British Journal of Psychology (2012), 103, 73–82

The British Psychological Society

www.wileyonlinelibrary.com

Integration of faces and voices, but not faces and names, in person recognition

Christiane O'Mahony and Fiona N. Newell*

School of Psychology and Institute of Neuroscience, Trinity College Dublin, Ireland

Recent studies on cross-modal recognition suggest that face and voice information are linked for the purpose of person identification. We tested whether congruent associations between familiarized faces and voices facilitated subsequent person recognition relative to incongruent associations. Furthermore, we investigated whether congruent face and name associations would similarly benefit person identification relative to incongruent face and name associations. Participants were familiarized with a set of talking video-images of actors, their names, and their voices. They were then tested on their recognition of either the face, voice, or name of each actor from bimodal stimuli which were either congruent or novel (incongruent) associations between the familiarized face and voice or face and name. We found that response times to familiarity decisions based on congruent face and voice stimuli were facilitated relative to incongruent associations. In contrast, we failed to find a benefit for congruent face and name pairs. Our findings suggest that faces and voices, but not faces and names, are integrated in memory for the purpose of person recognition. These findings have important implications for current models of face perception and support growing evidence for multisensory effects in face perception areas of the brain for the purpose of person recognition.

Person recognition is an essential but highly challenging cognitive task. Every person is unique in their appearance, yet people look very similar. Furthermore, humans are identifiable by both perceptual and semantic characteristics. A complex recognition system may have therefore evolved in the brain to make such close within-category discriminations (e.g., Kanwisher & Yovel, 2006; Yovel & Kanwisher, 2004) and to combine different types of information into an integrated memory of one person.

Both voices and faces are unique identity signatures (with the possible exception of many monozygotic twins [e.g., Brown, Carrel, & Willard, 1997]), they are physically inseparable from the individual, and they co-occur at close spatial and temporal contiguity. In contrast, semantic characteristics, such as names, are a less reliable signature to person identification, since they are not unique, are arbitrarily associated to the physical person and exist only once language develops although person recognition

*Correspondence should be addressed to Fiona Newell, Institute of Neuroscience, Lloyd Building, Trinity College, Dublin 2, Ireland (e-mail: fiona.newell@tcd.ie).

DOI:10.1111/j.2044-8295.2011.02044.x

occurs earlier (e.g., Johnson, 1994; Schweinberger, Herholz, & Stief, 1997). From an evolutionary perspective, it would therefore have been adaptive to develop a perceptual system that integrated face and voice information, but not arbitrary cues such as names, in order to maximize person recognition.

There is a growing body of literature in support of the hypothesis that visual and auditory information is integrated in the representation of a person (see, e.g., Campanella & Belin, 2007 for a review). Cross-modal identity matching can be seen as a consequence of the integration of faces and voices in memory and has been reported in several studies. For example, Lachs and Pisoni (2004) found that observers were able to easily match dynamic images of familiar faces with voice information. Others reported evidence for cross-modal matching even when different sentences were spoken across the two modalities (Kamachi, Hill, Lander, & Vatiliotis-Bateson, 2003; Munhall & Buchan, 2004) suggesting that temporal cues alone (i.e., movement of the lips and speech prosody) do not mediate cross-modal matching performance. Furthermore, Rosenblum, Smith, Nichols, Hale, and Lee (2006) found that voices can be efficiently matched to point-light displays of facial movement during speech.

Other studies have investigated cross-modal effects of voice and faces on learning and recognition memory. Sheffert and Olson (2004) found that voice learning was facilitated by facial information, suggesting mandatory integration of faces and voices in memory, even when information in one modality is irrelevant for the task. Moreoever, Schweinberger, Roberston, and Kaufmann (2007) recently reported that familiar voice recognition was facilitated by a simultaneous presentation of a time-synchronized articulating face only when the identity of the face corresponds to the voice. This effect was limited to familiar faces only, suggesting that the effects depend on the availabilty of a multisensory representation of the person in memory.

Several neuroimaging studies provide direct support for the hypothesis that faces and voices are integrated in person identification. In a functional Magnetic resonance Imaging (fMRI) study, Von Kriegstein, Kleinschmidt, Sterzer, & Giraud (2005) found that a voice recognition task activated the fusiform face area (FFA) previously shown to be involved in face recognition (Kanwisher, McDermott, & Chun, 1997). Von Kriegstein *et al*. later argued that this activation was achieved by low-level sensory binding of face and voice information at a processing stage that was prior to the person recognition stage (Von Kriegstein, Kleinschmidt, & Giraud, 2006). Von Kriegstein and Giraud (2006) also reported that cortical activation in the FFA during voice recognition was facilitated by previously learned face–voice pairings and but not by learned face–name pairings. Van Wassenhove, Gant, and Poeppel (2005), on the other hand, found evidence that visual information can affect auditory processing. They reported that visual speech speeded up the cortical processing of subsequent auditory speech, suggesting that the effects of visual processing and auditory processing for person recognition are bi-directional.

The Burton, Bruce, and Johnston (1990) interactive activation model (IAC) of person recognition provides a theoretical framework of the processes involved in person recognition. Specifically, the IAC model (which is a further elaboration of the Bruce and Young (1986) functional model of face perception) suggests separate units in memory for specific perceptual and semantic attributes related to a person, such as the face, voice, name, and other semantic information. For example, this model postulates separate units for the processing and storage of face information (i.e., Face Recognition Units or FRUs) and voice information (i.e., Voice Recognition Units or VRUs). Thus, each familiar face or voice is associated with a particular FRU or VRU. Other information about the person, such as biographical information, is processed and stored in what are called Semantic

Information Units (SIUs). Name information, however, is stored separately from other semantic information in what are called Name Recognition Units (NRUs). All of these units (i.e., FRUs, VRUs, NRUs, and SIUs) are connected only via the central point in the person recognition system; the Personal Identity Node (PIN). The PIN can therefore receive multisensory inputs from all parts of the system in order to make familiarity decisions.

The model makes several predictions about how persons are recognized. For example, since faces, voices, and names have no connecting links except via the PIN then face, voice and name recognition should occur within domain-specific systems; FRUs, VRUs, and NRUs (Burton *et al*., 1990). Within-modality priming studies have supported this prediction: Brunas, Young, and Ellis (1990) and Bruce and Valentine (1985) found long-lasting within-domain, but not cross-domain, priming effects, reflecting the robust nature of the structural changes within the recognition system. Bruce and Valentine (1986) later provided evidence that the processes involved in within-modal priming between semantically related familiar faces or familiar names are similar and therefore likely mediated by a common semantic (i.e., PIN) but not verbal (i.e., via the NRU when naming the individuals) context. Furthermore, according to the IAC model, cross-domain priming effects can only occur via the PIN, and are smaller and more temporary than within-domain priming (Burton *et al*., 1990). For example, the time to decide whether the person is familiar or not from their name can be facilitated by a face prime containing the image the same person or of a closely associated familiar person (see Young, Flude, Hellawell, & Ellis, 1994). These effects, according to Young *et al*., are likely mediated via the PIN and supported by semantic knowledge of the familiar person (i.e., whether they are a politician or comedian, etc.) and are compatible with the Burton *et al*. model (see, e.g., McNeill & Burton, 2002).

Interestingly, the IAC model makes no distinction between cross-domain priming between two perceptual input units (such as FRUs and VRUs) and cross-domain priming between a perceptual input unit and a semantic input unit (such as FRUs and NRUs). However, some evidence suggests that this assumption may be incorrect. For example, Schweinberger *et al*. (1997) found face but not name priming of famous voices with a 10-min interval between the face or name primes and voice targets. The IAC model could not account for the difference in priming across these conditions, since both conditions presumably activate the PIN, nor could it account for the effect of priming across long delays since the activation of the PIN is presumed to be short-lived. An alternative explanation, provided by Schweinberger *et al*. (1997) is that voice and face representations may be connected at a pre-semantic, that is, pre-PIN level which is the locus of the reported face–voice priming. This supports the hypothesis that faces and voices, but not names, are integrated in person recognition and, in turn, predicts that voice information but not name information would prime the recognition of a target face image.

In this study, we used an explicit memory paradigm to investigate whether faces and voices were integrated with each other, in comparison to faces and names, for the purpose of person recognition. The following study was based on Garner's constrained classification task (Garner, 1974): the assumption of this task is that if voice and face or name and face information is integrated in memory then recognition will occur more quickly and more accurately in conditions in which the identity of the person is congruent rather than incongruent across modalities. We predicted that if voices and faces are integrated in memory then the recognition of identity congruent face and voice pairs would be more efficient than the recognition of identity incongruent face

and voice pairs. Moreover, we predicted that name information would not be integrated with faces; therefore, no difference was expected between the recognition of identity congruent or incongruent face–name pairs.

Method

Participants

Twenty-four undergraduate students (23 female, 1 male) from the School of Psychology, Trinity College Dublin, participated in this study in exchange for course credit. Their ages ranged from 18 to 40 years. The study was approved by the School of Psychology Research Ethics Committee and, accordingly, all participants provided written consent prior to taking part in the study.

Stimuli and apparatus

Stimuli consisted of audio-visual (AV) recordings of 48 unfamiliar actors, 24 male, and 24 female, aged between 18 and 25. These actors were filmed using a digital camera whilst they counted from number '1' to '5' at a rate of one number per second. A constant rate of counting was maintained across the sample by instructing each person to count in time with the beat of a (visual) metronome that was positioned at eye level in front of each participant. Hair and clothes were covered and glasses and jewellery were removed to avoid identification by these characteristics (see Figure 1). Each actor was filmed against the same plain background.

Each AV clip was edited using Adobe Premiere 6.5 (Adobe Systems Inc. USA) to create face–voice combinations as stimuli. For each of these face–voice stimuli used throughout the experiment, we dubbed the voice of one actor onto the lip movements of another actor. As the auditory counting rate was equivalent for all actors, the synchrony between the auditory counting and the corresponding lip movement was ensured across all conditions and all stimuli. As a precaution, however, we asked a group of five independent judges to view all AV stimuli and to rate each for perceived synchrony between the actor's lip movements and the vocal information. They judged the congruent and incongruent stimuli as being fully synchronized with no differences

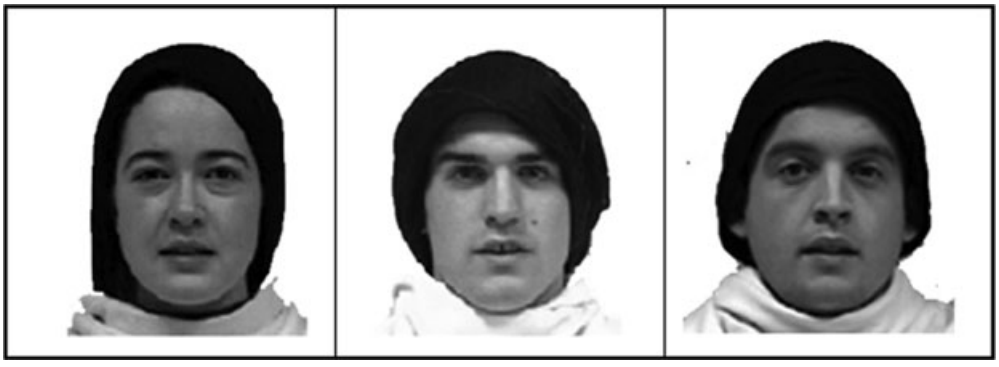

Figure 1. Example of the face stimuli used in the experiment. Stimuli were video clips of actors counting from 'one' to 'ten'. Secondary information such as clothing, hairstyle, and jewellery that could be used for identification purposes was masked. Female actors were required to remove make-up prior to filming.

between stimulus type. In addition, an auditory name clip was created for each actor. The name clip consisted of the voice of the experimenter (i.e., a voice not belonging to any of the test stimuli) which was recorded stating each actor's (pseudo) name, for example, 'This is David'). Apart from the name, no other semantic information was provided about these actors to the participants (see, e.g., Young *et al*., 1994 and McNeill & Burton, 2002 on semantic effects in person recognition).

During the training sessions and the main experiment, each stimulus was presented for 5 s. During training, a combined face–voice stimulus was preceded by a name presented aurally. Thus, each training trial consisted of an auditory presentation of the actor's name (e.g., 'This is David'), followed by the AV clip of the combined face and voice. Following training, each unimodal component, that is, face (silent), voice, or name, of the stimulus was presented in a trial. In the main experiment, face, voice, and name components were combined to create stimuli for each of the identity conditions; congruent; incongruent; and unfamiliar. Thus, identity congruent stimuli consisted of the familiarized combination of a face and voice or face and name. An identity 'incongruent' stimulus consisted of familiarized components but in a novel combination, for example, a familiarized face paired with a previously familiarized but non-matching voice or name. During the main experiment, for each of the face–name stimulus combinations, a silent visual face stimulus was simultaneously presented with a voice over indicating name only (presented aurally via headphones) without the actor's voice information.

The experiment was programmed using the DmDX software package. A Hewlett Packard, Pavilion ZE 4111s laptop computer was used to run the experiment and record data. The experiment was conducted in a testing lab in the School of Psychology.

Design

The experimental protocol consisted of four different blocks, each comprising three sessions; a training session, a familiarity test session, and the main experimental session. For the training session, AV clips of 32 different actors were randomly selected from the 48 actors as training stimuli. All participants were trained on the same set of 32 stimuli. In the experiment, these stimuli were randomly assigned to each of the congruent and incongruent identity conditions and the remaining untrained 16 AV clips were assigned to the 'unfamiliar' identity condition. Within each of the identity conditions, four actors were assigned to each of the cross-modal stimulus combinations which was then counterbalanced across participants.

The four experimental blocks were separated on the basis of the combined stimulus information and on each task. As such, each block comprised of either face–voice, voice–face, face–name, or name–face combinations and the task was to decide if the target stimulus in each combination, that is, the latter stimulus in each combination, was familiar or not. Furthermore, we manipulated the identity of each of these combinations to be either congruent, incongruent, or unfamiliar. Consequently, the main experiment was based on a 4×3 repeated measures design with cross-modal stimulus combination (face–voice; voice–face; face–name, or name–face) and identity (congruent, incongruent, or unfamiliar) as factors. Each experimental block contained 12 trials with equal numbers of congruent, incongruent, and unfamiliar trials. The presentation order of the four blocks was counterbalanced across participants and the order of the trials was randomly presented within blocks.

Procedure

Throughout the study, participants were seated in front of the laptop and wore headphones through which voice and name information was received. In the training session of each block, participants were presented with eight face–voice stimuli which they were required to learn. In this training session only, the name of the actor was provided immediately before each face–voice stimulus was presented in order to avoid any overlapping of actor's voice information with the spoken name of the actor. Each of the eight names with face–voice stimulus combinations was presented three times in a random order during training. Following training, participants were tested on their recognition of the trained stimuli and were required to name each of the eight faces and eight voices. A criterion performance was set at a minimum accuracy rate of 7/8 and if participants did not reach the minimum accuracy rate they were trained and tested on all stimuli from that block again. They were tested on face naming and voice naming separately to avoid any cross-domain facilitation effects and to ensure equal levels of stimulus familiarity across domains. The training procedure was the same for all experimental blocks.

As soon as criterion performance was reached, the main experimental session followed. In this session, participants were required to decide if either the face, voice, or name was familiar, depending on the experimental block. Each block consisted of identity congruent, incongruent, or unfamiliar trials. For the purpose of the experiment, an identity 'congruent' stimulus consisted of a familiarized, paired association between one actor's face and another actor's voice or name. In contrast, an identity 'incongruent' stimulus consisted of a familiarized actor's face paired with a different but previously familiarized actor's voice (i.e., the voice of another actor) or name. Finally, unfamiliar stimuli were those not previously presented in the training session.

During the main experimental session, each trial lasted 5 s with an inter-trial interval of 1 s. Prior to each testing block, participants were instructed to attend to all information present in the stimulus, but to make a familiarity decision about the face, voice, or name as appropriate, as quickly and accurately as possible by pressing a designated key on a keyboard corresponding to a 'familiar' or 'unfamiliar' decision. The position of the 'familiar' and 'unfamiliar' response keys was counterbalanced across participants. Response times and errors were recorded. Although a response key press did not terminate the movie clip, only the first response was recorded. During debriefing, participants were asked whether any of the faces used in the experiment were previously known to them. All participants confirmed that none of the faces were familiar prior to testing.

Results

We first performed separate analyses on the overall mean error rates and response times in the main experiment across each of the identity and cross-modal stimulus combination conditions using one-way, repeated measures analysis of variance (ANOVA). For the cross-modal stimulus combinations of voice–face, name–face, face–voice, and face–name, the mean percentage errors made were 2.29%, 2.43%, 4.17%, and 1.50% and the mean response times (in ms) were 1,252, 1,821, 2,787, and 2,677, respectively. The effect of stimulus combination on error rates failed to reach significance [$F(3,69) = 2.57$, $p < .061$]. There was a main effect of stimulus combination on response times [$F(3,69) = 39.62, p < .001$] with significantly faster response times to the voice–face combination than the other three (Newman–Keuls *post hoc* test, $p < .01$); faster response times to the

name–face combination than either of the face–voice and face–name conditions ($p < .01$) and no difference between the face–voice and face–name conditions. The relatively longer response times to the face–voice and face–name combinations possibly reflects the time taken to listen to the name or to a sufficient sample of the voice.

The mean percentage errors made to the congruent, incongruent, and unfamiliar identity conditions were 1.56%, 4.95%, and 1.22%, respectively. There was a main effect of identity condition on the error rates [$F(2,46) = 14.82$, $p < .01$] with significantly more errors made to the incongruent condition than either of the other two conditions (*post hoc* Newman–Keuls test, for both comparisons $p < .01$) and no difference between the congruent and unfamiliar conditions. The mean response times were 2,035, 2,223, and 2,146 ms for the congruent, incongruent, and unfamiliar conditions, respectively. Although there was a trend indicating faster response times for the congruent trials over the incongruent and unfamiliar trials, the overall effect of condition on response times failed to reach significance [$F(2,46) = 2.99$, $p = .06$]. As the responses to the unfamiliar trials were not of interest to this study, data from these trials were not used in our further analyses.

We then, for each participant, collapsed the data across the cross-modal stimulus combination conditions when the same stimuli were presented. For example, for each participant, we took the average response across stimuli involving the same face and voice pairing where a response was required either to the voice in one trial or the face in the other trial (i.e., we averaged the data across voice–face and face–voice stimulus pairs within each of the identity conditions), and to the stimuli involving the same face and name pairing where a response was required either to the face in one trial or the name in the other trial (i.e., name–face and face–name pairs). We then conducted a 2×2, within-subject ANOVA on the mean reaction times with identity condition (congruent or incongruent) and cross-modal stimulus combination (face/voice or face/names) as the main factors.

For error rates, we found a main effect of identity condition [$F(1,23) = 18.77$, $p < .001$], with more errors to the incongruent than congruent condition. There was no effect of stimulus combination [$F(1,23) = 2.54$, n.s.] and an interaction between the factors failed to reach significance [$F(1,23) = 3.01$, $p = .096$]. A two-way, within-subjects ANOVA on the response times revealed a significant effect of identity condition [$F(1,23) = 7.12$, $p < .05$] with faster response times to the congruent than incongruent identity condition. The effect of cross-modal stimulus combination was not significant [$F(1,23) = 2.55$, $p = .12$]. Most pertinently, however, we found an interaction between the identity and stimulus combination factors [$F(1,23) = 20.42$, $p < .001$] (see Figure 2). A *post hoc*, Newman–Keuls analysis revealed that response times to the congruent face–voice pairs were significantly faster ($p < .05$) than to the incongruent face–voice pairs and that there was no difference in response times between the congruent and incongruent face–name pairs.[1]

Discussion

Our results suggest that familiarity decisions about face–voice associations congruent for identity were facilitated relative to decisions about incongruent face–voice pairings.

[1] *A separate 2 × 2 ANOVA on response times to the non-collapsed face–name stimulus combinations, depending on whether the response was to the face or the name, further confirmed this finding in that there was no evidence of a main effect of congruency [F(1,23) < 1, nor an interaction between these factors [F(1,23) = 1.7, p = .2].*

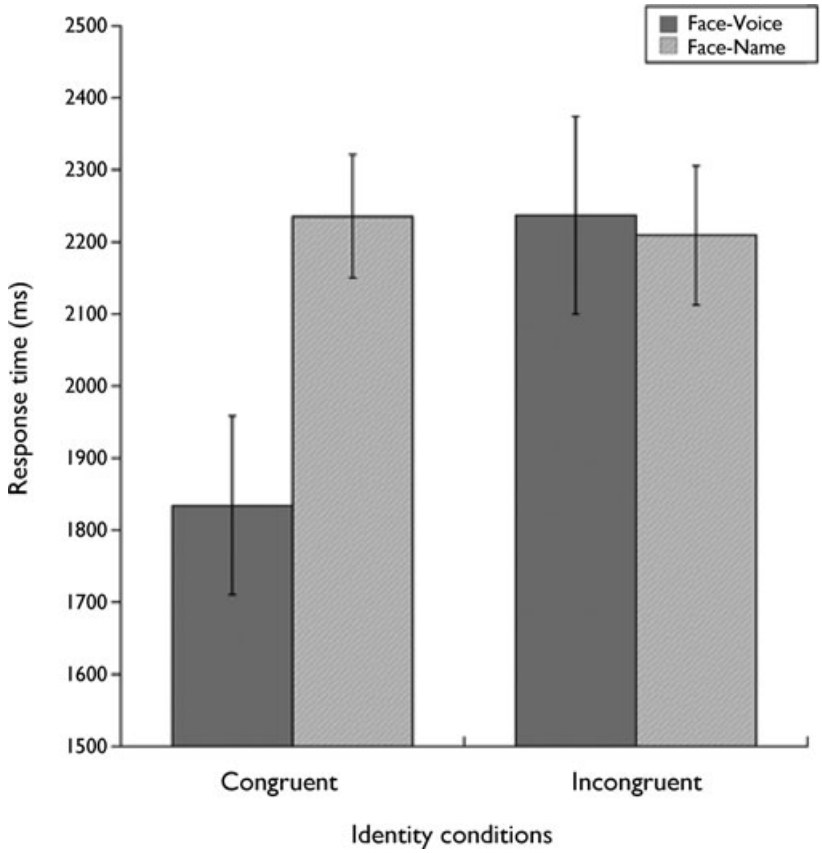

Figure 2. Plot showing the mean response times across each of the identity congruent and identity incongruent pairs for both the face–voice pairs and the face–name pairs. Error bars represent ± 1 standard error of the mean.

In contrast, decisions to face–name pairings congruent for identity were no faster than to incongruent face–name pairings. We reasoned that if two simultaneously presented dimensions are integrated in memory, then response times would be faster when they are congruent and than when they are incongruent (see Garner, 1974). The results of the present study therefore suggest that faces and voices are integrated in person recognition, but that faces and names are not.

Most functional models of face perception, such as the IAC model, fail to make a distinction between how faces, voices, and names relate to each other in memory (e.g., Burton *et al.*, 1990; Bruce & Young, 1986; Calder & Young, 2005). For example, the IAC model postulates two separate and independent systems of person recognition, the perceptual system, containing FRUs and VRUs and the semantic system, containing NRUs and SIUs. These perceptual and semantic domains are connected only via the PIN. As such, no particular difference in the relationship between faces and voices and between faces and names in person identification would be predicted. For example, it is not clear how faces would facilitate the recognition of a congruent voice (and vice versa) but not facilitate the recognition of a congruent name, since all connections between faces, voices, and names are via the PIN. Indeed, it would be assumed that there should be

no direct facilitation between faces and voices or names when there is an intermediate processing level involved such as the PIN. The results of the present study support the idea that faces and names may indeed only be connected via the PIN, given that the congruency of face and name information had no effect on response performance. In contrast, our results suggest that faces and voices, as perceptual attributes are directly linked and may be connected prior to the PIN, given that congruency of face and voice information had a significant benefit on performance. Our findings, together with those previously reported by Schweinberger *et al*. (2007), suggest an important revision to the IAC model, namely that the perceptual attributes of a person should be integrated at a stage prior to person recognition. In other words, we propose that there is a direct, pre-PIN link between faces and voices.

The results of the present study demonstrate that perceptual attributes, such as faces and voices, are more closely integrated than perceptual and semantic attributes, such as faces and names. Indeed, as mentioned previously, the integration of perceptual attributes, as opposed to semantic attributes, may have developed at different stages in evolutionary terms; a perceptual person recognition system, designed to process and integrate physical attributes such as faces and voices, may have developed prior to language for the purpose of optimizing person recognition. The semantic person recognition system would have developed, to a large extent, once language was acquired, because most semantic information about a person, such as their name, is communicated verbally. It is also possible that perceptual and semantic person recognition are separate processes that can occur sequentially during typical social encounters. For example, we can become acquainted with somebody based on their perceptual characteristics (i.e., face recognition or voice recognition) without yet acquiring any semantic knowledge about them. It therefore seems feasible that perceptual and semantic personal attributes are separately represented in memory. Interestingly, with the advent of internet chat rooms, the opposite order of personal information acquisition is increasingly more common: through chat rooms, we can get to know people on a semantic level without knowledge of their perceptual characteristics. It remains to be seen whether this type of social learning has any effect on the processes involved in subsequent person recognition.

It is the first study to our knowledge that systematically compared face, name, and voice recognition with multisensory conditions at the learning and test phases. The results provide evidence that faces and voices are more closely integrated in person recognition than faces and names. These findings, together with previous reports, have implications for models of person recognition in that the processes involved in the representation of individuals in memory are more multisensory than previously thought.

References

Brown, C. J., Carrel, L., & Willard, H. F. (1997). Expression of genes from the human active and inactive X chromosomes. *American Journal of Human Genetics*, *60*, 1333–1343.

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

Bruce, V., & Valentine, T. (1985). Identity priming in recognition of familiar faces. *British Journal of Psychology*, *76*, 373–383.

Bruce, V., & Valentine, T. (1986). Semantic priming of familiar faces. *Quarterly Journal of Experimental Psychology*, *38A*, 125–150.

Brunas, J., Young, A. W., & Ellis, A. W. (1990). Repetition priming from incomplete faces: Evidence for part to whole completion. *British Journal of Psychology*, *81*, 43–56.

Burton, A. M., Bruce, V., & Johnston R. A. (1990). Understanding face recognition with and interactive activation model. *British Journal of Psychology*, *81*, 361–380.

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

Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences*, *11*(12), 535–543.

Garner, W. R. (1974). The processing of information and structure. Maryland, US: Lawrence Erlbaum Associates, Potomac.

Johnson, M. H. (1994). Brain and cognitive development in infancy. *Current Opinion in Neurobiology*, *4*(2), 218–225.

Kanwisher, N., McDermott, J., & Chun, M. (1997). The fusiform face area: A module in the human extrastriate cortex specialized for the perception of faces. *Journal of Neuroscience*, *17*, 4302–4311.

Kanwisher, N., & Yovel, G. (2006). The fusiform face area: A cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society of London, B Biological Sciences*, *361*(1476), 2109–2128.

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*, 159–187.

McNeill, A., & Burton, A. M. (2002). The locus of semantic priming effects in person recognition. *Quarterly Journal of Experimental Psychology*, *55A*(4), 1141–1156.

Munhall, K. G., & Buchan, J. N. (2004). Something in the way she moves. *Trends in Cognitive Scienves*, *8*(2), 51–53.

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

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., 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 & Psychophysics*, *66*, 352–362.

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*, 367–376.

Von Kriegstein, K., Kleinschmidt, A., & Giraud, A. L. (2006). Voice recognition and crossmodal responses to familiar speakers' voices in prosopagnosia. *Cerebral Cortex*, *16*, 1314–1322.

Von Kriegstein, K., & Giraud, A. L. (2006). Implicit multisensory associations influence unimodal voice recognition. *PLoS Biology*, *4*(10), e326, 1809–1820.

Van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Science*, *102*(4), 1181–1186.

Young, A. W., Flude, B. M., Hellawell, D. J., & Ellis, A. W. (1994). The nature of semantic priming effects in the recognition of familiar people. *British Journal of Psychology*, *85*, 393–411.

Yovel, G., & Kanwisher, N. (2004). Face perception: Domain specific, not process specific. *Neuron*, *44*(5), 889–898.

Received 23 May 2009; revised version received 18 March 2011