Journal of Experimental Social Psychology 47 (2011) 259–263

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

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

FlashReport

When two become one: Temporally dynamic integration of the face and voice

Jonathan B. Freeman [*], Nalini Ambady

Tufts University, Medford, MA 02155, USA

ARTICLE INFO

Article history:
Received 16 April 2010
Revised 14 July 2010
Available online 7 September 2010

Keywords:
Person perception
Social categorization
Face perception
Voice perception
Temporal
Dynamics
Motor processes

ABSTRACT

In everyday interactions with others, people have to deal with the sight of a face and sound of a voice at the same time. How the perceptual system brings this information together over hundreds of milliseconds to perceive others remains unclear. In 2 studies, we investigated how facial and vocal cues are integrated during real-time social categorization by recording participants' hand movements (via the streaming *x*, *y* coordinates of the computer mouse) en route to "male" and "female" responses on the screen. Participants were presented with male and female faces that were accompanied by a same-sex voice morphed to be either sex-typical (e.g., masculinized male voice) or sex-atypical (i.e., feminized male voice). Before settling into ultimate sex categorizations of the face, the simultaneous processing of a sex-atypical voice led the hand to be continuously attracted to the opposite sex-category response across construal. This is evidence that ongoing results from voice perception continuously influence face perception across processing. Thus, social categorization involves dynamic updates of gradual integration of the face and voice.

Upon encountering others, we rapidly glean a variety of information. Most important, perhaps, are social categories, such as sex, race, and age (Macrae & Bodenhausen, 2000). Most research on social categorization has focused on visual features, such as facial cues, with little attention paid to auditory features, such as vocal cues. Recently, however, vocal cues were shown to give rise to categorical judgments and stereotypic inferences of others, and these inferences are sensitive to within-category variation (Ko, Judd, & Blair, 2006), as is seen with facial cues (Blair, Judd, Sadler, & Jenkins, 2002). Thus, like the face, the voice plays an important role in social categorization.

How information from the face and voice is combined to categorize others, however, remains poorly understood. Previous work has provided clear evidence that perceivers do combine facial and vocal inputs. For instance, when a face appears sad but is accompanied by a voice that sounds happy, perceivers consistently report seeing the face as more happy than it really is. This remains true even when participants are instructed to disregard the voice (de Gelder & Vroomen, 2000). Furthermore, congruency between vocal and facial features tends to make person perception more accurate and efficient (for review, Campanella & Belin, 2007). Very few studies, however, have examined face–voice integration in social categorization.

Recent research investigating the real-time social categorization process has found that facial cues trigger multiple partially-active social categories (e.g., male and female) that simultaneously compete over time to gradually stabilize onto ultimate construals (Freeman & Ambady, 2009; Freeman, Ambady, Rule, & Johnson, 2008; Freeman, Pauker, Apfelbaum, & Ambady, 2010). Such work suggests that social categorization is a dynamic, integrative process. Both the masculine and feminine cues on a particular face, for instance, are dynamically integrated over time to form ultimate judgments of sex-category (Freeman et al., 2008). Although multiple facial cues may be dynamically integrated over time, such cues are all within the same sensory modality. It remains unclear how cues from different sensory modalities, such as those bombarding the visual and auditory systems, integrate across the social categorization process.

In line with a dynamic continuity account (see Freeman et al., 2008), we propose that the biases of another person's sensory information (e.g., facial, vocal, and bodily cues) converge the moment they become available in the input to weigh in on multiple partially-active social category representations, and these parallel representations settle onto ultimate judgments across a process of continuous competition. Ongoing voice-processing results, therefore, should integrate with ongoing face-processing results over time. If true, processing of sex-specifying vocal cues should exert a temporally dynamic influence on face processing across construal. This would be consistent with evidence for recurrent interactions between the visual and auditory cortices and top-down feedback from higher-order multimodal cortices (e.g., Ghazanfar, Chandrasekaran, & Logothetis, 2008; Kreifelts, Ethofer, Grodd, Erb, & Wildgruber, 2007).

It is difficult to directly test the face–voice integration process in humans. Although electrophysiological studies have determined the earliest moments at which brain activity may be influenced by the presence of cross-modal information, it remains unclear how after these moments, the processing of another's face and voice is integrated over time. To shed light on this face–voice integration process, one would benefit from a technique that could monitor how

[*] Corresponding author. Department of Psychology, Tufts University, 490 Boston Avenue, Medford, MA 02155, USA.
E-mail address: jon.freeman@tufts.edu (J.B. Freeman).

0022-1031/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jesp.2010.08.018

the actual perception of a face–voice percept evolves over time and how each sensory source contributes to that evolution. Here, we use a computer mouse-tracking technique to track in real-time how voice processing weighs in on resolving facial ambiguities. We examine this in the context of perceiving a target's sex, whose face–voice integration was shown in prior work (Masuda, Tsujii, & Watanabe, 2005; Smith, Grabowecky, & Suzuki, 2007).

By measuring hand movements, we can assess how face and voice processing temporally integrate. Previously, hand movements were used to explore the processing dynamics of a face's sex (Freeman et al., 2008). On each trial, participants categorized the sex of a face by moving the cursor from the bottom-center of the screen to click on a MALE or FEMALE response located at the top-left and top-right corners. When presented with sex-atypical male and female faces (bearing a mixture of masculine and feminine cues), mouse (hand) movements showed a continuous attraction toward the opposite sex-category. This is evidence that the mixture of masculine and feminine cues triggered a continuous competition between male and female categories, which gradually settled onto ultimate judgments. Here, rather than examine how multiple facial cues are integrated, we test whether cues across different sensory modalities are integrated by such a dynamic competition process.

Study 1

By looking at instances in a mouse-tracking paradigm where there is a conflict between a category triggered by facial cues and an opposing category triggered by vocal cues—and how these might compete with one another over time—we can measure how information from the face and voice is merged together. So long as participants' hands are in motion while face–voice processing is still ongoing, the dynamic face–voice integration we predict would be evidenced by the hand's continuous attraction toward the opposite sex-category response in instances where, although the face is categorized as the correct sex, vocal cues partly suggest the opposite sex.

Method

Forty-one individuals participated for partial course credit. Face stimuli were 10 male and 10 female faces computer-generated to be slightly sex-ambiguous. Voice stimuli were 2000 ms clips from 10 males and 10 females selected to be naturalistic for a first-impression encounter (e.g., "My family's origins are pretty interesting."). These were morphed to be sex-typical (e.g., masculinized male voice) or sex-atypical (e.g., feminized male voice) by manipulating the formant ratio. Full details on the stimuli are provided in the Supplementary methods.

Each face was presented twice in the experiment, accompanied by a same-sex voice (once sex-typical and once sex-atypical). Voice stimuli were randomly paired with face stimuli (without replacement). Participants were instructed to categorize the face's sex (and only use the voice if it could help resolve the face's sex, as correct responses were based on the face). To begin each trial, participants clicked a START button at the bottom-center of the screen. The face then appeared in its place, and the voice began playing. Participants categorized by mouse-clicking either MALE or FEMALE located at the top-left and top-right corners (randomized across participants). Meanwhile, the *x*, *y* coordinates of the mouse were recorded using the freely-available MouseTracker package: http://mousetracker.jbfreeman.net (Freeman & Ambady, 2010).

Results and discussion

To permit averaging and comparison across trials, we normalized trajectories into 101 time-steps and remapped leftward trajectories rightward (inverted along the *x*-axis). To index trajectories' attraction toward the opposite sex-category, we computed the maximum deviation (MD): the largest perpendicular deviation from an idealized straight line between the trajectory's start and endpoints. See Freeman and Ambady (2010) for further details on mouse trajectory preprocessing and analytic techniques.

Participants were more likely to misinterpret the face as the opposite sex when faces were accompanied by sex-atypical ($M = 11.5\%$, $SE = 1.0\%$) relative to sex-typical ($M = 5.1\%$, $SE = 0.8\%$) voices, $t(40) = 5.38$, $p < .0001$, a finding often cited as evidence of face–voice integration (e.g., Hietanen, Leppänen, Illi, & Surakka, 2004). To examine the temporal dynamics of this integration, we examined trials that were correctly categorized.

Participants initiated movement early after face/voice onset, and this did not differ between conditions: sex-atypical ($M = 247$ ms, $SE = 16$ ms) and sex-typical ($M = 252$ ms, $SE = 18$ ms), $t(40) = 0.54$, $p = .59$. This ensures that participants' movements were on-line with face–voice processing. Expectedly, response times were longer for faces accompanied by sex-atypical ($M = 1375$ ms, $SE = 42$ ms) than sex-typical ($M = 1317$ ms, $SE = 37$ ms) voices, $t(40) = 2.88$, $p < .01$. More importantly, before participants settled into their ultimate categorizations, the hand was continuously attracted to the opposite sex-category while categorizing faces accompanied by sex-atypical voices ($M = 0.33$, $SE = 0.03$) relative to sex-typical voices ($M = 0.26$, $SE = 0.03$), $t(40) = 3.81$, $p < .001$, as indicated by MD (Fig. 1).

It is possible that this continuous-attraction effect was spuriously produced by averaging across some trajectories in the sex-atypical condition that headed straight to the correct sex-category and others that first headed straight to the opposite sex-category, which were then discretely redirected straight toward the correct sex-category. If true, the MD distribution in the sex-atypical condition would exhibit bimodality (Freeman & Ambady, 2010). However, the MD distribution for sex-atypical trials was within the bimodality-free zone ($b < .555$; SAS Institute, 1989), $b = .407$, as was the distribution for sex-typical trials, $b = .428$ (Fig. 2). Furthermore, the Kolmogorov–Smirnov test confirmed that the shapes of these two distributions were statistically indistinguishable ($D = .02$, $p = .99$), ruling out the possibility of latent bimodality. This ensures that the continuous-attraction effect was not fallaciously produced by a combination of discrete-like movements.

The continuous-attraction effect found in the present study is initial evidence for our hypothesis that category information from the face and voice is dynamically integrated over time. In Study 2, we solidify this evidence by addressing two limitations.

Study 2

It is possible that the continuous-attraction effect found in Study 1 may have reflected a less decisive movement toward the correct sex-category, rather than a genuine attraction to the opposite sex-category due to vocal cues dynamically biasing the processing of the face's sex. To rule out this possibility, here we include a control condition. The same trials of Study 1 were presented, except that the correct sex-category and an animal control word (farm/jungle) appeared as the response alternatives. Trials with farm and jungle animal face–voice pairs were presented as fillers. Participants also completed identical trials of Study 1. If the attraction effect was due to a less decisive movement toward the correct sex-category, it should persist regardless of the opposite alternative. If due to a genuine attraction toward the opposite sex-category, however, the effect should disappear when the opposite alternative is not that category.

Another possibility is that the attraction effect of Study 1 was an artifact of each face being repeated twice in the experiment. Although this ensured that any effects could not be due to differences in face stimuli between conditions, subsequent judgments may have been influenced by prior judgments in a way that might spuriously produce attraction. To rule out this possibility, each face was presented only once in the present study.

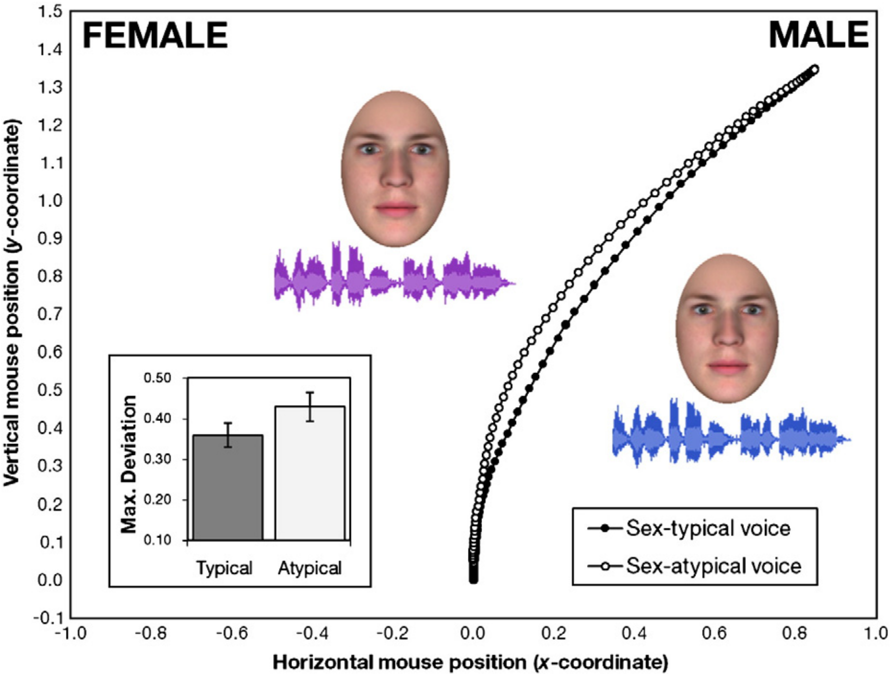

Fig. 1. Mean mouse trajectories of Study 1 (aggregated across male and female targets). In this figure, trajectories for all targets were remapped rightward, with the opposite sex-category on the left and the sex-category consistent with the face's sex on the right. A sample male face stimulus is displayed (all male and female face stimuli were somewhat sex-ambiguous). A voice stimulus typical for the face's sex (masculine) is shown on the right (audio waveform depicted in blue), next to the mean trajectory for sex-typical trials. Its atypical (feminine) counterpart is shown on the left, next to the mean trajectory for sex-atypical trials (audio waveform depicted in purple). During an actual trial, a single face was centered at the bottom of the screen while the voice stimulus played. The bar graph shows trajectories' maximum deviation toward the opposite sex-category separately for sex-typical and sex-atypical trials (error bars denote standard error of the mean).

Method

Twenty-five individuals participated in exchange for $10. Using the stimuli of Study 1, participants engaged in 20 experimental and 20 control trials (each containing 10 sex-typical and 10 sex-atypical face–voice pairs, half male and half female). Experimental trials were identical to those in Study 1, involving a MALE vs. FEMALE decision, whereas control trials involved a decision between the correct sex-category and FARM or JUNGLE. Ten farm and ten jungle animals were used as filler trials, which involved a decision between the correct animal-category and MALE or FEMALE. Additional 20 faces (10 males and 10 females) were generated so that each face could be presented only once across the experiment. Full details of the study's methods are provided in the Supplementary methods.

Results and discussion

Preprocessing was identical to that of Study 1. A repeated-measures ANOVA on MD values indicated a significant sex-typicality (typical, atypical) × condition (experimental, control) interaction, $F(1,24) = 5.28$, $p < .05$. When the unselected response alternative was the opposite sex-category (experimental trials), the attraction effect of Study 1 replicated, with trajectories for sex-atypical face–voice pairs ($M = 0.38$, $SE = 0.05$) showing higher MD than sex-typical pairs ($M = 0.28$, $SE = 0.03$), $t(24) = 3.26$, $p < .01$ (Fig. 3). Distributional analyses indicated that this continuous-attraction effect was not the spurious result of a combination of discrete-like movements (see Supplementary results). When the unselected response alternative was an animal control word (control trials), however, the attraction effect

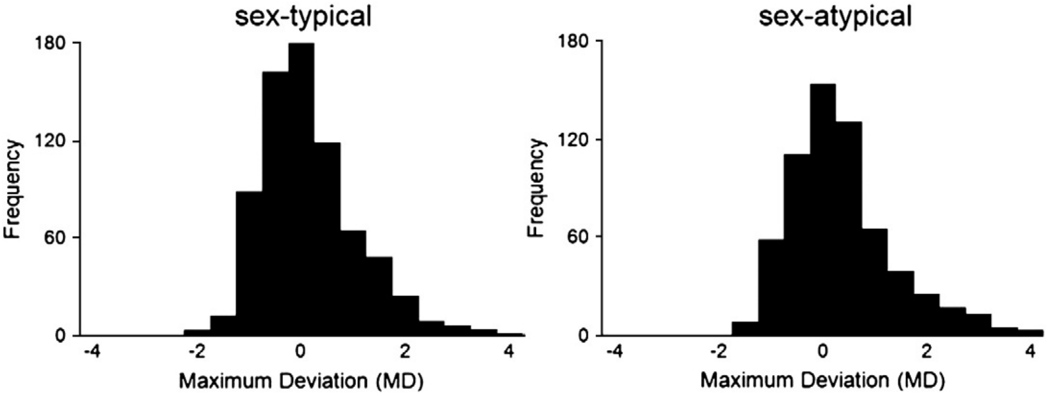

Fig. 2. Histograms of the z-normalized distribution of trajectories' maximum deviation (MD) values in the sex-typical and sex-atypical conditions. The plots illustrate unimodality and a lack of bimodality.

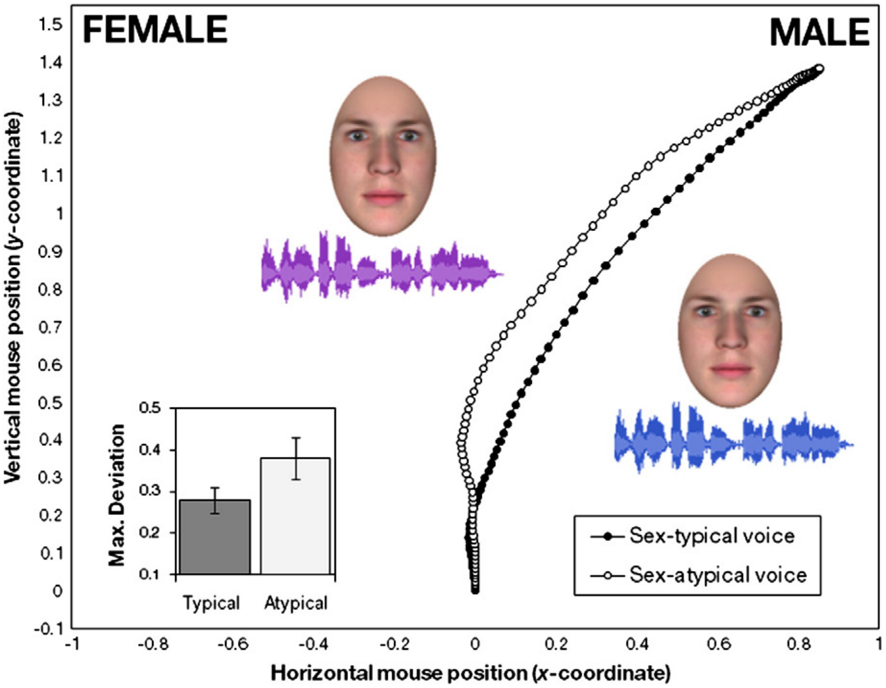

Fig. 3. Mean mouse trajectories in the experimental condition of Study 2 (aggregated across male and female targets). In this figure, trajectories for all targets were remapped rightward, with the opposite sex-category on the left and the sex-category consistent with the face's sex on the right. A sample male face stimulus and sample voice stimuli are displayed, as they were shown in Fig. 1. Unlike in Study 1, each face was only presented once in the experiment. The bar graph shows trajectories' maximum deviation toward the opposite sex-category separately for sex-typical and sex-atypical trials (error bars denote standard error of the mean).

disappeared, with MD in the sex-atypical condition ($M = 0.26$, $SE = 0.04$) no greater than MD in the sex-typical condition ($M = 0.23$, $SE = 0.04$), $t(24) = 1.02$, $p = .32$ (Fig. 4). The main effects of this ANOVA in addition to analyses of accuracy, initiation times, and response times appear in the Supplementary results.

We replicated the continuous-attraction effect found in Study 1 and demonstrated that it cannot be attributed to less decisive movements. Instead, we showed that the effect reflects a genuine parallel attraction to the opposite sex-category, solidifying our evidence for temporally dynamic face–voice integration. Moreover,

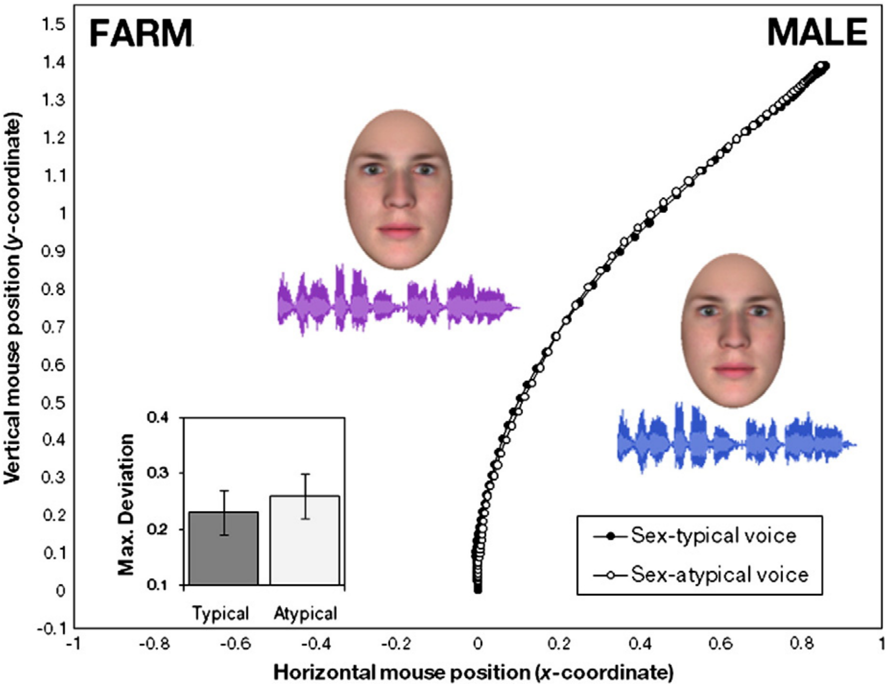

Fig. 4. Mean mouse trajectories in the control condition of Study 2 (aggregated across male and female targets). In this figure, trajectories for all targets were remapped rightward, with the animal control word (FARM or JUNGLE) on the left and the sex-category consistent with the face's sex on the right. A sample male face stimulus and sample voice stimuli are displayed, as they were shown in Figs. 1 and 3. Unlike in Study 1, each face was only presented once in the experiment. The bar graph shows trajectories' maximum deviation toward the animal control word separately for sex-typical and sex-atypical trials (error bars denote standard error of the mean).

by presenting faces only once, we assured that face repetition did not confound the attraction effect. In fact, the effect found in the present study was considerably larger than that found in Study 1, perhaps due to repeated judgments of the same face in Study 1 weakening the attraction.

General discussion

While participants' hands were moving en route to making a sex-categorization of the face, the simultaneous processing of a sex-atypical voice led the hand to travel closer to the opposite sex category, continuously across the course of construal. This is evidence that ongoing voice perception continuously biased face perception across processing and that information from the voice was gradually integrated with information from the face. At each moment during the categorization of sex-atypical pairs, mouse movements were neither in a discrete pursuit straight to the MALE response nor in a discrete pursuit straight to the FEMALE response. Rather, as seen with the conspicuous curving of the trajectory toward the opposite sex-category in Figs. 1 and 3, at each moment the location of the mouse was in a weighted combination of one pursuit consistent with face processing (e.g., male) and a simultaneous pursuit consistent with voice processing (e.g., female), while the mouse progressively stabilized onto ultimate interpretations of the face. Thus, mouse trajectories reflected a gradual incorporation of category information from the face and voice as it accumulated from each source in real-time.

Very few studies have examined face–voice interactions in perceiving social categories, such as sex (Masuda et al., 2005; Smith et al., 2007). The present work thus provides further support that social categorization involves face–voice integration. It also extends these studies by providing insight into the continuous temporal dynamics underlying the categorization process. It is also noteworthy that, in this process, face and voice information is represented in parallel, even when ultimate judgments reflected use of only one channel. On trials where participants correctly categorized a face's sex, vocal cues triggered a partially-active representation of the opposite sex-category, which coexisted in parallel with a representation of the correct sex-category triggered by facial cues. Thus, although ultimate judgments might reflect the use of only one sensory channel, here we show that multiple channels flexibly weigh in on construals and simultaneously receive processing in person perception, and these channels interact with one another over time.

The present studies have several real-world implications. For instance, face–voice interactions were recently shown to be impaired in schizophrenia (De Gelder et al., 2005) and alcoholism (Maurage et al., 2008). Thus, future research could exploit a mouse-tracking paradigm (Freeman & Ambady, 2010), like that used here, to investigate the real-time dynamics of impaired or abnormal face–voice interactions, potentially providing insight into how these dynamics go awry. Other studies in the social domain have found that important interpersonal outcomes may arise when different sensory channels present particular combinations of information, such as conflicting cues or “mixed messages” (e.g., LaPlante & Ambady, 2002). Future work could thus zoom in on the real-time processing by which these conflicting cues are integrated (as done with gendered facial and vocal cues here) because subtleties in this real-time processing could likely reveal important information about downstream social consequences.

In sum, through movements of the hand, we show how results from face and voice processing dynamically integrate over fractions of a second to gradually settle onto ultimate person construals.

Supplementary materials related to this article can be found online at doi:10.1016/j.jesp.2010.08.018.

Acknowledgments

This research was funded by research grant NSF BCS-0435547 (N.A.). We thank Jeffrey Brooks, Rebecca Sylvetsky, and Karen Castillo for their help in data collection.

References

Blair, I. V., Judd, C. M., Sadler, M. S., & Jenkins, C. (2002). The role of Afrocentric features in person perception: Judging by features and categories. *Journal of Personality and Social Psychology*, *83*, 5—25.

Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences*, *11*, 535—543.

de Gelder, B., & Vroomen, J. (2000). The perception of emotions by ear and by eye. *Cognition and Emotion*, *14*, 289—311.

De Gelder, B., Vroomen, J., de Jong, S. J., Masthoff, E. D., Trompenaars, F. J., & Hodiamont, P. (2005). Multisensory integration of emotional faces and voices in schizophrenics. *Schizophrenia Research*, *72*, 195—203.

Freeman, J. B., & Ambady, N. (2009). Motions of the hand expose the partial and parallel activation of stereotypes. *Psychological Science*, *20*, 1183—1188.

Freeman, J. B., & Ambady, N. (2010). MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods*, *42*, 226—241.

Freeman, J. B., Ambady, N., Rule, N. O., & Johnson, K. L. (2008). Will a category cue attract you? Motor output reveals dynamic competition across person construal. *Journal of Experimental Psychology: General*, *137*, 673—690.

Freeman, J. B., Pauker, K., Apfelbaum, E. P., & Ambady, N. (2010). Continuous dynamics in the real-time perception of race. *Journal of Experimental Social Psychology*, *46*, 179—185.

Ghazanfar, A. A., Chandrasekaran, C., & Logothetis, N. K. (2008). Interactions between the superior temporal sulcus and auditory cortex mediate dynamic face/voice integration in rhesus monkeys. *Journal of Neuroscience*, *28*, 4457—4469.

Hietanen, J., Leppänen, J., Illi, M., & Surakka, V. (2004). Evidence for the integration of audiovisual emotional information at the perceptual level of processing. *European Journal of Cognitive Psychology*, *16*, 769—790.

Institute, S. A. S. (1989). SAS/STAT user's guide. Cary, NC: Author.

Ko, S. J., Judd, C. M., & Blair, I. V. (2006). What the voice reveals: Within- and between-category stereotyping on the basis of voice. *Personality and Social Psychology Bulletin*, *32*, 806—819.

Kreifelts, B., Ethofer, T., Grodd, W., Erb, M., & Wildgruber, D. (2007). Audiovisual integration of emotional signals in voice and face: An event-related fMRI study. *Neuroimage*, *37*, 1445—1456.

LaPlante, D., & Ambady, N. (2002). Saying it like it isn't: Mixed messages from men and women in the workplace. *Journal of Applied Social Psychology*, *32*, 2435—2457.

Macrae, C. N., & Bodenhausen, G. V. (2000). Social cognition: Thinking categorically about others. *Annual Review of Psychology*, *51*, 93—120.

Masuda, S., Tsujii, T., & Watanabe, S. (2005). An interference effect of voice presentation on face gender discrimination task: Evidence from event-related potentials. *International Congress Series*, *1278*, 156—159.

Maurage, P., Philippot, P., Joassin, F., Pauwels, L., Pham, T., Prieto, E. A., et al. (2008). The auditory–visual integration of anger is impaired in alcoholism: An event-related potentials study. *Journal of Psychiatry and Neuroscience*, *33*, 111—122.

Smith, E. L., Grabowecky, M., & Suzuki, S. (2007). Auditory–visual crossmodal integration in perception of face gender. *Current Biology*, *17*, 1680—1685.