An audiovisual test of kinematic primitives for visual speech perception

Lawrence D. Rosenblum and Helena M. Saldaña
University of California, Riverside

Published as: Rosenblum, L.D. and Saldaña, H.M. (1996). An audiovisual test of kinematic primitives for visual speech perception. Journal of Experimental Psychology: Human Perception and Performance. 22(2), 318-331.

Abstract

Isolated kinematic properties of visible speech can provide information for lipreading. Kinematic facial information is isolated by darkening an actor's face and attaching dots to various articulators so that only moving dots can be seen with no facial features present. To test the salience of these images, experiments were conducted to determine whether they could visually influence the perception of discrepant auditory syllables. Results showed that these images can influence auditory speech and that this influence is not dependent on subjects' knowledge of the stimuli. In other experiments, single frozen frames of visible syllables were presented with discrepant auditory syllables to test the salience of static facial features. Results suggest that while the influence of the kinematic stimuli was perceptual, any influence of the static featural stimuli was likely based on subject misunderstanding or post-perceptual response bias.

Extracting visible articulatory information can be an important part of the speech perception process. It has long been known that, for those with hearing impairments, lip-read information can supplement the impoverished auditory signal. Moreover, it is likely that listeners with good hearing integrate visual speech information in a noisy environment. Research suggests that seeing a speaker can improve intelligibility to a degree equivalent to increasing the signal-to-noise ratio 15 dB (Sumby and Pollack, 1954 ; Erber, 1969; Middleweerd and Plomp, 1987; MacLeod and Summerfield, 1987; MacLeod and Summerfield, 1990). One way in which visual information enhances a degraded auditory signal is that aspects of segments which are difficult to hear are often relatively easy to see (Walden, Prosek, Montgomery, Scherr & Jones, 1977). For example, whereas auditory information for manner of articulation is relatively stable under acoustically degraded conditions, information for place of articulation is considerably more vulnerable. Visible speech, on the other hand, is often quite informative about place of articulation but does not provide much information about manner.

The importance of visual speech perception is further evidenced by examples involving non-degraded signals. It is known that even with a perfectly clear auditory source, integrating visual speech information can help in recovering a difficult message. Reisberg, McLean, & Goldfield (1987) have found that when one is listening to a speaker with a heavy foreign accent or to a passage with difficult semantic content (e.g., Kant's Critique of Pure Reason), seeing the speaker can help with comprehension.

The language acquisition literature also attests to the importance of visual speech perception. Research suggests that visually-impaired children have trouble acquiring certain phonemic distinctions (Mills, 1987). Visually impaired children make significantly more production errors across viseme boundaries (e.g., /m/ switched for /n/) than do sighted children. These children also attempt fewer words which begin with a visible phoneme. This evidence suggests that the acquisition of some segments is expedited by visual speech perception. Other experiments have demonstrated that infants are sensitive to audiovisual discrepancy in speech. Aronson and Rosenbloom (1971) showed that 10-day old infants show distress when their mother's voice is heard to emanate from a location distal from her face. Also, Dodd (1979) has reported that infants gaze longer at a speaking face where the audio and visual sources are synchronous and are

sensitive to asynchronies as small as 400 ms. Finally, Kuhl and Meltzoff (1982, 1984) demonstrated that 4 month old infants gaze longer at a video image that is vocalically compatible with an auditory signal than one that is vocalically incompatible. In all, these studies suggest that infants are attuned to visible speech information and recognize its relation to auditory speech.

Finally, the saliency of visual speech information is demonstrated by a striking laboratory phenomenon known as the McGurk effect. In the McGurk effect, visual information can sometimes combine with and even override conflicting auditory information, causing a perceiver to report hearing what he/she sees (e.g. McGurk & McDonald, 1976; Rosenblum and Saldaña, 1992; but see Sekiyama and Tohkura, 1991). In one strong version of the McGurk effect, the auditory syllable /ba/ is repeatedly dubbed onto a videotape of a speaker's lips producing the syllables /be/, /ve/, /de/. When shown this dubbed video tape, observers report *hearing* the syllables /ba/, /va/, /da/ (Liberman and Mattingly, 1985). In another example of the effect, a visual /ga/ presented with an auditory /ba/ can induce a 'heard' syllable of /da/ suggesting that various features from each modality can be integrated to produce a separate fused segment. For both visual dominance and fusion examples, the effect is quite striking: even when completely aware of the dubbing procedure, observers still report hearing a clear syllable which is (unconsciously) influenced by what they see. Furthermore, when subjects are explicitly told to attend either solely to the audio or to the video portions of the stimuli, they are still strongly influenced by the 'ignored' informational dimension (Massaro, 1987). The McGurk effect has served to highlight visual speech perception as a significant—and perhaps mandatory—part of the speech process. To quote Summerfield: ". . .any comprehensive account of how speech is perceived should encompass audiovisual speech perception. The ability to see as well as hear has to be integral to the design, not merely a retro-fitted after-thought." (1987, p. 47).

Visual speech information

Although we now know a great deal about auditory speech information, we know relatively little about the visual information for speech (see Summerfield, et al, 1989 for a review). Most analyses of visible speech have involved identifying static features such as (visual information for) place of constriction; open, closed, or rounded lips; and visible teeth (e.g., Montgomery and Jackson, 1983; Petajan, 1984; Summerfield and McGrath 1984; McGrath, 1985). As an example, Montgomery and Jackson (1983) attempted to characterize the information for lipread vowels using a multi-dimensional scaling space based on the static features of degree of lip spreading/rounding and tongue height. Also, Massaro and Cohen (1990) have suggested that degree of lip opening (potentially useful in distinguishing between /ba/, /da/, and / a/) could be a critical visual feature which acts as input to the audiovisual integration process. These descriptions of visual speech features are *static* in that they can be captured in a single still photograph. They are also 'pictorial' in requiring some graphic basis for demarcating facial features such as the *texture and/or color* of the skin on the lips -as distinguished from the surface of the cheeks, teeth, and tongue.

However, another description of visual speech primitives based on more time-varying attributes can be offered (see Summerfield, 1987; and Jackson, 1988). In one of the few studies designed to examine time-varying visible speech information, Brooke and Summerfield (1983) performed analyses of lip and jaw trajectories during production of /aCV/ syllables where the consonant was from the /m, p, b/ viseme category, and the (second) vowel was either /i/, /a/, or /u/. They found that vowel distinctions were observable in the kinematic differences (displacement over time, velocity, acceleration) of these articulators. This suggests that visible speech features can be described along kinematic as well as static dimensions. One question that arises concerns the degree to which visual speech *perception* can make use of these kinematic properties. To explore this question, isolated kinematic variables—without static featural cues—can be tested to determine whether they can be informative about articulation.

Outside the realm of speech perception, isolated kinematic properties, as reflected in optical information, can enable observers to identify events (e.g., Johansson, 1973; Runeson, 1977; Bingham, 1987). Kinematic properties can be visually isolated through a point-light technique. In a classic experiment (Johansson, 1973), small point-lights were placed on the major joints of a

darkened actor's body. The actor was then videotaped in the dark performing various activities (walking, running, jumping). These videotapes were then presented to subjects through a contrast-adjusted video monitor so that only the movement of the lights could be seen. Although these displays could not be identified when the image was frozen, once the image moved, they were quickly recognized as a human form performing specific activities. Subsequently, it has been shown that observers can also recognize actor gender (Kozlowski and Cutting, 1977), amount of weight lifted (Runeson and Frykholm, 1981; Bingham, 1987), and various inanimate events (Bingham, Rosenblum, & Schmidt, in press) through point-light specification.

Since point-light images cannot be recognized when frozen in time, it is likely that these stimuli lack adequate static pictorial cues. Accordingly, findings using point-lights have been interpreted as evidence that visually reflected kinematic form—isolated from static pictorial cues—is sufficient to specify events (e.g., Bingham, Rosenblum, and Schmidt, in press).

Returning to visual speech perception, it may be that the kinematic properties of a speaking face are salient for lipreading. In fact, recent research in our laboratory suggests that isolated kinematic visual information for articulation can be used to lipread (Johnson, Rosenblum, and Saldaña, 1994; Rosenblum, Johnson, and Saldaña, in preparation). Borrowing a methodology first established by Bassili (1978) and Summerfield (1979), we have applied the point-light methodology to an articulating face. Our facial point-light technique involves placing reflective dots on the lips, teeth, tongue, chin, and cheeks. The resultant displays are seen as bright, moving dots against a black background. Our results suggest that facial point-light displays do enable observers to distinguish various vowels, consonants, and sentences as well as *identify* many consonantal viseme categories (Rosenblum, et al, in preparation). We have also shown that these displays can improve the speech reception thresholds of sentences in noise (Johnson, et al, 1994). As with full-body point-light stimuli, we have found that when our images are frozen in time, observers cannot identify them as faces (see also Bassili, 1978; and Berry, 1990). This last finding indicates that our stimuli do not contain recognizable static facial feature information suggesting that the observed lipreading effects are based on kinematic form.

Beyond demonstrating that observers can lipread based on the kinematic dimensions of an articulating face, the point-light technique can help determine the salient visual information for lipreading. The technique affords careful control over the amount and type of information available to an observer at any one time. In our research (Johnson, Rosenblum, and Saldaña, 1994; Rosenblum, Johnson, and Saldaña, in preparation), we have begun to manipulate the placement of point-lights and have found that the degree of lip reading proficiency depends on the number and location of the points on the face. Additionally, point-light images enable straightforward kinematic analyses since there are relatively few visible regions to track (e.g., Bingham, Rosenblum, and Schmidt, in press). Following the moment-to-moment position of facial skin regions would be vastly more difficult. Finally, using point-light displays for both perceptual tests *and* kinematic analyses ensures that all of the perceptually relevant information observers use from such displays can be analyzed kinematically.

In these ways, the point-light technique can be used to uncover the salient visual information for speech in an analogous way to the early work on auditory speech cues (e.g., Delattre, Liberman, Cooper, and Gerstman, 1952; Cooper, Delattre, Liberman, Borst, and Gerstman, 1952; Delattre, Liberman, and Cooper, 1955; Harris, 1958). That research involved analysis of the acoustic signal, conjecture about the salient cues, isolation of those cues by signal modification and synthesis, and perceptual tests of these reduced signals to determine the relative salience of the cues. The point-light technique permits for a similar isolation and determination of the salient dimensions of visual speech.

Given the potential practical and conceptual utility of point-light speech stimuli, it is important to determine the degree to which these reduced images are treated as real visual speech. Although we have shown that speech segments can be recovered from these stimuli (e.g., Rosenblum, et al, in preparation), it is not clear whether this ability is truly perceptual and automatic in nature, or whether it is a result of some conscious, attention-demanding, post-perceptual analysis. One way to examine this issue is to test the degree to which the perception of these stimuli is automatic or mandatory. For example, we can test whether these visual stimuli influence perception when

observers are not asked to make judgments on the images themselves.

There is evidence that intact visual speech information (i.e., with fully-illuminated faces) does influence the speech process in a mandatory way. In the McGurk effect, visual speech affects perception even when observers are asked to base their judgments on the auditory component. In that the effect is impenetrable, it is generally considered a true perceptual phenomenon, and not one based on post-perceptual decision biases (Liberman and Mattingly, 1985; Rosenblum and Saldaña, 1992). Along these lines, Summerfield, McGrath and their colleagues (McGrath, 1985; Summerfield, MacLeod, McGrath, and Brooke, 1989) have used the McGurk effect as a tool to determine whether synthesized visual face stimuli (comprised of outlines of the major facial features) were recognized through perceptual—and not conscious-inference—processes. (Details of this experiment are presented in the Discussion section of Experiment 4.) Will point-light speech stimuli influence 'heard' speech in a similar fashion? If so, then evidence would be gained that there is enough information in these stimuli to be perceived as real visual speech. This finding would provide further support that the kinematic properties of visual speech are perceptually salient.

In two experiments, we test whether point-light syllables can influence heard speech to the extent that they 'sound' like the syllables conveyed visually. If so, we would have evidence that our point-llight stimuli are treated as real visual speech (meaning that it allows for segment recovery in a perceptual, automatic way and does not require conscious, post-perceptual analysis). Additional experiments test the degree to which static, fully-illuminated images can influence heard speech. These latter experiments were conducted to determine whether there is enough information available in static visual speech features to integrate with auditory speech. For the syllables used in our experiments, we selected an acoustic /ba/ (with an initial voiced bilabial plosive) paired with an optical /va/ (voiced labiodental fricative). We have found that this audiovisual pairing can result in a visually-influenced percept of /va/ over 96% of the time (with intact face stimuli) (Rosenblum & Saldaña, 1992; Saldaña and Rosenblum, 1993; Saldaña and Rosenblum, 1994).

The first experiment tests whether point-light images of a produced /va/ influence perception of the auditory /ba/. If point-light stimuli are treated as real visual speech, then this visual syllable should influence the heard speech even when observers are asked to base their judgments on what they hear. In order to ensure that any visual influence is due to the kinematics and not recognizable facial features, we also test whether these stimuli can be recognized as faces when frozen. Finally, fully-illuminated (dynamic) facial stimuli will be tested in order to provide a comparison to any point-light visual influence.

EXPERIMENT 1

Method

Subjects

Sixteen undergraduates at the University of California, Riverside, participated for partial fulfillment of a class requirement. All reported normal or corrected vision, and were native speakers of English.

Stimuli

Two types of video displays were prepared. The fully-illuminated (FI) display involved recording an actor in a fully lit room with no alteration to the actor's face. For the point-light (PL) display, the same actor's face was blackened with theatrical make-up, his teeth were blackened with theatrical tooth-black, and his tongue was darkened with food coloring (Scheinberg, 1980; Berry, 1990). Dots made up of retro-reflective tape were attached to the actor's face. The dots were 3mm in diameter. They were attached to the actor's skin using a medical cement and to the teeth and tongue using a dental adhesive. Twenty-eight dots were placed on the face and articulators. A schematic representation of the dot configuration is shown in Figure 1. Four dots were placed on the upper and lower central incisors, one dot was placed on the tongue tip, six dots were placed on the lips (one in each corner, two on the upper lip, and two on the lower lip), two dots were placed above the lips, four dots on the chin, eight dots on the cheeks, two dots on the jaw (on either side of the chin), and one dot on the nose tip. For the PL display, the actor was

videotaped producing syllables under low illumination. This resulted in a video display in which only the dots and their motions were visible.

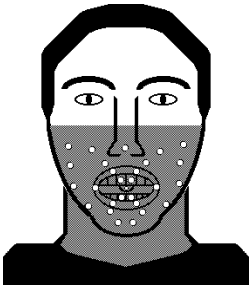

Figure 1. Schematic representation of point light configuration (see text for details).

A Panasonic PVS350 camcorder and SM57 microphone were used to record the initial audiovisual tape. For both FI and PL displays, the actor was seated five feet in front of the camera. His head was placed in a wire head brace to inhibit movement. The camera was centered on the actor's lips so that the recorded image consisted of the bottom of the actor's nose to the bottom of his chin. The actor was recorded articulating the syllables /ba/ and /va/ six times each in both the FI and PL conditions.

A Compaq 386/25 computer and two Panasonic video recorders were used for dubbing the audiovisual tokens. A good exemplar of an audio /va/ from the original tape was digitized. As a control we wanted both the /ba/ and /va/ audio tokens to have the same general acoustic characteristics (e.g., intensity, fundamental frequency contour). Therefore a hybrid /ba/ token was created by deleting the first 55 ms of the /va/ token.

The acoustic /va/ and /ba/ tokens were then dubbed synchronously onto the FI and PL /ba/ and /va/ tokens. To dub each token, the original tape was played so that its video signal was output to the recorder and its audio signal was output to a sound activated circuit that was interfaced with the computer. Upon sensing the audio signal, the sound activated circuit signaled the computer to output an audio token to the video recorder. Thus, the video token of the original tape and the audio token from the computer file were recorded simultaneously onto a second tape resulting in a new synchronous audiovisual token. Both consistent and discrepant tokens were dubbed using this same procedure so that the dubbing lag time was comparable for all audiovisual stimuli. The lag time for dubbing was found to be no more then 9.4 ms, well below the 80 ms range which is required for subjects to notice an audio-visual discrepancy (McGrath & Summerfield, 1985).

The audiovisual tokens for both the PL and FI conditions consisted of a consistent audiovisual /va/ token (audio /va/-visual /va/), a consistent audiovisual /ba/ token (audio /ba/-visual /ba/), and a discrepant audiovisual token (audio /ba/-visual /va/). Each audiovisual token was two seconds long allowing the entire articulatory event to be seen.

The presentation tape was set up in the following order: The first forty audiovisual tokens involved the PL visual displays. There were 20 consistent audiovisual /ba/ tokens, 10 consistent audiovisual /va/ tokens, and 10 discrepant audiovisual tokens. The order of these tokens was completely randomized and recorded on a third tape. This was followed by a PL visual alone condition (the auditory portion of the tape was turned off). This condition consisted of 10 /va/

articulations and 10 /ba/ articulations. The visual alone condition was followed by an audio alone condition which also consisted of 10 /va/ tokens and 10 /ba/ tokens. The FI stimuli followed the PL stimuli with the same general setup: 40 randomized audiovisual tokens, followed by 20 visual alone tokens, followed by 20 audio alone tokens. In all, the entire presentation tape consisted of 160 tokens. All tokens were separated by a 3 s ISI and blocks of conditions were separated by 10 s.

Procedure

Subjects were run in groups of two or three. They were seated at a table 5 feet in front of a video monitor. The audio stimuli were presented through a loudspeaker positioned directly beneath the monitor. The lights were turned off in the presentation room. The only sources of illumination were the television monitor and two small lights which were positioned and focused on the table so that subjects could see their response sheets. For the first half of the experiment (PL condition), the contrast was adjusted on the monitor so that no facial contour was visible.

Subjects in this experiment were completely naive regarding the nature of the stimuli. The experimenter first told subjects that they were going to see an image on the video screen and they were to write a description of what they thought the image was depicting. We stressed to subjects that we were not interested in a projective interpretation of the stimuli, simply a best guess of what the image was. The experimenter then presented a static version of the PL display. This was accomplished by freezing the first frame on the presentation tape. The first token on the presentation tape was an audiovisual /ba/ token; however because the image was frozen on the first frame, it consisted of a neutral articulatory position with lips slightly parted. Subjects were given as much time as they wanted to respond. Following this task, subjects were told that for the remainder of the experiment, they would be required to watch as well as listen to stimuli. They were told that after each token was presented, they were to quickly write down what they heard and then look back up to the monitor for the next presentation. They were told that it was important to watch each presentation; however they were to write down only what they heard. For the audio alone blocks, the video monitor was switched off and subjects were asked to write down what they heard. For the video alone blocks, the loudspeaker was shut off and subjects were asked to write down what they thought they might hear if the event they saw was producing some sound. Following the presentation of all the PL tokens (including audio and video alone blocks), subjects were asked to write down again what they thought the PL image was depicting.

The PL stimulus block was then followed by presentation of the FI audiovisual tokens. Again, subjects were told to watch the monitor but base their judgments on what was heard.

Results and Discussion

None of the sixteen subjects reported seeing the static point light image as a face. The most common description given by subjects was that the image was of a butterfly. Following the presentation of the moving point light stimuli, all but three subjects described the image as some sort of mouth (an animal mouth or a human mouth). These three subjects described the image as an owl, butterfly, and sheep.

The pooled percentage of 'correct' responses for each token type are shown in Table 1. A correct response was recorded whenever a subject responded with the initial consonant presented on the *audio* portion of the tape. In the case of the discrepant tokens, a visual influence is demonstrated if the percentage of correct responses is low and the response is instead based on the visual information presented. In the present experiment, all alternative responses consisted of either /va/ or /fa/ responses. These two responses can be considered visually- influenced because they consist of the same viseme.

Table 1. Pooled percentage of correct responses for each token type for Experiment 1.

Token		Percentage Correct (based on audio)	
		Point Light Face	Fully Illuminated Face
Audiovisual	/ba-ba/	87	99
	/va-va/	97	100

	/ba-va/	45	01
Audio-Alone	/ba/	89	91
	/va/	98	100
Video-Alone	/ba/	90	96
	/va/	91	98

In both the point-light and fully-illuminated conditions, the mean percentage of correct responses is noticeably lower for the discrepant condition than for the other types of tokens. A t-test revealed that the percentage of correct responses in the discrepant PL condition was significantly lower than the percentage of correct responses in the audio alone /ba/ condition $t(15)=5.693$, $p<.001$. These results indicate that the visual point-light stimuli did significantly influence the heard speech —even when subjects were asked to base their judgments on what they heard. Thus, like intact (fully-illuminated) visual speech, the point-light stimuli were influential to the extent that they affected auditory judgments. Based on the above arguments, this finding suggests that the point-light stimuli were treated as real visual speech. Also, because these images could not be identified as faces when frozen, it can be surmised that they did not involve *recognizable* static facial features. Thus, these initial findings provide support that the isolated kinematic properties of a moving face are treated as real visual speech information.

As expected, a significant visual influence was also observed for the fully-lit facial stimuli based on comparison to the audio alone /ba/ condition, $t(15)= 19.03$, $p<.001$. This result replicates findings observed by us and others (e.g., Repp, Manuel, Liberman, & Studdert-Kennedy, 1983; Rosenblum and Saldaña, 1992). A final comparison was performed between the results for the discrepant PL condition and the discrepant FI condition. A t-test revealed a significant difference $t(15)=6.809$, $p<.001$. This result indicates that the visual influence of the fully illuminated face was significantly greater than the visual influence of the point-light images.

This last finding can be given a number of interpretations. First, it may be that the arrangement of point-lights used in this experiment was not optimal for capturing all of the salient kinematic dimensions. It is known that the placement of point-lights is important for accurate portrayal of full-body activities such as walking and running (Runeson, 1993). Generally, when the points are placed on the major joints of the body such that there is a rigid limb segment between each, observers can easily recognize the performed activities. However, if the points are placed on the limbs between the joints, observers have more difficulty interpreting the displays. Since there are relatively few rigid portions of a human face, we could not guide our point-light placement on this principle. Instead, our placement of points was based mainly on intuition informed by other research using facial point-light stimuli (e.g., Bassili, 1978; Summerfield, 1979; Berry, 1990). Thus, although these point-light images did portray enough kinematic information to induce a visual influence, it is quite possible that the point arrangement did not provide the optimal information for audiovisual integration.

An alternative interpretation of our PL vs. FI results can be offered. It is possible that the point-light images did not induce a strong visual influence because they were not treated as speech to the same degree. It could be argued that it was not the kinematic information itself that influenced the heard speech. Perhaps the point-light images were only effective in inducing a visual influence to the degree that the points were matched to representations of static facial features stored in memory. Although observers could not recognize the array of dots as faces when the image was frozen, it might be that once the image moved—allowing subjects to recognize faces—the dots were interpreted as facial features. These features could then be recognized through, for example, a prototype description matching process and thereby induce a visual influence (cf., Massaro, 1987).

It is the case that 13 of 16 subjects did ultimately recognize the point-light stimuli as an articulating face. For these subjects, it is possible that the visual influence was based on a static featural interpretation of the point lights. However, three subjects never recognized the dynamic point-light images as a face yet *still* showed a significant visual influence (68% of the time). This may indicate that a visual influence is not necessarily dependent on facial recognition and

interpretation of the points as facial features. However, this construal is based on the subjective reports of three subjects and should be taken with caution.

In order to further explore whether interpretation of the point-light stimuli as a face affects the degree of visual influence, a second experiment was conducted. Experiment 2 made use of the same stimuli and general task of Experiment 1. In Experiment 2 however, subjects were fully informed about the nature of the point-light stimuli. Before judging the PL stimuli, subjects were told explicitly what the stimuli were, how they were made, and where the dots were placed on the face. Clearly, informing subjects about the nature of the stimuli allows them to see the images as articulating faces more easily. If the strength of the visual influence observed in Experiment 1 was based on the degree to which the points were interpreted as facial features, then detailed knowledge of the stimuli should allow observers to see facial features more easily and induce a greater visual influence. If on the other hand, the visual influence was *not* dependent on subjects' recognition of the point-lights as facial features, then having explicit knowledge about the displays should not necessarily increase the visual influence. Under the second hypothesis, our point-light stimuli are treated as real visual speech information, and kinematic variables are salient primitives for visual speech.

EXPERIMENT 2

Method

Subjects

Sixteen undergraduates at the University of California, Riverside served as subjects for partial fulfillment of a class requirement. All reported normal or corrected vision and were native speakers of English. None of the subjects in Experiment 2 participated in Experiment 1.

Stimuli

The same stimulus tape used in Experiment 1 was used in Experiment 2, with only the instruction set changed.

Procedure

The general procedure for Experiment 2 was the same as for Experiment 1. However, in Experiment 2 subjects were not asked to identify a static image of the point light display. Instead, subjects were made completely aware of the technique used to produce the visual displays. Additionally, the experimenter pointed out the placement of the dots on a frozen image of the point light face. The subjects were told that they would be presented with a video of an actor producing speech syllables. As in Experiment 1, they were instructed to watch as well as listen to the syllables and to respond by writing down only the syllable that they heard. The PL block of trials was again followed by FI, audio-alone, and video alone blocks.

Results and Discussion

The pooled percentage of correct responses for each token type are shown in Table 2. Again, a response is counted as correct if the subject wrote down the initial consonant that was presented auditorily. A visual influence is demonstrated in the discrepant conditions when the percentage of correct responses is low, and the response is based on the visual information presented. As in the first experiment, all alternative responses were either /va/ or /fa/.

Table 2. Pooled percentage of correct responses for each token type for Experiment 2.

Token		Percentage Correct(based on audio)	
		Point Light Face	Fully Illuminated Face
Audiovisual	/ba-ba/	91	99
	/va-va/	100	100
	/ba-va/	44	01
Audio-Alone	/ba/	96	98
	/va/	100	100
Video-Alone	/ba/	83	99
	/va/	89	92

A comparison between the audio-alone /ba/ token and the discrepant PL token revealed a significant difference, $t(15) = 5.317$, $p < .001$. As in Experiment 1 this indicates that the visual PL information significantly affected subjects' auditory judgments. A t-test also revealed a significant difference between the percentage correct for FI and audio-/ba/ tokens, $t(15) = 50.35$, $p < .001$, indicating a significant effect of FI video on auditory judgments. As in the first experiment, there was a significant difference in the pooled percentage of correct responses for the discrepant PL tokens versus the discrepant FI tokens $t(15) = 6.809$, $p < .001$. This indicates that the FI condition was more effective at influencing subjects' responses.

The critical question in Experiment 2 was whether informing subjects about the PL stimuli would induce a greater visual influence. In order to test this question, a comparison was conducted between the discrepant PL results of Experiment 1 and Experiment 2. Tables 1 and 2 show that the mean visual influence was very similar for both PL conditions (55% for Experiment 1 and 56% for Experiment 2). A t-test revealed no significant difference between the two PL conditions, $t(15) = .173$, $p > .05$. This finding suggests that explicit knowledge of the PL stimuli does not affect the degree of visual influence. Again, if a visual influence is dependent on subjects interpreting the point-lights as facial features, then explicit information about how the points were situated on the face should have induced a stronger influence. The fact that no such increase in effect size occurred provides evidence that static facial features are not necessary for a visual influence. This provides support that the visual influence observed in both experiments was based on kinematic properties and that our point-light stimuli are treated as real visual speech.

However, Experiment 2 serves to demonstrate that *explicit* information about the PL stimuli has no effect on the visual influence. This result does not preclude the possibility that *implicit* information about the stimuli as articulating faces allowed subjects to interpret the point-lights as facial features. It could be that in both experiments, subjects' experience with the stimuli allowed them to see the points as features. Accordingly, we decided to examine the basis of the point-light speech visual influence further by testing static facial stimuli in Experiment 3. Experiment 3 was also designed to determine whether there is enough information available in static visual speech features to integrate with auditory speech.

EXPERIMENT 3

Experiments 1 and 2 demonstrate that point-light articulating faces can influence heard speech. This finding is germane to the question of whether point-light stimuli are treated as real visual speech and whether kinematic primitives are salient for lipreading. Thus, barring the concerns mentioned above, Experiments 1 and 2 suggest that observers can use kinematic information alone for audiovisual speech integration. The question arises as to whether isolated static featural information—without any kinematic form—can be used for speech integration. It is known that observers can recognize various emotional expressions from static photographs of human faces (see Ekman, Friesen, and Ellsworth, 1972, for a review). Further, various speech gestures can be conveyed by photographs (e.g., Campbell, 1986). However, it is unclear to what extent static images are treated as real visual speech. To examine this question, static speech images can be tested in an audiovisual paradigm to determine whether they can induce a visual influence on heard speech. In other words, is there enough information in static visual images to integrate with auditory speech?

In Experiment 3, we test this question by implementing fully-illuminated static facial images in a McGurk-type paradigm. As in Experiments 1 and 2, we have chosen to test the influence of a visual /va/ on an auditory /ba/. Experiment 3 also involves tests of video alone static /va/ and /ba/. For both syllables, the static image was chosen based on what was considered the most characteristic articulatory position. Informal pilot studies determined that these selected static images could be easily interpreted as the phonemes /v/ and /b/. We reasoned that if static speech images are informative to the degree they are treated as real visual speech, then a visual influence would be observed for these stimuli. If on the other hand, a kinematic component is necessary,

then no visual influence should be observed.

Method

Subjects

Fourteen undergraduates at the University of California, Riverside served as subjects for partial fulfillment of a class requirement. All reported normal or corrected vision, and were native speakers of English. None of the subjects in Experiment 3 participated in Experiment 1 or 2.

Stimuli

The audio stimuli for the third experiment were the same as in Experiments 1 and 2. The visual stimuli consisted of a static fully illuminated (SFI) display and the (dynamic) fully illuminated (FI) display used in the first two experiments. The SFI displays were created by freezing a critical frame of the FI displays on a Panasonic video player and recording the single frame on a second tape for two seconds. The critical frames for both the /va/ and the /ba/ articulations were defined as the point just prior to release of the labio-dental and bilabial constriction, respectively. The audio syllables were then dubbed onto the frozen video display. The 600 and 545 ms audio tokens (/va/ and /ba/ tokens) were dubbed onto each video token so that there was an equal amount of time of video only token before the onset of the auditory syllable and offset of the auditory syllable. Again, the audiovisual tokens for both the SFI and FI conditions consisted of a consistent audiovisual /va/ token (audio /va/-visual /va/), a consistent audiovisual /ba/ token (audio /ba/-visual /ba/), and a discrepant audiovisual token (audio /ba/-visual /va/).

The presentation tape was organized in the same manner as the tape used in Experiments 1 and 2: The first forty audiovisual tokens involved the SFI visual displays. There were 20 consistent audiovisual /ba/ tokens, 10 consistent audiovisual /va/ tokens, and 10 discrepant audiovisual tokens. The order of these tokens was completely randomized and recorded on a third tape. This was followed by a SFI visual alone condition (the auditory portion of the tape was turned off). This condition consisted of 10 /va/ articulations randomized with 10 /ba/ articulations. The visual alone condition was followed by an audio alone condition which also consisted of 10 /va/ tokens and 10 /ba/ tokens. Again, the FI stimuli followed the SFI stimuli with the same general setup: 40 audiovisual tokens, followed by 20 visual alone tokens, followed by 20 audio alone tokens. The entire presentation tape consisted of 160 tokens. All tokens were separated by a 3 s ISI.

Procedure

The subjects were told that they would be presented with a video of an actor producing speech syllables. They were instructed to watch as well as listen to the syllables and to respond by writing down only the syllable that they heard. The subjects were warned ahead of time that the first 60 tokens would consist of a frozen frame of an actor producing syllables.

Results and Discussion

The pooled percentages of correct responses for each utterance type are shown in Table 3. First, the high mean percentages for the video alone results (100%) suggest that subjects could distinguish the frozen facial images as /va/ and /ba/. It should be kept in mind however, that subjects first saw these images paired with /va/ and /ba/ audio tokens. Accordingly, although subjects could *distinguish* these static images, it is unclear whether they could have interpreted the images as /va/ and /ba/ without the priming provided by this first condition.

Table 3. Pooled percentage of correct responses for each token type for Experiment 3.

Token		Percentage Correct(based on audio)	
		Static	Dynamic
Audiovisual	/ba-ba/	99	99
	/va-va/	98	100
	/ba-va/	65	02
Audio-Alone	/ba/	98	95
	/va/	100	98

Video-Alone	/ba/	100	100
	/va/	100	99

As in the first two experiments, all alternative responses for the audiovisual conditions were either /va/ or /fa/. A comparison between the audio alone /ba/ token and the discrepant SFI token revealed a significant difference $t(13) = 3.562$, $p < .001$. This indicates that the static visual information significantly affected subjects' auditory judgments. A t-test also revealed a significant difference between the percentage correct for the discrepant FI tokens and audio /ba/ tokens, $t(13) = 21.57$, $p < .001$. This indicates that there was again a significant effect of fully-illuminated dynamic visual information on auditory judgments. There was also a significant difference between the pooled percentage of correct responses for the discrepant SFI tokens versus the discrepant FI tokens, $t(13) = 7.284$, $p < .001$, indicating that the FI condition was more effective at influencing subjects' responses than the SFI. One possible reason for this difference is that the static frame chosen in the SFI condition was not optimal for conveying the critical place information. This, and other interpretations will be discussed in detail below. A final comparison was made between the discrepant PL results from Experiment 1 versus the discrepant SFI results from this Experiment. This result was non-significant, $t(28) = 1.705$, $p > .05$, indicating that the static displays and the point light displays were equally successful at influencing subjects' heard judgments.

Thus, the results from Experiment 3 suggest that static images of an articulating face can influence heard speech. This finding could be taken as evidence that static facial information is treated as real visual speech. Further, when considered with the results of Experiments 1 and 2, it would seem that observers can use either static featural or kinematic visual information for speech integration. Before drawing this general conclusion however, we need to consider one more interpretation of our results.

The instructions for all three experiments made explicit that subjects were to watch the video but perform judgments based on what they heard. However, informal observations of the subjects during the SFI condition of Experiment 3 suggested that subjects might not have fully understood the task. Upon seeing the SFI tokens, many subjects voiced some degree of confusion and asked that the instructions be repeated. Based on this observation, the possibility that our results are due to misunderstanding or experimental demand characteristics needs to be examined. To this end, further tests were performed to help ensure that the observed effects were based on true visual influences.

Previous research suggests that when a true visual influence occurs, it is robust to the extent that it is not hampered by prior experience with related stimuli. For example, we have shown that the visual influences observed with a dynamic full-facial /va/-/ba/ are not reduced by first experiencing video- or audio-alone /va/ and /ba/ trials (Rosenblum and Saldaña, 1992; Saldaña and Rosenblum, 1994). In Experiments 1-3, subjects always received the critical PL or SFI stimuli first. This design was implemented in order to test subjects' naïve impressions of the stimuli. However, this design strategy could have induced erroneous judgments in these first conditions based on subjects' confusion with the task. In order to preclude this possibility, Experiment 4 tests for visual influences with both the PL and SFI stimuli presented after the dynamic fully illuminated stimuli. We reasoned that if *true* visual influences were observed in Experiments 1-3 for both PL and SFI stimuli, then these influences should not be diminished by the change in stimulus ordering. If, on the other hand, the critical results of these experiments were based on subject confusion or experimental demand characteristics, then the change in condition ordering might dilute the effects in Experiment 4.

EXPERIMENT 4

Method

Subjects

Twenty two undergraduates at the University of California, Riverside served as subjects for

partial fulfillment of a class requirement. All reported normal or corrected vision, and were native speakers of English. None of the subjects in Experiment 4 participated in Experiment 1, 2, or 3.

Stimuli

The stimuli for Experiment 4 consisted of both the presentation tape from Experiment 1 (and 2) and the presentation tape for Experiment 3.

Procedure

Subjects were run in groups of two or three. Eleven of the subjects were shown the first presentation tape (point light and fully illuminated). However, the order of presentation was reversed so that subjects saw the PL conditions at the end of the experiment. Subjects were presented with the stimulus blocks in the following order: 1) FI audiovisual condition; 2) FI video-alone condition; 3) audio-alone condition; 4) PL audiovisual condition; and 5) PL video condition. Prior to the presentation of the PL condition, subjects were made aware of the video recording technique. The experimenter also pointed out the configuration of the points of light on the face.

The other eleven subjects were presented with the second presentation tape (SFI and FI). These subjects were presented with the stimulus blocks in the following order: 1) FI audiovisual condition; 2) FI video-alone condition; 3) audio-alone condition; 4) SFI audiovisual condition; and 5) SFI video condition.

All subjects were told to watch as well as listen to the syllables but to respond by writing down only what they heard.

Results and Discussion

The pooled percentage of correct responses for each token type is presented in Table 4. Again, all alternative responses for Experiment 4 were either /va/ or /fa/ as they were in the first three experiments. Surveying the PL and SFI audiovisual means, it is clear that the stimulus order manipulation had a differential effect on the two types of stimuli. First, for the PL stimuli, a large visual influence was induced with subjects in Experiment 4 reportedly hearing a visually-influenced syllable 87% of the time. As in Experiments 1 and 2, the PL stimuli did significantly influence the heard percept as compared to the audio-alone /ba/ condition $t(10)= 11.017$, $p<.001$. Thus, the ordering manipulation did not dilute the visual influence. This adds support to the interpretation that the PL stimuli induces a true McGurk-type effect. In fact, for Experiment 4, the PL visual influence was as strong as the visual influence of the FI stimuli, $t(10)= -.191$, $p>.05$. This indicates that first seeing the FI stimuli allowed the PL stimuli to be very effective in its visual influence. In fact, a test between the point-light visual influences found in Experiments 2 and 4 revealed that first seeing the FI stimuli significantly enhanced the PL influence, $t(25) = 5.19$, $p<.05$. Possible reasons for this enhancement of effect strength are discussed below.

Table 4. Pooled percentage of correct responses for each token type for Experiment 4.

Token		Percentage Correct (based on audio)	
		Point Light Face	Dynamic Fully Illum.
Audiovisual	/ba-ba/	93	100
	/va-va/	100	100
	/ba-va/	13	12
Audio-Alone	/ba/	93	92
	/va/	100	99
Video-Alone	/ba/	100	100
	/va/	100	100
		Static	Dynamic Fully Illum.
Audiovisual	/ba-ba/	100	99
	/va-va/	100	99
	/ba-va/	86	05
Audio-Alone	/ba/	91	96
	/va/	100	100

Video-Alone	/ba/	100	100
	/va/	100	100

The manipulation of stimulus presentation order affected the SFI condition differently. The visual influence of the SFI stimuli in Experiment 4 dropped to 14% /va/ responses. In fact, a t-test revealed that there was no significant visual influence of the SFI stimuli as tested relative to the audio-alone /ba/, $t(10)=1.419$, $p>.05$. Thus, judging the (dynamic) FI condition first nullified the visual influence of the SFI stimuli. (Also, a test between the visual influences of the SFI stimuli in Experiments 3 and 4 revealed a decrease in the effect which nearly reached statistical significance, $t(10)=4.704$, $p=.055$.) This suggests that the observed influence of the SFI stimuli is quite fragile. As mentioned, past research with McGurk-type stimuli demonstrates that true visual influences are impervious to changes in presentation ordering. It could be the case then, that the SFI stimuli does not induce a true visual influence and that the results of Experiment 3 were based simply on subject misunderstanding or demand characteristics. As in Experiment 3, the discrepant SFI condition was significantly less effective in its visual influence than the discrepant FI condition $t(10)=10.488$, $p<.05$.

Not surprisingly, a final t-test revealed a significant difference between the discrepant SFI and PL conditions $t(10)=7.86$, $p<.05$. This finding indicates that the PL images induced a significantly stronger visual influence than the SFI images.

Overall, the results of Experiment 4 suggest that the visual influences for the point-light and static fully-illuminated stimuli were of a different nature. While the initial presentation of the (dynamic) fully-illuminated stimuli nullified the influence of the static stimuli, it did just the opposite for the point-light presentations. As mentioned, McGurk effects with natural speech are not diminished by previous presentation of related auditory and visual stimuli. Following from this, we conclude that while the PL stimuli were perceptually-integrated to influence the heard speech, the influence of SFI stimuli reported in Experiment 3 were not truly perceptual in nature. It is likely that the observed influences of Experiment 3 subjects were post-perceptual—possibly based on conscious decision strategies. These results provide evidence that although the kinematic point-light displays were treated as visual speech so as to integrate with auditory speech, the static images were not.

However, other static facial stimuli might lend themselves to a more robust visual influence. The decision to select static frames from a point just prior to release of the dental-labial and bilabial constriction was based primarily on intuition. To us and to pilot subjects, these frames optimally distinguished the /b/ and /v/ segments. Although these frames were distinguishable, it could be that other static frames would have induced a more robust visual influence. Further research using frames of other static articulatory positions could be conducted to explore this question.

Finally, mention should be made of the fact that prior presentation of the fully-illuminated stimuli *enhanced* the visual influence induced by the point-light stimuli. Whereas the point-light stimuli induced a visual influence 55% and 56% of the time in Experiment 1 and 2 respectively, initial presentation of the fully-illuminated stimuli increased the point-light visual influence to 87% in Experiment 4. We do not know why this enhancement occurred. However, similar results were observed by McGrath (1985) who found that the strength of the visual influence of computer-animated facial stimuli increased for subjects who had previously seen analogous natural face stimuli. Using the animated schematic faces described earlier, McGrath tested four types of discrepant audiovisual tokens: audio /pa/-video /ga/; audio /ba/-video /ga/; audio /ka/-video /ba/; and audio /ga/-video /ba/. He found that for subjects who saw a similar *natural* set of audiovisual stimuli first, the animated displays induced a 44% visual influence. This contrasts with a 10% visual influence found for subjects presented the animated displays first. In explaining this natural display 'priming', Summerfield, McGrath and their colleagues (Summerfield, et al, 1989) highlight the point that a number of the tested audiovisual pairings induce the relatively unusual 'blend' percepts (e.g., audio /ga/-video /ba/ induces a perceived 'bga'). They conjecture that perhaps the greater clarity of the natural visual displays more effectively establishes the blend

response categories. Once these categories are established, the subsequent animated displays can then induce a more pronounced visual influence.

This explanation cannot account for our results however, because establishing unusual response categories was not needed for our audio /ba/-video /va/ stimuli. Possibly, seeing the natural visual stimuli first helped subjects attend to the salient kinematic information in the point-light stimuli, allowing it to induce a more compelling visual influence. Perhaps then, making use of kinematic facial information involves some degree of perceptual learning. This is not to say that learning to perceive the appropriate kinematic primitives is something special to point-light stimuli: it might, in fact, be a component of the process involved in improving general lip-reading ability. Also, perceptual learning need not be antithetical to mandatory perception. In supporting attunement to relevant informational properties, perceptual learning might help an observer detect information which was not detected previously. Once the relevant information is detected, its part in event recovery might occur in a mandatory fashion. This interpretation of perceptual learning is compatible with the Gibsonian or Ecological approach to perception (Gibson, 1979; Michaels and Carello, 1981).

Nonetheless, it is probable that the observed priming from our natural stimuli is not a result of its informing subjects that the point-light stimuli were of a face with points on particular features. All of this information was made explicit to subjects in Experiment 2 with no resultant increase in effect strength.

General Discussion

Taken together, these experiments suggest that point-light images are treated as real visual speech to the degree that they can integrate with and influence auditory speech. In all experiments testing the point-light stimuli, a significant visual influence was observed Furthermore, the results of Experiment 4 suggest that this influence was not based on a post-perceptual decision strategy. This contrasts with static facial images for which any influence seemed due to post-perceptual decision biases. This last result also suggests that the observed point-light influence was not based on a process that made reference to static facial features. When static features were provided to subjects (Experiments 3 and 4), no *real* visual influence was observed. It is likely then that the salient aspect of the point-light images is not any superficial property of 'faceness'. Rather, observers might simply extract the relevant articulatory kinematic information to integrate with and influence the heard speech. This notion is also supported by the results of the three subjects in Experiment 1 who displayed a visual influence (68%) without ever recognizing the point-light images as a face.

Kinematic primitives for visual speech perception

Our research has shown that point-light facial images can convey visible speech information (Johnson, Rosenblum, and Saldaña, 1994; Rosenblum, Johnson, and Saldaña, in preparation), and that these images provide enough information to integrate with auditory speech. Taken together, these findings indicate that the kinematic properties of visible speech are salient and should be considered when determining visual speech features.

The crucial role of kinematic information for visual speech has support from recent evidence in neuropsychology. Campbell (1992) has reported on the behavior of a brain-lesion patient with severe agnosia that extends to recognizing static facial images. Further, this patient cannot match static images of speech articulations to speech sounds, a task that is easily performed by normal subjects. However, when shown dynamic articulations, this patient lipreads normally. Campbell (1992) concludes that "Among the visual processes that need to be intact to support effective lipreading are those that allow the perception of events through seen movement." (p.43).

The proposal that the time-varying characteristics of visual speech are salient echoes recent developments in the *auditory* speech literature. For auditory speech, primitives have been traditionally thought to consist of discrete featural cues such as steady-state formant properties (thought to cue vowel color and consonantal voicing characteristics), formant transitions (which can cue manner and place of consonantal articulation), onset spectral composition such as noise

bursts (which can cue place of articulation), and low frequency murmurs (which can signal nasal manner) (e.g., Liberman and Studdert-Kennedy, 1978). Discrete featural cues are most often the assumed inputs whether the speech perception process is thought to be based on analysis-by-synthesis (Stevens and Halle, 1964), cue-weighting techniques (e.g., Massaro, 1987), or network interactions (e.g., McClelland and Elman, 1986).

However, an alternative perspective proffers that the acoustic primitives are better construed as more temporally-extended (e.g., Fowler, 1987). Evidence for this perspective involves the observation that signals that do not involve the traditional cues of formants, transitions, and noise bursts can still be recognized as speech (Remez, Rubin, Pisoni, and Carrell, 1981). To generate these signals, three or four sine-waves are synthesized to track the pitch and amplitude of the center formant frequencies of an utterance. To listeners who are not primed to hear speech, these sine-waves sound like simple computer bleeps and whistles. To those primed to hear speech however, the signals can be understood as speech to the extent that many listeners can transcribe sentences. Moreover, similar 'sine-wave speech' segments and syllables can induce perceptual effects (e.g., vowel normalization, consonantal context effects) characteristic of natural speech stimuli (Remez, Rubin, Nygaard, and Howell, 1987; Williams, Verbrugge, and Studdert-Kennedy, 1983). Clearly, the sinewave speech phenomena pose a significant challenge to models that assume discrete spectral cues as acoustic primitives.

There is further evidence that salient segmental information appears in the time-varying aspects of the speech signal. Fowler (1987) cites a number of studies that demonstrate that listeners gain more segmental information from the dynamic, *coarticulated* parts of the signal than from the discrete, segment 'nuclei'. A number of studies have tested the relative saliency of consonantal release spectra and coarticulatory-dependent transitions and have shown that this latter, time-varying cue more often determines stop consonant identifications (Blumstein, Isaacs, and Mertus, 1982; Walley and Carrell, 1983). Regarding vowels, Strange and her colleagues have shown that in CVC syllables, much of the vowel's center can be deleted without impairing its identification (see Strange, 1987 for a review). Thus, there is substantial segmental information in the dynamic, coarticulated margins surrounding segments. In all, the dynamical view of auditory speech information has evolved into an important alternative to the discrete featural perspective.

Clearly, our point-light speech findings are similar in suggesting the importance of time-varying features for speech. In fact, the point-light stimuli are analogous to auditory sinewave speech. Both types of stimuli demonstrate that the time-varying components of the speech signal, when dissociated from static features, can be informative.[1] Whether visual speech primitives are better construed as kinematic rather than static—as is currently being claimed for auditory speech—is an important question for future research.

The form of integrated audiovisual information

The present findings also have implications for issues of audiovisual speech integration. First, these results provide support for an informationally-encapsulated integration process. It has been argued that speech perception occurs through a behaviorally and anatomically distinct module (e.g., Liberman and Mattingly, 1985). This module is considered to be informationally-encapsulated in disallowing information available from outside the module to seep in and influence its mandatory operations (Fodor, 1983). Like the original McGurk effect, this is the case for our point-light effects. Subjects in Experiment 2 were aware of the nature of the point-light stimuli yet they displayed no less of a visual influence than the naïve subjects of Experiment 1. Additional evidence for an informationally-encapsulated integration process is provided by the three subjects of Experiment 1. For these subjects, speech integration occurred without *any* reference to extra-modular (conscious) knowledge of the nature of the visual information.

Next, our finding that kinematic primitives can act as visual input to the speech integration process has implications for understanding the form of the integrated information. One question central to audiovisual speech perception is in what metric or metrics are the visual and auditory streams represented at the point of integration. There is a good deal of evidence that categorization occurs after integration (e.g., Foster, 1982; Green and Miller, 1985). It is likely that the speech

system takes in all auditorily and visually-specified linguistic dimensions, integrates them, and then performs phonetic categorization. Summerfield (1987) has proposed that the integration metric might take an articulatorily-based, modality-neutral form (see also Studdert-Kennedy, 1989). This information could include time-varying kinematic patterns that are instantiated in a number of modalities which serve to specify articulatory dynamics (cf., Liberman and Mattingly, 1985; Fowler and Rosenblum, 1991). Thus, the mechanism ultimately responsible for integration would be sensitive to modality-neutral kinematic patterns and would provide for the recognition of articulatory objects of perception. In specifying the time-varying aspects of speech production, the articulatory dynamical metric is more appropriate for vowel productions which rarely achieve canonical positions. In being based on a modality-neutral form of information, integration could proceed without the extra step of translating the auditory and visual information into an integratable form.

Our evidence that extracted kinematic visual information can act as input for speech integration provides support for Summerfield's hypothesis that the form of integrated information is kinematic. A kinematic metric can more easily handle the important time-varying dimensions of visible articulation. Further, if the *auditory* primitives can also be considered time-varying/dynamical (see above), then a modality-neutral metric becomes tenable. Although many models of integration have only discussed discrete featural auditory primitives (e.g., Massaro and Cohen, 1990; Braida, 1991), the advances made in understanding time-varying primitives of auditory-alone speech can be applied to the audiovisual domain. One way to determine whether auditory primitives for integration can be time-varying/dynamical is to test sinewave speech in a McGurk-type paradigm. If sinewave speech can be shown to integrate with a visual image in a robust way, then support for time-varying/dynamic auditory primitives for integration would be gained. If it turns out that both auditory and visual primitives for integration can be kinematic, then Summerfield's modality-neutral metric would be plausible.

It should be mentioned that additional evidence for a dynamic metric of audiovisual integration has recently been provided in the results of Green and Gerdeman (in press). They found that audiovisual integration is sensitive to cross-modal discrepancies in time-varying coarticulatory information between an initial consonant and following vowel. They take this finding as evidence that information about dynamic aspects of articulation must be described in the underlying audiovisual representations for integration.

Lipreading and the point-light technique

Further mention should be made of the benefits of the point-light technique for lipreading research. The technique affords isolation of salient kinematic features as well as for efficient kinematic analyses of visible speech movements. Along these lines, the point-light technique has been used successfully to uncover salient kinematic information for various nonspeech events. For example, Bingham (1987) conducted kinematic analyses on human weight-curling point-light events and found that dynamic differences (in mass) were evident in the peak and average flexion velocity of the movements as well as duration of flexion. Significantly, Bingham also found that these kinematic variables best predicted judgments of perceived heaviness. Bingham, Rosenblum, and Schmidt (in press) found that observers could distinguish a naturally falling/bouncing spring from a hand-guided falling/bouncing spring through point-light specification. In performing kinematic analyses, they found differences distinguishing these events in the overall form of the phase trajectories. Specifically, the hand-guided spring displayed a temporal asymmetry between falling and bounce phases not evident for the naturally falling spring. Applying the point-light technique to lipreading research will permit similar determination of salient kinematic features. Kinematic analyses of visible speech are currently being conducted in our laboratory.

The point-light technique has also been used to study the salient information for reading sign-language. Poizner, Bellugi, and Lutes-Driscoll (1981) applied point-lights to the head, shoulders, elbows, wrists, and index finger-tips of a deaf signer. They found that from these displays, sign-readers can recover information about lexical and inflectional movements of American Sign Language. In a first attempt to uncover the salient information in the displays, Poizner and his colleagues systematically eliminated each light, one at a time, and tested word identification. They

found that only the points on the index fingertips were necessary for sign identification and that proficiency was directly related to how distal on the body the missing point-light was located. In a more elaborate demonstration of point-light sign specification, Tartter and Knowlton (1981) devised a pair of black gloves with point-lights on all finger tips, at the each of the major finger joints, and around the wrist. They found that two signers could carry-on lengthy discussions by watching each other's point-light signs on contrast-adjusted monitors.

Finally, the benefits of the point-light technique for telecommunication systems has also been considered. In their paper, Tartter and Knowlton (1981) suggest that, unlike fully-illuminated hand movements, the highly-reduced point-light images afford transmission through the low bandwidth of a telephone line. If so, the deaf could soon make use of existing telecommunications equipment to converse using the wealth of expression sign-language provides. Pearson (1981) and Massaro (1987) have both suggested that similar benefits might exist for point-light facial images. If these images are found to convey rich linguistic information, then point-light facial stimuli could be used to allow the deaf to telecommunicate more effectively with those who do not know sign language.

References

Aronson, E. & Rosenbloom, S. (1971). Space perception in early infancy: Perception within a common auditory-visual space. *Science, 172,* 1161-1163.

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

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

Bingham, G.P. (1987). Scaling and kinematic form: Further investigations on the visual perception of lifted weight. *Journal of Experimental Psychology: Human Perception and Performance, 13(2),* 155-177.

Bingham, G.P., Rosenblum, L.D., & Schmidt, R.C. (in press). Dynamics and the orientation of kinematic forms in visual event recognition. In press at *Journal of Experimental Psychology: Human Perception and Performance.*

Blumstein, S., Isaacs, E., & Mertus, J. (1982). The role of gross spectral shape as a perceptual cue to place of articulation in initial stop consonants. *Journal of the Acoustical Society of America, 72,* 43-50.

Braida, L.D. (1991). Crossmodal integration in the identification of consonant segments. *The Quarterly Journal of Experimental Psychology. 43,* 647-677.

Brooke, N.M., & Summerfield, A.Q. (1983). Analysis, synthesis and perception of visible articulatory movements. *Journal of Phonetics,* 11, 63-76.

Campbell, R. (1986). The lateralisation of lipread sounds: A first look. *Brain and Cognition, 5,* 1-21.

Campbell, R. (1992). The neuropsychology of lipreading. *Philisophical Transcriptions of the Royal Society of London, 335,* 39-45.

Cooper, F.S., Delattre, P.C., Liberman, A.L., & Gerstman, L.J. (1952). SOme experiments on the perception of synthetic speech sounds. *Journal of the Acoustical Society of America, 24,* 597-606.

Delattre, P., Liberman, A.M., & Cooper, F.S. (1955). Acoustic loci and transitional cues for consonants, *Journal of the Acoustical Society of America, 27,* 769-773.

Delattre, P., Liberman, A.M., Cooper, F.S., and Gerstman, L.J. (1952). An experimental study of the acoustic determinants of vowel color: Observations on one- and two-formant vowles synthesized from spectrographic patterns, *Word, 8,* 195-210.

Dodd, B. (1979). Lipreading in infants: Attention to speech presented in and out of synchrony. *Cognitive Psychology, 11,* 478-484.

Ekman, P., Friesen, W.V., & Ellsworth, P. (1982). *Emotion in the human face*. New York: Pergamon Press.

Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech and Hearing Research, 12,* 423-425.

Fodor, J. (1983) *The Modularity of Mind.* Cambridge, MA: MIT Press.

Foster, G.A. (1982). The integration of audio-visual speech stimuli as a function of temporal desynchronization. *Proceedings of the Autumn Conference of the Institute of Acoustics,* H3.1-H3.5.

Fowler, C.A. & Rosenblum, L.D. (1991). Perception of the phonetic gesture. In I.G. Mattingly and M. Studdert-Kennedy (eds.) *Modularity and the Motor Theory.* Lawrence Earlbaum Assoc., N.J.

Fowler, C.A. (1987). Perceivers as realists, talkers too: Commentary on papers by Strange, Diehl, et al., Rakerd and Verbrugge. *Journal of Memory and Language, 26,* 547-587.

Gibson, J. J. (1979). *The ecological approach to visual perception.* Boston, Houghton-Mifflin.

Green, K.P. & Gerdman, A. (in press). Phonetic incongruity and the McGurk effect. In press at *Journal of Experimental Psychology: Human Perception and Performance.*

Green, K.P. & Miller, J.L. (1985). On the role of visual rate information in phonetic perception. *Perception and Psychophysics, 38,* 269-276.

Harris, K.S. (1958). Cues for the discrimination of American English fricatives in spoken syllables. *Language and Speech, 1,* 1-7.

Jackson, P.L. (1988). Theoretical minimal unit for visual speech perception: Visemes and coarticulation. *The Volta Review, 90,* 99-115.

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

Johnson, J.A., Rosenblum, L.D., & Saldaña, H.M. (1994). The contribution of a reduced visual image to speech perception in noise. *Journal of the Acoustical Society of America, 95,* 3009.

Kozlowski, L.T., & Cutting, J. E. (1977). Recognizing the sex of a walker from a dynamic point light display. *Perception and Psychophysics, 21,* 575-580.

Kuhl, P.K. & Meltzoff, A.N. (1984). The intermodal representation of speech in infants. *Infant Behavior and Development, 7,* 361-381.

Kuhl, P.K., & Meltzoff, A.N. (1982). The bimodal development of speech in infancy. *Science, 218,* 1138-1141.

Liberman, A., & Studdert-Kennedy, M. (1978). Phonetic Perception. In R.Held, H. Leibowitz, & H.L. Teuber (eds.) *Handbook of Sensory Physiology, Vol. VIII: Perception.* New York: Springer-Verlag.

Liberman, A.M. & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21,* 1-36.

MacLeod, A. and Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology, 21,* 131-141.

MacLeod, A. and Summerfield, Q. (1990). A procedure for measuring auditory and audio-visual speech reception thresholds for sentences in noise: Rationale, evaluation, and recommendations for use. *British Journal of Audiology, 24,* 29-43.

Massaro, D.W. & Cohen, M.M. (1990). Perception of synthesized audible and visible speech. *Psychological Science, 1,* 55-63.

Massaro, D.W. (1987). *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry.* Hillsdale, New Jersey: Lawrence Earlbaum Assoc.

McClelland, J.L. & Elman, J.L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18,* 1-86.

McGrath, M. & Summerfield, A.Q. (1985). Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *Journal of the Acoustical Society of America, 77,* 678-685.

McGrath, M. (1985). *An examination of cues for visual and audio-visual speech perception using natural and computer-generated faces. Unpublished doctoral dissertation,* University of Nottingham, Nottingham, U.K.

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

Michaels, C.F., & Carello, C. (1981). *Direct perception*. Englewood Cliffs, NJ: Prentice-Hall.

Middleweerd, M. J. and Plomp, R. (1987). The effect of speechreading on the speech reception threshold of sentences in noise. *Journal of the Acoustical Society of America, 82,* 2145-2146.

Mills, A.E. (1987). The development of phonology in the blind child (pp. 145-162). In B. Dodd and R. Campbell Eds, Hearing by eye: The psychology of lip reading. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Montgomery, A.A. & Jackson, P.L. (1983). Physical characteristics of the lips underlying vowel lipreading performance. *Journal of the Acoustical Society of America, 73,* 2134-2144.

Pearson, D. (1981) Visual communication systems for the deaf. *IEEE Transactions on Communications, COM-29,* 1986-1992.

Petajan, E.D. (1984). Automatic lipreading to enhance speech recognition. *Proceedings of the IEEE Communications Society,* November 26-29, Atlanta Georgia.

Poizner, H., Bellugi, U., & Lutes-Driscoll, V. (1981). Perception of American Sign Language in dynamic point-light displays. *Journal of Experimental Psychology: Human Perception and Performance, 7,* 430-440.

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd and R. Campbell Eds), Hearing by eye: The psychology of lip reading. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

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

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 and Performance, 13,* 40-61.

Repp B.H., Manuel, S. Y., Liberman A.M., & Studdert-Kennedy M. (1983). Exploring the "McGurk effect". *Paper Presented at the 24th annual meeting of the Psychonomic Society in San Diego.*

Rosenblum, L. D., Johnson, J. A., and Saldaña. H. M. (in preparation). Determining the kinematic features for visual speech perception.

Rosenblum, L.D., and Saldaña, H.M. (1992). Discrimination tests of visually-influenced syllables. *Perception and Psychophysics. 52 (4),* 461-473.

Runeson, S. & Frykholm, G. (1981). Visual perception of lifted weight. *Journal of Experimental Psychology: Human Perception and Performance, 7,* 733-740.

Runeson, S. (1977). On the visual perception of dynamic events. *Acta Universitatis Upsaliensis: Studia Psychologica Upsaliensia.* (Series No. 9).

Runeson, S. (1993). Personal Communication, November 11.

Saldaña, H.M. and Rosenblum, L.D. (1993). Visual influences on auditory pluck and bow judgments. *Perception and Psychophysics. 54 (3),* 406-416.

Saldaña, H.M. and Rosenblum, L.D. (1994). Selective adaptation in speech perception using a compelling audiovisual adaptor. *Journal of the Acoustical Society of America, 95 (6)* 3658-3661.

Scheinberg, J.S. (1980). Analysis of speechreading cues using an interleaved technique. *Journal of Communication Disorders, 13,* 489-492.

Sekiyama, K. & Tohkura, Y. (1991). McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility, *Journal of the Acoustical Society of America, 90,* 1797-1805.

Stevens, K.N., & Halle, M. (1964). Remarks on analysis by synthesis and distinctive features. In W. Wathen-Dunn (ed.) *Proceedings of the AFCRL Symposium on Models for the Perception of Speech and Visual Form.* Cambridge, Mass: MIT Press.

Strange, W. (1987). Information for vowels in formant transitions. *Journal of Memory and Language, 26,* 550-557.

Studdert-Kennedy, M. (1989). Reading gestures by light and sound. *Handbook of Research on Face Processing,* Young and Ellis (eds.) Elsevier pp. 217-222.

Sumby, W.H. & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26,* 212-215.

Summerfield, Q. & McGrath, M. (1984). Detection and resolution of audio-visual incompatibility in the perception of vowels. *Quarterly Journal of Experimental Psychology, 36A,* 51-74.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception (pp. 3-51). In B. Dodd and R. Campbell Eds), Hearing by eye: The psychology of lip reading. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Summerfield, Q., MacLeod, P., McGrath, M., & Brooke, N.M. (1989). Lips, teeth, and the benefits of lipreading. In *Handbook of Research on Face Processing*, Young and Ellis (eds.) Elsevier, 1989, pp. 223-233.

Summerfield, Q.A. (1979). Use of visual information for phonetic perception. *Phonetica, 36,* 314-331.

Tartter, V.C. & Knowlton, K.C. (1981). Perception of sign language from an array of 27 moving spots. *Nature, 289,* 676-678.

Walden, B.E., Prosek, R.A., Montgomery, A.A., Scherr , C.K. & Jones, C.J., (1977). Effects of training on the visual recognition of consonants. Journal of Speech and Hearing Research, 20, 130-145.

Walley, A. & Carrell, T. (1983). Onset spectra and formant transitions in the adult's and child's perception of place of articulation in stop consonants. *Journal of the Acoustical Society of America, 73,* 1011-1022.

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

Footnotes

1 One difference between point-light and sinewave speech should be noted, however. Sinewave signals display characteristic speech effects only after listeners are told that the stimuli can be heard as speech (Remez, Rubin, Nygaard, and Howell, 1987; Williams, Verbrugge, and Studdert-Kennedy, 1983). In contrast, the point-light stimuli, in integrating with auditory speech, exhibited a visual speech effect without being primed to be seen as speech (e.g., Experiment 1). Further, Experiment 2 indicates that there was no difference in effect strength between subject who were and were not informed that the stimuli were speech. Finally, the results of the three aforementioned subjects of Experiment 1 suggest that the point-light images acted as visual speech even when *unrecognized* as speech. (However, like sinewave speech, the point-light stimuli do benefit from some speech information priming. In Experiment 4, seeing the fully-illuminated stimuli first did significantly strengthen the subsequent visual influence of the point-light stimuli.)

Author Notes

We gratefully acknowledge the assistance of Maria Aguilar, Theresa Osinga, and Jocelyn Corominas and the helpful comments of Ruth Campbell, Carol Fowler, Dominic Massaro, and Arthur Samuel.

This research was supported by NSF Grant DBS-9212225 awarded to Lawrence D. Rosenblum as well as an intramural grant from the University of California.

Requests for reprints should be sent to Lawrence D. Rosenblum, Department of Psychology, University of California, Riverside, Riverside, California, 92521.