Cognition 176 (2018) 195–208

Contents lists available at ScienceDirect

Cognition

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

Original Articles

Learning to recognize unfamiliar talkers: Listeners rapidly form representations of facial dynamic signatures

Alexandra Jesse[*], Michael Bartoli

Department of Psychological and Brain Sciences, University of Massachusetts, 135 Hicks Way, Amherst, MA 01003, USA

ARTICLE INFO

Keywords:
Face perception
Audiovisual speech perception
Talker identification

ABSTRACT

Seeing the motion of a talking face can be sufficient to recognize personally highly familiar speakers, suggesting that dynamic facial information is stored in long-term representations for familiar speakers. In the present study, we tested whether talking-related facial dynamic information can guide the learning of unfamiliar speakers. Participants were asked to identify speakers from configuration-normalized point-light displays showing only the biological motion that speakers produced while saying short sentences. During an initial learning phase, feedback was given. During test, listeners identified speakers from point-light displays of the training sentences and of new sentences. Listeners learned to identify two speakers, and four speakers in another experiment, from visual dynamic information alone. Learning was evident already after very little exposure. Furthermore, listeners formed abstract representations of visual dynamic signatures that allowed them to recognize speakers at test even from new linguistic materials. Control experiments showed that any potentially remaining static information in the point-light displays was not sufficient to guide learning and that listeners learned to recognize the identity, rather than the sex, of the speakers, as learning was also found when speakers were of the same sex. Overall, these results demonstrate that listeners can learn to identify unfamiliar speakers from the motion they produce during talking. Listeners thus establish abstract representations of the talking-related dynamic facial motion signatures of unfamiliar speakers already from limited exposure.

1. Introduction

Seeing a speaker typically improves the recognition of speech (for an overview see e.g., Massaro, 1998; Massaro & Jesse, 2007), as visual speech contributes information that is redundant and complementary to the information provided by auditory speech (Jesse & Massaro, 2010; Summerfield, 1987; Walden, Prosek, & Worthington, 1974). The realization of visual speech varies, however, across speakers; and listeners are sensitive to this variation during speech recognition (e.g., Heald & Nusbaum, 2014; Yakel, Rosenblum, & Fortier, 2000). The variability in speech production across speakers comes, however, with a certain consistency in articulation within a speaker such that seeing how a person produces speech is informative about the person's identity. In particular, the time-varying dynamic information contained in visual speech has been shown to be sufficient for recognizing personally highly familiar speakers (Rosenblum, Niehus, & Smith, 2007), suggesting the storage of this dynamic facial information in long-term representations of highly familiar speakers. Functional and neural frameworks of face recognition (e.g., Bernstein & Yovel, 2015; Bruce & Young, 1986; Haxby, Hoffman, & Gobbini, 2000; O'Toole, Roark, & Abdi, 2002) postulate the existence of representations solely dedicated to storing dynamic facial signatures, in addition to separate representations of the invariant aspects of faces. The current view is, however, that facial dynamic information only helps with the recognition of familiar speakers, and only under difficult viewing conditions (e.g., Knight & Johnston, 1997; Lander & Bruce, 2000, 2004; Lander, Bruce, & Hill, 2001). In contrast, dynamic facial information is assumed not to contribute to learning to recognize unfamiliar speakers (e.g., Natu & O'Toole, 2011; O'Toole et al., 2002). Results have been mixed as to whether seeing motion related to speaking has benefits for the learning of unfamiliar faces (Bennetts et al., 2013; Bonner, Burton, & Bruce, 2003; Christie & Bruce, 1998; Lander & Bruce, 2003; Skelton & Hay, 2008). However, this prior work on talking faces did not test whether dynamic information is indeed stored for newly encountered faces, but rather only tested whether seeing dynamic information *enhances* the formation of static face representations, as the recognition of static faces was assessed at test. In the present study, we provide a direct test of whether seeing facial dynamics of speaking can lead to the formation of representations for unfamiliar speakers. We furthermore assess the amount of exposure needed to form such representations and

[*] Corresponding author.
E-mail address: ajesse@psych.umass.edu (A. Jesse).

https://doi.org/10.1016/j.cognition.2018.03.018
Received 15 March 2017; Received in revised form 13 March 2018; Accepted 21 March 2018

whether these representations are abstract in nature, which is necessary to allow recognition of a speaker from new utterances. We focus entirely on how the facial dynamics related to speaking inform about identity, though facial dynamics can also convey information about expressions and emotions.

1.1. Recognizing speakers from dynamic information in auditory speech

The recognition of speakers is a crucial skill in our social lives. Recognizing people, and recalling abstract and episodic information about them, is easier from faces than from voices (for reviews see Barsics, 2014; Barsics & Brédart, 2012). However, in situations when a speaker can be heard and seen, identity information from voice and face is processed and even integrated to recognize the person (e.g., Belin, Bestelmeyer, Latinus, & Watson, 2011; Campanella & Belin, 2007). In addition, early crosstalk between these processes may exist (Blank, Anwander, & von Kriegstein, 2011; Schall, Kiebel, Maess, & von Kriegstein, 2013; von Kriegstein, Kleinschmidt, Sterzer, & Giraud, 2005).

The majority of research has focused on how speakers can be recognized from the static, invariant properties of their voices, such as perceived voice quality, and of their faces, such as from shape or configuration. However, speakers also show systematic idiosyncrasies in the realization of phonemes, words, and prosody, and in their speech habits (e.g., lexical and syntactical choices) that should as such be informative about the person’s identity. Indeed, for auditory speech, listeners can learn and recognize talkers from systematic phonetic variation, in the absence of acoustic cues to voice quality (Fellowes, Remez, & Rubin, 1997; Remez, Fellowes, & Rubin, 1997; Sheffert, Pisoni, Fellowes, & Remez, 2002). Artificially created sinewave speech discards the acoustic correlates of voice quality (e.g., fundamental frequency information) and only preserves spectrotemporal information, which still allows for speech recognition (e.g., Remez, Rubin, Pisoni, & Carrell, 1981). The time-varying auditory information contained in sinewave speech is also sufficient for recognizing familiar talkers (Remez et al., 1997). Fine-grained phonetic detail in auditory speech can thus be indexical, indicating that the same set of acoustic characteristics can serve both speech and talker recognition. These findings let to a departure from the long-held view that indexical properties of a talker have their own set of acoustic correlates (e.g., Bricker & Pruzansky, 1976; Hecker, 1971).

Furthermore, listeners can create novel talker representations for unfamiliar speakers solely on the basis of this time-varying information provided by sinewave speech (Sheffert et al., 2002). Importantly, the acquired talker representations are effective in that they allow listeners to recognize speakers from any utterance. Listeners extract and learn abstract properties of the speaker from sinewave speech as, once a speaker is learned, listeners are also able to identify that speaker from new sinewave replicas of speech (Sheffert et al., 2002). Furthermore, having learned to recognize a speaker from sinewave speech transfers to natural speech (Remez et al., 1997; Sheffert et al., 2002), suggesting the accessibility of the same time-varying information in natural speech. In line with this idea is also that the perceptual similarity between unfamiliar talkers in natural speech persists in sinewave replicas (Remez, Fellowes, & Nagel, 2007). Knowledge acquired about a speaker from the dynamic information contained in sinewave speech is therefore also available and used in natural speech. Time-varying attributes of a speaker in auditory speech thus contribute to recognizing familiar speakers and to learning about unfamiliar speakers.

1.2. Recognizing speakers from talking-related motion

Similar to time-varying information in auditory speech, the time-varying information contained in visual speech also contributes to the recognition of speech and of a (familiar) speaker's identity. The equivalent of sinewave replica in the visual modality are point-light displays (PLDs) that preserve kinematic information while eliminating static facial identity cues. To create PLDs of talking faces, the motion of fluorescent dots placed on critical articulators in a speaker's face is tracked in recorded videos and then applied to create animations of a similar set of dots. The resulting videos show a configuration of dots animated with the original motion of the talking face, but do not show the speaker's face. PLDs therefore primarily isolate biological motion, discarding static information (Johansson, 1973). PLDs of the faces of people engaged in communicative interactions can provide sufficient dynamic information to recognize a person’s age (e.g., Berry, 1990) and sex (e.g., Berry, 1991; Hill, Jinno, & Johnston, 2003). Furthermore, PLDs also preserve dynamic information needed to identify the emotions and facial expressions a person was instructed to produce (e.g., an actor who was not engaged in talking was told to portray happiness) (Bassili, 1978, 1979).

Point-light displays of faces producing speech provide, just as sinewave speech, dynamic information that is sufficient and beneficial for speech recognition (Rosenblum, Johnson, & Saldaña, 1996; Rosenblum & Saldaña, 1996). Furthermore, the talking-related facial dynamics preserved in PLDs also inform about the idiosyncratic realization of speech. This indexical information can be temporarily held in short-term memory to match visual speech samples to the same speaker. This dynamic talker information can be obtained from both PLDs and from fully illuminated talking faces. In a matching task, participants, who first saw a fully illuminated talking face, were able to identify which of two subsequently presented PLDs of new speech tokens was produced by the same speaker (Rosenblum, Yakel, Baseer, & Panchal, 2002). Matching was only possible when motion was presented, and best if the frames of these videos were presented in their original order and timing. Furthermore, even in the presence of fully illuminated faces, humans can identify speakers based on their idiosyncratic motion independent of facial form. Participants successfully matched samples to the same speaker based on idiosyncratic motion, even when the motion of all speakers was mapped onto the same avatar (Girges, Spencer, & O'Brien, 2015). Together, these results show that the dynamic information isolated in PLDs is also accessed in the presence of a full face.

1.3. Learning about unfamiliar speakers

Humans can therefore extract, and hold at least temporarily in working memory, visual dynamic signatures of unfamiliar speakers from point-light displays and from fully-illuminated faces. To perform well in a matching task, no speaker representation has to be formed (for a similar argument see Bennetts et al., 2013). In contrast, there is only limited evidence suggesting that information about speakers' visual dynamic signatures of talking is eventually stored in long-term memory as part of representations for familiar speakers. Participants can identify their friends from PLDs of them uttering a sentence, but not from seeing static frames of these PLDs (Rosenblum et al., 2007). The results of this study dovetail with prior work showing that participants can recognize their friends from PLDs showing their faces produce other types of motion (Bruce & Valentine, 1988), such as non-rigid motion related to expressing emotions (e.g., smiling) and rigid head motion (e.g., nodding), as well as from PLDs of body movements (e.g., Cutting & Kozlowski, 1977; Jacobs, Pinto, & Shiffrar, 2004; Loula, Prasad, Harber, & Shiffrar, 2005).

It is unclear whether information about speakers' visual dynamic signatures of talking is stored during the early formation of representations in long-term memory for newly encountered, unfamiliar speakers and whether this information can be sufficient for learning. On the one hand, such storage would be expected, paralleling, as described above, the formation of new speaker representations through access to the dynamic information contained in auditory speech (Sheffert et al., 2002). Corresponding results could thus be expected for visual dynamic talker information, especially as this information is similar to what can