Neuropsychologia 116 (2018) 205–214

Contents lists available at ScienceDirect

Neuropsychologia

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

The cognitive neuroscience of person identification

Irving Biederman[a,b,*], Bryan E. Shilowich[a], Sarah B. Herald[b], Eshed Margalit[b], Rafael Maarek[c], Emily X. Meschke[b], Catrina M. Hacker[b]

[a] Department of Psychology, University of Southern California, USA
[b] Neuroscience Program, University of Southern California, USA
[c] Biomedical Engineering, University of Southern California, USA

ARTICLE INFO

Keywords:
Voice recognition
Phonagnosia
Face recognition
Prosopagnosia
Face imagination
Voice imagination

ABSTRACT

We compare and contrast five differences between person identification by voice and face. 1. There is little or no cost when a familiar face is to be recognized from an unrestricted set of possible faces, even at Rapid Serial Visual Presentation (RSVP) rates, but the accuracy of familiar voice recognition declines precipitously when the set of possible speakers is increased from one to a mere handful. 2. Whereas deficits in face recognition are typically perceptual in origin, those with normal perception of voices can manifest severe deficits in their identification. 3. Congenital prosopagnosics (CPros) and congenital phonagnosics (CPhon) are generally unable to imagine familiar faces and voices, respectively. Only in CPros, however, is this deficit a manifestation of a general inability to form visual images of any kind. CPhons report no deficit in imaging non-voice sounds. 4. The prevalence of CPhons of 3.2% is somewhat higher than the reported prevalence of approximately 2.0% for CPros in the population. There is evidence that CPhon represents a distinct condition statistically and not just normal variation. 5. Face and voice recognition proficiency are uncorrelated rather than reflecting limitations of a general capacity for person individuation.

1. Introduction

Social competence requires that we distinguish members of our own species. The face offers a primary stimulus for individuation; voice provides another. Although other perceptual routes to person identification exist, such as body shape and movement, we here review the similarities and differences in face and voice recognition with special attention to their deficits when congenital in origin, prosopagnosia (CPros) and phonagnosia (CPhon), respectively. As face recognition and its deficits have garnered an outsized portion of human individuation science and have undergone recent, extensive reviews (e.g., Duchaine and Yovel, 2015), we will focus more on phenomena associated with voice recognition, considering face recognition as a standard for comparison.

Specifically, we will explore five propositions with respect to the characteristics of face and voice recognition. 1. There is little or no cost when a familiar face is to be recognized from an unrestricted set of possible faces, but the accuracy of familiar voice recognition declines precipitously when the set of possible speakers is increased from one to a mere handful. 2. Whereas deficits in face recognition are typically perceptual in origin—apparent on minimal, simultaneous match-to-sample tasks with no memory requirement– those with normal perception of voices can manifest severe deficits in their identification. 3. CPros report that they do not have imagery of any kind; CPhons only report an imagery deficit for voices. 4. The prevalence of CPhon at 3.2% is moderately higher than the approximately 2.0% reported for CPros, but only for CPhons do we have evidence that this rate exceeds what would be expected from normal variation. 5. Face and voice recognition proficiency are uncorrelated rather than reflecting limitations of a general capacity for person individuation. We review what is known about the cortical localization of these capacities.

Nomenclature. Prosopagnosics for whom there was no evidence of a lesion or neurological condition that could have led to a deficit in face recognition have been termed “Developmental Prosopagnosics” to distinguish them from “Acquired” Prosopagnosics, where a lesion or other neurological condition could have led to the deficit. However, “Developmental” implies that the origin of the deficit was a consequence of behavioral events in infancy or early childhood. There is no evidence, to our knowledge, for such causality. In the absence of either specific “acquired” lesions or differential childhood experience and given that there is higher concordance of prosopagnosia in identical than fraternal twins (Wilmer et al., 2010, PNAS), the more likely explanation is that such cases are congenital in origin so we use the term Congenital Prosopagnosia or Prosopagnosics (CPros). CPros have been

* Correspondence to: University of Southern California, Neuroscience Program, Hedco Neurosciences Bldg., 3641 Watt Way, Los Angeles, CA 90089-2520, USA.
E-mail address: bieder@usc.edu (I. Biederman).

https://doi.org/10.1016/j.neuropsychologia.2018.01.036
Received 4 April 2017; Received in revised form 22 January 2018; Accepted 24 January 2018
Available online 15 February 2018
0028-3932/

shown to have smaller receptive fields than controls in FFA (Fusiform Face Area) (Witthoft et al., 2016). Of course, one cannot confidently attribute a congenital origin without an identification of specific genetic markers but now we believe that the predominance of evidence favors a congenital explanation for those who are prosopagnosic without any evidence for acquired effects. By extension we use the term Congenital Phonagnosia (CPhon) to refer to marked and persistent deficits in voice recognition without a history of neurological insult.

In an age of nighttime lighting, caller ID, and the rarity of immersion in dense foliage or jungle, voice recognition undoubtedly plays a lesser role than it did in our evolutionary past, but it still is of value for those who are engaged in conversation with a person not in view, or with a small group of individuals who are not readily differentiated visually, either because the speakers are not in easy view or the listener has low vision. Indeed, people we know *expect* their voices to be recognized:

Knock, knock.
"Who's there?"
"It's me."

2. Uncertainty in the recognition of objects, faces, and voices

We can achieve near ceiling accuracy if we are asked to identify an image of a familiar object or a headshot of a well-known celebrity without any restriction of the set of possible individuals. Can this be done at RSVP rates? Can familiar voices be recognized when the set of possible speakers is large?

We will adopt object recognition as a yardstick against which to assess the recognition of familiar (celebrity) faces which, in turn, will provide a basis of comparison for voice recognition. Few high-level recognition tasks can be performed as quickly and as accurately as the basic-level recognition of familiar classes of objects, even under conditions where there are no restrictions of the set of possible objects.

2.1. Positive detection of target objects and faces in RSVP sequences

Subramaniam et al. (1995, 2000) compared the detection of objects and familiar celebrity faces in RSVP sequences. Each sequence was composed of either 32 line drawings of common objects, with a target specified by a basic-level name, e.g., "chair," or 32 gray-level headshots of celebrities (primarily politicians and entertainers), with the target specified by the celebrity's name, e.g., "Bill Clinton," well known to the college undergraduate participants at the time of testing. There was a .50 probability of the sequence containing the target which, if present, never appeared in the first six or last six positions. At 126 msec/image, accuracy in object detection averaged 95.0%; faces averaged 82.0%. Note that this task could not be accomplished by selectively monitoring for a simple, low-level feature as participants were uncertain as to what particular image of the object category or celebrity face would be presented. That is, the specific model of a chair and its orientation were unknown prior to its presentation as was the pose, lighting, hairstyle, and expression of the celebrity's face. The advantage in accuracy for objects is not at all surprising in that the object task was being performed at a basic level—any chair—whereas the face task was being performed at a subordinate level—a particular person's face–which would be physically much more similar to the surrounding faces than the surrounding objects to the target object.

It is possible to increase the uncertainty of the characteristics of the target exemplar still further for both faces and objects. Intraub (1981) employed a "negative detection" RSVP task in which subjects were to respond if the sequence contained an object that was, for example, *not* an animal. At 114 msec/picture, negative detection accuracy was markedly lower at 35% compared to the 71% accuracy when the target was specified by a name, e.g., "elephant." In the positive (name) detection condition, for a response to be deemed correct, a subject had to verbally report distinguishing perceptual features of the target, e.g., "leather easy chair." In the negative detection condition, the basic-level name would suffice.

Is detection possible for an unnamed celebrity's face among non-celebrity faces in RSVP sequences, and if the presence of a celebrity is detected, could the celebrity be identified? Meschke et al. (2017) had subjects view RSVP sequences of 32 colored photographs of either familiar objects, all but possibly one from the same category, e.g., tools, or high-quality headshots, all but possibly one, of non-celebrities. For the objects, subjects performed a negative detection task, similar to Intraub's, in detecting an object that was not a member of an object category, e.g., "Not a Tool." For the face detection task, subjects were to detect whether there was a celebrity in the sequence. (Like the object task, the face task could be regarded as a negative detection task as well, in that the subjects were to detect a face that was *not* that of a non-celebrity.) Both kinds of targets occurred on 50% of the sequences.

In requiring detection of an unspecified celebrity—*any* celebrity—among non-celebrity faces, the detection task was designed to assess the limits (if they could be found) of speeded face recognition under severe limitations of processing time, sequential attentional capacity, forward and backward masking of highly similar stimuli, and with high uncertainty as the set of possible individuals was likely in the hundreds, if not the thousands. Given variations in the 3D pose, lighting direction, expression, hair style, etc. of the faces, the effective image variation was essentially infinite.

The faces were presented at rates of 114–150 msec/image and the objects at rates of 76 msec/image. (At slower presentation rates pilot testing had established that accuracy for object recognition was close to ceiling for most of the participants.) Overall, negative detection of familiar objects at 76 msec/object was reliably higher than the negative detection of celebrity faces at 114–150 msec, 89–75%.

Although caveats are in order in comparing across experiments, the level of performance in the negative detection tasks is higher than what would be expected from the Subramaniam et al. (1995 experiment, 2000) *positive* detection RSVP tasks where at 126 msec/image, objects were detected at 95% accuracy; faces at 82% accuracy. In Meschke et al.'s negative detection experiment for objects presented at 76 msec/image, target objects were detected at 89% accuracy and celebrity faces, presented at an average rate of 123 msec/image, were detected at 74% accuracy. Almost all the errors on the face task were misses. There were only a few false alarms where the subject erroneously judged that there was a celebrity in the sequence. These results document a surprising robustness of face recognition performance under conditions of high uncertainty and extremely brief, masked exposures with foils (non-celebrities) that were highly similar to the targets.

Although the subject's main task in the Meschke et al. RSVP task with faces was to detect whether or not there was a celebrity in the sequence, following a positive detection response they were also instructed to identify the celebrity by name or other individuating information. Over 97% of the positive detections of faces were accompanied by sufficient individualizing information to clearly indicate that the subject knew who the celebrity was—most often with a voicing of the celebrity's name. These results suggest scant reliance on an "unconscious familiarity signal" that would indicate signal recognition without conscious awareness (e.g., Tranel and Damasio, 1985).

This somewhat lengthy review of the (often minimal) effects of uncertainty on object and celebrity face recognition is motivated by the marked deficit in the accuracy of voice recognition as the number of possible targets for a given voice is increased to a number markedly below what yields high accuracy for faces.

2.2. The effect of uncertainty on the recognition of newly learned and familiar celebrity voices

It would be unwieldly, if not impossible, to present RSVP sequences for voices with interpretable results. We will thus examine only the

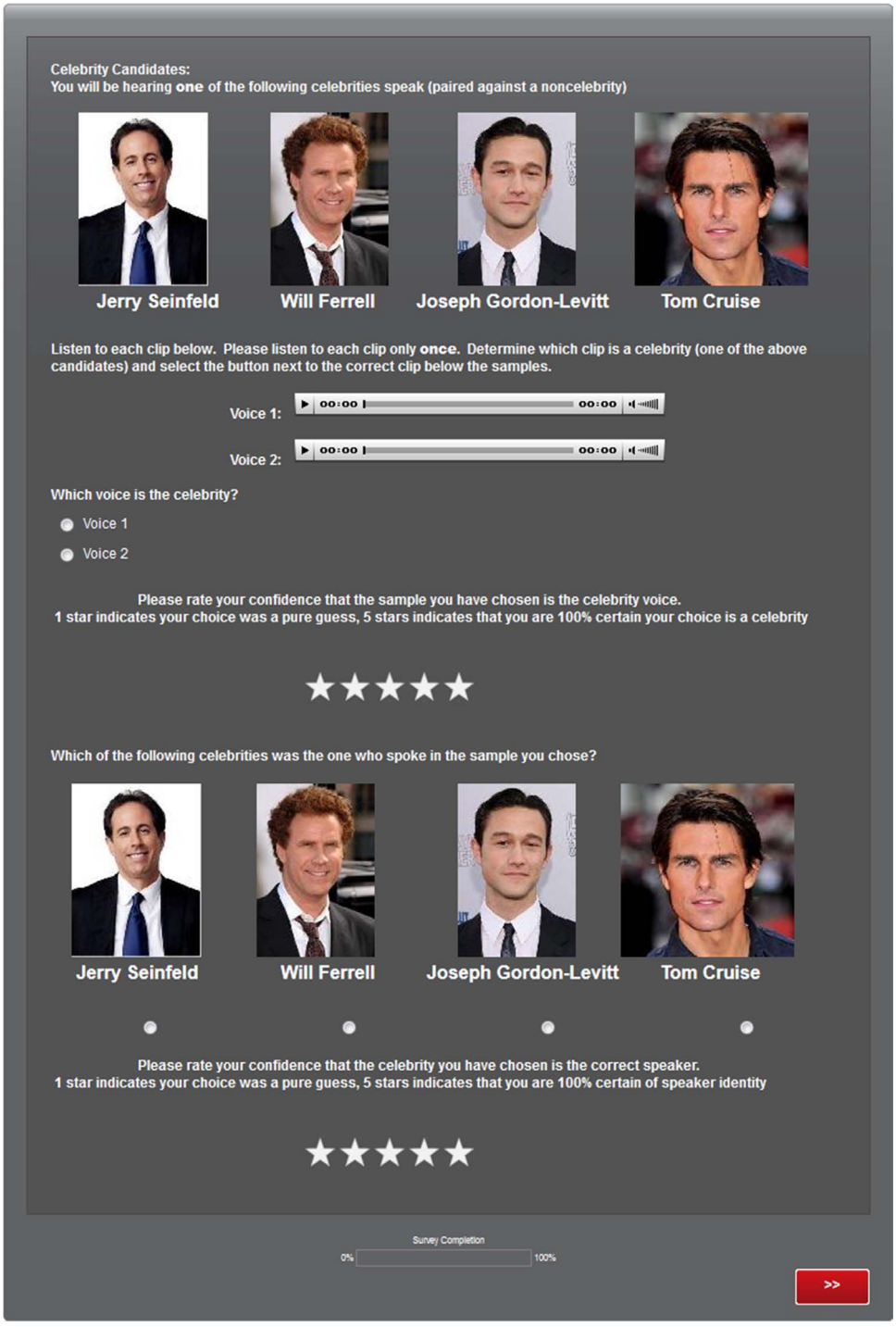

Fig. 1. Sample screen shot of a four-celebrity trial from the USC Celebrity Voice Recognition Test (USC-CVRT). Participants listened to the clips by pushing the "play" buttons and then selected the bubble for Voice 1 or Voice 2 to choose which voice was one of the celebrities. The specific identity of that choice was chosen with the bubble options under the headshots. Confidence ratings for both the voice and identity choices were made with the five-star scale. Trials with two celebrities had the same format but with only two celebrity pictures. Trials with one celebrity displayed only the upper portion of the screen shot shown in Fig. 1 as the voice choice also defined the one pictured celebrity.

effect of variation in the number of possible voices on recognition accuracy as was done by Legge et al. (1984) in his investigation of recognition memory of newly learned voices. Subjects heard recordings of 5, 10, or 20 females, age 18–35, reading brief passages from Grimm's Fairy Tales of varied lengths from 6 to 133 s. After a 15-min retention interval, the subjects heard two voices reading a passage from another of Grimm's Fairy Tales. One of the speakers was one of the original voices; the other was a new voice of the same sex, accent, and approximate age. Participants had to indicate which of the two was a voice from the earlier phase of the experiment. There was a marked decline in accuracy as a function of number of voices originally studied, with the group that had heard 20 voices each speaking clips of 6 s duration not significantly above the .50 chance level.

Whereas Legge et al. (1984) assessed recognition of newly learned voices, Xu et al. (2015) and Shilowich and Biederman (2016) employed a celebrity voice identification test (Fig. 1) to assess recognition accuracy for familiar voices. Subjects heard conversational samples, 7 s in length, spoken by two different individuals. On the screen were the

headshots of either 1, 2, or 4 celebrities, all the same sex, approximate age, and accent. One clip was that of the celebrity (or, with two or four choices, one of the pictured celebrities) with the content of the clip offering no hint as to the celebrity's identity. The other clip was that from a non-famous person. Subjects had to indicate which clip was that of a celebrity and then to indicate, when there were two or four celebrities, which celebrity it was. (With one celebrity, the selection of which clip was that of the celebrity voice also specified the identity of the celebrity.) To be scored as correct on a given trial, subjects had to both select which of the two voices was that of the celebrity and, with two or four celebrities, which was the celebrity whose voice they had just heard. For both choices (Which is the celebrity's voice? Which celebrity is it?), the subjects indicated their degree of confidence in the correctness of their choice.

Xu et al. (2015) and Shilowich and Biederman reported a highly significant decline in both accuracy and confidence on this test as the number of possible celebrities increased from one to four possibilities. For a non-phonagnosic subject, choosing the correct voice was associated with choosing the correct identity on 90% of the trials, and vice versa. If either the voice choice or identity choice were wrong, accuracy on the other question on that trial was at chance. Confidence for control subjects exceeded 4.5 on the 5-point scale for answers they got correct – when they were correct, they knew it. These effects were apparent even when the subjects indicated high familiarity both with the celebrity whose voice they heard (the target) as well as with the foils.

The result that RSVP detection of a celebrity face is almost always accompanied by successful identification (Meschke et al., 2017) would seem to parallel the voice recognition result that choosing the correct celebrity voice was almost always accompanied by successful identification and that being incorrect on one question was accompanied by being at chance on the other question. In both faces and voices, detecting the correct face or voice as the celebrity was almost always associated with correctly identifying to whom the face or voice belonged.

As suggested by the large decline in identification from one to four possible target celebrities, the recognition of celebrity voices is extraordinarily difficult with a large set of possibilities, e.g., “a celebrity” (See also Legge et al., 1984). Four individuals who performed well on the USC Voice Recognition Test (where the subjects were to choose from one to four possible celebrities on each trial) were unable to identify any celebrity voices when there was no restriction as to the set of possible voices. The effect of uncertainty is, perhaps, the most striking difference between face and voice recognition. With an unrestricted set of possible familiar celebrities, recognition of their faces is readily demonstrated under normal viewing conditions and remains highly accurate even at the extremely short, masked presentation durations in the RSVP negative detection tasks. Under comparable conditions of uncertainty but with normal speaking rates and a 7 s voice clip that allowed a comfortable sample, recognition of the voices of these same celebrities is virtually impossible. Below we discuss some of the theoretical implications of this difference.

We note an asymmetry in accuracy in correctly selecting which celebrity was speaking in the two- and four-celebrity trials when the subject was familiar with the target and not the foil(s) on the voice recognition test compared to the opposite case (familiar with the foil(s) but not the target). The effect is most easily described with the two-choice test: When the subject was familiar with the target but not the foil, accuracy was markedly higher at 81% at selecting the correct speaker than when the subject was familiar with the foil but not the target (62%), even though, logically, the uncertainty was the same. Having a model of the target's voice led to more accurate performance than a model of what the target's voice was not (Shilowich and Biederman, 2016).

3. Phonagnosia: a case study

Although many cases of CPros have been described in the literature, there had been no systematic studies of CPhon until Garrido et al.'s (2009) investigation of KH, a 60 year old woman who reported difficulty in recognizing voices. Xu et al. (2015) investigated AN, a 21-year old congenital phonagnosic at the time of testing, who had approached author IB, her instructor in her cognitive neuroscience course, with the statement that she could not recognize voices. She had only recently become aware of her deficit when alerted to an incident when she had failed to recognize the voice of her favorite singer. Until that episode and her experience in the research project, AN was unaware that it was even possible to recognize a person by voice without seeing that person's face.

She reported no neurological history of any kind nor displayed any remarkable features in her structural MRI. A T2 scan examined by a neuroradiologist reported no incidental findings. She has what appears to be perfectly normal conversational abilities with no hint of a hearing loss, as she was equal to controls in discriminating tones and chimes, enjoys music, and plays the guitar. She has a responsive and engaging demeanor. She presents no cognitive or social deficits. She had a 3.8 undergraduate GPA and is now in graduate school—and married to her long-term boyfriend. As a result of our study, she was interviewed by the BBC (link: http://www.bbc.co.uk/programmes/b0832fq5) where she presented herself as an articulate and responsive interviewee.

Testing on the USC Celebrity Voice Recognition Test (USC-CVRT) (Fig. 1) confirmed AN's severe deficit in voice recognition. The 100 celebrities that constituted the celebrity voices on the test were generated by AN as her top (i.e., most familiar) 100 celebrities. Fig. 2 shows AN's score in relation to 21 control subjects who took the USC-CVRT. Prior to taking the test each subject rated their familiarity with each celebrity's voice. There was a strong relation between the degree of familiarity with the celebrity voices and the accuracy of choosing the correct celebrity. On that basis, AN is most appropriately compared to the nine subjects who had high familiarity (> .95) with the voices of that set of celebrities. The severity of AN's deficit is underscored by those nine subjects achieving near ceiling accuracy compared to AN's 52%. Nonetheless, AN's score exceeded chance accuracy, which was 29%.

3.1. Does AN have a perceptual deficit in discriminating voices?

ANs difficulty with recognizing celebrity voices does not appear to be a consequence of a general failure to perceptually discriminate

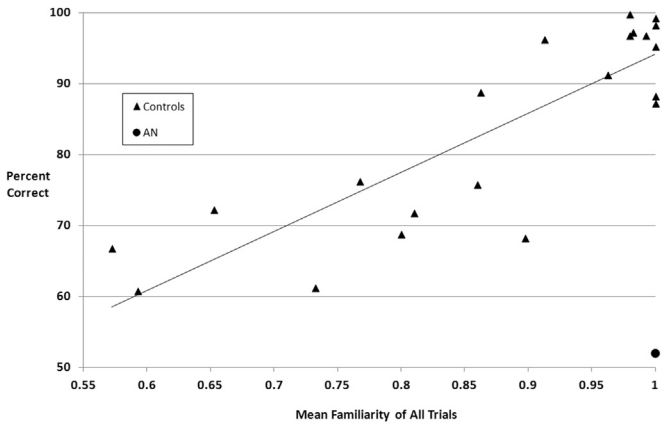

Fig. 2. Mean percent correct voice recognition as a function of mean familiarity with the celebrities. The individual scores for 21 control subjects are designated by black triangles; AN's by the black disk in the lower right corner. The nine subjects who had high familiarity ratings with the celebrity voices (> .95) scored over 85%. AN's score was at 52%. Chance on this test was 29%. Adapted from Xu et al. (2015) Fig. 3.

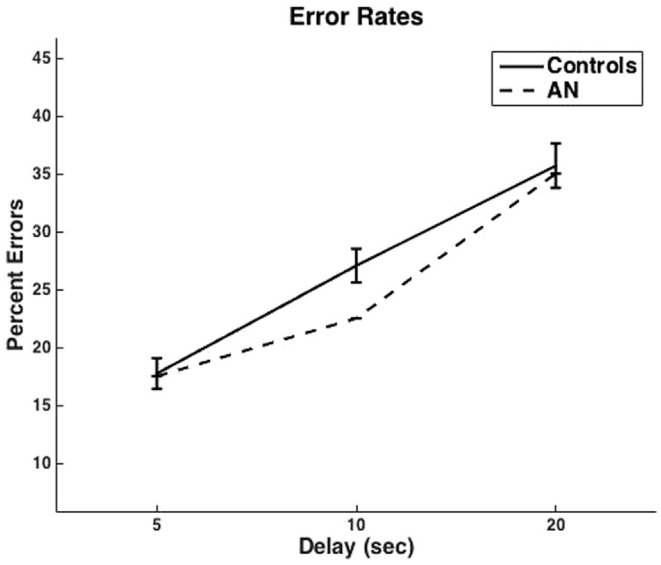

Fig. 3. Percent error in matching one of a pair of similar female voices as the voice originally heard reading a short sentence over a filled delay (counting backwards by 3 s) of 5–20 s. Chance performance on the test was 50%. AN was at the median of the controls (n = 9) and showed the same effect of delay. From Xu et al. (2015), Fig. 4.

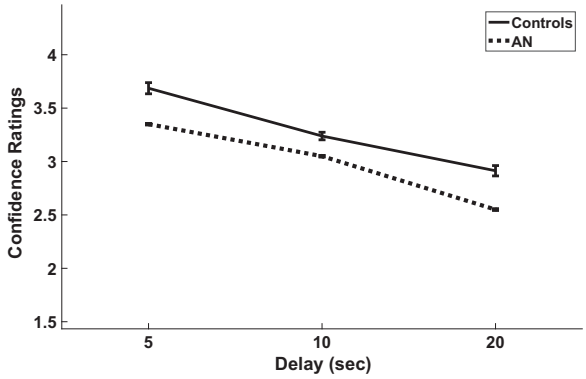

Fig. 4. Mean confidence ratings (on a five-point scale, with 5 designating the highest confidence) of accuracy on the voice-discrimination test shown in Fig. 3 of AN and controls over delay between the initial voice and the test voices.

voices. On each trial of a voice discrimination test AN and controls listened to a short sentence ("The clown had a funny face.") spoken by a novel female voice, approximately 20 years of age, without a discernible accent. After a filled interval (counting backwards aloud by 3 s from a 3-digit number) of 5–20 s, they then heard two female voices, one the original and one a similar but different voice, each speaking the same sentence ("The postman closed the gate.") but one which differed from the initial sentence. They then selected, by key press, whether the voice matching the original voice was the first or second test voice. AN's performance on this task was at the median of the controls, with both showing similar increases in error rates over delay length (Fig. 3).

After each trial, participants rated their confidence that they were correct on a 5-point scale (with 5 = High Confidence) in selecting the correct voice. AN's confidence was lower than that of the controls—beforehand she expressed low confidence that she could perform this task—she showed the same decline in confidence as the controls over the delay interval (Fig. 4).

After each trial, subjects rated the similarity of the speakers' voices on that trial. Although AN rated the similarity of the voices to be much higher than controls, like the controls, these ratings did not vary with delay (Fig. 5).

The pairs of test sentences differed in difficulty, likely reflecting the

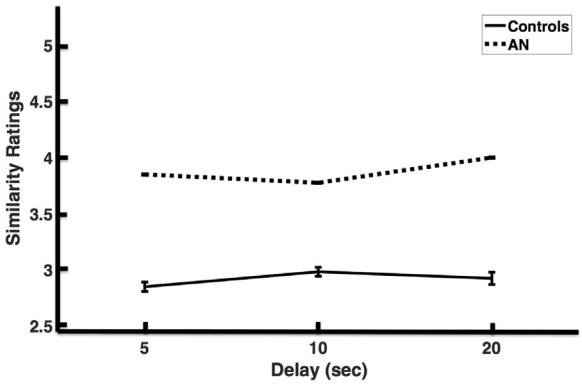

Fig. 5. Ratings of the similarity of the two voices by AN and Controls as a function of the delay between original and matching voices.

similarity of the voices. The correlation of ANs average accuracy with the average accuracy vector of the sentences was .29, at the lower end but within the range of the controls, .27–.63 (mean, $r = .41$, $p < .001$ vs. $r = 0$). This result suggests that AN does not construe voices differently from the controls (Xu et al., 2015). Although caution, of course, must be exercised in generalizing from a single case, AN represents a clear case of extreme congenital phonagnosia without a) a deficit in voice or non-voice sound perception or in her short-term memory of voices, and b) any history of neurological insult or fMRI evidence of brain abnormalities.

4. Cortical loci and function for face and voice recognition

The regions supporting face recognition and prosopagnosia have been the subject of considerable study, in both humans and macaques and will just be reviewed briefly here. In humans, three relatively posterior areas, termed the "core face processing system" (Haxby et al., 2000), consisting of the Occipital Face Area (OFA), FFA, and the Superior Temporal Sulcus (STS), are selective to faces, showing greater BOLD activity to images of faces compared to objects, scrambled faces, body parts, places, words, and numbers. Bilateral lesions to OFA and FFA, with sparing of STS, are sufficient to render an individual profoundly prosopagnosic (Xu and Biederman, 2014).

4.1. Cortical loci for voice recognition

Belin et al. (2000) and his associates have identified a bilateral temporal lobe region that shows greater activation to human vocalizations than to (non-human) animal or inanimate sounds. They termed this region the *temporal voice area* (TVA) which is localized by having participants passively listen to blocks of human sounds and non-human sounds, including inanimate objects, e.g., a cart rolling on a wooden floor, animal sounds, e.g., a songbird, and natural non-animal sounds, e.g., a babbling brook. The TVA is defined as the bilateral temporal lobe region where greater activity is observed when listening to the human sounds compared to sounds made by inanimate objects or natural entities or animals.

It is perhaps unfortunate that the blocks of human sounds in the TVA localizer include, in random appearing fashion, not just speech but also non-speech sounds, such as sneezing or laughing. The motivation for the inclusion of non-speech sounds was to demonstrate that activation of the TVA was not merely reflecting language. It would be preferable to have been able to analyze separately the activity elicited by speech from human non-speech vocalizations to determine their possible differential activation of the TVA. Moreover, to our knowledge, there has not been a demonstration that reasonably accurate speaker identification could even be achieved through human non-speech vocalizations. We certainly hear speech from a known individual more

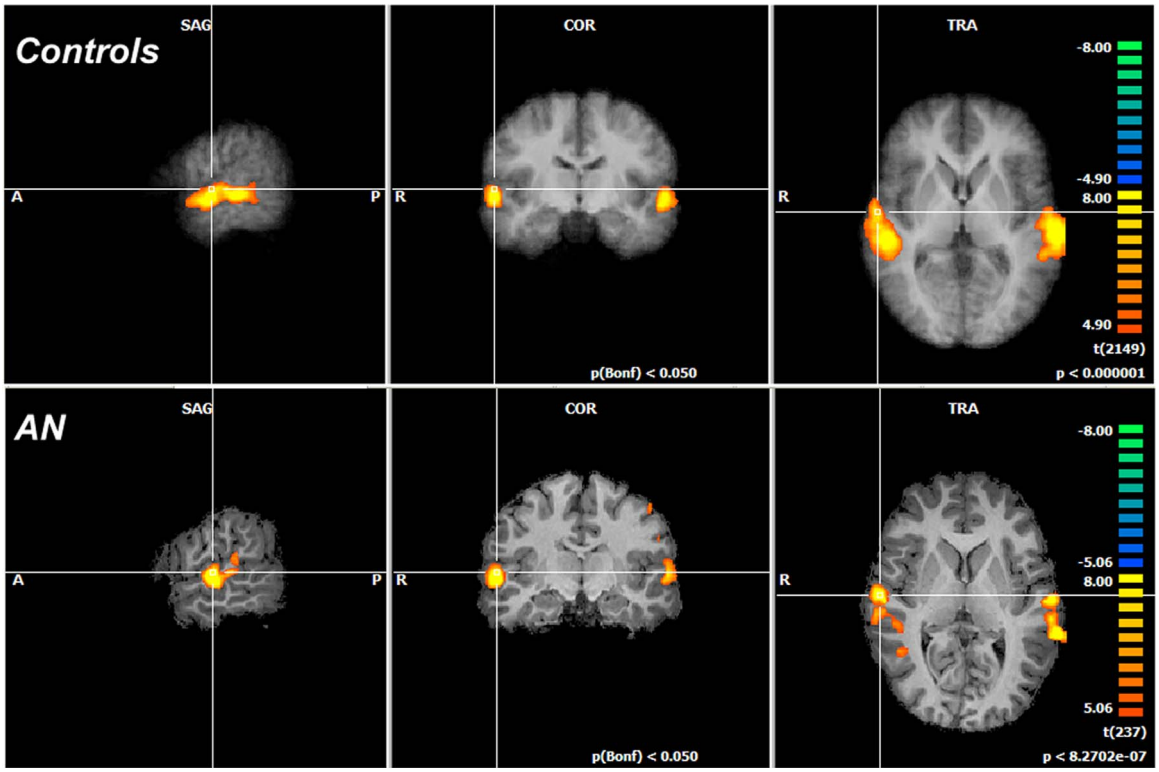

Fig. 6. Activation pattern of the contrast of listening to human versus non-human sounds. The Temporal Voice Areas (TVAs) are shown in both the controls (n = 9) and AN, superimposed on the average of the controls and AN's structural image, respectively. From Xu et al. (2015), Fig. 6.

frequently than we hear him or her laugh or sneeze, so we are probably better able to determine identity from speech than from other kinds of human vocalizations, although, this issue has not been put to test. Similarly, the information in an utterance that allows a listener to identify a speaker has not been isolated. Following Xu et al. (2015), we will refer to that information as Voice Individuating Cues (VICs). Such cues include both supersegmental variation such as prosody, as when a voice rises when posing a question, as well as instantaneously conveyed aspects of the voice, such as its fundamental frequency that would be apparent when saying a single vowel. Kreiman and Sidtis (2011) list 103 features of prosody that have been shown to vary across individuals.

More precise localization of sensitivity to differences in human voices were reported by Von Kriegstein and Giraud (2004) who had subjects listen to sentences spoken by either familiar or unfamiliar speakers. Recognition of voices activated rSTS more than recognition of verbal content, with the posterior rSTS more activated by novel than familiar voices. Consistent with this finding were the results of a fast event-related adaptation study by Belin and Zatorre (2003) who reported that rSTS showed a reduced BOLD response when a syllable was repeated by the same voice compared to a different voice.

It would not be implausible to hypothesize that VICs are extracted and activate representations of different voices in the superior temporal gyrus (STG) where the TVA is localized. Bethmann et al. (2012) performed fMRI scanning of participants while they listened to 2 s clips of short phrases of familiar celebrity and unfamiliar voices. They judged whether the voice was familiar or unfamiliar and attempted to name those that they judged as being familiar. Greater BOLD activation was observed for the familiar voices in the superior temporal lobe bilaterally, with the magnitude of the effect increasing in the more anterior regions of the temporal lobes (ATL) and whether the voice could be named (vs. not named). Consistent with the prior discussion of the difficulty of voice recognition with large and unknown sets of individuals, only an average of 11 of the 70 famous voices could be accurately identified. It was not specified whether the identified voices tended to be highly distinctive.

A potential shortcoming of the design was that subjects were actively attempting to identify the voices so there is the possibility that the enhanced BOLD response to familiar voices reflected the sometimes-successful activation of biographical associations of the speakers rather than the matching of the VICs to a stored representation of a speaker's voice. Consistent with this interpretation is that the enhanced BOLD response to familiar voices was also observed when the participants judged the voice to be familiar even when it was not that of a celebrity. It would be of interest to assess whether familiar voices would elicit enhanced activation with an orthogonal task in which the participant was judging, say, the age of the speaker.

Somewhat counter evidence to the role of right pSTS involvement in voice perception comes from Jiahui et al. (2017). These investigators tested a prosopagnosic patient with a lesion that included the right pSTS on a variety of voice perception tasks (e.g. identity, sex, age) and found no difference from controls. The authors conclude that the BOLD responses to voice stimuli in the right pSTS are likely reflecting higher-level integration of voice and face information rather than the processing of voice information that underlies voice perception and recognition. The same explanation was cited as a possible account of Bethmann et al.'s (2015) results.

4.2. Possibility that the loci for speaker identification may not be the same loci involved in language recognition

From another perspective, the region supporting phoneme decoding might *not* be expected to be the region involved in voice identification, given that we can extract phonemes independently of who is talking.

That there may be independent decoding of language and speaker is consistent with several observations. Kreiman and Sidtis (2011) describe individuals with language deficits so profound that they could not understand any linguistic aspect of speech yet, nonetheless, were readily able to identify the speaker. The lesion site for these individuals, unfortunately, was not reported. More generally, the three congenital

phonagnosics studied by Xu et al. (2015) evidenced no deficits in understanding speech. The one congenital phonagnosic (AN) run on the TVA localizer showed TVA activation indistinguishable from non phonagnosic controls (Xu et al., 2015), as shown in Fig. 6 and appears to have perfectly normal speech comprehension.

This does not necessarily mean that the STG is irrelevant for distinguishing voice identity. Voice identification, of course, requires an auditory input but it may be that the auditory areas of the superior temporal lobe may not be where individual voices are decoded. Consistent with the findings of other investigators, Van Lancker and Canter (1982) reported that deficits in voice identification tend to be produced by right hemisphere lesions. Deficits were defined as being 1.5 SDs below the mean of a non-lesioned control group in: a) matching familiar voices to named headshots of one of four celebrities or b) in discriminating unfamiliar female voices as same or different. Left hemisphere lesions produced no deficit in recognition but discrimination scores that were intermediate between controls and right hemisphere lesioned participants. A somewhat surprising finding was that the largest deficits in recognition were associated with lesions to the right parietal cortex. Consistent with the lesion results, Schelinski et al. (2016) have shown that the right precuneus is more strongly activated when listening for voice identity than when listening for language. This effect is absent in individuals high on the autism spectrum disorder scale. Such individuals have difficulty in recognizing voices.

4.3. Loci of phoneme discrimination

Mesgarani et al. (2014) used subdural recordings from five electrodes in the left superior temporal gyrus (STG) in six patients, prior to their undergoing surgery for epilepsy. From a total of 500 different sentences each spoken by 400 people, they determined the mean neural response at each electrode to every phoneme and were able to determine the coding of individual phonetic features, e.g., nasality, voicing, for the full set of English phonemes. That phoneme classification could be extracted from the recordings of so many different speakers/sentences documents the invariance over speaker that is achieved by the left STG. But what allows invariance over speakers may not be the same system that allows speaker individuation. We can view the lateral occipital complex (LOC) and the posterior parietal cortex as a parallel case from vision where different regions perform different computations on the same input, with the former specifying the shape of an object for recognition and the latter its position and shape for motor interaction. It would be of interest to determine if the identity of different speakers, all speaking the same sentences, could be decoded from intracranial recordings or fMRI MVPA at the same approximate location in left STG or, alternatively, if activity at some other region, e.g., right STG, as hypothesized by Von Kriegstein and Giraud (2004), and Bethmann et al. (2012), might be sufficient for decoding VICs.

Wherever the locus of VIC decoding, at least some cases of congenital phonagnosia may stem more from an inability to store, long term, the VICs of a familiar voice than a perceptual deficit in extracting the VICs. This was the tentative conclusion of Xu et al. (2015) regarding AN's phonagnosia insofar as she was equal to controls in discriminating voices over a brief filled interval. She evidences normal episodic knowledge of personally familiar people so the one underlying cause of her deficit that has not been excluded is her long-term memory of the VICs and their association to a particular individual.

5. Loci of associative networks of face and voice recognition

5.1. Evidence of a role of the Person Identification Node (PIN)

In 1986 Bruce and Young proposed a *Person Identification Node* (PIN), a convergence network that links all the information associated with a given individual, e.g., face, voice, name, occupation, sex, etc., which can be activated by any perceptual or semantic individuating probe. It is thus plausible that activation of the PIN (or some associative network that codes for PIN functions) by a non-perceptual cue may improve face or voice identification for the identity associated with the cue. Although depicted as a local network—a single box with visual (face) and auditory (voice) inputs and semantic associations including the person's name as outputs–there is no reason to exclude an implementation as a distributed network other than the general assumption that economy of wiring would suggest that highly associated percepts and concepts might be more closely represented—in terms of neural distances–in associative memory. Reviews by Gainotti and Marra (2011) and Fox et al. (2008) suggest the existence of such loci in associative cortex for various aspects of our knowledge and emotional responses to known individuals. Perhaps best documented are lesions to the left temporal pole that result in deficits in naming familiar people (e.g., Damasio et al., 1996; Gainotti and Marra, 2011). Evidence that the distributed loci for what can be conceptualized as PIN functions might extend beyond the temporal lobes is suggested by Leveroni et al. (2000) finding of greater activation in the ventromedial prefrontal cortex (vmPFC) from personally familiar faces compared to faces repeatedly shown throughout the experiment which had no associated biographical information.

5.2. Facilitation of face and voice recognition from arbitrary associations

Given the functions of a PIN, we would expect that the recognition of a person by face or voice would also provide access to non-perceptual associations to that person, such as her occupation or nationality. But could the activation of arbitrary individuating associations serve to facilitate recognition of a face or voice? In the same manner that a name might elicit an image of a familiar face, there would be no reason to exclude *arbitrary* associations to a face or voice that might serve to facilitate the subsequent recognition of that individual's face or voice. Schwartz and Yovel (2016) studied the effects on face recognition of accompanying the viewing of a face with a name or occupation. The later recognition of that face when depicted in a different pose and lighting direction was more accurate than when the face was originally accompanied by a non-biographical label (“table”), a symbol, a name incompatible with the sex of the face, or even when the face was viewed, without a label, at a variety of orientations and lighting directions during its initial study phase.

In their landmark study of the recognition of newly learned voices, Legge et al. (1984) assessed the effect of presenting a unique face when originally listening to individual voices and, for separate groups, also having the face present during the recognition testing, or no face at all. The largest benefit was having the face present during both initial encoding and retrieval testing, suggesting that faces could serve as a retrieval cue for voices. There was no significant benefit of having the faces only during the initial presentation period over not having a face present at all. It does appear, then, that the facilitation of accompanying unfamiliar faces or unfamiliar voices with a personalized context is dependent on having both present during recall rather than only during initial encoding.

6. Prevalence of prosopagnosia and phonagnosia

Thousands of people have taken the Famous Faces Test on faceblind.org which provided estimates of the prevalence of congenital prosopagnosia of approximately 2%, defined as being 2 S.D.s below the mean in face recognition tests (Holden, 2006; Kennerknecht et al., 2006). In the absence of published distributions and genetic testing, an array of behavioral and neural measures suggests a complex of not-always-consistent phenotypes across individuals who present face recognition deficits.

An issue has been whether what is termed congenital prosopagnosia represents just normal variation in proficiency or whether there is a detectable subgroup of individuals who are of low proficiency beyond

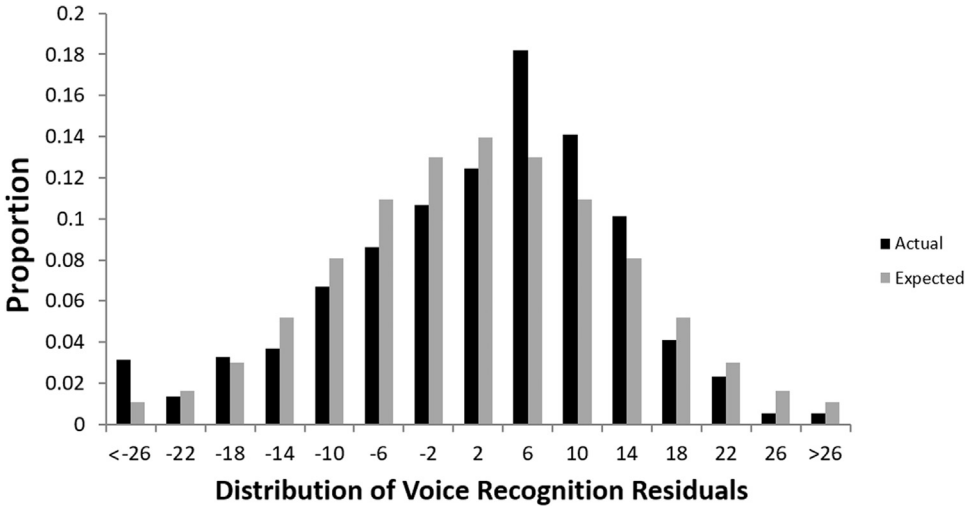

Fig. 7. Distribution of residuals on the USC Celebrity Voice Recognition Test (n = 730) with actual proportions in black and expected proportions from a normal distribution in gray. The residuals are the scores on the voice recognition test (Fig. 1) with the effect of celebrity familiarity on each trial partialled out. The ordinate is the proportion of the total sample for each bin. The values for the bins on the abscissa represent the highest (most positive) value for that bin, either positive or negative. The lowest bin was 2.28 SDs below the mean. 23 participants (3.2%) were in that bin; 8 (1.1%) would have been expected from a normal distribution, 99.5% Confidence Interval = ± .012, $z_{prop1\text{-}prop2}$, $p < .001$. From Fig. 7, Shilowich and Biederman (2016).

what would be expected from, say, an arbitrary cutoff from a normal distribution (Barton and Corrow, 2016). Although this issue has yet to be resolved in the case of congenital prosopagnosia, for congenital phonagnosia, Shilowich and Biederman (2016) employed a version of the celebrity voice recognition test depicted in Fig. 7 and obtained an estimate of prevalence of phonagnosia of 3.2%, which was 2.28 SDs below the mean, shown in Fig. 7. Using Barton and Corrow's criteria, they were able to reject the hypothesis that this estimate of the prevalence of phonagnosia was merely a consequence of normal variation in ability. The value CPhon is thus somewhat higher than the 2.0% prevalence reported for CPros (Kennerknecht et al., 2006).

6.1. Controlling for familiarity

A major challenge in assessing recognition of familiar (celebrity) faces or voices is the proper control of familiarity. In the USC celebrity voice and face recognition experiments, familiarity was assessed by having the subject rate his or her familiarity with each celebrity's voice or face, typically on a 5-point scale. In the voice recognition experiments, there were a varied number (1,2, or 4) possible celebrities on each trial. For every subject an average familiarity value was calculated for each trial. As there was a positive correlation between rated voice familiarity and accuracy, $r(728) = .57$, $p < .001$, separate residuals were computed for each subject's score for trials with 1, 2, and 4 alternative celebrities, based on the regression of that subject's average familiarity values. Each score was thus "corrected" for that subject's specific familiarity values (Shilowich and Biederman, 2016).

The distribution (Fig. 7) shows the residual voice recognition scores, corrected by each individual's familiarity ratings of the celebrities' voices on each trial of the USC Celebrity Voice Recognition Test. Of particular interest is that the bin with the lowest recognition residuals had three times the expected frequency that would be expected from a normal distribution. The inflation of the number of participants in that bin is all the more striking in that the distribution had significantly less kurtosis (i.e., more peakedness) than a normal distribution as is readily evident in Fig. 7, so there should have been fewer cases in the lowest bin than that expected from a normal distribution. By Barton and Corrow's (2016) criteria, this is highly suggestive of a definite syndrome rather than just normal random variation.

We took precautions to guard against the possibility that the increased incidence in the lowest bin from what would be expected from a Gaussian was not a function of the inclusion of subjects who could not be expected to perform well because of general unfamiliarity with American celebrities or who did not give the test serious effort (Shilowich and Biederman, 2016). After eliminating those who did not finish the test, we adopted exclusionary criteria of a) having lived in the U.S. less than five years, and b) being unfamiliar with President Obama's voice. We then excluded individuals who took the test too quickly (as estimated by a 45-min test run by one of the authors) to properly listen and respond to the voice clips. This removed 11 people, which was 1.1% of the 977 who began the test. Their times ranged from 23 to 43 min, with none of the scores being reliably above chance (29%).

Most importantly, subjects made confidence ratings after each choice: Which one is the celebrity voice? Which one is the celebrity (on trials with two or four possible celebrities)? Overall, these confidence ratings were strongly correlated with accuracy. When subjects recognized a voice they invariably knew it. Those 11 subjects who had times too fast for reasonable performance accuracy (and with scores that were not reliably above chance) had extremely low correlations between accuracy and confidence confirming the basis of their exclusion on grounds of high speed and low accuracy. By these criteria, the distribution of residual scores (regressed on individual familiarity ratings) thus excludes those subjects who might not have given an honest effort in taking the test.

None of the subjects in this lowest bin reported any hearing loss or difficulty in understanding speech. None reported any neurological deficits. We also confirmed a correlate with recognition accuracy that was revealed by our earlier research with three phonagnosics (Xu et al., 2015): Those 23 subjects who were in the lowest bin expressed significantly greater difficulty in imagining familiar voices, but not non-voice sounds, than those subjects who scored higher on the recognition test.

6.2. Imagery in prosopagnosia and phonagnosia

Grüter et al. (2009) assessed visual mental imagery in 53 congenital prosopagnosics using Mark's (2009) Vividness of Visual Imagery Questionnaire (VVIQ) which solicits subjective ratings as to vividness of mental imagery of scenes, such as a sunrise, faces, and other entities. The CPros had the lowest mental imagery scores of any group ever reported (Grüter et al., 2009). Notably, the CPros were deficient in imaging *all* stimuli, not just faces. `Although the VVIQ is in dire need of validation, the differential self-report between prosopagnosics and non prosopagnosics is, itself, a reliable behavioral marker. As noted above, congenital phonagnosics also show a marked deficiency in auditory mental imagery but it is confined to human voices, nothing else (Xu et al., 2015). This was assessed by Xu et al. using an online behavioral auditory imagery-rating test where participants could rate the quality of their imagery (on a scale from 1 to 5) of the celebrity voice and non-voice items. The items were composed of 50 highly popular (to college-age individuals) celebrities and 42 non-voice items (objects, natural sounds, musical instruments, and human non-speaking sounds, such as

sneezing). Many of the items were taken from Belin et al. (2000) TVA localizer. All three phonagnosics reported an inability to imagine human voices but generally reported no difficulty in imaging non-voice sounds with ratings equal to controls.

Significantly greater difficulty in imagining familiar human voices was also reported by the subjects in the lowest bin in the phonagnosia prevalence study (Compared to higher scoring subjects) in the Shilowich and Biederman (2016) prevalence study.

7. Super-recognizers

There are exceptional cases where individuals achieve feats of recognition despite enormous uncertainty and limitations of exposure duration and attention. For faces, “super-recognizers” have been discovered (Russell et al., 2009; Keefe, 2016) who can identify “almost every face they have ever seen, including waiters and salespeople encountered only briefly and months earlier” (CBS 60 min web description).

There have been no systematic attempts to detect whether some people are super-recognizers for voices. The distribution of voice recognition residuals (Fig. 7) shows an asymmetry in that, as previously discussed, there is an excess of cases in the lowest bin of the distribution relative to a normal distribution, suggesting some tendency toward phonagnosia. But there is not an excess of cases in the upper bins, suggesting that super-recognizers for voices, if they do exist as a definable syndrome, are sufficiently rare as to not exceed what is expected from normal variation. Nonetheless, there is a report of at least one exceptional individual. Ms. Hariott Daley served as the first telephone switchboard operator for the U.S. Congress from 1898 to 1945. She was reputed to be able to identify, by voice, all 96 senators and 394 representatives, as well as 300 reporters (Schulz, 2017). Given the era, one could reasonably infer that these voices were almost exclusively middle aged to elderly White males, though with variability in regional accent.

8. What is the relationship between voice and face recognition?

Bruce and Young's (1986) PIN account assumes that perceptual inputs from a person's voice and face independently converge on a “node” that is linked to semantic associations about that person. Whether one accepts the concept of a PIN, the presumed functions are clearly ubiquitous and cannot be denied: the recognition of a person quickly activates a vast storehouse of associations involving that person. Using functional and diffusion MRI, Blank et al. (2011) reported connections between voice–processing areas in the middle and anterior superior temporal sulcus (STS) and face processing areas in the superior temporal sulcus, as well as connections between FFA to STS. Blank et al. argue that unisensory face and voice information are integrated using the reciprocal interactions between voice and face processing areas, presumably prior to the associative links in a PIN. The evidence cited in support of this view derives from the finding of von Kriegstein et al. (2008) that voices are better recognized when subjects had audiovisual training with a video of the speakers.

However, Blank et al.’s study does not rule out the distinct possibility that this facilitation might derive from higher level associative connections. As noted above, Jiahui et al. (2017) found that rSTS played no role in voice recognition. Support for an associative basis for the facilitation of voice recognition by faces derives from several studies, such as that of Schwartz and Yovel (2016), who found that face recognition was facilitated by, for example, the presentation of sex-appropriate but not sex-inappropriate names to faces. Similarly, Legge et al. (1984) showed that presenting a picture of the face of a speaker when listening to her voice facilitated later identification of that speaker's voice.

Blank et al. (2011) did not expand upon the potential perceptual consequences of an early integration of voice and face. Would it be expected to yield some type of synesthetic experience where face and voice yielded unique perceptual consequences? To our knowledge, none have been reported.

9. Are abilities at recognition by face and voice correlated?

Is there a general deficit in individuation competence such that people who are at a particular level of ability in voice identification tend to be at that approximate level in face identification? Our own sampling of a general population without detectable lesions does not provide evidence for such a linkage. The correlations with modest sized samples between scores on the USC Celebrity Voice Recognition and non-celebrity measures of Face Recognition ability are all extremely low: .01 with the PI20 (df = 14), .16 with the CFMT (df = 14), and .02 with the USC match-to-sample face task (df = 16) (Biederman et al., 2017). These results are consistent with those of Liu et al. (2015) who reported that cPros were equivalent to controls in voice discrimination and recognition.

That individuals might be discovered who suffer decrements in both face and voice recognition has been argued by Gainotti and Marra (2011). They hypothesize that a right anterior temporal lesion could result in an *associative* (semantic) form of person recognition deficit that might manifest itself both when viewing faces and hearing voices. The details of such a hypothesis have not been fleshed out but presumably such a patient could have normal face and voice perceptual discrimination capacities but have difficulty in associating faces and/or voices with particular individuals. It would not be implausible that such an associative deficit would also produce a deficiency in the patient's semantic and episodic memory for familiar individuals. Phonagnosic AN could qualify as such a case in that she has normal discriminative and short-term memory capacities for voices but is deficient in her long-term association of voices to individuals. But she does not evidence any decrement in the episodic memories of people in her life.

This general picture of independence is reinforced by the contrast in the actual scores (Table 1) of an acquired prosopagnosic, MGH, and phonagnosic AN. The first six tests assess face recognition ability and MJH's scores on these face tests are markedly lower than AN's. The USC FPT (Face Perception Test), described previously, provides a direct measure of one's capacity for perceptually discriminating faces (Biederman et al., 2017). The Doppelgänger test (Meschke et al., 2017)

Table 1
Comparison of congenital phonagnosic AN and acquired prosopagnosic MJH on six face and one voice recognition tests.

Subject	PI20[a] (M = 40.8)	USC FPT[b] (M = 83.8%)	Famous Faces Faceblind.org (M = 83.4%)	USC Celebrity Recognition (M = 82.4%)	Doppelgänger Test[b] (M= 87.5%)	CFMT (M = 76.8%)	USC Voice Recognition[c] (M = 88.2%)
AN	25	97%	96.4%	100%	90%	89%	52.0%[d]
MJH	83	52%	3.0%	26%	49%	38%	84.0%

[a] Higher scores indicate greater difficulty with face recognition.
[b] Chance = 50%
[c] Chance = 29%
[d] Tied with the lowest score of 127 subjects who rated themselves as being highly familiar (> 90%) with the celebrities on the voice recognition test.

presents a celebrity head shot with a foil with similar facial characteristics and the subject is to select the celebrity. Performance on the test thus requires no recall or articulation of a name or identifying information. These two tests have a well-defined chance level of 50%, and MJH is at chance on both tests whereas AN is clearly in the superior range on those tests and at ceiling on the USC Celebrity Face Recognition Test. In Voice Recognition, the relative performance is reversed, with AN matching record-low performance levels whereas MJH is in the normal range.

10. Conclusions

Our review of individuation by face and voice identified a number of differences between these two major routes to person identification.

1. There is a marked cost of increasing the number of potential target individuals, i.e., an increase in uncertainty, on voice identification but little or no effect when engaged in face identification.
2. Congenital prosopagnosics are deficient in forming visual images of any kind, whether of faces or objects; congenital phonagnosics are only unable to imagine voices. This is a behavioral marker for the two conditions.
3. Prosopagnosics and those who, more generally, have difficulty in face recognition, show a clear perceptual deficit in their discrimination of faces, as evidenced by their performance on the USC Face Perception Test which requires no memory or invariance challenges. The most extensively investigated case of congenital phonagnosia, AN, evidences no perceptual deficit in voice discrimination or short-term memory for voices. Of course, caution has to be exercised in generalizing from a single case.
4. Performance on the USC Face Recognition Test is strongly correlated—accounting for much of the predictable variance—with the standard tests for face recognition performance, such as the CFMT, the PI20, The Famous Faces Test, and the USC Celebrity Face Recognition Test, suggesting that much of the deficit in prosopagnosia is perceptually based, rather than a limitation of memory or an inability to achieve invariance to orientation or noise.
5. The low correlations between the tests that assess face and voice recognition abilities suggest that variation in these abilities are largely independent rather than reflecting a general capacity for person identification.

Acknowledgements

Supported by NSF BCS 0617699 and Dornsife Research Fund to IB. We thank Jordan J. Juarez for his assistance in data collection in several of the experiments on face recognition and discrimination.

References

Barton, J.J.S., Corrow, S.L., 2016. The problem of being bad at faces. Neuropsychologia 89, 119–124.
Bethmann, A., Scheich, H., Brechmann, A., 2012. The temporal lobes differentiate between voices of famous and unknown people: an event-related fMRI study on speaker recognition. PLoS One 7 (10), e47626. http://dx.doi.org/10.1371/journal.pone.0047626.
Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P., Pike, B., 2000. Voice-selective areas in human auditory cortex. Nature 403, 309–312. http://dx.doi.org/10.1038/35002078.
Biederman, I., Margalit, E., Maarek, R.S., Meschke, E.X., Shilowich, B.S., Hacker, C.M., Juarez, J.S., Seamans, T.J., Herald, S.B., 2017. What is the nature of the perceptual deficit in congenital prosopagnosia? Poster presented at In: Proceedings of the Annual Meeting of the Vision Sciences Society, St. Pete Beach, FL. May. 〈http://geon.usc.edu/~biederman/presentations/PerceptualDeficit_VSS_17.pdf〉.
Blank, H., Anwander, A., von Kriegstein, K., 2011. Direct structural connections between voice- and face-recognition areas. J. Neurosci. 31, 12906–12915.
Bruce, V., Young, A., 1986. Understanding face recognition. Br. J. Psychol. 77, 305–327.
Damasio, H., Grabowski, D.J., Tranel, D., Hichwa, R.D., Damasio, A.R., 1996. A neural basis for lexical retrieval. Nature 380, 499–505.
Duchaine, B., Yovel, G., 2015. A revised neural framework for face processing. Annu. Rev. Vision. Sci. 1, 393–416.
Fox, C.J., Iaria, G., Barton, J.J., 2008. Disconnection in prosopagnosia and face processing. Cortex 44, 996–1009.
Gainotti, G., Marra, C., 2011. Differential contribution of right and left temporo-occipital and anterior temporal lesions to face recognition disorders. Front. Human. Neurosci. 5 (55), 1–11.
Garrido, L., Eisner, F., McGettigan, C., Stewart, L., Sauter, D., Hanley, J.R., Schweinberger, S.R., Warren, J.D., Duchaine, B., 2009. Developmental phonagnosia: A selective deficit of vocal identity recognition. Neuropsychologia 47, 123–131.
Grüter, T., Grüter, M., Bell, V., Carbon, C.C., 2009. Visual mental imagery in congenital prosopagnosia. Neurosci. Lett. 453, 135–140.
Haxby, J.V., Hoffman, E.A., Gobbini, M.I., 2000. The distributed human neural system for face perception. Trends Cogn. Sci. 4, 223–233.
Holden, C., 2006. Have We Met?". Sci. Mag. 312, 1449.
Intraub, H., 1981. Rapid conceptual identification of sequentially presented pictures. J. Exp. Psychol.: Human. Percept. Perform. 7, 604–610.
Jiahui, G., Garrido, L., Liu, R.R., Susilo, T., Barton, J.J., Duchaine, B., 2017. Normal voice processing after posterior superior temporal sulcus lesion. Neuropsychologia 105, 215–222. http://dx.doi.org/10.1016/j.neuropsychologia.2017.03.008.
Keefe, P.R., 2016. Total recall. The New Yorker, Aug. 22, 2016.
Kennerknecht, I., Grüeter, T., Welling, B., Wentzek, S., Horst, J., Edwards, S., Grüeter, M., 2006. First report of prevalence of non-syndromic hereditary prosopagnosia (HPA). Am. J. Med. Genet. 140A, 1617–1622.
Kreiman, J., Sidtis, D., 2011. Foundations of Voice Studies: an Interdisciplinary Approach to Voice Production and Perception. Wiley-Blackwell, West Sussex, U.K.
Legge, G.E., Grosman, C., Pieper, C.M., 1984. learning unfamiliar voices. J. Exp. Psychol.: Learn. Mem. Cogn. 10, 298–303.
Leveroni, C.L., Seidenberg, M., Mayer, A.R., Mead, L.A., Binder, J.R., Rao, S.M., 2000. Neural systems underlying the recognition of familiar and newly learned faces. J. Neurosci. 20, 878–886.
Liu, R.R., Corrow, S.L., Pancaroglu, R., Duchaine, B., Barton, J.J.S., 2015. The processing of voice identity in developmental prosopagnosia. Cortex 71, 390–397. http://dx.doi.org/10.1016/j.cortex.2015.07.030.
Meschke, E., Hacker, C., Juarez, J., Maarek, R., Biederman, I., 2017. Can familiar faces be negatively detected at RSVP rates? J. Vision. 17, 1027. http://dx.doi.org/10.1167/17.10.1027.
Mesgarani, N., Cheung, C., Johnson, K., Chang, E.F., 2014. Phonetic feature encoding in human superior temporal gyrus. Science 343, 1006–1010. http://dx.doi.org/10.1126/science.1245994.
Russell, R., Duchaine, B., Nakayama, K., 2009. Super-recognizers: people with extraordinary face recognition ability. Psychon. Bull. Rev. 16 (2), 252–257.
Schelinski, S., Borowiak, K., von Kriegstein, K., 2016. Temporal voice areas exist in autism spectrum disorder but are dysfunctional for voice identity recognition. Social. Cogn. Affect. Neurosci. 11 (11), 1812–1822.
Schulz, K., 2017. Call and response. The New Yorker, March 6, 26-32.
Schwartz, L., Yovel, G., 2016. The roles of perceptual and conceptual information in face recognition. J. Exp. Psychol.: General. 145 (11), 1493–1511. http://dx.doi.org/10.1037/xge0000220.
Shilowich, B.E., Biederman, I., 2016. An estimate of the prevalence of developmental phonagnosia. Brain Lang. 159, 84–91. http://dx.doi.org/10.1016/j.bandl.2016.05.004.
Subramaniam, S., Biederman, I., Kalocsai, P., Madigan, S.R., 1995. Accurate identification, but chance forced-choice recognition for RSVP pictures. Investig. Ophthalmol. Vis. Sci. 36, 377.
Subramaniam, S., Biederman, I., Madigan, S.A., 2000. Accurate identification but no priming and chance recognition memory for pictures in RSVP sequences. Vis. Cogn. 7, 511–535.
Tranel, D., Damasio, A.R., 1985. Knowledge without awareness: an automatic index of facial recognition by prosopagnosics. Science 228, 1453–1455.
Van Lancker, D., Canter, G.J., 1982. Impairment of voice and face recognition in patients with hemispheric damage. Brain Cogn. 1, 185–198.
von Kriegstein, K., Dogan, O., Grüter, M., Giraud, A.L., Kell, C.A., Grüter, T., Kleinschmidt, A., Kiebel, S.J., 2008. Simulation of talking faces in the human brain improves auditory speech recognition. PNAS 105, 6747–6752.
Von Kriegstein, K., Giraud, A.-L., 2004. Distinct functional substrates along the right superior temporal sulcus for the processing of voices. NeuroImage 22, 948–955.
Witthoft, N., Poltoratski, S., Nguyen, M., Golarai, G., Liberman, A., LaRocque, K., Smith, M.E., Grill-Spector, K. 2016. Reduced spatial integration in the ventral visual cortex underlies face recognition deficits in developmental prosopagnosia bioRxiv, http://dx.doi.org/10.1101/051102.
Xu, X., Biederman, I., Shah, M.S., 2014. A neurocomputational account of the face configural effect. J. Vision. 14, 1–9. http://dx.doi.org/10.1093/cercor/bht005.
Xu, X., Biederman, I., Shilowich, B.E., Herald, S.G., Amir, O., Allen, N.E., 2015. Developmental phonagnosia: neural correlates and a behavioral marker. Brain Lang. 149, 106–117.