The effect of inserting an inter-stimulus interval in face-voice matching tasks

Journal:	*Quarterly Journal of Experimental Psychology*
Manuscript ID	QJE-STD 16-079.R2
Manuscript Type:	Standard Article
Date Submitted by the Author:	09-Sep-2016
Complete List of Authors:	Smith, Harriet; Nottingham Trent University, Psychology Dunn, Andrew; Nottingham Trent University, Psychology Baguley, Thomas; Nottingham Trent University, Psychology Stacey, Paula; Nottingham Trent University, Psychology
Keywords:	face-voice matching, static face, inter-stimulus interval, person perception

The effect of inserting an inter-stimulus interval in face-voice matching tasks

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

Nottingham Trent University, UK

Correspondence concerning this article should be addressed to Harriet M. J. Smith, Psychology Division, Nottingham Trent University, Burton Street, Nottingham, NG1 4BU. Telephone number: +44 (0) 115 941 8418. Email: harriet.smith02@ntu.ac.uk

Acknowledgement: This work was supported by a Ph.D. studentship from the Division of Psychology, Nottingham Trent University.

Abstract

Voices and static faces can be matched for identity above chance level. No previous face-voice matching experiments have included an inter-stimulus interval (ISI) exceeding 1 second. We tested whether accurate identity decisions rely on high-quality perceptual representations temporarily stored in sensory memory, and therefore whether the ability to make accurate matching decisions diminishes as the ISI increases. In each trial, participants had to decide whether an unfamiliar face and voice belonged to the same person. The face and voice stimuli were presented simultaneously in Experiment 1, there was a 5 second ISI in Experiment 2, and a 10 second interval in Experiment 3. The results, analysed using multilevel modelling, revealed that static face-voice matching was significantly above chance level only when the stimuli were presented simultaneously (Experiment 1). The overall bias to respond *same identity* weakened as the interval increased, suggesting that this bias is explained by temporal contiguity. Taken together, the findings highlight that face-voice matching performance is reliant on comparing fast-decaying, high-quality perceptual representations. The results are discussed in terms of social functioning.

Keywords: face-voice matching, static face, inter-stimulus interval, person perception

The effect of inserting an inter-stimulus interval in face-voice matching tasks

Whilst some studies have found that unfamiliar face-voice matching accuracy depends on pairing visually encoded articulatory movement to auditory speech (Kamachi, Hill, Lander & Vatikiotis-Bateson, 2003; Lachs & Pisoni, 2004a; Lander, Hill, Kamachi & Vatikiotis-Bateson, 2007), others have observed that voices and static faces can be accurately matched above-chance level (Krauss, Freyberg & Morsella, 2002; Mavica & Barenholtz, 2013; Smith, Dunn, Baguley & Stacey, 2016a). Based on the results of 3 experiments Smith, Dunn, Baguley and Stacey (2016b) concluded that source identity information is shared by voices and faces regardless of whether the faces are static or dynamic (i.e. articulating but muted). The balance of evidence suggests that voices and static faces do provide sufficient concordant identity information (Smith et al., 2016a) so that it is possible to accurately match an unfamiliar face to a voice (Smith et al., 2016b).

All previous tests of face-voice matching have presented faces and voices close together in time, with a maximum 1-second (s) inter-stimulus interval (ISI) (Kamachi et al., 2003; Krauss et al., 2002; Lachs & Pisoni, 2004a, 2004b; Lander et al., 2007; Mavica & Barenholtz, 2013; Smith et al., 2016a, 2016b). Whilst this has been insightful, in everyday social interactions faces and voices belonging to the same person might be separated by longer intervals of time. For example, in a crowded place it could feasibly take significantly longer than 1 s to shift attention towards an unfamiliar speaker. Furthermore, any bias affecting performance may be dependent on time-course because the face and voice of the same person tend to be experienced close together in time (and space). With the aim of further understanding the cognitive processes underlying face-voice matching decisions we addressed this issue in a series of 3 experiments. To provide a baseline, static faces and voices were presented simultaneously in Experiment 1. In the next 2 experiments we temporally offset face and voice stimuli by 5 s (Experiment 2) and 10 s (Experiment 3) to measure the effect of the temporal offset on matching accuracy and response bias.

Sensory memory and face-voice matching accuracy

Our aim to test whether temporally separating faces and voices undermines matching accuracy is motivated by the sensory memory literature. Presenting to-be-compared stimuli within a short time frame likely facilitates appraisals based on high-quality (i.e. detailed and accurate) perceptual representations of faces and voices. Precise representations of both visual and auditory information in sensory memory degrade quickly. Iconic memory typically lasts for a few hundred milliseconds (ms) (Coltheart, 1980; Neisser, 1967; Sperling, 1960), although recent evidence has been put forward for the existence of an intermediate, high capacity visual store which enables highly detailed visual information to persist for up to 4s with the help of afterimages (Sligte, Scholte & Lamme, 2008, 2009). The time-course of auditory representation decay is longer: echoic memory persists for longer than iconic memory (Crowder & Morton, 1969; Penney, 1985), up to a period of about 5 s (Glanzer & Cunitz, 1966; Lu, Williamson & Kaufman, 1992; Treisman, 1964; Wickelgren, 1969). Short ISIs of 500ms (Kamachi et al., 2003; Krauss et al., 2002; Lachs & Pisoni, 2004a; Lander et al., 2007; Mavica & Barenholtz, 2013) and 1 s (Smith et al., 2016a, 2016b) are likely within the limits of both iconic and echoic memory, meaning that high-quality representations of faces and voices can be compared for source-identity information. This might facilitate accurate identity matches.

Response biases in face-voice matching

Inserting a longer (>1 s) ISI in novel face-voice matching tasks may also affect response bias (i.e. an overall tendency to respond that faces and voices belong to the same or different identities). Assumptions of common identity should be more likely when faces and voices are presented within a brief time frame. When two events are presented close together in time, attributions of causality tend to be inferred; a 2 second window appears to be the crucial time period within which stimuli are bound together in this way (Reed, 1992; Shanks, Pearson & Dickinson, 1989).

As well as being relevant to causality judgements, temporal contiguity is clearly also important in face and voice processing (Stevenage, Neil & Hamlin, 2014). The research on audio-visual speech perception suggests that face-voice speech integration occurs when faces and voices are presented within a short temporal window (Munhall, Gribble, Sacco & Ward, 1996; Robertson & Schweinberger, 2010; Van Wassenhove, Grant & Poeppel, 2007). There might be a corresponding temporal window during which people exhibit a bias to attribute a novel face and voice to the same identity.

The hypothesis that biases are influenced by the time-course of stimulus presentation is supported by previous face-voice matching studies. Using a same-different task, with a 1 s interval between presentation of the face and voice, Smith et al. (2016a) demonstrated that response bias is prominent characteristic of face-voice matching performance. In each trial, the participants had to decide whether a face and voice belonged to the same or different identities. The results pointed to the existence of a bias to respond *same identity*, particularly when participants saw a face before hearing a voice. Smith et al. (2016a) also found that matching accuracy was higher on same identity than different identity trials, hinting at a general overall bias to respond *same*. Such a response bias might be dependent on the face and voice being presented close together in time. This hypothesis can be tested using 2AFC methodologies. The participants see a single face and have to decide which 1 of 2 voices belongs to the same identity, or they hear a single voice and have to decide which 1 of 2 faces belongs to the same identity. Consistent with the conclusion that response bias depends on temporal proximity, Smith et al. (2016b) found an effect of temporal position; in 2AFC face-voice matching tasks, people tended to accept the faces and voices presented closest together in time as belonging to the same identity. The participants were also more likely to accept the first of two face-voice combinations they encountered as sharing a common identity.

Aims

In 3 experiments, we addressed how face-voice matching performance operates when faces and voices are presented simultaneously (Experiment 1), when there is an inter-stimulus of 5 s (Experiment 2) and when there is an interval of 10 s (Experiment 3). In order to measure response bias, a same-different procedure was adopted in all 3 experiments (Stanislaw & Todorov, 1999).

Experiment 1

Faces and voice were presented simultaneously in Experiment 1 to provide a baseline of performance when the matching task does not impose a load on memory, and also to test static face-voice matching accuracy in the light of the previous contradictory results (Kamachi et al., 2003; Krauss et al., 2002; Lachs & Pisoni, 2004a; Lander et al., 2007; Mavica & Barenholtz, 2013; Smith et al., 2016a, 2016b). Taking the existing evidence together as a whole, particularly Smith et al.'s (2016a) observation of above chance performance when a same-different procedure featured a 1 s ISI, we expected static face-voice matching to be significantly above chance level (50%). We also expected for there to be an overall bias to attribute the face and voice to a common identity.

Method

Design. In Experiment 1, identity (same or different) was manipulated within subjects. For the matching accuracy analysis, the dependent variable was accuracy. For the matching response analysis, which addressed response bias, the dependent variable was a *same identity* response.

Participants. There were 6 male and 18 female participants (*N* = 24), with an age range of 18 – 32 years (*M* = 20.79, *SD* = 4.0). They were recruited from the Nottingham

Trent University Psychology Division's Research Participation Scheme. Participants received research credits in return for participation. Ethical approval for all 3 experiments was granted by the university's Business, Law and Social Science College Research Ethics Committee (ref: 2013/37).

Apparatus and materials. The experiment featured 18 speakers (9 male and 9 female) from the GRID audio-visual sentence corpus (Cooke, Barker, Cunningham & Shao, 2006), which contains videos of British adults, each saying a unique 6-word nonsense sentence. The speakers are only visible from the shoulders up. The speakers selected from the corpus were white, British, between the ages of 18 and 30, and spoke with an English accent. The stimuli (static faces and voices) were the same as those used in previous face-voice matching studies (Smith et al., 2016a, 2016b). Two videos (.mpegs) for each speaker were selected at random from numbered files using an online research randomiser (Urbaniak & Plous, 2013). One of the 2 videos was used to create static pictures of faces, which were 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. Each of the static images measured 368 x 288 pixels and was presented in colour. The voices played from the second .mpeg file with the face not visible (audio quality: 256 kbits per second, 44,100 Hz, 16 bit). All of the stimuli (static images and voices) were each presented for 2 s in total.

The experiment was run using Psychopy v1.77.01 (Peirce, 2009). Participants completed the experiment on an Acer Aspire laptop (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 cm away from the edge of the desk at which the participants were seated. Voice recordings were presented binaurally at a comfortable volume through Sennheiser (HD205) headphones, which suppress external and ambient noise. The volume of the voice recordings ranged between 70 – 75 dB, and was measured using a Svantek (977) sound level meter, with the headphones placed over a G. R. A. S. (RA0039) artificial ear simulator. The sound intensity was kept constant across participants.

To maximise generalizability, a research randomiser (Urbaniak & Plous, 2013) was used to create 4 versions of the experiment; across versions, different combinations of faces and voices were encountered in same identity and different identity trials. Each of the 18 stimulus faces and voices only appeared once in a version, so each version consisted of 18 trials in total. There were 9 same identity trials, and 9 different identity trials. On different identity trials, both stimuli were matched for sex. Although the order of trials was always different, each individual trial (within a version) was the same.

Procedure. The participants were randomly allocated to one of the 4 versions of the experiment. The procedure used in Experiment 1 is illustrated in Figure 1. Participants saw a face and heard a voice presented simultaneously. The face-voice combination was presented for 2 s. After the combination had been presented, the participants were instructed to press '1' if they thought the face and voice belonged to the same identity, and '0' if they thought they were from different identities. The response buttons were not counterbalanced across participants because assigning responses in this way is intuitive. Whilst '1' corresponds to a positive response (i.e. identifying a match), '0' corresponds to identifying no match. The participants used the digit keys ('0' and '1') that appear horizontally above the letter keys. They were instructed to press '1' with their left index finger and '0' with their right index finger. No time pressure was imposed while they made this decision.

[FIGURE 1 ABOUT HERE]

Data analysis. This was a fully crossed design, with each participant encountering all stimuli (18 faces, 18 voices) throughout the experiment. Accounting for the variance associated with stimuli is crucial when investigating face-voice matching performance, because some people look and sound more similar than others (see Mavica & Barenholtz, 2013, Smith et al., 2016b). In order that both participants and stimuli could be treated as random effects, the data were analysed using multilevel models. This is the most appropriate analysis because it takes into account the variability associated with individual performance as well as different face and voice stimuli. This is superior to the common alternative of undertaking separate by-participant and by-item analyses (see Raaijmakers, 2003; Raaijmakers, Schrijnemakers, & Gremmen, 1999). The main advantages of multilevel modelling are that it avoids aggregating data (see Wells, Baguley, Sergeant & Dunn, 2013; Smith et al., 2016a, 2016b) and reduces the probability of committing a Type 1 error (Clark, 1973; Baguley, 2012; Judd, Westfall & Kenny, 2012).

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. This not appropriate with the current set of data, where it was necessary to avoid aggregation (Wright, Horry & Skagerberg, 2009). We took the hit rate (accuracy on same identity trials) and true negative rate (accuracy on different identity trials) as respective measures of sensitivity and specificity. The observed accuracy across same identity and different identity trials was compared against chance level performance (50%) in order to separate the signal from the noise. To measure the response bias, the percentage of *same identity* responses across all trials was compared against chance level.

Results

The overall accuracy (panel A) and the overall pattern of responses (panel B) for Experiment 1 (0 s ISI) are illustrated in Figure 2 by the left-most data points in each panel. This figure also presents data from Experiment 2 (5 s ISI) and Experiment 3 (10 s ISI).

[FIGURE 2 ABOUT HERE]

Matching accuracy. Overall accuracy was above chance level, *M* = 60.7%, 95% CI [54.6, 66.5]. The matching accuracy analysis was conducted using multilevel logistic regression with the lme4 version 1.06 package in R (Bates, Maechler, Bolker & Walker, 2014). Two nested models were compared, and both were fitted using restricted maximum likelihood. The dependent variable was accuracy (0 or 1). The first model included a single intercept, and the second model included the main effect of identity. Setting up the model in this way involves testing for individual effects in a similar way to *t*-tests or ANOVA. However, in all 3 experiments we report likelihood ratio tests provided by lme4 because these are generally more robust. In Experiment 1, the likelihood ratio test was obtained by dropping the null model from the main effect model. This revealed a significant effect of identity (b = 1.184, SE = 0.232, G^2 = 28.437, p<.001). In the main effect model the estimate of *SD* of the face random effect was 0.127 while for voice it was 0.142. The estimated *SD* for the participant effect was less than 0.001. A similar pattern held for the null model. Variability associated with the stimuli was much greater than variability at the level of individual differences.

Figure 3 shows the means and 95% confidence intervals for accuracy (%) in both conditions. Confidence intervals were obtained by simulating the posterior distributions of cell means in R (arm package, version 1.6) (Gelman & Su, 2013).

[FIGURE 3 ABOUT HERE]

Figure 3 reveals that the hit rate (same identity trials), *M* = 74.14%, 95% CI [67.2, 80.1] was consistently higher than the true negative rate (different identity trials), *M* = 46.57%, 95% CI [39.3, 54.21].

Matching response. The matching response analysis was conducted using the same method as the accuracy analysis. Overall, faces and voices were attributed to the same identity above chance level, *M* = 64.1%, 95% CI [56.8, 70.8].

Discussion

Face-voice matching accuracy was above chance level. This result replicates previous findings, and provides additional evidence for accurate static face-voice matching (Krauss et al., 2002; Mavica & Barenholtz, 2013; Smith et al., 2016a, 2016b). Higher accuracy on same identity than different identity trials is consistent with previous studies using a same-different face-voice matching procedure (Smith et al., 2016a). In line with predictions informed by the results of Smith et al. (2016a, 2016b), there was an overall bias to respond *same identity* when the face and voice were presented simultaneously.

Experiment 2

In Experiment 2, we used a same-different procedure (as in Experiment 1), but this time the face and the voice were separated by 5 s. An interval of 5 s 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).

Experiment 2 also differed from Experiment 1 in that we included a manipulation of stimulus presentation order. Previous sequential face-voice matching studies have either presented the face first (visual-auditory (V-A) condition) or the voice first (auditory-visual (A-V) condition) (Kamachi et al., 2003; Lachs & Pisoni, 2004a, 2004b; Lander et al., 2007; Smith et al., 2016a, 2016b). Although an effect of order has never been detected in terms of sensitivity (Kamachi et al., 2003; Lachs & Pisoni, 2004a, 2004b; Lander et al., 2007; Smith et al., 2016a, 2016b), people do seem to exhibit more of a bias to respond *same identity* when the face is presented first (V-A condition) (Smith et al., 2016a).

Possible order effects warrant further investigation, particularly when including intervals of an unprecedented duration (>1 s). The rationale for manipulating the order of stimulus presentation expressed in other studies (see Lachs & Pisoni, 2004a) focuses on face-voice asymmetries in terms of speech information, but it is also possible that differential memory for faces and voices will affect performance when the ISI is longer than 1 s. Voices are less well remembered (Stevenage, Hugill & Lewis, 2012; Stevenage & Neil, 2014), and more sensitive to interference (Stevenage, Howland & Tippelt, 2011) than faces. Therefore, it might be the case that performance is less accurate in the A-V condition when it is necessary to remember the voice for longer than the face.

Although we are unable to derive a strong prediction about the expected outcome based on the available literature, we did not anticipate that matching accuracy would improve as the interval increased to 5 s. Rather, if accurate face-voice matching relies on the ability to compare highly detailed representations of faces and voices, the accuracy levels observed in Experiment 1 are likely to be compromised when there is an ISI of 5 s. 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 diminish towards chance level.

Method

Apart from the following exceptions, the methods were identical to Experiment 1.

Design. The study employed a 2 x 2 within subject factorial design. The factors were identity (same or different) and order (visual to auditory (V-A) or auditory to visual (A-V)). For the matching accuracy analysis, the dependent variable was accuracy. For the matching response analysis, the dependent variable was a *same identity* response.

Participants. There were 24 participants (22 females and 2 males), with an age range of 18 to 35 years (*M* = 19.8, *SD* = 3.7). None had taken part in previous face-voice matching experiments undertaken in our lab.

Apparatus and materials. In Experiment 2, we used identical experiment versions to Experiment 1. As previous results indicate that some people look and sound more similar than others (Smith et al., 2016b), it was important to avoid confounds relating to new stimulus combinations.

Procedure. There were two counterbalanced experimental blocks. Each consisted of a practice trial, followed by 8 randomly ordered experimental trials. The procedure is illustrated in Figure 4. In the V-A block, participants saw the face first, and in A-V block they heard the voice first. All of the stimuli were presented for 2 s, and there was a 5 s ISI. In each trial, participants pressed '1' if they thought the face and voice belonged to the same identity, and '0' if they thought they belonged to different identities. They were not allowed to make a decision until they had seen both stimuli, and no time pressure was imposed.

[FIGURE 4 ABOUT HERE]

Results

The overall accuracy (panel A) and the overall pattern of responses (panel B) for Experiment 2 (5 s) are illustrated by the middle data points in Figure 2.

Matching accuracy. Overall accuracy was at chance level, *M* = 57.7%, 95% CI [49.7, 65.3] (see Figure 2, panel A). Performance was at chance level on both the A-V, *M* = 57.68%, 95% CI [47.93, 66.76] and V-A trials, *M* = 57.67 %, 95% CI [48.02, 66.66]. As in Experiment 1, the matching accuracy analysis was conducted using multilevel logistic regression. The dependent variable was accuracy (0 or 1). There were 2 factors, so 3 nested models were compared: the first model included a single intercept, the second model included the main effects (identity and order), and the third model added the two-way interactions. Table 1 reports the likelihood chi-square statistic (G^2) and *p* value associated with dropping each effect, as well as the coefficients (*b*) and standard errors (on a log odds scale) (*SE*) for each effect in the three-way interaction model. In the two-way model, the estimate of *SD* of the face random effect was 0.352 while for voice stimulus it was 0.303. The estimated *SD* for the participant effect was less than 0.313. A similar pattern was observed in the null model. Table 1 shows that there was a significant main effect of identity and a significant interaction between identity and order.

[TABLE 1 ABOUT HERE]

The cell means and 95% confidence intervals for matching accuracy in each condition are shown in Figure 5. The main effect of identity reveals that the hit rate, *M* = 65.0%, 95% CI [56.3, 72.9], was reliably higher than the true negative rate, *M* = 49.4%, 95% CI [40.5, 58.6]. The interaction between identity and order reflects less of a difference between the true positive rate (same identity trials) and the true negative rate (different identity trials) in the A-V condition (panel B) than in the V-A condition (panel A).

[FIGURE 5 ABOUT HERE]

Matching response. Overall, *same identity* responses were not made significantly above chance level, *M* = 58.5%, 95% CI [49.4, 67.0] (see Figure 2, panel B). Faces and voices were attributed to the same identity above chance level in the V-A trials, *M* = 61.9%, 95% CI [51.6, 71.1], but not in the A-V trials, *M* = 54.9%, 95% CI [44.6, 64.8].

Discussion

The results of the matching accuracy analysis show some evidence of degraded performance in comparison to previous results. Although overall matching accuracy was only just at chance level in Experiment 2, it is noteworthy that performance was significantly above chance when the face and voice were presented simultaneously (Experiment 1). In keeping with the interpretation that performance is compromised by longer ISIs (5 s), Smith et al. (2016a, Experiment 2) observed above chance level accuracy using an ISI of 1 s.

There was no overall bias to accept a face and voice as belonging to the same person when the stimuli were separated by 5 s. *Same identity* matching responses were not made above chance level. This finding supports the hypothesis that biases in face-voice matching are explained by temporal contiguity (Buehner & May, 2003; Ginns, 2006; Reed, 1992; Shanks et al., 1989). As displayed in Figure 2, when faces and voices were presented simultaneously (0 s ISI) in Experiment 1, participants made *same identity* responses above chance level.

Experiment 2 showed the same pattern of results as Smith et al. (2016a, Experiment 2, 1 s ISI), with a main effect of identity and 2-way interaction between order and identity. Figure 5 illustrates that whilst sensitivity did not differ across conditions, the true negative rate (specificity) was lower in the V-A condition. Both experiments therefore highlight the existence of a stronger bias to respond *same identity* when the face is presented before the voice. Experiment 2 shows that the bias endures over a 5 s ISI. 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.

Experiment 3

In Experiment 3 we investigated face-voice matching performance with a longer ISI. When there is a 10 s ISI, 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). Our interpretation of the results of Experiment 2 informed our hypothesis that overall accuracy would deteriorate to chance level, and that there would be no bias to accept a face and voice as belonging to the same person.

Method

Apart from the following exceptions, the methods were identical to Experiment 2.

Participants. There were 24 participants (22 females and 2 males), with an age range of 18 to 45 years (M = 23.6, SD = 8.0).

Procedure. The ISI was 10 s.

Results

These data were analysed using the same methods as Experiment 2. The overall accuracy (panel A) and the overall pattern of responses (panel B) for Experiment 3 (10 s) are illustrated in Figure 2 by the right-most data points in each panel.

Matching accuracy. Overall matching accuracy was at chance level, *M* = 52.5%, 95% CI [44.9, 59.9] (see Figure 2, panel A). Performance was at chance level on the A-V trials, *M* = 53.54%, 95% CI [44.19, 62.76] as well as the V-A trials, *M* = 51.57 %, 95% CI [42.27, 60.95]. The data were analysed using the same procedure as Experiment 2. The likelihood chi-square statistic (G^2) and *p* value associated with dropping each effect are reported in Table 2, as are the coefficients (*b*) and standard errors (on a log odds scale) (*SE*) for each effect in the two-way interaction model. In the two-way model the estimate of *SD* of the face random effect was 0.288 while for voice stimulus it was 0.391. The estimated *SD* for the participant effect was less than 0.001. The pattern was similar in the null model. As in Experiment 1, the variability associated with stimuli was greater than the variability at the participant level.

[TABLE 2 ABOUT HERE]

There was a main effect of identity. There was also a significant interaction between identity and order. The cell means and 95% confidence intervals for matching accuracy are shown in Figure 6.

[FIGURE 6 ABOUT HERE]

As displayed in Figure 6, the significant main effect of identity revealed that the hit rate, *M* = 60.3%, 95% CI [50.8, 69.2], was higher than the true negative rate, *M* = 44.4%, 95%CI [35.0, 54.2]. The interaction between identity and order shows that there is a much smaller difference between the true positive rate (same identity trials) and the true negative rate (different identity trials) in the A-V condition (panel B) than the V-A condition (panel A).

Matching response. Overall, faces and voices were not attributed to the same identity significantly above chance level, *M* = 57.6%, 95% C I[47.7, 66.8] (see Figure 2, panel B). Although *same identity* responses were made above chance level in V-A trials, *M* = 62.6%, 95% CI [51.6, 72.5], they were at chance level in A-V trials, *M* = 52.4%, 95% CI [41.4, 63.4].

Discussion

When the ISI was extended to 10 s, overall face-voice matching accuracy was at chance level. Taken together with the results from Experiment 1 and 2, this finding supports the hypothesis that accurate performance degrades as the ISI increases (see Figure 2, panel A).

As in Experiment 2, there was a significant main effect of identity, and a significant interaction between identity and order. As indicated by the matching response analysis, when there is a 10 s ISI, this interaction translates into a significant bias to respond that a face and voice belong to the same person in the V-A condition. In keeping with the predictions based on the results of Experiment 1 and 2, participants did not exhibit an overall bias to respond *same identity*.

General Discussion

In this paper we tested the effect of inserting longer ISIs on face-voice matching performance. No previous face-voice matching studies have included an ISI longer than 1 s, and few have investigated how bias operates. The findings show that face-voice matching is possible when faces and voices are presented simultaneously (Experiment 1), but performance is at chance level when an ISI of 5 s or more is introduced (Experiment 2 and 3). This supports the conclusion that the task involves guessing when traces for faces and voices have decayed. Our investigation of response bias revealed that the tendency to attribute common identity to faces and voices reduces as their temporal separation increases.

The pattern of variance observed in all 3 experiments shows that people differ in the extent to which they look and sound similar. Indeed, in Experiments 1 and 3, the variance associated with the face and voice stimuli was much greater than that associated with individual differences in matching performance. These results of the multilevel modelling analysis replicate those of Smith et al. (2016b), and support the explanation that characteristics of stimulus sets help to explain previous contradictions in the literature (Kamachi et al., 2003; Krauss et al., 2002; Lachs & Pisoni, 2004a; Mavica & Barenholtz, 2013; Smith et al., 2016a). Future face-voice matching studies using other stimulus sets should also employ multilevel modelling (Baguley, 2012; Judd et al., 2012).

In Experiments 1 and 3, the multilevel modelling analysis showed that the *SD* of the participant random effect was minimal (<0.001). In Experiment 2 it was larger (0.313), indicating that the participants were not responding uniformly to the stimuli in each trial. Characteristics such as the participants' age and gender did not appreciably differ across groups in Experiments 2 and 3, but it is feasible that the increased level of variance is attributable to individual differences in sensory memory. By 5 seconds, detailed representations may persist in some but not other people's echoic (Glanzer & Cuniz, 1966; Treisman, 1964; Wickelgren, 1969; Lu, Williamson & Kaufman, 1992) or iconic memory (Sligte et al., 2008; 2009).

Matching accuracy. Consistent with previous studies showing that static face-voice matching might be possible when faces and voices are presented within 1 s of each other (Krauss et al., 2002; Mavica & Barenholtz, 2013; Smith et al., 2016a, 2016b), above chance static face-voice matching was observed in Experiment 1. In both Experiments 2 and 3, performance was only above chance level in one condition: same identity V-A. However, as explained below, performance in this condition is likely to be driven by the existence of a bias to respond *same identity* in the V-A condition. Therefore, the overall results of Experiments 2 and 3 suggest that it is difficult to perform this task when the ISI is 5 s (Experiment 2) or 10 s (Experiments 3). It seems that access to common source identity information in static faces and voices is relatively transient. These results fit with the interpretation that above-chance matching accuracy depends on being able to compare high-quality perceptual representations of static faces and voices, which are temporarily stored in echoic and iconic memory. These representations are likely to have significantly decayed after 5 s (Coltheart, 1980; Glanzer & Cunitz, 1966; Lu et al., 1992; Neisser, 1967; Sligte et al., 2008, 2009; Sperling, 1960; Treisman, 1964; Wickelgren, 1969).

The overall 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 it is easier to accurately attribute common identity when faces and voices are presented within a short time frame. Being able to accurately link faces and voices that are significantly temporally offset would perhaps incur an unnecessary cost in terms of cognitive load.

Matching response. The bias to respond *same identity* is influenced by faces and voices being presented close together in time. Although an overall bias operates when a face and voice are presented simultaneously (Experiment 1), as well as when the ISI is 1 s (Smith et al., 2016a, Experiment 2), it does not manifest when the voice is presented 5 s (Experiment 1) or 10 s (Experiment 2) before the face in the A-V condition. This sits well with the predictions informed by temporal contiguity research, which point to associative inferences being more likely when stimuli are presented close together in time (Buehner & May, 2003; Ginns, 2006; Reed, 1992; Shanks et al., 1989).

Taken together with the results of Smith et al. (2016a), the results of Experiment 2 and 3 add to evidence of a stronger response bias in the V-A condition than in the A-V condition. In Experiment 2 (5 s interval) and 3 (10 s interval), there was less of a difference between accuracy on same identity and different identity trials when the voice was presented before the face (A-V condition). The matching response analyses also showed that whilst the overall bias to accept faces and voices in each trial as belonging to the same identity does not persist at a 5 s or 10 s intervals in the A-V condition, it does persist in the V-A condition. The order effect according to bias is perhaps attributable to the strength of identity information associated with faces and voices (Damjanovic & Hanley 2007; Hanley & Turner 2000; Stevenage et al., 2011, 2012; Stevenage, Neil, Barlow, Dyson, Eaton-Brown & Parsons, 2013; Stevenage & Neil, 2014). Faces provide more reliable cues to identity than voices, so voices could 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 is a reasonable strategy to rely on the face as a cue to identity, and preferentially accept a subsequent voice as belonging to the same person.

The pattern of results reported in these three experiments support the argument that the bias to attribute common identity to faces and voices provides a useful foundation for successful audio-visual speech integration. Therefore, beyond a short time frame, the overall lack of a bias to respond *same identity* is perhaps unsurprising. In speech perception, audio-visual 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). Furthermore, the order asymmetry in face-voice matching operates in a parallel pattern to biases in audio-visual speech integration. It has been shown that integration occurs from an auditory lead (comparable to the A-V condition) of up to around 100ms, and an auditory lag (comparable to the V-A condition) of around 300ms (Munhall et al., 1996; Robertson & Schweinberger, 2010; Van Wassenhove et al., 2007).

Conclusion. These 3 experiments demonstrate that face-voice matching performance is dependent on the time-course of stimuli presentation. The results help to clarify how cognitive processes driving matching decisions affect performance, emphasising how both accuracy and bias are reliant on comparing fast-decaying, high-quality perceptual representations. Finally the results offer potential clues as to the function of accurate face-voice matching. This ability may help people to navigate the complex social world during multi-speaker conversations and support speech integration to aid communication.

References

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

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

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

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

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

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

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

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., & 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

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

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

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

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., 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

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

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

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

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

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

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

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

Raaijmakers, J. G. W. (2003). A further look at the "language-as-fixed- effect fallacy." *Canadian Journal of Experimental Psychology*, *57*(3), 141-151. doi: 10.1037/h0087421

Raaijmakers, J. G. W., Schrijnemakers, J.M.C., & Gremmen, F. (1999). How to deal with the "language-as-fixed-effect fallacy": Common misconceptions and alternative solutions. *Journal of Memory & Language*, *41*(3), 416-426. doi: 10.1006/jmla.1999.2650

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

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

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

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

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

Smith, H. M. J., Dunn, A. K., Baguley, T., & Stacey, P.C. (2016b). Matching novel face and voice identity using static and dynamic facial images. *Attention, Perception, & Psychophysics*, 1-12. doi: 10.3758/s13414-015-1045-8

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

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., 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. (2014). 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

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

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

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

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

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

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

Formatted: English (U.S.)

~~Table 1~~

~~*Parameter estimates (b) and likelihood tests for the 2x2 factorial analysis, Experiment 2: 5 s inter-stimulus interval*~~

~~*Source*~~	~~*df*~~	~~*b*~~	~~*SE*~~	~~*G*^{2}~~	~~*p*~~
~~Intercept~~	~~1~~	~~0.272~~	~~0.265~~	~~-~~	~~-~~
~~Identity~~	~~1~~	~~1.245~~	~~0.362~~	~~7.51~~	~~.006~~
~~Order~~	~~1~~	~~0.474~~	~~0.322~~	~~0.02~~	~~.901~~
~~Identity x Order~~	~~1~~	~~1.136~~	~~0.495~~	~~5.61~~	~~.018~~

Figure Captions

Formatted: Font: Bold, Not Italic

~~**Table 2**~~

~~**Parameter estimates (b) and likelihood tests for the 2x2 factorial analysis, Experiment 3: 10 s inter-stimulus interval**~~

~~**Source**~~	~~**df**~~	~~**b**~~	~~**SE**~~	~~**G^2**~~	~~**p**~~
~~Intercept~~	~~1~~	~~0.457~~	~~0.254~~	~~-~~	~~-~~
~~Identity~~	~~1~~	~~1.092~~	~~0.329~~	~~7.53~~	~~.006~~
~~Order~~	~~1~~	~~0.491~~	~~0.324~~	~~0.28~~	~~.867~~
~~Identity x Order~~	~~1~~	~~0.951~~	~~0.460~~	~~4.22~~	~~.040~~

Formatted: Font: Bold

Formatted: Font: Bold, Not Italic

Formatted: Font: Bold

Formatted: Font: Bold, Not Italic

Figure 1. Illustration of the procedure used in Experiment 1

Figure 2: Overall matching accuracy and *same identity* responses for 0 s (i.e. simultaneous face-voice presentation), 5 s and 10 s inter-stimulus intervals in a same-different task. Error bars show 95%CI for the condition means

Figure 3: Matching accuracy for simultaneous face-voice presentation. Error bars show 95%CI for the condition means

Figure 4: Illustration of the procedure used in Experiment 2

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

Figure 6: Face-voice matching accuracy on V-A (panel A) and A-V (panel B) trials with a 10 s inter-stimulus interval. Error bars show 95% CI for the condition means

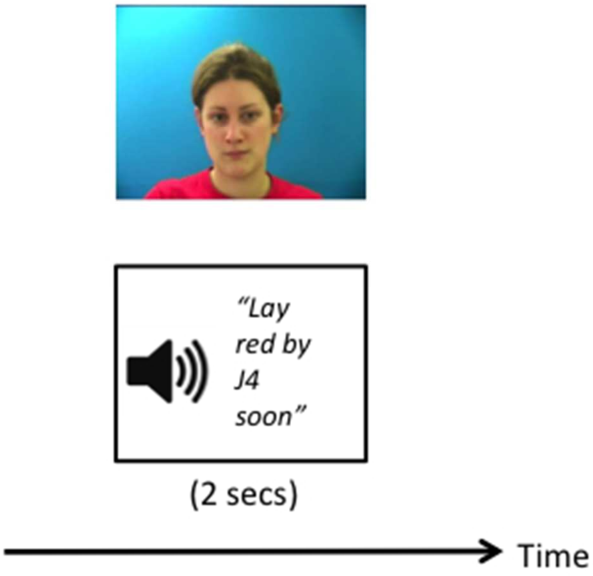

Figure 1: Illustration of the procedure used in Experiment 1
Figure 1
112x112mm (72 x 72 DPI)

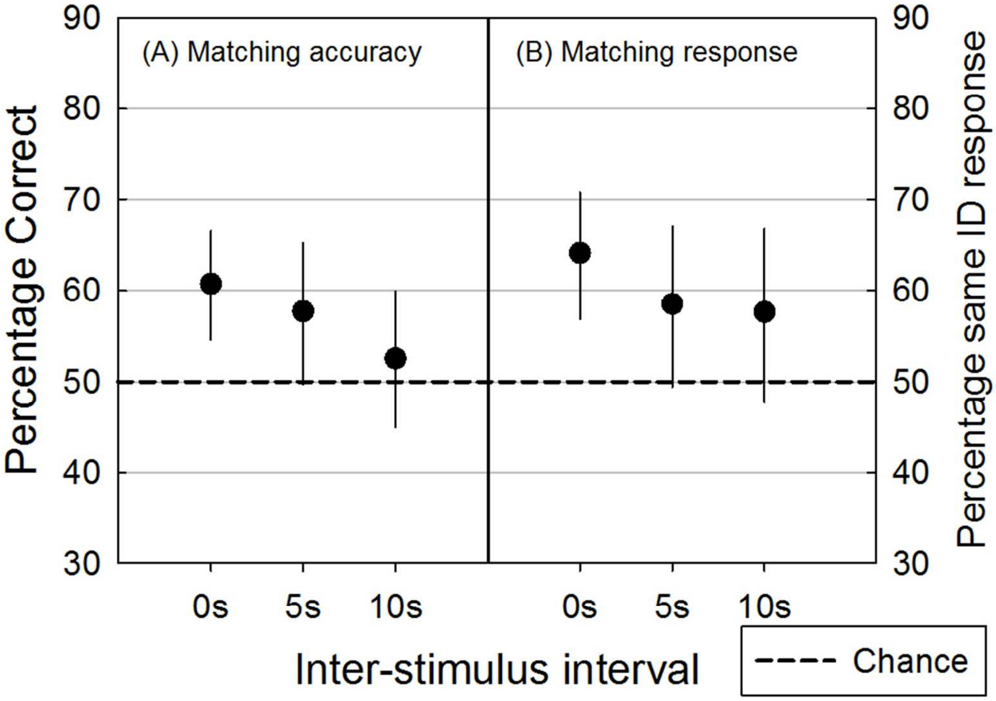

Figure 2: Overall matching accuracy and same identity responses for 0s (i.e. simultaneous face-voice presentation), 5s and 10s inter-stimulus intervals in a same-different task. Error bars show 95%CI for the condition means
Figure 2
164x117mm (150 x 150 DPI)

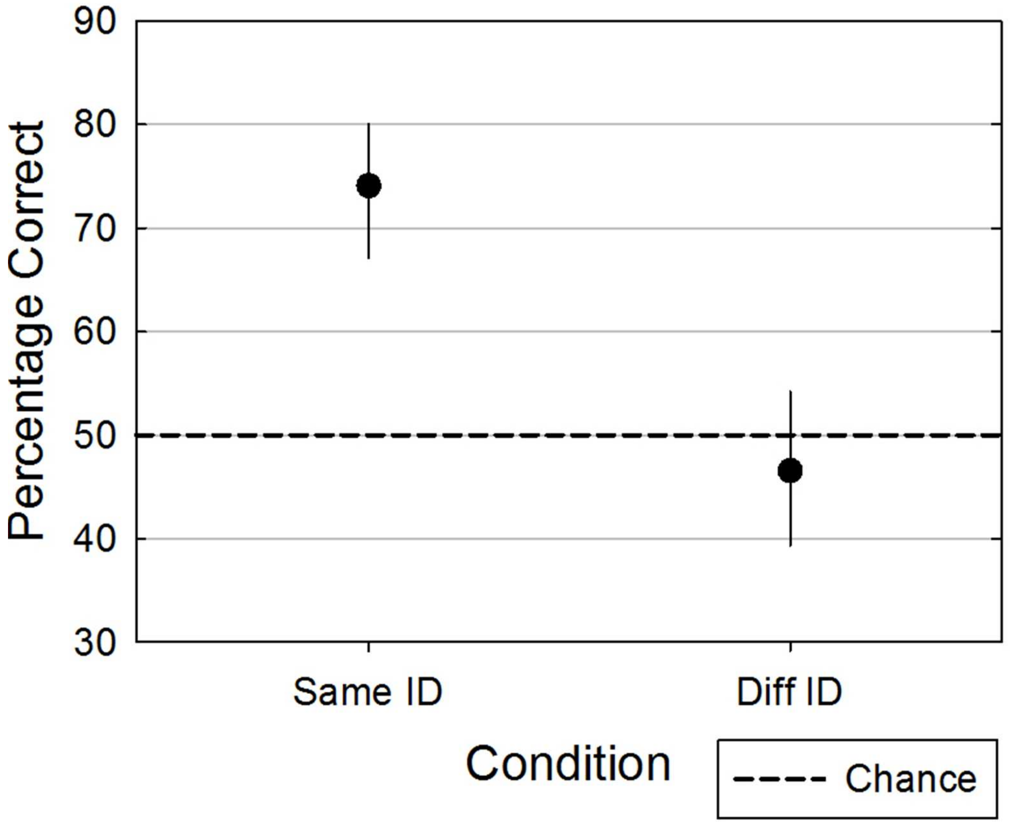

Figure 3: Matching accuracy for simultaneous face-voice presentation. Error bars show 95%CI for the condition means
Figure 3
144x119mm (150 x 150 DPI)

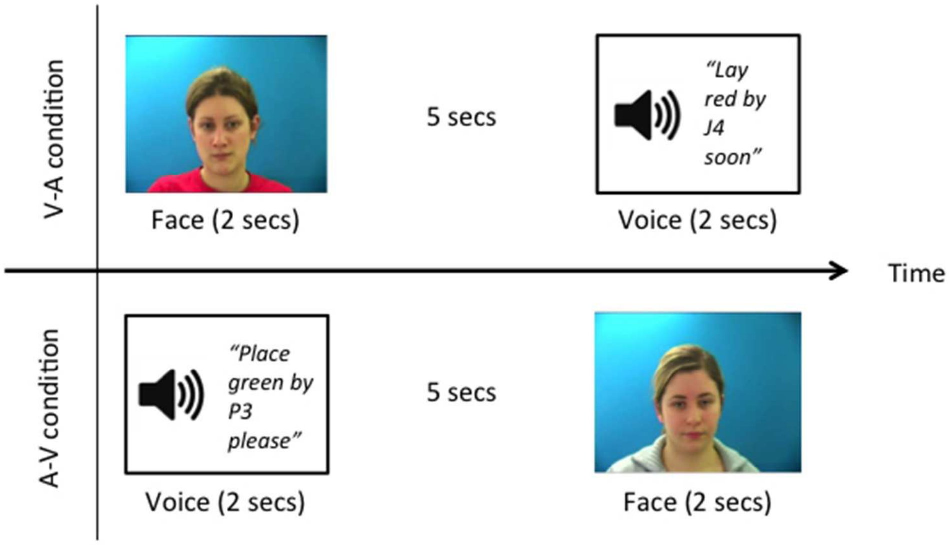

Figure 4: Illustration of the procedure used in Experiment 2
Figure 4
218x127mm (72 x 72 DPI)

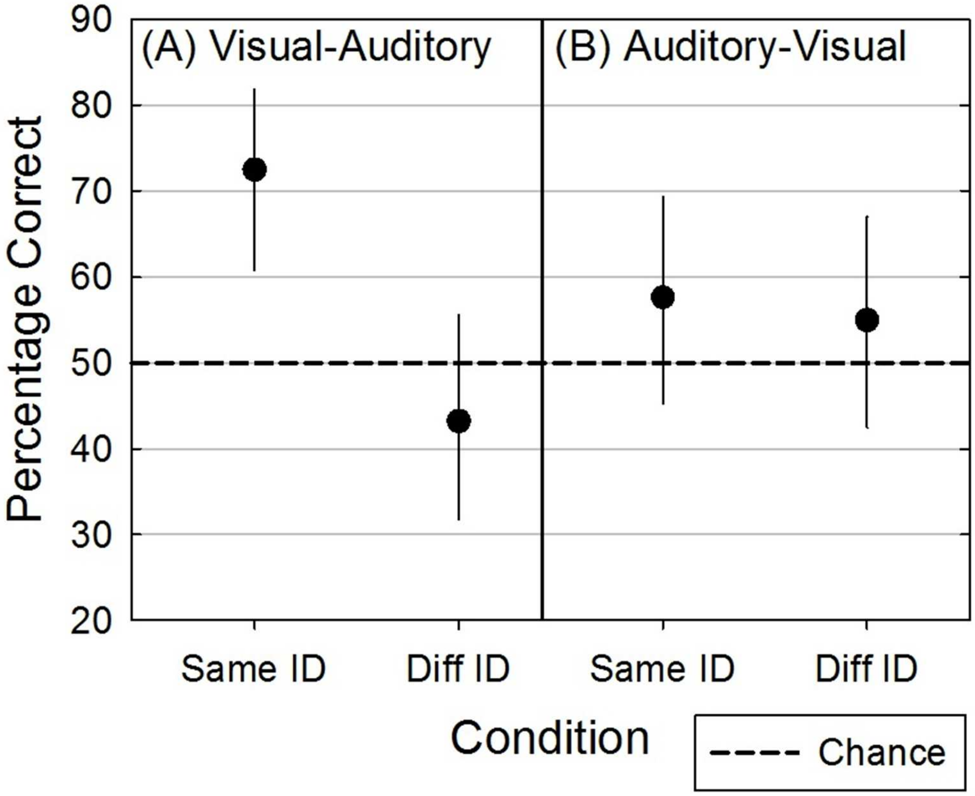

Figure 5: Face-voice matching accuracy on V-A (panel A) and A-V (panel B) trials with a 5s inter-stimulus interval. Error bars show 95% CI for the condition means
Figure 5
149x123mm (150 x 150 DPI)

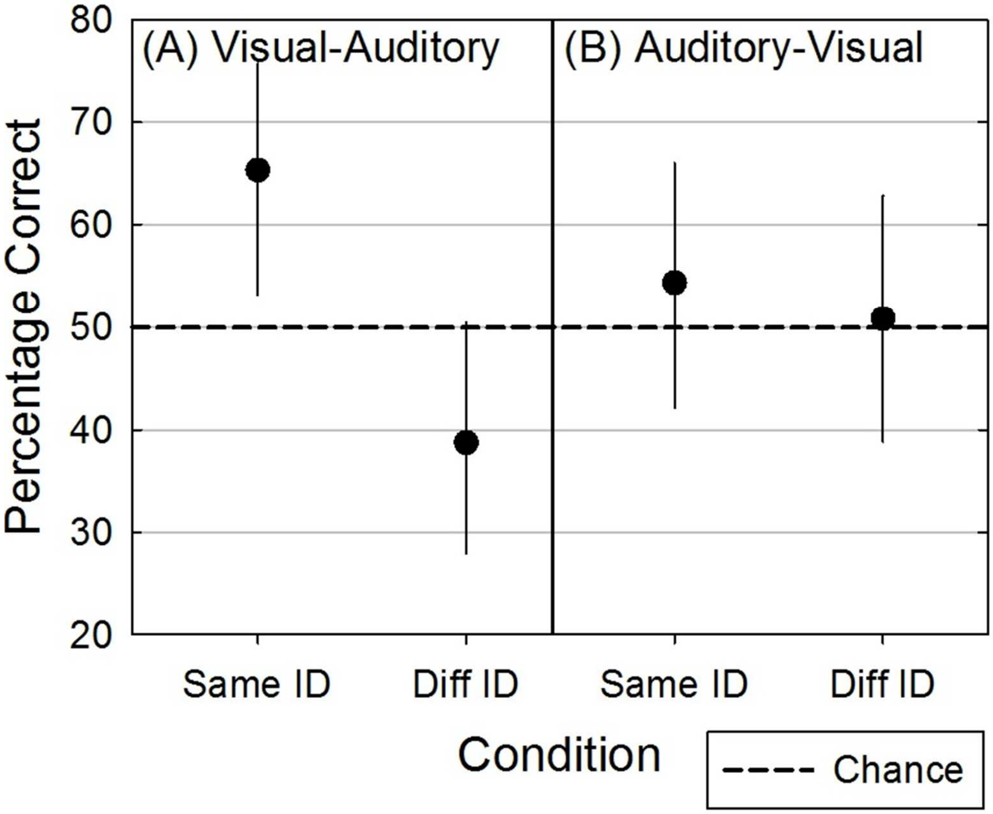

Figure 6: Face-voice matching accuracy on V-A (panel A) and A-V (panel B) trials with a 10s inter-stimulus interval. Error bars show 95% CI for the condition means
Figure 6
146x120mm (150 x 150 DPI)

Table 1

Parameter estimates (b) and likelihood tests for the 2x2 factorial analysis, Experiment 2: 5 s inter-stimulus interval

Source	*df*	*b*	*SE*	G^2	*p*
Intercept	1	0.272	0.265	.	.
Identity	1	1.245	0.362	7.51	.006
Order	1	0.474	0.322	0.02	.901
Identity x Order	1	1.136	0.495	5.61	.018

Table 2

Parameter estimates (b) and likelihood tests for the 2x2 factorial analysis, Experiment 3: 10 s inter-stimulus interval

Source	*df*	*b*	*SE*	G^2	*p*
Intercept	1	0.457	0.254	.	.
Identity	1	1.092	0.329	7.53	.006
Order	1	0.491	0.324	0.28	.867
Identity x Order	1	0.951	0.460	4.22	.040