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

be obtained from auditory speech. This similarity is evident in that listeners can match PLDs of talking faces cross-modally with successively presented natural speech samples (Rosenblum, Smith, Nichols, Hale, & Lee, 2006) and sinewave samples (Lachs & Pisoni, 2004b) to the same speaker. Listeners can also link (fully illuminated) talking faces with natural voices (Kamachi, Hill, Lander, & Vatikiotis-Bateson, 2003; Lachs & Pisoni, 2004a), even though the matching of still images of faces to voices seems to be also possible under some circumstances (Krauss, Freyberg, & Morsella, 2002; Mavica & Barenholtz, 2013; Smith, Dunn, Baguley, & Stacey, 2016). The link between the two modalities can thus be established through joint talker-specific dynamics, originating from the articulatory activity that produces both speech signals (e.g., Vatikiotis-Bateson, Munhall, Hirayama, Lee, & Terzopoulos, 1996; von Kriegstein et al., 2008; Yehia, Rubin, & Vatikiotis-Bateson, 1998). This link explains, for example, the benefit observed for recognizing a speaker from auditory speech (and for speech recognition itself) when listeners have previously not only heard but also seen the speaker (von Kriegstein et al., 2008). We therefore predict that seeing the indexical dynamic information in visual speech guides the early formation of talker representations.

On the other hand, the formation of new representations for unfamiliar speakers may not rely on facial dynamical information, though in the long-term this information may be stored for familiar speakers. Different processes and representations seem to underlie the recognition of familiar and unfamiliar speakers, as patient studies and neuroimaging studies suggest a double dissociation between the processing of unfamiliar and of familiar faces (for a review see e.g., Johnston & Edmonds, 2009; Malone, Morris, Kay, & Levin, 1982; Rossion, Schiltz, Robaye, Pirenne, & Crommelinck, 2001; Young, Newcombe, de Haan, Small, & Hay, 1993). Representations formed when encountering unfamiliar speakers could thus differ from the well-established long-term representations of familiar speakers. One option is that these representations differ in the type of information they store: As listeners become more familiar with a speaker, representations may become more complex and include more and/or different information (Bonner et al., 2003; Ellis, Shepherd, & Davies, 2016).

Another option is that the type of representations created for unfamiliar and familiar speakers differ. Functional and neural frameworks of face recognition (Bernstein & Yovel, 2015; Bruce & Young, 1986; Foley, Rippon, Thai, Longe, & Senior, 2012; Haxby et al., 2000; O'Toole et al., 2002) assume separate representations for static and dynamic information about faces (e.g., Bate & Bennetts, 2015; Lander, Humphreys, & Bruce, 2004; Longmore & Tree, 2013; Pitcher, Dilks, Saxe, Triantafyllou, & Kanwisher, 2011; von Kriegstein et al., 2008; cf. Calder & Young, 2005). However, cross-talk between these two types of representations may exist (Turk-Browne, Norman-Haignere, & McCarthy, 2010; Zhang, Tian, Liu, Li, & Lee, 2009). O'Toole and colleagues (Natu and O'Toole, 2011; O'Toole et al., 2002, page 265) have argued that processing of dynamic information in the dorsal pathway does not contribute to the recognition of unfamiliar faces, but only matters for the recognition of faces that are to some extent familiar, if and only if, viewing conditions are poor (e.g., due to blurring). In these difficult situations, recognizing the face of a familiar person can be easier when seeing a talking face than when seeing a static face (Knight & Johnson, 1997; Lander & Bruce, 2000, 2004; Lander et al., 2001; for an overview of recognizing people in motion see Yovel & O'Toole, 2016).

This motion benefit arises to some extent from seeing the face from more viewpoints when it is moving than when it is static (*representation enhancement hypothesis*, O'Toole et al., 2002); but the motion benefit also seems to arise from the dynamic identity information provided by seeing moving faces (Knight & Johnston, 1997; O'Toole et al., 2002). This *supplemental information hypothesis* is supported by three types of evidence: First, seeing talking faces benefits face recognition above and beyond seeing the same videos with the same frames presented in random order (Lander & Bruce, 2000; Lander, Christie, & Bruce, 1999). Second, when the dynamics were altered (e.g., by changing the rhythm, tempo, or direction of the motion (Lander & Bruce, 2000; Lander et al., 1999)), the motion benefit, though not eliminated, decreased. Last, those speakers who had been rated independently to have distinctive motion provided participants with a larger motion advantage for face recognition (Lander & Chuang, 2005). In contrast to this motion benefit for recognizing familiar faces, the evidence as to whether or not seeing (fully illuminated) faces talk can support the learning of face representations is mixed (Bennetts et al., 2013; Bonner et al., 2003; Christie & Bruce, 1998; Lander & Bruce, 2003; Skelton & Hay, 2008), supporting the claim that dynamical representations in the dorsal pathway do not contribute to the recognition of unfamiliar speakers.

A likely alternative explanation for the difficulties of prior studies in finding a motion benefit for learning novel faces could be that those studies were not testing whether dynamic signatures are stored. Rather, these studies assessed whether these dynamic signatures enhance static face representations sufficiently to benefit the later recognition of static faces in the absence of dynamic information. While participants had been presented with static vs. moving faces during exposure, they were always tested later on the recognition of static faces. That is, these studies were trying to find behavioral evidence for the crosstalk between dynamic and static representations and not whether dynamic information was stored. In line with this alternative explanation, learning studies testing participants on moving faces found a motion benefit for unfamiliar faces. However, the contribution of motion in these latter studies is inconclusive as a motion benefit could at least in part be due to an increase in attention to moving faces during training. In contrast, our present study is directly designed to test whether dynamic representations are formed for the recognition of unfamiliar speakers.

1.4. The present study

The question addressed in the present study is whether or not talking-related facial dynamic information already plays a role in the early formation of speaker representations in long-term memory. The goals of the present study were twofold. The first goal was to establish whether listeners can learn to recognize an unfamiliar speaker based solely on the dynamic information that visual speech provides. That is, we tested whether listeners form representations of unfamiliar speakers' dynamic signatures of talking in long-term memory. PLDs of speakers uttering short sentences were created. To ensure that only dynamical information was available and no structural information, we normalized the configuration of the point lights across speakers and created PLDs by mapping the motion of the dot arrays of each speaker onto this average model of all speakers. During training, participants watched these PLDs and were asked to select the name of a speaker from several options. Feedback was given on each trial. Four different tokens of two sentences were shown per speaker and repeated three times across training blocks. Analyses of performance by training block allowed to track the amount of exposure necessary for learning the facial dynamics of unfamiliar speakers. Participants were presented with two speakers in Experiment 1 and with two additional speakers in Experiment 3. We predicted that listeners can learn to identify speakers from visual dynamic information alone. If that is the case, then performance should be above chance level by the end of training. Additional control experiments (Experiments 2a/2b/2c) with static frames selected from the PLDs ruled out the possibility that learning was due to any potentially remaining static information in the PLDs. In Experiment 4, we assessed whether learning can occur when speakers are of the same sex. If listeners learned to identify the speakers, rather than only to recognize their sex, then learning should occur.

The second goal of this study was to examine the nature of learning in participants who had succeeded in learning to recognize the unfamiliar speakers from their talking-related facial dynamics. During a subsequent test phase without feedback, we assessed whether the newly

acquired knowledge that allowed participants to recognize these speakers from facial dynamics generalized to new utterances of the same sentences and to new sentences. If so, then this would reflect the learning of speakers' abstract facial dynamics rather than a token-specific learning of the PLD tokens' surface features. Together, the overarching aim was to test whether listeners are able to form representations of the facial dynamics of unfamiliar speakers.

2. Experiment 1

2.1. Method

2.1.1. Participants

Twenty-four undergraduate students from the University of Massachusetts Amherst participated for course credit (mean age = 19.33 years; four men). All of them were native, monolingual speakers of American English and reported to have no hearing, vision, language, or attention deficit.

2.1.2. Materials

Four short sentences were created (similar to Rosenblum et al., 1996): (A) *The friend played in the shed.*; (B) *The spy picked out the theme.*; (C) *The pig slipped through the fence.*; (D) *The thief shrugged at the pearls*. All sentences had a subject-verb-object structure and the same number of words and syllables. All sentences were meaningful but were semantically weakly constraining. Each sentence contained monosyllabic content words with initial phonemes from a different selection of three of the five visually most salient viseme categories ({p}, {f}, {s}, {sh}, {th}; Massaro, 1998). The selection of these viseme categories and their order of occurrence within a sentence differed non-systematically across sentences.

In this first experiment, we tested for learning under the favorable conditions of two speakers. In a later experiment (Expt. 3), we will test for the learning of more talkers. One male and one female native speaker of American English were recorded with a SONY EVI-HD7V camera and a Shure KSM44A microphone, producing eight tokens of each sentence. Following the method suggested in Thomas and Jordan (2001), 23 3-mm dots of white card paper were attached to the face of each speaker before the recording. Fig. 1 shows the placement of these dots on the face (similar to Rosenblum & Saldaña, 1996). Speakers were illuminated with ultraviolet and halogen lights. Videos were recorded as h.264 at 25 fps (1280 × 720). To verify the correct production of these sentences, audio was also recorded (in mono at 48 kHz with a 16-bit sampling rate). To create PLDs, the motion of the dots was tracked in Adobe After Effects CS5 and this tracking was verified by hand. The tracking data was then used to animate an average dot configuration, which had been generated by averaging the coordinates of the dots in the first frame of one video per speaker across speakers. Using configuration-normalized PLDs eliminated any potential speaker differences in the size and shape of the faces and/or in the relative configuration of dots, thus leaving only dynamic visual information available (see also Hill et al., 2003). The final stimuli were 64 configuration-normalized PLDs, consisting of eight tokens for each of four sentences spoken by each speaker. These 64 PLDs were organized into four sets of 16 PLDs each (2 speakers × 2 sentences × 4 tokens). Each set consisted of four tokens for each of two (A and B or C and D) of the four sentences per speaker. PLDs were presented without sound.

2.1.3. Design and procedure

Participants were tested individually in a sound-attenuated booth. The experiment consisted of a training phase, followed immediately by a test phase. On each training trial, participants saw a PLD of a speaker before choosing one of two displayed names as response. Half of the participants received *Sarah* and *Tyler* as options; the other half received *Megan* and *Ryan*. Upon answering, participants were always shown the correct response along with their own response. On trials with incorrect responses, participants were next required to confirm the feedback by selecting the correct response from the response options. This was not done on trials with correct responses. Subsequently, independent of the accuracy of the participant's original response, the same video was always shown again, but this time with the correct name of the speaker printed underneath. No response was collected; rather, the next trial began automatically. The amount of exposure to each speaker was thus held the same across participants, independent of their performance. Overall, each participant received a total of 48 training trials (3 repetitions × 4 tokens × 2 speakers × 2 sentences), split into three blocks of four PLDs per speaker for each of two sentences. The assignment of PLD set to training was counterbalanced across participants. The order of presentation within each block was random.

During the test phase, participants saw on each trial one PLD of a speaker, before choosing the name of the speaker from two options. No feedback was given here. All participants saw the same 64 PLDs, each presented once. The order of presentation was randomized. Out of these 64 trials, one set of sixteen PLDs was an exact repetition from training (*familiar token, familiar sentence condition*), another set of sixteen PLDs contained new tokens of sentences that were familiar from training (*new token, familiar sentence condition*), and two sets of a total of thirty-two PLDs consisted of new sentences (*new token, new sentence condition*). The assignment of set to condition was counterbalanced across participants.

2.2. Results and discussion

All data from the experiments reported here can be accessed from the Open Science Framework database (https://osf.io/9zgxy/?view_only=36eccaa0e24d47d5b91e544fc735db94here; Jesse, 2018). Fig. 2 shows the distribution of participants' proportion-correct scores by training block. Three participants performed with 0.44 in block 3 slightly below chance (< 0.5) but had performed at a level above 0.5 in at least one previous block (result patterns of all three participants across blocks were: 0.56, 0.63, 0.44; 0.63, 0.50, 0.44; 0.56, 0.56, 0.44). A one-sample *t*-test compared the proportion of correct responses of all participants during the final block of the training phase to chance level performance (0.5). Results show that participants recognized speakers better than what would be expected by chance alone ($M = 0.76$, $SD = 0.17$, $t(23) = 7.48$, $p < .00001$; $D = 1.53$). Participants were

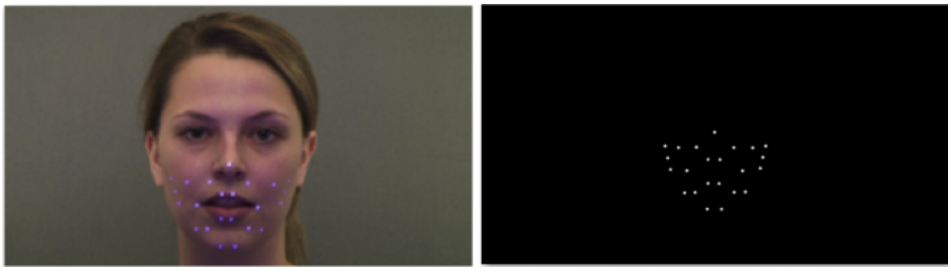

Fig. 1. Sample frame of an original video and its corresponding configuration-normalized point-light display.

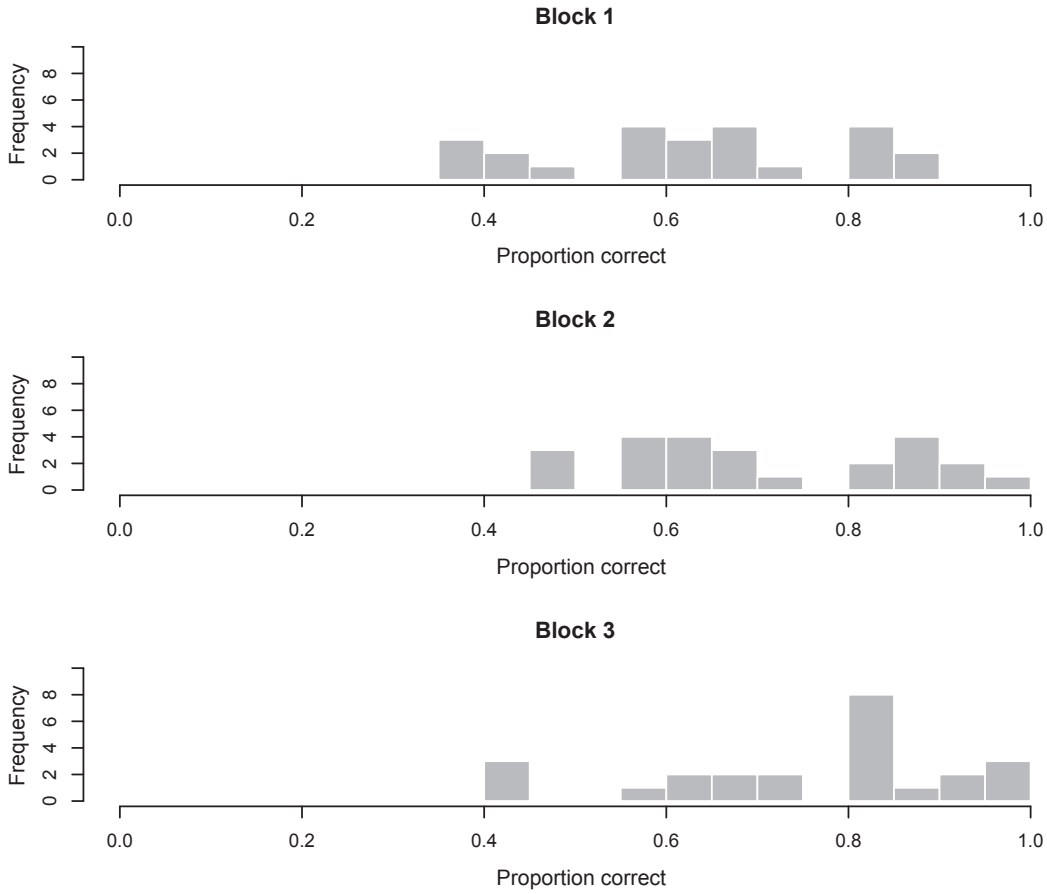

Fig. 2. Histograms showing the distribution of participants' proportion-correct scores by training block in Experiment 1.

thus able to learn to recognize speakers from their talking-related facial dynamics.

To examine the build-up of learning during training, one sample *t*-tests comparing performance to chance level showed learning for block 1 ($M = .63$, $SD = .16$, $t(23) = 4$, $p < .001$; $D = 0.82$) and for block 2 ($M = .71$, $SD = .16$, $t(23) = 6.57$, $p < .00001$; $D = 1.34$). Planned paired two-sample *t*-tests revealed that learning improved from block 1 to block 2 ($t(23) = 2.35$, $p = .02$; $D = 0.51$), but did not significantly increase further in block 3 ($t(23) = 1.69$, $p = .10$; $D = 0.3$). Participants had therefore already learned to recognize the speakers within the first eight trials per speaker but benefitted from further training.

Fig. 3 shows histograms of participants' proportion-correct scores for the three test conditions. The results shown in Fig. 3 suggest that participants were able to recognize speakers also from new materials. Data from the three participants who had performed below the chance level of 0.5 in the third training block were excluded, as generalization of learning can only be tested when there is learning. However, the patterns of the statistical results were not affected by this exclusion. One-sample *t*-tests showed that speaker identification was above chance, no matter whether participants were asked to recognize speakers from the same PLDs as presented during training ($M = 0.76$, $SD = 0.15$, $t(20) = 8.07$, $p < .00001$; $D = 1.76$), from new PLD tokens of the same sentences previously presented during training ($M = 0.68$, $SD = 0.18$, $t(20) = 4.68$, $p < .001$; $D = 1.02$), or from PLDs of new sentences ($M = 0.69$, $SD = 0.14$, $t(20) = 6.16$, $p < .00001$; $D = 1.34$). Planned comparisons across the three test conditions only showed a difference in speaker recognition when the original training PLDs were presented at test compared to when PLDs of completely new sentences were presented ($t(20) = 2.81$, $p < .01$; $D = 0.47$). All other comparisons did not reveal significant results (all $p > .05$). Overall, these results suggest that participants learned to recognize speakers. Participants did not simply learn to recognize speakers from surface detail of the PLDs, but rather formed abstract speaker representation that allowed for speaker recognition even from utterances with new linguistic content.

3. Experiment 2

Experiment 1 provided evidence that listeners can learn to recognize unfamiliar speakers from their facial dynamics. Participants learned to identify two speakers from only seeing the motion they produced while talking. Furthermore, rather than learning low-level features of specific PLD samples, abstract facial dynamic representations were established that allowed participants to recognize speakers at test also from novel speech samples. While PLDs isolate dynamic information, some static information could remain available. Importantly, any speaker differences in the size or shape of the faces and/or in the relative configuration of dots were eliminated by normalizing the PLD configurations across speakers. That is, the same average PLD configuration was animated to follow the motion of each speaker. To further ensure that learning was not driven by spurious static information, in Experiment 2, we tested whether participants could learn to identify the speakers from still frames taken from the PLDs. In Experiment 2a, participants were presented with one still frame randomly selected from each PLD, shown for the same duration as the PLD it had been taken from. Such a one-frame static condition has often sufficed as a control for experiments where faces had been recorded from the same viewpoint (e.g., Bennetts et al., 2013; Knight & Johnston, 1997; Lander & Davies, 2007; Roark, O'Toole, Abdi, & Barrett, 2006), as is the case here. In addition, we also conducted two multistatic experiments with three and five randomly-selected frames,

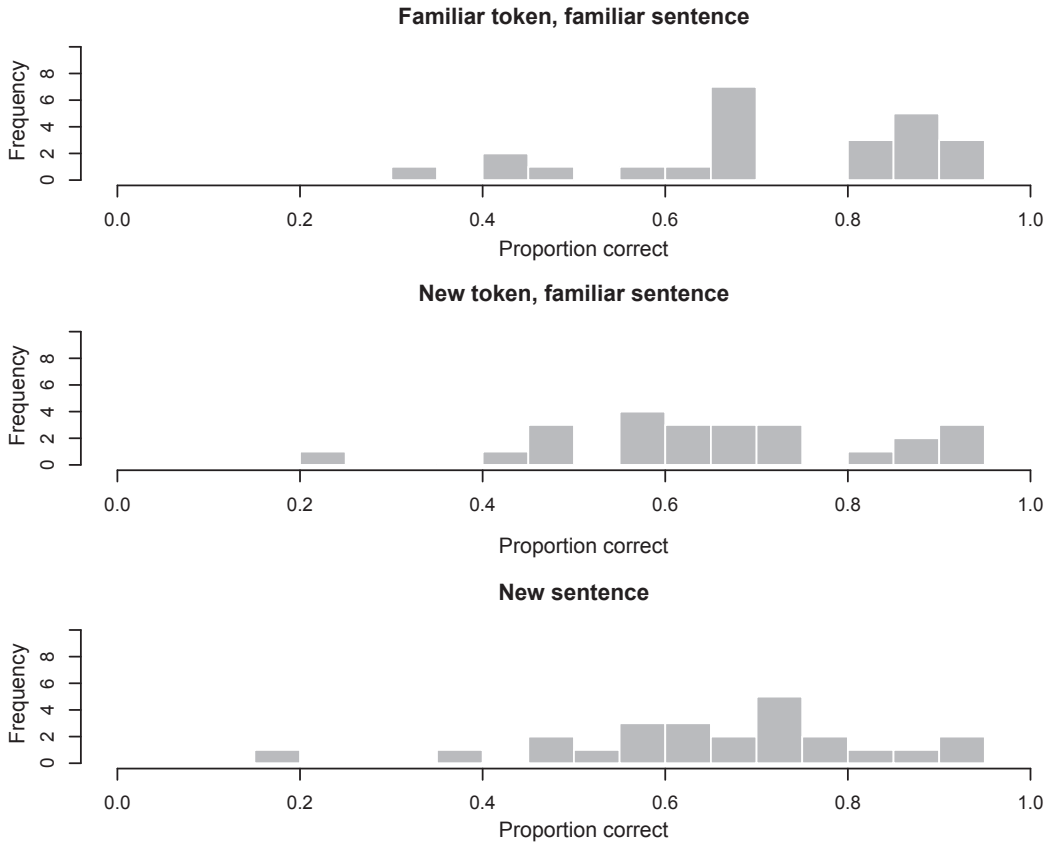

Fig. 3. Histograms showing the distribution of participants' proportion-correct scores for each test condition in Experiment 1. Participants were tested on exact repetitions of training sentences (familiar token, familiar sentence), on new tokens of the training sentences (new token, familiar sentence), and on new sentences.

respectively, to provide participants with potentially more static information (Bonner et al., 2003; Bruce & Valentine, 1988; Christie & Bruce, 1998; Lander & Bruce, 2003; Loula et al., 2005; Skelton & Hay, 2008). These frames were shown sequentially, in total for the same duration as the PLDs they came from. Learning should be decreased in all of these static conditions to the extent that participants had relied on dynamic information in Experiment 1. If participants had, however, relied on static information in Experiment 1, and/or on durational differences in the PLDs of the different speakers, then learning should still be found with exposure to stills in Experiment 2.

3.1. Method

3.1.1. Participants

Twenty-four new participants (Experiment 2a: mean age = 19.54 years, 3 men; Experiment 2b: mean age = 19.58 years, 3 men; Experiment 2c: mean age = 19.87 years, 3 men) took part in each of the three versions of Experiment 2, sampled from the same population as done for Experiment 1. None of them had participated in Experiment 1.

3.1.2. Materials

The materials were the same as for Experiment 1. One, three, or five still frames were randomly selected from each PLD.

3.1.3. Design and procedure

The same design and procedure was used as in Experiment 1. However, still frames were shown in place of the PLD videos they had been taken from. Participants were informed that these stills had been selected from PLDs of the speakers. In the multistatic versions, stills were presented in succession. Stills within each trial were separated by a black screen shown for one second, which is sufficient to prevent the perception of apparent motion across the still frames (Thornton, Pinto, & Shiffrar, 1998). Stills were shown for the same (cumulative) duration as the original PLD.

3.2. Results and discussion

For each experiment, one-sample *t*-tests compared the performance of all participants during each block at training, and in each of the three conditions at test, to chance level performance (0.5). Tables 1 and 2 provide the results. In summary, no evidence of learning was found in any of the conditions in any of the three experiments.

4. Experiment 3

Experiment 2 corroborates the finding of Experiment 1, that participants learned to identify speakers from their talking-related visual

Table 1
Mean proportions and standard deviations (*SD*) of correct responses during the three blocks of the training phase in Experiment 2 and statistical comparisons to chance level performance.

Experiment	Training block	Mean (*SD*)	*t*(23)	*p*	Cohen's D
2a (1 still)	1	0.48 (0.16)	−0.62	.54	0.13
	2	0.54 (0.14)	1.44	.16	0.29
	3	0.53 (0.13)	0.96	.35	0.2
2b (3 stills)	1	0.48 (0.11)	−0.67	.51	0.14
	2	0.54 (0.14)	1.43	.17	0.29
	3	0.51 (0.10)	0.38	.71	0.08
2c (5 stills)	1	0.51 (0.11)	0.45	.66	0.09
	2	0.53 (0.12)	1.03	.31	0.1
	3	0.53 (0.10)	1.37	.18	0.3

Table 2
Mean proportions and standard deviations (*SD*) of correct responses in the three conditions of the test phase in Experiment 2 and statistical comparisons to chance level performance.

Experiment	Condition	Mean (*SD*)	*t*(23)	*p*	Cohen's *D*
2a (1 still)	Familiar token, familiar sentence	0.54 (0.14)	1.51	.14	0.31
	New token, familiar sentence	0.51 (0.11)	0.22	.82	0.05
	New sentence	0.49 (0.09)	−0.43	.67	0.09
2b (3 stills)	Familiar token, familiar sentence	0.55 (0.13)	1.86	.08	0.4
	New token, familiar sentence	0.51 (0.13)	0.5	.62	0.10
	New sentence	0.48 (0.07)	−1.53	.14	0.31
2c (5 stills)	Familiar token, familiar sentence	0.51 (0.13)	0.48	.64	0.1
	New token, familiar sentence	0.43 (0.12)	−2.78	.01[a]	0.57
	New sentence	0.51 (0.09)	0.63	.54	0.13

[a] Effect is not in the predicted direction.

dynamic signatures. Static information had not driven the learning effects observed in Experiment 1. In Experiment 3, we tested whether participants could also form speaker representations of talking-related facial dynamics when encountering four unfamiliar speakers. Furthermore, we examined the kinematic profiles of these four speakers and the dynamic cues that may guide learning.

4.1. Method

4.1.1. Participants

Twenty-four participants (mean age = 19.79 years; six men), sampled from the same population as for the other experiments, completed Experiment 3. None of them had participated in Experiments 1 or 2.

4.1.2. Materials

The materials were the same as for Experiment 1. One more male and one more female native speaker of American English had been recorded along with the speakers for Experiment 1 and 2. Their videos were edited in the same fashion as for Experiment 1.

4.1.3. Design and procedure

The same design and procedure was used as in Experiment 1. However, in Experiment 3, participants identified four speakers. Training phase and test phase were therefore overall twice as long as in Experiment 1, but the number of trials presented per speaker was the same across experiments. The response options were always *Sarah*, *Megan*, *Tyler*, and *Ryan*. The assignment of a name to a speaker of the same gender was counterbalanced across participants.

4.2. Results and discussion

Fig. 4 shows histograms of participants' distribution of proportion-correct scores for each training block. The results depicted in Fig. 4 show that all but two participants (0.12 and 0.16) performed above chance level (0.25) in block 3. One of these two participants never demonstrated learning (proportion correct across blocks was 0.22, 0.22, and 0.12), while the other participant had shown learning in block 1 (0.34) and in block 2 (0.44) before performing poorly in block 3 (0.16). One-sample *t*-tests comparing the performance of all participants during each training block to chance level performance (0.25) suggests that learning already took place during the first block of training ($M = 0.3$; $SD = 0.07$, $t(23) = 3.03$, $p < .01$; $D = 0.61$) and was found at each subsequent block (block 2: $M = 0.32$, $SD = 0.1$, $t(23) = 3.47$, $p < .01$; $D = 0.71$; block 3: $M = 0.36$, $SD = 0.11$, $t(23) = 5.57$, $p < .0001$; $D = 1.14$). Learning did not improve between block 1 and block 2 ($t(23) = 1.08$, $p = .29$; $D = 0.3$) or between block 2 and block 3 ($t(23) = 1.47$, $p = .16$; $D = 0.43$). Participants were therefore able to learn to recognize four speakers from their facial dynamics with very little exposure.

Next, we examined the kinematic profile of each speaker. Absolute velocity (speed) and acceleration were computed for each dot for each frame. Total distance was calculated as the integral over the speed of each dot over frames. Displacement was calculated as the radial distance between the position of each dot on any given frame compared to its position in the standard facial configuration shown at the beginning of each PLD. The maximum, mean, and standard deviation per PLD video were calculated for each measure (see Table 3). As these characteristics are by definition related to each other, high inter-correlations were found. We therefore conducted a principal component analysis (PCA) to identify which measures form independent composites that best explain the variance in the proportion of correct responses given in block 3 of training. All assumptions of the PCA were met. The Kaiser-Meyer-Olkin measure of sampling adequacy was .77, above the commonly recommended value of at least .60 (Tabachnick & Fidell, 2012). Bartlett's test of sphericity was significant ($\chi^2(66) = 2698.6$, $p < .00001$). The first two components were selected, as they had eigenvalues greater than 1 (Kaiser's criterion; Costello & Osborne, 2005). Together, these two components explained 84.90% of the total variance in the accuracy scores (66.94% and 17.95%, respectively). The first factor captured the measures of velocity, acceleration, and total distance. The second factor captured the displacement measures. A best-subset regression analysis using the leaps package (Lumley & Miller, 2004) within in the R computing program (R Core Team, 2016) with these two factors identified a model with only the first factor ($\beta = 2.93$) as the best model (Adjusted $R^2 = .143$; $F(1,126) = 22.2$, $p < .00001$). All assumptions of linear regression were met (checked using the gvlma package (Edsel & Slate, 2014). Together, these results confirm that participants relied in their recognition of the speakers on their kinematic profiles related to velocity, acceleration, and total distance.

Fig. 5 shows the distribution of participants' proportion correct scores by test condition in Experiment 3. We excluded data from the two participants who had proportion-correct scores below chance (0.25) in the third block of the training phase from all analyses. Whether or not data from these two participants was included did, however, not change the pattern of the statistical results. During test, participants recognized speakers better than chance from familiar PLDs ($M = 0.35$, $SD = 0.12$, $t(21) = 3.71$, $p < .01$; $D = 0.79$), from new PLDs of familiar training sentences ($M = 0.31$, $SD = 0.11$, $t(21) = 2.78$, $p = .01$; $D = 0.59$), and from PLDs of new sentences ($M = 0.31$, $SD = 0.08$, $t(21) = 3.38$, $p < .01$; $D = 0.72$). There was no difference in performance as a function of whether a PLD was familiar, a new version of a training sentence, or a new sentence (all $p > .05$). Participants thus learned to recognize four speakers from their facial dynamics, and even from new linguistic material.

5. Experiment 4

The experiments presented so far provide strong evidence that listeners can learn to recognize unfamiliar speakers from their talking-related facial dynamics. Prior work has shown, however, that the sex of

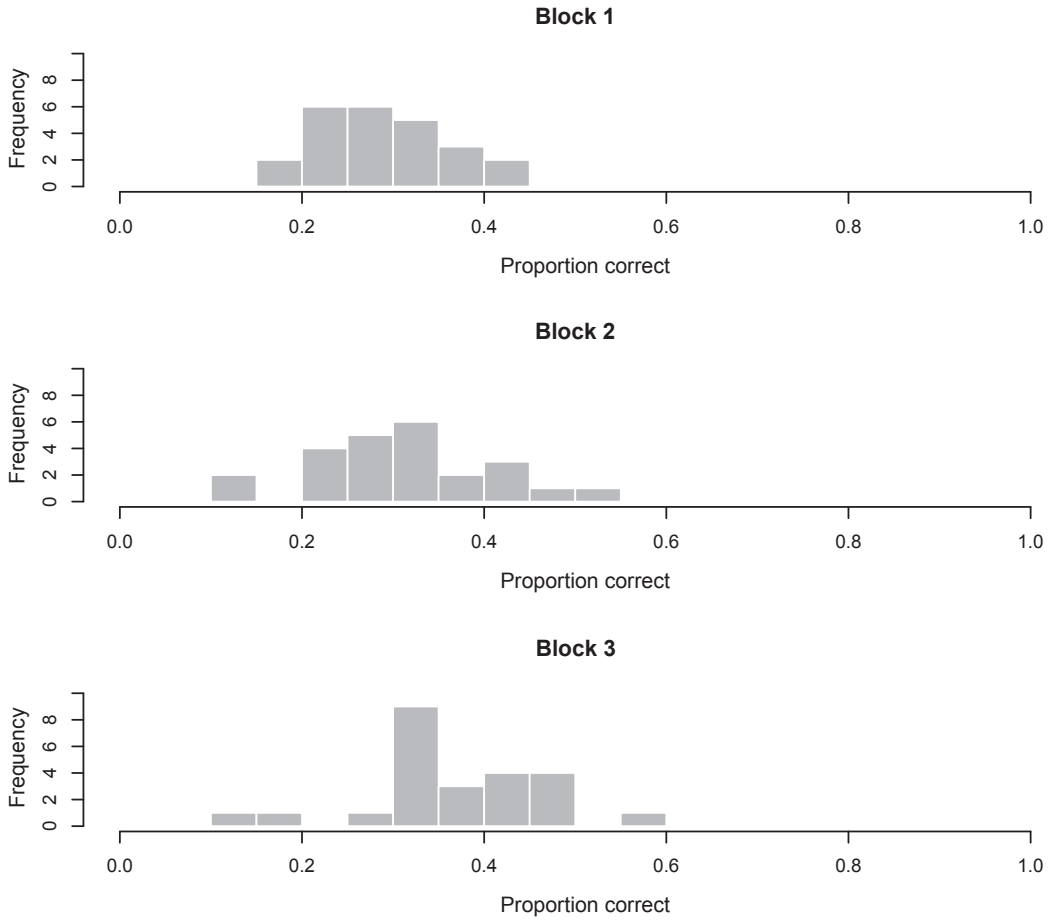

Fig. 4. Histograms showing the frequency distribution of participants' proportion-correct scores by training blocks in Experiment 3.

the speaker is detectable from dynamic talking-related motion (Berry, 1990, 1991; Hill et al., 2003). In Experiment 1, the two speakers were of different sex. To the extent that sex differences may exist in the facial dynamics of talking, listeners could have therefore learned to identify the sex of a speaker rather their identity. Even in Experiment 3, where participants learned about two male and two female speakers, cues to the sex of the speakers could have contributed. It is thus possible, that cues to the sex of the speakers could have at least contributed to, if not driven, the results. In Experiment 4, we hence tested whether participants can learn the identity of a speaker from their idiosyncratic visual speech dynamics, even when speakers are of the same sex.

5.1. Method

5.1.1. Participants

Twenty-four new participants (mean age = 19.5 years; four men) came from the same population as sampled from for the other experiments. None of them had participated in any of the other experiments.

5.1.2. Materials

Materials were taken from the two female speakers in Experiment 3.

5.1.3. Design and procedure

The same design and procedure was used as in Experiment 1. However, two female speakers were shown and only the two female speaker names (*Megan*, *Sarah*) were used as response options.

5.2. Results and discussion

Histograms of participants' distribution of proportion-correct scores for each training block are depicted in Fig. 6. The results show that all but four participants (0.31, 0.38, 0.38, and 0.44) performed above chance level (0.5) in block 3. Only one of these participants never demonstrated learning in an earlier block (max. 0.50). The other three participants showed above chance performance on prior blocks (max. 0.81). One-sample *t*-tests compared the performance of all participants during each training block to chance level performance (0.5). These tests showed that learning had occurred during the first block of

Table 3
Mean, standard deviation (SD), and maximum (Max) of acceleration, velocity, total distance, and displacement for the four speakers, averaged across tokens.

Speaker	Acceleration			Total Distance			Velocity			Displacement		
	M	SD	Max	M	SD	Max	M	SD	Max	M	SD	Max
Male 1	1.16	1.19	9.00	65.92	38.03	149.24	1.64	1.73	13.25	5.71	3.81	23.61
Male 2	1.15	1.11	8.58	66.88	32.56	142.93	1.64	1.55	11.73	7.32	4.37	24.83
Female 1	1.44	1.43	10.89	92.84	40.61	190.15	2.21	2.03	15.72	11.79	5.98	32.85
Female 2	1.05	1.15	9.52	70.52	40.14	172.27	1.53	1.70	14.26	6.8	4.16	25.43

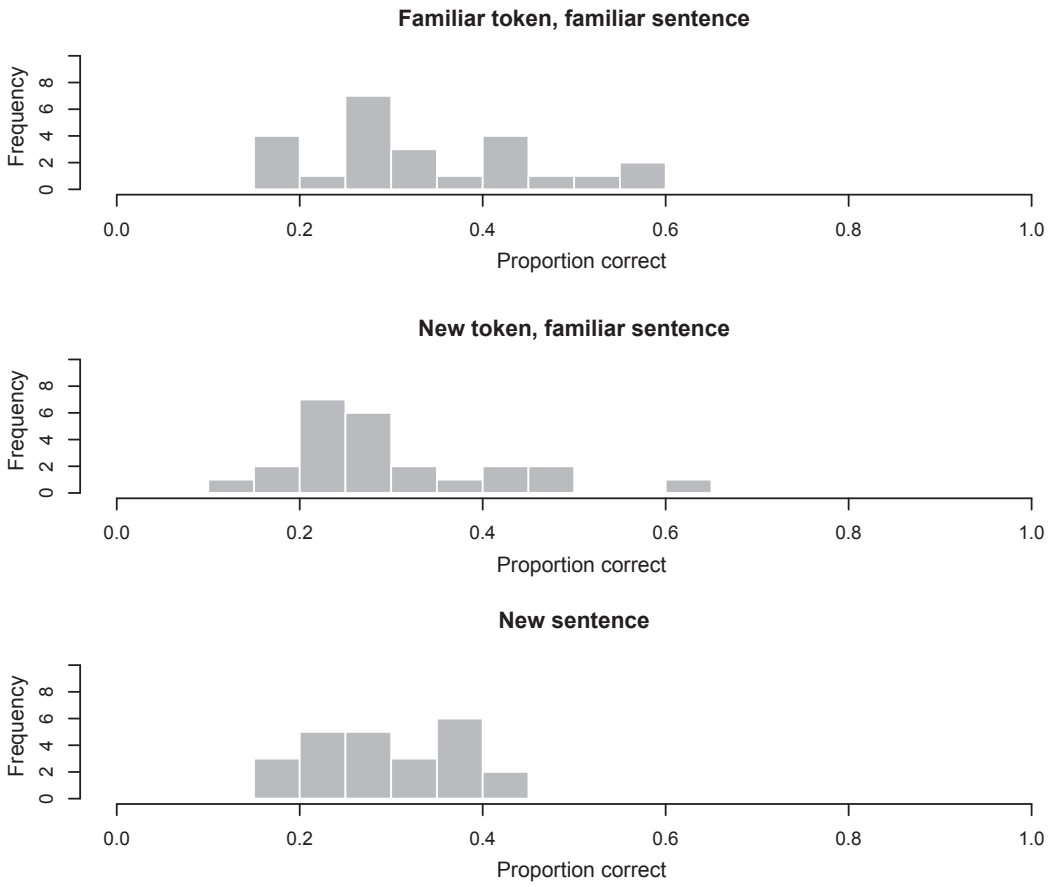

Fig. 5. Histogram showing the distribution of participants' proportion-correct scores for each of the three test conditions in Experiment 3. Participants were tested on the training sentences (familiar token, familiar sentence), on new tokens of the familiar training sentences (new token, familiar sentence), and on new sentences.

training ($M = 0.58$; $SD = 0.12$, $t(23) = 3.47$, $p < .01$; $D = 0.71$). Furthermore, learning was evident at each subsequent block (block 2: $M = 0.68$, $SD = 0.18$, $t(23) = 5.08$, $p < .0001$; $D = 1.04$; block 3: $M = 0.67$, $SD = 0.19$, $t(23) = 4.35$, $p < .001$; $D = 0.89$). Learning improved between block 1 and block 2 ($t(23) = 2.76$, $p = .01$; $D = 0.68$), but not further between block 2 and block 3 ($t(23) = 0.33$, $p = .74$; $D = 0.07$). Participants were thus able to learn to recognize speakers, without the aid of any potential cues about their sex. The build-up of learning here shows the same pattern as found in Experiment 2 for the learning of speakers of different sex.

The distribution of participants' proportion correct scores by test condition in Experiment 4 is shown in Fig. 7. We excluded data from the four participants who had accuracy scores below chance (0.5) in the third block of the training phase. This exclusion did, however, not affect the pattern of the statistical results. One-sample t-tests revealed that participants identified speakers better than at chance level from the same PLDs as presented during training ($M = 0.73$, $SD = 0.18$, $t(19) = 5.8$, $p < .0001$; $D = 1.3$), from new PLDs of the same sentences previously presented during training ($M = 0.73$, $SD = 0.15$, $t(19) = 6.75$, $p < .00001$; $D = 1.51$), and from PLDs of new sentences ($M = 0.66$, $SD = 0.13$, $t(19) = 6.16$, $p < .0001$; $D = 1.28$). Planned comparisons across test conditions showed that speaker recognition only differed when presented with new sentences vs. new tokens of the familiar training sentences ($t(19) = 2.55$, $p = .02$; $D = 0.48$). Together, these results suggest again that participants did not learn to recognize the identity of speakers from surface detail of the PLDs, but rather formed abstract speaker representation that allowed for speaker recognition even from utterances with new linguistic content. Participants can therefore learn the individual identity of speakers from their talking-related motion and do not need potential cues to the sex of a talker to contribute to this learning.

6. General discussion

The aim of this study was to test whether listeners can establish representations of unfamiliar speakers' facial dynamic signatures of talking. The results of all experiments provide strong supporting evidence, as the talking-related dynamic information provided by the configuration-normalized PLDs of speakers alone was sufficient for participants to learn to recognize these novel speakers. Participants were able to learn about two and about four speakers simultaneously. Effect sizes for comparisons of recognition levels to chance level performance were large. Point-light displays discard static information and only preserve dynamic information. Our conclusion that participants relied on facial dynamics in their learning is supported in three ways. First of all, the same average point-light display configuration had been animated for all speakers, eliminating any differences in the size and shape of the faces of the speakers as well as any potential differences in the placement of the dots across speakers. Secondly, in control experiments, participants did not learn when presented only with static frames taken from the PLDs. These control experiments suggest that it is unlikely that any potentially remaining static information in the PLDs could have driven the effects and exclude the possibility that durational differences between the PLDs of the speakers alone are responsible for learning. Thirdly, a regression analysis showed that listeners learned to recognize talkers based on their kinematic profile in relation to velocity, acceleration, and total distance. Talking-related dynamic identity information is therefore stored for newly encountered unfamiliar speakers and allows for their recognition.

Furthermore, participants indeed stored *identity* representations for

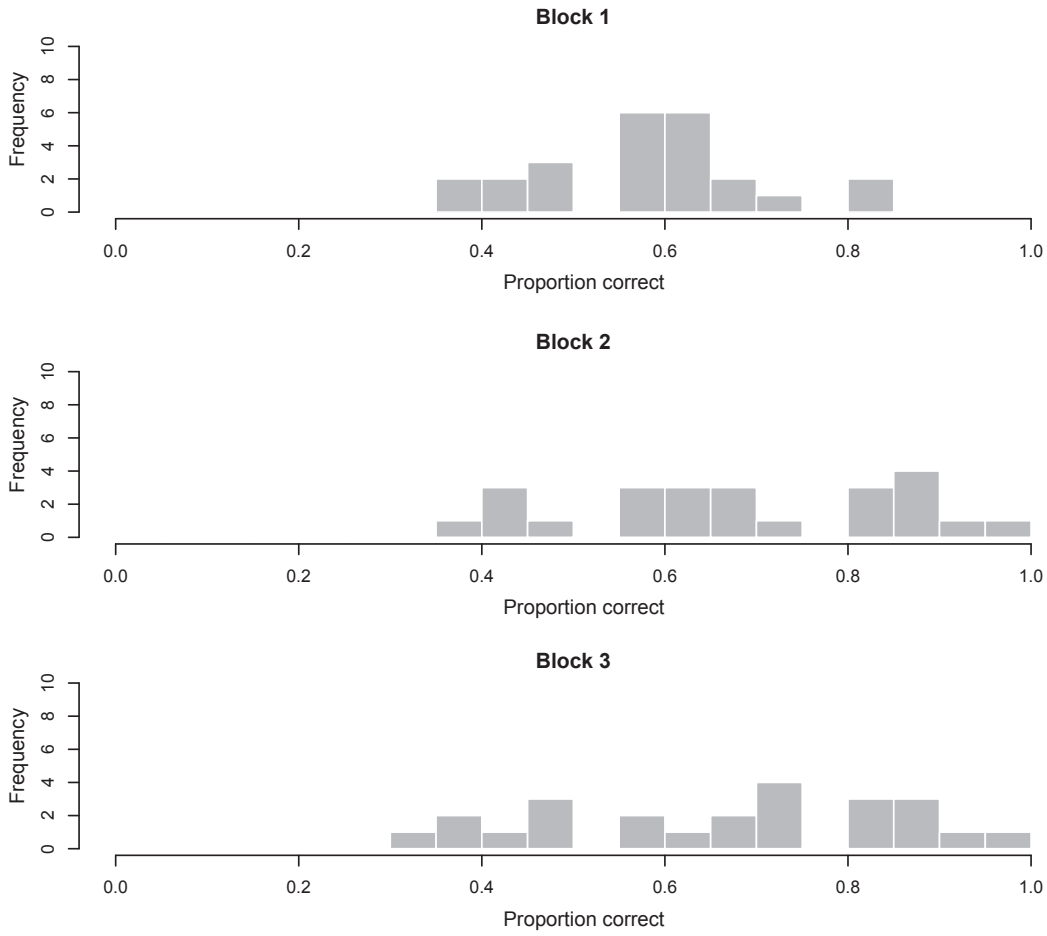

Fig. 6. Histograms showing the frequency distribution of participants' proportion-correct scores by training blocks in Experiment 4.

these unfamiliar talkers. As speakers of different sex were presented in Experiments 1 and 3, dynamic cues to sex differences (Berry, 1991; Hill et al., 2003) could have driven, or at least contributed to, the learning. However, this possibility was ruled out by showing that learning was still found in Experiment 4 with two speakers of the same sex. Participants thus indeed learned to recognize the individual talkers, not just their sex. While not necessary for successful learning, it remains, however, possible that sex differences can further help with this identification when indeed encountering talkers of different sex.

Importantly, listeners did not learn the surface details of the specific PLD tokens presented during training, but rather formed abstract representations of facial dynamics, that allowed for the recognition of speakers even from new utterances. During test, participants recognized speakers from PLDs of the same tokens and new tokens of the training materials as well as from PLDs of entirely new sentences, suggesting that abstract representations were acquired. Participants were better at recognizing speakers from training sentences rather than new sentences only in Experiments 2 and 4, when they were learning to identify two speakers. When simultaneously learning four speakers, learning was fully generalized to new materials. Our experiments cannot currently speak as to why this difference may emerge. One possibility is that generalization occurs when the dynamics are processed more deeply because finer distinctions between talkers have to be made (see also Loebach, Bent, & Pisoni, 2008), as it is the case for four compared to two speakers. Further experimentation is needed to determine the circumstances under which a better abstraction of dynamic signatures, and hence full generalization, can be achieved. That the representations acquired about the dynamic signatures of speakers are abstract in nature in that - once learned - speakers can be recognized even from utterances with new linguistic content, parallels the abstract nature of representations for auditory voice recognition (Legge, Grosmann, & Pieper, 1984; Nygaard & Pisoni, 1998; Sheffert & Olson, 2004; Sheffert et al., 2002; Zäske, Volberg, Kovács, & Schweinberger, 2014). The ability to establish abstract representations of facial dynamics could have been mediated by visual exposure to a variety of phonemes during training. The new materials at test showed a large amount of phonemic overlap with the training materials. To the extent that idiosyncrasies in the dynamic realization of phonemes persisted across these different contexts, this segmental overlap between training and test materials could have allowed for, or at least contributed to, the generalization of learning to new materials. It is also possible that idiosyncrasies arising at other levels of speech production (e.g., idiosyncrasies in encoding prosodic structure in articulatory and head movement) persisted across materials, thus contributing to generalization. Lexical or syllabic overlap was minimal (only the preposition "*the*" was repeated across sentences), and thus unlikely to have contributed much to transfer.

Last, learning of facial dynamic identity information for two and for four speakers was evident after very limited exposure. In all three experiments, participants showed learning already in block 1, that is, learning occurred from less than eight presentations of a speaker (i.e., from less than four tokens of two sentences per speaker). Furthermore, only learning about two (but not four) speakers improved with further training in block 2. The current results thus convincingly show that learning about the dynamic signature of a speaker occurs and is already completed after limited exposure (Lander & Davies, 2007; Zäske et al., 2014).

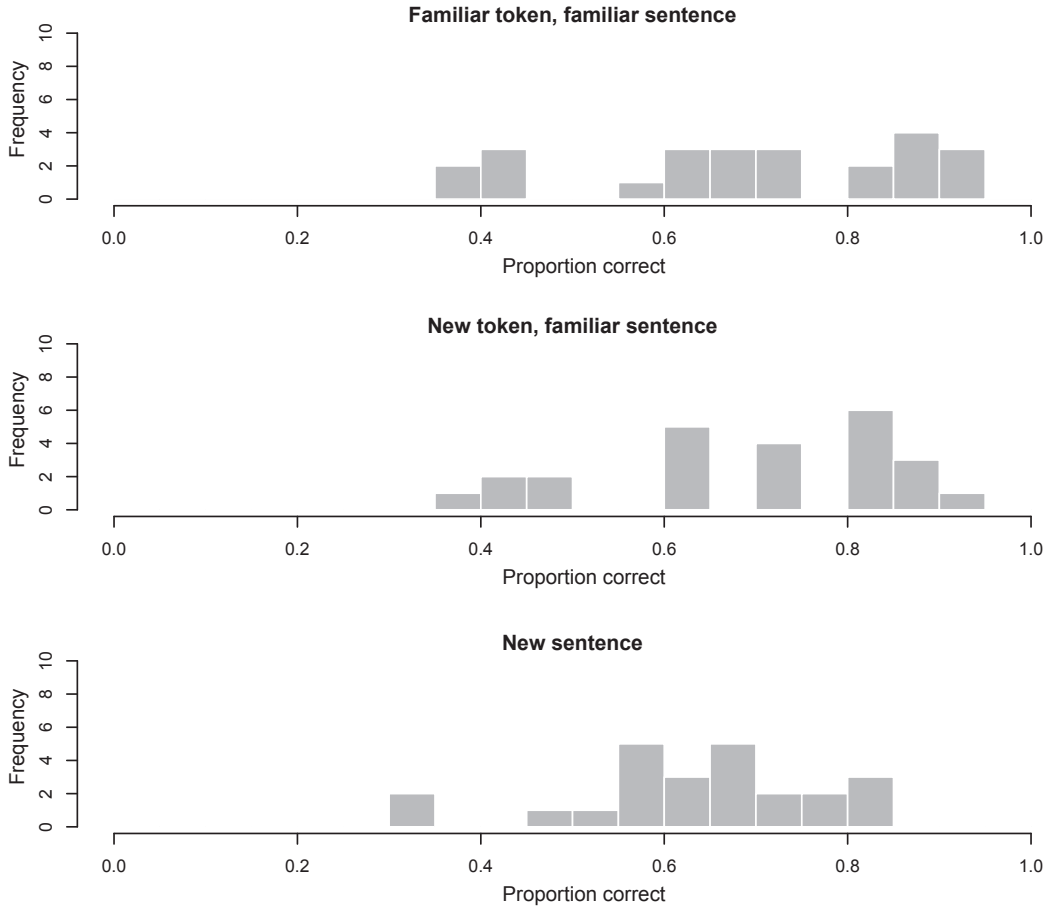

Fig. 7. Histograms showing the distribution of participants' proportion-correct scores for each test condition in Experiment 4. Participants were tested on exact repetitions of training sentences (familiar token, familiar sentence), on new tokens of the training sentences (new token, familiar sentence), and on new sentences.

6.1. The storage of facial dynamic information for speakers

The results of the present study suggest that abstract representations can be formed for unfamiliar speakers based on seeing their dynamical signature of talking and that these representations contribute to person recognition. Such dynamic visual representations have been postulated in functional and neural frameworks of face recognition (Bernstein & Yovel, 2015; Bruce & Young, 1986; Foley et al., 2012; Haxby et al., 2000; O'Toole et al., 2002) to be separate from representations of static identity information about faces (e.g., Bate & Bennetts, 2015; Lander et al., 2004; Longmore & Tree, 2013; Pitcher et al., 2011; von Kriegstein et al., 2008; cf. Calder & Young, 2005). For example, in O'Toole et al.'s (Natu & O'Toole, 2011; O'Toole et al., 2002) modification of Haxby et al.'s (2000) neural framework for face perception, static information about the face (e.g., its structure, size) is processed along a ventral stream leading to the static face representations in the face-responsive area in the fusiform gyrus (fusiform "face" area; Kanwisher, McDermott, & Chun, 1997; cf. Schultz & Pilz, 2009), whereas dynamic facial information (from speaking, expressing, etc.) is processed along a dorsal stream, passing information through motion processing areas to dynamic face representations in the superior temporal sulcus (STS). The STS is a functionally diverse area that seems to be primarily implicated in the processing of auditory and visual information in social tasks (Allison, Puce, & McCarthy, 2000; Hein & Knight, 2008; Lahnakoski et al., 2012; Redcay, 2008; Watson, Latinus, Charest, Crabbe, & Belin, 2014), such as in speech perception and audiovisual binding (e.g., Callan et al., 2003; Calvert & Campbell, 2003; Calvert, Campbell, & Brammer, 2000; Calvert et al., 1999; Macaluso, George, Dolan, Spence, & Driver, 2004; Riedel, Ragert, Schelinski, Kiebel, & von Kriegstein, 2015; Sekiyama, Kanno, Miura, & Sugita, 2003; Wright, Pelphrey, Allison, McKeown, & McCarthy, 2003), biological motion processing (e.g., Beauchamp, Lee, Haxby, & Martin, 2003; Bonda, Petrides, Ostry, & Evans, 1996; Fox, Iaria, & Barton, 2009; Grossman et al., 2000; Howard et al., 1996), and voice processing (Andics et al., 2010; Belin & Zatorre, 2003; Belin, Zatorre, & Ahad, 2002; Belin, Zatorre, Lafaille, Ahad, & Pike, 2000; Grossmann, Oberecker, Koch, & Friederici, 2010; Nakamura et al., 2001; Shultz, Vouloumanos, & Pelphrey, 2012; von Kriegstein, Eger, Kleinschmidt, & Giraud, 2003). Perhaps most importantly, for familiar speakers, areas of the STS that are sensitive to facial motion are causally related to auditory speech recognition (Riedel et al., 2015).

As discussed in the Introduction, the current understanding has been that the processing of the dynamic information in the dorsal pathway only contributes to the recognition of familiar faces, if and only if viewing conditions are poor, but not to the recognition of unfamiliar faces (Natu & O'Toole, 2011; O'Toole et al., 2002 page 265). In these difficult situations, moving faces might provide an advantage as they can provide different viewpoints that may lead to an enhanced representation (O'Toole et al., 2002) but also as the additional dynamic information can supplement the then more limited static information (Knight & Johnston, 1997; O'Toole et al., 2002).

In contrast, studies addressing whether seeing speaking boosts the processing or learning of *static* faces of unfamiliar speakers failed to provide unequivocal evidence for a role of facial dynamic information (Bennetts et al., 2013; Bonner et al., 2003; Christie & Bruce, 1998; Lander & Bruce, 2003; Skelton & Hay, 2008). These mixed results were interpreted as supporting the claim that dynamical representations in the dorsal pathway do not contribute to the recognition of unfamiliar

speakers. The results of the present study provide, however, strong evidence that representations of talking-related dynamic facial information can be rapidly formed for unfamiliar speakers (see also von Kriegstein et al., 2008). As we had proposed in the Introduction, these prior studies were testing whether dynamic signatures enhance static face representations sufficiently to benefit the later recognition of static faces in the absence of dynamic information, and not whether dynamic signatures themselves are stored. By testing participants' learning only with static faces, these studies tried to provide behavioral evidence for the crosstalk between dynamic and static representations, that is postulated by functional and neural frameworks of face recognition (Bernstein & Yovel, 2015; Bruce & Young, 1986; Foley et al., 2012; Haxby et al., 2000; O'Toole et al., 2002) and supported by neuroimaging work (Turk-Browne et al., 2010; Zhang et al., 2009). The few studies that assessed participants' learning on moving faces found a motion benefit for unfamiliar faces (Lander & Davies, 2007; Roark et al., 2006). Nevertheless, these latter studies cannot speak to the contribution of motion to the learning of unfamiliar faces. The motion benefit could at least in part be due to an increase in attention to moving faces during training.

In comparison, our study was directly designed to test whether dynamic representations are formed for the recognition of unfamiliar speakers. And indeed, our results provide strong evidence that early representations for unfamiliar speakers can be formed on their talking-related dynamic signatures alone, and already from very little exposure. Dynamic representations built based on the idiosyncratic realization of visual speech can therefore aid the recognition of newly encountered speakers. One limitation of our study is however that dynamic information was presented in isolation by using PLDs. Future studies thus have to show whether representations of facial talking-related dynamic signatures are also acquired from full faces. This result would be predicted based on prior work showing that viewers extract the same talking-related dynamic information that is isolated in PLDs also from fully illuminated faces (Rosenblum et al., 2002). Furthermore, even when the same avatar face is shown for all speakers, participants can still successfully match samples to the same speaker based on their distinct motion (Girges et al., 2015). Last, listeners can identify their friends from PLDs showing them uttering a sentence (Rosenblum et al., 2007). The knowledge guiding this recognition had to come from fully illuminated faces in everyday social interactions.

7. Conclusions

In conclusion, our findings show that dynamic facial signatures of talking provide sufficient information to rapidly build abstract identity representations for unfamiliar speakers that allow the recognition of these speakers, even from utterances with new linguistic content. As we meet new speakers face to face, we create representations of their facial dynamic signatures.

Acknowledgements

We thank Katharina von Kriegstein for helpful comments on an earlier version of this manuscript and Emilee Bates, Chantel Brennan, Deanna Ferrante, Sarah Hammond, Melissa Karp, Smriti Karwa, Emma Leahey, and Bernadette Ojukwu for their help with the data collection. Experiment 3 was part of the second author's undergraduate thesis under the supervision of the first author. Parts of this work were presented at the 57th Annual Meeting of the Psychonomic Society, Boston (Jesse & Bartoli, 2016).

References

Allison, T., Puce, A., & McCarthy, G. (2000). Social perception from visual cues: Role of the STS region. *Trends in Cognitive Sciences, 4*(7), 267–278. http://dx.doi.org/10.1016/S1364-6613(00)01501-1.

Andics, A., McQueen, J. M., Petersson, K. M., Gál, V., Rudas, G., & Vidnyánszky, Z. (2010). Neural mechanisms for voice recognition. *NeuroImage, 52*(4), 1528–1540. http://dx.doi.org/10.1016/j.neuroimage.2010.05.048.

Barsics, C. (2014). Person recognition is easier from faces than from voices. *Psychologica Belgica, 54*(3), 244–254. http://dx.doi.org/10.5334/pb.ap.

Barsics, C., & Brédart, S. (2012). Access to semantic and episodic information from faces and voices: Does distinctiveness matter? *Journal of Cognitive Psychology, 24*(7), 789–795. http://dx.doi.org/10.1080/20445911.2012.692672.

Bassili, J. N. (1978). Facial motion in the perception of faces and of emotional expression. *Journal of Experimental Psychology: Human Perception and Performance, 4*(3), 373–379.

Bassili, J. N. (1979). Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face. *Journal of Personality and Social Psychology, 37*(11), 2049–2058.

Bate, S., & Bennetts, R. J. (2015). The independence of expression and identity in face-processing: Evidence from neuropsychological case studies. *Frontiers in Psychology, 6*. http://dx.doi.org/10.3389/fpsyg.2015.00770 pp. 427-427.

Beauchamp, M. S., Lee, K. E., Haxby, J. V., & Martin, A. (2003). FMRI responses to video and point-light displays of moving humans and manipulable objects. *Journal of Cognitive Neuroscience, 15*(7), 991–1001. http://dx.doi.org/10.1162/089892903770007380.

Belin, P., Bestelmeyer, P. E. G., Latinus, M., & Watson, R. (2011). Understanding voice perception. *British Journal of Psychology, 102*(4), 711–725. http://dx.doi.org/10.1111/j.2044-8295.2011.02041.x.

Belin, P., & Zatorre, R. J. (2003). Adaptation to speaker's voice in right anterior temporal lobe. *Neuroreport, 14*(16), 2105–2109. http://dx.doi.org/10.1097/01.wnr.0000091689.94870.85.

Belin, P., Zatorre, R. J., & Ahad, P. (2002). Human temporal-lobe response to vocal sounds. *Cognitive Brain Research, 13*(1), 17–26. http://dx.doi.org/10.1016/S0926-6410(01)00084-2.

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature, 403*(6767), 309–312. http://dx.doi.org/10.1038/35002078.

Bennetts, R. J., Kim, J., Burke, D., Brooks, K. R., Lucey, S., Saragih, J., & Robbins, R. A. (2013). The movement advantage in famous and unfamiliar faces: A comparison of point-light displays and shape-normalised avatar stimuli. *Perception, 42*(9), 950–970. http://dx.doi.org/10.1068/p7446.

Bernstein, M., & Yovel, G. (2015). Two neural pathways of face processing: A critical evaluation of current models. *Neuroscience & Biobehavioral Reviews, 55*, 536–546. http://dx.doi.org/10.1016/j.neubiorev.2015.06.010.

Berry, D. S. (1990). What can a moving face tell us? *Journal of Personality and Social Psychology, 58*(6), 1004–1014.

Berry, D. S. (1991). Child and adult sensitivity to gender information in patterns of facial motion. *Ecological Psychology, 3*(4), 349–366. http://dx.doi.org/10.1207/s15326969eco0304_3.

Blank, H., Anwander, A., & von Kriegstein, K. (2011). Direct structural connections between voice- and face-recognition areas. *The Journal of Neuroscience, 31*(36), 12906–12915. http://dx.doi.org/10.1523/JNEUROSCI.2091-11.2011.

Bonda, E., Petrides, M., Ostry, D., & Evans, A. (1996). Specific involvement of human parietal systems and the amygdala in the perception of biological motion. *Journal of Neuroscience, 16*(11), 3737–3744. http://dx.doi.org/10.1162/089892900562417.

Bonner, L., Burton, A. M., & Bruce, V. (2003). Getting to know you: How we learn new faces. *Visual Cognition, 10*(5), 527–536. http://dx.doi.org/10.1080/13506280244000168.

Bricker, P. D., & Pruzansky, S. (1976). Speaker recognition. In N. J. Lass (Ed.). *Contemporary issues in experimental phonetics* (pp. 295–326). New York: Academic Press.

Bruce, V., & Valentine, T. (1988). When a nod's as good as a wink: The role of dynamic information in facial recognition. In M. M. Gruneberg, P. Morris, & R. N. Sykes (Vol. Eds.), *Memory in everyday life: . Vol. 1. Practical aspects of memory: Current research and issues* (pp. 169–174). New York: Wiley.

Bruce, V., & Young, A. W. (1986). Understanding face recognition. *British Journal of Psychology, 77*(3), 305–327. http://dx.doi.org/10.1111/j.2044-8295.1986.tb02199.x.

Calder, A. J., & Young, A. W. (2005). Understanding the recognition of facial identity and facial expression. *Nature Reviews. Neuroscience, 6*(8), 641–651. http://dx.doi.org/10.1038/nrn1724.

Callan, D. E., Jones, J. A., Munhall, K., Callan, A. M., Kroos, C., & Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement by visual speech gestures. *Neuroreport, 14*(17), 2213.

Calvert, G. A., Brammer, M. J., Bullmore, E. T., Campbell, R., Iversen, S. D., & David, A. S. (1999). Response amplification in sensory-specific cortices during crossmodal binding. *Neuroreport, 10*(12), 2619.

Calvert, G. A., & Campbell, R. (2003). Reading speech from still and moving faces: The neural substrates of visible speech. *Journal of Cognitive Neuroscience, 15*(1), 57–70. http://dx.doi.org/10.1162/089892903321107828.

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology, 10*(11), 649–657.

Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences, 11*(12), 535–543. http://dx.doi.org/10.1016/j.tics.2007.10.001.

Christie, F., & Bruce, V. (1998). The role of dynamic information in the recognition of unfamiliar faces. *Memory & Cognition, 26*(4), 780–790. http://dx.doi.org/10.3758/BF03211397.

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research Evaluation, 10*, 1–9.

Cutting, J. E., & Kozlowski, L. T. (1977). Recognizing friends by their walk: Gait perception without familiarity cues. *Bulletin of the Psychonomic Society, 9*(5), 353–356.
Edsel, A. P., & Slate, E. H. (2014). Gvlma: Global validation of linear models assumptions. R package version 1.0.0.2. < https://CRAN.R-project.org/package=gvlma > .
Ellis, H. D., Shepherd, J. W., & Davies, G. M. (2016). Identification of familiar and unfamiliar faces from internal and external features: Some implications for theories of face recognition. *Perception, 8*(4), 431–439. http://dx.doi.org/10.1068/p080431.
Fellowes, J. M., Remez, R. E., & Rubin, P. E. (1997). Perceiving the sex and identity of a talker without natural vocal timbre. *Perception & Psychophysics, 59*(6), 839–849.
Foley, E., Rippon, G., Thai, N. J., Longe, O., & Senior, C. (2012). Dynamic facial expressions evoke distinct activation in the face perception network: A connectivity analysis study. *Journal of Cognitive Neuroscience, 24*(2), 507–520. http://dx.doi.org/10.1162/jocn_a_00120.
Fox, C. J., Iaria, G., & Barton, J. J. S. (2009). Defining the face processing network: Optimization of the functional localizer in fMRI. *Human Brain Mapping, 30*(5), 1637–1651. http://dx.doi.org/10.1002/hbm.20630.
Girges, C., Spencer, J., & O'Brien, J. (2015). Categorizing identity from facial motion. *Quarterly Journal of Experimental Psychology, 68*(9), 1832–1843. http://dx.doi.org/10.1080/17470218.2014.993664.
Grossman, E., Donnelly, M., Price, R., Pickens, D., Morgan, V., Neighbor, G., & Blake, R. (2000). Brain areas involved in perception of biological motion. *Journal of Cognitive Neuroscience, 12*(5), 711–720. http://dx.doi.org/10.1162/089892900562417.
Grossmann, T., Oberecker, R., Koch, S. P., & Friederici, A. D. (2010). The developmental origins of voice processing in the human brain. *Neuron, 65*(6), 852–858. http://dx.doi.org/10.1016/j.neuron.2010.03.001.
Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences, 4*(6), 223–233. http://dx.doi.org/10.1016/S1364-6613(00)01482-0.
Heald, S. L. M., & Nusbaum, H. C. (2014). Talker variability in audio-visual speech perception. *Frontiers in Psychology, 5*(655), 698. http://dx.doi.org/10.3389/fpsyg.2014.00698.
Hecker, M. H. L. (1971). Speaker recognition: An interpretive survey of the literature. *ASHA Monographs, 16*, 1–103.
Hein, G., & Knight, R. T. (2008). Superior temporal sulcus—it's my area: Or is it? *Journal of Cognitive Neuroscience, 20*(12), 2125–2136. http://dx.doi.org/10.1162/jocn.2008.20148.
Hill, H., Jinno, Y., & Johnston, A. (2003). Comparing solid-body with point-light animations. *Perception, 32*(5), 561–566.
Howard, R. J., Brammer, M. J., Wright, I., Woodruff, P. W., Bullmore, E. T., & Zeki, S. (1996). A direct demonstration of functional specialization within motion-related visual and auditory cortex of the human brain. *Current Biology, 6*(8), 1015–1019. http://dx.doi.org/10.1016/S0960-9822(02)00646-2.
Jacobs, A., Pinto, J., & Shiffrar, M. (2004). Experience, context, and the visual perception of human movement. *Journal of Experimental Psychology: Human Perception and Performance, 30*(5), 822–835. http://dx.doi.org/10.1037/0096-1523.30.5.822.
Jesse, A. (2018). Data from "Learning to recognize unfamiliar talkers: Listeners rapidly form representations of facial dynamic signatures". Available from Open Science Framework: < https://osf.io/9zgxy/?view_only=36eccaa0e24d47d5b91e544fc735db94 > .
Jesse, A., & Massaro, D. W. (2010). The temporal distribution of information in audiovisual spoken-word identification. *Attention, Perception, & Psychophysics, 72*(1), 209–225. http://dx.doi.org/10.3758/APP.72.1.209.
Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics, 14*(2), 201–211.
Johnston, R. A., & Edmonds, A. J. (2009). Familiar and unfamiliar face recognition: A review. *Memory, 17*(5), 577–596. http://dx.doi.org/10.1080/09658210902976969.
Kamachi, M., Hill, H., Lander, K., & Vatikiotis-Bateson, E. (2003). Putting the face to the voice. *Current Biology, 13*(19), 1709–1714. http://dx.doi.org/10.1016/j.cub.2003.09.005.
Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience, 17*(11), 4302–4311.
Knight, B., & Johnston, A. (1997). The role of movement in face recognition. *Visual Cognition, 4*(3), 265–273. http://dx.doi.org/10.1080/713756764.
Krauss, R. M., Freyberg, R., & Morsella, E. (2002). Inferring speakers' physical attributes from their voices. *Journal of Experimental Social Psychology, 38*, 618–625. http://dx.doi.org/10.1016/S0022-1031(02)00510-3.
Lachs, L., & Pisoni, D. B. (2004a). Crossmodal source identification in speech perception. *Ecological Psychology, 16*(3), 159–187. http://dx.doi.org/10.1207/s15326969eco1603_1.
Lachs, L., & Pisoni, D. B. (2004b). Specification of cross-modal source information in isolated kinematic displays of speech. *Journal of the Acoustical Society of America, 116*(1), 507–518. http://dx.doi.org/10.1121/1.1757454.
Lahnakoski, J. M., Glerean, E., Salmi, J., Jääskeläinen, I. P., Sams, M., Hari, R., & Nummenmaa, L. (2012). Naturalistic FMRI mapping reveals superior temporal sulcus as the hub for the distributed brain network for social perception. *Frontiers in Human Neuroscience, 6*, 233. http://dx.doi.org/10.3389/fnhum.2012.00233.
Lander, K., & Bruce, V. (2000). Recognizing famous faces: Exploring the benefits of facial motion. *Ecological Psychology, 12*(4), 259–272. http://dx.doi.org/10.1207/S15326969ECO1204_01.
Lander, K., & Bruce, V. (2003). The role of motion in learning new faces. *Visual Cognition, 10*(8), 897–912. http://dx.doi.org/10.1080/13506280344000149.
Lander, K., & Bruce, V. (2004). Repetition priming from moving faces. *Memory & Cognition, 32*(4), 640–647. http://dx.doi.org/10.3758/BF03195855.
Lander, K., Bruce, V., & Hill, H. (2001). Evaluating the effectiveness of pixelation and blurring on masking the identity of familiar faces. *Applied Cognitive Psychology, 15*(1), 101–116. http://dx.doi.org/10.1002/1099-0720(200101/02)15:1<101::AID-ACP697>3.0.CO;2-7.
Lander, K., Christie, F., & Bruce, V. (1999). The role of movement in the recognition of famous faces. *Memory & Cognition, 27*(6), 974–985.
Lander, K., & Chuang, L. (2005). Why are moving faces easier to recognize? *Visual Cognition, 12*(3), 429–442. http://dx.doi.org/10.1080/13506280444000382.
Lander, K., & Davies, R. (2007). Exploring the role of characteristic motion when learning new faces. *Quarterly Journal of Experimental Psychology, 60*(4), 519–526. http://dx.doi.org/10.1080/17470210601117559.
Lander, K., Humphreys, G., & Bruce, V. (2004). Exploring the role of motion in prosopagnosia: Recognizing, learning and matching faces. *Neurocase, 10*(6), 462–470. http://dx.doi.org/10.1080/13554790490900761.
Legge, G. E., Grossmann, C., & Pieper, C. M. (1984). Learning unfamiliar voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(2), 298–303.
Loebach, J. L., Bent, T., & Pisoni, D. B. (2008). Multiple routes to the perceptual learning of speech. *Journal of the Acoustical Society of America, 124*(1), 552–561. http://dx.doi.org/10.1121/1.2931948.
Longmore, C. A., & Tree, J. J. (2013). Motion as a cue to face recognition: Evidence from congenital prosopagnosia. *Neuropsychologia, 51*(5), 864–875. http://dx.doi.org/10.1016/j.neuropsychologia.2013.01.022.
Loula, F., Prasad, S., Harber, K., & Shiffrar, M. (2005). Recognizing people from their movement. *Journal of Experimental Psychology: Human Perception and Performance, 31*(1), 210–220. http://dx.doi.org/10.1037/0096-1523.31.1.210.
Lumley, T., & Miller, A. (2004). *Leaps: Regression subset selection (R package version).* Vienna, Austria: R Foundation.
Macaluso, E., George, N., Dolan, R., Spence, C., & Driver, J. (2004). Spatial and temporal factors during processing of audiovisual speech: A PET study. *NeuroImage, 21*(2), 725–732. http://dx.doi.org/10.1016/j.neuroimage.2003.09.049.
Malone, D. R., Morris, H. H., Kay, M. C., & Levin, H. S. (1982). Prosopagnosia: A double dissociation between the recognition of familiar and unfamiliar faces. *Journal of Neurology, Neurosurgery & Psychiatry, 45*(9), 820–822. http://dx.doi.org/10.1136/jnnp.45.9.820.
Massaro, D. W. (1998). Perceiving talking faces: From speech perception to a behavioral principle. Cambridge, MA.
Massaro, D. W., & Jesse, A. (2007). Audiovisual speech perception and word recognition. In M. G. Gaskell (Ed.). *The Oxford handbook of psycholinguistics* (pp. 19–35). Oxford: Oxford University Press.
Mavica, L. W., & Barenholtz, E. (2013). Matching voice and face identity from static images. *Journal of Experimental Psychology: Human Perception and Performance, 39*(2), 307–312. http://dx.doi.org/10.1037/a0030945.
Nakamura, K., Kawashima, R., Sugiura, M., Kato, T., Nakamura, A., Hatano, K., et al. (2001). Neural substrates for recognition of familiar voices: A PET study. *Neuropsychologia, 39*(10), 1047–1054. http://dx.doi.org/10.1016/S0028-3932(01)00037-9.
Natu, V., & O'Toole, A. J. (2011). The neural processing of familiar and unfamiliar faces: A review and synopsis. *British Journal of Psychology, 102*(4), 726–747. http://dx.doi.org/10.1111/j.2044-8295.2011.02053.x.
Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics, 60*(3), 355–376. http://dx.doi.org/10.3758/BF03206860.
O'Toole, A. J., Roark, D. A., & Abdi, H. (2002). Recognizing moving faces: A psychological and neural synthesis. *Trends in Cognitive Sciences, 6*(6), 261–266.
Pitcher, D., Dilks, D. D., Saxe, R. R., Triantafyllou, C., & Kanwisher, N. (2011). Differential selectivity for dynamic versus static information in face-selective cortical regions. *NeuroImage, 56*(4), 2356–2363. http://dx.doi.org/10.1016/j.neuroimage.2011.03.067.
R Core Team (2016). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing < https://www.R-project.org/ > .
Redcay, E. (2008). The superior temporal sulcus performs a common function for social and speech perception: Implications for the emergence of autism. *Neuroscience & Biobehavioral Reviews, 32*(1), 123–142. http://dx.doi.org/10.1016/j.neubiorev.2007.06.004.
Remez, R. E., Fellowes, J. M., & Nagel, D. S. (2007). On the perception of similarity among talkers. *Journal of the Acoustical Society of America, 122*(6), 3688–3696. http://dx.doi.org/10.1121/1.2799903.
Remez, R. E., Fellowes, J. M., & Rubin, P. E. (1997). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance, 23*(3), 651–666.
Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science, 212*(4497), 947–949.
Riedel, P., Ragert, P., Schelinski, S., Kiebel, S. J., & von Kriegstein, K. (2015). Visual face-movement sensitive cortex is relevant for auditory-only speech recognition. *Cortex, 68*, 86–99. http://dx.doi.org/10.1016/j.cortex.2014.11.016.
Roark, D. A., O'Toole, A. J., Abdi, H., & Barrett, S. E. (2006). Learning the moves: The effect of familiarity and facial motion on person recognition across large changes in viewing format. *Perception, 35*(6), 761–773. http://dx.doi.org/10.1068/p5503.
Rosenblum, L. D., Johnson, J. A., & Saldaña, H. M. (1996). Point-light facial displays enhance comprehension of speech in noise. *Journal of Speech, Language, and Hearing Research, 39*(6), 1159–1170. http://dx.doi.org/10.1044/jshr.3906.1159.
Rosenblum, L. D., Niehus, R. P., & Smith, N. M. (2007). Look who's talking: Recognizing friends from visible articulation. *Perception, 36*(1), 157–159. http://dx.doi.org/10.1068/p5613.
Rosenblum, L. D., & Saldaña, H. M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 22*(2), 318–331.
Rosenblum, L. D., Smith, N. M., Nichols, S. M., Hale, S., & Lee, J. (2006). Hearing a face: Cross-modal speaker matching using isolated visible speech. *Perception &*

Psychophysics, 68(1), 84–93. http://dx.doi.org/10.3758/BF03193658.
Rosenblum, L. D., Yakel, D. A., Baseer, N., & Panchal, A. (2002). Visual speech information for face recognition. *Perception & Psychophysics, 64*(2), 220–229. http://dx.doi.org/10.1007/978-3-642-34500-5_57.
Rossion, B., Schiltz, C., Robaye, L., Pirenne, D., & Crommelinck, M. (2001). How does the brain discriminate familiar and unfamiliar faces?: A PET study of face categorical perception. *Journal of Cognitive Neuroscience, 13*(7), 1019–1034.
Schall, S., Kiebel, S. J., Maess, B., & von Kriegstein, K. (2013). Early auditory sensory processing of voices is facilitated by visual mechanisms. *NeuroImage, 77*, 237–245. http://dx.doi.org/10.1016/j.neuroimage.2013.03.043.
Schultz, J., & Pilz, K. S. (2009). Natural facial motion enhances cortical responses to faces. *Experimental Brain Research, 194*(3), 465–475. http://dx.doi.org/10.1007/s00221-009-1721-9.
Sekiyama, K., Kanno, I., Miura, S., & Sugita, Y. (2003). Auditory-visual speech perception examined by fMRI and PET. *Neuroscience Research, 47*(3), 277–287. http://dx.doi.org/10.1016/S0168-0102(03)00214-1.
Sheffert, S. M., & Olson, E. (2004). Audiovisual speech facilitates voice learning. *Perception & Psychophysics, 66*(2), 352–362.
Sheffert, S. M., Pisoni, D. B., Fellowes, J. M., & Remez, R. E. (2002). Learning to recognize talkers from natural, sinewave, and reversed speech samples. *Journal of Experimental Psychology: Human Perception and Performance, 28*(6), 1447–1469. http://dx.doi.org/10.1037//0096-1523.28.6.1447.
Shultz, S., Vouloumanos, A., & Pelphrey, K. (2012). The superior temporal sulcus differentiates communicative and noncommunicative auditory signals. *Journal of Cognitive Neuroscience, 24*(5), 1224–1232. http://dx.doi.org/10.1162/jocn_a_00208.
Skelton, F., & Hay, D. (2008). Do children utilize motion when recognizing faces? *Visual Cognition, 16*(4), 419–429. http://dx.doi.org/10.1080/13506280701577496.
Smith, H. M. J., Dunn, A. K., Baguley, T., & Stacey, P. C. (2016). Concordant cues in faces and voices testing the backup signal hypothesis. *Evolutionary Psychology, 14*(1), http://dx.doi.org/10.1177/1474704916630317 pp. 1474704916630317.
Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd, & R. Campbell (Eds.). *Hearing by eye the psychology of lipreading* (pp. 3–51). London: Lawrence Earlbaum Associates.
Tabachnick, B. G., & Fidell, L. S. (2012). *Using multiavariate statistics* (6th ed.). New Jersey: Allyn & Bacon.
Thomas, S. M., & Jordan, T. R. (2001). Techniques for the production of point-light and fully illuminated video displays from identical recordings. *Behavior Research Methods, Instruments, & Computers, 33*(1), 59–64.
Thornton, I. M., Pinto, J., & Shiffrar, M. (1998). The visual perception of human locomotion. *Cognitive Neuropsychology, 15*(6–8), 535–552. http://dx.doi.org/10.1080/026432998381014.
Turk-Browne, N. B., Norman-Haignere, S. V., & McCarthy, G. (2010). Face-specific resting functional connectivity between the fusiform gyrus and posterior superior temporal sulcus. *Frontiers in Human Neuroscience, 4*, 1–15. http://dx.doi.org/10.3389/fnhum.2010.00176.
Vatikiotis-Bateson, E., Munhall, K. G., Hirayama, M., Lee, Y. V., & Terzopoulos, D. (1996). The dynamics of audiovisual behavior in speech. In D. G. Stork, & M. E. Hennecke (Eds.). *Speechreading by humans and machines: Models, systems, and applications* (pp. 221–232). Berlin, Heidelberg: Springer. http://dx.doi.org/10.1007/978-3-662-13015-5_16.
von Kriegstein, K., Dogan, Ö., Grüter, M., Giraud, A.-L., Kell, C. A., Gruter, T., ... Kiebel, S. J. (2008). Simulation of talking faces in the human brain improves auditory speech recognition. *Proceedings of the National Academy of Sciences, 105*(18), 6747–6752. http://dx.doi.org/10.1073/pnas.0710826105.
von Kriegstein, K., Eger, E., Kleinschmidt, A., & Giraud, A.-L. (2003). Modulation of neural responses to speech by directing attention to voices or verbal content. *Cognitive Brain Research, 17*(1), 48–55. http://dx.doi.org/10.1016/S0926-6410(03)00079-X.
von Kriegstein, K., Kleinschmidt, A., Sterzer, P., & Giraud, A.-L. (2005). Interaction of face and voice areas during speaker recognition. *Journal of Cognitive Neuroscience, 17*(3), 367–376. http://dx.doi.org/10.1162/0898929053279577.
Walden, B. E., Prosek, R. A., & Worthington, D. W. (1974). Predicting audiovisual consonant recognition performance of hearing-impaired adults. *Journal of Speech, Language, and Hearing Research, 17*(2), 270–278.
Watson, R., Latinus, M., Charest, I., Crabbe, F., & Belin, P. (2014). People-selectivity, audiovisual integration and heteromodality in the superior temporal sulcus. *Cortex, 50*(C), 125–136. http://dx.doi.org/10.1016/j.cortex.2013.07.011.
Wright, T. M., Pelphrey, K. A., Allison, T., McKeown, M. J., & McCarthy, G. (2003). Polysensory interactions along lateral temporal regions evoked by audiovisual speech. *Cerebral Cortex, 13*(10), 1034–1043. http://dx.doi.org/10.1093/cercor/13.10.1034.
Yakel, D. A., Rosenblum, L. D., & Fortier, M. A. (2000). Effects of talker variability on speechreading. *Perception & Psychophysics, 62*(7), 1405–1412. http://dx.doi.org/10.3758/BF03212142.
Yehia, H., Rubin, P., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication, 26*, 23–43.
Young, A. W., Newcombe, F., de Haan, E. H. F., Small, M., & Hay, D. C. (1993). Face perception after brain injury. *Brain, 116*(4), 941–959. http://dx.doi.org/10.1093/brain/116.4.941.
Yovel, G., & O'Toole, A. J. (2016). Recognizing people in motion. *Trends in Cognitive Sciences, 20*(5), 383–395. http://dx.doi.org/10.1016/j.tics.2016.02.005.
Zäske, R., Volberg, G., Kovács, G., & Schweinberger, S. R. (2014). Electrophysiological correlates of voice learning and recognition. *The Journal of Neuroscience, 34*(33), 10821–10831. http://dx.doi.org/10.1523/JNEUROSCI.0581-14.2014.
Zhang, H., Tian, J., Liu, J., Li, J., & Lee, K. (2009). Intrinsically organized network for face perception during the resting state. *Neuroscience Letters, 454*(1), 1–5. http://dx.doi.org/10.1016/j.neulet.2009.02.054.