Perception & Psychophysics
2006, 68 (1), 84-93

Hearing a face: Cross-modal speaker matching using isolated visible speech

LAWRENCE D. ROSENBLUM, NICOLAS M. SMITH, SARAH M. NICHOLS,
STEVEN HALE, and JOANNE LEE
University of California, Riverside, California

An experiment was performed to test whether cross-modal speaker matches could be made using isolated visible speech movement information. Visible speech movements were isolated using a point-light technique. In five conditions, subjects were asked to match a voice to one of two (unimodal) speaking point-light faces on the basis of speaker identity. Two of these conditions were designed to maintain the idiosyncratic speech dynamics of the speakers, whereas three of the conditions deleted or distorted the dynamics in various ways. Some of these conditions also equated video frames across dynamically correct and distorted movements. The results revealed generally better matching performance in the conditions that maintained the correct speech dynamics than in those conditions that did not, despite containing exactly the same video frames. The results suggest that visible speech movements themselves can support cross-modal speaker matching.

Historically, it has been assumed that recognizing an individual and recognizing what he or she is saying are two separate perceptual functions. However, recent evidence in both the auditory and the visual perception literatures has challenged this assumption (for reviews, see Nygaard, 2005; Rosenblum, 2005). For example, whether speech is perceived through auditory, visual (lipread), or audiovisual means, familiarity with a speaker facilitates recognition (e.g., Nygaard & Pisoni, 1998; Nygaard, Sommers, & Pisoni, 1994; Schweinberger & Soukup, 1998; Sheffert & Olson, 2004; Sommers, Nygaard, & Pisoni, 1994; Yakel, Rosenblum, & Fortier, 2000). Moreover, there is evidence that, within both modalities, speech and speaker perception can make use of common informational dimensions (e.g., Remez, Fellowes, & Rubin, 1997; Rosenblum et al., 2002).

For audition, it is now known that isolated phonetic information can be used to recognize speakers (Fellowes, Remez, & Rubin, 1997; Remez et al., 1997; Sheffert, Pisoni, Fellowes, & Remez, 2002). Phonetic information (where *phonetic* pertains to the acoustic consequences of the articulation behind a *specific* utterance; Remez et al., 1997) can be isolated using sine wave speech stimuli (Remez, Rubin, Pisoni, & Carrell, 1981). Sine wave speech is a resynthesis technique in which speech signals are reduced to three simultaneously modulating sine waves that track the center formant frequencies of a natural utterance. Over 15 years of research has shown that phonetic information can be recovered from sine wave speech and that these signals can be treated as real speech in supporting classic speech phenomena (Remez, Rubin, Nygaard, & Howell, 1987; Williams, Verbrugge, & Studdert-Kennedy, 1983). More recent research has demonstrated that sine wave speech can also provide speaker information despite the fact that these signals lack fundamental frequency and complex spectral structure—dimensions usually thought necessary to convey speaker identity. Familiar speakers can be identified from sine wave sentences, and sine wave sentences derived from unfamiliar speakers can be matched to the speakers' natural utterances (Remez et al., 1997; Sheffert et al., 2002).

These findings have been interpreted as evidence that the isolated phonetic dimensions available in sine wave speech actually provide articulatory style information specific to speakers (Remez et al., 1997; Sheffert et al., 2002). Thus, although sine wave speech deletes the acoustic dimensions commonly thought to inform about a speaker, a speaker's speaking style is retained in the phonetic structure conveyed by these signals. It has been speculated that the use of common phonetic information for auditory speech and speaker recognition might help account for the contingencies observed between the two functions (Remez et al., 1997).

Analogous observations and arguments have been discussed in the *visual* speech literature (Rosenblum et al., 2002). In research modeled on the sine wave speech work, we found evidence that isolated visible articulatory information can be informative about speakers. To isolate visible articulation, we use a point-light technique (e.g., Bassili, 1978; Berry, 1990; Rosenblum & Saldaña, 1996). This technique involves applying luminous dots to the teeth, tongue, and face of speakers and then filming the speakers

This research was supported by NSF Grant BCS-0111949, awarded to the first author. We gratefully acknowledge the assistance of Stephanie Lin, Olga Sagalovich, and Yen To, as well as the helpful comments of Sonya Sheffert and two anonymous reviewers. Correspondence concerning this article should be sent to L. D. Rosenblum, Department of Psychology, University of California, Riverside, CA 92521 (e-mail: rosenblu@citrus.ucr.edu).

articulating in the dark. The resultant images show only the luminous points moving against a black background; no facial features are visible. Research has shown that visual speech information can be recovered from these images (Rosenblum, Johnson, & Saldaña, 1996) and that they are treated as "true" visual speech stimuli in being automatically integrated with auditory speech (McGurk & MacDonald, 1976; Rosenblum & Saldaña, 1996). Our more recent research has shown that the isolated visible speech information contained in point-light stimuli can also inform about speaker identity (Rosenblum, Smith, & Niehus, 2006; Rosenblum et al., 2002). Despite the fact that these stimuli do not contain the facial feature and configuration information assumed necessary for recognition, point-light speech can be used for face recognition in both matching (Rosenblum et al., 2002) and identification contexts (Rosenblum et al., 2006; see also Bruce & Valentine, 1988). These findings are consistent with other reports showing the salience of dynamic information for face recognition (Christie & Bruce, 1998; Knappmeyer, Thornton, & Bülthoff, 2003; Knight & Johnston, 1997; Lander, Christie, & Bruce, 1999).

In explaining how observers can recognize faces from isolated visible speech, we suggest that like sine wave speech, point-light speech conveys "phonetic" information specific to speakers. Thus, point-light speech might also provide information about the articulatory style specific to individual talkers. Speaking style is, after all, a property of articulation and can potentially structure light as well as sound. Although speaking style is usually considered something to be heard, our point-light speaker recognition experiments suggest that it can also be seen.

If it is true that speaking style can be conveyed in auditory and visual speech signals, an interesting prediction arises. Observers should be able to make cross-modal speaker matches across heard and seen utterances; that is, they should be able to match voices to speaking faces. Intuitively, this might seem to be a difficult task. We are all aware of anecdotal examples in which a person's face does not seem to match his or her voice. Striking instances include the boxer Mike Tyson and the late chef Julia Child. However, these examples might be the exceptions that prove the rule: We are struck by the apparent mismatch for these individuals because, for most people we encounter, voices and faces generally *do* match, especially along dimensions of speaking style. Potentially, then, observers should be able to make cross-modal matches on the basis of speaking style information.

In fact, initial support for this prediction has been provided by four very recent articles reporting research conducted in parallel with our own study (discussed below; Kamachi, Hill, Lander, & Vatikiotis-Bateson, 2003; Lachs & Pisoni, 2004a, 2004b, 2004c). These four articles reported that, with an XAB methodology, observers were able to successfully match voices to faces and faces to voices. On a typical trial, the observers were presented a single voice followed by two silently articulating faces, one of which was that of the same speaker whose voice had been presented (or were presented a single articulating face followed by two voices). The observers were asked to choose which of the faces was that of the speaker whose voice they had heard. All of these studies reported a matching performance that was significantly better than chance for voice-to-face and face-to-voice matching.

The authors of each of these studies explained cross-modal speaker matching in the same general way. They proposed that in both auditory and visible speech signals, there is speaker-specific articulatory information that specifies both phonetic message and speaker (Kamachi et al., 2003; Lachs & Pisoni, 2004a, 2004b, 2004c). It was thought that this articulatory information conveys an idiosyncratic speaking style that can be specified in the time-varying dimensions of both auditory and visual signals (see also Rosenblum, 2004, 2005). Cross-modal matching, then, is based on the recognition of this modality-neutral, common speaking style information available in both modalities. Clearly, this explanation is consistent with those discussed above pertaining to unimodal speaker recognition from isolated auditory and visual speaker information (e.g., Remez et al., 1997; Rosenblum et al., 2002).

All four of the articles reporting cross-modal speaker matching also reported follow-up experiments testing the salience of *audible* speaker-specific phonetic information. These auditory tests involved either deleting the phonetically extraneous information from the signal (e.g., sine wave speech; Kamachi et al., 2003; Lachs & Pisoni, 2004b, 2004c) or testing whether acoustic distortions that hinder phonetic perception also hinder cross-modal matching (Lachs & Pisoni, 2004a). The results of these tests support the conclusion that the idiosyncratic phonetic information contained in the auditory signal is salient for cross-modal matching.

With regard to the salience of *visible* speaking style for cross-modal matching, three of the four studies (Kamachi et al., 2003; Lachs & Pisoni, 2004b, 2004c) reported experiments designed to test whether visible articulatory movements are important for these purposes. In normal speaking face displays, there is necessarily more information than just articulatory movement information available that could support matching judgments. Normal videos of speaking faces also contain pictorial face information that could potentially support voice matching (e.g., along dimensions of age, size, or attractiveness; see Collins & Missing, 2003). To help rule out this explanation of their results, Kamachi et al. (2003), as well as Lachs and Pisoni (2004b), tested video stimuli composed of static faces, as well as of articulating faces presented in reverse. In finding that neither of these control stimuli supported matching, both sets of authors concluded that the successful matching they observed with normal speaking faces was likely based on visible articulatory movement information (Kamachi et al., 2003; Lachs & Pisoni, 2004b).

However, it can be argued that both the static and the reversed display control tests conducted by Kamachi et al. (2003) and Lachs and Pisoni (2004b) were incomplete. With regard to static faces, it could be that poor matching performance was based not on a lack of speech movements in the displays, but on the fact that each of the displays was composed of only a single frame. In contrast, normal dynamic video clips are composed of 30 frames per second (fps). Thus, differences in performance between static and dynamic video displays might be based on the specific frames and/or the number of frames available, rather than on any portrayed movement per se. This argument will be elaborated later (see also Lander et al., 1999; Rosenblum et al., 2002). With regard to the displays shown in reverse, both Kamachi et al.'s (2003) and Lachs and Pisoni's (2004b) studies also reversed the auditory stimuli in these tests. Consequently, the observed poor matching could be attributable to the distortions induced by reversing the visual, the auditory, or both types of information. (Although there is research showing that reversed auditory speech can retain speaker information [e.g., Bricker & Pruzansky, 1966; Sheffert et al., 2002; Van Lancker, Kreiman, & Emmorey, 1985], it is unclear whether reversed auditory speech is sufficient for cross-modal matching.) Thus, it can be argued that these studies failed to thoroughly establish whether visible speech *movements* play a critical role in cross-modal speaker matching.

Another recent project (Lachs & Pisoni, 2004c) implemented a point-light methodology to examine the importance of visible speech movements for cross-modal speaker matching. Lachs and Pisoni (2004c) applied 28 point-lights to the faces of 4 female speakers. The arrangement of point-light positions was the same for all the speakers. The 4 speakers were filmed producing a series of words under the lighting and videotaping conditions discussed above for point-light stimuli. These video clips were then presented for XAB speaker matching, along with the audio words derived from the same utterances as those that were videotaped. The results showed that the observers could match the point-light faces to the audio words, whether the audio stimuli were presented intact or were transformed to sine wave replicas.

Although these results are consistent with the proposed importance of visible speech movements, they cannot be conclusive in this regard. For example, although the point-light methodology certainly helps to isolate visible movement, Lachs and Pisoni (2004c) did not conduct control conditions to ensure that matches were not made on the basis of extraneous, nonmovement information still available in point-light images. It is known that there can be some structural/static information available in point-light stimuli, especially when the same general point-light arrangement is used for all speakers (Rosenblum et al., 2002). Furthermore, even if the images contain no useful structural information, research in our laboratory has shown that using a single arrangement of points can allow observers to perform matches by noting repeated pairings of stimuli (Rosenblum et al., 2002). For example, in a cross-modal speaker-matching experiment, a specific speaker's voice is presented together (in an XAB trial) with that speaker's face substantially more often than with any other face. If a facial image contains a specific arrangement of points that is maintained across the experiment, the subjects can note that this arrangement is paired most often with a specific voice and then make matches accordingly. In such cases, results showing successful speaker matching would be attributable to sensitivity to repeated pairings of stimuli, on the basis of superficial image features, rather than to articulatory dynamics.

A final reason that the extant findings are inconclusive is that when observers are asked to match a speaking face to an audio token derived from exactly the same recording, these matches can be made on the basis of some superficial dimensions of each utterance (e.g., duration). For all of these reasons, the point-light experiments of Lachs and Pisoni (2004c), like the control conditions used by Kamachi et al. (2003) and Lachs and Pisoni (2004b), fail to provide conclusive support that visible speaker movement information is sufficient for cross-modal speaker matching.

To summarize, although there is some initial evidence suggesting that cross-modal speaker matches can be made on the basis of visible speaking style, more evidence is clearly needed. This is particularly important because, as was discussed above, the extant theoretical accounts of matching all propose a basis in common audible *and visible* articulatory style information.

In the experiment reported below, conditions were designed to more thoroughly test the salience of visible speech movements for cross-modal speaker matching. As in Lachs and Pisoni's (2004c) study, a point-light methodology was implemented. However, because our main interest was in testing the salience of visible movement information, we included a number of critical control conditions and image manipulations to isolate the contribution of movement.

Five conditions were tested. The first condition involved a straight examination of whether voices can be matched to unaltered video clips of 10 silently speaking point-light faces in an XAB context. Next, in order to provide an initial test of the utility of point-light movements, a second condition involved testing whether voices could be matched to single, static frames taken from the point-light video clips. Historically, static frame control stimuli have been used as a way to determine whether there is salient pictorial information retained in point-light videos (e.g., Bassili, 1978; Johansson, 1973; Kamachi et al., 2003; Lachs & Pisoni, 2004b).

However, as has been mentioned, a static frame control is not sufficient to rule out the possible utility of the static information contained in dynamic stimuli (Lander et al., 1999; Rosenblum et al., 2002). Specifically, static stimuli are composed of a single frame, whereas dynamic stimuli are composed of 30 fps. Even if the presentation durations of a static and a dynamic stimulus are equalized, a dynamic stimulus contains 30 times more static frames than does a static stimulus. Thus, a dynamic video condition is distinguished from a static condition not just by the presence of movement information, but also by the number of individual frames presented. In this sense, observing greater matching performance on a dynamic versus static condition cannot, in and of itself, demonstrate the informativeness of speaker-specific articulatory movements. As has been stated, this confounding factor was not fully considered in the cross-modal matching experiments reported by Kamachi et al. (2003) and Lachs and Pisoni (2004b, 2004c).

For these reasons, three additional conditions were tested in our experiment. Borrowing from previous research (Lander et al., 1999; Rosenblum et al., 2002), in these three conditions, the specific frames making up each set of stimuli were equated, but the conditions differed in frame ordering and relative frame duration. In order to make these frame manipulations manageable, the frame rate of the original videos was reduced from 30 to 10 fps. This new frame rate was chosen on the basis of previous research showing that although the temporal quality of such video images degrades somewhat, these images can still convey salient facial movement information (e.g., Blokland & Anderson, 1998; Rosenblum et al., 2002; Vitkovitch & Barber, 1994). Reducing the original videos to this frame rate while keeping the frame ordering and relative timing intact produced a *reduced dynamic* condition. It was thought that these reduced dynamic stimuli would maintain much of the characteristic articulatory movement information useful for cross-modal matching.

In the fourth stimulus condition, the frames and frame durations were exactly the same as those in the reduced dynamic condition, but these frames were presented in a random order. This *jumbled* condition was designed as a comparison control condition to test the importance of the dynamics maintained in the reduced dynamic condition stimuli. In a fifth condition, again exactly the same frames as those in the reduced dynamic condition were used, as well as the frame ordering of that condition, but the relative duration for which each frame was shown varied (quasirandomly). Thus, these tokens presented the correct sequence of movement but, because of the varied timing, lacked much of the correct dynamic information (Rosenblum et al., 2002).

If there is information in the characteristic dynamics of visible speech movements that supports cross-modal speaker matching, superior performance would be expected in stimulus conditions that maintain these dimensions. Specifically, it would be expected that matching performance would be superior (and greater than chance) in the dynamic and reduced (but dynamically appropriate) conditions, relative to the static, jumbled, and staggered conditions.

In addition to these important control conditions, a number of necessary image manipulations were implemented through the point-light application process itself (see also Rosenblum et al., 2002). These manipulations (discussed below) were designed to prevent any matching based on (1) structural/static information available in the point-light arrangements, (2) a strategy noting repeated pairings of superficial dimensions, and (3) superficial commonalities across the audio and video tokens derived from the same recording. As has been mentioned, these possible matching strategies were not fully addressed in the point-light studies conducted by Lachs and Pisoni (2004c).

METHOD

Subjects

The subjects were 85 undergraduates from the University of California, Riverside. Their ages ranged from 18 to 24 years. All were given credit in order to fulfill a requirement for an introductory psychology course. All were native speakers of English and reported normal hearing and normal or corrected-to-normal vision.

Materials

The stimuli were spoken by 5 female and 5 male native American English speakers (20–25 years old) articulating the sentence "The football game is over." The use of 10 speakers (rather than 4; Lachs & Pisoni, 2004c) should help reduce a subject strategy of noting repeated pairings on the basis of superficial visual features. The particular sentence was chosen because previous research had found it to be particularly easy to lip-read (Rosenblum et al., 1996). The speakers were instructed to articulate this sentence in as natural a manner as possible and were not directly informed of what these recordings were designed to test.

Point-light images were created by affixing 30 reflective dots on the face of each speaker. These dots were made of 3-mm-diameter construction paper covered with fluorescent yellow paint (see Rosenblum et al., 2002, for further details). The dots were affixed to the skin with a medical adhesive and to the teeth and tongue with a dental adhesive (see below). The dots were small enough so as to not interfere with articulation but were large enough so that, when filmed under fluorescent black lights, the movements of the dots could be clearly seen. The speakers were filmed under black light illumination using a 24-in., vertical 10-W black fluorescent light bulb positioned about 3 ft away. This filming technique produced video images in which only the dots and their movements could be seen against a black background; no facial features were visible in the resultant images.

The selection of dot positions on the articulators was based on a number of considerations. First, it was important that the dot positions should convey good visual speech information. Following previous research in which salient dot placement has been examined (e.g., Rosenblum et al., 1996), 15 dots were placed on the cheeks, jaw, and forehead of each speaker's face. In addition, 15 dots were placed on various locations of the teeth, tongue, and lips (see Rosenblum et al., 2002, for further detail). Second, dot positions were chosen so as not to convey facial structure information when the images were shown statically. For these purposes, many of the points were placed at random positions on the face, lips, and teeth. These random positions were different for each speaker. To further make identifications on the basis of static/structural point-light image information more difficult, the speakers were videotaped with their faces positioned in a wooden frame that was covered by a random array of dots. Finally, in order to prevent a strategy based on noting repeated pairings of a speaker's voice with a particular arrangement of dots, nine different dot position arrangements were used for each speaker.

A Sony digital video camcorder (DRC-TRV11), positioned 6 ft in front of the speakers, was used to video and audio record all utterances. For recording of the audio stimuli, a Shure SM57 microphone, positioned 1 ft in front of and below the speakers' mouths, was connected directly to the camcorder. Each speaker was asked to repeat the test sentence "The football game is over" multiple times under each dot arrangement. Ultimately, nine different visible utterances (one for each dot arrangement) and one auditory utterance of the test sentence were used for each speaker. For each speaker, the auditory token was derived from a different utterance than was any of the nine point-light stimuli used in the experiment. Moreover, these audio tokens were derived from utterances recorded at least 1 h earlier than the point-light stimuli recordings. Using different utterances of the test sentence for the audio and visual stimuli for each speaker prevented cross-modal matching strategies based on superficial properties of specific utterances, such as exact duration (e.g., Remez et al., 1997; Rosenblum et al., 2002).

For each speaker, one audio and nine video stimuli were digitally captured onto a Macintosh G4 computer for editing and establishment of presentation order. Adobe Premiere video editing software was used for this procedure. For all 10 speakers, the duration of each point-light and auditory utterance was less than 3 sec, allowing all the video stimuli to be edited to a presentation length of 3 sec.

The stimuli for all five presentation conditions were ordered for XAB presentation. Each XAB triad involved the presentation of an auditory utterance followed by a 1-sec interstimulus interval (ISI), which was followed by the presentation of two successive point-light stimuli (separated by a 1-sec ISI), one of which depicted the articulation of the speaker who produced the auditory sentence.[1] Each triad was separated by a 3-sec ISI. During all ISIs and auditory presentations, the video screen was black.

Each stimulus condition series involved 10 auditory utterances (1 for each speaker) and 90 point-light images (9 for each speaker). Each speaker's auditory utterance was presented in combination with each of the other speakers' 9 point-light images twice: once when it matched the 1st point-light image in a triad, and once when it matched the 2nd point-light image in the triad. This resulted in 180 XAB (10 speakers × 9 sentences × 2 orderings) randomized triad presentations for each of the five stimulus conditions.

The five stimulus conditions were derived from the digitized video clips described above. In the dynamic condition, the visual stimuli were composed of the moving point-light video clips, edited and digitized in the manner described. In the static stimulus condition, the visual stimuli were composed of a single frame from the point-light clips. Static frames were chosen to depict a neutral mouth position (Rosenblum et al., 2002) and were shown for the same duration as the moving stimuli from which they were derived (3 sec).

The reduced dynamic stimulus condition involved frame-rate-reduced versions of the dynamic presentation condition stimuli. Adobe Premiere software was used to edit the dynamic stimuli on a Macintosh G4 computer. The dynamic stimuli were edited so that the 2nd and 3rd of every 3 frames were deleted and replaced by repetitions of the 1st frame of the 3. This resulted in dynamic stimuli that were composed of 10 sets of 3 repeated frames. This editing technique produced new dynamic sentences that changed at a frame rate of 10 fps, as opposed to the usual 30 fps. The duration of each token was 3 sec, so that each visual stimulus consisted of 90 frames total. These reduced stimuli preserved the correct sequence and timing of the movement but displayed that movement at a reduced frame rate.

The jumbled stimuli were created directly from the reduced dynamic stimuli. For the jumbled stimuli, exactly the same frames as those for the reduced stimuli were displayed, but in a random order. Thus, in the jumbled condition, exactly the same frames and frame rate were displayed as in the reduced stimuli condition, but with an incorrect sequence of movements.

The staggered stimuli were also created from the reduced dynamic stimuli. For these stimuli, exactly the same frames as those used for the reduced stimuli were manipulated so that the duration of each frame varied from 30 to 90, 150, 210, or 360 msec. This was accomplished by presenting each frame once or by repeating these frames in clusters of 3, 5, 7, or 9, so that the full duration of these stimuli were 3 sec. Thus, the *average* frame duration and frame rate of these stimuli were the same as those of the reduced dynamic stimuli. This resulted in tokens that displayed the same frames in the same order as in the reduced dynamic stimuli condition, but the frames were displayed for varying degrees of time. Thus, these tokens presented the correct sequence of movement but, because of the varied timing, likely lacked much of the correct dynamic information.

The dynamic and static condition stimuli were recorded (in random order) onto videotapes, one for each of the presentation conditions. The reduced, jumbled, and staggered stimuli were presented (in random order) directly from the computer through the same video monitor (and with the same image size and spatial resolution) as the dynamic and static stimuli. General video and audio quality was judged to be comparable across the videotaped and computer-presented conditions. All the presentation sequences contained the same ISIs and total durations, regardless of whether they were presented by video or computer.

Procedure

The subjects were randomly assigned to the dynamic (18 subjects), static (18 subjects), reduced dynamic (15 subjects), jumbled (16 subjects), or staggered condition (18 subjects).

All the subjects were tested individually in an 11 × 9 ft room while seated at a table facing a 20-in. Panasonic Color Video Monitor (CT-2010Y) at a distance of 5 ft. On this monitor, the contrast was adjusted, and the color was turned off in order to maximize the appearance of the point-light images. Auditory speech stimuli were presented over loudspeakers positioned on each side of the video monitor.

The subjects were instructed to closely attend to each triad and then to decide which of two point-light images matched the sentence they had heard. They were told that the spoken sentence would be "The football game is over." The subjects in the dynamic and static presentation conditions were instructed to record their answers by circling the number "1" or "2" (first or second point-light image) on an answer sheet. The subjects in the reduced, jumbled, and staggered presentation conditions responded by pressing a key on a computer keyboard. These subjects were instructed to press a key labeled "1" if the auditory token matched the first point-light stimulus of the triad and to press a key labeled "2" if the auditory token matched the second point-light stimulus.

The subjects in all five presentation conditions were reminded to wait until the end of the second point-light presentation before recording their choice. They were told that the task would be difficult but to give their best guess throughout the experiment. Before the critical portion of the experiment was administered, the subjects were presented with 10 practice trials consisting of one triad for each of the 10 speakers. These practice trials involved dynamic, static, reduced, jumbled, or staggered visual stimuli, depending on the condition to which the subject was assigned. No feedback was provided for these or any other trials in the experiment. The experiment lasted approximately 1 h 20 min for each subject.

RESULTS

Means were calculated for the correct responses for the conditions and speakers. The mean percentages correct and standard errors (pooling over the speakers and subjects) for the five presentation conditions can be seen in Figure 1. The ranges of the 10 speaker means for the presentation conditions can be seen in Table 1. It should be mentioned that these performance means are fully comparable to the performance means observed in the other experiments conducted on cross-modal speaker matching (Kamachi et al., 2003; Lachs & Pisoni, 2004a, 2004b, 2004c).

An ANOVA was conducted on the factors of stimulus condition (5 levels) and speaker (10 levels). The ANOVA revealed a significant effect of presentation condition [$F(4,80) = 11.66$, $MS_e = 363.58$, $p < .0001$] and speaker [$F(9,720) = 22.49$, $MS_e = 193$, $p < .0001$], as well as a significant interaction of these factors [$F(36,720) = 4.9$, $MS_e = 193$, $p < .0001$]. Post hoc Fisher (PLSD) tests conducted at the level of presentation condition revealed a significant difference between the dynamic stimulus con-

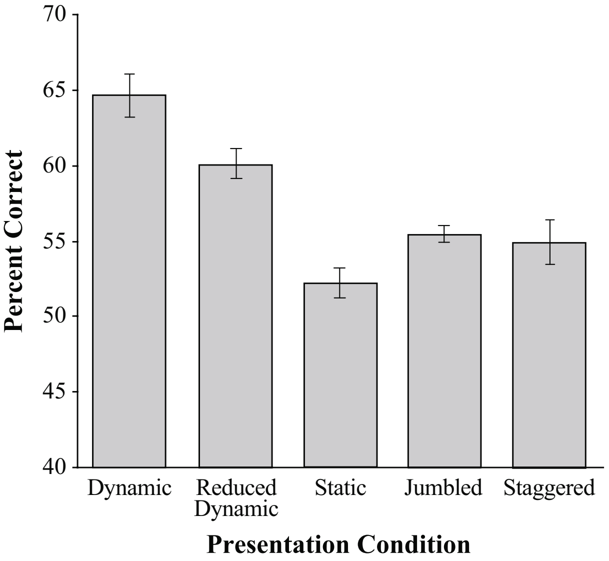

Figure 1. Mean percentage correct (and standard error) values for dynamic, reduced dynamic, static, jumbled, and staggered conditions.

dition and each of the other four conditions ($p < .05$). In addition, this analysis revealed a significant difference between the reduced dynamic stimulus condition and each of the static, jumbled, and staggered conditions. No significant difference was observed between the static, jumbled, and staggered conditions. The results of this analysis are supportive of our hypothesis that matching performance with point-light stimuli that maintain intact movement information (the dynamic and reduced dynamic conditions) should be better than performance with stimuli not containing this information (static, jumbled, and staggered conditions). Furthermore, in revealing significant differences between the reduced dynamic and the jumbled and staggered conditions, these results suggest that this superior performance was not based on any static frame information contained in these stimuli.

Additional analyses (one-sample *t* tests) were conducted to determine under which conditions and for which speakers the subjects were able to make matches at greater than chance levels (50% correct; Remez et al., 1997; Rosenblum et al., 2002). For the dynamic condition, the subjects matched 8 of 10 speakers at greater than chance levels ($p < .05$); for the static condition, 2 of 10 speakers were matched at better than chance levels ($p < .05$), with a third matched at better than chance at the $p = .055$ level. For the reduced dynamic condition, 6 of 10 speakers were matched at better than chance levels; for the jumbled condition, 2 of 10 speakers were matched at better than chance levels, and for the staggered condition, 5 of 10 speakers were matched at better than chance levels.

These *t* test results are also *generally* supportive of our hypotheses that dynamically intact point-light stimuli would support superior matching performance. However it should be noted that on the basis of this analysis, the staggered stimulus condition also seemed to provide some information for speaker matching. Although overall performance on the staggered condition was inferior to that of the dynamic and reduced dynamic conditions (on the basis of the post hoc test), half of the speakers in the staggered condition were recognized at better than chance levels. Thus, although there seems to have been more information for cross-modal matching in the dynamic and reduced dynamic displays, there was some useful information available in the staggered stimuli to support matching for half of the speakers.

What type of information available in the staggered stimuli supported matching, and might this information be similar to that available in the dynamic and reduced dynamic stimuli? The staggered stimuli maintained the same frames and frame ordering as those in the reduced dynamic condition, but not the same frame timing. It seems unlikely that the information available in the individual frames themselves could support successful matching, given that the jumbled condition also contained exactly these same frames but elicited above-chance performance only for 2 of the 10 speakers. Instead, the inclusion of correct frame ordering, despite variation in frame timing, may have allowed the staggered stimuli to capture some aspect of the dynamics for some speakers. Although the frame duration was designed to vary randomly across each utterance, it could be that, for some speakers, particularly

Table 1
Range of Percentages of Correct Means for Speakers and Presentation Conditions

	Presentation Condition				
Speaker	Dynamic	Reduced Dynamic	Static	Jumbled	Staggered
1	44.4–88.9	38.9–77.8	38.9–77.8	22.2–72.2	38.9–77.8
2	38.9–94.4	27.8–83.3	33.3–88.9	33.3–94.4	22.2–66.7
3	44.4–94.4	50.0–94.4	33.3–77.8	22.2–88.9	27.8–77.8
4	38.9–88.9	38.9–83.3	33.3–88.9	27.8–77.8	44.4–83.3
5	44.4–94.4	50.0–94.4	27.8–72.2	27.8–72.2	61.1–100.0
6	16.7–72.2	16.7–77.8	11.1–66.7	16.7–88.9	16.7–66.7
7	55.6–94.4	33.3–88.9	11.1–66.7	38.9–83.3	44.4–83.3
8	44.4–83.3	38.9–72.2	27.8–66.7	44.4–94.4	33.3–61.1
9	27.8–83.3	22.2–61.1	11.1–72.2	22.2–88.9	27.8–61.1
10	55.6–100.0	33.3–100.0	44.4–83.3	33.3–94.4	22.2–88.9

salient parts of the utterance were presented with correct enough timing to allow cross-modal matches.

In fact, there is evidence in the point-light perception literature that event recognition can be relatively tolerant of variations in stimulus presentation rate (Pavlova, 1995), as well as interframe interval (Thornton, Pinto, & Shiffrar, 1998). At the same time, gender recognition of point-light walkers can be disrupted by substantially slowing stimulus presentation rate (Barclay, Cutting, & Kozlowski, 1978). More relevant to the present study, exactly the same staggered frame manipulation proved to be more effective at disrupting performance in a task in which fully illuminated faces were matched to point-light faces (Rosenblum et al., 2002). Thus, depending on the task and stimulus type, variations in stimulus video frame rate can have a greater or lesser influence on event recognition.

Returning to the present findings, an analysis was conducted to examine whether the staggered stimuli might offer observers the same type of salient information—or support the same type of strategies—as the dynamic and reduced dynamic stimuli. For these purposes, a regression test was performed on the speaker-matching means across all of the stimulus conditions. This analysis revealed a significant relationship between the speaker means across the dynamic and the reduced dynamic conditions [$r = .86$; $F(1,8) = 23.58$, $p = .0013$]. In addition, this analysis revealed a significant relationship between the means of the dynamic and the staggered conditions [$r = .75$; $F(1,8) = 10$, $p = .013$], as well as between the reduced dynamic and the staggered condition means [$r = .91$; $F(1,8) = 39.63$, $p = .0002$]. No other speaker mean correlations between conditions proved to be significant at the $p < .05$ level. These regression tests show that the *relative* matching performance across the speakers was similar for the dynamic, reduced dynamic, and staggered conditions. This could mean that despite the fact that overall matching performance was better for the dynamic and reduced dynamic conditions, matches with the staggered stimuli might have made use of information and/or strategies similar to those used in those two superior conditions.

In future research, the degree to which video stimuli with variable frame rates can still provide movement information for cross-modal speaker matching can be examined. However, it should be emphasized that overall performance with the dynamic and reduced dynamic stimuli was significantly better than that with the staggered stimuli and that overall performance with the staggered stimuli was statistically no better than that with the jumbled and static stimuli. Furthermore, on the basis of poor overall and speaker mean performance in the jumbled condition, it is unlikely that the useful information in the staggered condition was contained in individual frames.

DISCUSSION

The results of these experiments suggest that visible speech movements can support cross-modal speaker matching. The results revealed that full-frame-rate videos of point-light speakers, as well as reduced-frame-rate, dynamically intact videos, could be matched to voices for a majority of the speakers tested. In addition, the results showed that these two dynamically intact stimulus sets could be matched to voices at levels significantly greater than could stimuli that contained the same image frames but lacked dynamical information.

These results buttress the recent evidence that speakers can be matched cross-modally with some success and that visible speech movements can support this matching (Kamachi et al., 2003; Lachs & Pisoni, 2004a, 2004b, 2004c). As has been discussed, although in three of the studies reporting cross-modal matching (Kamachi et al., 2003; Lachs & Pisoni, 2004b, 2004c) an attempt was also made to demonstrate a visible speech basis, each of them was incomplete in isolating this property of the stimuli. The point-light technique used here, along with the critical frame-matched control conditions, offer stronger evidence that visible speech movements can be used for cross-modal matches. Whereas Lachs and Pisoni (2004c) also used point-light stimuli in their study, the present experiment added a number of important control measures to prevent the use of static/structural information often contained in point-light displays. These measures included the use of different utterances for the audio and video components for each speaker, to prevent cross-modal matching based on superficial stimulus characteristics (e.g., duration; Rosenblum et al., 2002; see also Remez et al., 1997). In addition, in the present study, multiple, and "random" point configurations were used for each speaker, in order to prevent matching based either on static/structural face information or on the noting of repeated pairings of voice and face stimuli on the basis of superficial image features (Rosenblum et al., 2002). Finally, the use of 10, rather than 4, speakers (as in Lachs & Pisoni, 2004c) for our stimulus set would seem to be an important addition, especially in light of the speaker differences observed across the image conditions. The fact that our results still generally show a performance advantage for dynamically intact visible speech stimuli suggests that these control measures were effective in preventing matches based on static/structural image information (which also would have been available in the dynamically distorted stimuli).

At the same time, it must be acknowledged that intrinsic to this type of matching task, subjects could perform judgments not by matching on similarity, but by using a process of elimination for which the *least* similar token is judged. Thus, in the present experiment, the subjects could, on each trial, look for the point-light stimulus that was *less* like the voice along any number of dimensions. In this strategy, subjects base judgments on *dissimilarity*, which is not equivalent to matching based on voice–face correspondences. Although the present experiment cannot preclude this possibility, it should be noted that this dissimilarity strategy could account for data in any matching experiment of this nature, including the experiments reported by Kamachi et al. (2003) and Lachs and Pisoni (2004a, 2004b, 2004c).

Informational Support for Cross-Modal Matches

On what basis can subjects make cross-modal speaker matches? One possibility is that the observers were recognizing visible speakers by extracting general facial-structure-from-motion information (e.g., Braunstein, Hoffman, & Pollick, 1990; Christie & Bruce, 1998; Hildreth, Grzywacz, Adelson, & Inada, 1990; Pollick, 1997) and then matching it to the heard voice that best fit that facial structure. Although the results reported above are consistent with this interpretation, a post hoc analysis revealed results less consistent with a structure-from-motion basis. This analysis examined whether speaker gender information might have played a role in matching judgments. Our point-light design allowed for a target face to be presented in a trial with any of the remaining nine faces as distractors, including faces of opposite gender. On the basis of this design, a test was conducted to determine whether the subjects performed more accurately on trials that involved distractors of opposite versus same gender faces. If it is assumed that the gender of a face is relatively easy to extract from facial structure (e.g., Burton, Bruce, & Dench, 1993), it was reasoned that if a structure-from-motion strategy was used, our subjects would perform better on trials with opposite versus same gender distractors. However, an analysis (*t* test) on the responses from the dynamic presentation condition revealed no such performance difference ($p > .05$). This result suggests that our subjects did not base their matching judgments on the apparent gender of the point-light faces and provides indirect evidence against the structure-from-motion explanation. These results are also consistent with the sine wave speech findings of Remez et al. (1997), who reported confusion data indicating that listeners did not base speaker recognition judgments on gender information.

The structure-from-motion explanation for cross-modal matching is also less consistent with previous findings that fully illuminated, stationary face images cannot be easily matched to voices (Kamachi et al., 2003), as well as with findings that sine wave speech stimuli, which have little anatomically related acoustic dimensions, *can* be used for cross-modal matches (Kamachi et al., 2003; Lachs & Pisoni, 2004c). Finally, the structure-from-motion explanation would seem relatively unparsimonious in requiring multiple translations across different representational forms (see also Lander et al., 1999, and Knappmeyer et al., 2003, for evidence against the structure-from-motion interpretation).

A more parsimonious interpretation of cross-modal speaker matching has been offered by the authors of the four previously published articles on the topic (Kamachi et al., 2003; Lachs & Pisoni, 2004a, 2004b, 2004c). As was stated above, these authors suggested that cross-modal speaker matches are based on the recognition of phonetically realized speaking style information contained in both the auditory and the visual signals. As support, each of the prior cross-modal speaker matching studies provided some demonstration of the salience of speaker-specific acoustic information available at the phonetic level. Furthermore, three of these studies, in addition to our own, have provided demonstrations of the utility of visible speech movements for cross-modal matching. What has not yet been determined is whether the salient information contained in visible speech movements takes the form of idiosyncratic speaking style per se. Although the present study helps establish that speech movements can be used for matching, it cannot directly address whether the salient information in those movements is contained in visible speaking style.

Directly addressing this issue will likely require a thorough motion analysis of visible speech stimuli, which is a very time intensive endeavor and goes beyond the scope of the present project. Moreover, such an analysis will involve designing stimuli for exactly this purpose—for example, point-light stimuli that provide easily tracked points positioned similarly across all utterances and speakers. In contrast, the visual stimuli used in the present study involved multiple and varying point positions for each speaker, making detailed visible motion analyses unmanageable.

Still, the present stimuli and data do lend themselves to a preliminary examination of a speaking style basis for cross-modal matching. If observers are basing cross-modal matches on audible and visible speaking style information, some speaker-related informational dimensions should be observable across visual and auditory media. Furthermore, speakers whose signals portray these dimensions more dramatically should, in principle, be easier to match. As has been intimated, a thorough search for these common cross-modal dimensions would involve stimuli designed for these purposes (e.g., ones allowing point-tracking). For the present study, however, this issue can be examined by looking at some of the more coarse-grained and conspicuous cross-modal signal commonalities and whether they are predictive of speaker-matching performance.[2] Along these lines, we conducted two sets of simple analyses examining potential correspondences across each speakers' auditory and visual stimuli and whether these correspondences predicted success of matching the different speakers.

First, it could be that observers relied upon the general durational properties of the speakers' sentences to make cross-modal matches. Although efforts were made to ensure that the *exact* duration of a speaker's visible and auditory utterances were not the same, there were natural durational differences between and among the 10 utterances (9 visual, 1 auditory) of each speaker. Possibly, then, the observers could have made matches on the basis of interspeaker durational differences. If this were the case, it would be expected that the speakers whose nine visible utterance durations were more similar to their auditory sentence duration and/or less variable overall would be more easily matched than would the speakers whose sentence durations did not show these consistencies. To examine this possibility, a series of simple regression analyses were conducted using speakers' auditory and visual sentence durations, as well as their mean matching scores from the dynamic condition. The details of this analysis are reported on the Web site (www.faculty.ucr.edu/%7Erosenblu/HearingFaceAnalyses). Not surprisingly, a significant correlation was observed between the auditory and the mean visual sentence duration for the speakers ($p < .05$). However, neither the degree of cross-modal duration similarity nor the consistency of these values correlated significantly with matching accuracy across the speakers. Thus, speakers who were more easily identified were not necessarily those whose auditory and mean visible sentences durations were more similar or more consistent.

We chose next to evaluate whether the audible and visible extent of speakers' articulations might provide a basis for the cross-modal matching. Potentially, a speaker's characteristic *extent*, or general magnitude of articulation, could be reflected cross-modally in the acoustic amplitudes and visible articulatory displacements. If correspondences in extent measures did exist for the speakers, observers might have been able to make cross-modal matches on the basis of this dimension. To address this question, a series of extent/magnitude measurements were made on the audible and visible sentence stimuli, and then tests were conducted to determine whether these relationships predicted cross-modal matching performance for the speakers in the dynamic condition. The details of this analysis are reported on the Web site (www.faculty.ucr.edu/%7Erosenblu/HearingFaceAnalyses). The results revealed no significant correlation between the speakers' mean visual extent and their (overall and peak) auditory sentence amplitudes. Furthermore, none of these acoustic and visual extent measures, on their own, was correlated significantly with cross-modal matching performance for the speakers. Finally, a simple extent variability measure for the visible sentences also failed to significantly correlate with the dynamic condition's matching means for the speakers. Thus, on the basis of these preliminary measures of visible and audible extent, it seems that cross-modal information for this dimension does not underlie successful matching.

Failing to find predictive power in the duration and extent dimensions should not be interpreted as evidence against a visible speaking style basis for cross-modal matches. There are a multitude of articulatory style dimensions that could appear in the cross-modal signals, many of which exist on a more fine-grained scale than do the two measures examined above. For example, both Kamachi et al. (2003) and Lachs and Pisoni (2004a, 2004b, 2004c; see also Remez et al., 1997; Rosenblum et al., 2002; Sheffert et al., 2002) have speculated that the relevant information is contained in the idiosyncratic fine-grained phonetic realizations of a speaker's articulations. As has been stated, there is mounting evidence that the *acoustic* information for cross-speaker matching lies at the phonetic level (Kamachi et al., 2003; Lachs & Pisoni, 2004b, 2004c). In explaining sine wave speaker recognition effects, Remez et al. (1997) speculated that the salient dimensions lie at the segmental level, possibly in the coarticulatory assimilation (e.g., of consonants) specific to speaking style (see also Amerman & Daniloff, 1977; Bladon & Al-Bamerni, 1976; Sheffert et al., 2002).

If the speculation that the relevant information lies at the fine-grained phonetic level also applies to visual speech, it would not be surprising that our duration and extent measures were not predictive of matching performance. These two measures are neither fine-grained nor phonetically relevant.[3] However, as has been stated, our choice of analyses was constrained by visual stimuli that lend themselves only to more macroscopic tests. Future research conducted with visible stimuli designed for more fine-grained analyses (e.g., of coarticulatory assimilation) should be revealing about whether the salient cross-modal dimensions for speaker matching truly lie with modality-neutral information for articulatory style (Kamachi et al., 2003; Lachs & Pisoni, 2004a, 2004b, 2004c; Rosenblum, 2004, 2005).

REFERENCES

Amerman, J. D., & Daniloff, R. G. (1977). Aspects of lingual coarticulation. *Journal of Phonetics*, **5**, 107-113.

Barclay, C. D., Cutting, J. E., & Kozlowski, L. T. (1978). Temporal and spatial factors in gait perception that influence gender recognition. *Perception & Psychophysics*, **23**, 145-152.

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

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

Bladon, R. A. W., & Al-Bamerni, A. (1976). Coarticulation resistance in English /l/. *Journal of Phonetics*, **4**, 137-150.

Blokland, A., & Anderson, A. H. (1998). Effect of low frame-rate video on intelligibility of speech. *Speech Communication*, **26**, 97-103.

Braunstein, M. L., Hoffman, D. D., & Pollick, F. E. (1990). Discriminating rigid from nonrigid motion: Minimum points and views. *Perception & Psychophysics*, **47**, 205-214.

Bricker, P. D., & Pruzansky, S. (1966). Effects of stimulus content and duration on talker identification. *Journal of the Acoustical Society of America*, **40**, 1441-1449.

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. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory: Current research and issues. Vol. 1: Memory in everyday life* (pp. 169-174). New York: Wiley.

Burton, A. M., Bruce, V., & Dench, N. (1993). What's the difference between men and women? Evidence from facial measurement. *Perception*, **22**, 153-176.

Christie, F., & Bruce, V. (1998). The role of dynamic information in the recognition of unfamiliar faces. *Memory & Cognition*, **26**, 780-790.

Collins, S. A., & Missing, C. (2003). Vocal and visual attractiveness are related in women. *Animal Behaviour*, **65**, 997-1004.

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**, 839-849.

Hildreth, E. C., Grzywacz, N. M., Adelson, E. H., & Inada, V. K. (1990). The perceptual buildup of three-dimensional structure from motion. *Perception & Psychophysics*, **48**, 19-36.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, **14**, 201-211.

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.

Knappmeyer, B., Thornton, I. M., & Bülthoff, H. H. (2003). The use of facial motion and facial form during the processing of identity. *Vision Research*, **23**, 1927-1936.

Knight, B., & Johnston, A. (1997). The role of movement in face recognition. *Visual Cognition*, **4**, 265-273.

Lachs, L., & Pisoni, D. B. (2004a). Cross-modal source information and spoken word recognition. *Journal of Experimental Psychology: Human Perception & Performance*, **30**, 378-396.

Lachs, L., & Pisoni, D. B. (2004b). Crossmodal source identification in speech perception. *Ecological Psychology*, **16**, 159-187.

Lachs, L., & Pisoni, D. B. (2004c). Specification of cross-modal source information in isolated kinematic displays of speech. *Journal of the Acoustical Society of America*, **116**, 507-518.

Lander, K., Christie, F., & Bruce, V. (1999). The role of movement in the recognition of famous faces. *Memory & Cognition*, **27**, 974-985.

McGurk, H., & MacDonald, J. W. (1976). Hearing lips and seeing voices. *Nature*, **264**, 746-748.

Nygaard, L. C. (2005). Perceptual integration of linguistic and nonlinguistic properties of speech. In D. B. Pisoni & R. E. Remez (Eds.), *Handbook of speech perception* (pp. 390-414). Malden, MA: Blackwell.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, **60**, 355-376.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, **5**, 42-46.

Pavlova, M. A. (1995). Biological motion perception under various presentation rates. *Perception*, **24**(Suppl.), 112.

Pollick, F. E. (1997). The perception of motion and structure in structure-from-motion: Comparisons of affine and Euclidean formulations. *Vision Research*, **37**, 447-466.

Remez, R. E., Fellowes, J. M., & Rubin, P. E. (1997). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception & Performance*, **23**, 651-666.

Remez, R. E., Rubin, P. E., Nygaard, L. C., & Howell, W. A. (1987). Perceptual normalization of vowels produced by sinusoidal voices. *Journal of Experimental Psychology: Human Perception & Performance*, **13**, 40-61.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science*, **212**, 947-950.

Rosenblum, L. D. (2004). Perceiving articulatory events: Lessons for an ecological psychoacoustics. In J. G. Neuhoff (Ed.), *Ecological psychoacoustics* (pp. 219-248). Amsterdam: Elsevier.

Rosenblum, L. D. (2005). The primacy of multimodal speech perception. In D. Pisoni & R. Remez (Eds.), *Handbook of speech perception* (pp. 51-78). Malden, MA: Blackwell.

Rosenblum, L. D., Johnson, J. A., & Saldaña, H. M. (1996). Point-light facial displays enhance comprehension of speech in noise. *Journal of Speech & Hearing Research*, **39**, 1159-1170.

Rosenblum, L. D., & Saldaña, H. M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception & Performance*, **22**, 318-331.

Rosenblum, L. D., Smith, N., & Niehus, R. P. (2006). *Look who's talking: Recognizing friends from visible articulation*. Manuscript submitted for publication.

Rosenblum, L. D., Yakel, D. A., Baseer, N., Panchal, A., Nodarse, B. C., & Niehus, R. P. (2002). Visual speech information for face recognition. *Perception & Psychophysics*, **64**, 220-229.

Schweinberger, S. R., & Soukup, G. R. (1998). Asymmetric relationships among perceptions of facial identity, emotion, and facial speech. *Journal of Experimental Psychology: Human Perception & Performance*, **24**, 1748-1765.

Sheffert, S. M., & Olson, E. (2004). Audiovisual speech facilitates voice learning. *Perception & Psychophysics*, **66**, 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 & Performance*, **28**, 1447-1469.

Sommers, M. S., Nygaard, L. C., & Pisoni, D. B. (1994). Stimulus variability and spoken word recognition: Effects of variability in speaking rate and overall amplitude. *Journal of the Acoustical Society of America*, **96**, 1314-1324.

Thornton, I. M., Pinto, J., & Shiffrar, M. (1998). The visual perception of human locomotion. *Cognitive Neuropsychology*, **15**, 535-552.

Van Lancker, D., Kreiman, J., & Emmorey, K. (1985). Familiar voice recognition: Patterns and parameters. I: Recognition of backward voices. *Journal of Phonetics*, **13**, 19-38.

Vitkovitch, M., & Barber, P. (1994). Effect of video frame rate on subjects' ability to shadow one of two competing verbal passages. *Journal of Speech & Hearing Research*, **37**, 1204-1210.

Williams, D. R., Verbrugge, R. R., & Studdert-Kennedy, M. (1983). Judging sinewave stimuli as speech and nonspeech. *Journal of the Acoustical Society of America*, **74**, S66.

Yakel, D. A., Rosenblum, L. D., & Fortier, M. A. (2000). Effects of talker variability on speechreading. *Perception & Psychophysics*, **62**, 1405-1412.

NOTES

1. Whereas other studies in which cross-modal speaker matching has been examined have tested XAB matches of both (1) one audio token to two videos and (2) one video token to two audio tokens, we chose to use the former ordering only. This choice to use one ordering was made because of the large number of manipulations and conditions implemented in our experiment. The selection of this specific ordering was made on the basis of our previous research (Rosenblum et al., 2002).

2. We thank Sonya Sheffert and an anonymous reviewer for suggesting these types of analyses.

3. Although articulatory extent can be a phonetically relevant dimension at the segment level, its measurement in the present analyses was of average or peak amplitude over the course of a full sentence.

(Manuscript received August 27, 2003;
revision accepted for publication March 25, 2005.)