Distinctiveness helps when matching static faces and voices.

Sarah V Stevenage, Iain Hamlin and Becky Ford

Department of Psychology, University of Southampton, UK

Correspondence:

\*Please address all correspondence to Sarah Stevenage at:

Department of Psychology, University of Southampton, Highfield, Southampton, Hampshire, UK, SO17 1BJ

Email: [svs1@soton.ac.uk](mailto:svs1@soton.ac.uk); Fax: 02380 594597; Tel : 02380 592973

Acknowledgements :

This work was supported by the Engineering and Physical Sciences Research council (EPSRC) Grant (EP/J004995/1 SID: An Exploration of SuperIdentity) awarded to the first author. Colleagues on this grant are thanked for helpful contributions to the current work. In particular, the authors would like to thank Dr Greg Neil for assistance in obtaining distinctiveness ratings.

Abstract

Three experiments are reported which examined the capacity to match a voice with a static image of a face. When using a simultaneous same/different matching task, performance was significantly better than chance (Experiments 1 and 2). However, it did not appear to depend either on sex of speaker, sex of listener, stimulus distinctiveness, or self-reported strategies (Experiment 2). Concerns over floor effects as well as a significant response bias prompted a change of task, and when performance was examined through matching a voice to a face lineup, a more interesting pattern emerged. Again, performance was significantly better than chance, but in addition, it was demonstrably affected by the distinctiveness of the speaker’s voice. These results are considered in the context of theoretical discussions regarding face-voice integration, and in the context of more applied considerations regarding multimodal benefits in witness scenarios.

Keywords:

Face-voice matching, face-voice integration, multimodal person perception, witness performance

Distinctiveness helps when matching faces and voices.

Recent literature has suggested that the face and voice may offer overlapping cues about a person, and that these can be perceived even when that person is unfamiliar. These overlapping cues may support the perception of personality traits such as extraversion and conscientiousness (Allport & Cantril, 1934), and the perception of physical characteristics such as masculinity/femininity, health and height (Smith, Dunn, Baguley & Stacey, 2016a). In addition, the face and voice overlap in the presentation of speech-related cues such that the sight of a talking face can help to disambiguate speech perception amidst noise (Sumby & Pollack, 1954). Overlapping cues have been suggested as a basis for the capacity to match a face to the corresponding voice from the same unfamiliar individual (Mavica & Barenholtz, 2013; Smith, Dunn, Baguley & Stacey, 2016b). However, the extent to which this is based on dynamic speech-based cues has been thrown into question through the demonstration of successful face-voice matching even when faces are static. The purpose of the present paper is to replicate recent demonstrations of unfamiliar face-voice matching with static faces using an optimal design. More importantly, the work reported here provides an exploration of speaker and listener characteristics that may optimise matching performance.

*Face-Voice Matching*

The capacity to match an unfamiliar face and voice has been explored in a number of studies. The first was presented by Kamachi, Hill, Lander and Vatikiotis-Bateson (2003) who used the faces and voices of 40 unfamiliar Japanese targets as stimuli with native Japanese speakers as participants. The participants completed a delayed match-to-sample (XAB) task in which stimulus modality was changed between presentation of the target (X) and the two alternatives (A and B). Thus, in the face-voice condition, participants saw a face (X) prior to hearing two voices presented sequentially (A and B), whilst in the voice-face condition, participants heard a voice (X) prior to seeing two faces presented sequentially (A and B). In both cases, the participants’ task was to indicate which of the two alternatives (A or B) matched the identity of the target (X), and successful matching was revealed if performance exceeded chance levels of 50%. The manipulation of stimulus order (face-voice; voice-face) enabled exploration of asymmetries in the strength of each modality when signalling identity. More importantly, Kamachi et al. explored matching performance when faces were both static (presented as a photograph) and dynamic (presented as a silent articulating video).

Kamachi et al.’s results were clear in revealing a capacity to match unfamiliar faces and voices when faces were dynamic, and when face and voice clips were played forwards. Additionally, performance remained above chance in the face-voice condition even when the voices had been manipulated to provide sinewave speech (which removes identity-relevant cues such as fundamental frequency and vocal quality). This finding was consistent with the interpretation that the face acted as a stronger signaller of identity than the manipulated voice.

Against these results, Kamachi et al. showed that performance dropped to chance level when static faces were used, or when sequences were played backwards, leading to the conclusion that successful face-voice matching depended on coarse information that was time-varying and direction-dependent. As a result, successful face-voice matching with static faces was neither anticipated nor demonstrated.

These results were replicated by Lachs and Pisoni (2004) who also used a delayed match-to-sample (XAB) task. In contrast to Kamachi *et al.* (2003) who used 40 speakers uttering short sentences, Lachs and Pisoni (*ibid)* used only 8 speakers who provided voice clips consisting of single words. Nevertheless, their results echoed those of Kamachi *et al.* in demonstrating successful face-voice matching when face and voice clips were played forwards. Similarly, this capacity did not depend on preserved information about fundamental frequency as it was still possible to match faces and voices when the voice clips had been noise-band filtered. Again, however, matching was not possible when face and voice clips were played backwards or when the face was static.

Given the demonstration of matching from dynamic but not static faces, both sets of authors concluded that face-voice matching depended on spatiotemporal information related to a common source – the act of articulation. Indeed, Kamachi et al. considered that the fabric of the face, and the characteristics of the voice, were each affected in perceptible ways by the common source event – the production of speech. As such, face-voice matching with static faces was considered unlikely.

The results of Lander, Hill, Kamachi and Vatikitotis-Bateson (2007) confirmed the importance of speech-based information. Lander et al. used 48 speakers as targets, and again used the delayed match to sample (XAB) task with dynamic face and voice clips. However, the speakers used different styles of speech ranging from short sentences providing statements, questions, conversational speech, slow speech (clearly articulated) and rushed speech (casually articulated). Consequently, both content (what was said) and style (how it was said) could be varied between study (X) and test (A and B). Lander et al.’s results suggested that matching was possible, as long as content and style did not *both* vary. However, performance was significantly affected by the style of speech (statement versus question, conversational versus clear (slow), and conversational versus casual (rushed)). In contrast, performance was not affected by artificial linear speeding or slowing of the voice clip. As such, Lander et al. (2007) emphasized the importance of the style of speaking in face-voice matching given the individual differences in prosody that may exist. Consequently, the basis for face-voice matching was again linked to the common source event, speech production, making matching between a voice and a static face unlikely.

Against these results, a number of studies now exist which do however demonstrate successful face-voice matching even when the faces are presented as static images. The first was provided by Mavica and Barenholtz (2013) who used 64 voice clips in a modified version of the delayed match-to-sample (XAB) in which memory demands were avoided. In their simultaneous match-to-sample task, participants saw two test faces as static photographs (A and B) whilst simultaneously listening to the target voice (X). As before, the participants’ task was to indicate which of the two faces matched the speaker’s voice. This change to a simultaneous match-to-sample task may be key. Indeed, evidence exists to suggest that dynamic facial information may be retained better in memory than static facial information (Christie & Bruce, 1998, Lander & Chuang, 2005). As such, a matching task which imposes a memory load may bias against successful performance when the face is static. The removal of memory demands in Mavica and Barenholtz’s (ibid) study meant that participants were able to demonstrate above-chance matching of static faces and voices (Experiment 1). This demonstration was important as it contradicted the previous assumption that matching was based on the perception of common speech cues from the voice and the dynamic face. Instead, Mavica and Barenholtz suggested that the voice and static face carried sufficient common and overlapping information to support matching even without the addition of speech cues.

Surprisingly, however, Mavica and Barenholtz (2013) demonstrated above-chance matching to static faces even when the delayed match-to-sample (XAB) task of previous studies was used (Experiment 2). These latter results were unexpected if the previous demonstration of static face-voice matching relied on the removal of a memory load. However, it is possible that Mavica and Barenholtz (2013) were successful in revealing above-chance face-voice matching with static faces because of their use of a considerable set of 64 speakers. This raises the suggestion that whilst task demands are important in the demonstration of static face-voice matching, so too are the items used.

The work of Smith, Dunn, Baguley and Stacey (2016a) sought to address these methodological and stimulus-based factors. They used 18 targets who provided abstract 6-word voice clips (i.e., ‘Place blue at J9 now’) along with either static or dynamic face stimuli. In Experiment 1, they asked participants to rate faces and voices for a series of characteristics and their results suggested a positive correlation between the ratings made from faces and voices of the same target regardless of whether the faces were static or dynamic. In Experiment 2, they explored face-voice matching using a same-different matching task in which participants heard a voice before seeing a single face, or saw a face before hearing a single voice. Consequently, the participants’ memory load was reduced by only having to hold the target (rather than the target and both alternatives) in memory. The participants’ task was to say whether the second stimulus came from the same person as the first. With memory demands lowered, the results showed an above-chance ability to match faces and voices. Echoing Mavica and Barenholtz (2013), this emerged for static as well as dynamic faces. The authors suggested that the correlation between face and voice ratings (Expt 1) demonstrated that the two modalities presented overlapping signals and that these overlapping signals may have been the basis of the successful performance in the matching task (Expt 2).

The final study of relevance is provided by Smith, Dunn, Baguley and Stacey (2016b). Using the same stimuli as Smith et al. 2016a, the authors again demonstrated the capacity to match faces and voices at above-chance levels even when faces were presented as static images. However, this effect was highly sensitive to the methodology, and to the items used. In particular, static face-voice matching was not possible when the delayed match-to-sample task was used (Expt 1), or when the memory load was reduced by asking participants to judge which of two face-voice combinations represented a matching pair (Expt 2). However, when participants were asked to listen to a voice and then choose which of two simultaneously presented faces was of the same identity, static face-voice matching was revealed. Here, the authors again noted the importance of reducing the memory load, specifically through the simultaneous presentation of alternatives to match to the target.

These findings combine to suggest that static face-voice matching is possible but it appears to be fragile and particularly sensitive to methodological conditions. Indeed, successful static face-voice matching seems most likely when the memory demands in the task are reduced or removed. However, the results of Smith et al. (2016a, 2016b) highlighted two concerns which are worth noting. First, the change in methodology enabled Smith et al. (2016a) to examine not only the ability to *correctly match* a face to a voice, but also the ability to *correctly detect a mismatch*. In this regard, the results revealed an overall tendency to say ‘match’. This was suggestive of a response bias, although the method of analyses meant that this was not formally examined. Second, Smith et al.’s (2016a, b) application of multilevel modelling enabled transparency regarding the differences between items as well as participants. Indeed, their results revealed considerable item effects suggesting that some voices were easier to match to their faces than others. Given the use of only 18 voices in their study, and indeed the use of only 8 voices by Lachs and Pisoni (2004), Smith et al. (2016a) advocated the use of a larger speaker set to be sure that results were not weakened by an unfortunate choice of stimuli.

With these considerations in mind, the current studies used a large database of voice clips, dynamic face clips (video), and static face images (photographs), and employed two different methodologies each of which completely removed memory demands. The purpose, given this optimal design, was to replicate the demonstration of above-chance face-voice matching with static faces, and then to start to explore the strategies or stimulus characteristics that may underpin performance.

Experiment 1

Experiment 1 was designed to replicate previous demonstrations of successful face-voice matching from static faces. Importantly, however, the present study addresses the limitations of previous work through the use of a large database of 72 targets each of whom provided static facial images, dynamic articulating face clips, and voice clips uttering short sentences. In addition, the present study removed all memory load associated with the matching task to enable demonstration of static face-voice matching to its fullest extent. Finally, the use of a same/different matching task enabled exploration of the capacity to identify a mismatch as well as the capacity to determine a match (as in Smith et al., 2016a). However, it also enabled formal exploration of the suggested response bias in performance (Smith et al., 2016a, b). Given the demonstration of successful static face-voice matching in the literature when these issues were individually addressed, it was predicted that participants would be able to match static faces to voices here with increased levels of performance.

*Design*

Participants took part in a simultaneous same/different matching task in which they decided whether a face and voice belonged to the same person or not. A within-participants design was used in which trial type (match, mismatch) was varied. Accuracy was recorded and represented the dependent variable.

*Participants*

A total of 88 participants (37 males, 51 females) took part, with ages varying between 18 and 61 years (*M* = 33.9 years, *SD* = 12.3). Participants were drawn largely from a US population and were recruited via the Mechanical Turk online crowdsourcing website (MTurk)[[1]](#footnote-1). They received a small monetary reward for their time. Participants confirmed normal, or corrected to normal, hearing and vision, and all were unfamiliar with the stimuli used in the current task.

*Materials*

Stimuli consisted of the faces and voices of 72 targets (36 males, 36 females). They were drawn from a set of 117 targets represented in the SuperIdentity Stimulus Database ([www.southampton.ac.uk/superidentity](http://www.southampton.ac.uk/superidentity)). The database was comprised of targets between 18 and 30 years of age, however, those selected for the current study were all drawn from an 18-24 year old age bracket meaning that all displayed the facial and vocal characteristics that follow puberty, and none showed significant signs of ageing as may arise in middle years. In addition, none had distinguishing vocal features such as accents, stutters or lisps. Example male and female stimuli are provided as supplementary data although, for the protection of the targets’ privacy, the identity of faces and voices do not correspond.

For each of the 72 targets, a high quality facial image was obtained, depicting a full-face image of the target with a neutral expression, seated in front of a plain grey curtain. Images were edited within Photoshop to be presented within a standard frame of 369 x 369 pixels. For each of the 72 targets, a good quality voice clip was also obtained in which the target uttered a standard phrase (‘The smell of freshly ground coffee never fails to entice me into the shop’). The phrase took between 4 and 5 seconds to utter, and all clips were edited within Audacity 3.1.0 to remove unnecessary pauses, and to standardise volume.

From these stimuli, 36 matching trials and 36 mismatching trials were created. In matching trials, the voice and face of the same individual were presented whereas, in mismatching trials, the voice of one target was paired with the face of a same-sex foil. Care was taken to counterbalance the identity of targets in each trial type such that the targets presented in matching trials for half the participants were presented in mismatching trials for the remaining participants.

Stimuli were presented and data were recorded using iSurvey ([www.soton.ac.uk/isurvey](http://www.soton.ac.uk/isurvey)), a bespoke online survey engine, which allowed the presentation of multimedia trials, and enabled participants to listen to the voice as many times as they wished.

*Procedure*

Following the provision of informed consent, participants received on-screen instructions describing the nature of the face-voice matching task. Given concerns over the quality of data that may be obtained from online participants, special care was given both when recruiting and instructing the participants. Specifically, participants were asked to set their screen display settings to 1024 x 768 pixels, and were asked to use headphones for optimal listening conditions. They were also asked to avoid breaks or distractions from other devices, and were advised of the likely length of the experiment at the outset. Whilst it was not possible to ensure compliance with all requests, the total participation time was recorded and did not indicate substantial deviation from the anticipated study length and thus, no participants were dropped on this basis. During the experiment itself, audibility of the voice clips was optimised through the use of a repeated-play function allowing participants to adjust and set their volume as necessary at the start of the experiment without risk of missing the stimuli. Finally, following the experiment, the data were screened to minimise any instances of repeat-participation through cross-referencing of the IP addresses used by participants.

The request for online consent, and the provision of instructions, was followed by the presentation of a randomised order of 72 experimental trials. All trials took the same format and consisted of the simultaneous presentation of a face whilst hearing a target voice. An on-screen question asked ‘Do you think this face and voice belong to the same person?’ and participants were directed to respond by pressing ‘Y’ for ‘yes’ and ‘N’ for ‘no’. Importantly, the face remained visible until response and, prior to response, the participants could listen to the voice as many times as they wished, via on-screen ‘play’ and ‘stop’ buttons. No feedback was provided and the entire experiment lasted an average of 21 minutes after which participants were thanked and debriefed.

Results and Discussion

Given the possibility of response bias in the same/different matching task, accuracy for matching and mismatching trials was combined according to the signal detection framework. This provided a measure of sensitivity of discrimination (d’) and response bias (C) and enabled a formal evaluation of response bias given the concerns of Smith et al. (2016a, b). In addition, the demonstration of item effects in previous work meant that variance across items could not be ignored. Whilst a large number of items were used in the current study to minimise the likelihood of unfortunate selection of stimuli, item effects were also formally examined through the conduct of analyses by-items. These are reported alongside the by-participants analyses. Primary analyses were conducted on sensitivity of discrimination and response bias, with secondary analyses provided to explore accuracy of performance. All measures are summarised in Table 1.

*Sensitivity of Discrimination and Response Bias*

A one-sample *t*-test was used to compare sensitivity of discrimination (d’) to the chance level of zero. This revealed above-chance performance overall (by-participants: *t*(87) = 4.44, p < .001; by-items: *t*(71) = 2.55, *p* = .013). Thus, participants were able to match faces with voices despite the use of static faces in the current study. As above, a one-sample *t-*test was also used to compare bias (C) to zero. This revealed a significant bias to say ‘match’ (by-participants: *t*(87) = 7.62, *p* < .001; by-items: *t*(71) = 11.00, *p* < .001), confirming the suggestion of a response bias by Smith *et al*. (2016a, b).

(Please insert Table 1 about here)

*Accuracy of Performance*

In order to determine whether trial type affected performance, a paired samples *t-*test was conducted to compare accuracy for matching and mismatching trials. In line with the evidence of a response bias above, this revealed significantly better performance on matching trials (67.6%) than on mismatching trials (37.7%) (by-participants: *t*(87) = 9.78, p < .001; by-items: *t*(71) = 11.51, *p* < .001). Moreover, Bonferroni-corrected one-sample *t*-tests confirmed that performance on matching trials was significantly better than chance (by-participants: *t*(87) = 11.50, p < .001; by-items: *t*(71) = 10.48, *p* < .001) whilst performance on mismatching trials was significantly worse than chance (by-participants: *t*(87) = -6.98, p < .001; by-items: *t*(71) = -7.34, *p* < .001). Consequently, the good overall discrimination reported above rested on substantially better performance in matching than mismatching trials.

Finally, and in light of concerns over item effects (Mavica & Barenholtz, 2013; Smith et al. 2016a, b) the data were explored by calculating the face-voice matching accuracy for each of the 72 speakers. This showed substantial variation in performance across items, with overall accuracy ranging from 33% to 77%. However, there was an overall trend towards better-than-chance performance, with 43/72 voices (60%) showing performance over chance levels. Whilst the overall pattern of performance was consistent whether analyses were conducted by-participants or by-items, the figures nevertheless echo the concerns over item effects expressed in previous studies.

Taken together, the results of Experiment 1 demonstrated above chance levels of performance in a simultaneous face-voice matching task. This confirmed previous findings in that participants were able to determine whether a face and a voice belonged to the same individual even when the face was static. Interestingly, successful face-voice matching did not appear to rely on the speakers and listeners being drawn from the same demographic. Whilst the MTurk participants may have been familiar with the English accent, and whilst all declared fluency in English as a condition of participation, the MTurk participants were drawn largely from the US. Yet, they were able to match faces and voices successfully despite these voices potentially speaking with a relatively unfamiliar accent.

In considering the successful face-voice matching results here, it may be helpful to remember that the use of targets from a similar and relatively narrow age range meant that performance was unlikely to be based on the matching of perceived age from both face and voice. With age being readily perceived from faces and voices (see Zäske & Schweinberger, 2011; Zäske, Skuk, Kaufmann & Schweinberter, 2013), it could have provided a basis for face-voice matching in a more heterogeneous sample of targets and, whilst this would not undermine the capacity for successful matching, it may mask other more interesting bases for this performance. As such, the present design appeared to provide a robust yet controlled test of face-voice matching. This suggested that the present methodology was appropriate to enable exploration of the characteristics that may underpin performance. This was the purpose of Experiment 2.

Experiment 2

Experiment 2 replicated the face-voice matching task from Experiment 1. However, in addition, it examined whether performance may be accounted for by speaker variables, listener variables, or listener strategies. Following the results of Experiment 1, and the observations of Lander et al*.,* (2007), Mavica and Barenholtz (2013) and Smith et al. (2016a, b), it was anticipated that items effects may be important. Given the importance of item distinctiveness when processing faces (Bruce, Burton & Hancock, 1995) and voices (Mullenix, Ross, Smith, Kuykendall, Conard & Barb, 2011), and given Smith et al.’s (2016b) suggestion that some faces and voices may be more easily matched than others, Experiment 2 manipulated vocal and facial distinctiveness to determine their influence on performance. In addition, Experiment 2 explored the influence of speaker sex and listener sex. Whilst previous research had not revealed an effect of either variable on face-voice matching (Mavica & Barenholtz, 2013), we explored here the possibility of an own-sex matching effect in which participants may perform better with targets of their own sex. Finally, participants within Experiment 2 were asked how they made their decisions in order to determine whether they held any insight into their performance. In sum, the purpose of Experiment 2 was to start to explore the characteristics that may enable successful static face-voice matching.

*Design*

As in Experiment 1, participants took part in a simultaneous same/different matching task in which they decided whether a face and voice belonged to the same person or not. Here, however, a 2 x 2 x 2 mixed design was used in which trial type (matching, mismatching) and sex of speaker (male, female) were varied for male and female participants. As before, accuracy was recorded and represented the dependent variable.

*Participants*

A total of 100 participants (46 males, 54 females) took part, with ages varying between 21 and 64 years (*M* = 37.9 years, *SD* = 11.9). As in Experiment 1, participants were recruited via MTurk and received a small monetary reward for their time. Participants confirmed normal, or corrected-to-normal, hearing and vision, and all were unfamiliar with the stimuli used in the current task. In addition, screening of the IP addresses minimised the likelihood that anyone had taken part in Experiment 1.

*Materials*

The experimental task was identical to that used in Experiment 1. However, for this study, all target voices and faces were pre-rated by a set of 6 judges on a 10-point scale for distinctiveness. The ratings showed that the stimuli represented a spread along the distinctiveness scale, with distinctiveness of voice clips ranging from 3.63 to 8.33 (*M* = 6.30, *SD* = .94) and distinctiveness of faces ranging from 2.6 to 7.6 (*M* = 5.34, *SD* = 1.20). Descriptive data suggested that there was a similar level of variance across judges when rating distinctiveness of each face (average *SD* = 1.88) and each voice (average *SD* = 2.01). Moreover, there was no significant difference in the variance shown by judges for faces and voices (*t*(142) = 1.19, *p* = .24). These observations suggested that the consistency of judges’ ratings did not differ substantially when assessing facial and vocal distinctiveness.

A questionnaire was also administered in Experiment 2 to elicit self-reported strategies underpinning participant performance. This questionnaire was developed following the face-to-face testing of a small group of 20 university-aged individuals (18-20 years, *M* = 19.2 years, *SD* = .83). Each of these individuals completed the matching task as described in Experiment 1, and analysis confirmed that they showed an above-chance level of performance (*d’*: *t*(19) = 4.14, *p* = .001) and an equivalent level of discrimination to those tested in Experiment 1 (*d’*: *t*(106) = 1.55, *p* = .13). Importantly, these participants also outlined how they made their decisions. The experimenters extracted a set of 49 non-overlapping strategies from these responses, and these were compiled into a checklist for the current study (see Appendix 1).

*Procedure*

The matching task was administered exactly as in Experiment 1. In addition, however, participants completed a post-experimental stage during which they indicated how they had made their decisions. In this regard, participants provided a rating against each of the 49 strategies described in the pilot phase, allowing them to indicate the extent to which they used each of the strategies. Ratings ranged from 1 (never used) to 5 (very frequently used). The entire process lasted an average of 25 minutes after which participants were thanked and debriefed.

Results and Discussion

As in Experiment 1, accuracy on matching and mismatching trials was combined to provide measures of sensitivity of discrimination (d’) and bias (C). In addition, and as in Experiment 1, analyses were conducted both by-participants and by-items so that both sources of variance could be addressed. Primary analyses were conducted on sensitivity of discrimination and response bias, with secondary analysis being conducted on accuracy. All measures are summarised in Table 2 and Figure 1.

*Sensitivity of Discrimination*

A series of Bonferroni-corrected one-sample *t*-tests was used to compare sensitivity of discrimination (d’) to the chance level of zero. As before, these revealed above-chance performance overall (by-participants: *t*(99) = 6.11, *p* < .001; by-items: *t*(71) = 2.81, *p* = .006). Direct comparison to Experiment 1 showed no significant difference in levels of performance (*t*(186) < 1, *ns*).

In order to see whether sex of speaker and sex of listener influenced performance, a 2 x 2 mixed Analysis of Variance (ANOVA) was conducted on d’, with sex of speaker and sex of listener as variables. This revealed neither a main effect of sex of speaker (by-participants: *F*(1, 98) = 2.36, p = .13; by-items: *F*(1, 70) = 1.41, *p* = .24), nor of sex of listener (by-participants: *F*(1, 98) = 2.21, p = .14; by-items: *F*(1, 70) = 1.10, *p* = .30). Moreover, no interaction emerged between the two variables (by-participants: *F*(1, 98) < 1, *p* = .59; by-items: *F*(1, 70) < 1, *p* = .85). Taken together, these results revealed that participants were able to match faces with voices but, in common with the results of Mavica and Barenholtz (2013), there was no evidence to indicate any effect of speaker or listener sex. Moreover, there was no evidence to indicate an own-sex bias in performance.

(Please insert Table 2 and Figure 1 about here)

*Response Bias*

Again, a series of Bonferroni-corrected one-sample *t*-tests was used to compare bias (C) to zero. These revealed a significant bias to say ‘match’ when evaluating performance overall (by-participants: *t*(99) = 7.62, *p* < .001; by-items: *t*(71) = 11.13, *p* < .001). In order to see whether sex of speaker and sex of listener influenced response bias, a 2 x 2 mixed ANOVA was conducted as above. This revealed tentative evidence for a main effect of sex of speaker (by-participants: *F*(1, 98) = 15.48 *p* < .001, partial η2 = .14; by-items: *F*(1, 70) = 2.10, *p* = .15), with a greater response bias for female speakers than for male speakers. In comparison, however, there was no effect of sex of listener (by-participants: *F*(1, 98) < 1, *p* = .65; by-items: *F*(1, 70) < 1, *p* = .69), and there was no interaction to indicate any own-sex influence on response bias (by-participants: *F*(1, 98) < 1, *p* = .54; by-items: *F*(1, 70) < 1, *p* = .89).

*Accuracy of Performance*

In order to determine whether performance differed across trial type, a 2 x 2 x 2 mixed ANOVA was conducted on accuracy using trial type (matching, mismatching), sex of speaker (male, female), and sex of listener (male, female) as variables. This revealed no significant effect of either sex of speaker (by-participants: *F*(1, 98) = 3.63,*p* = .06; by-items: *F*(1, 70) < 1, *p* = .37) or sex of listener (by-participants: *F*(1, 98) = 1.61, p = .21; by-items: *F*(1, 70) < 1, *p* = .40). However, a main effect of trial type did emerge (by-participants: *F*(1, 98) = 176.71, *p* < .001, partial η2 = .64; by-items: *F*(1, 70) = 138.70, *p* < .001, partial η2 = .67), with performance being better for matching trials than for mismatching trials. This was qualified by a tentative interaction between sex of speaker and trial type (by-participants: *F*(1, 98) = 13.86, *p* < .001, partial η2 = .12; by-items: *F*(1, 70) = 2.26, *p* = .14). Tests of simple main effects confirmed that performance when collapsed across items was better for matching than mismatching trials for both male speakers (*t*(99) = 10.42, *p* < .001) and female speakers (*t*(99) = 14.01, *p* < .001), but this effect was greater for female speakers (*t*(99) = 3.70, *p* < .001). No further interactions emerged to qualify these findings (by-participants: all *F*s(1, 98) < 1, *ns*; by-items: all *F*s(1, 70) < 1, *ns*).

As a whole, these results suggested that participants were able to match faces and voices at above-chance levels. This ability was unaffected by sex of speaker and sex of listener, and did not suggest an own-sex bias. Moreover, performance suggested both a perceptual factor through d’, and a decisional factor through response bias, accounting for greater accuracy on matching rather than mismatching trials.

*Item Effects and Effects of Vocal Distinctiveness*

As in Experiment 1, it was notable that performance varied across items, with accuracy ranging between 31% and 78%. However, again, there was a clear trend towards better-than-chance performance, with 44/72 voices (61%) eliciting overall accuracy levels above 50%. To determine whether these item effects reflected variation in vocal distinctiveness, a correlational approach was taken. Pearson’s correlations were calculated between vocal distinctiveness, facial distinctiveness and performance as measured through sensitivity of discrimination, accuracy on matching trials, and accuracy on mismatching trials as calculated for each item. These revealed no association between vocal distinctiveness and facial distinctiveness (*r2* = -.18, n = 72, *p* = .14) suggesting that the distinctiveness of an individual’s voice was not associated with the distinctiveness of their face. Of more interest, however, the results revealed no association between vocal distinctiveness and d’, overall accuracy, or accuracy on matching and mismatching trials (all *r2*s < 1, *p*s > .49), and no association between facial distinctiveness of targets and accuracy on matching trials, or of facial distinctiveness of foils and accuracy on mismatching trials (all *r2*s < 1, *p*s > .69). Consequently, performance could not be accounted for by sex of speaker or listener (above) or by distinctiveness of face or voice (here).

*Effectiveness of Self-Reported Strategies*

Finally, in order to determine whether participants’ strategies explained their above-chance performance, an analysis was conducted on the data from the post-experimental stage. In this regard, a principal axis factor analysis was performed on the participants’ strategy endorsements. An examination of the scree plot indicated the presence of three factors and, given no theoretical reason to assume that the factor structure would be orthogonal, direct oblimin was chosen as the method of rotation. The resultant factors cumulatively accounted for 38.7% of the variance in the data. The factors were identified based on the items that loaded on them. As such, Factor 1 indicated the use of emotion when matching faces and voices, with participants indicating that they looked for congruence of emotion when making their matching judgements. This factor accounted for 24.4% of the variance in the data. Factor 2 indicated the use of general physical attributes when making matching decisions by assuming, for example, that a man with a deep voice may be tall as indicated by neck width. This factor accounted for 7.7% of the variance. Finally, Factor 3 indicated the use of specific facial features or specific vocal characteristics when making matching decisions by assuming a correlation between physical and acoustic features (i.e., someone with bigger lips may place more emphasis on the ‘p’ of ‘shop’). This factor accounted for 6.6% of the overall variance (see Appendix for a list of strategies loading on each factor).

Whilst the factor analysis revealed the dominant strategies that participants used, the final analysis determined whether any of these strategies were associated with better performance. In this regard, a multiple regression was conducted to determine the degree to which sensitivity of discrimination may be predicted by the factor scores for each participant. Surprisingly, the overall regression equation failed to account for a significant proportion of the variance in performance (Adjusted *R2* = -.025, *F*(3, 96) < 1, *ns*). Moreover, none of the factors were significant predictors of performance (Emotion: *beta* = -.06; General Attributes: *beta* = .04; Specific Characteristics: *beta* = -.04, all *t*s(1) < 1, *ns*).

As a whole, the results of Experiment 2 suggested that participants were again able to match static faces to voices at above-chance levels, and this confirmed the robustness of the methodology when examining static face-voice matching. However, sex of speaker, sex of listener, stimulus distinctiveness, and self-reported strategies all failed to account for this level of performance. In considering the results both of Experiments 1 and 2 however, it was clear that whilst performance was above chance, it was still far from perfect. Moreover, a notable response bias was revealed in both studies suggesting that performance may, at least in part, be attributable to a liberal criterion leading participants to report that the faces and voices matched. Given this, Experiment 3 employed a change of methodology in the hope of providing a more appropriate test of face-voice matching in which potential floor effects may be avoided and the response bias may be addressed. Specifically, it was reasoned that if performance could be improved by such a change, then the factors guiding performance may emerge.

Experiment 3

Experiment 3 employed a 4AFC target-present lineup task in which participants were asked to listen to a voice whilst viewing the faces of four individuals. The target face was always present, and the participants’ task was to indicate which of the four faces they thought corresponded to the speaker. The move to a lineup task provided a more applied test of face-voice matching. Additionally, however, it provided a method in which the previous response bias may be addressed (as the target was always present).

*Design*

A 4-person target-present lineup task was used in which participants listened to a voice whilst looking at four simultaneously presented faces. Their task was to indicate which face belonged to the voice that they heard. Accuracy and self-rated confidence were recorded and represented the dependent variables.

*Participants*

A total of 48 university-aged participants (8 males, 40 females) took part in the present study in return for course credit. Ages ranged from 18 to 49 years (*M* = 21.4 years, *SD* = 6.4) and all participants reported normal hearing and vision, and no familiarity with the stimuli used as targets and foils. No participants had taken part in either of the previous studies.

*Materials*

A total of 64 target voices were used in the current study, representing a subset of those used in Experiments 1 and 2. These were selected as before from the SuperIdentity Stimulus Database ([www.southampton.ac.uk/superidentity](http://www.southampton.ac.uk/superidentity)) on the basis of the distinctiveness ratings provided by the experimenters using a 7-point scale. As before, all targets were drawn from a similar demographic, with ages ranging from 18 to 24 years. This resulted in 32 distinctive sounding speakers (all stimuli were rated > 4 on the 7-point scale; *M* = 6.07, *SD* = .74) and 32 typical sounding speakers (all stimuli were rated below 3.5 on the 7-point scale; *M* = 2.62, *SD* = .51). An independent samples *t*-test confirmed a significant difference in rated distinctiveness across the two voice sets (*t*(62) = 21.76, *p* < .001). Within each voice set, half the stimuli were male and the remainder were female. In addition, none had distinguishing visual features such as facial hair or spectacles, and none had distinguishing vocal features such as accents, stutters or lisps. The entire voice set was divided into two groups (A and B). Half the participants completed the task with stimulus set A whilst the remainder used stimulus set B. This measure was taken to ensure as large a stimulus pool as possible whilst limiting the length of the participants’ task.

All voices were processed within Audacity 3.1.0 as in Experiments 1 and 2. The faces of these target speakers also served as stimuli within the current study, along with the faces of 32 foils who represented the alternates in the 4AFC lineup trials. All faces were processed within Photoshop as in Experiments 1 and 2.

Given these stimuli, 32 4AFC lineups were constructed using 16 distinctive and 16 typical stimuli as targets. Care was taken to use same-sex foil faces to avoid a trivially easy and biased lineup task. In addition, each target was yoked to another same-sex target, so that each served as a foil for the other. As a result, every target appeared in two lineup trials (once as the target and once as a foil). Importantly, however, every foil face was also shown in two lineup trials so that frequency of exposure was equivalent for all faces and could not be used as a basis for selection or rejection in the lineup task.

As in Experiments 1 and 2, stimuli were presented and data were recorded using iSurvey allowing the presentation of multimedia trials, and enabling the participants to listen to the voice as many times as they wished.

*Procedure*

Following the provision of informed consent, participants received on-screen instructions introducing them to the lineup task. This was followed by the presentation of a randomised order of 32 trials during which participants could simultaneously view the faces of 4 individuals whilst listening to a target voice. The participants’ task was to indicate, by means of a button press, which face they thought belonged to the voice they heard, and they had the capacity to replay the voice as many times as they wished prior to response. Participants then responded by pressing the numbered key corresponding to the numbered position of the chosen face in the lineup. A forced-choice task was used meaning that participants had to make a lineup selection, and no feedback was provided. Finally, participants provided a self-rated indication of confidence in their decision by pressing a numbered key between 1 (not at all confident) and 7 (very confident indeed).

The entire experiment lasted no more than 20 minutes after which participants were thanked and debriefed.

Results and Discussion

Given the use of a 4AFC lineup task, the concerns over response bias were reduced in this study. Consequently, accuracy of performance with distinctive and with typical sounding voices represented the dependent variable. At the outset, analyses were conducted to test for any effects of stimulus set. These revealed no difference in performance according to whether participants had heard targets voices from set A or set B (all *t*s(46) < 1.00, *ns*). Consequently, all subsequent analyses were conducted collapsing across this factor. The data are summarised in Table 3. As in previous studies, given concern over variance across the stimulus set, analyses were conducted both by-participants and by-items.

*Accuracy of Performance and Effects of Distinctiveness*

With a 4AFC target-present lineup, the likelihood of choosing the correct face by chance alone was .25. Comparison to this standard by means of a one-sample *t*-test confirmed that performance overall was significantly better than chance (*M* = .32; by-participants: *t*(47) = 5.70, *p* < .001; by-items: *t*(63) = 3.39, *p* = .001). However, as with the previous studies, performance varied across the entire set of 64 speakers, ranging from 0% to 79% (*M* = 32%, *SD* = .17), with 42/64 voices (65.6%) eliciting a line-up performance above chance. Analysis via a *t*-test revealed no difference in performance when the target voice was male (*M* = .33) than when female (*M* = .32) (by-participants: *t*(47) < 1, *p* = .69; by-items: *t*(62) < 1, *p* = .83 ). However it did reveal significantly better performance when the target voice was distinctive (*M* = .37) than when typical (*M* = .28) (by-participants: *t*(47) = 3.89, *p* < .001; by-items: *t*(62) = 2.14, *p* = .04). Indeed, performance was no better than chance when the target voice was typical (by-participants: *t*(47) = 1.61, *p* = .11; by-items: *t*(31) = 1.12, *p* = .27).

(Please insert Table 3 about here)

*Self-Rated Confidence*

Self-rated confidence was examined following responses to distinctive voice lineups and typical voice lineups. A repeated-measures *t*-test confirmed that not only was performance better, but participants were also more confident when voices were distinctive than when typical (by-participants: *t*(47) = 3.00, *p* = .004: by-items: *t*(62) = 2.48, *p* = .016). This was important given that self-rated confidence may determine the likelihood with which someone is either willing to report to the police or is willing to testify in a court setting.

Taken together, these data suggested that when task demands were altered to maintain a zero memory load, reduce floor effects and address the response bias, static face-voice matching was again significantly above chance levels. Additionally, it was seen to be affected by the distinctiveness of the target voice, with performance being more accurate, and participants indicating more confidence, when the target voice was distinctive than when typical. This finding may go some way to explaining the demonstrations of item effects in previous studies.

General Discussion

The experiments reported here have each tested static face-voice matching using a large set of speakers uttering 4-5 second sentences. In addition, each experiment was designed to remove memory demands entirely. As such, the limitations associated with previous studies were addressed here providing an optimal test of static face-voice matching. Across all studies, the results consistently revealed a capability to match an unfamiliar voice to its face even when the face was presented as a static photograph. Moreover, using the optimal design with a large set of stimuli and no memory demands, performance in matching trials reached 70% which is comparable with the highest levels of performance reported in previous studies. In evaluating these results, it was considered unlikely that successful matching was achieved on the basis of matching the perceived age of the face and voice, as the stimuli were drawn from a relatively narrow age range. These results assume importance because they confirm that an explanation of face-voice matching cannot solely rest on the use of spatiotemporal information associated with speech events. Indeed, any explanation must also concede that the voice and the static face share sufficient information to support matching.

The other notable finding within the current paper is the demonstration that, given the right experimental conditions, the characteristics affecting face-voice matching can start to be revealed. In this regard, Experiment 3 highlighted the importance of vocal distinctiveness both in terms of accuracy of matching and confidence of the matcher. This result sits well alongside a growing literature describing a distinctiveness-advantage when processing voices. For example, distinctive voices are more easily recognised (Mullenix, et al., 2011), especially when they are of the opposite gender to the listener (Skuk & Schweinberger, 2013). Distinctive voices also serve as better cues for semantic and episodic retrieval (Barsics & Brédart, 2012), better withstand interference from distractors (Neil & Stevenage, 2014) and may better survive the effects of facial overshadowing (Tomlin, 2015). The current results add another dimension to this distinctiveness advantage in that distinctive voices also appear to be more easily matched to their faces compared to their typical sounding counterparts. Importantly, this may facilitate cross-modal integration of faces and voices such that multimodal person perception is advantaged when the voice is distinctive.

When considering the effect of distinctiveness on face-voice matching performance, some consideration should be given the discrepancy in results across the same/different task (Experiment 2) and the lineup task (Experiment 3). Whilst both methodologies held the memory load constant and low, through the use of a simultaneous presentation of voice and candidate face(s), there was evidence of a response bias in the same/different task which echoed the concerns of Smith et al. (2016b). The lineup task may have reduced the effect of such a bias, and this difference may have allowed the distinctiveness effect to emerge. However, it is also possible that the demographic of the participants was better matched to that of the stimuli in Experiment 3 allowing vocal distinctiveness to be better appreciated. Such an explanation suggests that a distinctiveness effect may indeed emerge in other tasks when speaker and listener demographics are better aligned, and further work is encouraged to explore this possibility.

*Face-Voice Matching through Overlapping Cues*

Several explanations may be offered to account for the current demonstration of static face-voice matching. The remainder of this discussion is given to a consideration of two accounts. The first rests on the suggestions of Mavica and Barenholtz (2013) and Smith et al. (2016a, b) that matching may be based on the overlapping cues that faces and voices present. Static face-voice matching may be served by overlapping cues when perceiving physical or personality based characteristics (perhaps along the lines indicated by Smith et al*.* (2016a)). Dynamic face-voice matching may be served by the addition of another complementary set of overlapping cues which draw on the common act of articulation. Distinctive targets may facilitate this face-voice matching through making those overlapping cues more obvious.

According to this overlapping cues account, one might expect that face-voice matching would be optimal when both static and dynamic sets of overlapping cues are available (see also Smith et al., 2016b). Unfortunately, however, available tests of static versus dynamic face-voice matching leave this issue unanswered as they have either incorporated too few speakers, or have imposed a memory demand, each of which may have biased against performance in the static face-voice matching condition. The only study which addressed all of these issues was provided by Mavica and Barenholtz (2013). However, their use of static facial images only means that the question of whether there is indeed any advantage of dynamic over static images in a face-voice matching task, using a good methodological approach, remains open.

*Face-Voice Matching through Face-Voice Integration*

The second account of successful face-voice matching rests on the possibility of face-voice integration and sits within the larger context of person-perception. This face-voice integration account is not necessarily incompatible with the overlapping cues account described above. Indeed, the overlapping cues may support face-voice integration. Nevertheless, a distinct literature exists on face-voice integration, and it is useful to reflect on this.

Recent reviews are provided by Campanella and Belin (2007) and Mathias and von Kriegstein (2014) who both lay out the behavioural and the neurological evidence supporting face-voice integration. Behavioural evidence for audio-visual integration comes from studies of cross-modal priming in which prior presentation of a familiar face can speed the recognition of the corresponding voice, and vice versa (Ellis, Jones & Mosdell, 1997; Schweinberger, Herlholz & Stief, 1997; see also Stevenage, Hugill & Lewis, 2012). In addition, the presence of a face enhances voice learning (Sheffert & Olson, 2004; Zäske, Mühl & Schweinberger, 2015), and the co-presentation of face and voice enhances familiarity judgements compared to the presentation of the voice alone (Joassin, Maurage, Bruyer, Crommelinck & Campanella, 2004). More recently, Schweinberger, Robertson and Kaufmann (2007) used a speeded recognition task to demonstrate costs as well as benefits in face-voice integration when the face and voice were co-presented, relative to when the voice alone was heard. Costs emerged when the face and voice came from different identities whilst benefits emerged when they came from the same identity. The current results add to this behavioural evidence for integration

All of these results point to face-voice integration in person perception. However, they do not answer the issue of *when* this integration occurs. Two possibilities exist: integration may occur at an early stage through direct communication between unimodal areas processing the face and the voice. Alternatively, integration may occur at a later, supramodal stage once unimodal recognition has been achieved. In support of supramodal integration, Joassin et al. (2004) measured Event Related Potentials (ERPs) to reveal early activity in the visual cortex (110ms) and in the auditory cortex (170ms) followed by later activity in an area best understood as a region of multimodal convergence (270ms). These may equate to activation of the face representation the voice representation and a region reflecting the supramodal integration respectively.

In contrast, compelling neural evidence for early face-voice integration is presented by von Kriegstein and colleagues. Using fMRI, she demonstrated activation in the area believed to be specialised for face processing, in response to the presentation of a visually familiar speaker’s voice (von Kriegstein & Giraud, 2006; von Kriegstein, Kleinschmidt, Sterzer & Giraud, 2005 see also Schall, Kiebel, Maess & von Kriegstein, 2013). In addition, functional connectivity analysis was able to demonstrate direct coupling between the face and the voice processing areas. Perhaps most importantly, the time-course of activation in these two areas was strongly correlated, and the variance in activation of one area was largely accounted for by that of the other area. In a similar vein, Schweinberger, Kloth and Robertson (2011) used ERP recordings to show that early neural markers were affected by co-presentation of face and voice. Indeed, in comparison to unimodal presentation of face *or* voice, the face-voice combination elicited a faster early response and elicited a larger neural marker of recognition – the N170.

Together, these results strongly indicate an early, direct integration of face and voice processing rather than a later supramodal integration. Added to these demonstrations of functional connectivity, Blank, Anwander and von Kriegstein (2011) used diffusion fMRI to investigate the existence of a structural link between the face and voice areas of the brain. Their results are the first to reveal white matter connections providing a possible structural architecture to support direct face-voice integration.

Taken together, there is now a body of behavioural evidence to suggest that faces and voices may be integrated at an early stage to support multimodal person perception. Successful face-voice matching as demonstrated here forms part of this evidence base but cannot contribute to a determination of whether that integration occurs early or late. This said, there is also a compelling body of neural evidence to support an early integration of face and voice through direct functional and structural connections between the face and voice areas of the brain. Optimal processing occurs when all areas are connected and fully functional, with benefits seen in the difficult task of voice learning, and in tasks of face and voice recognition either in combination or in isolation.

Implications

Given the benefits that can occur for person perception through the perception of overlapping cues or through the integration of multiple modalities, the current results may hold value in applied contexts such as witness identifications. If a witness has heard the voice of a perpetrator, they may be invited to take part in voice lineup. However, presently, there is very little research detailing how that lineup should be conducted. The best practice guidelines draw heavily on the processes underpinning a visual lineup despite the fact that voices differ from faces in important ways. Currently, the available research highlights the importance of holding constant the *manner* in which lineup members speak (Lander et al., 2007), and notes that it may be more important to hold the manner of speech constant than to hold the message constant. The present data add to these recommendations through highlighting care over voice lineups if the face of the speaker is also available. Under such circumstances, the present data may suggest a benefit if the face and (witnessed) voice appear to come from the same person. By extension, there may be a cost to performance if the face and (witnessed) voice appear not to come from the same person. Future work would be well-directed to examine the influence of face-voice combination on aspects such as hit and false alarm rates during lineup tasks, in order that factors which help performance can be identified, and factors which hinder performance may be avoided.

*Conclusions*

Static face-voice matching has been explored here across three experiments and across two tasks. Importantly, these tasks removed all memory demands, and used a good number of targets to minimise the influence of item effects. In every case, static face-voice matching was demonstrated at above-chance levels, confirming the previous literature. Here, however, the results suggested a role for vocal distinctiveness in facilitating performance. In accounting for these results, it was possible that more distinctive voices enabled the better perception of overlapping cues, or the better integration of face and voice at functional and neural levels. These results may be set in the context of theoretical discussions regarding multimodal integration in person perception. However, they may also hold value for more applied situations in which optimal earwitness performance is sought.

References

Alport, G.W., & Cantrill, H. (1934). Judging personality from voice. *Journal of Social Psychology, 5,* 37-55.

Barsics, C., & Brédart, S. (2012). Access to semantic and episodic information from faces and voices: Does distinctiveness matter? *Journal of Cognitive Psychology, 24(7)*, 789-795. doi: 10.1080/20445911.2012.692672

Blank, H., Anwander, A., & von Kriegstein, K. (2011). Direct structural connections between voice- and face-recognition areas. *The Journal of Neuroscience, 7*, 12906-12915.

Bruce, V., Burton, A.M., & Hancock, P. (1995). Missing dimensions of distinctiveness. In T. Valentine (Ed.). *Cognitive and Computational Aspects of Face Recognition*. London: Routledge.

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

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

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, 143-156.

Hauser, D.J., & Schwarz, N. (2016). Attentive turkers: MTurk participants perform better on online attention checks than subject poor participants. *Behavior Research Methods, 48(1)*, 400-407. doi:10.3758/s13428-015-0578-z

Joassin, F., Maurage, P., Bruyer, R., Crommelinck, M., & Campanella, S. (2004). When audition alters vision: An event-related potential study of the cross-modal interactions between faces and voices. *Neuroscience4 Letters, 369*, 132-137.

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.

Lachs, L., & Pisoni, D.B. (2004). Crossmodal source identification in speech perception. *Ecological Psychology, 16(3)*, 159-187. doi: 10.1207/s15326969eco1603\_1

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.G., & 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

Mathias, S.R., & von Kriegstein, K. (2014). How do we recognise who is speaking? *Frontiers in Bioscience, S6, 92-109.*

Mavica, L.M., & 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

Mullenix, J.W., Ross, A., Smith, C., Kuykendall, D., 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

Neil, G.J., & Stevenage, S.V. (2014). Protecting voices from the effects of interference. Presentation at the British Psychological Society Cognitive Section Annual Conference, Nottingham Trent University, 3-5th September 2014.

Paolacci, G., & Chandler, J. (2014). Inside the turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science, 23,* 184-188.

Schall, S., Kiebel, S.J., Maess, B., & von Kriegstein, K. (2013). Early auditory sensory processing of voices is facilitated by visual mechanisms. *Neuroimage, 77*, 237-347.

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

Schweinberger, S.R., Kloth, N., & Robertson, D.M.C. (2011). Hearing facial identities: Brain correlates of face-voice integration in person identification. *Cortex, 47(9)*, 1026-1037. doi: 10.1016/j.cortex.2010.11.011

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

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

Skuk, V.G., & Schweinberger, S.R. (2013). Gender differences in familiar voice identification. *Hearing Research, 296*, 131-140.

Smith, H.M.J., Dunn, A.K., Baguley, T.S., & Stacey, P.C. (2016a). Concordant cues in faces and voices: Testing the backup signal hypothesis. *Evolutionary Psychology, 14*, 1-10. doi: 10.1177/1474704916630317

Smith, H.M.J., Dunn, A.K., Baguley, T.S., & Stacey, P.C. (2016b). Matching novel face and voice identity using static and dynamic facial images *Attention, Perception and Psychophysics, 78*, 868-879. doi: 10.3758/s13414-015-1045-8

Stevenage, S.V., Hugill, A., & 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

Sumby, W.H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of Acoustic Society of America, 26*, 212-215.

SuperIdentity Stimulus Database: [www.southampton.ac.uk/superidentity](http://www.southampton.ac.uk/superidentity)

Tomlin, R. (2015). Tests of voice recognition with ecologically valid conditions. *Unpublished MSc Dissertation,* University of Southampton.

Von Kriegstein, K., & Giraud, A-L. (2006). Implicit multisensory associations influence voice recognition. *PLoS Biology, 4*, e326.

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.

Zäske, R., Mühl, C., & Schweinberger, S.R. (2015). Benefits for voice learning caused by concurrent faces develop over time. *PLoS ONE*, 10(11), e0143151. doi: 10.1371/journal.pone.0143151

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

Zäske, R., Skuk, V.G., Kaufmann, J.M., & Schweinberger, S.R. (2013). Perceiving vocal age and gender: An adaptation approach. *Acta Psychologica, 144(3)*, 583-593.

Table 1:

Mean sensitivity of discrimination (d’) and response bias (C) on the simultaneous matching task in Experiment 1, together with proportion of accurate decisions on ‘matching’ and ‘mismatching’ trials (standard deviation in parenthesis).

|  |  |
| --- | --- |
|  | Performance on simultaneous  matching task |
| Sensitivity of Discrimination (d’) | .16 (.35) |
| Response Bias (C) | -.48 (.59) |
| Proportion Accurate on ‘matching’ trials | .68 (.14) |
| Proportion Accurate on ‘mismatching’ trials | .38 (.16) |

Table 2:

Mean sensitivity of discrimination (d’) and response bias (C) on the simultaneous matching task in Experiment 2, together with proportion of accurate decisions on ‘matching’ and ‘mismatching’ trials for male and female speakers and listeners (standard deviation in parenthesis).

|  |  |  |
| --- | --- | --- |
|  | Male Speakers | Female Speakers |
| Sensitivity of Discrimination (d’)  Male Listeners  Female Listeners | .25 (.48)  .20 (.39) | .18 (.39)  .06 (.45) |
| Response Bias (C)  Male Listeners  Female Listeners | -.36 (.38)  -.42 (.48) | -.50 (.44)  -.52 (.49) |
| Proportion Accurate on ‘matching’ trials  Male Listeners  Female Listeners | .67 (.13)  .68 (.15) | .70 (.13)  .68 (.15) |
| Proportion Accurate on ‘mismatching’ trials  Male Listeners  Female Listeners | .42 (.16)  .40 (.16) | .36 (.14)  .34 (.15) |

Table 3:

Mean accuracy in the 4AFC target-present lineup in Experiment 3 for distinctive and typical speakers (standard deviation in parentheses).

|  |  |  |
| --- | --- | --- |
|  | Distinctive Speakers | Typical Speakers |
| Accuracy of Performance (chance = .25) | .37 (.17) | .28 (.12) |
| Self-rated Confidence (out of 7) | 4.33 (1.02) | 4.16 (1.10) |

Figure 1: Mean proportion of accurate decisions for static face-voice matching in ‘match’ trials and in ‘mismatch’ trials across Experiment 1, and across male and female listeners and male and female speakers in Experiment 2 (error bars show standard deviation). Dotted lines indicated chance level performance of 0.5.

Appendix 1:

List of participant strategies from Experiment 2, together with their loadings on each of the three factors of Emotion, General Physical Attributes, and Specific Physical Attributes.

|  |  |  |  |
| --- | --- | --- | --- |
| Participant Strategies | Emotion | General Attributes | Specific Attributes |
| I looked to see whether a positive face might match with a positive voice | .74 |  |  |
| I tried to see if they had a friendly face and see if the voice matched in terms of friendliness (i.e. jolly and quite high pitched) | .72 |  |  |
| I expected that someone who looked confident would have a lively, energetic or enthusiastic sounding voice | .69 |  |  |
| I looked at the expression on the person's face and tried to match it to the expression I could hear in their voice | .68 |  |  |
| I paid attention to the warmth of their facial expressions | .64 |  |  |
| When the person was smiling, I expected a more 'exciting' voice | .63 |  |  |
| I thought that a confident pose in the photograph would probably correlate with how confident the voice sounded | .61 |  |  |
| When the voice was slower in pace, I expected a photo of a person who wasn't very happy or excited | .60 |  |  |
| I paid attention to the speed of their speech | .58 |  |  |
| I thought that how confident a person looked would be related to the volume and speed with which they spoke | .57 |  |  |
| I used facial expressions | .55 |  |  |
| I paid attention to how loud and how fast their speech was compared to how prominent their facial features were | .50 |  |  |
| I expected that the more shy the person looked, the quieter the voice may be | .45 | .35 |  |
| I expected that someone who looked shy would have a quieter or more monotone voice | .39 | .33 |  |
| I expected that females with delicate features would have higher-pitched voices |  | .73 |  |
| For men, I looked at how masculine they looked compared to their voice |  | .70 |  |
| I thought that some of the larger-set people seemed to 'fit' deeper voices better than the smaller, petite people |  | .66 |  |
| I paid attention to their build |  | .64 |  |
| I tried to link the tone of the voice to the size of the person |  | .64 |  |
| I expected that men with 'manly' faces (such as with chiselled jaw or possibly facial hair) would have a rougher tone of voice |  | .62 |  |
| If a man had strong facial features, then I would expect a voice to be unique and possibly deeper |  | .45 |  |
| When the facial characteristics were softer, I expected a high pitched voice |  | .44 |  |
| I imagined them speaking with that voice |  | .42 |  |
| I gave a rough estimate of age and tried to see if the tone of voice matched, i.e. younger people might have a higher pitched or more 'jolly' voice |  | .37 |  |
| I used existing stereotypes that I had of what voices should match what faces |  | .37 |  |
| I paid attention to whether the tone of voice matched the facial characteristics |  | .36 |  |
| I paid attention to the pitch of the voice |  | .35 |  |
| I expected that an older person would have a rougher sounding voice |  | .32 |  |
| I paid attention to the thickness of their lips |  |  | .84 |
| I expected that the appearance of the lower lip might relate to how timid or clear the voice sounded |  |  | .73 |
| I expected that someone with bigger lips might say the word 'shop' with an emphasis on the 'p' |  |  | .70 |
| I paid attention to their mouth shape |  |  | .63 |
| I thought that nasal sounding voices would go with people that had smaller noses |  |  | .67 |
| I expected that the more styled a man's hair was, the higher their pitch of voice may be |  |  | .55 |
| I paid attention to the size of their jaw |  |  | .51 |
| I expected that men with a 'rounder' face would have a more friendly, high-pitched voice |  |  | .51 |
| I paid attention to the Adam's apple and the neck |  |  | .47 |
| I tried to determine whether they were British or not based on their facial features. If not, a very English accent would not be theirs |  |  | .43 |
| I paid attention to whether they looked foreign or had a slight accent |  |  | .42 |
| I expected that age would affect their pronunciation of words |  | .36 | .36 |

Note: Factor loadings below .32 are not displayed.

1. Whilst concerns have been expressed over the quality of data obtained from MTurk participants (see Paolacci & Chandler, 2014), valid data may be obtained even for study designs that are traditionally laboratory-based. Indeed, Hauser and Schwarz (2016) demonstrated that participants recruited via MTurk were more attentive to task demands than subject-pool participants. Within the current studies, performance was screened to ensure compliance with the task demands, and no participants needed to be dropped from analyses due to their identification as outliers. Importantly, performance of the MTurk participants in Experiment 1 showed no statistical difference to that of the 20 pilot participants tested with the same task under face-to-face conditions in Experiment 2. These measures gave us confidence that the data provided via MTurk were reliable. [↑](#footnote-ref-1)