Cognition 137 (2015) 9–21

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Distinctive voices enhance the visual recognition of unfamiliar faces

I. Bülthoff[a,*], F.N. Newell[b]

[a] Max Planck Institute for Biological Cybernetics, Spemannstr. 38, D-72076 Tübingen, Germany
[b] School of Psychology and Institute of Neuroscience, Lloyd Building, Trinity College Dublin, Dublin 2, Ireland

ARTICLE INFO

Article history:
Received 17 June 2013
Revised 16 December 2014
Accepted 18 December 2014
Available online 12 January 2015

Keywords:
Face recognition
Multisensory
Distinctiveness
Voices
Audio–visual

ABSTRACT

Several studies have provided evidence in favour of a norm-based representation of faces in memory. However, such models have hitherto failed to take account of how other person-relevant information affects face recognition performance. Here we investigated whether distinctive or typical auditory stimuli affect the subsequent recognition of previously unfamiliar faces and whether the type of auditory stimulus matters. In this study participants learned to associate either unfamiliar distinctive and typical voices or unfamiliar distinctive and typical sounds to unfamiliar faces. The results indicated that recognition performance was better to faces previously paired with distinctive than with typical voices but we failed to find any benefit on face recognition when the faces were previously associated with distinctive sounds. These findings possibly point to an expertise effect, as faces are usually associated to voices. More importantly, it suggests that the memory for visual faces can be modified by the perceptual quality of related vocal information and more specifically that facial distinctiveness can be of a multi-sensory nature. These results have important implications for our understanding of the structure of memory for person identification.

1. Introduction

In the domain of face recognition research, numerous studies have identified circumstances under which recognition or identification of faces is enhanced. This previous research, mainly conducted using static images of faces, has led to important insights into the structure of face memory. We know, for example, that faces are better recognised when presented in an upright orientation than when inverted (Yin, 1969) and that distinctive faces are more recognisable than typical faces (e.g., Shepherd, Gibling, & Ellis, 1991; Valentine, 1991; Bruce, Burton, & Dench, 1994).

A useful framework that can account for the distinctiveness effect in faces has been proposed by Valentine (Valentine, 1991) and is known as the ‘face space’ model of memory for faces. In his framework, each face is encoded as a point in a multidimensional space defined by dimensions relevant for discriminating between faces. Typical faces are located in the densely populated centre of the face space while distinctive faces are located in the less densely populated area on the outskirts of the face space (see Fig. 1a). As such, the more distant a face is from the centre, the more distinctive it is and the more distant two faces are from each other in face space, the more they differ from each other and the easier it is to distinguish between them. This framework can account for various effects reported in the literature including that unfamiliar typical faces are more often misclassified as familiar than unfamiliar distinctive faces, and that unfamiliar distinctive

[*] Corresponding author. Tel.: +49 7071 601 611.
E-mail addresses: isabelle.buelthoff@tuebingen.mpg.de (I. Bülthoff), fiona.newell@tcd.ie (F.N. Newell).

http://dx.doi.org/10.1016/j.cognition.2014.12.006
0010-0277/© 2015 The Authors. Published by Elsevier B.V.

faces are better memorised than unfamiliar typical faces (see Bruce et al., 1994; Valentine & Bruce, 1986a, 1986b). Furthermore, the face space model can account for a variety of other behavioural effects such as face adaptation (Leopold, O'Toole, Vetter & Blanz, 2001), face inversion and viewpoint (Newell, Chiroro, & Valentine, 1999), attractiveness (Deffenbacher, Vetter, Johanson, & O'Toole, 1998; Potter & Corneille, 2008), caricaturing (Lee, Byatt, & Rhodes, 2000) and the recognition of other-race faces (Byatt & Rhodes, 2004). This model has also received some support from neurophysiology. For example, Leopold, Bondar, and Giese (2006), reported evidence of a direct relationship between face-induced activity in face-selective neurons in macaques and the level of caricature of the face stimuli presented, suggesting a norm-based representation of faces at the level of the single neuron.

Apart from the role of visual information in person recognition, others have investigated how information from other modalities can affect person recognition. For example, some recent studies have shown that better identification of voices is achieved when voices were presented with faces during learning than when voices were learned alone (Sheffert & Olson, 2004). The effect of visual information from faces on auditory perception is also well documented in the speech perception literature (McGurk & MacDonald, 1976; Calvert, Brammer, & Iversen, 1998). Whilst many research projects have investigated the role of visual information on speech perception only a few studies so far have addressed questions that are related to the interaction of voices and faces for the purpose of person recognition (see reviews by Belin, Shirley, & Bedard, 2004; Campanella & Belin, 2007; Yovel & Belin, 2013). Although voices take longer to identify than faces (Hanley, Smith, & Hadfield, 1998; Hanley & Turner, 2000) one might assume that the voice of a speaker might help recognise their face, for example, when the speaker is sufficiently distant to make face identification difficult or when face recognition is compromised, as in prosopagnosia. However, recent studies have suggested that voice information is not simply an alternative 'cue' to person recognition but is information that may be integrated in memory with facial information for person recognition. For example, both O'Mahony and Newell (2012) and Schweinberger, Robertson, and Kaufmann (2007) found that congruent pairings of familiar voices and faces resulted in better person recognition than incongruent pairings of the voices and faces. Moreover, O'Mahony and Newell (2012) found that voice identity was more likely to influence face recognition than other semantic information such as the person's name. Interestingly these effects were reduced when faces were static images in contrast to when dynamic, articulating faces were presented (Schweinberger et al., 2007). The authors of both of these studies argued that their findings suggest that faces and voices are integrated into a multisensory representation for the purpose of person identification.

Evidence from neuroimaging studies also supports the idea that voices and faces are integrated in the brain. For example, de Gelder and her colleagues investigated the interplay between emotional voices and emotional faces (De Gelder, Pourtois, & Weiskrantz, 2002) using electroencephalography (EEG). Their results indicated that facial expressions influence the way emotional voices are processed. The localisation of crossmodal voice–face interactions in the brain is, however, still a topic of investigation. Heteromodal cortical regions, such as the bilateral posterior superior temporal sulcus, that are well known as cortical regions of sensory convergence, have also been implicated as regions where the integration of voices and faces occurs (e.g. Calvert, Campbell, & Brammer, 2000). However, other cortical regions previously believed to be involved primarily in the unimodal processing of faces, such as the fusiform face area (or FFA), or voices, such as the temporal voice area (or TVA), appear to be activated by associated stimuli from a different sensory modality (von Kriegstein & Giraud, 2006). In particular, stronger connections between TVA and FFA were found following the acquisition of associations between faces and voices than associations between other kinds of audio–visual stimuli (von Kriegstein, Kleinschmidt, & Giraud, 2006), suggesting that these cortical areas are specialised for the processing of person-related information.

The effect of the distinctiveness of an item on recognition memory has been well documented. In general, distinctiveness effects are considered to be within-domain such that an item is distinctive relative to other items within the same category. For example, many studies have shown that when memorising a list of words, words that stand out because of their visual or conceptual qualities

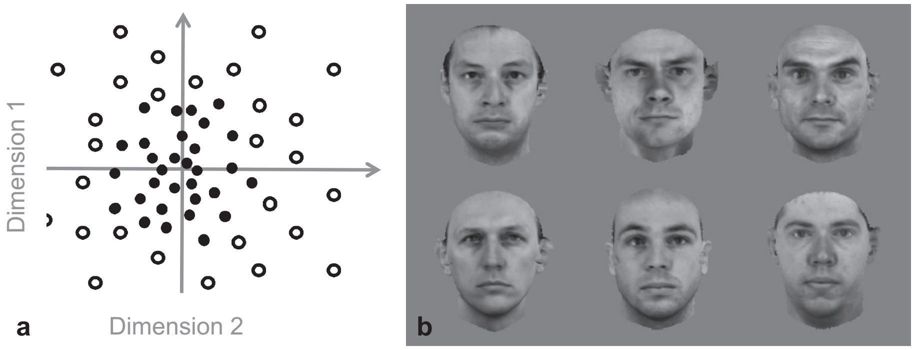

Fig. 1. (a) Schematic representation of multidimensional face space represented along two arbitrary dimensions describing facial properties. Each dot represents one face, with open circles representing distinctive faces, and black circles representing typical faces. (b) Examples of face images used as stimuli in the experiments.

(for example, they are written in a different colour or they refer to a different theme than all other words in the list) are better remembered in subsequent recognition tasks than other more typical words in the list (von Restorff, 1933; Wallace, 1965). In the field of face recognition, many studies have shown that distinctive faces are more easily recognised than typical faces (e.g. Bruce et al., 1994; Light, Kayra-Stuart, & Hollander, 1979; Bartlett, Hurry, & Thorley, 1984). In most studies, faces are distinct because they contain an unusual feature or combination of features (large nose, elongated face) or faces can be rendered distinct using computer graphics which can caricature the face (e.g. Deffenbacher, Vetter, Johanson & O'Toole, 1998).

With the exception of studies on speech perception (e.g. Rossi-Katz & Arehart, 2009), to our knowledge all previous studies on the effects of distinctiveness in person recognition were based on either facial information or vocal information (Latinus, McAleer, Bestelmeyer, & Belin, 2013; Mullennix et al., 2011) separately. Our goal here was to investigate whether face distinctiveness can be modulated by crossmodal interactions with auditory stimuli. Specifically we asked whether unfamiliar faces can become perceptually more distinct, and therefore better remembered, following a learned association with stimuli (voices) in another modality that are themselves distinctive.

Our study was designed to test whether voices enhance explicit memory for faces, and whether these effects were limited to voice–face interactions only. To this end, we first tested whether associating voices with static images of unfamiliar faces during learning would affect the subsequent recognition of those faces. We chose recordings of different voices as auditory stimuli because of the obvious natural association between voices and faces: we often view the face of a person whilst hearing them speak at the same time. Moreover, previous studies have shown that matching voices to videos of unfamiliar faces can be achieved to a level of performance greater than chance in delayed matching-to-sample tasks (Kamachi, Hill, Lander, & Vatikiotis-Bateson, 2003), and that the presence of articulating dynamic faces benefits performance in subsequent voice recognition tasks (Schweinberger et al., 2007), suggesting that person recognition involves direct associations between voices and faces. Since facial distinctiveness is known to be associated with improved recognition, we tested whether voices, previously rated as distinctive, when paired with unfamiliar faces would enhance the subsequent visual recognition of the face relative to when typical voices were paired with those faces. If visual and auditory representations are processed independently for the purpose of person recognition, then we would not necessarily expect any crossmodal benefit from the voice information on face recognition.

In view of the literature suggesting that there is sensory convergence of various inputs for person recognition, and the natural correspondence between these sources of information, we also asked whether voices and faces represent a special case for crossmodal memory enhancement or whether this effect might also be observed with arbitrary auditory and face associations. In sum, the results of our study should help provide insight into the nature of the face space, or norm-based frameworks of face and voice representations in memory (such as those proposed by Valentine, 1991; Latinus et al., 2013 and many others) and, more specifically, elucidate whether the aspect of face distinctiveness should be expanded to include multi-sensory person information.

2. Experiment 1a

In this experiment we tested whether voice information associated with static images of unfamiliar faces during learning would affect the subsequent recognition of those faces.

2.1. Methods

2.1.1. Participants

We recruited 21 undergraduate students (10 female) from the Eberhard-Karls University of Tübingen, Germany for this study. All observers (age range between 18 and 40 years) in this and all following experiments were paid volunteers and naive to the purpose of the experiments. None participated in more than one experiment. All participants had normal or corrected-to-normal vision and none reported any auditory impairment. The procedure was approved by local IRB, and written consent was obtained from each participant before the experiment.

2.1.2. Visual face stimuli

Grey-scale, static images of 60 faces were derived from 3D laser-scans of individual male heads (Blanz & Vetter, 1999). All faces were presented from a full-face view under identical conditions of illumination. The size and the luminance of all face images were equated (Graf & Wichmann, 2002). Faces were devoid of hair, beards and jewellery (see Fig. 1b). Each digital face image (256 × 256 pixels in size) subtended approximately 7° × 7° of visual angle at a viewing distance of about 70 cm. The face images were presented on a computer monitor against a black background at the centre of the screen.

2.1.3. Auditory stimuli

Auditory stimuli consisted of 29 recordings of male and female voices. We created two separate sets of auditory stimuli based on typical and distinctive voice patterns. The set of typical stimuli consisted of recordings of voices of male speakers only. Each speaker was recorded while saying the same text in German, except for the use of an individual first name. The German text was as follows: "*Hallo, ich heiße Rainer (Kurt, Christian…). Das ist ein Foto von mir. Merke Dir bitte mein Gesicht!*" (English translation is: "*Hello, my name is Rainer (Kurt, Christian…). This is a photo of me. Please remember my face*"). Different first names were used to ensure maximal naturalness of the voice–face pairing for the participants of the experiment. The distinctive set of stimuli consisted mainly of recordings of voices of males but we also recorded two female speakers specifically for this set. The meaning of the sentence uttered for all distinctive voice stimuli was the same as in the typical recordings, but the distinctive voice recordings were further defined as each voice having some unique quality. For example, some of the distinctive voices differed from a typical voice in the language or dialect used while others differed in intonation, wording (involving synonym changes only such that e.g. the word "picture" was substituted by "image") or other distinctive qualities which were relative to the typical stimuli. We recorded, for example, a voice speaking in Japanese, in German with a Swiss German accent, a female voice speaking in German with a French accent, and a chanting voice. This corresponds to the concept of distinctiveness in faces, where a face may be distinctive because of different characteristics from the rest of the stimuli in the set, for example because of its elongated shape, or because of its large nose, very blue eyes, etc. All auditory stimuli lasted between 3 and 5 s and were grossly equated for loudness. Participants heard the auditory stimuli via a set of headphones.

In order to validate our distinctive and typical voice categories, we conducted a rating task on all voice stimuli. Ten independent judges were instructed to categorise each stimulus according to how distinctive they thought the vocalisation was. As such the raters evaluated of how different each stimulus was perceived from the other stimuli within the same category (i.e. voices). Participants responded using 3 keys on a button box. Specifically, they were instructed to press the left button if they thought the voice was 'typical'; press the right button the voice was 'distinctive' and the middle button if the voice was 'neither distinctive nor typical'. The responses were arbitrarily scored as follows: left button responses were scored as '1', middle button responses were scored as '3' and right button responses were scored as '5'. On average, a rating of 4.00 (SE = .19) and 2.29 (SE = .19) was given to the voice stimuli in our distinctive and typical category respectively, and a paired *t*-test found that those ratings were significantly different from each other [*t* (9) = 3.27, *p* = .01, *r* = 0.74].[1] Throughout this study, we will name those auditory stimuli *distinctive* or *typical* "voices".

2.1.4. Design

The experiment was based on a one-way, within-subjects design with *voice distinctiveness level* (typical or distinctive) paired with faces as the main factor.

We randomly divided the set of unfamiliar faces into two sets; set A and set B. For any participant, distinctive voices were paired with all faces in set A (or set B) and typical voices were paired with faces in set B (or set A). Distinctive voices and face set pairings were counterbalanced across participants. Within each set we pseudo-randomly paired voices with faces across participants (each face was paired with one of five different voices), thus controlling for the potential effect of superior memorability for certain faces or audio–visual pairings over others.

2.1.5. Procedure

The experiment comprised of a learning session followed by an old/new recognition test session. During the learning session participants were presented with 24 unfamiliar face–voice pairs. Of these, 12 faces were presented with typical voices (*t pairs*) and the other 12 faces were paired with distinctive voices (*d pairs*). Each face–voice stimulus was presented for 5 s and was preceded by a fixation cross for 250 ms. At the end of each trial in the learning session, participants pressed a key of the button box which triggered the onset of the subsequent face–voice stimulus. Participants were instructed to remember the faces and their associated voices. Each face–voice pair was repeated three times during the learning session and stimuli were presented in a random order across participants.

The test session immediately followed the learning session. There was a short training block before the test session using voice and face stimuli not used during the test session. In the test session, participants were presented with faces only, without any auditory pairings. All 24 faces that had been viewed during the learning phase (*i.e. old* faces) were randomly presented among an equal number of distractor (i.e. *new*) faces. Each old and new face was shown once during the test session. Participants were instructed to classify each face as old or new as fast and as accurately as possible by pressing the associated left or right keys on a button box. Button press responses were counterbalanced across participants. Each trial started with a fixation cross which was shown for 250 ms followed by a face image which was shown for 3 s followed by a blank screen. A response could be made at any time from the onset of the presentation of the face. No feedback was provided. A new trial started 500 ms after a response was made. All experiments were conducted in a dimly lit room. The stimuli were presented on a PC and the experiment was programmed using Eprime software (Psychological Software Tools, Inc.).

2.2. Results

The results of two participants were eliminated because debriefing revealed that they had not understood the procedure. All 19 remaining participants (10 females) performed better than chance at classifying all faces (i.e. as either old or new): mean correct response 80.4%, (SE = 2.6). They needed an average of 1741 ms (SE = 216) to respond to each face.

Prior to analyses, we removed outliers in the reaction time data which were determined as greater or less than 3 standard deviations from the mean for each participant. For each participant, the mean RTs were calculated from correct responses only.

We were specifically interested in comparing recognition performance between *old* faces that were previously associated with *distinctive* voices and *old* faces that were previously associated with *typical* voices (i.e. faces of *d pairs* versus *t pairs* in the old group). The mean number of correct responses for faces of the *d pairs* (*M* = 84.7%, SE = 2.9) was greater than the mean number of correct responses for *t pairs* (*M* = 78.4%, SE = 2.9) as shown in Fig. 2. Using a paired *t*-test, we found that this difference was significant [*t* (18) = 2.34, *p* = .03, *r* = 0.48], indicating that target faces previously paired with distinctive voices were better remembered. Although reaction times were 45 ms faster to faces of the *d pairs* (*M* = 1520 ms, SE = 192) than to *t pairs*

[1] See results of Experiment 2 in which we compare the recognition of faces paired to voices which were categorised as very 'typical' or very 'distinctive' only.

(*M* = 1565 ms, SE = 195), as shown in Fig. 2, this difference failed to reach significance [*t* (18) = 0.63, *p* = .54].

Finally, for completeness, we briefly report separately performance for *old* and *new* faces: mean correct response 81.7%, SE = 3.9 and 79.4%, SE = 2.5 for the old and new faces, respectively. Participants needed an average of 1709 ms (SE = 200, *old* faces) and 1773 ms (SE = 235, *new* faces) to respond to the face images. There was no significant difference in the number of correct responses [*t* (18) = .56, *p* = .58] and in response times [*t* (18) = .92, *p* = .37] between the *old* and *new* face types.

2.3. Discussion

In this experiment, we found that the previous association of a face with a distinctive voice during learning resulted in more accurate recognition of that face relative to a face that was previously associated with a typical voice. These findings suggest that the perceptual quality of the voice, i.e. its distinctiveness, can interact with visual stimuli to modify their perceptual quality, thus rendering unfamiliar faces relatively more distinct.

Interestingly, participants were only 45 ms faster at identifying target faces previously associated with distinctive than typical voices and this difference was not significant, suggesting that the crossmodal voice information has a greater benefit on response accuracy than the speed of recognition. These findings are in contrast to many previous studies using faces previously rated as typical or distinctive where a facilitation is found for recognising distinctive faces in response times only (e.g. Valentine & Bruce, 1986a, 1986b) or both response times and accuracy (Bartlett et al., 1984). Although it is not completely clear why we do not find a significant effect with response times in our own study, previous studies have suggested that the relative benefit on response times or accuracy may depend on the nature of the task used (Shepherd et al., 1991). It is possible, therefore, that crossmodal effects may manifest more in accuracy than on speed responses with our design. For example the failure to find a significant effect on response times may have been due to the face image remaining on the screen until the participant responded. This methodology may have encouraged large individual differences in reaction time performance across participants (from less than 1 s to 3.6 s) thus reducing the overall difference across the voice conditions.

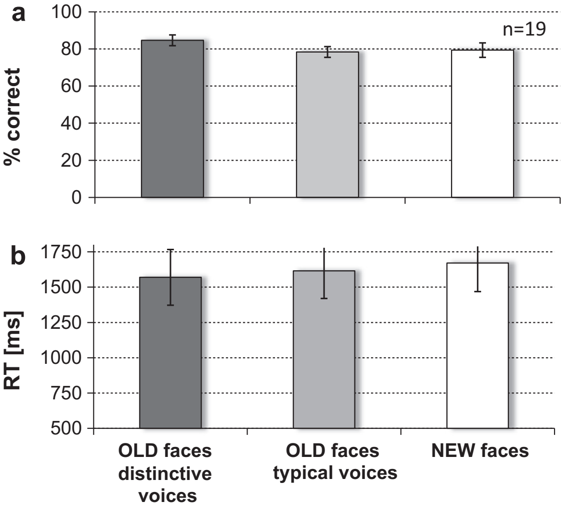

Fig. 2. Plots showing the results of Experiment 1a. (a) Mean percent correct responses to old faces is shown separately for faces that had been previously paired with distinctive voices (dark grey bar) or with typical voices (light grey bar). (b) Mean response time (ms) to old faces is shown separately for faces that had previously been paired with distinctive voices (dark grey bar) or with typical voices (light grey bar). In both a and b, the mean response performance for new faces (white bars) is shown for comparison only. All error bars represent ±1 standard error of the mean.

Although our findings suggest a crossmodal benefit of person relevant information on recognition, an alternative explanation is that faces that were paired to distinctive voices during learning were better remembered because of relative differences in arousal levels induced by the distinctive versus typical voices. As such, some face stimuli were better remembered because they were paired with distinctive voices which induced higher levels of attention on the faces, thus affecting encoding of the faces only, rather than because of any crossmodal enhancement on the representation of unfamiliar faces in memory. To confirm the specificity of the face–voice association and to test the alternative explanation mentioned above, we used other auditory stimuli in the next experiment.

3. Experiment 1b

In this experiment we tested whether arbitrary sounds associated with static images of unfamiliar faces during learning would also affect the subsequent recognition of those faces. We used the same face images as in the previous experiments as visual stimuli but here we used non-speech sounds as auditory stimuli that are not naturally associated to faces. We repeated the explicit recognition memory paradigm described in Section 2.1.5 (Experiment 1a). If voice and face pairings are special, or if the acquisition of associations between faces and voices is facilitated due to the natural correspondence between these sources of information (O'Mahony and Newell, 2012), then we expected no particular benefit on recognition memory for unfamiliar faces when paired with arbitrary sounds during learning.

3.1. Methods

3.1.1. Participants

Twenty-two participants (11 females, age range: 21–35 years) were recruited under the same conditions as mentioned in Experiment 1a.

3.1.2. Visual and auditory stimuli

The same face stimuli as used in the previous experiments were used. To create the auditory stimuli we used a synthesiser (Triton LE from Korg Inc.) that offers a broad palette of instruments ranging from standards such as piano to various exotic synthetic sounds (with 48-kHz sampling). Sounds were chosen over melodies or reversed speech stimuli, as such stimuli have been reported to have similar acoustic properties to speech (Patel, 2005) and speech-like stimuli might engage the same mechanisms for associations as voices. For the *typical* sounds, we recorded twelve, 3-note chords alternatively in minor (for example A–C–E) and major (for example C–E–G) keys using an instrumental sound labelled ‘acoustic piano’. The chords were distributed over the whole keyboard. For the *distinctive* sounds, several instrument options (e.g. pan flute, chime) and various electronic sounds (including rotor noise of a helicopter, old fashion telephone ring, distorted electric guitar) were used and chords were played where possible. All sound stimuli were audible for between 3 and 5 s and were grossly equated for loudness. Twelve typical sounds and twelve distinctive sounds were used. As the auditory stimuli had been created for the purpose of the experiment, we conducted a rating task to establish the distinctiveness of each stimulus. An independent group of 27 judges rated all sound stimuli using a similar rating procedure as described in Section 2.1.3 (Experiment 1). The mean ratings provided were 3.94 (SE = 0.07) and 2.31 (SE = 0.12) for the categorised *distinctive* and *typical* sounds respectively and the ratings differed significantly from each other [t (26) = 11.91, $p < 0.001$, $r = 0.92$].[2] For convenience, we named those auditory stimuli *distinctive* or *typical* “sounds”.

3.1.3. Design and procedure

The same experimental design and procedure described in Sections 2.1.4 and 2.1.5 (Experiments 1a) was used here, with the exception that we used 24 arbitrary sounds (12 distinctive and 12 typical sounds) rather than voices as auditory stimuli. As in the previous experiments we controlled for the effect of potential superior memorability of certain faces or pairs by pseudo-randomly pairing the sound and face stimuli across participants during learning (using 4 different learning sets of sound-face pairs). Trials were randomly presented across participants in both the learning and test sessions. As in Experiment 1a, only face stimuli were presented during the test.

3.2. Results

The data from one participant was eliminated because a technical problem occurred during the learning phase with the result that some of the stimuli were not presented. Accuracy performance at classifying all faces (i.e. as either old or new) for the other 21 participants (10 female) was better than chance: mean correct response 80.8% (SE = 2.4). Participants needed an average of 1627 ms (SE = 208) to respond to the face stimuli.

Prior to analyses, we removed outliers in the reaction time data which were determined as greater or less than 3 standard deviations from the mean for each participant. For each participant, the mean RTs were calculated from correct responses only. The mean accuracy rates and RTs across the distinctive and typical sound-face pairs with the correctly identified old faces are plotted in Fig. 3. We conducted a paired-samples t-test on the correct responses between faces of the *t pairs* and *d pairs* and found no significant difference [t (20) = 0.64, p = .53], indicating that there was no advantage in accuracy performance for identifying faces previously paired with distinctive sounds (for *d pairs* the mean correct response was 77.8%, SE = 3.7) than typical sounds (i.e. *t pairs*, mean correct response 76.1.6%, SE = 3.1). Moreover, reaction times were very similar to faces of the *d pairs* (1442 ms, SE = 183) and *t pairs* (1448 ms, SE = 191), and these differences were not significant [t (20) = .99, p = .92].

Finally, for completeness, we briefly report separately performance for *old* and *new* faces: mean correct response 77.0%, SE = 3.1 and 84.9%, SE = 2.8 for the old and new faces respectively. Although participants were significantly better at correctly recognising new than old faces [t (20) = 2.25, p = .036, r = 0.45], they needed less time to respond to the old faces (M = 1445 ms, SE = 185) than to new faces (M = 1585 ms, SE = 208) [t (20) = 2.33, p = .03, r = 0.46].

3.3. Discussion

Although rated as distinctive, arbitrary sounds learned with unfamiliar faces did not affect the subsequent recognition of those faces relative to typical sound pairings with faces. The results of this experiment contrast to those reported in Experiment 1a where distinctive voices were associated with a greater benefit on accurate face recognition performance than typical ones. These findings suggest that arbitrary sounds do not seem to share the same accessibility to faces in memory representations, since these sounds do not affect the subsequent recognition of the faces.

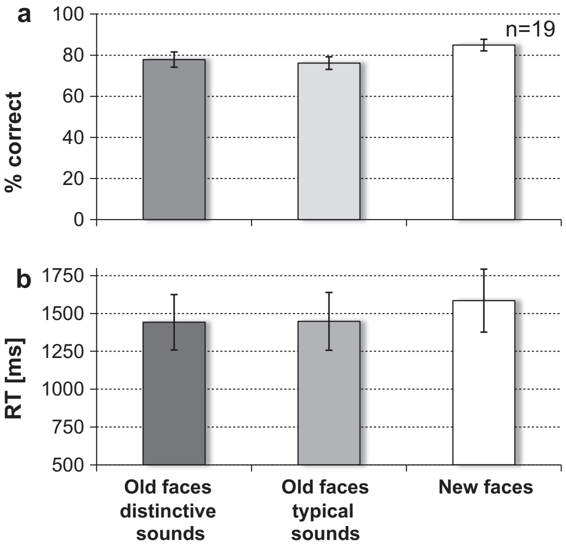

Fig. 3. Plots showing the results of Experiment 1b. (a) Mean percent correct responses to old faces is shown separately for faces that had been previously paired with distinctive sounds (dark grey bar) and with typical sounds (light grey bar). (b) Mean response time (ms) to old faces is shown separately for faces that had been paired to distinctive sounds (dark grey bar) and to typical sounds (light grey bar). For both plots, the mean response performance to new faces (white bars) is shown for comparison only. All error bars represent ±1 standard error of the mean.

[2] See results of Experiment 2 in which we compare the recognition of faces paired to sounds which were categorised as very ‘typical’ or very ‘distinctive’ only.

The accuracy performance is equivalent across experiments, suggesting that the difference between the effect of voices and arbitrary sounds on face recognition is not due to differences in the difficulty of the task. However, a second possible reason could be that the difference between experiments was due to a greater disparity between distinctiveness and typicality for the voices than for the sounds. The results of our rating tasks on the distinctiveness of voices (Experiment 1a) or sounds (Experiment 1b) suggest that this is also not a sufficient explanation for the discrepancy between the experiments. For example, distinctive voices were given an average rating of 4, whereas distinctive sounds were given a very similar average rating of 3.95. Likewise, typical voices were given an average rating of 2.29 and typical sounds were given a similar average rating of 2.28. Moreover, these ratings were collected from different groups of participants and were not significantly different from each other for the distinctive stimuli [unpaired *t*-test, t (35) = 0.24, ns] or the typical stimuli [unpaired *t*-test, t (35) = −0.08, ns]. Therefore, we would argue that it is unlikely that the differences in results between voice–face associations (Experiment 1a) and sound-face associations (Experiment 1b) are due to differences in the perceptual quality of the distinctive and typical stimuli.

Furthermore these findings support our hypothesis that the higher recognition performance in Experiment 1a was obtained for faces learned with distinctive voices because of the perceptual quality of those voices, and not because of a higher arousal level induced by them compared to typical faces. However, to ensure that the benefit on face recognition was specific to voices, and that our effects were replicable using a different design, in the following experiment we combined Experiment 1a and Experiment 1b and tested, using a within-subjects design, a naive group of participants with face stimuli learned with both types of auditory stimuli (i.e. voices and sounds).

4. Experiment 2

Here we tested participants' recognition memory for faces that had been previously learned with either voices or with arbitrary sounds. Our main goal was again to test whether voices and faces represent a special case for cross-modal memory enhancement or whether this effect also occurs with other sound and face associations while directly comparing the influence of distinctive voices and sounds. A secondary aim was to determine whether an advantage could be found on recognition times to faces previously learned with distinctive voices over those learned with typical voices. To that end, we provided feedback on participants' response times following each test trial.

4.1. Methods

4.1.1. Participants

We recruited 24 participants (17 female, age range between 19 and 59 years) under the same protocol conditions as described in Experiment 1a.

4.1.2. Visual face stimuli

Twenty additional face images (taken from the same original database of 3D laser scanned faces) were added to the face set to total 80 faces. All other details are the same as in the previous experiments, except that here all face images were presented against a grey background, and the presentation time of the test stimulus was reduced.

4.1.3. Voice stimuli

The voice stimuli comprised of a set of twenty recordings of male and female voices (taken from our larger set of 24 voice clips previously used in Experiment 1a). All other details are the same as in the previous experiments. On average, a rating of 4.16 (SE = .35) and 2.11 (SE = .34) was given to the voice stimuli in our distinctive and typical categories respectively, and a paired *t*-test found that those ratings were significantly different [t (9) = 3.35, p = .009, r = 0.74].

4.1.4. Sound stimuli

Twenty recordings of sounds (taken from the original 24 sounds) were used. The mean ratings for those sounds were 4.30 (SE = 0.09) and 2.10 (SE = 0.14) for the categorised *distinctive* and *typical* sounds respectively and the ratings differed significantly from each other [t (26) = 12.89, $p < 0.001$, r = 0.92]. All other details are the same as in the previous experiments.

4.1.5. Design

The experiment was based on a two-way, within-subjects design with *type of auditory stimuli* (voice or sound) *and distinctiveness level* (typical or distinctive) as main factors.

The experiment was divided in two blocks. In one block, during learning, faces were associated with voices and in the other block the faces were associated with sounds. Block order was counterbalanced across participants. As in the previous experiments, each block comprised a learning and a test phase. All unfamiliar faces had been divided, pseudo-randomly, into four sets of 20 faces set. Care was taken to distribute face stimuli used in the previous experiment (Experiments 1a and 1b) and new faces evenly across all conditions. For the learning session, participants learned the faces of one set, i.e. the 'target' set. These target faces were presented together with an auditory stimulus during this session. For each learning session, 10 of the 20 faces were paired with distinctive auditory stimuli (*d pair faces*), the other 10 were paired with typical auditory stimuli (*t pair faces*). For the test phase, participants saw all 20 learned faces from the target set, which were randomly presented and intermixed with faces of another, distractor face set (new faces). The task for each participant was the same old/new recognition test described in the previous experiment. Across participants and in the learning sessions, each face was paired equally often to a distinctive or a typical auditory stimulus and each face set was used equally often as either the target set or the distractor (i.e. new) set. In this way, we controlled for any potential effect of superior memorability for certain faces or audio–visual pairings over others. Trials were randomly presented across participants in both the learning and test sessions.

4.1.6. Procedure

The experiment comprised of two blocks. Each block consisted of a learning session which was immediately followed by an old/new recognition test session. Each block differed according to the nature of the auditory stimuli which were used during the learning session. As such, in each block, either voice stimuli or arbitrary sounds were used. In each learning session, participants were required to learn 20 faces which were paired with typical auditory stimuli (*t pair faces*, 10 pairs) and with distinctive auditory stimuli (*d pair faces*, 10 pairs). All other details of the learning session are as described before.

In the test session, only face stimuli were presented and the exposure time of the test faces was shortened and response time was limited: participants were warned that they had to respond within 2 s. A short training session familiarised them with the experimental procedure before each experimental block. At test, each trial started with a fixation cross shown for 250 ms followed by a face image shown for 1 s followed by a blank (grey) screen shown for 1 s. A response could be made at any time from the onset of the presentation of the face and during the presentation of the blank screen. After the blank screen, written feedback appeared on the screen for 1500 ms congratulating them for responding on time when participants responded within the given time window. When participants did not enter a response during the given response time window, both written feedback and an auditory beep alerted them to their failure to respond within the time limit. A new trial started 500 ms after the feedback. Faces were not repeated across blocks. Block order was counterbalanced across participants.

4.2. Results

For both blocks (voice and sound) the percentage of trials where a response was not made within the time limit out of all trials was negligible (less than 0.01% in both cases). In the sound block, overall accuracy performance was 75.3%, (SE = 2.6). For the voice block, accurate performance was 77.0%, (SE = 2.4).

For each participant, the mean RTs were calculated from trials which were correctly responded to only. In the sound block, participants needed an average of 847 ms (SE = 28 ms) to respond. In the voice blocks, faces were responded to by 892 ms (SE = 44) on average.

Our main purpose was to compare recognition performance between *old* faces that were previously associated with *distinctive* auditory stimuli (i.e. voices or sounds) and *old* faces that were previously associated with *typical* auditory stimuli (i.e. *d pair faces* versus *t pair faces* in the old group).

We analysed the data of both blocks together to directly compare the effect on face recognition of the two types of auditory stimuli during learning. See Figs. 4 and 5 for plots of the results (accuracy and RT) for each condition and Table 1 for a list of the values. Separate, repeated-measures ANOVAs with *type of auditory stimuli* (sound or voice) by *distinctiveness level* (typical or distinctive) as main factors, with partial eta squared ($\eta_p{}^2$) as an index of effect size (Cohen, 1988) were conducted on the accuracy and the RT data. The analysis of the recognition accuracy data revealed no main effects of type of auditory stimuli or distinctiveness level (both [$F(1,23) < 1$, ns). More pertinently, the interaction between these factors was significant [$F(1,23) = 12.61$, $p = .002$, $\eta_p{}^2 = .35$]. Post-hoc analyses using Tukey HSD revealed that auditory distinctiveness modulated face recognition differently depending on the type of auditory stimuli used. Specifically *d pair faces* were better recognised than *t pair faces* only when voices were used ($p = .024$) but not when sounds were used ($p = .248$). A 2 × 2 ANOVA conducted on the reaction times data also yielded no main effects of auditory stimuli [$F(1,23) = 2.11$, $p = .16$] nor distinctiveness level [$F(1,23) = 2.84$, $p = .11$]. There was no interaction between the factors [$F(1,23) < 1$, ns]. Given these null effects, no further analyses of the response times were conducted.

In order to ensure that the results were not simply due to our categorisation of the distinctive and typical auditory

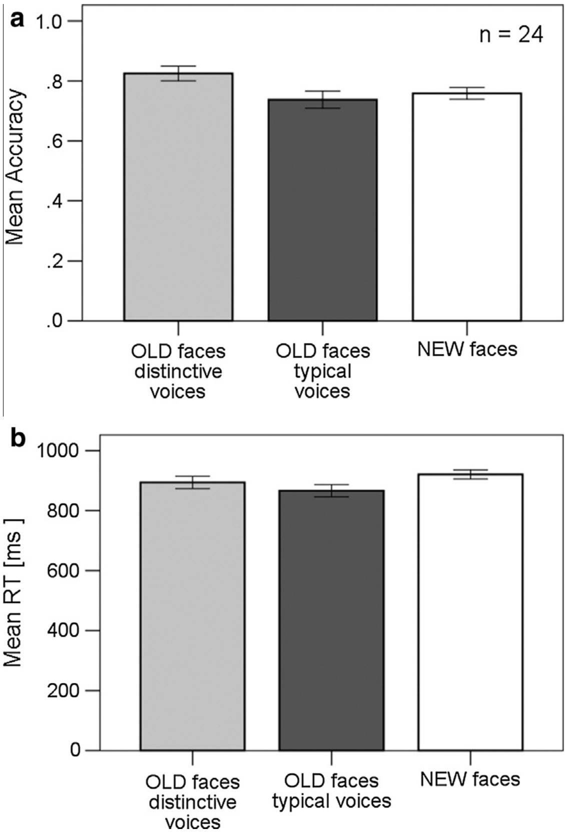

Fig. 4. Plots showing the results of the voice block in Experiment 2. (a) Mean response performance (percent correct) for old faces is shown separately for faces that had been previously paired with distinctive voices (dark grey bar) or with typical voices (light grey bar). (b) Mean response time (ms) to old faces is shown separately for faces that had previously been paired with distinctive voices (light grey bar) or with typical voices (dark grey bar). In both plots the mean response performance for new faces (white bars) is shown for comparison. All error bars represent ±1 standard error of the mean.

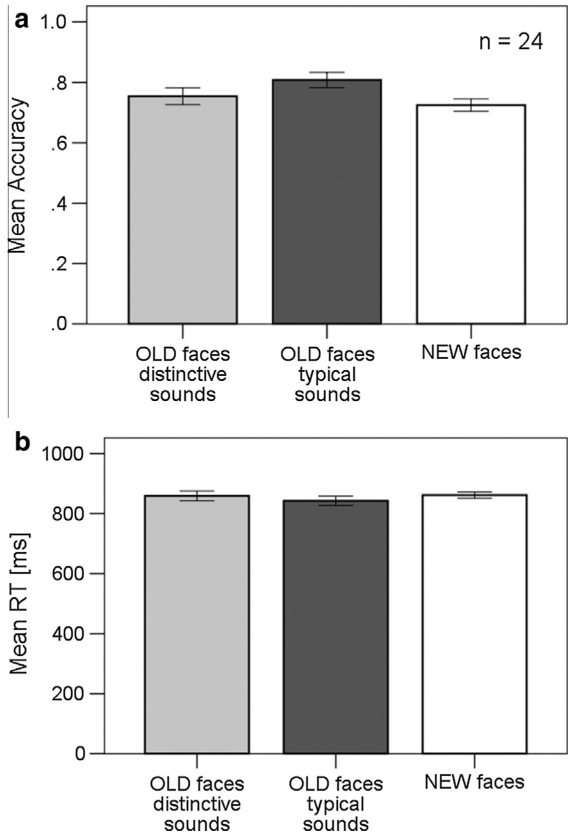

Fig. 5. Results of the sound block in Experiment 2. (a) Mean percent accuracy in responding to old faces is shown separately for faces that had been paired to distinctive sounds (light grey bar) and to typical sounds (dark grey bar) during learning. (b) Mean response time (ms) to old faces is shown separately for faces that had been paired to distinctive sounds (light grey bar) and to typical sounds (dark grey bar). For both plots, the mean response performance to new faces (white bars) is shown for comparison. All error bars represent ±1 standard error of the mean.

stimuli, we reanalysed the data using only those stimuli which were previously categorised by naïve observers as being very 'distinctive' or 'typical' (see validity studies under Sections 2.1.3 and 3.1.2 above). To that end we set a threshold of 25% around the "middle" rating of 3 (with 1 being "most typical" and 5 "most distinctive) and removed data from trials related to voices and sounds that had received the most ambiguous ratings, that is voices and sounds whose ratings were between 2.25 and 4.0. With this threshold, we removed 3 of the 10 stimuli of the distinctive and typical voice groups and in the typical sound group and 2 stimuli for the distinctive sound group. This exclusion of trials resulted in an increase in the average distinctive ratings, and a decrease in the average typical ratings for the remaining stimuli relative to the original ratings: For distinctive and typical voices: 4.43 SE = 0.08 and 1.86 SE = 0.08 respectively; for distinctive and typical sounds 4.51 SE = 0.10 and 1.91 SE = 0.12 respectively. Note also that the mean distinctiveness levels within each auditory type differ more from each other than in the original analysis. The accuracy performance values for the reduced trial sets are indicated in parentheses in Table 1. A repeat of the 2 (*type of auditory stimuli*) by 2 (*distinctiveness level*) repeated-measures ANOVA on this stimulus set revealed no significant effect of auditory type [$F(1,23) = 1.28$, $p = .269$, $\eta_p^2 = .05$]. However, a main effect of distinctiveness was found [$F(1,23) = 4.42$, $p = .047$, $\eta_p^2 = .16$]. This main effect emerged because performance for faces coupled to distinctive voices was significantly higher than in the original analyses with all trials included ($p = .001$), while there was no difference between performance across data sets for all other auditory pairings (all $ps > .091$). More importantly, the interaction between distinctiveness level and auditory type was significant [$F(1,23) = 14.56$, $p = .001$, $\eta_p^2 = .39$]. Again, post hoc analyses using Tukey HSD revealed that auditory distinctiveness differentially modulated face recognition depending on the type of auditory stimuli used, with *d pair faces* better recognised than *t pair faces* only when voices were used ($p = .0010$) while for sounds no difference was found ($p = .794$). Thus, despite the increased distinctiveness differences between the typical and the distinctive category, most notably for the sounds, this new analyses revealed no advantage for learning faces paired with distinctive sounds.

We argue that while arbitrary sounds are not easily associated to faces, voices are naturally associated to face representations and thus this association may result in a new dimension added to the visual face space. This association should therefore result in changed distinctiveness for the faces with a concomitant modulation of the recognisability of that face. To test this idea further, we correlated the distinctiveness ratings previously provided to the different sounds and voices with the recognition accuracy performance obtained in the current experiment. The results are shown in scatter plot form in Fig. 6. We found that this correlation was significant for voices [Pearson correlation; $r = .501$, $p = .029$] but not for sounds [$r = -.208$, $p = .380$]

Table 1
Mean accuracy and response times taken to respond to each of the face conditions relative to the paired voice or sound. The values in parentheses relate to the analysis of performance based on trials in which only voices or sounds rated high in either distinctiveness or typicality were included (see text for details).

Accuracy	Old faces		New faces
	d pairs	*t* pairs	
Voices	82.5%, SE = 2.3 (*88.2%, SE = 2.5*)	73.8%, SE = 4.1 (*72.6%, SE = 4.4*)	75.8%, SE = 3.1
Sounds	75.4%, SE = 1.8 (*74.7%, SE = 3.9*)	80.8%, SE = 1.4 (*77.9%, SE = 3.9*)	72.5%, SE = 3.8
Response times			
Voices	898 ms, SE = 45	865 ms, SE = 48	905 ms, SE = 45
Sounds	844 ms, SE = 33	826 ms, SE = 31	869 ms, SE = 27

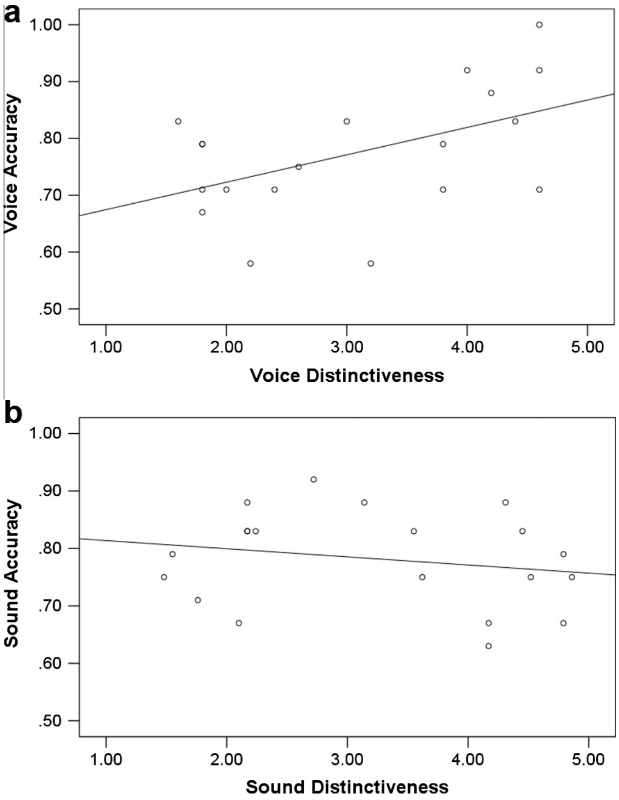

Fig. 6. Scatter plots showing the correlation (i.e. best linear fit) between the distinctiveness of (a) voice stimuli and (b) sound stimuli and the accuracy with which the faces were subsequently recognised in Experiment 2. Only the correlation shown in (a) was significant (see text for details).

indicating that the more distinctive a voice was, the more recognisable was its associated face pair. This result was not the case for sounds: faces associated with very distinctive sounds were not subsequently better recognised during the test phase.

Finally, for completeness, we performed separate, repeated-measures ANOVAs with *type of auditory stimuli* (sound or voice) by *face* (old or new) as main factors for reaction time and accuracy data. For the reaction time data, there was a main effect of face type [$F(1,23) = 5.757$, $p = .025$ $\eta_p^2 = .2$], with old faces more quickly recognised than new faces. There was no main effect of type of auditory stimuli [$F(1,23) = 1.808$, $p = .192$] and no interaction [$F(1,23) < 1$, ns]. Thus in both blocks, old faces were responded to faster than new faces, and the novelty of the faces influenced response times in the same way, independently of the type of the auditory stimuli used. For accuracy data, there was no main effects of face type or auditory stimuli [all Fs < 1] and no interaction between the factors [$F(1,23) = 1.922$, $p = .179$]. Thus performance measures were equivalent across both blocks, indicating that the difficulty of both tasks was comparable.

4.3. Discussion

Using a within-subject design, in this experiment we investigated whether the perceptual quality of a voice or a sound associated to a face during learning can affect the subsequent recognition of that face. Our results confirm the findings of the first two experiments which tested each type of auditory stimulus in separate groups of participants. First, we found again that accuracy and response time data for old and new faces were similar across blocks suggesting that the difference between the effect of voices and arbitrary sounds on face recognition is not due to differences in the difficulty of the task. Second, more importantly, we replicated our previous finding that the association of a face with a distinctive voice during learning subsequently benefited recognition of that face relative to a face that was previously associated with a typical voice. Moreover, we found that no such effects of distinctiveness on face recognition were found for arbitrary sounds, even when we reduced the analysis to those trials involving the most typical and distinctive sound and voice stimuli. Third, we again failed to find any significant effect on response times, despite a design that was slightly modified to encourage speedy responses from our participants.

In this experiment we used a subset of the auditory stimuli used in the previous experiment, therefore it was necessary to ensure that the stimuli were equally distinctive across voices and sounds. Prior to Experiment 2, the distinctive voices used as stimuli were given an average rating of 4.16, whereas distinctive sounds were given a slightly higher average rating of 4.30. Typical voices were given an average rating of 2.10 and typical sounds were given a similar average rating of 2.18. As in the previous experiments, these ratings were not significantly different from each other for the distinctive and typical stimuli [unpaired *t*-tests, all $ps > 0.57$]. Based on this result and the similar result from the previous experiment, therefore, we would argue that a difference between voices and sounds in terms of their perceptual distinctiveness is unlikely to explain our results.

Most importantly, in Experiment 2 we found that not only was recognition more accurate for faces previously learned with distinctive voices but not sounds, but we could also show that for the same participant, auditory distinctiveness impacted recognition accuracy statistically differently depending on the type of auditory stimuli used. Additionally, the differing influence of voices and sounds on face recognition was not only observed at the category level (distinctive or typical): there was also a positive correlation between the distinctiveness of the voice (as given by its mean distinctiveness rating) and the recognisability of its associated face at test. This correlation was not observed for sounds. We also noticed that some of the sounds were easily identified (for example helicopter rotor noise and old fashion phone ring) and were associated with high distinctiveness ratings, but, this had no effect on face recognition since their associated faces were not the best recognised. In fact our results suggest that slightly better performance was associated with target faces which were previously learned with typical sounds (e.g. piano chords) as suggested by the scatter plot in Fig. 6b (although this was not significant).

5. General discussion

The aim of our study was to investigate whether voices and faces are integrated in memory for the purpose of person identification and whether the nature of a person's voice can affect the subsequent recognition of their face. We found that the quality of the voice to which a face was paired affected subsequent face recognition such that faces initially paired with distinctive voices during learning were better remembered. Moreover, our findings support our hypothesis that the higher recognition performance obtained for faces learned with distinctive voices was specific to voices only and did not generalise to other arbitrary sounds. Furthermore, the benefit on face recognition occurred because of the perceptual quality of the associated voices during learning, and not because of a higher arousal level induced by distinctive compared to typical voices.

In contrast to previous studies which investigated the interactions between voices and faces in person recognition, we investigated whether crossmodal interactions were affected by the perceptual nature of the associated auditory stimuli by manipulating distinctiveness. Distinctiveness effects are well known in face perception, and more recently in voice perception (Latinus et al., 2013), and *norm-based* coding models have proposed to take account of these findings. For example, according to Valentine (1991), face space models that implement either a norm-referenced (e.g. Leopold, O'Toole, Vetter, & Blanz, 2001) or an examplar-based encoding of faces can account for the effect of face distinctiveness. In a purely visual model of face space, visually distinctive faces are located further away from the centre of the multi-dimensional space and further apart from other faces than more typical ones and are thus more easily recognised in a recognition task because there are fewer opportunities for them to be mixed up with nearby neighbours. Our results may be accommodated into such a model if face space was not purely visual, but took into account all relevant information for person recognition. For example, voice distinctiveness could be accommodated as another dimension in a multisensory representation framework, as a very simplistic assumption, and this added dimension may result in visually typical faces becoming distinctive. Thus we would expect these faces to be easier to recognise while this would not happen for typical faces associated with typical voices. A representation of this multisensory face space could be easily illustrated with Fig. 1 by having one of the two dimensions represented in the illustration to symbolise voice distinctiveness.

Recent studies have suggested that motion may be an important cue to facial identity (Lander & Bruce, 2000; Lander & Chuang, 2005; Hill & Johnston, 2001 and Knappmeyer, Thornton, & Bülthoff, 2003) and face detection (Pilz, Thornton, & Bülthoff, 2006). Thus, those studies and the results of the present study suggest that frameworks of face memory, such as the *face space* model, should accommodate other dimensions than the visual static dimensions describing a face. There are other aspects to faces which the model does not take into account, such as short term idiosyncratic changes due to expressions or speech or long-term changes due to ageing (Craw, 1995; Lewis & Johnston, 1999; Preminger, Blumenfeld, Sagi, & Tsodyks, 2009). Moreover, these studies suggest that information from another source, provided it is related to person identity, can affect face recognition and consequently the structure of face space (see also Schweinberger et al., 2007). Therefore, our results further imply that a more complex multisensory representation which includes not only visual characteristics but also other associated perceptual qualities from other relevant modalities, is required in any proposed model of person recognition.

Other functional models of face perception have been proposed to account for the hierarchy of processing stages involved in identifying a person from their face (e.g. Bruce & Young, 1986; Burton, Bruce, & Johnston, 1990; Calder & Young, 2005). In their review on voice perception, Belin and colleagues (Belin et al., 2004) proposed to adapt the Bruce & Young model to build a similar functional organisation for voice processing. As such, the interaction between voices and faces in person recognition could be accounted for by assimilating both voice and face recognition models. Moreover, our findings that voices can enhance face distinctiveness supports the idea that voices directly access face information at an early perceptual stage rather than at a later more cognitive stage (see also von Kriegstein, Kleinschmidt, Sterzer, & Giraud, 2005; von Kriegstein & Giraud, 2006).

Our findings raise the obvious question as to why face recognition benefits from learned associations with voices but not other sounds. We suggest that a general effect of cross-modal expertise, or learned correspondences across modalities, may underpin this effect. Specifically, we have developed a life-time's exposure to unique pairings of faces with voices (see e.g. Lewkowicz & Kraebel, 2004) and acquired relatively early on in life the processes involved to optimally encode both vocal and facial information together (Lewkowicz & Hansen-Tift, 2012). Thus it may be that the processing of audio–visual information for the purpose of person identification may be highly efficient relative to the processing of novel audio–facial pairings. Moreover, it might be possible that through the course of development and associated experience, as well as a result of an innate predisposition to certain cortical connectivity patterns (Wilmer et al., 2010; Zhu et al., 2010), a specific cortical network exists for the processing of person-specific visual and auditory information for the purpose of identification (see e.g. Kanwisher & Yovel, 2006). Since this network would rely on audio–visual information that is unique to the individual to be recognised, it is unlikely to be involved when arbitrary associations are learned between faces and non-specific information such as random sounds or names (O'Mahony and Newell, 2012; von Kriegstein & Giraud, 2006). For example, von Kriegstein and Giraud found that specific crossmodal associations increase brain activation in areas such as the fusiform cortex, including area FFA. These activations were specific to voice and face associations and not other associations such as voices with written names or cell (mobile) phones with ring tones.

Our results are in accordance with the results of von Kriegstein and Giraud, in that we found that face recognition benefited from the perceptual qualities of associated voices but not from associations with other more arbitrary sounds. In the study of von Kriegstein and Giraud, however, speech information was presented together with a moving face, thus it is possible that spatiotemporal redundancies between the movements of the mouth and the emitted sounds contributed to their observed activations (Lewkowicz & Kraebel, 2004). Indeed, Kamachi et al., 2003) have shown that matching an unfamiliar voice with the correct one of two articulating faces is better than chance, suggesting that person-specific identity information can be carried by both the visual and auditory modalities. Moreover, this surprising effect occurs even when the visually articulated sentence and heard sentences differ from each other, suggesting that similarity in structure or content across modalities is not necessary for crossmodal identity matching, although the effect can be disrupted with changes in voice intonation (Lander, Hill, Kamachi, & Vatikiotis-Bateson, 2007).

The participants in our studies learned to associate voices with faces that were contrived associations in the sense that we did not present the voices with faces that matched in the real world. The learned face and voice associations were randomly paired, yet our findings nevertheless suggest a benefit for these more ecological associations over other face and arbitrary sound associations. However, it is possible that our voice–face findings may have been further strengthened if we had used dynamic, articulating faces rather than static faces. More specifically, consistent effects on response times may be more likely to be found with dynamic faces. In any case, our results suggest that the specificity and robustness of face–voice associations is independent of physical congruency between the content of the speech in the voices and the related facial movements. In other words, we would argue that the relative efficiency at which faces and voices are associated occurs because we are expert at pairing voices to faces.

As previously mentioned, growing interest in multisensory perception has resulted in numerous studies investigating the locus of cross-modal interactions for faces and voices in the brain (e.g. von Kriegstein et al., 2005) or the effect of face distinctiveness on brain activity (Loffler, Yourganov, Wilkinson, & Wilson, 2005). As our study has shown, the quality of information in one modality, i.e. voice distinctiveness, can affect crossmodal interactions such that the perception of associated visual information is improved. Our findings raise an interesting question regarding the neuro-anatomical substrates underlying face perception, in that, it remains to be seen how crossmodal influences that enhance the representation of a stimulus in one modality might be implemented at the cortical level.

Acknowledgments

We thank Daniel Berger for help creating the auditory stimuli, Franziska Hausmann, Karin Bierig, Saskia Kühnhold and Silke Wittkowski for help in programming the experiments and testing participants, and Corrina Maguinness for help with data analyses. We are grateful to our colleagues for allowing us to record their voices for use as stimuli. This research was supported by the Max Planck Society and by funding from Science Foundation Ireland (Grant No. 10/IN.1/I3003) awarded to FNN.

References

Bartlett, J., Hurry, S., & Thorley, W. (1984). Typicality and familiarity of faces. *Memory and Cognition, 12*, 219–228.

Belin, P., Shirley, F., & Bedard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences, 8*, 129–135.

Blanz, V. & Vetter, T. (1999). A morphable for the synthesis of 3D faces. In *Symposium on Interactive 3D Graphics—Proceedings of SIGGRAPH'99* (pp. 187–194).

Bruce, V., Burton, A. M., & Dench, N. (1994). What's distinctive about a distinctive face. *Quarterly Journal of Experimental Psychology Section A – Human Experimental Psychology, 47*, 119–141.

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

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

Byatt, G., & Rhodes, G. (2004). Identification of own-race and other-race faces: Implications for the representation of race in face space. *Psychonomic Bulletin and Review, 11*, 735–741.

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

Calvert, G. A., Brammer, M. J., & Iversen, S. D. (1998). Crossmodal identification. *Trends in Cognitive Sciences, 2*, 247–253.

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology, 10*, 649–657.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Craw, I. (1995). A manifold model of face and object recognition. In T. Valentine (Ed.), *Cognitive and computational aspects of face recognition: Explorations in face space* (pp. 183–203). London: Routledge.

De Gelder, B., Pourtois, G., & Weiskrantz, L. (2002). Fear recognition in the voice is modulated by unconsciously recognised facial expressions but not by unconsciously recognised affective pictures. *Proceedings of the National Academy of Science, 99*, 4121–4126.

Deffenbacher, K. A., Vetter, T., Johanson, J., & O'Toole, A. J. (1998). Facial aging, attractiveness, and distinctiveness. *Perception, 27*, 1233–1243.

Graf, A., & Wichmann, F. (2002). Gender classification of human faces. In H. H. Bülthoff, S.-W. Lee, T. A. Poggio, & C. Wallraven (Eds.), *Biologically Motivated Computer Vision, LNCS 2525* (pp. 491–501). New York: Springer.

Hanley, J., Smith, S., & Hadfield, J. (1998). I recognise you but I can't place you: An investigation of familiar-only experiences during tests of voice and face recognition. *Quarterly Journal of Experimental Psychology Section A – Human Experimental Psychology, 51*, 179–195.

Hanley, J., & Turner, J. (2000). Why are familiar-only experiences more frequent for voices than for faces? *Quarterly Journal of Experimental Psychology Section A – Human Experimental Psychology, 53*, 1105–1116.

Hill, H., & Johnston, A. (2001). Categorizing sex and identity from biological motion of faces. *Current Biology, 11*, 880–885.

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.

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

Knappmeyer, B., Thornton, I. M., & Bülthoff, H. H. (2003). The use of facial motion and facial form during the processing of identity. *Vision Research, 43*, 1921–1936.

Lander, K., & Bruce, V. (2000). Recognizing famous faces: Exploring the benefits of facial motion. *Ecological Psychology, 12*, 259–272.

Lander, K., & Chuang, L. (2005). Why are moving faces easier to recognize? *Visual Cognition, 12*, 429–442.

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

Latinus, M., McAleer, P. M., Bestelmeyer, P. E., & Belin, P. (2013). Norm-based coding of voice identity in human auditory cortex. *Current Biology, 23*(12), 1075–1080.

Lee, K., Byatt, G., & Rhodes, G. (2000). Caricature effects, distinctiveness, and identification: Testing the face-space framework. *Psychological Science, 11*, 379–385.

Leopold, D. A., Bondar, I., & Giese, M. (2006). Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature, 442*, 572–575.

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

Lewis, M. B., & Johnston, R. A. (1999). Are caricatures special? Evidence of peak shift in face recognition. *European Journal of Cognitive Psychology, 11*, 105–117.

Lewkowicz, D. J., & Hansen-Tift, A. M. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech. *Proceedings of the National Academy of Science, 109*, 1431–1436.

Lewkowicz, D. J., & Kraebel, K. (2004). The value of multimodal redundancy in the development of intersensory perception. In G. A. Calvert, C. Spence, & B. Stein (Eds.), *Handbook of multisensory processing* (pp. 655–678). Cambridge: MIT Press.

Light, L., Kayra-Stuart, F., & Hollander, S. (1979). Recognition memory for typical and unusual faces. *Journal of Experimental Psychology. Human Learning and Memory, 5*, 212–228.

Loffler, G., Yourganov, G., Wilkinson, F., & Wilson, H. (2005). FMRI evidence for the neural representation of faces. *Nature Neuroscience, 8*, 1386–1390.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746–748.

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

Newell, F. N., Chiroro, P., & Valentine, T. (1999). Recognizing unfamiliar faces: The effects of distinctiveness and view. *Quarterly Journal of Experimental Psychology Section A – Human Experimental Psychology, 52*, 509–534.

O'Mahony, C., & Newell, F. N. (2012). Crossmodal integration of faces and voices, but not faces and names, in the recognition of unfamiliar persons. *British Journal of Psychology, 103*, 73–82.

Patel, A. D. (2005). The relationship of music to the melody of speech and to syntactic processing disorders in aphasia. *Annals of the New York Academy of Sciences, 1060*, 59–70.

Pilz, K. S., Thornton, I. M., & Bülthoff, H. H. (2006). A search advantage for faces learned in motion. *Experimental Brain Research, 171*, 436–447.

Potter, T., & Corneille, O. (2008). Locating attractiveness in the face space: Faces are more attractive when closer to their group prototype. *Psychonomic Bulletin and Review, 15*, 615–622.

Preminger, S., Blumenfeld, B., Sagi, D., & Tsodyks, M. (2009). Mapping dynamic memories of gradually changing objects. *Proceedings of the National Academy of Sciences of the United States of America, 106*, 5371–5376.

Rossi-Katz, J., & Arehart, K. H. (2009). Message and talker identification in older adults: Effects of task, distinctiveness of the talkers' voices, and meaningfulness of the competing message. *Journal of Speech, Language and Hearing Research, 52*, 435–453.

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

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

Shepherd, J. W., Gibling & Ellis, H. (1991). The effects of distinctiveness, presentation time and delay on face recognition. *European Journal of Cognitive Psychology, 3*, 137–145.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Quarterly Journal of Experimental Psychology Section A – Human Experimental Psychology, 43*, 161–204.

Valentine, T., & Bruce, V. (1986a). Recognizing familiar faces: The role of distinctiveness and familiarity. *Canadian Journal of Psychology, 40*, 300–305.

Valentine, T., & Bruce, V. (1986b). The effect of race, inversion and encoding activity upon face recognition. *Acta Psychologica, 61*, 259–273.

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

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

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 Restorff, H. (1933). 'Über die Wirkung von Bereichsbildung im Spurenfeld' [On the Effect of Field Formation on the Trace Field]. *Psychologische Forschung [Psychology Research], 18*, 299–342.

Wallace, W. P. (1965). Review of the historical, empirical, and theoretical status of the von Restorff phenomenon. *Psychological Bulletin, 63*, 410–424.

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Williams, M., Loken, E., et al. (2010). Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of Sciences of the USA, 107*, 5238–5241.

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology, 81*, 141–145.

Yovel, G., & Belin, P. (2013). A unified coding strategy for processing faces and voices. *Trends in Cognitive Sciences, 17*(6), 263–271.

Zhu, Q., Song, Y., Hu, S., Li, X., Tian, M., Zhen, Z., et al. (2010). Heritability of the specific cognitive ability of face perception. *Current Biology, 20*, 137–142.