Journal of Experimental Psychology:
Human Perception and Performance
2013, Vol. 39, No. 2, 307–312

© 2012 American Psychological Association
0096-1523/13/$12.00 DOI: 10.1037/a0030945

OBSERVATION

Matching Voice and Face Identity From Static Images

Lauren W. Mavica and Elan Barenholtz
Florida Atlantic University

Previous research has suggested that people are unable to correctly choose which unfamiliar voice and static image of a face belong to the same person. Here, we present evidence that people can perform this task with greater than chance accuracy. In Experiment 1, participants saw photographs of two, same-gender models, while simultaneously listening to a voice recording of one of the models pictured in the photographs and chose which of the two faces they thought belonged to the same model as the recorded voice. We included three conditions: (a) the visual stimuli were frontal headshots (including the neck and shoulders) and the auditory stimuli were recordings of spoken sentences; (b) the visual stimuli only contained cropped faces and the auditory stimuli were full sentences; (c) we used the same pictures as Condition 1 but the auditory stimuli were recordings of a single word. In Experiment 2, participants performed the same task as in Condition 1 of Experiment 1 but with the stimuli presented in sequence. Participants also rated the model's faces and voices along multiple "physical" dimensions (e.g., weight,) or "personality" dimensions (e.g., extroversion); the degree of agreement between the ratings for each model's face and voice was compared to performance for that model in the matching task. In all three conditions, we found that participants chose, at better than chance levels, which faces and voices belonged to the same person. Performance in the matching task was not correlated with the degree of agreement on any of the rated dimensions.

Keywords: voice recognition, face recognition, multisensory

Faces and voices both carry information that may be used to infer physical characteristics of unfamiliar people. People can infer properties such as the height or age of unfamiliar people based on pictures of their faces or recordings of their voices (Allport & Cantril, 1934; Lass & Colt, 1980). Similarly, people are able to determine personality traits, such as extraversion and conscientiousness of unfamiliar people from pictures of their faces or recordings of their voices (Allport & Cantril, 1934; Berry, 1991; Borkenau & Liebler, 1992).

The presence of common information across vocal and facial characteristics suggests that people might be able to directly match an unfamiliar voice to the face of the person of the same identity. Indeed, people learn mappings between certain facial and vocal properties early in development; infants can match faces and voices based on emotional expression at seven months (Walker-Andrews, 1986) and based on gender at eight months (Patterson & Werker, 2002). Over a lifetime of experience, these mappings could become specific enough to match facial and vocal identity. However, while people are able to accurately match voices with full-body photographs of unfamiliar people (Krauss, Freyberg & Morsella, 2002), two previous studies found that people could not match voice recordings to pictures of faces alone. Both Lachs and Pisoni (2004) and Kamachi, Hill, Lander, and Vatikiotis-Bateson (2003) found that people could match voices to dynamically articulating faces but not static photographs, suggesting that matching depended on dynamic properties of articulating faces, not their static, structural properties.

These negative results suggest that people's ability to learn mappings between static face structure and voices may be isolated to broad categories, such as gender, and may not extend to the more subtle mappings needed to match identities. However, here we report evidence that people can, in fact, match voices to static pictures of faces at greater than chance levels. In Experiment 1, we presented participants with photographs of two faces simultaneously, along with a recording of the voice of one of the people pictured in the photographs; participants had to choose which of the two photographs they thought belonged to the same person as the recorded voice. We included three experimental conditions with varying degrees of visual and auditory information (see Figure 1). In all three conditions, we found that participants were able to match the faces and voices at greater than chance levels.

We also conducted a follow-up experiment, designed to determine whether the discrepancy between our results and previous studies may have been due to the fact that they used sequential presentation of stimuli (which requires matching based on memory), while we presented the stimuli simultaneously. Here too we found better than chance performance.

As noted, previous studies have found that people can correctly assess certain basic characteristics, such as height and extraversion, based on both faces and voices presented independently. Thus, one way in which participants could theoretically have matched faces and

This article was published Online First December 31, 2012.

Lauren W. Mavica and Elan Barenholtz, Department of Psychology, Florida Atlantic University.

Correspondence concerning this article should be addressed to Lauren W. Mavica, Department of Psychology, Florida Atlantic University, 777 Glades Road, Boca Raton, FL 33431. E-mail: lkogelsc@fau.edu

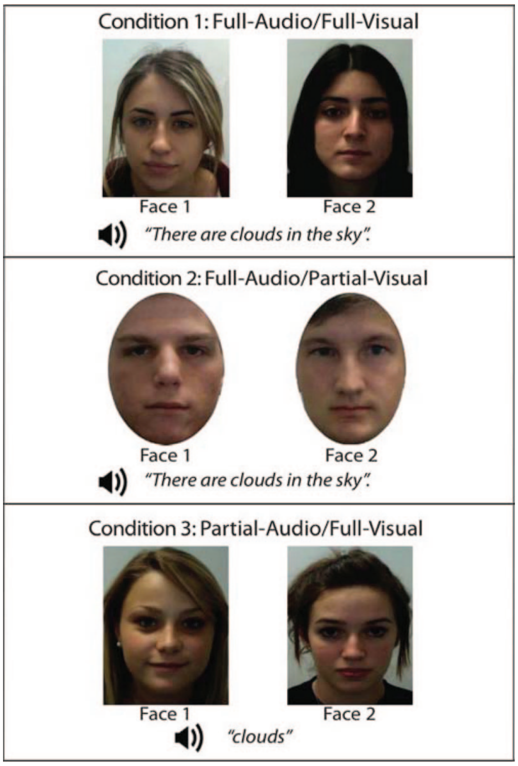
Condition 1: Full-Audio/Full-Visual
Face 1
Face 2
"There are clouds in the sky".
Condition 2: Full-Audio/Partial-Visual
Face 1
Face 2
"There are clouds in the sky".
Condition 3: Partial-Audio/Full-Visual
Face 1
Face 2
"clouds"

Figure 1. Example stimuli in each of the three experimental conditions in Experiment 1. Participants were presented with two pictures of faces and a recording of a voice. Their task was to choose which face belonged to the same person as the voice. In Condition 1, the visual stimuli presented were frontal headshots, including the head, hair, neck, and shoulders of the models, similar to those used in previous studies (Kamachi et al., 2003; Lachs & Pisoni, 2004); the auditory stimuli were recordings of full sentences, spoken by the same people. In Condition 2, the visual information only included a cropped face; the auditory stimuli in this condition were identical to the first condition. In Condition 3, we presented full headshots (as in Condition 1) but the auditory stimuli consisted of isolated English words, rather than full sentences. This condition aimed to reproduce the auditory stimuli used in Lachs and Pisoni (2004), who used single words as auditory stimuli.

voices in our experiment is by comparing them with regard to one or several of these dimensions. To test this possibility, after completing the matching task participants rated the faces and voices along six "physical dimensions"—height, weight, age, attractiveness, socioeconomic status (SES), and masculinity/femininity—and five "personality" dimensions—openness, conscientiousness, extraversion, agreeableness and calmness. We then tested whether the degree of agreement between the ratings of a model's face and voice along any of these physical or personality dimensions predicted performance in the matching task.

Experiment 1

Method

Participants. Participants were 75 English-speaking undergraduate students (25 per condition), naïve to the purposes of the experiment enrolled in an introductory psychology course at Florida Atlantic University, who received course credit for participation. All participants reported having normal or corrected-to-normal vision and hearing. After completion of the experiment, participants were asked if they recognized any of the models. Trials containing the face or voice of a recognized person (either as the correct identity or foil) were excluded from analysis. In Condition 1 three participants each recognized one model, and in Condition 3 four participants each recognized one model.

Stimuli. Stimuli consisted of photographs of faces and audio recordings of voices derived from 64 (32 male and 32 female) self-identified Caucasian undergraduate students from Florida Atlantic University. Figure 1 and the accompanying caption show the three experimental conditions.

To control for speed and varying articulation patterns, the stimulus models listened, through headphones, to prerecorded sentences, played in a loop. They were instructed to speak along with the sentences into a microphone between five and 10 times. A single recording of each sentence was selected, based on matching the original prerecorded sentence, for use as a stimulus.

Design and procedure. The study used a between-subjects design to compare the three experimental conditions. In the three conditions, participants were tested on all 64 models, repeated across three sequential blocks, with each block employing one of the three auditory recordings. Each block was subdivided into a male and female subblock, within which the face and voice stimuli of all 32 models of that gender were presented in random order; each participant was randomly assigned to have either the male or female stimuli shown first across all three blocks.

On each trial, participants were presented with photographs of two faces of the same gender, while they simultaneously listened to a recording of a voice (see Figure 1). One of the presented faces (randomly presented on the left as "Face 1" or the right as "Face 2") was a photograph of the same person whose voice recording was being played. The other, "foil" face consisted of one of the other 31 models of the same gender, which had been randomly paired with the correct face for that trial only. The participant's task was to guess which face belonged to the same person as the recorded voice. Each face appeared twice within each block, once as the correct face and once as a foil for a different, randomly selected face. The pairings between faces were randomized across each block. No feedback was given on whether their choice was correct or not, precluding learning of the face−voice pairings.

After completing the matching task, participants in Conditions 1 and 2 completed a ratings task in which they judged the faces and voices (in separate blocks) used as experimental stimuli. Participants in Condition 1 judged each face and voice along six "physical" dimensions: height (in inches), weight (in pounds), age (in years); a scale of 1–5 was used to judge each face and voice in terms of SES, masculinity or femininity, and attractiveness. Participants in Condition 2 judged the faces and voices along five "personality" dimensions on a scale of 1–5 for: openness, conscientiousness, extraversion, agreeableness, and calmness. We calculated the difference between the rating of each model's face and voice for each dimension to compute a "difference score" to compare to performance in the matching task.

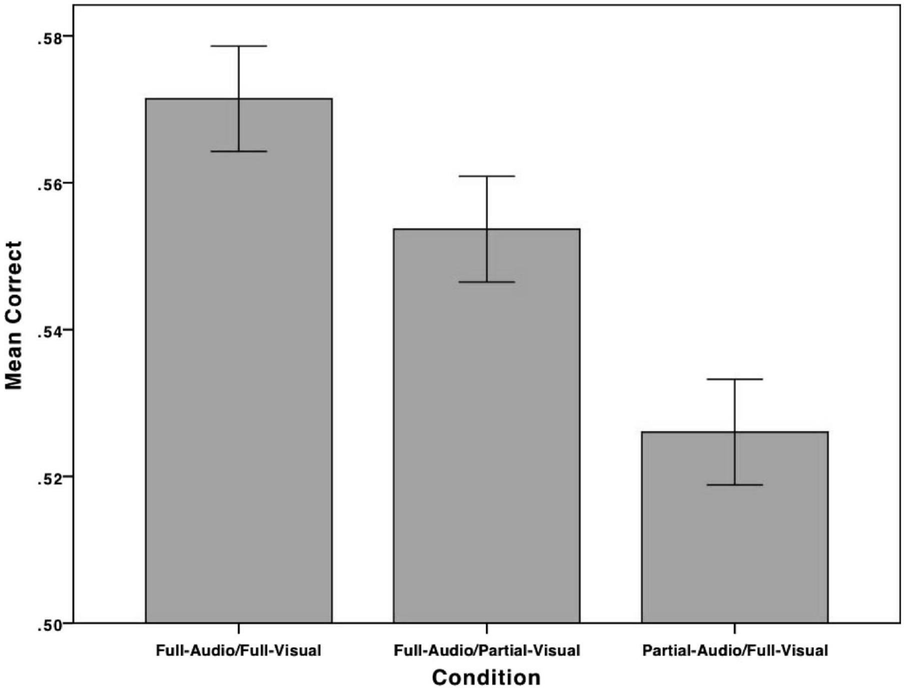

Figure 2. Average performance in each of the three experimental conditions. Error bars indicate +/−1 standard error of the mean.

Results. Across all three experimental conditions, there were no significant differences in performance across the three experimental blocks (i.e., for the three different sentences/ words); in subsequent analyses the data from the three blocks were collapsed. Figure 2 shows the average performance across participants for each of the three experimental conditions. Individual *t* tests found significantly better than chance performance (50%) in each of the conditions: Condition 1: Full Audio/Full Visual [$M = 0.57$, $SD = 0.038$, $t(24) = 9.554$, $p <$ .001, d = *1.91*], Condition 2: Full Audio/Partial Visual [$M =$ 0.55, $SD = 0.04$, $t(24) = 6.710$, $p < .001$, $d = 1.34$], and Condition 3: Partial-Audio/Full-Visual [$M = 0.53$, $SD = .04$, $t(24) = 3.607$, $p < .001$, $d = .72$]. An ANOVA comparing performance in the three conditions found a significant difference in performance, $F(2, 74) = 9.607$, $p< .001$. Tukey's HSD post hoc analysis showed that performance in Condition 1 and Condition 2 were both significantly better than in Condition 3 but that the first two conditions did not significantly differ from one another. Thus, viewing the full head and a portion of the shoulders did not yield better performance than viewing the face alone. However, hearing a full sentence yielded better performance than hearing a single word.

Additional *t* tests comparing performance of male and female subjects found no significant differences (females: $M = .56$, $SD =$ .04, males: $M = .54$, $SD = .04$, $p > .05$) or between performance for male versus female models, ($M = .56$, $SD = .06$ for both, $p >$ .05).

To assess performance for each of the individual models, we calculated the percentage of trials on which participants chose the correct response whenever that model's face appeared, either as the correct match to the voice or as the foil. This measured the ability to correctly match a face to its voice as well as reject matching that face to a different person's voice. Figure 3 shows the average performance, collapsed across all three conditions, for each of the 64 models. Performance varied widely for the different models, ranging between 70% for the best-performance model to 35% for the worst-performance model. However, there was a clear trend toward better-than-chance performance, with 56 of the 64 models (87.5%) yielding performance above 50% and only eight (12.5%) yielding performance at or below 50%. Using Bonferroni adjusted alpha levels of .0008 (.05/64), nine models yielded statistically significant greater-than-chance performance, while no models yielded worse-than-chance performance.

As shown in Figure 4, performance on the different individual models was positively correlated across each of the conditions [Condition 1 and Condition 2: $r(64) = .393$, $p < .001$; Conditions 1 and 3: $r(64) = .562$, $p < .001$; Conditions 2 and 3: $r(64) = .427$, $p < .001$]. Thus, there was a high degree of consistency with regard to which of the models yielded better or worse results in the matching task across conditions.

For the ratings, we calculated the average difference score for each dimension of every model as well as the average performance in the matching task for each model. We then performed correlations between all 11 (six physical and five personality) difference scores as independent variables and performance in the matching task. None of the factors were significantly correlated with performance.

Discussion

Overall, we found that participants were able to match an unfamiliar voice to a static image of the face of the person from

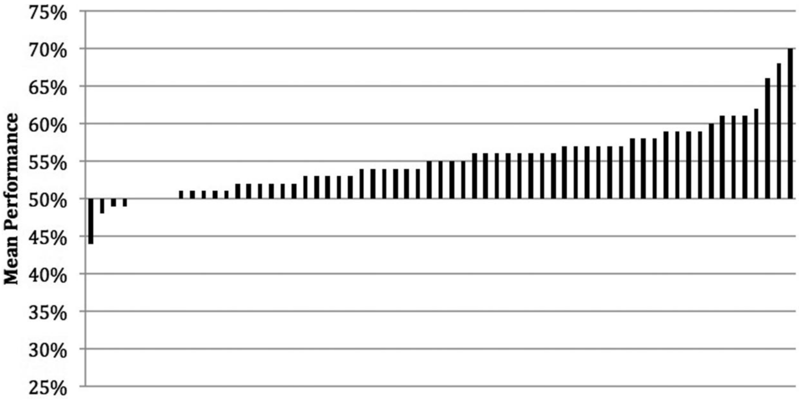

Figure 3. Mean performance in the matching task for each of the models, collapsed across the three conditions. Each bar represents one of the 64 models, ordered from left to right based on performance. The midline (50%) represents chance performance.

whom the voice was recorded at significantly better than chance levels in all three experimental conditions. Performance was better for full sentences than individual words (which still yielded performance significantly better than chance) but did not significantly vary for full or cropped facial pictures. The results of the ratings data did not support the view that participants' ability was due to matching faces and voices along any of the dimensions they rated, as none of the factors significantly correlated with performance.

The current results contradict the conclusions of several previous studies, which report that people could not perform at above chance levels in matching recordings of voices to static pictures of faces (Kamachi, Hill, Lander, & Vatikiotis-Bateson, 2003; Lachs & Pisoni, 2004). Power analysis did not support the possibility that this discrepancy may be due to a lack of statistical power in previous studies. One possible reason for this discrepancy was that these previous studies used a match-to-sample task, in which the faces and voices were presented sequentially. Thus, participants

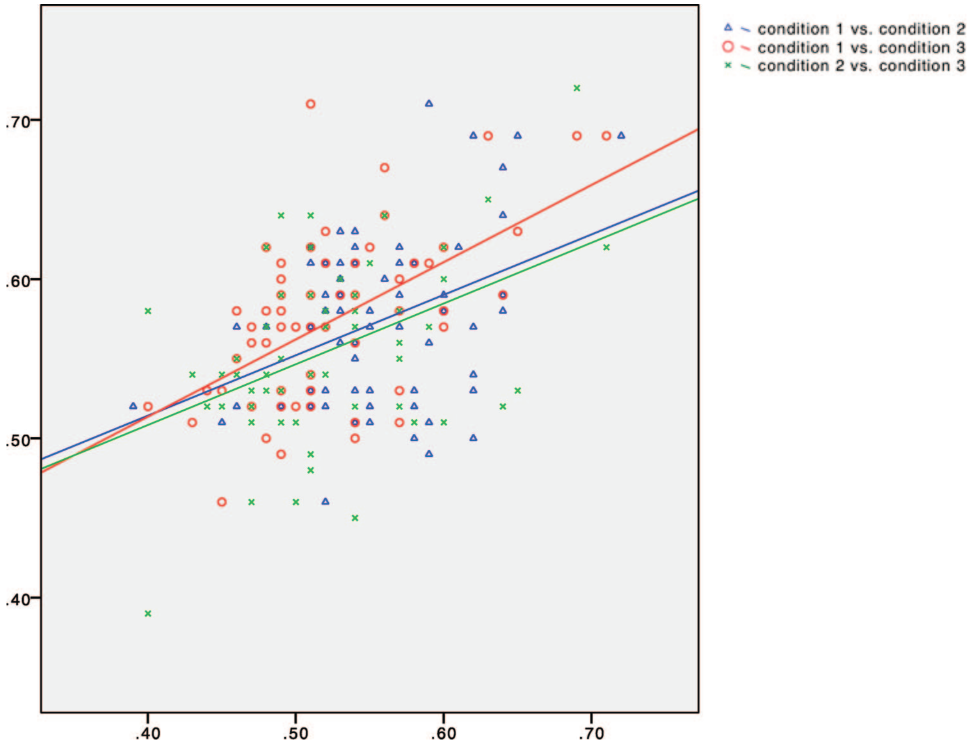

Figure 4. Scatterplot showing the relationship in performance across conditions for the different face-voice models. Each point represents the pairwise mean performance for a single model in two of the experimental conditions while the lines represent the best linear fit of that data.

had to hold the stimuli in memory while trying to match them rather than comparing them simultaneously. Thus, we conducted a follow-up experiment using the same stimuli but employing a sequential match-to-sample task to determine whether the matching ability would persist.

Experiment 2

Method

Participants. Twenty-five psychology undergraduates at Florida Atlantic University who did not participate in Experiment 1, and were naïve to the purpose of the experiment, participated for course credit.

Stimuli and procedure. Stimuli were identical to Condition 1 (Full Visual/Full Audio) in Experiment 1. However, rather than being presented simultaneously, participants first heard the voice recording followed by a 500 millisecond gap, and then were shown both the correct and foil face in random sequential order. The faces were displayed for 1.5 seconds with an intervening gap of 500 milliseconds. Participants chose which face, the first or second, they thought matched the voice they had heard.

Results and discussion. Accuracy was at 55%. A *t* test found that this was significantly above chance, $M = 0.55$, $SD = 0.05$, $t(24) = 5.035$, $p < .001$, $d = 1.03$. Thus, participants maintained the ability to match the faces and voices even when doing so relied on memory, as in previous studies. To determine whether statistical power contributed to these different results, we calculated the minimum sample size needed to observe an effect, and found that it was 15, compared with 15 participants in Kamachi et al. (2003), 20 participants in Lachs and Pisoni (2004), and 25 (per condition) in our study.

General Discussion

The current study found that people could match the faces and voices of unfamiliar people with greater than chance accuracy in both a simultaneous (Experiment 1) and sequential (Experiment 2) matching task. These results contradict the findings of several previous studies, which found no such ability for static images of faces (Kamachi et al., 2003; Lachs & Pisoni, 2004). While the best performance of the three conditions in our study was still not very good (57% correct), this was not far below performance reported in previous studies using dynamic stimuli, which ranged between 60 and 65% correct. In addition, this overall performance measure obscures the presence of systematic relationships between faces and voices. The overwhelming majority of models yielded performance above 50% (although the large number of comparisons reduced the statistical significance of many of these) and some models yielded performance *well* above 50%. In addition, there was a high degree of consistency as to which models yielded good and bad performance across the three conditions. This suggests that participants maintained shared expectations about which facial and vocal properties "belonged together;" what varied was the degree to which each model conformed to those expectations.

What accounts for the overall difference between our results and those of earlier studies? Our effect-size analyses suggest that it is not likely due just to lack of statistical power in their studies. One potentially important difference between our study and that of Lachs and Pisoni (2004) was the small number of models used in their study (eight models) compared with ours (64 models); since our results showed that performance varied widely across models, a small sample could have led to "unlucky" results. However, this factor is less likely to account for the negative results of Kamachi et al. (2003), who used 40 models. One obvious difference between our study and Kamachi et al.'s is that we used self-identified Caucasians as models while their study used Japanese models. Japanese people have been found to spend less time fixating on faces during social interaction compared with other ethnic groups (Argyle & Cook, 1976). This might result in different gaze behavior during an experiment or even reduced expertise in subtle facial characteristics. Alternatively, there may be different vocal information in Japanese voices; for example, Japanese females have been found to use high-pitched voices when speaking in Japanese (Loveday, 1986).

Several theories of speech perception hold that people are specifically sensitive to properties of speech that convey "amodal" information—that is, information concerning physical speech events, such as articulations, that lead to stimuli in multiple sensory domains (Bahrick & Lickliter, 2000; Fowler, 1986; Lachs & Pisoni, 2004; Rosenblum, 2008). The ability to match faces and voices based on gender (Patterson & Werker, 2002) and emotion (Walker-Andrews, 1986) may similarly suggest sensitivity to speech characteristics that convey amodal properties of sex or emotional state. The current results suggest that these sensitivities extend to a set of amodal "identity" properties that manifest in both facial and vocal characteristics. Determining the nature of these properties awaits future research.

References

Allport, G. W., & Cantril, H. (1934). Judging personality from voice. *Journal of Social Psychology, 5,* 37–55. doi:10.1080/00224545.1934.9921582

Argyle, M., & Cook, M. (1976). *Gaze and mutual gaze*. Cambridge, UK: Cambridge University Press.

Bahrick, L. E., & Lickliter, R. (2000). Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Developmental Psychology, 36,* 190–201. doi:10.1037//0012-1649.36.2.190

Berry, D. S. (1991). Accuracy in social perception: Contributions of facial and vocal information. *Journal of Personality and Social Psychology, 61,* 298–307. doi:10.1037/0022-3514.61.2.298

Borkenau, P., & Liebler, A. (1992). The cross-modal consistency of personality: Inferring strangers' traits from visual or acoustic information. *Journal of Research in Personality, 26,* 183–204. doi:10.1016/0092-6566(92)90053-7

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics, 14,* 3–28.

Kamachi, M., Hill, H., Lander, K., & Vatikiotis-Bateson, E. (2003). Putting the face to the voice: Matching identity across modality. *Current Biology, 13,* 1709–1714. doi:10.1016/j.cub.2003.09.005

Krauss, R. M., Freyberg, R., & Morsella, E. (2002). Inferring speakers' physical attributes from their voices. *Journal of Experimental Social Psychology, 38,* 618–625. doi:10.1016/S0022-1031(02)00510-3

Lachs, L., & Pisoni, D. B. (2004). Crossmodal source identification in speech perception. *Ecological Psychology, 16,* 159–187. doi:10.1207/s15326969eco1603_1

Lass, N. J., & Colt, E. G. (1980). A comparative study of the effect of visual and auditory cues on speaker height and weight identification. *Journal of Phonetics, 8,* 277–285.

Loveday, L. (1986). *Explorations in Japanese Sociolinguistics*. Amsterdam, The Netherlands: John Benjamins.

Patterson, M. L., & Werker, J. F. (2002). Infants' ability to match dynamic phonetic and gender information in the face and voice. *Journal of Experimental Child Psychology, 81,* 93–115. doi:10.1006/jecp.2001.2644

Rosenblum, L. D. D. (2008). Speech perception as a multimodal phenomenon. *Current Directions in Psychological Science: A Journal of the American Psychological Society, 17,* 405–409. doi:10.1111/j.1467-8721.2008.00615.x

Walker-Andrews, A. S. (1986). Intermodal perception of expressive behaviors: Relation of eye and voice? *Developmental Psychology, 22,* 373–377. doi:10.1037/0012-1649.22.3.373

Received July 3, 2012
Revision received October 9, 2012
Accepted October 15, 2012 ■