Neuropsychologia xxx (xxxx) xxx–xxx

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

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