Published in final edited form as:

J Acoust Soc Am. 2004 July ; 116(1): 507–518.

Specification of cross-modal source information in isolated kinematic displays of speech^a

Lorin Lachs^b and

Department of Psychology, 5310 North Campus Drive, California State University, Fresno, California 93740

David B. Pisoni

Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, Indiana 47405

Abstract

Information about the acoustic properties of a talker's voice is available in optical displays of speech, and vice versa, as evidenced by perceivers' ability to match faces and voices based on vocal identity. The present investigation used point-light displays (PLDs) of visual speech and sinewave replicas of auditory speech in a cross-modal matching task to assess perceivers' ability to match faces and voices under conditions when only isolated kinematic information about vocal tract articulation was available. These stimuli were also used in a word recognition experiment under auditory-alone and audiovisual conditions. The results showed that isolated kinematic displays provide enough information to match the source of an utterance across sensory modalities. Furthermore, isolated kinematic displays can be integrated to yield better word recognition performance under audiovisual conditions than under auditory-alone conditions. The results are discussed in terms of their implications for describing the nature of speech information and current theories of speech perception and spoken word recognition.

I. INTRODUCTION

Optical information about vocal articulation has been shown to have substantial effects on speech perception. In the absence of auditory stimulation, visual information is sufficient to support accurate spoken word recognition (Bernstein *et al.*, 2000). Combined with acoustic information, visual stimulation can also enhance speech intelligibility in noise by +15 dB (Sumby and Pollack, 1954; Summerfield, 1987). Conflicting information in the acoustic and optic displays of an audiovisual speech stimulus can also interact to form illusory percepts in speech perception [the "McGurk" effect (McGurk and MacDonald, 1976)].

These core phenomena in audiovisual speech perception are robust and reliable (Summerfield, 1987), and provide strong evidence that some components of the speech perception process must also be involved in combining different sources of information across sensory modalities. A great deal of research, especially over the past two decades, has focused on the nature of the perceptual integration process—how sensory stimulation from disparate and seemingly incommensurate modalities can influence the perception of a multimodal stimulus. Theories of this process vary considerably. For example, Braid's (1991) prelabeling model represents audiovisual stimuli in a multidimensional space

^aThis work was previously submitted by Lorin Lachs to Indiana University in partial fulfillment of the requirements for a Ph.D.

© 2004 Acoustical Society of America

^bllachs@csufresno.edu.

consisting of acoustic- and optic-specific dimensions. Alternatively, Massaro's (1998) fuzzy logical model of perception represents multimodal perception as the outcome of a decision process that evaluates the degree of support from each modality for unimodal prototypes.

Another class of theory for explaining the process of multimodal integration proposes that auditory and visual inputs are somehow evaluated in terms of a common representational format. According to Schwartz *et al.* (1998), there are two alternatives for common format models. In one, information from one modality is translated or recorded into a format that is compatible with another, more familiar or "dominant" modality. Schwartz *et al.* assert that models based on this type of a representation necessarily predict that visual influences on speech perception will only be observed when optical information conflicts with the auditory stimulus, or when the auditory stimulus is degraded. This prediction was disconfirmed by Sumby and Pollack (1954), who found that auditory-alone performance improved as the signal-to-noise ratio increased. After scores in the audiovisual condition were normalized relative to this increasing baseline, Sumby and Pollack discovered that the amount by which performance improved due to audiovisual stimulation remained constant over the entire range of signal-to-noise ratios tested. This important finding indicates that the effect of the additional visual information is not conditional on the degree of ambiguity in the auditory signal. Remarkably, this finding was reconfirmed by Reisberg *et al.* (1987), who showed that concurrently presented visual information facilitated the repetition of foreign-accented speech and semantically complex sentences presented with no background noise at all. These findings demonstrate that visual information about speech is useful and informative, and is not simply compensatory in situations where auditory information is insufficient to support perception.

The currently available evidence therefore indicates that translation of information into a dominant modality format is not a viable theoretical construct. Instead, Schwartz *et al.*'s alternative "common format" model assumes that acoustic and optic information about speech is analyzed with reference to some kind of multimodal, amodal, or modality-neutral representational space. In one theoretical approach of this type, the modality-neutral perceptual space relates to the vocal tract articulation that underlies the production of auditory or visual speech signals (Fowler, 1986; Liberman and Mattingly, 1985; Summerfield, 1987).

Convincing evidence for a modality-neutral form of speech information comes from a study of the McGurk effect using auditory and *tactile* information about speech. Fowler and Dekle (1991) had naive participants listen to spoken syllables synthesized along a /ba/-/pa/ continuum while using their hands to obtain information about the articulation, in much the same way that deaf-blind users of the Tadoma method do (Schultz *et al.*, 1984). The tactile information on every trial was either a /ba/ or a /ga/, articulated by a talker who was unable to hear the auditory syllable. Observers were then asked to categorize the heard stimulus as either a /ba/ or a /ga/ in a forced choice task. Fowler and Dekle found that the auditory perceptual boundary for /ba/ and /ga/ was shifted by the phonetic content of the tactile pattern. Interestingly, this effect was found with subjects who had no training in Tadoma at all, indicating that stored associations between tactile and auditory speech gestures could not be involved in the integration results.

In a second experiment, observers were simultaneously presented with the acoustic syllables and *orthographic* displays of "BA" or "PA." Because all of the observers were literate college undergraduates, Fowler and Dekle assumed that they had stored robust associations between the orthographic symbols and their phonetic counterparts in memory. However, the visual orthographic displays (with visual masking designed to bring performance down to a level equitable with performance in the tactile condition) did not affect the position of the

category boundary along the continuum at all. Fowler and Dekle interpreted these findings as evidence that the ability to “integrate” information about speech is *not* based on matching features to learned representations, but is rather based on the detection of amodal information about speech articulation. Their results also demonstrate that some degree of useful information about speech can be obtained through sensory modalities other than audition and vision.

Further support for the modality-neutral form of speech information comes from a recent series of studies using a novel cross-modal matching task (Lachs and Pisoni, in press). In this cross-modal matching task, participants view a visual-alone, dynamic display of a talker speaking an isolated English word and are then asked to match the pattern with one of two auditory-alone displays. One of the alternative auditory displays is the acoustic specification of the *same articulatory events* that produced the visual-alone pattern. The other alternative is the acoustic specification of a different talker saying the same word that was spoken in the visual-alone pattern. When matching a visual-alone token to one of two auditory-alone tokens, the “order” is said to be “V-A.” In contrast, when participants hear an auditory-alone display first, and then are asked to match to one of two visual-alone alternatives, this is referred to as the “A-V” order.

Lachs and Pisoni (in press) found that participants performed above chance ($M=0.60$ for A-V; $M=0.65$ for V-A) when asked to match the same phonetic events across different sensory modalities, regardless of the order in which the modalities were presented. That is, observers could match a speaking face with one of two voices, or match a voice with one of two speaking faces. Lachs and Pisoni argued that the results provided evidence for the existence of “cross-modal source information,” that is, modality-neutral, articulatory information about the source or speaker of an utterance that is specified by sensory patterns of both acoustic or optic energy. In a series of further experiments, Lachs and Pisoni extended these findings using a series of manipulations to the acoustic and optical displays that were presented for cross-modal matching. One experiment found that participants were not able to match across sensory modalities when optical and acoustic patterns were played backwards in time, indicating that talker-specific information is sensitive to the normal temporal order of spoken events. Another experiment showed that static visual displays of faces could not be matched to dynamic acoustic displays. This result suggested that talker-specific information is specified in the dynamic, time-varying structure of visual displays, and not in their static features. Finally, an additional experiment showed that noiseband stimuli (Smith *et al.*, 2002) preserved talker-specific information, despite the elimination of a traditional acoustic cue to vocal identity: fundamental frequency (f_0). This result was interpreted as evidence that talker-specific information in speech is specified in the pattern of formant resonances as they evolve over time, irrespective of the f_0 used to excite them.

In another set of experiments, Lachs and Pisoni (2004) subjected the acoustic speech signal to various acoustic transformations, which were then used as the auditory stimulus in a cross-modal matching task (using fully illuminated faces) and a word identification task. Several of these transformations affected the frequency domain of the recorded words. Lachs and Pisoni found that, in general, transformations that preserved the linear relationship between formant frequencies of a speech utterance also preserved the ability to perform the matching and word identification tasks. However, there were transformations that destroyed this relationship. For example, frequency rotation around a central frequency value (Blessner, 1972) also destroyed participants’ ability to perform both tasks, as did a nonlinear scaling transformation of the frequency components of the utterance. Because the formant frequencies associated with speech are, to a rough approximation, markers of the instantaneous configuration of the vocal tract as it evolves over time, Lachs and Pisoni claimed that only acoustic patterns that carry information about the kinematics of

articulation are informative for the completion of both word identification and cross-modal matching.

The results of this series of experiments suggest that cross-modal matching can be carried out because of a common articulatory basis for acoustic and optic displays of speech. Lachs and Pisoni (in press) proposed that, because both visual and auditory displays of speech are lawfully structured by the same underlying articulatory events (Vatikiotis-Bateson *et al.*, 1997), matching is accomplished by comparison to a common, underlying source of information about vocal tract activity.

The proposal that audiovisual integration reflects a common underlying articulatory basis for audiovisual speech integration is also supported by a correlational analysis of the stimuli used in Grant's (2001) recent detection experiments. Grant showed that the rms amplitude of the acoustic signal (especially in the bandpass filtered F2 region) was strongly related to the area of the opening circumscribed by the lips. He proposed that this correlation between an acoustic variable and an optical variable may have been responsible for his finding that auditory speech detection thresholds were reduced under concurrent visual stimulation. It is interesting to note that the relationship reported by Grant between the auditory and visual variables is also related to a common, underlying kinematic variable based on jaw and mouth opening. Further work on the correspondence between auditory and visual speech signals has also been carried out by Vatikiotis-Bateson and colleagues (e.g., Vatikiotis-Bateson *et al.*, 2002; Yehia *et al.*, 1998).

An articulatory foundation for modality-neutral information is also consistent with the observation that the visual correlates of spoken language arise as a direct result of the movements of the vocal tract that are necessary for producing linguistically significant speech sounds (Munhall and Vatikiotis-Bateson, 1998). An articulating vocal tract structures acoustic and optic energy in space and time in highly constrained ways. As a consequence, the auditory and visual sensory information present in those displays is lawfully structured by a common unitary articulatory event. Thus, the seemingly disparate patterns of acoustic and optic energy are necessarily and lawfully related to each other by virtue of their common origin in articulation.

One avenue of investigation, then, is to discover the ways in which optical and acoustical structure specify the informative properties of vocal tract articulation. A growing body of research suggests that both the acoustic or optical properties of speech carry *kinematic* or *dynamic* information about vocal tract articulation (Rosenblum, 1994), and that these sources of information drive the perception of linguistically significant utterances (Fowler, 1986; Summerfield, 1987). Kinematic variables refer to position and its time derivatives: velocity, acceleration, etc.; dynamic variables refer to forces and masses (Bingham, 1995). These variables are necessarily grounded in spoken events: they do not refer to specific properties of visual or auditory stimulation, but rather to the physical events that structure light or sound. As such, these variables are prime theoretical candidates for the modality-neutral, articulatory, properties of events.

The point-light display (PLD) technique developed by Johansson (1973) has been used extensively to investigate the perception of kinematic variables (e.g., Kozlowski and Cutting, 1978; Runeson and Frykholm, 1981). By placing small reflective patches at key positions on a talker's face and darkening everything else in the display, researchers have been able to theoretically isolate the kinematic properties of visual displays of talkers articulating speech (Rosenblum *et al.*, 1996; Rosenblum and Saldaña, 1996). These "kinematic primitives" have been shown to behave much like unmodified, fully illuminated visual displays of speech (Rosenblum and Saldaña, 1996), albeit with smaller effect sizes. Indeed, Rosenblum and

Saldaña found that the McGurk illusion can be induced by dubbing visual point-light displays onto phonetically discrepant auditory syllables (Rosenblum and Saldaña, 1996). In addition, Rosenblum *et al.* (1996) also demonstrated that providing point-light information about articulation in conjunction with auditory speech embedded in noise can result in increased speech intelligibility scores, just as fully illuminated visual displays can improve speech intelligibility (Sumbly and Pollack, 1954).

As with optic displays, kinematic information can also be theoretically isolated and examined in acoustic displays of speech. “Sinewave speech replicas” (Remez *et al.*, 1981) are time-varying acoustic signals that contain three sinusoidal tones generated at the frequencies traced by the centers of the three lowest formants in a speech signal. Because the formants in speech are resonances of the vocal tract, that is, bands of energy at frequencies specified by the evolving configuration of the vocal tract transfer function, these “sinewave speech replicas” can be said to isolate kinematic information in the auditory domain. Pioneering experiments by Remez and his colleagues using sinewave speech replicas have demonstrated that listeners can perceive linguistically significant information from these minimal kinematic displays in many of the same ways that they can from untransformed acoustic displays (Remez *et al.*, 1981).

Two studies using unimodal kinematic primitives have indicated that indexical information in speech signals may be carried in the very same “kinematic details” that support phonetic identification (Runeson, 1994). In the visual domain, Rosenblum *et al.* (2002) presented participants with point-light displays of a talker speaking a sentence and then asked them to match the point-light display with one of two fully illuminated faces speaking the same sentence. One of the fully illuminated faces belonged to the same talker who had generated the point-light display. Their results showed that visual point-light displays of a talker articulating speech could be matched accurately to fully illuminated visual displays of the same utterances. This finding suggests that some talker-specific details are contained in the visually isolated kinematics of vocal tract activity. Analogously, in the auditory domain, Remez *et al.* (1997) presented listeners with sinewave replicas of English sentences and asked them to match the token with one of two untransformed auditory tokens. As above, one of the untransformed auditory tokens was spoken by the same talker who had spoken in the sinewave replica. The results showed that sinewave speech replicas carry enough idiosyncratic phonetic variation to support correct matching between untransformed, natural utterances and their sinewave speech replicas.

Several parallel lines of investigation have also investigated spoken word recognition and the *integration* of these minimal, unimodal stimulus displays with untransformed auditory stimuli or fully illuminated visual stimuli. For example, in one recent study, Remez *et al.* (1999) demonstrated that the intelligibility of sinewave replicas of sentences was significantly enhanced when presented in conjunction with a full visual display of the talker. In another study Rosenblum *et al.* (1996) showed that point-light displays of a talker speaking sentences could enhance recognition of untransformed auditory displays embedded in noise.

In summary, kinematic information about speech articulation is carried by both acoustic and optic energy, and these sources of information appear to be sufficient to support both speech perception and talker identification in each modality alone. In the present investigation, we extend the seminal work of Rosenblum and Remez in a study investigating talker-identification and speech perception using fully isolated displays of speech, that is, stimuli consisting of both point-light displays and sinewave speech replicas.

For the set of experiments reported here, point-light displays (PLDs) of four talkers speaking isolated English words were recorded, and the accompanying acoustic displays were converted into sinewave speech replicas. In the first experiment, we examined whether theoretically isolated kinematic displays of speech can carry indexical information about a talker. Participants were asked to match point-light displays and sinewave speech replicas of the same talker using a cross-modal matching task. In the second experiment, we investigated whether these minimal, “skeletonized” versions of speech could also be integrated across sensory modalities in an open set word recognition task and whether they would display the audiovisual enhancement effects: combined audiovisual stimulation leads to higher speech intelligibility scores than auditory-alone stimulation.

II. EXPERIMENT 1: CROSS-MODAL MATCHING OF KINEMATIC PRIMITIVES

Experiment 1 was designed to test the hypothesis that talker-specific information is specified in the kinematic behavior of the vocal tract. To accomplish this goal, we used a cross-modal matching task and asked participants to match point-light displays with sinewave speech replicas. If the object of perception is modality-neutral and based on the articulation of the vocal tract, we would expect that observers should be able to match faces and voices with only the minimal, theoretically isolated kinematic information about articulation available in point-light displays and sinewave replicas of speech.

A. Method

1. Participants—Participants were 40 undergraduate students enrolled in an introductory psychology course who received partial credit for participation. All of the participants were native speakers of English. None of the participants reported any hearing or speech disorders at the time of testing. In addition, all participants reported having normal or corrected-to-normal vision. None of the participants in this experiment had any previous experience with the audiovisual speech stimuli used in this experiment.

2. Stimulus materials

a. Point-light displays of speech: To capture and record the movement of the articulators, it was necessary to videotape a set of four talkers. Four females between the ages of 23 and 30 volunteered as talkers for the recording session. To create a point-light display, the motion of the points selected must be recorded directly; that is, the motion of the articulators cannot simply be extracted in postproduction from a recording of a fully illuminated display of the talker’s face (Runeson, 1994). Before recording, therefore, each talker glued glow-in-the-dark dots, each approximately 3 mm in diameter, to her face in the pattern outlined in Fig. 1. For dots on the outside of the mouth, the adhesive used was a spirit gum commonly used for adhesion of latex masks (Living Nightmare® Spirit Gum). Dots on the lips, teeth, and tongue were affixed with an over-the-counter dental adhesive (Fixodent® Denture Adhesive Cream).

Stimuli were recorded using a digital video camera and microphone (Sony AKGC414) in a sound-attenuated room onto Sony DVCam PDV-124ME digital media tapes. Each talker was seated approximately 56 in. from the camera lens, and the zoom control on the camera was adjusted such that the visual angle subtended by the distance between the talker’s ears was equal across talkers. The camera lens and face of the talker were placed at a height of approximately 44 in. Two black lights (15.7 cm long, 15 W each) were secured on permanent fixtures placed 13.5 in. on either side of the camera lens, on an axis perpendicular to the line between the talker’s face and the camera lens. The two black lights were the only source of illumination used during videotaping, and did not significantly illuminate the skin of the talker being recorded. However, the glow in the dark dots reflected the black light in

the normal visible spectrum. The video track of the recorded movies thus recorded only the movement of the glow-in-the-dark points in isolation of the face to which they were affixed.

During videotaping, talkers read a list of 96 English words off a teleprompter configured with a green font and black background so that it provided no additional ambient illumination in the recording session. There was a 3-s delay between the presentation of successive words on the teleprompter. The digitized stimuli were segmented later such that each word was preceded and followed by ten silent frames ($10 \times 29.97 \text{ fps} = 333.66 \text{ ms}$) with no speech sounds.

b. Sinewave speech replicas: The audio track of each utterance was extracted from the videotape and stored in a digital audio file for transformation into sinewave replicas of speech (Remez *et al.*, 1981). Sinewave speech transformations were created by tracking the frequencies and amplitudes of the first four formants as they varied over time.

These acoustic measurements were obtained in a two-step process. First, each sound file (originally recorded at a sampling rate of 22.050 kHz) was resampled to 10 kHz. The resampled sound file was then broken into windows of 10 ms each. Each window was the subject of an eight-order LPC analysis, and the four coefficients with the highest magnitudes were then converted to frequencies and magnitudes and stored in a data file. Each sound file thus had an associated data file with eight parameters (four frequencies and four magnitudes) per 10-ms window, corresponding to the change in the formants in the original sound file over time.

Sometimes the LPC analysis would output erroneous or spurious noises in the original sound file. To deal with this problem, the automated output file was scrutinized by hand to determine the accuracy of the automated process. Overall, the automated process proved to be an excellent starting point, although some slight adjustments in frequency and amplitude were necessary in several windows for most of the stimuli. The automated process had some difficulty in tracking formants during unvoiced portions of the speech waveform corresponding to periods of aspiration, consonant transitions, and frication. Adjustments in frequency and amplitude were made by hand using home-grown software constructed within MATLAB for this purpose.

The output of the two-stage process was a new data file specifying eight parameters (four frequencies and four associated amplitudes) for each 10-ms window in the original sound file. These data were then submitted to a synthesis routine that produced four sinusoidal tones that varied over time according to the parameter's output by the measurement process.

Figure 2 shows examples of the spectrograms of an original sound file from the stimulus set and its corresponding sinewave speech replica. As shown in the figure, the sinewave replica eliminated all extraneous information from the sound file other than the variation of the formants over time.

c. Familiarization phase: There was also a short familiarization phase so that participants could become accustomed to the unnaturalness of the sinewave speech replicas. The stimulus materials used during the familiarization phase were 20 isolated, monosyllabic English words that were not used in the final test list. These 20 words were spoken by a different talker who was not used to create the stimuli used in the point-light display database. The 20 familiarization words were converted into sinewave speech replicas using the same methods outlined earlier.

3. Procedure—During the familiarization phase, participants heard a sinewave speech replica of a spoken word and then the original, untransformed utterance from which that replica was constructed. This sinewave-original pair was presented three times in succession. After the third presentation of the stimulus pair, participants were asked to rate on a three-point scale how closely the sinewave replica matched the natural utterance. No feedback was provided. The familiarization task gave participants some exposure to the unusual nature of sinewave speech replicas without explicitly instructing them on how to identify words from them.

Figure 3 shows a schematic description of the cross-modal matching task. Participants in the “V-A” condition were first presented with a visual-alone point-light display video clip of a talker uttering an isolated, English word. Shortly after seeing this video display (500 ms), they were presented with two auditory-alone, *sinewave replicas* of speech. One of the clips was the same talker they had seen in the video (taken from the same utterance), while the other clip was a different talker. Participants were instructed to choose the audio clip that matched the talker they had seen (“first” or “second”). Similar instructions were provided for participants in the “A-V” condition. These participants heard the audio clip first, and had to make their decision based on two video displays.

On each trial of the experiment, the test stimulus was either the video or audio portion of one movie token. Each movie token displayed an isolated word spoken by a single talker. The order in which the target and distracter choices were presented was randomly determined on each trial. For each participant, talkers were randomly paired with each other, such that each talker was contrasted against one and only one other talker for all trials in the experiment. For example, “Mary” was always contrasted with “Jane,” regardless of whether “Mary” or “Jane” was the target alternate on the trial. Responses were entered by pressing one of two buttons on a response box and transferred to a log file for further analysis.

A short training period (eight trials) preceded each participant’s test session. During the training period, the participant was presented with a cross-modal matching trial and asked to pick the correct alternate. During training only, the response was followed immediately by feedback. The feedback consisted of playing back the entire audiovisual movie clip of the test word. The talkers and words used in the familiarization phase were different from those used in the testing period.

Half of the participants matched point-light displays with the original, untransformed sound files; the other half of the participants matched point-light displays with sinewave replicas of speech.

B. Results

Figure 4 shows performance on the cross-modal matching task when point-light (PL) displays were matched to untransformed auditory-only (AO) speech tokens (PL–AO) or sinewave speech (SWS) tokens (PL–SWS). The figure shows that cross-modal matching performance with point-light displays was quite poor, although statistically different from chance (0.5). Table I provides a summary of the descriptive statistics for each of the conditions pictured in Fig. 4. As shown in the table, performance in both matching conditions and both orders differed significantly from chance, indicating that cross-modal matching judgments can be carried out using only isolated kinematic displays in both the visual and auditory sensory domains.

The data were submitted to a 2×2 repeated measures ANOVA (order, matching condition). The ANOVA did not reveal any significant main effects or interactions.

C. Discussion

Point-light displays and sinewave speech replicas of isolated spoken words can be matched at above chance levels of performance across sensory modalities in a cross-modal matching task. This finding suggests that theoretically isolated kinematic displays of speech (visual and auditory) contain some information for cross-modal specification. The results are consistent with the proposal that cross-modal matching performance is based on the ability to detect vocal tract movement information from unimodal displays. Furthermore, the results are consistent with the hypothesis that cross-modal comparisons are made using a common, modality-neutral metric based on the kinematics of articulatory activity. Because these extremely simplified stimulus displays were designed to eliminate all traditional cues to visual identity (e.g., configuration of facial features, shading, shape, etc.) and to auditory vocal identity (e.g., f_0 , average long-term spectrum, etc.), but preserve only the isolated movement of the vocal articulators, the results suggest that the common talker-specific information is sensory information that is related to articulatory kinematics of vocal tract activity.

It is interesting to note here that, compared to the performance on a similar cross-modal matching task (Lachs and Pisoni, in press) using fully illuminated faces and untransformed speech, performance was low. In those experiments, Lachs and Pisoni found that performance was 0.60 in the A-V condition and 0.65 in the V-A condition. There are two alternative explanations for this decrement in performance. First, it is possible that perceivers use more than just visible kinematics to successfully match identity across modality. However, because static images and dynamic video displays played backward do not allow for cross-modal matching above chance at all, this possibility seems unlikely. A second explanation is that the point-light displays used in experiment 1 failed to capture all of the relevant kinematic activity of the articulators. This seems plausible, as previous work has shown that the number and placement of points on a face during recording of point-light images affects performance on speech reception thresholds in noise (Rosenblum *et al.* 1996).

III. EXPERIMENT 2: WORD IDENTIFICATION WITH KINEMATIC PRIMITIVES

Taken together with the findings of experiment 1, the available evidence about the perception of indexical information from isolated kinematic primitives supports the hypothesis that talker-specific information is carried in parallel with phonetic information in the fine-grained kinematic details of a talker's articulator motions over time. These kinematic details are inherently modality-neutral because they refer to the common articulatory event, not the surface patterns of acoustic or optic energy impinging on the eyes or ears. As discussed above, an articulatory, modality-neutral form for speech information has also been proposed to account for audiovisual speech phenomena (Rosenblum, 1994), like the McGurk effect and audiovisual enhancement to speech intelligibility. Is the kinematic information available in sinewave speech and point-light displays also sufficient to support accurate word recognition?

Experiment 2 was designed to show that multimodal kinematic primitives are also integrated in a classic audiovisual integration task. Participants were asked to identify spoken words under two presentation conditions. In the auditory-alone condition, they were presented with sinewave replicas of isolated English words. In the audiovisual condition, they were presented with *multimodal* kinematic primitives: point-light displays of speech paired with sinewave replicas of the same utterance. Based on the earlier findings reported by Rosenblum *et al.* (1996) and Remez *et al.* (1999), we expected that word identification performance under combined, multimodal stimulation would show enhancement and would be better than performance under unimodal, auditory-alone stimulation. This finding would provide evidence in support of the proposal that auditory and visual displays of speech are

“integrated” because they both carry information about the underlying kinematics of articulation.

A. Method

1. Experimental design—Experiment 2 used a word recognition task to measure speech intelligibility for isolated words. One within-subjects factor was manipulated. “Presentation mode” consisted of two levels: sinewave speech alone (SWS) and point-light display plus sinewave speech (PL+SWS). The levels of this factor were presented in blocks, which were counterbalanced for order of presentation across participants. In addition, an equal number of participants were presented with each word in each presentation mode.

2. Participants—Participants were 32 undergraduate students enrolled in an introductory psychology course who received partial credit for participation. All of the participants were native speakers of English and none of them reported any hearing or speech disorders at the time of testing. In addition, all participants reported having normal or corrected-to-normal vision. None of the participants in this experiment had any previous experience with the audiovisual speech stimuli used in this experiment.

3. Stimulus materials—The stimulus materials used during the familiarization phase were 20 isolated, monosyllabic English words that were not used in the final test list. These 20 words were spoken by a different talker who was not used to create the stimuli used in the point-light display database. The 20 familiarization words were converted into sinewave speech replicas using the same methods outlined in experiment 1.

The stimulus materials used during testing were identical to those used in experiment 1. However, only the sinewave replicas of the spoken words were used as auditory tokens in this experiment.

4. Procedures—Each session began with a short familiarization phase identical to the one outlined earlier in experiment 1 so that participants could become accustomed to the unusual nature of the sinewave speech replicas. It should be emphasized here that the familiarization task only asked for judgments of “goodness” and provided no feedback at all.

During the testing phase, each participant heard all 96 words spoken by one of the four point-light talkers. Every participant heard each talker speak on an equal number of trials, and no participant ever heard a word twice over the course of the experiment. The talker who spoke a given word was counterbalanced across participants. For each participant, 48 words were presented under audio-only (AO:SWS) conditions and the other 48 words were presented under audiovisual (AV:PL+SWS) conditions. A separate analysis, using another group of participants, revealed that the visual-alone intelligibility of the 96 words was 1.6% (Bergeson *et al.*, 2003). The assignment of a presentation context to each list and the order in which presentation contexts were viewed was counterbalanced across participants.

After each stimulus item was presented, the participant simply typed in the word he/she heard on a standard keyboard and clicked the mouse button to advance to the next trial.

B. Results

Each participant’s responses were hand-screened for typing and spelling errors by two reviewers who worked independently. A typing error was defined as a substituted letter within one key on a standard keyboard of the target key or an inserted letter within one key of an adjacent letter in the response. Spelling errors were only accepted if the letter string did not form a word in its own right. Using this conservative method of assessment, the two

reviewers had a 100% agreement rate on classifying responses as typing and spelling errors. Responses on this task were scored correct if, and only if, they were homophonous with the target word in a standard American English dialect (e.g., “bare” for “bear,” but not “pin” for “pen”).

A paired-samples *t*-test comparing scores on the SWS and PL+SWS conditions revealed a significant difference between auditory-only and audiovisual performance, $t(31)=7.69$, $p<0.05$. Figure 5 shows that performance with the audiovisual (PL+SWS) stimuli was higher than performance with the auditory-only (SWS) stimuli. The mean difference between the SWS and PL+SWS conditions was 12.4% (SE = 1.6%).

a. Audiovisual gain (R scores)—In order to examine individual differences in the extent to which combined audiovisual stimuli enhanced word intelligibility, the scores in the SWS and PL+SWS conditions were combined into a single metric to obtain the measure *R*, the relative gain in speech perception due to the addition of visual information (Sumbly and Pollack, 1954). *R* was computed using the following formula:

$$R_a = \frac{AV - A}{100 - A}$$

where AV and A represent the speech intelligibility scores obtained in the audiovisual and auditory-alone conditions, respectively. From this formula, one can see that *R* measures the gain in accuracy in the AV condition relative to the accuracy in the A condition, normalized relative to the amount by which speech intelligibility could have possibly improved above the auditory-alone scores. The *R* score can be used as an effective measure for comparing audiovisual gain across participants, because it effectively normalizes for the baseline, auditory-alone performance.

Figure 6 shows the *R* scores for all 32 participants in experiment 2 ($M=0.154$, $SE=0.020$). A few participants ($N=4$) showed no gain at all over auditory-alone performance. However, most of the participants showed some advantage in identifying words in the audiovisual condition relative to the auditory-alone condition. One participant showed a gain of 41% due to multimodal presentation. The *R* scores were significantly different from zero [$t(31)=7.67$, $p<0.05$] indicating that the point-light displays facilitated the intelligibility of all words, despite the fact that they were virtually unintelligible under visual-alone conditions (Bergeson *et al.*, 2003).

C. Discussion

The present set of results using point-light displays and sinewave replicas of speech demonstrate that some information for the combined perception of visual and auditory speech exists in highly impoverished displays that isolate the kinematics of articulatory activity that underlie the production of speech. Even with limited, minimal information in the two sensory modalities, participants were able to perceive point-light displays and sinewave speech as integrated patterns, and they were able to exploit the two sources of sensory information to aid them recognizing isolated spoken words. On average, additional point-light information increased intelligibility by 12%, almost doubling the intelligibility of the sinewave speech tokens observed in isolation. Furthermore, this performance advantage was well above the reported intelligibility of the PL stimuli in visual-alone conditions, indicating that the two sources of information were indeed integrated, and not simply linearly combined (Grant and Seitz, 1998; Massaro, 1998).

1. Individual differences in perceptual integration—In a recent study, Grant *et al.* (1998) compared observed audiovisual gain scores to the gains predicted by several models of sensory integration. These models predicted optimal AV performance based on observed recognition scores for consonant identification in auditory-alone and visual-alone conditions. Grant *et al.* (1998) found that the models either over- or underpredicted the observed audiovisual data, indicating that not all individuals integrate auditory and visual information optimally. The present results also revealed individual differences in the extent to which sources of auditory and visual information could be combined and integrated together to support spoken word recognition. Audiovisual gain scores in the present experiment ranged from 0% to 41%.

Without question, the informative value of optic and acoustic arrays for specifying linguistic information must be interpreted in terms of a perceiver's sensitivity to relevant phonetic contrasts in the language. It is well known that speechreading performance in hearing-impaired and normal-hearing listeners is subject to wide variation (Bernstein *et al.*, 2000). It is not unreasonable to assume that the speechreading of point-light displays will be subject to at least the same degree of variation, and that such variation might have extensive effects on the ability to integrate auditory and visual displays of speech. Investigations into the role of individual differences in multisensory perception may provide important insights into the perceptual integration of audiovisual speech information and its development (Lewkowicz, 2001).

Despite the individual variation, however, the results of the present study are consistent with the hypothesis that auditory and visual speech information are combined and integrated with reference to the underlying kinematic activity of the vocal tract. With dynamic movement information available from either sensory modality, perceivers appear to have evaluated the two kinds of information together, and exhibited the audiovisual gain finding that has been replicated many times in the literature (see Grant and Seitz, 1998). The present results extend and complement the recent findings of Rosenblum *et al.* (1996) and Remez *et al.* (1999) who found that theoretically isolated kinematic displays could facilitate word identification in conjunction with untransformed displays in the other sensory modality. The present results also provide additional support for the proposal that the evaluation of optical and acoustic information conceptualized along a common, kinematic and articulatory metric underlies classic audiovisual integration phenomena such as audiovisual enhancement (Summerfield, 1987).

IV. GENERAL DISCUSSION

Taken together, the results of these experiments using multimodal kinematic primitives reveal that talker-specific and phonetic information may be specified in the kinematics of a talker's vocal tract activity. A common form of phonetic representation, focusing on the articulatory origin of spoken language, may therefore be needed to account for the robust findings in the audiovisual speech perception literature (Schwartz *et al.*, 1998; Summerfield, 1987), as well as investigations into the indexical properties of spoken language (Remez *et al.*, 1997; Rosenblum *et al.*, 2002).

Our conclusion is drawn based on the unique nature of the stimuli used in the present experiments. We claim that our results do not simply imply that perceivers are sensitive to two impoverished signals in disparate sources of sensory information. The important point here is that the signals in our experiments are impoverished in a principled way that theoretically *isolates* important kinematic information about the underlying articulation. The sinewave speech transformation is a systematic way of removing all information from an acoustic speech signal, while at the same time preserving information about the resonances

of the front and back cavities of the vocal tract (Remez *et al.*, 1981). Similarly, point-light displays of speech do not simply make visual displays of speech harder to recognize. Rather, they specifically attempt to isolate the motion of points (or near-points) on the talker's face while eliminating all other sources of visual information (Rosenblum and Saldaña, 1996). The results presented here dovetail with the experiments presented in Lachs and Pisoni (2004), in which it was found that only acoustic transformations that preserve the linear relationship between formant frequencies preserved the ability to match faces and voices cross-modally. In that series of experiments, as in the present studies, the data were consistent with the assertion that preservation of articulatory information about speech is important for specifying the underlying link between the optical and acoustic correlates of speech.

One potential problem with the results from experiment 1 is that performance levels, although statistically above chance, were not very high. However, we point out that our main goal in the present experiment was to demonstrate that it was, in practice, *possible* for participants to perform this extremely difficult cross-modal matching task with minimal feedback, training, or other prior experience. As discussed above, the difficulty of the general cross-modal matching task was demonstrated in the fully illuminated, full bandwidth version of the cross-modal matching task (Lachs and Pisoni, in press), in which performance was only between 60% and 65% accurate (depending on the condition). Whether manipulations such as training, feedback, or other types of experience can raise performance (as we suspect they will) remains an empirical question we have chosen to leave to another series of experiments. Further evidence for the reliability of the basic cross-modal matching results reported here using kinematic primitives has also been established by two independent research groups working with different stimulus materials. Kamachi *et al.* (2003) investigated cross-modal matching ability with sentence-length stimuli. Their results showed that fully illuminated faces could be matched to sinewave speech with an average accuracy of 61%. In another recent study, Rosenblum *et al.* (under review) also used sentence-length stimuli and showed that point-light faces could be matched to untransformed speech at above chance levels of performance. Rosenblum *et al.* also compared performance in this condition to conditions in which the visual point-light information was either jumbled or staggered. In the jumbled condition, video frames from the original recording were presented out of their originally recorded temporal sequence. In the staggered condition, frames were presented in the correct order, but were presented such that the duration of the frame's presentation was varied. In both of these control conditions, Rosenblum *et al.* found that performance in the cross-modal matching task was inferior to performance with the unaltered point-light stimuli. The authors concluded from this pattern of results that the evolving dynamics of visible articulatory activity are implicated in the ability to perform this task.

One alternative explanation of the present findings could appeal to the use of an abstract (rather than articulatory) modality-neutral representational format. For example, in experiment 2, stimulation from the various sensory modalities could be converted, in isolation, to some kind of amodal, symbolic, phonetic description, and then these two phonetic descriptions could be evaluated and integrated, without recourse to articulatory behavior. However, such a representational format would not account for the results of experiment 1, where no phonetic information could possibly have helped match talkers' faces and voices (because the utterance used for both the target and distracter was identical). To account for these findings, it is necessary to propose a secondary set of abstract representations for talker identity to account for the cross-modal matching findings of experiment 1. We propose that our articulatory account is simpler and more parsimonious than the alternative, abstractionist account.

It is also possible that the results of experiment 1 could be explained by perceptual sensitivity to some kind of cross-correlation between information presented in the auditory and visual domains, without reference to articulation or any sort of articulatory metric. This might be a plausible explanation for the results of experiment 1, given the growing evidence that correlations between surface features of visual and auditory signals can explain reductions in detection thresholds for auditory speech in noise (Grant and Seitz, 2000) and speech band-pass filtered to isolate F1 or F2 (Grant, 2001). Indeed, the results of experiment 1 emphasize the need for models of audiovisual integration to recognize and incorporate sensitivity to cross-modal information. However, it is not clear how an account of cross-modal information based on cross-correlations of surface features in the acoustic and optic signals could account for the findings of experiment 2, in which the combination of isolated kinematics in both sensory modalities was used to enhance word recognition performance. While it seems reasonable to suggest that perceptual sensitivity to correlations between surface features would aid in lowering detection thresholds and in explicitly identifying cross-modal relationships, it is unclear how such sensitivity to multimodal correlations might assist in recognizing the meaningful linguistic content of an utterance.

The results presented here are consistent with direct-realist theories of speech perception (Fowler, 1986; Gibson, 1979). These “ecological” theories of perception are based on the findings that patterns in light and sound are structured and lawfully related to the physical events that caused them. For example, the ambient light in a room is reflected off multiple surfaces, like the floor, walls, tables, chairs, etc., before it stimulates the retina. The pattern of light observable from the point-of-view of an active observer, then, is highly *structured* with respect to the relationship between the observer and his or her environment. Furthermore, this structure is lawfully determined by the physical properties of light. In the same way, events in the environment that have acoustic consequences structure patterns in pressure waves by virtue of the physical properties of sound (Gibson, 1979). Direct realists therefore assert that direct and unmediated perceptual access to the causes of ecologically important events is possible simply by virtue of the lawful relations between the patterns of energy observable by a perceiver and the events that caused them.

According to this theoretical perspective, the question for audiovisual speech research is not to determine how auditory and visual signals are translated into a common representational format, but rather to discover the ways in which optical and acoustical signals structure or specify the informative properties of vocal tract articulation. The present results add to the now growing body of research suggesting that both the acoustic and optical properties of speech carry *kinematic* or *dynamic* information about the articulation of the vocal tract (Rosenblum, 1994), and that these sources of information drive the perception of linguistically significant utterances (Summerfield, 1987). As demonstrated in the present investigation, kinematic parameters provide enough detailed information to support spoken word recognition and talker identification. The kinematic description for speech information may then dovetail with attempts to recast computational models of the mental lexicon in terms that incorporate both linguistic and indexical information in memory for spoken words (Goldinger, 1998). Indeed, a modality-neutral basis for lexical knowledge has been implicated in recent investigations of speechreading that show an underlying similarity in the factors relevant to visual-alone and auditory-alone spoken word recognition. For example, Auer (2002) found that phonological neighborhood density (the number of words in the lexicon that are phonologically similar to a target word) influences speechreading for isolated words. Lachs and Pisoni (submitted) also reported similar results, showing that word frequency and phonological neighborhood frequency (i.e., the average frequency of phonological neighbors) also affect speechreading performance.

Such a reconceptualization of the mental lexicon might also help to integrate direct realist theories of speech perception (Fowler, 1986) with research on the traditionally higher-order, “cognitive” mechanisms involved in spoken word recognition and language processing (see Luce and Pisoni, 1998). At first glance, these abstract symbol processing approaches seem inherently incompatible with ecologically oriented approaches to psychology. However, some recourse to lexical knowledge must be incorporated into any theory of speech perception, direct-realist or otherwise. By grounding perceptual information about speech in a modality-neutral, ecologically plausible framework, it may be possible to reformulate traditional cognitive architectures in a format that is more amenable to ecological principles (see also Brancazio, 1998).

We should note that the modality-neutral form of talker-specific information suggested here is restricted to sensory patterns that are lawfully structured by articulatory events, and not to static visual features that are commonly implicated in models of face recognition [e.g., configuration of facial features, color, shading, and shape; see Bruce (1988)]. This observation raises an interesting dichotomy: the difference between visual kinematic information about *facial* identity versus visual kinematic information about *vocal* identity. The point-light displays used in the present experiment contained sufficient detail about the idiosyncratic speaking style (or “voice”) of the talker to specify the auditory form of the utterance. Thus, the kinematic motions that were useful to participants in the present experiment could be called “visual voice” or “visual indexical” information. In contrast, visual motion information can be useful for identifying the person portrayed in the visual display, *irrespective of* that person’s unique vocal qualities or articulatory behaviors.

Bruce and Valentine (1988), for example, showed that participants could identify familiar faces when shown point-light displays of familiar persons making different facial expressions and moving their heads. This kind of dynamic information reflects the identity of the person but is not related to the idiosyncratic speaking style of the talker, because the faces to be identified in their study were not speaking. Moreover, visual identity information can also be obtained from point-light displays of talkers *speaking*, as discussed above (Rosenblum *et al.*, 2002).

Evidence for the dichotomy between visual face and visual indexical information has also been found in neuropsychological investigations of prosopagnosia and aphasia. For example, Campbell *et al.* (1986) reported on two patients: Mrs. “D.” had the ability to speechread and was susceptible to the McGurk effect, despite having a profound prosopagnosia that rendered her incapable of discriminating the identity of various faces. In contrast, another patient, Mrs. “T.,” showed no ability to lipread, yet she could recognize faces quite well.

It should be emphasized that the present results are only relevant to the ability of optic displays to convey voice information; modality-neutral information for voice perception appears to be conveyed by kinematic movements of the articulators. It remains to be seen whether such information is also useful for discriminating faces in the absence of speech. For example, it is not clear from these results whether training with point-light displays of a person engaged in the act of speaking would facilitate the subsequent recognition of that person’s face engaged in *other* types of facial motion (e.g., laughing, smiling, frowning, nodding, etc.). What is clear from the present results, however, is that the idiosyncratic speaking style of a talker is partially conveyed in modality-neutral form and is comparable across visual and auditory sensory modalities as long as the underlying kinematics of the time-varying pattern of articulation is preserved.

The present findings demonstrate that theoretically isolated kinematic information in both the optical and acoustic displays of speech are able to specify talker-specific information that can be successfully combined during the process of speech perception. Future work on audiovisual speech perception may need to formulate more rigorous theoretical descriptions of the nature of audiovisual speech integration effects by recasting speech information in terms of its dynamic, articulatory, and ultimately modality-neutral form.

Acknowledgments

This research was supported by NIH-NIHCD Training Grant No. DC00012 and NIDCD Research Grant No. DC-00111 to Indiana University. The authors would like to thank Luis R. Hernandez, Ralph Zuzolo, Alan Mauro, Jeff Karpicke, and Jeff Reynolds for invaluable assistance during all phases of this research. In addition, the authors thank Geoffrey Bingham, Robert Port, Thomas Busey, and four anonymous reviewers for their valued input and advice.

References

- Auer ET Jr. The influence of the lexicon on speechread word recognition: Contrasting segmental and lexical distinctiveness. *Psychonom. Bull. Rev.* 2002; 9(2):341–347.
- Bergeson, TR.; Pisoni, DB.; Lachs, L.; Reese, L. Audiovisual integration of point light displays of speech by deaf adults following cochlear implantation. *Proceedings of the 15th International Conference on Phonetic Sciences; Causal Productions, Adelaide, Australia.* 2003. p. 1469-1472.
- Bernstein LE, Demorest ME, Tucker PE. Speech perception without hearing. *Percept. Psychophys.* 2000; 62(2):233–252. [PubMed: 10723205]
- Bingham, GP. Dynamics and the problem of visual event recognition. In: Port, RF.; Van Gelder, T., editors. *Mind as Motion: Explorations in the dynamics of cognition.* Cambridge, MA: The (MIT, Cambridge); 1995. p. 403-448.
- Blessner B. Speech perception under conditions of spectral transformation: I. Phonetic characteristics. *J. Speech Hear. Res.* 1972; 15(1):1–41.
- Braida LD. Crossmodal integration in the identification of consonant segments. *Q. J. Exp. Psychol.* 1991; 43A(3):647–677.
- Brancazio, L. University of Connecticut; 1998. Contributions of the lexicon to audiovisual speech perception. unpublished Ph.D., dissertation
- Bruce, V. *Recognising Faces.* Hove, UK: Erlbaum; 1988.
- Bruce, V.; Valentine, T. When a nod's as good as a wink: The role of dynamic information in facial recognition. In: Gruneberg, MM.; Morris, PE.; Sykes, RN., editors. *Practical Aspects of Memory: Current research and issues: Vol. 1. Memory in everyday life.* New York: Wiley; 1988. p. 169-174.
- Campbell R, Landis T, Regard M. Face recognition and lipreading: A neurological dissociation. *Brain.* 1986; 109:509–521. [PubMed: 3719288]
- Fowler CA. An event approach to the study of speech perception from a direct-realist perspective. *J. Phonetics.* 1986; 14:3–28.
- Fowler CA, Dekle DJ. Listening with eye and hand: Cross-modal contributions to speech perception. *J. Exp. Psychol. Hum. Percept. Perform.* 1991; 17(3):816–828. [PubMed: 1834793]
- Gibson, JJ. *The Ecological Approach to Visual Perception.* Boston: Houghton-Mifflin; 1979.
- Goldinger SD. Echoes of echoes? An episodic theory of lexical access. *Psychol. Rev.* 1998; 105(2): 251–279. [PubMed: 9577239]
- Grant KW. The effect of speechreading for masked detection thresholds for filtered speech. *J. Acoust. Soc. Am.* 2001; 109:2272–2275. [PubMed: 11386581]
- Grant KW, Seitz PF. Measures of auditory-visual integration in nonsense syllables and sentences. *J. Acoust. Soc. Am.* 1998; 104:2438–2450. [PubMed: 10491705]
- Grant KW, Seitz PF. The use of visible speech cues for improving auditory detection of spoken sentences. *J. Acoust. Soc. Am.* 2000; 108(3):1197–1208. [PubMed: 11008820]

- Grant KW, Walden BE, Seitz PF. Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *J. Acoust. Soc. Am.* 1998; 103:2677–2690. [PubMed: 9604361]
- Johansson G. Visual perception of biological motion and a model for its analysis. *Percept. Psychophys.* 1973; 14:201–211.
- Kamachi M, Hill H, Lander K, Vatikiotis-Bateson E. ‘Putting the face to the voice’: Matching identity across modality. *Curr. Biol.* 2003; 13:1709–1714. [PubMed: 14521837]
- Kozlowski LT, Cutting JE. Recognizing the gender of walkers from point-lights mounted on ankles: Some second thoughts. *Percept. Psychophys.* 1978; 23:459.
- Lachs L, Pisoni DB. Crossmodal source information and spoken word recognition. *J. Exp. Psychol. Hum. Percept. Perform.* 2004a; 30(2):378–396. [PubMed: 15053696]
- Lachs L, Pisoni DB. Crossmodal source identification in speech perception. *Ecol. Psychol.* (in press).
- Lachs L, Pisoni DB. Spoken word recognition without audition. *Percept. Psychophys.* (submitted).
- Lewkowicz DJ. Infants’ perception of the audible, visible and bimodal attributes of multimodal syllables. *Child Dev.* 2001; 71(5):1241–1257. [PubMed: 11108094]
- Lieberman A, Mattingly I. The motor theory revised. *Cognition.* 1985; 21:1–36. [PubMed: 4075760]
- Luce PA, Pisoni DB. Recognizing spoken words: The neighborhood activation model. *Ear Hear.* 1998; 19:1–36. [PubMed: 9504270]
- Massaro, DW. *Perceiving Talking Faces: From speech perception to a behavioral principle.* Cambridge, MA: MIT; 1998.
- McGurk H, MacDonald J. Hearing lips and seeing voices. *Nature (London).* 1976; 264:746–748. [PubMed: 1012311]
- Munhall, KG.; Vatikiotis-Bateson, E. The moving face during speech communication. In: Campbell, R.; Dodd, B.; Burnham, D., editors. *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-visual Speech.* East Sussex, UK: Psychology; 1998. p. 123-139.
- Reisberg, D.; McLean, J.; Goldfield, A. Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In: Dodd, B.; Campbell, R., editors. *Hearing by Eye: The psychology of lip reading.* Hillsdale, NJ: Erlbaum; 1987. p. 97-114.
- Remez RE, Fellowes JM, Rubin PE. Talker identification based on phonetic information. *J. Exp. Psychol. Hum. Percept. Perform.* 1997; 23(5):651–666. [PubMed: 9180039]
- Remez RE, Rubin PE, Pisoni DB, Carrell TD. Speech perception without traditional speech cues. *Science.* 1981; 212:947–950. [PubMed: 7233191]
- Remez RE, Fellowes JM, Pisoni DB, Goh WD, Rubin PE. Multimodal perceptual organization of speech: Evidence from tone analogs of spoken utterances. *Speech Commun.* 1999; 26(1–2):65–73. [PubMed: 21423823]
- Rosenblum LD. How special is audiovisual speech integration? *Curr. Psychol. Cognition.* 1994; 13(1): 110–116.
- Rosenblum LD, Saldaña HM. An audiovisual test of kinematic primitives for visual speech perception. *J. Exp. Psychol. Hum. Percept. Perform.* 1996; 22(2):318–331. [PubMed: 8934846]
- Rosenblum LD, Johnson JA, Saldaña HM. Point-light facial displays enhance comprehension of speech in noise. *J. Speech Lang. Hear. Res.* 1996; 39:1159–1170.
- Rosenblum LD, Smith NM, Nichols SM, Hale S, Lee J. Hearing a face: Cross-modal speaker matching using isolated visible speech. *Percept. Psychophys.* (under review).
- Rosenblum LD, Yakel DA, Baseer N, Panchal A, Nodarse BC, Niehaus RP. Visual speech information for face recognition. *Percept. Psychophys.* 2002; 64(2):220–229. [PubMed: 12013377]
- Runeson, S. Perception of biological motion: The KSD-principle and the implications of the distal versus optimal approach. In: Jansson, G.; Bergström, SS.; Epstein, W., editors. *Perceiving Events and Objects.* Hillsdale, NJ: Erlbaum; 1994. p. 383-405.
- Runeson S, Frykholm G. Visual perception of lifted weight. *J. Exp. Psychol. Hum. Percept. Perform.* 1981; 7:733–740. [PubMed: 6457088]
- Schultz M, Norton S, Conway-Fithian S, Reed C. A survey of the use of the Tadoma method in the United States and Canada. *Volta Rev.* 1984; 86:282–292.

- Schwartz, J-L.; Robert-Ribes, J.; Escudier, P. Ten years after Summerfield: A taxonomy of models for audio-visual fusion in speech perception. In: Campbell, R.; Dodd, B.; Burnham, D., editors. *Hearing by Eye II*. East Sussex, UK: Psychology; 1998. p. 85-108.
- Smith ZM, Delgutte B, Oxenham AJ. Chimaeric sounds reveal dichotomies in auditory perception. *Nature (London)*. 2002; 416:87–90. [PubMed: 11882898]
- Sumby WH, Pollack I. Visual contribution of speech intelligibility in noise. *J. Acoust. Soc. Am.* 1954; 26:212–215.
- Summerfield, Q. Some preliminaries to a comprehensive account of audio-visual speech perception. In: Dodd, B.; Campbell, R., editors. *Hearing by Eye: The Psychology of Lip-Reading*. Hillsdale, NJ: Erlbaum; 1987. p. 3-51.
- Vatikiotis-Bateson, E.; Hill, H.; Kamachi, M.; Lander, K.; Munhall, K. The stimulus as basis for audiovisual integration. paper presented at the International Conference on Spoken Language Processing; Denver, CO. 2002.
- Vatikiotis-Bateson, E.; Munhall, KG.; Hirayama, M.; Lee, YV.; Terzopoulos, D. The dynamics of audiovisual behavior in speech. In: Stork, DG.; Hennecke, ME., editors. *Speechreading by Humans and Machines*. Berlin: Springer-Verlag; 1997. p. 221-232.
- Yehia H, Rubin PE, Vatikiotis-Bateson E. Quantitative association of vocal-tract and facial behavior. *Speech Commun.* 1998; 26(1–2):23–43.

FIG. 1.

Dot configuration for point-light displays. All dots were of a uniform diameter. Five dots are not visible due to occlusion by the lips: upper teeth (two dots), lower teeth (two dots), and tongue tip (one dot).

FIG. 2.

The top panel shows a spectrogram of talker F1 speaking the word “WAIL.” The bottom panel shows a spectrogram of a synthetic sinewave replica of this token.

FIG. 3.

Schematic of the cross-modal matching task. The top row illustrates the task in the "V-to-A order." The bottom row illustrates the task in the "A-to V order." Faces denote stimuli that are presented visual-alone. Waveforms denote stimuli that are presented auditory-alone. Δt is always 500 ms.

FIG. 4. Average proportion correct in the point-light to sinewave speech (PL+SWS) and point-light to untransformed audio (PL-AO) conditions of experiment 1. Error bars show standard errors.

FIG. 5.

Average proportion correct in the auditory-alone (SWS) and audiovisual (PL+SWS) conditions of experiment 2. Error bars show standard errors.

FIG. 6.

R scores for each participant in experiment 2, sorted by performance. *R* scores represent the gain over auditory-alone performance due to additional visual information. Auditory stimuli were sinewave speech replicas and visual stimuli were point-light displays. The first four participants in the graph did not show any audiovisual gain.

TABLE I

Descriptive statistics for cross-modal matching performance when visual stimuli were point-light displays (PL) and auditory stimuli were either untransformed, natural utterances (AO) or sinewave speech replicas (SWS).

Matching condition	Order.	Mean	SE	<i>t</i> vs. 0.5
PL-AO	A-V	0.538	0.017	2.17 ^a
	V-A	0.547	0.015	3.19 ^b
PL-SWS	A-V	0.541	0.016	2.58 ^a
	V-A	0.541	0.015	2.78 ^a

^a $p < 0.05$.

^b $p < 0.01$.