Face-voice space: Integrating visual and auditory cues in judgments of person distinctiveness

Joshua R. Tatz¹ · Zehra F. Peynircioğlu¹ · William Brent²

Published online: 21 July 2020
© The Psychonomic Society, Inc. 2020

Abstract

Faces and voices each convey multiple cues enabling us to tell people apart. Research on face and voice distinctiveness commonly utilizes multidimensional space to represent these complex, perceptual abilities. We extend this framework to examine how a combined face-voice space would relate to its constituent face and voice spaces. Participants rated videos of speakers for their dissimilarity in face only, voice only, and face-voice together conditions. Multiple dimensional scaling (MDS) and regression analyses showed that whereas face-voice space more closely resembled face space, indicating visual dominance, face-voice distinctiveness was best characterized by a multiplicative integration of face-only and voice-only distinctiveness, indicating that auditory and visual cues are used interactively in person-distinctiveness judgments. Further, the multiplicative integration could not be explained by the small correlation found between face-only and voice-only distinctiveness. As an exploratory analysis, we next identified auditory and visual features that correlated with the dimensions in the MDS solutions. Features pertaining to facial width, lip movement, spectral centroid, fundamental frequency, and loudness variation were identified as important features in face-voice space. We discuss the implications of our findings in terms of person perception, recognition, and face-voice matching abilities.

Keywords Person distinctiveness · Faces · Voices · Multisensory integration · Multiple dimensional scaling

Introduction

Most of us can distinguish, recognize, and identify other people from just their faces or voices. We rapidly and perhaps even automatically extract such characteristics as age, gender, and race from people's faces and voices (e.g., Belin, Fecteau,

& Bédard, 2004; Wiese, Schweinberger, & Neumann, 2008). The information that faces and voices convey may be considered complementary because the involved modalities convey similar – but not entirely redundant – information about individuals (Campanella & Belin, 2007). For instance, a person's face and voice may both inform us that the person is female. However, other features are unique. We can only get eye color from people's faces and accents from people's voices. Of course, in everyday life, we frequently encounter other people through their faces and voices concurrently (if not along with other person-specific sources of information such as body position). Subsequently, faces and voices may be an ideal model with which to compare how the brain represents complementary information from different sensory modalities as well as what bimodal presentations of information add beyond the information present in each constituent sensory modality.

Although the literature combining faces and voices remains small (especially when compared to the wealth of studies on faces and voices separately), one apparent trend in both unimodal and bimodal studies on faces and voices is that faces seem to be more informative than voices in terms of people's identities, emotional states, and characteristics. For instance, voices are notoriously recognized less accurately than faces in episodic recognition tasks, and the same holds true in more

Significance Statement

Although face distinctiveness versus voice distinctiveness measures were only slightly correlated with one another, combined face-voice distinctiveness was best predicted by multiplying separate face and voice distinctiveness measures. This indicates that face and voice cues are combined in an interactive manner that makes people more distinct. Because we routinely encounter others through faces and voices, it is relevant to consider the implications of these findings for more general models of person recognition and perception. We suggest that person perception may in part be based on representing others based on whether our expectations for how faces and voices go together are confirmed or violated.

✉ Joshua R. Tatz
Joshua.tatz@gmail.com

¹ Department of Psychology, American University, Washington, DC 20016, USA

² Department of Performing Arts, American University, Washington, DC 20016, USA

semantic tasks, as in the identification of personally familiar individuals (e.g., Damjanovic, 2011). Our perception of emotional states, too, appears to be impacted more heavily by facial than vocal cues, although a greater reliance on vocal cues has been found under significant visual distortion (e.g., Collignon et al., 2008). Similarly, faces may impact gender identification more than voices because individuals who were perfect at identifying singers' gender (albeit indirectly because participants judged voice parts such as soprano or tenor) from voices singing a gender-neutral pitch fell below chance when voices were dubbed with task-irrelevant visual information depicting a singer of the opposite sex (Peynircioğlu, Brent, Tatz, & Wyatt, 2017). In both unimodal and bimodal settings, age estimation seems to rely more on faces than voices (e.g., Moyse, 2014; but see Smith, Dunn, Baguley, & Stacey, 2016a, for similar unimodal abilities). Although more direct comparisons are needed, identification of race from faces is quite accurate (e.g., Valentine & Endo, 1992; Zhao & Bentin, 2011), whereas identification of race from voice is considerably poorer (e.g., Walton & Orlikoff, 1994). Similarly, Rezlescu et al. (2015) found that faces had more impact on attractiveness judgments than voices. However, faces and voices contributed about equally to trustworthiness judgments, and voices had more impact in dominance judgments. Of current interest is a different person attribute, that of distinctiveness, which perhaps has more direct implications to person perception and recognition memory.

To date, distinctiveness has been examined separately in the face and voice literature using similar methodology but with slightly different goals and interpretations. In both, commonly, participants rate pairs of stimuli for (dis)similarity and then the resulting data are analyzed with some member of a family of multivariate statistical techniques known as multiple dimensional scaling (MDS). MDS techniques are designed to reduce the dimensionality of the data and represent the underlying structure as distances in a perceptual space (Lattin, Carroll, & Green, 2003). Although there is some early work in which MDS was used to create face spaces (e.g., Pedelty, Levine, & Shevell, 1985; Shepherd & Deregowski, 1981), arguably, the most influential work in this domain did not employ an MDS analysis but instead proposed that a similar mechanism might underly human face processing and explain several recognition effects. Valentine (1991) suggested that we may represent faces in relation to some average (or exemplar) face as points in a multidimensional space where each dimension is thought to reflect some different attribute (or set of attributes) that may be used to individuate faces. Importantly, this framework fared well in explaining the effects of distinctiveness, inversion, and race on face recognition. In general, the number of dimensions involved in representing faces are thought to be many and situationally dependent (Valentine, 1991). Following this notion, studies that have employed MDS analyses to (dis)similarity data for

faces have usually sought to maximize the number of dimensions included in their solutions: we know of four studies that selected MDS solutions with four or more dimensions (Byatt & Rhodes, 2004; Johnston, Milne, Williams, & Hosie, 1997; Lee, Byatt, & Rhodes, 2000; Potter, Corneille, Ruys, & Rhodes, 2007), two studies that selected three-dimensional solutions (Nishimura, Maurer, & Gao, 2009; Wallraven, 2014), and one study that selected a two-dimensional solution (Papesh & Goldinger, 2010). Perhaps because of the emphasis on distinctiveness, as reflected by the psychological *distance* from an average face (situated approximately at the origin of the face space), less emphasis has been placed on determining what attribute(s) correspond to which dimension – if any are proposed. Instead, they have been based on the experimenters' intuition. Some dimensions that have been suggested across studies are face width, forehead size, and age (e.g., Wallraven, 2014). Of course, consistent with Valentine's (1991) notion that individuals will exploit whatever attributes are available, if a categorical attribute such as race is available, then this will likely be the first (and therefore most important) dimension in face space (Byatt & Rhodes, 2004; Papesh & Goldinger, 2010).

The voice tradition differs somewhat from the face tradition in that, rather than viewing MDS as an end in itself, the main objective is to use correlational analyses to identify the attribute(s) that correlate most strongly with the dimensions returned by MDS (Baumann & Belin, 2010; Chhabra, Badcock, Mayberry & Leung, 2012; Kreiman, Gerratt, Precoda, & Berke, 1992; Murry & Singh, 1980; Singh & Murry, 1978; Walden, Montgomery, Gibeily, Prosek, & Schwartz, 1978). Thus, the voice approach is more feature driven, and the number of dimensions selected is usually more manageable (i.e., usually two or three). Commonly implicated dimensional correlates include features pertaining to fundamental frequency (f_0), speech duration, various formants, and age. As with the face studies using MDS, when participants can discriminate on the basis of socially important, categorical information, these seem to be the most important features in the resulting solution. For instance, this is the case when speakers differ with respect to their gender (Murry & Singh, 1980; Singh & Murry, 1978). Despite this similarity to Valentine's (1991) idea of "average faces," there has been little consideration of whether voices might be represented similarly. One exception is a functional imaging study in which activity in a putative voice region showed positive correlation with Euclidean distance from the average voice in an acoustic space involving fundamental frequency, formant dispersion, and harmonic-to-noise ratio as its three dimensions (Latinus, McAleer, Bestelmeyer, & Belin, 2013).

At present, we entertain the possibility that both faces and voices may contribute to a more general "person space," a combined Gestalt-like space influenced by the two modalities of information. That is, we are primarily interested in how

facial and vocal cues combine to differentiate unfamiliar individuals from one another. We ask this question at three levels: First, in a broad sense, we ask to what extent face-voice space resembles its constituent face and voice spaces. Given the common observation of face dominance in person perception and recognition, we expect the space to resemble faces more, but – because faces and voices contain complementary information and are frequently encountered together – we still expect voices to have at least some bearing on the face-voice space. Second, in a more granular sense, we can explore which features in the face and voice spaces are retained in their combined face-voice space. Indeed, Wallraven (2014) used a similar approach in demonstrating that similarity ratings for face molds explored visually versus haptically resulted in quite similar perceptual spaces (although with some small differences in the implicated features). Third, from a more computational standpoint, we can examine whether the combination of face and voice cues is better characterized by additive or interactive integration. For instance, Fujisaki, Goda, Motoyoshi, Komatsu, and Nishida (2014) found that participants' ratings for how well the sound and image of an object fit a material category (e.g., wood) were best predicted by multiplying (rather than adding) the weights of ratings from the individual modalities. In the same way, we can consider how distinctiveness is combined across faces and voices, although, in this case, we use the distinctiveness inferred from the MDS solution rather than the ratings themselves.

Method

Participants

Twenty-four American University students (14 female, 10 male, $M_{Age} = 20.1$ years, age range: 18–30 years) participated in the study for research credit in psychology courses. We targeted this sample size based on experimental design considerations (a multiple of six was required) and Wallraven (2014), who compared visual and haptic face spaces with 22 participants (which is comparatively well powered for a typical MDS study). Sixteen of the participants self-identified as White/Caucasian, seven as Asian (two of whom also self-identified as White/Caucasian), and four as Hispanic/Latinx (one of whom also self-identified as White/Caucasian). All reported normal or corrected-to-normal hearing and vision, and all provided informed consent prior to participation. The study was approved by the American University institutional review board.

Materials

The audiovisual speech stimuli included videos of 24 speakers (12 female, 12 male, $M_{Age} = 26.1$ years, age range: 20–35

years, all White/Caucasian individuals) from the CREMA-D database (Cao et al., 2014). A between-subject t-test did not indicate any age difference between male and female speakers ($t < 1$). The speakers each uttered three different statements matched for their number of syllables (“I’m on my way to the meeting,” “I would like a new alarm clock,” “That is exactly what happened”). Visually, the stimuli depicted dynamic (i.e., moving) speakers with approximately frontal viewpoints. The speakers were filmed in the same environment wearing similar gray shirts against a green background (Cao et al., 2014). The selected speakers did not have facial hair, jewelry, or glasses. We did not obscure the speakers’ hair from the video. Each video had been previously judged to feature emotionally neutral content (Cao et al., 2014).

Using VideoPad Video Editor (<https://www.nchsoftware.com/videopad/index.html>) face-only (i.e., speech articulations and movements were visible, but no audio was present), voice-only (which included a central fixation cross on a blank screen), and face-voice versions of each spoken statement were created and rendered as .mov files (640 × 480-pixel resolution, 44.1 kHz audio sampling rate). Stimuli were delivered via MATLAB 2018a (The Mathworks, Inc., Natick, MA, USA) with Psychophysics Toolbox 3 on an HP Envy Laptop (1,920 × 1,080-pixel resolution, 39.6 cm display) and Sony MDR-7506 headphones.

Participants provided their ratings by clicking and dragging the mouse horizontally to update an integer value on a 101-point, rectangular scale. The scale was initialized at the mid-point value of 51 and interactively displayed the participant’s selection. Centered above either end of the scale were the words “0 no different” and “100 most different.”

Design and procedure

There were three conditions: face-voice, face-only, and voice-only conditions. Stimuli were presented in pairs, one after the other. Comparisons were made for all possible pairings of speakers except that we did not have participants compare speakers across genders or across conditions. The stimulus pairs always featured different sentences (even when they were for the same speaker). This served to introduce greater speech variability to the stimuli so that the voice was of primary importance rather than the speech contents (e.g., Lavan, Burton, Scott, & McGettigan, 2019) and to limit the total length of the experiment so that attention would not wane. The participants heard every sentence spoken by each speaker multiple times, but the particular sentences used for each stimulus pairing was randomly determined beforehand and counterbalanced across six groups of participants so that each possible sentence pairing was encountered equally often. The presentation order of the trials (each consisting of one stimulus pair) was randomly determined for each participant. The presentation order of the stimuli within a pair was also

randomized in each trial (i.e., some trials were AB and others BA). Participants rated a total of 468 pairs (78 comparisons within a given gender and condition). Due to an initial programming error, two participants did not rate the full 468 pairs but instead rated 404 or 437 pairs. Fortunately, the type of MDS analyses we employed is well equipped to handle missing data (De Leeuw & Mair, 2011; Lattin et al., 2003). We therefore included the available data from these participants.

Before the experiment, participants were informed that they would be rating pairs of speakers based on how different their faces and/or voices were. They were further instructed to provide a rating of zero only if they were certain the pairs featured the same individual and a rating of 100 only if they thought the pairs were as different as two individuals of similar ages and the same race and gender could theoretically be. Three practice trials featuring different individuals from those in the main experiment served to familiarize participants with the procedure in each condition. In the main experiment, each video varied from about 1 s to 2.5 s in duration. The items within a pair were separated by 1 s of silence. Participants had up to 7 s to provide their ratings but were able to advance more quickly if they desired. Another second elapsed after their response before the next trial began. Thus, each trial lasted approximately 8 s, and the entire session lasted approximately 1.5 h.

Analyses

Distinctiveness measures First, dissimilarity judgments were analyzed using individual differences scaling (INDSCAL) MDS for interval data using the *R* package *smacof* (De Leeuw & Mair, 2011). It is important to note that the INDSCAL model assumes that individuals base their judgments on similar attributes but may differ with respect to how important those attributes are. Thus, the INDSCAL model provides a map of perceptual space on the group level and accounts for individual differences through noting the transformations that individual perceptual spaces must undergo to conform to the group space.

Importantly, in the resulting individual perceptual spaces, the Euclidean distance from the origin may be used to indicate the distinctiveness of one speaker relative to all others in the space (e.g., Johnston et al., 1997). This is specified by the Euclidean formula, where X_i refers to the distance (d) from the origin in the i th dimension, and n represents the total number of dimensions:

$$d = \sqrt{\sum_{i=1}^n (x_i - 0)^2} \quad (1)$$

Using these Euclidean distances as a measure of distinctiveness, we then examined whether audiovisual distinctiveness was better predicted by a multiplicative or additive

operation (cf. Fujisaki et al., 2014). If the operation is multiplicative, audiovisual distinctiveness should best be predicted by multiplying normalized auditory distinctiveness by normalized visual distinctiveness. That is, auditory and visual distinctiveness should interact such that the resulting audiovisual distinctiveness should exceed the sum of its constituent parts. This outcome would be indicated by a greater regression coefficient to an interaction term (β_3) than either of the regression coefficients corresponding to auditory only (β_1) or visual only (β_2) distinctiveness. If the operation is additive, however, the interaction coefficient should be near zero, and it should be sufficient to predict audiovisual distinctiveness through visual-only and auditory-only distinctiveness terms (i.e., a model without the interaction). The critical models to compare are thus as follows:

$$AV = \beta_{\text{Intercept}} + \beta_1 A_{\text{Only}} + \beta_2 V_{\text{Only}} + \beta_3 A_{\text{Only}} V_{\text{Only}} \text{ (Interactive)} \quad (2)$$

$$AV = \beta_{\text{Intercept}} + \beta_1 A_{\text{Only}} + \beta_2 V_{\text{Only}} \text{ (Additive)} \quad (3)$$

Feature extraction As another, largely exploratory goal, we also examined potential auditory and visual feature correlates with the dimensions resulting from each MDS solution. The full list of auditory and visual features we examined are included in Appendix 1. Since we were interested in relatively stable features that were specific to the speakers' identity (rather than speech variation), we averaged each auditory and visual feature measure across the three sentence types. We considered all the auditory features that Baumann and Belin (2010) examined using the same PRAAT software (P. Boersma & D. Weenink, <http://www.fon.hum.uva.nl/praat/>). In addition, we also examined auditory features pertaining to formant bandwidth (using PRAAT) as well as descriptive statistics pertaining to spectral centroid and mel-frequency cepstral coefficients (using the Auditory Toolbox in MATLAB 2019a). The visual features we considered included distances between several points on the speakers' faces and the movement of those face features. The distances between face features (also known as second-order, configurational relationships) included 30 distances originating from 36 points that Cox, Ghosn, and Yianilos (1996) used to train an algorithm to recognize once-encountered faces in new images. Following Cox et al. (1996), the distances were normalized by the speakers' pupil distances. We additionally considered four points about the forehead to quantify forehead width, height, and area. The experimenter manually determined the x- and y-pixel coordinates for the face-feature points. For this purpose, an image frame was selected from each video in which the speaker's lips were closed and the eyes were open. We next tracked the x- and y-position change of 33 face-feature points (see Appendix 1) across each video frame using the Kanade-

Lucas-Tomasi (KLT) feature-tracking algorithm (using the Computer Vision Toolbox in MATLAB 2019a), an algorithm optimized for tracking face points based on their inferred three-dimensional properties (Tomasi & Kanade, 1991). One difference from the KLT feature-tracking algorithm is that we tracked the face features we were interested in rather than the unconstrained points recommended by the algorithm. To ensure accurate feature tracking, each video was monitored with the overlaid tracking points. We quantified feature motion by summing the change in Euclidean distance across each frame and dividing the resulting values by the respective speaker's pupil distance. Of course, with this metric, longer videos could contain more motion simply because they have more frames. Nevertheless, we focused on total feature motion because using average feature motion (i.e., further dividing total feature motion by the number of frames in each video) could sacrifice important motion information resulting from timing differences and/or inter-frame variability.

Results

We used MDS analyses with INDSCAL to generate individual and group perceptual spaces from participants' dissimilarity judgments (De Leeuw & Mair, 2011). Respectively, these perceptual spaces were derived to support subsequent analyses aimed at (1) characterizing audiovisual integration and (2) identifying important features in person distinctiveness. We performed separate INDSCAL analyses as a function of condition and speaker gender. Data were treated as interval. *STRESS* values indicated that four-dimensional solutions provided fair model fits for each analysis (Male face only: .180, Male voice only: .178, Male Face-Voice: .178, Female face only: .177, Female voice only: .179, Female Face-Voice: .179). For convenience, we depict only the first three dimensions of each four-dimensional solution in Fig. 1 (see also Appendix 2 for average dissimilarity matrices).

Audiovisual integration

We next used regression analyses to examine whether audiovisual distinctiveness would be better described by an additive or multiplicative (i.e., interactive) operation. To this end, the normalized Euclidean distances corresponding to each speaker in each participants' voice-only and face-only perceptual spaces (and the interaction) were used to predict the observed face-voice space (see Equations 2 and 3). Thus, the 24 participants each provided distinctiveness measures for each of the 24 speakers, resulting in a total of 576 (24×24) observations. Note that half of these observations originated from the distinctiveness measures pertaining to female speakers and half from male speakers. Initially, a hierarchical linear model was considered using the main conditions of interest, speaker

gender as a dummy-coded predictor, and participant identity as a random-effect term. The random-effect term partitioned very little variance from the model ($\sigma^2 = .0013$). We therefore considered a multiple regression model (i.e., without the random-effect term). In this model, speaker gender was not found to be a significant predictor, $\beta = -0.002$, $t(571) = 0.21$, $p = .834$. This term was therefore also dropped from the model. Subsequently, we examined the additive and interactive models using multiple regression, following the model Fujisaki et al. (2014) used for material category judgments. An *F*-test for overall significance indicated that the interactive model (Equation 2) was significant, $F(3, 572) = 121.0$, $p < .001$, $R^2 = .39$, as was the interaction term, $\beta_3 = 1.13$, $t(572) = 3.74$, $p < .001$. More importantly, an *F*-test for nested models indicated that the improvement of the interactive model over the simpler additive model was also significant, $F(1, 572) = 14.0$, $p < .001$. As with Fujisaki et al. (2014), we also examined the extent to which each predictor in the interactive model separately correlated with AV ratings. Figure 2 depicts these plots. The interaction term correlated most closely with AV ratings ($r = .59$, $R^2 = .35$) followed by V_{Only} ratings ($r = .54$, $R^2 = .29$) and trailed by A_{Only} ratings ($r = .38$, $R^2 = .14$) (all $ps < .001$). A small but statistically reliable correlation was found between auditory- and visual-only ratings ($r = .19$, $R^2 = .04$, $p < .001$). Given that these relationships were all positive, the proportion of variance explained (R^2) is perhaps the more informative number. Very little variability in face distinctiveness can be accounted for by voice distinctiveness (or vice versa). At most, this suggests only a small degree of redundancy in the information that these sources of information convey. By contrast, these relatively independent sources of information enabled better person discrimination than either source independently (as indicated by the multiplicative interaction term). Moreover, the additional variance accounted for by the interaction term over the V_{Only} term is much more than the variance shared by the A_{Only} and V_{Only} terms. This indicates that the improved predictability of the AV ratings with the interaction term cannot be accounted for by redundant cues present in the faces and voices. Rather, encountering the faces and voices together must have prompted cues to be used somewhat differently in distinctiveness judgments.

Same-speaker comparisons

Wilcoxon signed-rank tests were used to compare participants' aggregate distinctiveness ratings across face, voice, and both conditions for the same speakers. These correspond to the diagonals of the dissimilarity matrices (see Appendix 2). The comparisons included speakers from both genders, and the statistics were Bonferroni corrected for three tests. The results indicated that the distinctiveness ratings for the same speakers in the voice-only condition were higher ($Mdn = 14.94$, $SD = 6.43$) than those for the speakers with faces only

Male speakers

Fig. 1 The first three dimensions of four-dimensional solutions for male and female speakers in each condition (Voices, Faces, and Faces-voices). The dimension units are arbitrary. Distinctiveness is reflected by Euclidean distance from the origin

($Mdn = 10.67$, $SD = 3.74$; $W = 224.0$, $p = .028$, $r_s = .75$) as well as faces and voices together ($Mdn = 8.79$, $SD = 2.15$; $W = 292$, $p < .001$, $r_s = .97$). Likewise, the ratings for faces were higher than those for faces and voices ($W = 130.5$, $p = .011$, $r_s = .17$). Insofar as these results represent uncertainty for whether the two individuals presented were of the same identity, these findings parallel those from different speaker comparisons in suggesting that, although faces are more informative and more discriminable than voices, voices add something substantial to faces.

Dimension correlates

As an exploratory analysis, we next examined which auditory and visual features correlated with the dimensions in face, voice, and combined face-voice spaces. This was completed separately for male and female MDS solutions. Tables 1, 2, and 3 indicate the features that significantly correlated with one of the four dimensions in each space. We did not correct for multiple comparisons (i.e., we kept $\alpha = .05$) on these tests because doing so would have been too conservative given the low power of these tests (e.g., Baumann & Belin, 2010). Given the high potential for false positives with our multiple

comparisons, we concentrate on correlations that either replicated dimension features in the literature or replicated across male and female speakers in our experiment:

- Age correlated with various dimensions in face space and voice space for both males and females (though for male faces the correlation was only nearly significant) but did not correlate with any dimensions in face-voice space. This is consistent with participants' ability to extract similar age information using separate modality cues present in faces and voices (e.g., Smith et al., 2016a), but it is unclear why age was not implicated for males or females in the combined face-voice space.
- Mean formants (and their relative spacing) relate closely to vocal tract length and appear to provide important cues in judging the sex, size, or age of speakers (e.g., Smith & Patterson, 2005) as well as similarity judgments (e.g., Murry & Singh, 1980). In voice space, F4 was the only feature consistently implicated across both genders. In face-voice space, most formants including formant dispersion (F1-F5) were implicated. Interestingly, Baumann and Belin (2010) have noted that higher formants such as F4 are less dependent on vowels and could thus be a useful feature for discrimination although the energy may

Fig. 2 The correlations among distinctiveness scores for each speaker (i.e., the normalized distances from the origin in each perceptual space) in the Voice (A_{Only}), Face (V_{Only}), and Face-Voice (AV) spaces as well as the Interaction of Face and Voice distinctiveness (reflecting a

multiplication of the normalized A_{Only} and V_{Only} distinctiveness scores). Pearson's correlations (r) and the slope estimate (β) of the linear equation are included in the lower right of each plot

be too low for differentiation in female speakers. Partly consistent with this notion, our results suggest that, at least with varying sentences as stimuli, higher formants and their relative spacing may play an important role in differentiating speakers.

- Mean f_0 , which has consistently been implicated as an important feature in voice space (e.g., Baumann & Belin, 2010), was found in female voice space but not in male voice space. Mean f_0 was a significant correlate in both face-voice spaces.
- Spectral centroid is closely related to the perceived brightness/darkness of a sound and is frequently found alongside differences in f_0 in hearing research (e.g., Oxenham, 2018). Mean and minimum spectral centroid correlated with dimensions in voice space (note that the mean centroid correlation was nearly significant for male voices). However, standard deviation and maximum spectral centroid correlated with dimensions in face-voice space (note that standard deviation of the centroid correlation was nearly significant in males). Interestingly, Murry and Singh (1980) implicated a related timbre cue – that of perceived nasality – as an important dimensional

correlate in four-dimensional voice space for males and females. Together, these findings suggest that aspects of timbre are also important in differentiating people and may be captured by spectral centroid parameters.

- Various features pertaining to loudness were also implicated, the most consistent being loudness standard deviation (found in both face-voice spaces, but only for females in voice space). This finding may not be surprising given that we used longer speech stimuli (i.e., sentences) than are typically used (e.g., Baumann & Belin, 2010), which may have allowed for more variations in relative intensity.
- As has been suggested in the face literature (e.g., Wallraven, 2014), we found multiple significant features pertaining to facial width in the face and face-voice spaces, the most consistent features being the width of the face at the eyes and the forehead (these replicated across both face and face-voice spaces except that forehead width was not significant in male face-voice space).
- Finally, the amount of mouth movement appeared to be an important feature for males and females in face space but was found only for males in the face-voice space.

Table 1 Auditory feature correlates of the four dimensions from male and female voice spaces

	Male				Female			
	D1	D2	D3	D4	D1	D2	D3	D4
Age				-.0647*			-.586*	
F0								-.609*
F0 max								-.637*
F1 bandwidth					.740**			
F2	-.701*							
F2 bandwidth							-.612*	
F3					-.790**			
F4		-.576*					-.776**	
F5							-.629*	
F1–F5 Dispersion							-.624*	
F4–F5 Dispersion						-.577*		
Duration						-.649*		
Unvoiced		-.873**						
Jitter				-.586*				.589*
Loudness mean		.712**						
Loudness SD							.592*	
Loudness min			-.581*					
Loudness max							.762**	
Centroid mean			.524 ⁺				.584*	
Centroid min	.729**				.613*			

* Significant at the .05 level (two-tailed)

** Significant at the .01 level (two-tailed)

⁺ Approached significance $p < .10$ (two-tailed, only included if significant for the other gender)

Discussion

MDS techniques have been used to derive separate perceptual spaces for faces and voices. We utilized MDS to examine how face and voice spaces would compare to one another as well as to their combined face-voice space. We concentrate here on three main findings that were supported by each of our MDS, regression, and same-speaker rating measures (in both genders).

First, only minimal overlap was found when separately considering face and voice distinctiveness. Second, faces generally dominated person distinctiveness in that face-voice space more closely resembled face space, face distinctiveness was more predictive of face-voice distinctiveness, and the ratings for the same individuals with faces only were lower than with voices only (indicating higher reliability if not higher confidence). Third, and despite the two findings above, voice and face distinctiveness still interacted such that the resulting person distinctiveness was “different from the sum of its constituent parts” (von Ehrenfels, as quoted in Fuchs & Milar, 2003, p. 17). Indeed, people became more distinctive from each other (i.e., further separation from one another in perceptual space) when information was available from both

their faces and voices, a multisensory integration rule that is best described as multiplicative (see Fujisaki et al., 2014). Interestingly, this outcome was paralleled by ratings for the same individuals in which responses improved (i.e., were lower) for same-speaker face-voice comparisons relative to the same-speaker comparisons with faces only. We discuss these three findings in more detail followed by potential extensions of this work.

With our stimuli, we found only a small amount of redundancy between faces and voices in terms of their distinctiveness. Although there was less overlap than anticipated, any overlap may still accord with the general notion of faces and voices as providing complementary information (i.e., providing both redundant and unique information) about people (e.g., Campanella & Belin, 2007). Of course, it would not be surprising to find more redundancy with familiar individuals. Likewise, our limiting speaker comparisons to individuals of the same gender and race as well as a small age range probably also contributed to the low redundancy. Interestingly, studies with similar constraints find people to be above chance (though far from perfect) at matching unfamiliar people’s faces and voices (e.g., Kamachi, Hill, Lander, & Vatikiotis-Bateson, 2003). Most often, dynamic visual speech

Table 2 Visual feature correlates of the four dimensions from male and female face spaces

	Male				Female			
	D1	D2	D3	D4	D1	D2	D3	D4
Age		.559 ⁺						-.576 [*]
Eye height					.711 ^{**}			
Width of face at eyes	.589 [*]				.557 [*]			
Width of face at nose	.842 ^{**}							
Mouth width					.646 [*]			
Chin height					.551 [*]			
Upper cheek edge		-.606 [*]						
Eye to nose		-.649 [*]						
Outer eye to eyebrow		-.572 [*]			.684 [*]			
Eyebrow width					.563 [*]			
Inner eyebrow width					.553 [*]			
Outer eyebrow width					.569 [*]			
Right eye to nose		-.570 [*]						
Inner eye to mouth		-.620 [*]						
Outer eye to mouth		-.640 [*]						
Forehead width		-.581 [*]			-.587 [*]			
Forehead area					-.565 [*]			
Bottom nose						.577 [*]		
Left nose edge						.694 [*]		
Right nose edge						.615 [*]		
Outer right eye						.516 ⁺		.565 [*]
Right eye edge						.591 [*]		
Right pupil								.716 ^{**}
Bottom right eye								.680 [*]
Top right eye								.673 [*]
Bottom right brow								.641 [*]
Top right brow								.580 [*]
Outer right brow								.578 [*]
Inner right brow						.557 ⁺		
Outer left eye						.579 [*]		
Inner left eye								.586 [*]
Bottom left eye								.621 [*]
Top left eye								.523 ⁺
Bottom left brow								.637 [*]
Top left brow								.614 [*]
Inner left brow								.588 [*]
Outer left brow								.531 ⁺
Left nostril								.589 [*]
Right nostril								.500 ⁺
Left mouth								.714 ^{**}
Bottom upper lip								.713 ^{**}
Top bottom lip								.731 ^{**}
Right mouth								.704 [*]
Right mouth edge								.561 ⁺
Left mouth edge			-.574 [*]					.522 ⁺
Top upper lip								.722 ^{**}
Bottom lower lip			-.501 ⁺					.747 ^{**}

* Significant at the .05 level (two-tailed)

** Significant at the .01 level (two-tailed)

+ Approached significance $p < .10$ (two-tailed, only included if significant for one gender, or in the case of motion features, if significant for a symmetrical portion of the face)

information is required for this ability to be reliably above chance (Lachs & Pisoni, 2004), although it seems plausible that, in some cases, people may also be above chance at matching voices to static faces (e.g., Mavica & Barenholtz, 2013; Smith, Dunn, Baguley, & Stacey, 2016b). Our results suggest that the degree of redundancy between face and voice

cues may be similarly low in the distinctiveness of unfamiliar people, even if it is not the same set of cues (or weightings) used to judge sameness.

Our findings also point to visual over auditory dominance in person distinctiveness. As noted in the introduction, many comparisons of face and voice findings would predict this

Table 3 Auditory and visual feature correlates of the four dimensions from male and female face-voice spaces

	Male				Female			
	D1	D2	D3	D4	D1	D2	D3	D4
F0				.653*			-.586*	
F0 SD								-.609*
F1	-.639*							-.637*
F2	-.647*				.740**			
F3		-.681*					-.611*	
F3 Bandwidth					-.790**			
F4							-.776**	
F5			.592*				-.629*	
F1–F5 Dispersion			.596*				-.624*	
F4–F5 Dispersion						-.577*		
Duration						-.649*		
Shimmer	.692*							
Harmonic-to-noise				.700*				
Loudness mean	-.577*							
Loudness SD		-.687*					.592*	
Loudness min		.730**						
Loudness max							.762**	
Centroid SD		-.562 ⁺					-.580*	
Centroid max		-.603*					-.574*	
Eye height	-.653*							
Width of face at eyes		.631*				.551 ⁺		
Mouth width						.583*		
Lower lip height		-.630*						
Eye to nose		-.600*						
Outer eye to eyebrow						.777**		
Outer eyebrow width						.580*		
Forehead width						-.611*		
Left mouth edge	-.637*							
Bottom lower lip	-.609*							

* Significant at the .05 level (two-tailed)

** Significant at the .01 level (two-tailed)

⁺ Approached significance $p < .10$ (two-tailed, only included if significant for one gender, or in the case of motion features, if significant for a symmetrical portion of the face)

outcome, but there are instances in multisensory processing generally (as well as with faces and voices) in which audition is the more influential sensory modality (e.g., Rezlescu et al., 2015). Presently, face-only distinctiveness was about twice as predictive of face-voice distinctiveness as voice-only distinctiveness. As with our identifying “the degree” of overlap between face- and voice-only distinctiveness, this finding should be viewed as one observed value for a naturally varying parameter rather than some fixed quantification. Also, there is reason to suspect that our stimuli might have prompted more weight being given to faces than would naturally be the case. Specifically, we had participants compare speakers across different speech stimuli to better capture the natural variability in

speech (e.g., Lavan et al., 2019) so as to prompt the utilization of cues more specific to speaker identity. Although visual speech variation (along with other motion cues) was reflected in the dynamic visual stimuli we employed, we certainly did not match the variation in viewpoints (all videos depicted frontal views) and luminance with which we typically encounter faces (Burton, Schweinberger, Jenkins, & Kaufmann, 2015). Also, emotional content may represent important dynamic cues to consider, especially because faces and voices that match in terms of their emotional content can improve identification speed and accuracy (e.g., Collignon et al., 2008). These limitations notwithstanding, a finding of visual dominance is important for considerations of how

distinctiveness might impact person perception and recognition in general.

Finally, we found that face-voice distinctiveness was best predicted by combining individual face distinctiveness and voice distinctiveness in a multiplicative fashion. This parallels Fujisaki et al. (2014), who found a multiplicative operation to best predict likeliness ratings for various material categories (e.g., wood). Our finding of greater multisensory distinctiveness may seem surprising given the low-overlap (i.e., non-redundancy) between the separate face and voice cues as well as our finding of visual-dominance. Potentially, having both face and voice cues might have resulted in more ways to use cues, less consensus among raters, and thus less discriminability among speakers at the group level. Or, participants might have relied almost exclusively on vision. Indeed, in contrast to their finding of multiplicative integration for material categories, Fujisaki et al. found weighted-average models (i.e., without the audiovisual interaction term) to best describe audiovisual ratings for material properties (e.g., hardness). With these ratings, the dominant modality (e.g., for hardness, audition) generally predicted audiovisual ratings best and was not improved by considering the other modality separately or multiplicatively. To see a multiplicative integration rule best describe person distinctiveness may be surprising, then, given the non-categorical nature of person distinctiveness. That is, our comparisons were made within the single category of people, and we took precautions to prevent comparisons based on some of the major salient categories we use to differentiate people (e.g., gender). However, considering one hypothesis from Fujisaki et al. affords a consistent interpretation. They propose that their participants' category ratings may have been based on aggregate considerations of the material properties. At least consistent with this idea was their finding that – when encountering one item visually and another auditorily resulted in the perception of a third material category – the audiovisual property ratings were consistent with those of the third material category. Under this hypothesis, a multiplicative integration rule would seem plausible for our stimuli given that several different perceptual properties from faces and voices likely contribute to person distinctiveness. Presently, we show that features present in people's faces and voices combine in a fashion that enables us to better differentiate them.

What those specific features are will remain an important question for future research. Based on our exploratory, correlational analyses, some important features in face-voice space appeared to include facial width, fundamental frequency, various formants, spectral centroid, loudness variation, and (for males) lip movement. Others have suggested the importance of facial width (e.g., Wallraven, 2014) to face space as well as fundamental frequency and various formants (e.g., Baumann & Belin, 2010) to voice space. We add that spectral centroid (an aspect of timbre relating to the darkness/brightness of a sound) and lip movement may also be important to person

distinctiveness (at least with stimuli involving more speech variation). Still, all of these (or related) features present in face-voice space were also implicated in the constituent face and voice spaces. Perhaps the only striking difference was that age, which was implicated in the separate face and voice spaces, was not implicated in face-voice space. This was surprising, particularly because face and voice cues can be integrated to provide better age estimates (e.g., Moyse, 2014).

How is it that face-voice space could result in greater person distinctiveness even while some features are maintained and others seem to disappear? One possibility is that we did not take into account enough dimensions in the face-voice space. We considered only a four-dimensional solution to enable valid comparisons of distinctiveness across modalities, and because the solution was optimal according to the observed STRESS value. Another possibility is that the face and voice cues themselves may have been used interactively. Of course, our limited correlational analyses ignore these more complicated relationships. One interesting and still open question is whether these multisensory interactions are dependent on statistical correspondences between faces and voices (e.g., Spence, 2011). For instance, Ernst (2007) observed a cost in discriminability when participants' learned associations between luminance (visual) and stiffness (haptic) were violated. More recently, Brang (2019) showed that, in matching one voice to one of two concurrently presented dynamic faces, identity information could override spatiotemporal information. Importantly, the speakers could be matched over spatiotemporal cues using identity information even with the same gender, implicating a likely role for pitch and timbre being matched with various face features. To see if person distinctiveness (or some other attribute) might similarly rely on multisensory interactions, it would be important to systematically vary features thought to have some natural correspondence. One intriguing possibility is that when the goal is to differentiate people, violations in expected relationships might prompt more robust differences in person distinctiveness than confirming their agreement (e.g., the perceptual equivalent of such remarks as “I would not expect such a low voice from her”/ “He looks much younger than he sounds”). Indeed, Smith et al. (2016a) have implicated gender, age, health, height, and weight as redundant cues that people can extract from faces or voices. Thus, when the aim of perception is to highlight differences, perhaps these correspondences are unremarkable to the observer. Such understandings may be particularly important for considering the initial memorability of people more generally (e.g., through their faces and voices). Indeed, Bülthoff and Newell (2015) showed that faces encoded with distinct voices were better recognized at test, indicating that distinctiveness need not be task-relevant to affect memory performance. More basically, it is shown that face and voice cues jointly influence person distinctiveness, a consideration that may be useful in the memory sphere.

Another relevant area to consider along with the present findings might include ensemble coding, namely, when perception relies on average representations of a feature(s). For instance, Haberman and Whitney (2009) have shown that the average emotion of faces is readily extracted from a set of faces, even while information pertaining to the emotion of individuals is lost. Thus, these average representations may involve a complex and holistic representation of multiple features. Similarly, in the auditory modality, Piazza, Sween, Wessel, Silver, and Whitney (2013) have demonstrated that participants represented the average tone in a sequence without remembering specific tones or their position in the sequence. Of course, the present study does not assume that faces and voices are necessarily related to average faces and voices, merely that they are centered around an approximately average face or voice through comparisons. Also, comparing two individuals at a time might emphasize featural differences more than if they were rated for distinctiveness in isolation, which might be expected to show a greater influence of ensemble coding. Still, MDS or related techniques could prove useful in assessing how close an average face and/or voice (created by morphing techniques) is to the center of perceptual space and whether the average face/voice becomes more central as additional faces and voices are encountered (as would be predicted with ensemble coding). Interestingly, in a three-dimensional acoustic space, Latinus et al. (2013) found that the average voice (achieved by morphing) was not located near the center. Thus, compared to spaces derived from the stimulus features, it would be possible to see if it is the participants themselves who are effectively centering the faces and voices they encounter about the origin, which might expand our knowledge of how people use summary statistics.

We also cannot rule out the potential for memory to have impacted the results. One reason we favored a design in which participants rated the dissimilarity of pairs of speakers rather than speakers in isolation was to emphasize comparisons of stimulus features. We suspect that this task involved rapid perceptual judgments given that information from the first speaker in the pair had to be maintained in working memory while the second member was presented. Presumably, this higher cognitive load combined with the emphasis on perceptual characteristics over more meaning-based encoding techniques (with the assumption that perceptual characteristics represent shallower memory encoding strategies) would decrease recognition of faces and voices (Craik & Lockhart, 1972). Also, it may be important to consider the relatively large number of pairs that were compared, which could also decrease recognition. We therefore do not believe memory played a substantial role in our findings.

We also note that stimulus presentation duration may be an important consideration. Stimulus duration matters in other domains, such as ensemble coding (e.g., Haberman & Whitney, 2009). Our stimulus durations depended on the length of the speech segment (which varied from 1 to 2.5 s). Prior MDS studies on voice have often used similar, variable durations (e.g., Singh & Murry, 1978), whereas durations for faces have often been fixed at 1 s (e.g., Johnston, 1997). Since faces and likely voices are thought to be rapidly processed and holistically represented, perhaps shorter stimulus durations (≤ 500 ms) would be interesting to consider. In ensemble coding, decreasing duration in this range prompts more variable representations of an average face (e.g., Haberman & Whitney, 2009). Potentially, decreasing stimulus duration in an MDS study might have a similar effect, although without including average faces and voices it would be unclear whether ensemble coding is the mechanism as opposed to less time to process and compare stimulus features. Of course, one advantage with the longer, dynamic stimuli we employed in this study was the variability in speech and face movement that was introduced to the stimuli – both of which appeared to represent important dimension correlates.

Lastly, to our knowledge, MDS techniques have only been used to compare spaces across sensory modalities in instances where a high degree of overlap was predicted and found. This includes various studies comparing visual, haptic, and, in two cases, visuo-haptic spaces for the same objects (e.g., Cooke, Jäkel, Wallraven, & Bühlhoff, 2007; Erdogan, Yildirim, & Jacobs, 2015; Gaißert, Wallraven, & Bühlhoff, 2010). Insofar as our study was the first to combine MDS and regression techniques to compare very different spaces across modalities and to multisensory spaces, we suggest that this framework may be useful in other domains. For instance, one might examine whether various auditory, visual, and tactile properties contribute to how well an object fits a material category (as hypothesized by Fujisaki et al., 2014). Alternatively, one may use this technique to explore instances in which one population is expected to show greater redundancy across the senses than others, such as looking at how vision and audition influence music perception in expert and non-expert populations or correspondences between sensory stimuli and synesthetic experiences in synesthetes and non-synesthetes. Finally, although MDS techniques have been used in psychological research for over half a century (Shepard, 1962), it is likely that much can still be learned using this technique, especially in the domain of multisensory integration.

Open Practices Statement The data and code for all experiments are available upon request to the first author. For the stimulus materials, please refer to Cao et al. (2014) to access their open-source database, CREMA-D. This experiment was not preregistered.

Appendix 1

Table 4 List of auditory and visual features used in the exploratory correlational analyses

Audiovisual correlates

-Age

Auditory correlates

-Mean F0

-F0 standard deviation

-F0 minimum

-F0 maximum

-F1

-F1 bandwidth

-F2

-F2 bandwidth

-F3

-F3 bandwidth

-F4

-F4 Bandwidth

-F5

-F5 Bandwidth

-Dispersion between F1 and F5

-Dispersion between F4 and F5

-Duration

-Duration of voiced speech

-Standard deviation of duration of voiced speech

-Fraction of unvoiced frames

-Jitter

-Shimmer

-Mean harmonics-to-noise ratio

-Mean loudness

-Loudness standard deviation

-Loudness minimum

-Loudness maximum

-Spectral centroid mean

-Spectral centroid standard deviation

-Spectral centroid minimum

-Spectral centroid maximum

-Mel-frequency cepstrum coefficient mean

-Mel-frequency cepstrum coefficient standard deviation

-Mel-frequency cepstrum coefficient minimum

-Mel-frequency cepstrum coefficient maximum

Visual correlates (spatial)

-Mean width of eyes

-Mean height of eyes

-Width of face at eyes

-Width of face at nose

-Width of face at mouth

-Width of chin

Table 4 (continued)

-Mean lip edge to outer chin distance

-Width of the mouth

-Height of the upper lip

-Height of the lower lip

-Height of chin at the center of the mouth

-Distance from top of upper lip to the bottom of the nose

-Nose width at outer nostrils

-Mean distance along the outer edge of the upper half of the cheeks

-Mean distance along the outer edge of the lower half of the cheeks

-Mean distance along the outer edge of the jaw

-Mean distance from inner corner of the eye to edge of the nostril

-Distance between inner ends of eyebrows

-Mean distance from the outer corner of the eyebrow to the outer corner of the eye

-Mean width of the eyebrows

-Mean distance from inner corner of the eyebrow to the inner corner of the eye

-Mean height of the eyebrows at the midpoint of the eye

-Mean width of inner eyebrows

-Mean width of outer eyebrows

-Distance from the inner corner of the right eye to the bottom of the nose-

Distance from the inner corner of left eye to bottom of the nose

-Mean distance from the inner corner of eye to the outer corner of mouth

-Mean distance from the inner outer of eye to the outer corner of mouth

-Forehead height

-Forehead width

-Forehead area

Visual correlates (movement)

-Left corner, right eye

-Right corner, right eye

-Right hairline (mid-eye level)

-Right pupil

-Bottom right eye

-Top right eye

-Bottom mid-right eyebrow

-Top mid-right eyebrow

-Outer right eyebrow

-Inner right eyebrow

-Left corner, left eye

-Right corner, left eye

-Left hairline (mid-eye level)

-Left pupil

-Bottom left eye

-Top left eye

-Bottom mid-left eyebrow

-Top mid-left eyebrow

-Inner left eyebrow

-Outer left eyebrow

-Left nostril

-Right nostril

-Bottom tip of the nose

-Left edge of the face (level with the bottom of the nose)

-Right edge of the face (level with the bottom of the nose)

-Left corner of the mouth

-Center top lip

-Center bottom lip

-Right corner of the mouth

-Right edge of the face (at mid-lip level)

-Left edge of the face (at mid-lip level)

-Mid-upper lip

-Mid-bottom lip

Appendix 2

Average dissimilarity matrices as a function of condition and gender

Table 5 The average dissimilarity matrix (and standard deviations) of 24 participants for the 12 male speakers in the audio-only condition

Speaker	1	2	3	4	5	6	7	8	9	10	11	12
1	15.21 (23.21)	---	---	---	---	---	---	---	---	---	---	---
2	44.87 (25.33)	5.71 (9.59)	---	---	---	---	---	---	---	---	---	---
3	44.29 (22.69)	52.88 (21.83)	29.13 (23.74)	---	---	---	---	---	---	---	---	---
4	55.21 (24.1)	46.58 (26.16)	58.38 (21.05)	11.74 (16.28)	---	---	---	---	---	---	---	---
5	55.63 (27.62)	69.46 (20.82)	57.88 (25.44)	65.67 (22.54)	9.65 (15.84)	---	---	---	---	---	---	---
6	45.17 (30.74)	64.96 (20.72)	35.88 (25.82)	53.48 (23.82)	49.17 (26.66)	16.25 (19.86)	---	---	---	---	---	---
7	64.83 (19.98)	74.88 (16.14)	71.17 (16.00)	61.42 (24.68)	66.13 (21.32)	60.17 (24.3)	7.64 (12.58)	---	---	---	---	---
8	50.48 (25.89)	67.42 (22.27)	55.83 (22.65)	45.91 (18.22)	55.38 (21.82)	55.25 (25.38)	56.54 (23.02)	16.46 (18.92)	---	---	---	---
9	27.29 (22.94)	64.58 (25.03)	29.83 (25.00)	53.25 (23.62)	53.04 (22.54)	41.63 (25.98)	69.54 (19.57)	48.79 (25.91)	16.3 (17.29)	---	---	---
10	49.58 (24.35)	69.7 (21.25)	53.92 (30.28)	53.21 (24.53)	50.33 (19.99)	42.74 (24.97)	54.78 (26.07)	47.38 (30.11)	48.96 (29.11)	12.57 (19.95)	---	---
11	46.83 (21.84)	66.29 (23.02)	56.92 (25.43)	50.38 (24.96)	27.25 (24.05)	53.00 (23.69)	62.17 (21.83)	50.08 (25.01)	55.46 (24.53)	43.29 (26.95)	9.43 (20.14)	---
12	62.17 (22.05)	70.54 (17.89)	49.67 (24.76)	58.54 (22.44)	56.04 (23.54)	55.46 (22.93)	61.39 (23.87)	47.50 (25.84)	58.63 (22.55)	56.08 (27.77)	66.35 (18.97)	13.04 (22.88)

Table 6 The average dissimilarity matrix (and standard deviations) of 24 participants for the 12 female speakers in the audio-only condition

Speaker	1	2	3	4	5	6	7	8	9	10	11	12
1	12.00 (16.22)	---	---	---	---	---	---	---	---	---	---	---
2	55.46 (24.54)	36.25 (27.39)	---	---	---	---	---	---	---	---	---	---
3	64.91 (22.42)	56.42 (27.24)	13.13 (13.44)	---	---	---	---	---	---	---	---	---
4	50.54 (28.32)	52.5 (26.72)	58.88 (18.5)	13.21 (18.59)	---	---	---	---	---	---	---	---
5	49.71 (27.92)	50.83 (27.02)	51.35 (25.3)	46.29 (25.55)	19.42 (21.17)	---	---	---	---	---	---	---
6	70.57 (19.27)	50.42 (29.49)	48.42 (30.96)	51.38 (22.4)	47.63 (26.69)	10.08 (15.31)	---	---	---	---	---	---
7	52.91 (27.2)	44.58 (20.61)	58.46 (26.37)	59.75 (19.75)	41.54 (25.34)	61.17 (23.56)	16.04 (16.88)	---	---	---	---	---
8	43.39 (22.94)	60.71 (20.85)	69.67 (22.97)	60.21 (23.44)	27.29 (23.86)	52.78 (22.15)	48.29 (26.94)	17.5 (25.45)	---	---	---	---
9	64.13 (25.42)	57.21 (28.24)	60.67 (22.9)	23.25 (24.5)	56.96 (28.5)	55.96 (26.1)	58.13 (23.46)	59.22 (19.38)	14.67 (19.44)	---	---	---
10	40.83 (27.61)	45.04 (25.16)	60.08 (20.65)	52.25 (24.28)	47.00 (26.04)	61.71 (24.26)	37.13 (30.85)	29.96 (23.19)	42.43 (19.21)	16.29 (20.96)	---	---
11	50.00 (28.51)	57.38 (26.89)	66.67 (24.86)	56.79 (23.57)	57.92 (26.79)	58.92 (29.66)	65.95 (20.06)	59.71 (22.55)	63.42 (20.94)	66.79 (20.37)	16.83 (24.25)	---
12	60.96 (22.42)	55.08 (28.09)	44.42 (24.86)	55.46 (27.6)	51.22 (24.49)	60.96 (23.56)	46.79 (23.33)	57.96 (24.87)	64.21 (25.06)	52.33 (22.85)	55.38 (29.45)	16.33 (24.00)

Table 7 The average dissimilarity matrix (and standard deviations) of 24 participants for the 12 male speakers in the visual-only condition

Speaker	1	2	3	4	5	6	7	8	9	10	11	12
1	10.91 (23.56)	---	---	---	---	---	---	---	---	---	---	---
2	64.96 (26.75)	8.21 (19.37)	---	---	---	---	---	---	---	---	---	---
3	67.33 (20.29)	59.13 (23.58)	18.67 (30.6)	---	---	---	---	---	---	---	---	---
4	59.00 (23.91)	55.46 (25.44)	70.17 (21.23)	15.5 (25.73)	---	---	---	---	---	---	---	---
5	67.29 (22.05)	68.08 (23.03)	73.33 (20.44)	70.42 (20.75)	9.91 (20.66)	---	---	---	---	---	---	---
6	56.96 (24.83)	58.46 (23.41)	63.54 (20.84)	42.25 (29.28)	70.61 (23.72)	14.00 (26.25)	---	---	---	---	---	---
7	67.21 (23.06)	67.71 (20.23)	60.25 (25.22)	72.3 (18.28)	69.00 (23.03)	64.92 (22.94)	16.83 (27.3)	---	---	---	---	---
8	66.88 (21.45)	67.88 (22.81)	70.08 (18.17)	59.67 (26.89)	66.87 (24.36)	61.92 (23.95)	69.17 (21.53)	16.54 (26.34)	---	---	---	---
9	73.83 (22.36)	69.95 (26.67)	69.42 (20.99)	74.00 (24.52)	67.00 (24.54)	64.79 (22.88)	66.17 (23.52)	65.91 (24.24)	9.46 (22.93)	---	---	---
10	64.57 (25.4)	69.13 (22.34)	67.13 (22.65)	68.92 (23.29)	68.83 (23.86)	64.17 (24.15)	63.96 (22.22)	72.00 (18.00)	62.29 (25.06)	9.29 (19.75)	---	---
11	71.04 (14.54)	61.63 (21.08)	60.92 (22.99)	73.88 (17.73)	74.63 (16.19)	65.75 (23.94)	67.29 (19.92)	69.83 (16.51)	68.38 (19.42)	59.3 (24.11)	9.25 (16.34)	---
12	58.38 (21.4)	64.88 (24.57)	67.71 (22.6)	68.88 (21.22)	64.33 (22.22)	66.13 (24.03)	56.63 (25.67)	67.75 (20.3)	65.21 (21.69)	60.54 (25.62)	62.58 (20.08)	10.21 (22.73)

Table 8 The average dissimilarity matrix (and standard deviations) of 24 participants for the 12 female speakers in the visual-only condition

Speaker	1	2	3	4	5	6	7	8	9	10	11	12
1	10.17 (23.09)	---	---	---	---	---	---	---	---	---	---	---
2	66.38 (22.62)	20.54 (26.7)	---	---	---	---	---	---	---	---	---	---
3	77.83 (17.63)	56.88 (23.96)	13.88 (25.10)	---	---	---	---	---	---	---	---	---
4	71.29 (23.02)	58.88 (24.15)	59.83 (19.50)	12.96 (24.29)	---	---	---	---	---	---	---	---
5	75.25 (20.64)	63.25 (21.67)	66.25 (16.96)	52.04 (24.61)	9.58 (19.01)	---	---	---	---	---	---	---
6	75.25 (19.89)	62.17 (20.67)	59.00 (25.28)	64.42 (21.69)	64.75 (24.83)	8.75 (20.59)	---	---	---	---	---	---
7	63.63 (21.24)	54.58 (21.88)	75.33 (14.91)	72.04 (16.15)	70.88 (17.83)	61.08 (24.63)	14.67 (25.39)	---	---	---	---	---
8	65.79 (19.75)	62.74 (21.4)	70.58 (23.65)	68.96 (21.73)	72.46 (16.34)	67.79 (19.09)	56.50 (26.22)	17.5 (28.94)	---	---	---	---
9	74.46 (16.46)	54.38 (24.79)	60.13 (24.96)	64.79 (21.66)	69.92 (15.31)	63.21 (22.42)	50.25 (23.41)	64.17 (24.83)	8.78 (19.23)	---	---	---
10	62.00 (22.01)	57.74 (21.69)	74.17 (18.82)	66.63 (23.46)	74.30 (12.10)	67.83 (13.91)	51.29 (22.44)	50.00 (29.31)	54.50 (22.87)	11.13 (20.82)	---	---
11	68.50 (21.77)	52.87 (23.67)	71.46 (18.57)	63.63 (21.83)	65.92 (24.28)	69.42 (19.37)	54.71 (22.06)	66.21 (19.42)	53.38 (26.07)	52.26 (25.15)	10.43 (21.4)	---
12	66.83 (26.3)	62.33 (21.16)	61.83 (25.54)	48.58 (23.22)	61.79 (23.3)	56.79 (24.86)	68.50 (22.72)	72.00 (20.66)	66.50 (21.86)	70.17 (22.16)	65.83 (20.49)	6.88 (14.0-0)

Table 9 The average dissimilarity matrix (and standard deviations) of 24 participants for the 12 male speakers in the audiovisual condition

Speaker	1	2	3	4	5	6	7	8	9	10	11	12
1	8.50 (16.82)	---	---	---	---	---	---	---	---	---	---	---
2	70.08 (20.65)	7.83 (20.13)	---	---	---	---	---	---	---	---	---	---
3	59.71 (25.83)	57.63 (24.69)	8.42 (18.74)	---	---	---	---	---	---	---	---	---
4	54.58 (24.92)	52.96 (25.64)	72.58 (21.76)	10.54 (19.21)	---	---	---	---	---	---	---	---
5	66.54 (21.61)	73.92 (20.81)	70.67 (22.56)	80.87 (12.93)	7.96 (18.36)	---	---	---	---	---	---	---
6	59.79 (26.5)	65.57 (19.36)	68.38 (23.45)	50.88 (30.76)	76.09 (15.39)	9.57 (23.52)	---	---	---	---	---	---
7	65.17 (22.1)	71.42 (22.12)	67.58 (23.03)	74.42 (16.50)	70.71 (21.24)	69.92 (23.66)	6.78 (17.93)	---	---	---	---	---
8	59.17 (25.96)	72.67 (10.51)	75.79 (16.65)	67.83 (22.41)	71.04 (19.02)	60.42 (22.13)	62.87 (23.04)	8.83 (20.78)	---	---	---	---
9	71.00 (19.27)	68.00 (20.41)	70.50 (17.67)	77.25 (18.53)	68.04 (24.58)	67.38 (21.07)	70.43 (18.08)	59.75 (26.11)	6.33 (14.73)	---	---	---
10	69.46 (23.58)	60.96 (22.29)	66.71 (16.86)	64.7 (21.78)	67.92 (25.53)	67.70 (24.93)	65.79 (26.33)	65.04 (22.25)	66.58 (24.37)	8.92 (21.03)	---	---
11	69.09 (17.41)	71.29 (22.4)	62.71 (22.06)	74.04 (19.39)	70.75 (18.28)	68.04 (21.01)	64.63 (21.17)	75.00 (14.50)	60.04 (24.64)	60.75 (21.11)	9.33 (18.74)	---
12	62.00 (22.76)	67.17 (26.43)	67.38 (23.63)	66.21 (21.06)	69.38 (21.96)	71.13 (19.65)	63.21 (24.54)	68.04 (20.12)	73.92 (17.16)	68.67 (14.59)	64.79 (20.57)	11.21 (22.22)

Table 10 The average dissimilarity matrix (and standard deviations) of 24 participants for the 12 female speakers in the audiovisual condition

Speaker	1	2	3	4	5	6	7	8	9	10	11	12
1	8.71 (21.26)	---	---	---	---	---	---	---	---	---	---	---
2	74.75 (16.76)	11.79 (23.14)	---	---	---	---	---	---	---	---	---	---
3	79.92 (17.74)	63.96 (21.73)	8.75 (18.71)	---	---	---	---	---	---	---	---	---
4	73.78 (21.78)	65.96 (22.29)	68.42 (23.29)	7.04 (14.89)	---	---	---	---	---	---	---	---
5	70.92 (21.4)	62.54 (20.61)	62.29 (23.28)	45.29 (26.27)	13.04 (24.93)	---	---	---	---	---	---	---
6	81.63 (13.61)	55.17 (25.19)	60.54 (23.39)	67.83 (18.62)	65.00 (19.59)	10.13 (20.67)	---	---	---	---	---	---
7	65.21 (26.00)	54.22 (24.29)	71.67 (19.12)	70.25 (21.35)	64.38 (19.04)	68.54 (21.46)	7.13 (13.29)	---	---	---	---	---
8	65.92 (22.23)	71.13 (15.1)	76.75 (18.10)	71.75 (24.47)	75.46 (13.85)	74.48 (19.38)	57.74 (24.6)	4.25 (10.60)	---	---	---	---
9	68.92 (20.53)	58.17 (24.25)	70.04 (21.93)	54.50 (24.95)	71.67 (19.77)	68.08 (20.46)	54.30 (26.08)	62.54 (25.99)	9.18 (20.36)	---	---	---
10	69.75 (16.91)	65.83 (18.89)	77.71 (12.4)	64.71 (22.66)	71.67 (14.09)	70.00 (20.8)	51.71 (22.82)	52.13 (26.42)	53.13 (26.08)	13.04 (24.95)	---	---
11	68.17 (24.71)	60.61 (21.59)	72.22 (20.21)	72.17 (20.57)	71.17 (20.81)	69.88 (23.42)	54.96 (26.59)	66.25 (23.99)	56.25 (23.43)	60.13 (25.23)	6.46 (13.89)	---
12	74.78 (18.01)	62.17 (23.97)	59.67 (25.16)	57.13 (24.05)	57.83 (23.67)	59.09 (26.96)	69.17 (22.82)	78.71 (15.51)	64.33 (22.51)	65.87 (20.04)	72.83 (19.98)	11.38 (20.62)

References

- Baumann, O., & Belin, P. (2010). Perceptual scaling of voice identity: Common dimensions for different vowels and speakers. *Psychological Research PRPF*, 74(1), 110. <https://doi.org/10.1007/s00426-008-0185-z>
- Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences*, 8(3), 129–135. <https://doi.org/10.1016/j.tics.2004.01.008>
- Brang, D. (2019). The Stolen Voice Illusion. *Perception*, 48(8), 649–667. <https://doi.org/10.1177/0301006619858076>
- Bülthoff, I., & Newell, F. N. (2015). Distinctive voices enhance the visual recognition of unfamiliar faces. *Cognition*, 137, 9–21. <https://doi.org/10.1016/j.cognition.2014.12.006>
- Burton, A. M., Schweinberger, S. R., Jenkins, R., & Kaufmann, J. M. (2015). Arguments against a configural processing account of familiar face recognition. *Perspectives on Psychological Science*, 10(4), 482–496. <https://doi.org/10.1177/1745691615583129>
- Byatt, G., & Rhodes, G. (2004). Identification of own-race and other-race faces: Implications for the representation of race in face space. *Psychonomic Bulletin & Review*, 11(4), 735–741. <https://doi.org/10.3758/BF03196628>
- Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences*, 11(12), 535–543. <https://doi.org/10.1016/j.tics.2007.10.001>
- Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. (2014). CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4), 377–390. <https://doi.org/10.1109/TAFEC.2014.2336244>
- Chhabra, S., Badcock, J. C., Maybery, M. T., & Leung, D. (2012). Voice identity discrimination in schizophrenia. *Neuropsychologia*, 50(12), 2730–2735. <https://doi.org/10.1016/j.neuropsychologia.2012.08.006>
- Collignon, O., Girard, S., Gosselin, F., Roy, S., Saint-Amour, D., Lassonde, M., & Lepore, F. (2008). Audio-visual integration of emotion expression. *Brain Research*, 1242, 126–135. <https://doi.org/10.1016/j.brainres.2008.04.023>
- Cooke, T., Jäkel, F., Wallraven, C., & Bülthoff, H. H. (2007). Multimodal similarity and categorization of novel, three-dimensional objects. *Neuropsychologia*, 45(3), 484–495. <https://doi.org/10.1016/j.neuropsychologia.2006.02.009>
- Cox, I. J., Ghosn, J., & Yianilos, P. N. (1996). Feature-based face recognition using mixture-distance. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 96, 209–216.
- Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11(6), 671–684. [https://doi.org/10.1016/S0022-5371\(72\)80001-X](https://doi.org/10.1016/S0022-5371(72)80001-X)
- Damjanovic, L. (2011). The face advantage in recalling episodic information: Implications for modeling human memory. *Consciousness and Cognition*, 20(2), 309–311. <https://doi.org/10.1016/j.concog.2010.06.002>
- De Leeuw, J., & Mair, P. (2011). Multidimensional scaling using majorization: SMACOF in R.
- Erdogan, G., Yildirim, I., & Jacobs, R. A. (2015). From sensory signals to modality-independent conceptual representations: A probabilistic language of thought approach. *PLoS Computational Biology*, 11(11), e1004610. <https://doi.org/10.1371/journal.pcbi.1004610>
- Ernst, M. O. (2007). Learning to integrate arbitrary signals from vision and touch. *Journal of Vision*, 7(5), 1–14. <https://doi.org/10.1167/7.5.7>
- Fuchs, A. H., & Milar, K. S. (2003). Psychology as a science. In D. K. Freedheim & I. B. Weiner (Eds.), *Handbook of psychology*, Vol. 1: *History of psychology* (pp. 1–26). Hoboken, NJ: John Wiley & Sons, Inc.
- Fujisaki, W., Goda, N., Motoyoshi, I., Komatsu, H., & Nishida, S. (2014). Audiovisual integration in the human perception of materials. *Journal of Vision*, 14(4), 1–20. <https://doi.org/10.1167/14.4.12>
- Gaißert, N., Wallraven, C., & Bülthoff, H. H. (2010). Visual and haptic perceptual spaces show high similarity in humans. *Journal of Vision*, 10(11), 1–20. <https://doi.org/10.1167/10.11.2>
- Haberman, J., & Whitney, D. (2009). Seeing the mean: ensemble coding for sets of faces. *Journal of Experimental Psychology: Human Perception and Performance*, 35(3), 718–734. <https://doi.org/10.1037/a0013899>
- Johnston, R. A., Milne, A. B., Williams, C., & Hosie, J. (1997). Do Distinctive Faces Come from Outer Space? An Investigation of the Status of a Multidimensional Face-Space. *Visual Cognition*, 4(1), 59–67. <https://doi.org/10.1080/713756748>
- Kamachi, M., Hill, H., Lander, K., & Vatikiotis-Bateson, E. (2003). Putting the face to the voice: Matching identity across modality. *Current Biology*, 13(19), 1709–1714. <https://doi.org/10.1016/j.cub.2003.09.005>
- Kreiman, J., Gerratt, B. R., Precoda, K., & Berke, G. S. (1992). Individual differences in voice quality perception. *Journal of Speech, Language, and Hearing Research*, 35(3), 512–520. <https://doi.org/10.1044/jshr.3503.512>
- Lachs, L., & Pisoni, D. B. (2004). Crossmodal source identification in speech perception. *Ecological Psychology*, 16(3), 159–187. https://doi.org/10.1207/s15326969eco1603_1
- Latinus, M., McAleer, P., Bestelmeyer, P. E., & Belin, P. (2013). Norm-based coding of voice identity in human auditory cortex. *Current Biology*, 23(12), 1075–1080. <https://doi.org/10.1016/j.cub.2013.04.055>
- Lattin, J., Carroll, J. D., & Green P. E. (2003). *Analyzing multivariate data*. Belmont, CA: Brooks/Cole.
- Lavan, N., Burton, A. M., Scott, S. K., & McGettigan, C. (2019). Flexible voices: Identity perception from variable vocal signals. *Psychonomic Bulletin & Review*, 26(1), 90–102. <https://doi.org/10.3758/s13423-018-1497-7>
- Lee, K., Byatt, G., & Rhodes, G. (2000). Caricature effects, distinctiveness, and identification: Testing the face-space framework. *Psychological Science*, 379–385.
- Mavica, L. W., & Barenholtz, E. (2013). Matching voice and face identity from static images. *Journal of Experimental Psychology: Human Perception and Performance*, 39(2), 307–312. <https://doi.org/10.1037/a0030945>
- Moyse, E. (2014). Age Estimation from Faces and Voices: A Review. *Psychologica Belgica*, 54(3), 255–265. <https://doi.org/10.5334/pb.aq>
- Murry, T., & Singh, S. (1980). Multidimensional analysis of male and female voices. *The Journal of the Acoustical Society of America*, 68(5), 1294–1300. <https://doi.org/10.1121/1.385122>
- Nishimura, M., Maurer, D., & Gao, X. (2009). Exploring children's face-space: A multidimensional scaling analysis of the mental representation of facial identity. *Journal of Experimental Child Psychology*, 103(3), 355–375. <https://doi.org/10.1016/j.jecp.2009.02.005>
- Oxenham, A. J. (2018). How we hear: The perception and neural coding of sound. *Annual Review of Psychology*, 69, 27–50. <https://doi.org/10.1146/annurev-psych-122216-011635>
- Papesh, M. H., & Goldinger, S. D. (2010). A multidimensional scaling analysis of own- and cross-race face spaces. *Cognition*, 116(2), 283–288. <https://doi.org/10.1016/j.cognition.2010.05.001>
- Pedely, L., Levine, S. C., & Shevell, S. K. (1985). Developmental changes in face processing: Results from multidimensional scaling. *Journal of Experimental Child Psychology*, 39(3), 421–436. [https://doi.org/10.1016/0022-0965\(85\)90050-5](https://doi.org/10.1016/0022-0965(85)90050-5)
- Peynircioğlu, Z. F., Brent, W., Tatz, J. R., & Wyatt, J. (2017). McGurk effect in gender identification: Vision trumps audition in voice judgments. *The Journal of General Psychology*, 144(1), 59–68. <https://doi.org/10.1080/00221309.2016.1258388>

- Piazza, E. A., Sweeny, T. D., Wessel, D., Silver, M. A., & Whitney, D. (2013). Humans use summary statistics to perceive auditory sequences. *Psychological Science*, 24(8), 1389–1397. <https://doi.org/10.1177/0956797612473759>
- Potter, T., Corneille, O., Ruys, K. I., & Rhodes, G. (2007). “Just another pretty face”: A multidimensional scaling approach to face attractiveness and variability. *Psychonomic Bulletin & Review*, 14(2), 368–372. <https://doi.org/10.3758/BF03194079>
- Rezlescu, C., Penton, T., Walsh, V., Tsujimura, H., Scott, S. K., & Banissy, M. J. (2015). Dominant Voices and Attractive Faces: The Contribution of Visual and Auditory Information to Integrated Person Impressions. *Journal of Nonverbal Behavior*, 39(4), 355–370. <https://doi.org/10.1007/s10919-015-0214-8>
- Shepard, R. N. (1962). The analysis of proximities: multidimensional scaling with an unknown distance function. *Psychometrika*, 27(2), 125–140. <https://doi.org/10.1007/BF02289630>
- Shepherd, J. W., & Deregowski, J. B. (1981). Races and faces—A comparison of the responses of Africans and Europeans to faces of the same and different races. *British Journal of Social Psychology*, 20(2), 125–133. <https://doi.org/10.1111/j.2044-8309.1981.tb00485.x>
- Singh, S., & Murry, T. (1978). Multidimensional classification of normal voice qualities. *The Journal of the Acoustical Society of America*, 64(1), 81–87. <https://doi.org/10.1121/1.381958>
- Smith, D. R., & Patterson, R. D. (2005). The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. *The Journal of the Acoustical Society of America*, 118(5), 3177–3186. doi:<https://doi.org/10.1121/1.2047107>
- Smith, H. M., Dunn, A. K., Baguley, T., & Stacey, P. C. (2016a). Concordant cues in faces and voices: Testing the backup signal hypothesis. *Evolutionary Psychology*, 14(1), 1474704916630317.
- Smith, H. M., Dunn, A. K., Baguley, T., & Stacey, P. C. (2016b). Matching novel face and voice identity using static and dynamic facial images. *Attention, Perception, & Psychophysics*, 78(3), 868–879. <https://doi.org/10.3758/s13414-015-1045-8>
- Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics*, 73(4), 971–995. <https://doi.org/10.3758/s13414-010-0073-7>
- Tomasi, C., & Kanade, T. (1991). Detection and tracking of point features. Pittsburgh, PA: School of Computer Science, Carnegie Mellon University.
- Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology Section A*, 43(2), 161–204. <https://doi.org/10.1080/14640749108400966>
- Valentine, T., & Endo, M. (1992). Towards an exemplar model of face processing: The effects of race and distinctiveness. *The Quarterly Journal of Experimental Psychology Section A*, 44(4), 671–703. <https://doi.org/10.1080/14640749208401305>
- Walden, B. E., Montgomery, A. A., Gibeily, G. J., Prosek, R. A., & Schwartz, D. M. (1978). Correlates of psychological dimensions in talker similarity. *Journal of Speech and Hearing Research*, 21(2), 265–275. <https://doi.org/10.1044/jshr.2102.265>
- Wallraven, C. (2014). Touching on face space: Comparing visual and haptic processing of face shapes. *Psychonomic Bulletin & Review*, 21(4), 995–1002. <https://doi.org/10.3758/s13423-013-0577-y>
- Walton J.H., & Orlikoff, R.F. (1994). Speaker race identification from acoustic cues in the vocal signal. *Journal of Speech, Language, and Hearing Research*, 37(4), 738–745. <https://doi.org/10.1044/jshr.3704.738>
- Wiese, H., Schweinberger, S. R., & Neumann, M. F. (2008). Perceiving age and gender in unfamiliar faces: Brain potential evidence for implicit and explicit person categorization. *Psychophysiology*, 45(6), 957–969. <https://doi.org/10.1111/j.1469-8986.2008.00707.x>
- Zhao, L., & Bentin, S. (2011). The role of features and configural processing in face-race classification. *Vision Research*, 51(23–24), 2462–2470. <https://doi.org/10.1016/j.visres.2011.10.001>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.