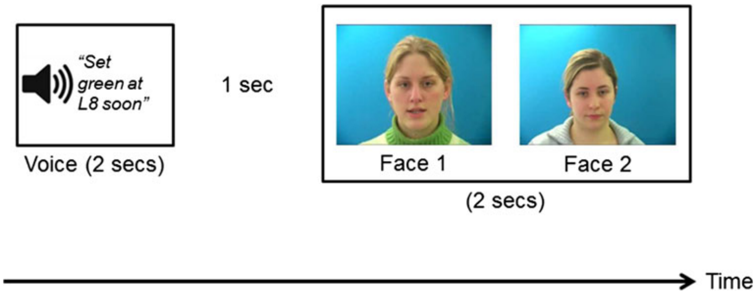


Fig. 5 Procedure used in Experiment 3

face–voice matching is possible. The results also replicate the findings of Krauss et al. (2002), but using headshots rather than full-length images. When we consider these alongside the results presented in Experiments 1 and 2, it appears that static face–voice matching performance is sensitive to procedure, thus offering one possible explanation for the contradictions between previous studies.

Experiments 1 and 2 showed that there is a temporal position bias when test options are presented sequentially. However, Experiment 3 suggests that there is no corresponding spatial position bias; when the test options are presented simultaneously, the position bias is negligible.

General discussion

In an attempt to resolve the discrepancies across previous face–voice matching studies, the three experiments presented here tested whether crossmodal source identity information is exclusively dependent on encoding visual articulatory patterns, or whether static faces and voices offer sufficient concordant information to facilitate above-chance performance. Taken together, the results are consistent with the conclusion that, although articulatory movement might be important in facilitating face–voice matching (Exps. 1 and 2), it is also possible to match static faces and voices when a 2AFC procedure facilitates comparisons between the alternatives (Exp. 3). Therefore, it seems that the procedural differences between previous studies offer a possible explanation for the discrepant results in the literature. Furthermore, as was shown by the variance associated with the stimuli in the multilevel modeling analysis, people vary in the extent to which they look and sound similar. This offers a complementary explanation for the contradictions in previous studies, because results may be highly dependent on the particular stimuli used.

Table 3 Parameter estimates (*b*) and likelihood ratio tests for the analysis, Experiment 3: Simultaneously presented alternatives

Source	*df*	*b*	*SE*	G^2	*p*
Intercept	1	0.446	0.147	–	–
Spatial Position	1	0.199	0.203	0.98	.329

Static versus dynamic face–voice matching

In Experiments 1 and 2, we presented the test alternatives in the 2AFC task sequentially. The results replicated those of audiovisual speech perception studies, showing that although dynamic faces and voices can be matched at a level significantly above chance, static faces and voices cannot (Kamachi et al., 2003; Lachs & Pisoni, 2004a). However, static face–voice matching was very close to being above chance level, and there was no significant difference between the facial stimulus conditions. These results hint at the existence of a trend toward accurate static face–voice matching across all three experiments. As was shown by the results of Experiment 3, and in keeping with the hypothesis that static faces and voices also offer concordant source identity information (Feinberg et al., 2005; Krauss et al., 2002; Mavica & Barenholtz, 2013; Saxton, Caryl, & Roberts, 2006; Smith et al., 2015), when the alternatives were presented simultaneously, performance was significantly above chance. The overall results are therefore not consistent with the conclusion that dynamic articulatory movement is exclusively responsible for explaining crossmodal matching (e.g., Kamachi et al., 2003; Lachs & Pisoni, 2004a), although they do not rule out the audiovisual speech perception argument that visual articulatory movement shares source identity information with voices (Kamachi et al., 2003; Lachs & Pisoni, 2004a, b; Rosenblum et al., 2006).

The lack of a statistical difference between static and dynamic face–voice matching in Experiments 1 and 2 corresponds with the results of previous findings using a same–different procedure (Smith et al., 2015). This warns against overstating the importance of visual articulatory movement in accounting for crossmodal matching accuracy. That said, the lack of an effect of facial stimulus type is not necessarily at odds with the results of studies that have detected accurate face–voice matching when movement was isolated using point-light displays and static information was unavailable (Lachs & Pisoni, 2004b; Rosenblum et al., 2006). Dynamic point-light displays could offer sufficient information to inform accurate face–voice matching, independently of the structural information available in static images.

Procedural differences

On both static and dynamic facial stimulus trials, we observed a uniform position effect in Experiment 2 when the memory load was reduced. This finding suggests that the discrepant pattern of results across previous studies is not a consequence of differential memory effects for static and dynamic faces. Rather, our findings are more consistent with the conclusion that the position effect is attributable to the nature of the 2AFC task (García-Pérez & Alcalá-Quintana, 2011; Yeshurun et al., 2008) when the two test alternatives are presented sequentially. In keeping with this argument, the position effect disappeared when the static alternatives were presented simultaneously, in Experiment 3.

Alternatively, the position effect might have manifested because faces and voices are most commonly perceived simultaneously during social interactions. Therefore, participants may have exhibited a bias to accept a face and voice presented in relative temporal proximity (Exp. 1) or the combination presented first (Exp. 2) as coming from the same person. This explanation would disproportionately support matching accuracy when the matching other-modality stimulus appears in Position 1, in line with the position bias observed in both Experiment 1 and 2.

In comparing the results of Experiments 1 and 2 to those of Experiment 3, it appears that static face–voice matching is sensitive to the procedure employed. The similarity of the results across Experiments 1 (sequential face–voice presentation) and 2 (simultaneous face–voice presentation) suggest that the contradictions between previous studies are not attributable to superior performance when faces and voices are presented simultaneously. This may occur because the more critical comparison to make in facilitating matching accuracy is between alternatives, rather than between the face and the voice. When the two alternatives are presented simultaneously, as in Experiment 3, the key comparison, a relative judgment (Wells, 1984), is easier to make.

At this point, it should be noted that in previous face–voice matching experiments using a crossmodal matching procedure, a standard interstimulus interval of 500 ms has been used (e.g., Lachs & Pisoni, 2004a, b; Mavica & Barenholtz, 2013), which is half as long as the interval featured in the experiments we report. With 1-s intervals in Experiment 1, we observed chance-level static face–voice matching when the stimuli were presented sequentially. Using 500-ms intervals, Mavica and Barenholtz (2013, Exp. 2) observed above-chance-level matching accuracy. It is necessary to consider the possible impact of this methodological dissimilarity. It could be argued that a longer interval might increase the load on auditory and visual sensory memory, making the task more difficult. The results that we report support the argument that sensory memory pressures do not account for the chance-level static facial stimulus results in Experiment 1. Experiment 2, in which faces and voices were presented simultaneously, was designed to alleviate memory load, and the results were very similar to those of Experiment 1: Static face–voice matching was still at chance level.

Variability associated with the stimuli

An explanation based on procedural differences does not accommodate all of the results in the previous literature. Mavica and Barenholtz (2013) observed above-chance static face–voice matching using sequential presentation of alternatives in the A–V condition of the standard crossmodal matching task (Lachs, 1999). Alongside procedural differences, our set of three experiments also highlights the importance of stimulus variability in providing an additional, but complementary, explanation for the contradictions between previous studies. Other studies have used varying numbers of face–voice pairs when testing crossmodal matching. For example, Lachs and Pisoni (2004a) used eight pairs of stimuli, but Kamachi et al. (2003) used 40. Our multilevel modeling analysis revealed that some people look and sound more similar than others; relatively high levels of variance associated with the stimuli were observed for the 18 face–voice pairs used here, and in all three experiments, the overall variance associated with stimuli was far greater than that associated with participants. Consistent with this, Mavica and Barenholtz reported that for their stimuli, levels of matching accuracy varied widely, between 35 % and 70 %, across 64 face–voice pairs. Overall, Mavica and Barenholtz's stimulus pairings of voices and static faces may have been easier to match than the pairings featured in our study, or than those featured in previous studies (Kamachi et al., 2003; Lachs & Pisoni, 2004a).

A key strength of the present research is our use of multilevel modeling. Although Mavica and Barenholtz (2013) ran a power analysis indicating that the discrepancies between previous studies were not due to lack of statistical power, simultaneously accounting for variance associated with stimuli and participants is a problem that can only be appropriately dealt with by running a multilevel model (Baguley, 2012; Judd et al., 2012). This statistical approach allows generalizations to be made across both stimuli and participants, and is generally more conservative than traditional analyses such as ANOVA, which aggregate over one or the other variable. However, multilevel modeling has not been previously used when investigating face–voice matching, reducing confidence in the generality of the findings in this field.

No order effects in 2AFC tasks

In line with other studies (Kamachi et al., 2003, forward and backward conditions; Lachs & Pisoni, 2004a; Lander et al., 2007), neither Experiment 1 nor 2 showed an effect of order. Although some asymmetries were found between V–A and A–V conditions in Smith et al.'s (2015) same–different procedure, the results suggested that these asymmetries were owing to a response bias on A–V trials. We would not expect such an effect to manifest in a 2AFC paradigm, which tests sensitivity rather than response bias.

Conclusion

The results of the three experiments reported here suggest that source identity is shared by dynamic articulating faces and voices, as well as by static faces and voices. Our findings help resolve previous uncertainty about whether static face–voice matching is possible, presenting two complementary explanations for the apparent contradictions. The data suggest that static face–voice matching is more likely to be above chance level when the alternatives in a 2AFC task are presented simultaneously. In addition, the variance associated with stimuli indicates that some people look and sound more similar than others, an issue that has not been properly accounted for by the analyses undertaken in previous research, but that helps explain why the static face–voice matching performance across previous studies might be inconsistent. Our results therefore support the conclusion that dynamic visual information about articulatory patterns facilitates accuracy (Kamachi et al., 2003; Lachs & Pisoni, 2004a, b; Lander et al., 2007; Rosenblum et al., 2006), but that it alone cannot explain the existence of shared source identity information with voices. Crossmodal source identity information is available in both static and dynamic faces.

Author note This research was supported by a Ph.D. studentship from the Division of Psychology, Nottingham Trent University.

References

Abitbol, J., Abitbol, P., & Abitbol, B. (1999). Sex hormones and the female voice. *Journal of Voice, 13,* 424–446. doi:10.1016/S0892-1997(99)80048-4

Baguley, T. (2012). *Serious stats: A guide to advanced statistics for the behavioral sciences*. Basingstoke: Palgrave.

Bates, D, Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.0-6. Retrieved January 30, 2015, from http://CRAN.R-project.org/package=lme4

Beckford, N. S., Rood, S. R., & Schaid, D. (1985). Androgen stimulation and laryngeal development. *Annals of Otology, Rhinology and Laryngology, 94,* 634–640.

Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences, 11,* 535–543. doi:10.1016/j.tics.2007.10.001

Christie, F., & Bruce, V. (1998). The role of dynamic information in the recognition of unfamiliar faces. *Memory & Cognition, 26,* 780–790. doi:10.3758/BF03211397

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12,* 335–359. doi:10.1016/S0022-5371(73)80014-3

Clark, S. E., Howell, R. T., & Davey, S. L. (2008). Regularities in eyewitness identification. *Law and Human Behavior, 32,* 187–218. doi:10.1007/s10979-006-9082-4

Collins, S. A., & Missing, C. (2003). Vocal and visual attractiveness are related in women. *Animal Behaviour, 65,* 997–1004. doi:10.1006/anbe.2003.2123

Cooke, M., Barker, J., Cunningham, S., & Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. *Journal of the Acoustical Society of America, 120,* 2421–2424. doi:10.1121/1.2229005

Ebbesen, E. B., & Flowe, H. D. (2002). *Simultaneous v. sequential lineups: What do we really know?* Unpublished manuscript.

Feinberg, D. R., Jones, B. C., DeBruine, L. M., Moore, F. R., Law Smith, M. J., Cornwell, R. E., & Perrett, D. I. (2005). The voice and face of woman: One ornament that signals quality? *Evolution and Human Behavior, 26,* 398–408. doi:10.1016/j.evolhumbehav.2005.04.001

García-Pérez, M. A., & Alcalá-Quintana, R. (2011). Improving the estimation of psychometric functions in 2AFC discrimination tasks. *Frontiers in Psychology, 2,* 96. doi:10.3389/fpsyg.2011.00096

Gelman, A. E., & Su, Y. S. (2013). arm: Data analysis using regression and multilevel/hierarchical models (R package version 1.6-05). Retrieved January 30, 2015, from http://CRAN.R-project.org/package=arm

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103,* 54–69. doi:10.1037/a0028347

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. doi:10.1016/j.cub.2003.09.005

Knappmeyer, B., Thornton, I. M., & Bülthoff, H. H. (2003). The use of facial motion and facial form during the processing of identity. *Vision Research, 43,* 1921–1936. doi:10.1016/S0042-6989(03)00236-0

Kneller, W., Memon, A., & Stevenage, S. (2001). Simultaneous and sequential lineups: Decision processes of accurate and inaccurate eyewitnesses. *Applied Cognitive Psychology, 15,* 659–671. doi:10.1002/acp.739

Krauss, R. M., Freyberg, R., & Morsella, E. (2002). Inferring speakers' physical attributes from their voices. *Journal of Experimental Social Psychology, 38,* 618–625. doi:10.1016/S0022-1031(02)00510-3

Kuhl, P. K., & Meltzoff, A. N. (1984). The intermodal representation of speech in infants. *Infant Behavior and Development, 7,* 361–381. doi:10.1016/S0163-6383(84)80050-8

Lachs, L. (1999). A voice is a face is a voice: Cross-modal source identification of indexical information in speech. In *Research on spoken language processing* (Progress Report No. 23, pp. 241–258).

Bloomington, IN: Indiana University, Department of Psychology, Speech Research Laboratory.

Lachs, L., & Pisoni, D. B. (2004a). Crossmodal source identification in speech perception. *Ecological Psychology, 16,* 159–187. doi:10.1207/s15326969eco1603_1

Lachs, L., & Pisoni, D. B. (2004b). Specification of cross-modal source information in isolated kinematic displays of speech. *Journal of the Acoustical Society of America, 116,* 507–518. doi:10.1121/1.1757454

Lander, K., & Chuang, L. (2005). Why are moving faces easier to recognize? *Visual Cognition, 12,* 429–442. doi:10.1080/13506280444000382

Lander, K., Hill, H., Kamachi, M., & Vatikiotis-Bateson, E. (2007). It's not what you say but the way you say it: Matching faces and voices. *Journal of Experimental Psychology: Human Perception and Performance, 33,* 905–914. doi:10.1037/0096-1523.33.4.905

MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Perception & Psychophysics, 24,* 253–257. doi:10.3758/BF03206096

Mavica, L. W., & Barenholtz, E. (2013). Matching voice and face identity from static images. *Journal of Experimental Psychology: Human Perception and Performance, 39,* 307–312. doi:10.1037/a0030945

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746–748. doi:10.1038/264746a0

O'Connor, J. J., Re, D. E., & Feinberg, D. R. (2011). Voice pitch influences perceptions of sexual infidelity. *Evolutionary Psychology, 9,* 64–78.

O'Toole, A. J., Roark, D., & Abdi, H. (2002). Recognizing moving faces: A psychological and neural synthesis. *Trends in Cognitive Science, 6,* 261–266. doi:10.1016/S1364-6613(02)01908-3

Peirce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics, 2*(10), 1–8. doi:10.3389/neuro.11.010.2008

Penton-Voak, I. S., & Chen, J. Y. (2004). High salivary testosterone is linked to masculine male facial appearance in humans. *Evolution and Human Behavior, 25,* 229–241. doi:10.1016/j.evolhumbehav.2004.04.003

Perrett, D. I., Lee, K. J., Penton-Voak, I., Rowland, D., Yoshikawa, S., Burt, D. M., & Akamatsu, S. (1998). Effects of sexual dimorphism on facial attractiveness. *Nature, 394,* 884–887. doi:10.1038/29772

Pisanski, K., Mishra, S., & Rendall, D. (2012). The evolved psychology of voice: Evaluating interrelationships in listeners' assessments of the size, masculinity, and attractiveness of unseen speakers. *Evolution and Human Behavior, 33,* 509–519. doi:10.1016/j.evolhumbehav.2012.01.004

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15,* 351–357. doi:10.2307/2087176

Rosenblum, L. D., Smith, N. M., Nichols, S. M., Hale, S., & Lee, J. (2006). Hearing a face: Cross-modal speaker matching using isolated visible speech. *Perception & Psychophysics, 68,* 84–93. doi:10.3758/BF03193658

Saxton, T. K., Caryl, P. G., & Roberts, C. S. (2006). Vocal and facial attractiveness judgments of children, adolescents and adults: The ontogeny of mate choice. *Ethology, 112,* 1179–1185. doi:10.1111/j.1439-0310.2006.01278.x

Smith, H. M. J., Dunn, A. K., Baguley, T., & Stacey, P. C. (2015). Concordant cues in faces and voices: Testing the back-up signal hypothesis. *Evolutionary Psychology* (in press).

Steblay, N. K., Dysart, J. E., & Wells, G. L. (2011). Seventy-two tests of the sequential lineup superiority effect: A meta-analysis and policy discussion. *Psychology, Public Policy, and Law, 17,* 99–139. doi:10.1037/a0021650

Thornhill, R., & Grammer, K. (1999). The body and face of woman: One ornament that signals quality? *Evolution and Human Behavior, 20,* 105–120. doi:10.1016/S1090-5138(98)00044-0

Urbaniak, G. C., & Plous, S. (2013). Research randomizer (Version 4.0) [Computer software]. Accessed 22 Nov 2014, at www.randomizer.org

Wells, G. L. (1984). The psychology of lineup identifications. *Journal of Applied Social Psychology, 14,* 89–103. doi:10.1111/j.1559-1816.1984.tb02223.x

Wells, T., Baguley, T., Sergeant, M., & Dunn, A. (2013). Perceptions of human attractiveness comprising face and voice cues. *Archives of Sexual Behavior, 42,* 805–811. doi:10.1007/s10508-012-0054-0

Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human Behavior, 22,* 603–647. doi:10.1023/A:1025750605807

Yehia, H., Rubin, P., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication, 26,* 23–43. doi:10.1016/S0167-6393(98)00048-X

Yeshurun, Y., Carrasco, M., & Maloney, L. T. (2008). Bias and sensitivity in two-interval forced choice procedures: Tests of the difference model. *Vision Research, 48,* 1837–1851. doi:10.1016/j.visres.2008.05.008