

RESEARCH ON SPOKEN LANGUAGE PROCESSING

Progress Report No. 26 (2003-2004)

Indiana University

**Some Effects of Feedback on the Perception of Point-Light and Full-Face
Visual Displays of Speech: A Preliminary Report¹**

Stephen J. Winters and David B. Pisoni

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

¹ This research was supported by the NIH-NIDCD Research Grant DC-00111 and NIH-NIDCD Training Grant DC-00012 to Indiana University. We would like to thank Jeff Reynolds for his work on the point-light portion of this study and Luis Hernandez for conceptual inspiration and technical assistance.

Some Effects of Feedback on the Perception of Point-Light and Full-Face Visual Displays of Speech: A Preliminary Report

Abstract. This study investigated the effects of feedback on the perception of words from point-light and fully-illuminated displays of speech. Participants attempted to identify words in these displays and were later given feedback about the identity of the words they had just seen. Feedback was presented to the participants in three forms: audio-visual, audio-only, and orthographic representations of the stimulus word. A control group of participants also received no feedback on the stimuli. It was expected that dynamic feedback—as found in the audio or audio-visual signals—would improve participant performance on the perceptual task more than static, orthographic feedback, or no feedback at all, due to the extra, event-based information about the original stimuli inherent in the dynamic representations. In general, we found that participants improved their ability to perform the perceptual task over the course of the experiment; however, their level of improvement did not depend on the type of feedback they received. This finding suggests that participants may not have improved their visual-only perception skills through attending to the feedback information but rather by simply practicing the experimental task and relying on what they already knew about the lawful acoustic consequences of articulatory gestures. Participants also had more success identifying the full-face stimuli than the point-light stimuli. However, they also exhibited different patterns of misidentification for some of the point-light and full-face stimuli, suggesting that the point-light stimuli were not merely impoverished representations of the full-face displays. Instead, the two representations of speech must be, to some extent, perceptually independent of one another.

Introduction

This study investigated the extent to which observers can extract phonetic information from visual-only point-light and fully-illuminated displays of speech. Point-light displays (PLDs) are animated sequences of illuminated dot patterns. They are produced by attaching luminescent dots to various points on a person's body and then filming that person while they perform certain motions under a black light (Johansson, 1973). Figure 1 shows an example sequence of frames from a point-light movie; the person filmed in this particular point-light movie executed a placekick.

Point-light displays are valuable research tools because observers can perceive meaningful motion in them, even though they generally do not perceive anything more than random dot patterns when presented with any of the individual point-light frames in isolation. This suggests that perceivers are able to extract meaningful information from dynamic, time-varying properties in the sequences of the point-light patterns which do not exist in any of the individual, static point-light patterns per se. These findings are interesting because they might reflect aspects of event perception under normal conditions. For instance—does the visual perception of events ever rely on the recognition of individual static patterns? Or is event perception always based solely on dynamic, time-varying information in the visual array?

Several theorists have argued that the visual perception of speech involves the perception of meaningful articulatory events (e.g., Fowler & Rosenblum 1991). Lachs and Pisoni (2004) have shown that people can accurately perceive such events in visual-only, fully-illuminated displays of speech. Lachs and Pisoni (submitted) have also shown that people can perceive such articulatory events in PLDs of speech. Furthermore, Rosenblum, Johnson and Saldaña (1996) found that PLDs of speech provide an

intelligibility gain to the perception of speech in noise, similar to the gain found for fully-illuminated visual displays in Sumbly and Pollack (1954). Rosenblum and Saldaña (1996) also found that PLDs may induce a “McGurk Effect” on the perception of audio-visually mismatched stimuli, as was found for fully-illuminated visual displays in McGurk and McDonald (1976).

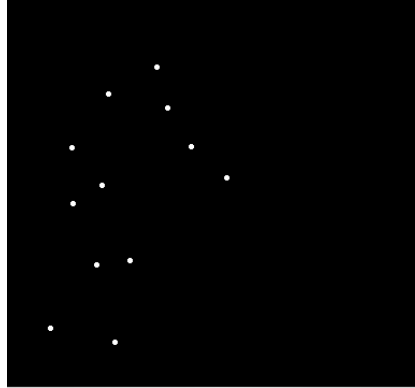


Figure 1a. Frame from a point-light display movie of a person executing a placekick.



Figure 1b. Frame from a point-light display movie of a person executing a placekick.

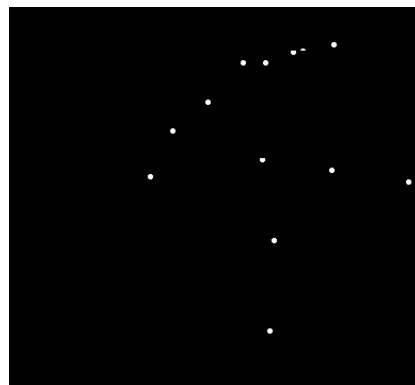


Figure 1c. Frame from a point-light display movie of a person executing a placekick (courtesy of <http://astro.temple.edu/~tshipley/mocap/dotMovie.html>).

Even though people can extract meaningful information from both visual-only point-light and fully-illuminated displays of speech, there are at least three reasons to believe that it might be more difficult for people to perceive visually presented words in PLDs than in fully-illuminated displays. Point-light displays are, first of all, visually impoverished versions of full-face displays; they contain only dynamic cues and no static visual information. Furthermore, PLDs do not necessarily contain all of the dynamic cues inherent in the full-face displays, since they only represent articulatory motions with a handful of illuminated dots attached to particular spots on a speaker's chin, lips, cheeks, teeth or tongue. The motions of such articulators in between these dots may provide meaningful information to perceivers which the PLDs cannot transmit. Finally, people might have difficulty interpreting PLDs of speech if only because they have little, if any, prior experience attempting to perceive speech in them.

Since PLDs present all of these perceptual challenges to human observers, it is remarkable that people can perceive speech in them at all. How do they perform this perceptual feat? One possible answer is that the PLDs preserve fundamental dynamic cues which are important to the visual perception of speech in full-face displays. Another possible answer is that they perceive meaningful events in the PLDs based on their knowledge of the physically lawful relationships which hold between articulation and acoustics. The perception of speech in PLDs might be direct, as it were, in the technical sense put forth by Fowler (1986). In this sense, speech perception does not need to be mediated by abstract knowledge of a symbolic system; instead, it is specified by visually apparent attributes in the speech signal itself. If this is the case, then it is unlikely that increased experience with the symbolic representations of speech events or gestures (as in, e.g., the written word) will improve a person's ability to perceive those events. By the same token, however, people may be able to perceive speech in representations with which they have had little if any experience if those representations happen to be based in physically lawful ways on the same articulatory events which constitute the original speech signal.

Fowler and Dekle (1991) made a compelling case for the "direct" nature of speech perception with an ingenious experiment that involved the perception of speech through the sense of touch. In this experiment, Fowler and Dekle briefly introduced their participants to the Tadoma method, which involves the placing of the fingers of one's hand on the lips, jaw and neck of a speaker in order to perceive what they are saying through (directly) feeling their articulators move (cf. Reed et al., 1989). Fowler and Dekle then presented their participants with McGurk-like stimuli which were mismatched between the auditory and tactile domains. The participants may have heard the syllable "ga," for instance, while they felt the syllable "ba." Fowler and Dekle also presented their participants (who were Ivy League undergraduates, with extensive reading experience) with stimuli that were similarly mismatched in audio and orthographic representations. They asked their participants to report what they had heard and what they had read/felt independently. Interestingly, Fowler and Dekle found that the participants' reports of what they had "heard" were not strictly independent of what they had felt in the mismatched-stimuli conditions. Participants who had heard "ga" but felt "ba," for instance, were more likely to report that they had heard "ba" than when they had heard the same "ga" stimulus in conjunction with a felt "ga" syllable. Mismatched audio and orthographic syllables did not induce similar cross-modal perceptual biases; in those cases, the participants' reports of what they had heard were not biased one way or another by what they had read.

Fowler and Dekle's (1991) participants experienced such McGurk-like effects in the combined auditory and tactile stimuli, even though they had never perceived speech through the Tadoma method before. The authors suggested that it was possible to induce this kind of McGurk effect for their participants in this way because of the physically lawful relationships which held between the movements the participants could feel and the acoustic sound patterns they could hear. The participants could extract information from both tactile and auditory modalities about the underlying speech events because those

speech events structured the auditory and tactile modalities in a lawful way. Such lawful relationships between the two modalities are important to perception because they are not arbitrary, and are not learned by association. In contrast, the relationship between an auditory stimulus and its orthographic representation is arbitrary. Hence, Fowler and Dekle's participants experienced no McGurk-like effects for conflicting audio and orthographic stimuli, despite all of their experience associating the two kinds of representations with one another. Thus, Fowler and Dekle concluded that people perceive the events which produce the speech, through the lawfully structured consequences of those events in the acoustic, visual, or tactile domains. They do not perceive speech through reference to arbitrary relationships between sound and symbol which have been learned through experience.

In the present study we hypothesized that people might be able to perceive speech in PLDs—even without prior experience of them—because of the lawful relationships they have with the articulatory events that produced them. We investigated the effects that different kinds of feedback might have on participants' ability to improve their perception of speech in PLDs over the course of the perception experiment. Specifically, we reasoned that informing participants of which words they had seen being spoken under visual-only presentation might help improve their ability to perceive the words in subsequent visual-only stimuli. While we expected that providing such feedback information would help improve participants' perceptual performance in general, we also expected that some forms of feedback would be more helpful to participants if it came to them in the form of dynamic information that was lawfully based on the articulatory events they had seen in the visual-only stimuli. With feedback, participants might better learn how to perceive the visual-only stimuli if they actually heard what was being said in the stimuli tokens, rather than simply reading in orthographic form which words the speaker had said in the stimuli. We further hypothesized that auditory, event-based feedback would help the participants' perception of these events (whether they are represented in PLDs or fully-illuminated faces) because this kind of feedback would specify to the participants what the lawfully-based, acoustic consequences of those events are. Written words, on the other hand, would only inform participants of a static, abstract, idealized linguistic representation of those articulatory events, which—as Fowler and Dekle (1991) have shown—is unlikely to influence the direct perception of those events, no matter how much experience participants might have had with orthography.

The following experiment was designed to test this hypothesis about the nature of perceiving speech from PLDs by providing participants with different kinds of feedback on their performance in a visual-only speech perception experiment. The participants' primary task in this experiment was to watch a short video clip of a talker saying one English word at a time, in the form of a consonant-vowel-consonant (CVC) syllable. Half of the participants saw these videos in the form of PLDs, while the other half saw them as fully-illuminated displays of the face of the speaker; in neither case, however, did the participants hear what the speaker in the video was saying. Their task was to attempt to identify the word that the speaker had said, based only on what they could see in each video. After attempting to identify each word, the participants received feedback about their response. In some conditions of the experiment, they were then told what word had they had seen being spoken. Some participants received this information by seeing the video again, this time along with the original audio track; another group of participants just heard the audio track, while a third group of participants simply read an orthographic representation of the stimulus word on a computer screen. A fourth group in the control condition was not told which words they had seen, but simply moved on to the next video stimulus in the sequence after attempting to identify each stimulus word.

We expected that the participants who saw the full-face displays would be much better at identifying the stimulus words than the participants who saw the PLDs. We also expected that the performance of all participants on this task would improve over the course of the experiment, but that the extent of this improvement would vary depending on the type of feedback the participants received. The

participants who received no feedback on their responses, for instance, were not expected to improve as much on the task as those who did receive feedback. Likewise, those groups who received static, symbolic feedback were not expected to improve as much as those who received dynamic, event-based feedback, in the form of either audio (A) or audio-visual (A+V) representations of the original visual stimulus. Such event-based feedback was expected to improve the participants' performance more than the symbolic feedback because it would not only provide the participants with information about the acoustic consequences of the dynamic events they were trying to perceive, but also reinstate the processing operations during the initial perception of the visual-only stimulus. Since audio-visual feedback would provide the participants with even more redundant information about the lawful relationships between articulation and acoustics than audio feedback alone, participants in the audio-visual feedback condition were also expected to show more improvement in the task than those in the audio feedback group. Finally, feedback of all kinds was expected to have more of an impact on the improvement of the participants' ability to identify words in the PLDs than in the full-face displays, since no participant had seen PLDs of speech prior to the experiment. Participants might thus depend more heavily on the feedback information they received during the experiment in order to interpret what they might be seeing in the PLDs. The participants' ability to identify words in the visual-only PLDs might also be closer to floor levels at the start of the experiment, which would leave considerably more room for improvement over the course of a short perception experiment.

Methods and Procedures

Participants. All of the participants in this study were college undergraduates at Indiana University in Bloomington, Indiana, taking part for either course credit or a small fee. The experiment used a between-subjects design; participants took part in only one of the four feedback conditions, for either of the stimulus types. The exact number of participants in each of the feedback conditions is given below in Table 1.

Full-face	N	Point-light	N
AVF	21	AVF	21
AF	22	AF	20
OF	20	OF	24
NF	20	NF	16

Table 1. Number of participants in each experimental group, by feedback condition and stimulus type.

Stimulus Materials. The individual full-face video stimuli in this experiment were first produced for the Hoosier Audio-visual Multi-talker Database (Sheffert, Lachs, & Hernandez, 1996; Lachs & Hernandez, 1998). The point-light stimuli were originally produced for use in Lachs (submitted). Both point-light and full-face stimuli consisted of short (between one and two seconds long) videos of a female, native speaker of English saying a single monosyllabic, consonant-vowel-consonant (CVC) word. Different speakers were filmed in the point-light and full-face videos. Stimuli of both types consisted of close-ups of the speaker's face, from just above the shoulders up. In the point-light videos, luminescent dots were attached to the speaker's face according to the pattern seen in Figure 2, which shows an isolated frame from one of these videos. The visible dots in Figure 2 were attached to the cheeks, lips, nose and chin of the speaker. Two dots were also placed on both the upper and lower teeth of the speaker, as well as another dot on the blade of the speaker's tongue. Figure 3 shows a corresponding example frame from one of the fully-illuminated videos.



Figure 2. Example frame from a point-light display stimulus video.



Figure 3. Example frame from a full-face display stimulus video.

Although originally recorded on videotape, all stimuli were digitized and transferred to Macintosh G3 computers for presentation in this experiment. All videos had a 640 x 480 aspect ratio and completely filled the entire monitor screen when they were presented to the participants. Each word in the stimulus set was a CVC, monosyllabic English word. There were 96 words in all; following Lachs (2002), half of these were “easy” words to identify in the sense that they were high-frequency items selected from low-density lexical neighborhoods, whereas the other half were “hard” in that they were low-frequency items selected from dense lexical neighborhoods. For example, easy words included “wife,” “road,” and “teeth,” while hard words included “hag,” “dame,” and “toot.”

Design and Procedure. Either a customized Psycoscope routine (ver. 1.2.5.PPC) or SuperCard stack (ver. 4.1.1) were used to present the stimuli videos to the participants. On each individual trial, these programs presented visual stimuli—without sound—to the participants and then prompted them to type into the computer what word they thought the speaker in the video had said. The words were presented in random order to each participant. After the participants typed in their response to each stimulus, the

customized programs then informed the participants in the feedback groups what word had actually been spoken in the video they had just seen. For the orthographic feedback group (OF), the programs presented the feedback to the listeners in the form of a written word, centered on the screen. For the audio feedback group (AF), the programs played the audio clip of the word to the listener over Beyer Dynamic DT-100 headphones, just as it was spoken in the original video stimulus. For the audio-visual feedback group (AVF), the programs played the original video stimulus to the participants again, together this time with the original audio track. The programs did not inform the participants in the control group—the no feedback group (NF)—which words had been spoken in the video stimuli; these participants just moved on to the next stimulus once they had finished typing in each of their responses.

All of the feedback conditions for the point-light stimuli, as well as the no feedback condition for the full-face stimuli, were run using a Psyscope program; it was not possible to use this program to run the orthographic feedback, audio feedback and audio-visual feedback conditions for the full-face stimuli because of computer memory limitations. Instead, these three conditions were run with the SuperCard stack. The Psyscope and SuperCard implementations of the experimental paradigm were essentially identical except for a few minor details, which were mostly aesthetic in nature. The most significant of these differences was that the SuperCard program presented the orthographic feedback to the participants for a full second (1000 milliseconds), whereas the Psyscope program only presented this information to the participants for 500 milliseconds. For both types of stimuli, audio and audio-visual feedback was only given to the participants for the inherent duration of the CVC word in the recording, which ranged between 1000 and 2000 milliseconds.

Prior to the experiment, the participants were informed that all of the words they would see being spoken were one-syllable, English words. They were also told that some words might be harder to identify than others; hence, they were encouraged to make guesses as to the identity of each word, even if they had no idea what the word was.

Data Analysis

Since all stimuli were of the form consonant-vowel-consonant, the observers' responses could be scored not only in terms of whether or not they had identified the whole word correctly, but also in terms of whether or not they correctly identified each segmental portion of the stimulus: the onset (initial consonant), the nucleus (the vowel) and the coda (final consonant). In order to make such phoneme-by-phoneme evaluations, all stimuli and all responses were first converted into phonetic transcriptions by matching them up with entries in the Carnegie Mellon pronouncing dictionary (version 0.6; <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>). Each entry in this dictionary consists of an English word listed in normal orthography along with a corresponding phonetic transcription encoded in an ASCII-based phonetic alphabet. All phonetic transcriptions in this dictionary offset each phoneme in the word with spaces and uniquely mark vowels with a number indicating their level of stress in the word. The phonetic transcription for the entry "hag," for instance, is /hh ae1 g/. For each stimulus word, the entire CMU dictionary was searched for a corresponding orthographic entry. A perl script was written that searched the CMU dictionary for orthographic entries corresponding to each of the stimulus items. Once such a match had been found, its corresponding phonetic transcription was segmented into an "onset," a "nucleus" and a "coda." Since all the words were of the form CVC, the vowel of each word was considered to be the "nucleus," the initial consonant the "onset," and the final consonant the "coda." In the case of "hag," for example, the /hh/ formed the onset, the /ae1/ vowel formed the nucleus, and the /g/ formed the coda.

Even though all participants were informed, prior to the task, that they would only see monosyllabic words, many of their responses had two or more syllables in them. For all responses—no

matter how many syllables they contained--the “nucleus” was considered to be the vowel with the highest stress level in the response. All segments—including any consonants or vowels—which preceded this response nucleus were then taken to be the “onset” of the response, and all segments which succeeded it were taken to be the response’s “coda.” For example, one participant gave the response “camera” to the point-light stimulus “thumb.” The phonetic transcription for “camera” in the CMU pronouncing dictionary is /k ae1 m ax0 r ax0/. Since the /ae1/ vowel has the highest stress level in the word, it was taken to be the “nucleus” of the response. Thus, the /k/ which preceded it formed the response “onset,” while the final /m ax0 r ax0/ sequence formed the “coda.”

Response onsets, nuclei, or codas—as determined in this fashion--were only considered to be correct identifications of their counterparts in the original stimuli if the two matched perfectly. Thus, response onsets or codas which contained more than one segment were considered to be incorrect even if one of those segments formed the original stimulus onset or coda. Thus, the /m ax0 r ax0/ coda of “camera” did not count as a correct identification of the /m/ coda in the “thumb” stimulus, even though an /m/ formed part of the response coda. Many of the participants’ responses could not be matched to any entry in the CMU pronouncing dictionary. Those that were obvious misspellings (e.g., “cheif”) were simply corrected in the original data file and then matched with the corresponding dictionary entry, while those responses that were not obviously English words (e.g., “rith”) were given onset-nucleus-coda transcriptions by hand and then scored accordingly.

Participant responses were also scored in terms of whether or not the participants had correctly identified the place of articulation and viseme category of the stimulus coda and onset. Each coda and onset consonant in the stimulus was thus classified in terms of both its place of articulation (alveolar, bilabial, interdental, glottal, labio-dental, labio-velar, palato-alveolar, velar), and its viseme type (bilabial, interdental, dorso-lingual, glottal, lateral, labio-dental, labio-velar, palato-alveolar, retroflex, /s/). “Visemes” is a concept that was first introduced by Walden et al. (1977) in order to account for the broad categories of consonantal phonemes that can be consistently identified in visual-only speech perception. Most viseme categories correspond primarily to a particular place of articulation, but they also include a few categories which are determined by manner of articulation (e.g., lateral /l/, retroflex /r/, and alveolar fricative /s/. Similar classifications were also made for the place of articulation and viseme category of the consonants in the response onsets and codas. Those response onsets and codas which contained more than one segment were considered to have “mixed” places of articulation or viseme categories--unless all of the segments in those onsets and codas happened to agree in either viseme type or place of articulation. In this case the common viseme category or place of articulation was then taken to be the appropriate classification for that portion of the response.

The place of articulation or viseme type of the onsets, nuclei and codas of the responses were only counted as correct identifications if they exactly matched the corresponding sub-phonemic features of the stimulus. One participant, for instance, gave the response “damp” to the “dame” stimulus. In “damp,” the coda /mp/ was considered to have the bilabial place of articulation, since both /m/ and /p/ are bilabial consonants. This was scored as a correct identification of the place of the stimulus coda consonant, since the coda /m/ in “dame” also has a bilabial place of articulation. Another participant, however, identified the same “dame” stimulus as “table.” Since the coda of “table” includes both /b/ and /l/ segments, which have bilabial and alveolar places of articulation, respectively, it was categorized as having a “mixed” place of articulation. This response was therefore scored as an incorrect identification of the bilabial place of articulation in the stimulus coda /m/.

Results and Discussion

Prior to the experiment, we expected to find two general trends in the response data. First, we expected the percentage of correct responses to be much higher for the full-face stimuli than for the point-light stimuli. Second, we expected the participants' perceptual performance to improve more with event-based feedback than with either symbolic, orthographic feedback or no feedback at all. In order to quantify the amount of improvement participants made in the perceptual task over the course of the experiment—and thereby test this second prediction—percent correct scores at all levels of analysis (whole words, phonemes, and visemes/places of articulation) were tallied independently for the responses in the first and the second halves of the experiment. Comparing the differences between the percent correct scores across the two halves of the experiment thus provided a rough but straightforward way to gauge the amount of improvement participants made in identifying the visual-only stimuli over the course of the experiment.

Words Correct. Table 2 lists the mean percentages of whole words correctly identified in both halves of the experiment. This table lists these percentages separately by feedback condition and stimulus type; each set of scores thus reflects the performance of a different group of participants. The amount of “improvement” participants made over the course of the experiment can be assessed by subtracting their mean percentage correct scores in the first half of the experiment from their mean percentage correct scores in the second half of the experiment. This difference is listed for each experimental condition under the “Improvement” column in Table 2. The final column—labeled “% Improvement”—normalizes this improvement score by dividing it by the difference between 100% and the mean percentage correct score in the first half of the experiment for that condition—i.e., the potential amount the participants' mean percentage correct score could improve after their first-half performance.

Point-Light	N	First Half	Second Half	Average	Improvement	% Improvement
audio-visual	21	2.3%	3.3%	2.8%	1.0%	1.0%
audio	20	2.1%	2.8%	2.4%	0.7%	0.7%
orthographic	24	2.2%	1.7%	2.0%	-0.4%	-0.4%
none	16	1.7%	1.8%	1.8%	0.1%	0.1%

Full-Face	N	First Half	Second Half	Average	Improvement	% Improvement
audio-visual	21	22.2%	23.9%	23.1%	1.7%	2.2%
audio	22	21.9%	23.3%	22.6%	1.4%	1.8%
orthographic	20	19.9%	22.6%	21.3%	2.7%	3.4%
none	20	20.2%	22.4%	21.3%	2.2%	2.7%

Table 2. Mean percentages of whole words correctly identified by listeners in all four feedback conditions, for each stimulus type.

The data summarized in Table 2 confirm the first of the general predictions—the mean percentage of whole words correctly identified was much higher for the full-face stimuli (around 20%) than it was in the point-light stimuli (from 2% to 3%). However, the data does not provide correspondingly convincing evidence to confirm the second set of predictions—that participants' performance would improve more in the audio-visual and audio feedback conditions than in either of the other two feedback conditions. Independent, two-way Analyses of Variance (ANOVA) were run for the percentage correct data from each stimulus condition in order to test the effects that feedback type (audio-

visual, audio, orthographic, none)—a between-subjects factor—and experiment half (first, second)—a within-subjects factor—had on the raw percentages of whole words correctly identified. Neither of these ANOVAs revealed any significant effects of either feedback type or experiment half. Only the experiment half factor in the ANOVA for the full-face stimuli came marginally close to significance ($F = 3.848$; $df = 1,79$; $p = .053$).

It is perhaps not surprising that participants showed no significant improvement in their ability to correctly identify whole words in this experimental paradigm. This experiment provided participants with feedback on individual word stimuli they had just seen, but it never presented any of those stimuli to the participants again during the rest of the experiment. Feedback on an incorrect response to a particular stimulus would only be likely to help improve a participant's ability to identify that stimulus if the participant saw that same stimulus again, later on in the experiment (cf. Pashler, Cepeda, Wixted & Rohrer, in press). In the present experiment, participants never saw the same word stimulus twice. They did, however, see different tokens of the same phoneme more than once. The phoneme /p/, for instance, appeared in the onset position of five different stimulus words: pool, peace, pet, push, and page. Any participant who might have misidentified the /p/ in "pet," then, would have been informed of the correct identity of this consonant (in the feedback conditions) before getting another chance at identifying the same phoneme, in the same syllabic position, in a different word—e.g., "push." Learning the correct identities of previously misidentified phonemes in this way should have helped improve the participants' ability to identify those phonemes in subsequent stimuli. The same holds true for the sub-phonemic features of viseme type and place of articulation. The anticipated effects of feedback on the improvement of visual-only perception may be more likely to emerge in a more detailed analysis of the percentages of phonemes and features correctly identified.

Phonemes Correct. The results of the analyses of the number of phonemes correctly identified in the responses did indeed support this prediction. Tables 3 and 4 break down the mean percentages of phonemes correctly identified across the various experimental conditions and halves in the same way that Table 2 did for whole words. Table 3 lists the correct percentage scores for onset phonemes while Table 4 lists the same scores for phonemes in coda position.

Point-Light	N	First Half	Second Half	Average	Improvement %	Improvement
audio-visual	21	19.9%	22.0%	21.0%	2.1%	2.6%
audio	20	17.6%	23.5%	20.6%	5.9%	7.2%
orthographic	24	17.0%	23.2%	20.1%	6.2%	7.4%
none	16	19.1%	18.1%	18.6%	-1.0%	-1.3%

Full-Face	N	First Half	Second Half	Average	Improvement %	Improvement
audio-visual	21	46.8%	51.7%	49.3%	4.9%	9.1%
audio	22	44.5%	47.2%	45.8%	2.7%	4.8%
orthographic	20	44.3%	47.6%	45.9%	3.3%	6.0%
none	20	44.4%	48.1%	46.3%	3.8%	6.7%

Table 3. Mean percentages of onset phonemes correctly identified by listeners in all four feedback conditions, for each stimulus type.

Point-Light	N	First Half	Second Half	Average	Improvement %	Improvement
audio-visual	21	9.8%	11.6%	10.7%	1.8%	2.0%
audio	20	9.4%	11.4%	10.4%	2.0%	2.2%
orthographic	24	9.1%	8.9%	9.0%	-0.2%	-0.2%
none	16	8.3%	10.0%	9.2%	1.7%	1.8%

Full-Face	N	First Half	Second Half	Average	Improvement %	Improvement
audio-visual	21	34.7%	36.7%	35.7%	2.0%	3.0%
audio	22	36.4%	37.3%	36.8%	0.9%	1.5%
orthographic	20	35.1%	39.1%	37.1%	4.0%	6.1%
none	20	33.9%	37.1%	35.5%	3.2%	4.9%

Table 4. Mean percentages of coda phonemes correctly identified by listeners in all four feedback conditions, for each stimulus type.

These numbers reflect the same general pattern seen in the whole-word data: the scores for the full-face stimuli were much higher than those for the PLDs. Also, the scores tended to be much higher, in general, for onset phonemes than they are for coda phonemes. Independent two-way ANOVAs were run for the data from each stimulus type and syllabic position in order to test the effects of feedback type and experiment half on the percentages of phonemes correctly identified. All four of these ANOVAs revealed significant effects of experiment half on the phoneme percentage correct scores. For the ANOVA on onset phoneme identification in full-face stimuli, experiment half was significant at $p < .006$ ($F = 7.916$; $df = 1, 79$). Likewise, experiment half was significant for the coda phoneme data in the full-face stimuli ($F = 5.889$; $df = 1, 79$; $p = .018$), as well as for the point-light onset phonemes ($F = 15.7$; $df = 1, 77$; $p < .001$) and point-light coda phonemes ($F = 4.234$; $df = 1, 77$; $p = .043$).

The ANOVA for the point-light onset phoneme identification scores also revealed a significant feedback by half interaction ($F = 4.034$; $df = 3, 77$; $p = .010$). This is the only significant feedback by half interaction that emerged in the analysis of the data. Figure 4 presents this interaction graphically; it shows that the improvements made by the audio and orthographic feedback groups were significantly greater than those made by the audio-visual and no feedback groups.

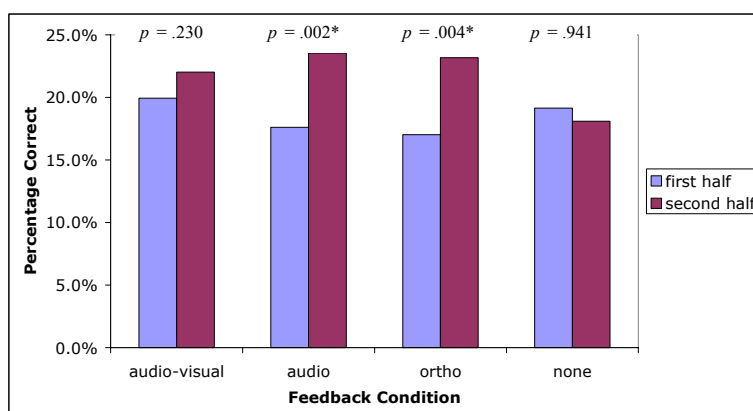


Figure 4. Percentage of onset phonemes correctly identified, point-light stimuli (p values are from t-test of significance for percentage correct scores between first and second halves of the experiment in that particular feedback condition).

It is not immediately clear why audio-visual feedback did not help participants' visual-only perception at least as much as audio feedback. One possible explanation is that the onset of the participants' perceptual improvement in the audio-visual feedback condition occurred earlier than it did in the audio or orthographic groups; the percentage of correct onset phoneme identifications is, in fact, higher in the first half of the experiment for the AVF group (19.6%) than it is for either the AF (17.6% or OF (17.0%) groups. The decision to compare correct percentage scores between the first and second halves of the experiment is arbitrary, and grouping the response data together in halves may have thus glossed over improvements the participants made during the first half of the experiment alone. However, this explanation may be confounded by the fact that different groups of participants took part in each feedback condition, and any one group of participants may have entered the task with a greater or lesser ability to perceive speech from visual-only PLDs than any of the other groups. For instance, the no feedback group performed relatively well in the first half of this condition—scoring 19.1% onset phonemes correctly identified, which was nearly as good as the percentage posted by the AVF group. However, the no feedback group's performance then declined to 18.1% in the second half of the experiment. Their relatively high percentage correct score in the first half of the experiment is probably, therefore, not due to significant improvement during the first half alone. On the other hand, the AVF group may, on the other hand, not have been able to improve as much as the AF or OF groups because their performance started out closer to ceiling level in this task, which may be near 23% or 24% correct. Since the OF and AF groups started with a lower baseline performance, their ability to attain near-ceiling performance in the second half of the experiment produced a significant effect of experiment half in the Analysis of Variance, even though similarly high levels of performance by the AVF group in the second half were not significantly different from their higher performance baseline in the first.

It is also important to emphasize that the three groups who received feedback also showed improvement between the first and second halves of the experiment that were in the expected (positive) direction. For the group that did not receive feedback, however, participant performance on onset phoneme identification dropped between the first and the second halves of the experiment. This pattern of results suggests that feedback does have a positive effect on the participants' ability to perform this visual-only speech perception task. The fact that the feedback by experimental half interaction only emerged among the groups who saw point-light stimuli is also suggestive, since they were predicted to be more susceptible to the positive effects of feedback, due to their unfamiliarity with that type of stimulus. The broad trend for the full-face stimuli, on the other hand, was improvement between first and second halves for all feedback groups, including the group that received no feedback at all. This suggests that the participants in the full-face conditions could simply get better at the experimental task through practice, rather than having to depend on the information they received through feedback in order to interpret the visual-only stimuli.

The fact that the participants did not show significant amounts of improvement—in all experimental conditions—in their ability to correctly identify the vowel phonemes in the stimuli is also suggestive. Table 5 shows the mean percentages of vowels correctly identified for all feedback groups and both stimulus types, broken down by experiment half.

Independent, two-way (feedback type x experimental half) ANOVAs were run on the percentages of vowels correctly identified in both the full-face and point-light stimuli, considering feedback type and experimental half as factors. Neither of these ANOVAs revealed any significant factors or interactions in the vowel identification scores. Nonetheless, the percentages of vowels correctly identified in Table 5 reveal some interesting trends. In general, the percentages of vowels the participants correctly identified were much higher for the full-face stimuli (about 50%) than it was for the point-light stimuli (about 20%). The participants' ability to identify vowels correctly for the full-face stimuli was also marginally better

than their ability to identify consonant phonemes in the same stimuli in either onset or coda position. For the full-face stimuli, participants identified between 50% and 55% of the vowels correctly, while their percentages for the onset phonemes in the same stimuli were between 45% and 50%. While this difference may seem significant at first glance, it may simply reflect the fact that participants had to identify vowels from a smaller set of responses (15) than they had for consonants (22 alternatives in onset position).

Point-Light	N	First Half	Second Half	Average	Improvement%	Improvement
audio-visual	21	19.0%	22.2%	20.6%	3.2%	3.9%
audio	20	19.3%	23.1%	21.2%	3.9%	4.8%
orthographic	24	20.3%	19.8%	20.1%	-0.5%	-0.7%
none	16	19.7%	17.8%	18.8%	-1.8%	-2.3%

Full-Face	N	First Half	Second Half	Average	Improvement%	Improvement
audio-visual	21	56.3%	56.3%	56.3%	0.0%	0.0%
audio	22	53.5%	53.5%	53.5%	0.0%	0.0%
orthographic	20	49.2%	51.9%	50.5%	2.7%	5.3%
none	20	54.0%	54.6%	54.3%	0.6%	1.4%

Table 5. Mean percentages of vowel phonemes correctly identified by listeners in all four feedback conditions, for each stimulus type.

Seen in this light, the fact that nearly equivalent percentages of vowels and onset phonemes were identified correctly in the point-light stimuli ($\approx 20\%$) reflects poorly on the participants' vowel identification skills. However, the patterns of improvement for vowel identification in the point-light conditions revealed an interesting trend: participants in the AVF and AF groups actually got better (improvements of 3.2% and 3.9% correct, respectively), while the OF and NF groups got marginally worse (decreases of .5% and 1.8% correct, respectively). Even though these levels of improvement are not statistically significant, this pattern roughly confirms the prediction that dynamic, event-based feedback could help improve participant performance more than either static, symbolic feedback or no feedback at all. However, the full-face vowel identification data exhibits the exact opposite pattern—participants showed no improvement at all between halves in the AF and AVF conditions (0% for both, exactly), while they showed marginal improvement in the OF and NF conditions (2.7% and 0.6%, respectively). Future work might be able to better clarify the importance of these patterns of improvement by increasing the number of stimuli in the experiment. This could enable some of these suggestive—but conflicting—trends in the improvement of vowel identification to reach statistical significance.

Place of articulation and visemes correct. Participants' failure to significantly improve their vowel identification scores may have, however, reflected a lack of clearly defined place of articulation cues in vowels. In contrast to their performance on vowel identification, participants in all the experimental conditions improved their ability to correctly identify both place of articulation and viseme type between the two halves of the experiment. Tables 6 and 7 show the mean percentage scores for visemes correctly identified in both halves of the experiment, broken down by the eight participant groups in the study.

The scores here mirror those in the phoneme correct identification scores, except that the overall percentage means were considerably higher. For the full-face stimuli, for instance, viseme scores were around 70% in onset position and 65% in coda position, while they were only between 45 to 50% for onset phonemes and around 35% correct for coda phonemes. Independent, two-way ANOVAs—using

experiment half and feedback type as factors—were carried out on the raw percentages of visemes correctly identified in both syllabic positions, for both stimulus types. These ANOVAs yielded similar significant effects to those found in the ANOVAs for the phoneme identification data. Experiment half was a significant factor for onset visemes in the full-face stimuli ($F = 15.98$; $df = 1,79$; $p < .001$), as well as for the coda visemes in the same condition ($F = 8.287$; $df = 1,79$; $p = .005$). The same factor was also significant for both onset visemes ($F = 21.26$; $df = 1, 77$; $p < .001$) and coda visemes ($F = 4.487$; $df = 1,77$; $p = .037$) in the point-light stimuli. All of these factors were also significant in the corresponding ANOVAs for the phoneme identification data; the only difference between the Analyses of Variance at the two different levels of linguistic structure was that the feedback by half interaction did not reach significance for the identification of point-light onset visemes, even though it did for the identification of point-light onset phonemes.

Point-Light	N	First Half	Second Half	Average	Improvement%	Improvement
audio-visual	21	37.5%	44.6%	41.1%	7.1%	11.4%
audio	20	37.4%	43.2%	40.3%	5.8%	9.3%
orthographic	24	34.3%	41.2%	37.8%	6.9%	10.6%
none	16	35.5%	37.4%	36.5%	1.8%	2.8%

Full-Face	N	First Half	Second Half	Average	Improvement%	Improvement
audio-visual	21	69.5%	73.5%	71.5%	4.0%	13.0%
audio	22	67.5%	69.7%	68.6%	2.2%	6.7%
orthographic	20	67.5%	72.4%	69.9%	4.9%	15.1%
none	20	67.4%	71.0%	69.2%	3.6%	11.2%

Table 6. Mean percentages of onset visemes correctly identified by listeners in all four feedback conditions, for each stimulus type.

Point-Light	N	First Half	Second Half	Average	Improvement%	Improvement
audio-visual	21	27.5%	28.8%	28.1%	1.3%	1.8%
audio	20	26.7%	33.2%	29.9%	6.6%	8.9%
orthographic	24	28.4%	27.4%	27.9%	-1.0%	-1.3%
none	16	26.3%	27.9%	27.1%	1.6%	2.1%

Full-Face	N	First Half	Second Half	Average	Improvement%	Improvement
audio-visual	21	62.1%	64.9%	63.5%	2.8%	7.3%
audio	22	64.4%	66.1%	65.2%	1.7%	4.8%
orthographic	20	65.4%	67.7%	66.6%	2.3%	6.6%
none	20	61.1%	65.3%	63.2%	4.2%	10.7%

Table 7. Mean percentages of coda visemes correctly identified by listeners in all four feedback conditions, for each stimulus type.

The close parallels between the phoneme and viseme data seem to indicate that the overall phoneme scores depended to a large extent on the viewers' ability to identify each phoneme's viseme type. The 20% discrepancy between viseme scores and phoneme scores in most conditions may thus be attributed largely to the difficulty the participants had in identifying the sub-phonemic features of the onset and coda which are not integrated into the categorization of visemes—i.e., voicing and, to a lesser extent, manner of articulation. The difficulty the participants had in identifying manner and voice in the

point-light stimuli may also account for the aforementioned differences in the results at the two different levels of linguistic structure. The significant feedback by half interaction which emerged in the point-light onset phoneme correct scores indicated that scores improved in the conditions where participants received feedback but got worse in the condition where participants received no feedback at all. A similar feedback by half interaction failed to emerge for the viseme correct data because these scores improved—at least marginally—in all four feedback conditions. This pattern of results suggests that the participants could improve their ability to perceive particular viseme categories even without feedback but that their ability to pick up manner and voice information in the point-light stimuli could only improve if they received feedback about those stimuli.

	Point Light		Full Face			Point Light		Full Face	
	1st	2nd	1st	2nd		1st	2nd	1st	2nd
Alveolars					Labio-velars				
audio-visual	44.9%	52.6%	40.9%	46.1%	audio-visual	57.1%	72.9%	95.9%	91.9%
audio	36.3%	43.3%	36.1%	46.3%	audio	73.4%	80.2%	98.7%	93.6%
orthographic	32.2%	38.2%	40.9%	50.8%	orthographic	73.7%	80.4%	98.5%	94.7%
none	34.2%	23.7%	43.0%	46.5%	none	35.1%	5.5%	95.7%	95.7%
Bilabials					Laterals				
audio-visual	65.6%	67.3%	90.9%	92.3%	audio-visual	18.4%	19.6%	71.4%	89.3%
audio	72.4%	71.9%	87.0%	90.6%	audio	10.4%	15.5%	86.8%	84.2%
orthographic	70.9%	70.4%	89.6%	94.7%	orthographic	1.6%	12.1%	84.3%	85.7%
none	61.6%	48.3%	89.9%	92.9%	none	13.2%	9.5%	94.1%	93.9%
Glottals					Palato-alveolars				
audio-visual	15.9%	42.5%	33.3%	42.2%	audio-visual	15.3%	30.4%	93.5%	97.7%
audio	12.0%	16.0%	38.2%	42.4%	audio	14.9%	22.0%	94.2%	96.6%
orthographic	32.6%	37.7%	61.0%	61.5%	orthographic	14.1%	26.8%	90.0%	94.0%
none	15.6%	3.1%	47.1%	56.5%	none	10.3%	7.3%	95.0%	96.7%
Interdentals					Retroflex				
audio-visual	0.0%	2.9%	92.3%	91.9%	audio-visual	1.9%	7.3%	59.5%	70.0%
audio	0.0%	0.0%	90.2%	92.0%	audio	0.7%	2.9%	48.7%	55.3%
orthographic	2.9%	0.0%	96.6%	80.6%	orthographic	3.9%	5.2%	45.3%	54.4%
none	0.0%	0.0%	82.1%	87.5%	none	2.4%	3.9%	40.6%	51.3%
Labio-dentals					Velars				
audio-visual	42.2%	60.6%	86.0%	90.3%	audio-visual	7.6%	8.8%	37.2%	29.7%
audio	33.3%	49.2%	76.5%	82.8%	audio	7.6%	18.5%	32.0%	31.7%
orthographic	24.4%	39.8%	83.0%	88.3%	orthographic	9.9%	17.3%	29.2%	30.0%
none	21.2%	29.0%	88.3%	84.9%	none	10.0%	5.3%	28.9%	23.2%

Table 8. Percent hits, by place of articulation, across feedback condition, experiment half, and stimulus type for phonemes in onset position.

Interestingly, the discrepancy between the percentage of visemes correctly identified and the percentage of phonemes correctly identified increased to about 30% for the coda segments in the full-face stimuli. This shift in correct identification scores may have been the result of a lack of phonetic balance for the various places of articulation in the stimuli codas combined with the fact that the participants' ability to identify place of articulation was not uniform across all place categories.

Tables 8 and 9 show the percentages of correct identifications participants made for each individual place of articulation, for both point-light and full-face stimuli, across both halves of the experiment. Table 8 shows these percentages for the various places of articulation in onset position, while Table 9 shows the corresponding percentages for the coda places of articulation.

	Point Light		Full Face			Point Light		Full Face	
	1st	2nd	1st	2nd		1st	2nd	1st	2nd
Alveolars					Laterals				
audio-visual	31.1%	35.5%	58.4%	63.1%	audio-visual	12.3%	14.7%	34.1%	40.0%
audio	30.9%	34.5%	59.6%	62.3%	audio	4.5%	8.9%	39.8%	45.2%
orthographic	33.5%	30.0%	63.4%	65.5%	orthographic	7.3%	6.3%	43.5%	44.0%
none	33.8%	34.4%	53.5%	58.4%	none	6.3%	7.7%	38.3%	35.4%
Bilabials					Palato-Alveolars				
audio-visual	46.2%	34.4%	59.3%	65.0%	audio-visual	11.9%	25.4%	92.2%	88.7%
audio	37.4%	40.9%	57.4%	65.6%	audio	23.6%	26.9%	79.7%	88.9%
orthographic	40.0%	34.9%	60.0%	64.2%	orthographic	31.0%	31.5%	81.8%	83.3%
none	30.9%	17.1%	56.5%	80.0%	none	7.1%	4.4%	87.3%	87.7%
Interdentals					Retroflex				
audio-visual	1.9%	9.6%	70.7%	80.9%	audio-visual	8.0%	9.1%	41.5%	40.4%
audio	12.5%	4.9%	70.9%	80.0%	audio	8.8%	10.3%	59.1%	36.4%
orthographic	4.4%	4.0%	70.6%	81.6%	orthographic	12.3%	12.7%	30.4%	38.9%
none	---	7.5%	63.3%	66.7%	none	0.0%	4.5%	44.9%	49.0%
Labio-dentals					Velars				
audio-visual	16.7%	27.0%	73.7%	77.5%	audio-visual	9.9%	6.3%	26.2%	20.1%
audio	17.8%	17.6%	75.0%	79.7%	audio	8.5%	14.0%	17.2%	17.9%
orthographic	5.7%	23.8%	71.6%	78.1%	orthographic	9.2%	14.8%	22.8%	19.7%
none	1.3%	2.8%	81.4%	80.0%	none	4.3%	2.8%	21.6%	16.4%

Table 9. Percent hits, by place of articulation, across feedback condition, experiment half, and stimulus type for phonemes in coda position.

The data in these tables indicates that, in general, those places of articulation nearer the front of the mouth (bilabial, labio-dental, labio-velar) are easier for participants to identify than those which are further back (e.g., velar, glottal). The alveolar consonants—which have a place of articulation in between the labials and the velars—show a moderately high rate of correct identification in onset position. For the

full-face stimuli, for instance, participants correctly identified alveolars in onset position between 40% and 50% of the time. This falls in between the percentage of bilabials they correctly identified (90%) and the percentage of velars they correctly identified (30%). In coda position, however (see Table 9), the percentage of correct alveolar identifications (about 60%) was much closer to that of bilabials (60% to 65%) than it was to that of velars (about 20%). This increase in the participants' success at identifying alveolars in coda position reflects both an inherent response bias in the participants, as well as an imbalance in the places of articulation in the stimuli codas. Table 10 shows the distribution of the various places of articulation in both the onsets and the codas of the experimental stimuli.

Onset	N	%	Coda	N	%
Bilabials	21	21.9%	Bilabials	9	9.4%
Labio-dentals	10	10.4%	Labio-Dentals	7	7.3%
Labio-velars	7	7.3%			
Interdentals	3	3.1%	Interdentals	5	5.2%
Alveolars	19	19.8%	Alveolars	42	43.8%
Laterals	5	5.2%	Laterals	8	8.3%
Retroflex	11	11.5%	Retroflex	5	5.2%
Palato-Alveolars	5	5.2%	Palato-Alveolars	6	6.3%
Velars	11	11.5%	Velars	14	14.6%
Glottals	4	4.2%			
Total	96		Total	96	

Table 10. Distribution of places of articulation for onset and coda consonants in experimental stimuli.

Table 10 clearly shows that nearly half of the stimuli codas were alveolars. This predominance of alveolars in coda position reflects a general trend in the English language. 48.2% (61,234 out of all 127,006 items in the CMU pronouncing dictionary end in alveolar consonants). This is, in part, due to the predominance of word-final alveolar inflections in English (e.g., plural /-s/, past-tense /-d/, etc.), but it holds true of monomorphemic items, as well—32.4% (194 out of 598) of the monomorphemic CVC words in the CMU pronouncing dictionary also end in alveolars. Thus, the ability of the participants in this study to identify alveolar coda consonants well probably reflects their realization that they should expect this place of articulation to appear often in the coda position of English words. It may also reflect a tendency on the part of the participants to simply choose words from the English lexicon—at random, even—which happened to end in alveolar consonants. Such a bias towards responding with coda consonants which end in alveolar consonants may have converged with the greater likelihood of this place of articulation in the stimuli codas to increase the overall hit rate for coda alveolars across all experimental conditions. To the extent to which this increased hit rate was due to response bias, it does not reflect an increased sensitivity to the visual cues for the alveolar place of articulation in coda position. Nonetheless, this increase in the percentage of hits for alveolars in coda position probably accounts in large part for the increased difference between the percentage of correct viseme identifications and the percentage of correct phoneme identifications for the coda consonants in the full-face conditions.

Breaking down correct identification scores by individual places of articulation also reveals a number of interesting differences between the perception of full-face and point-light stimuli. In general, the identification of place in the full-face stimuli was much better than it was in the point-light stimuli. For example, the average percent correct for bilabial place of articulation in onset position, across all four feedback conditions, hovered around 90% for the full-face stimuli, but was only a modest 70% in most of the point-light conditions. For other places of articulation, however, the point-light groups did far worse

than might be expected. In particular, the participants had almost no ability to identify the interdental place of articulation in the onset of the point-light stimuli, scoring little better than 0% correct for interdental consonants in all feedback conditions. The corresponding percentages correct for the full-face groups were, on the other hand, close to 90%. The complete inability of the point-light groups to identify a place of articulation which was relatively easy for the full-face groups to perceive probably reflects significant gaps in the particular pattern of fluorescent dots that were placed on the speaker's faces, lips, teeth and tongues during the production of the point-light stimuli. While one dot was placed on the blade of each speaker's tongue, along with two dots each on both rows of teeth, there were no dots placed on the edge or tip of the speaker's tongue. Such dots may not have provided salient cues to the interdental place of articulation. The lack of such cues in the point-light stimuli meant the participants could perceive nothing in them that indicated that the speaker had produced an interdental consonant. This seems to have been especially true in onset position; for the interdentals in coda position, on the other hand, the participants' performance improved modestly to between 5% and 10% correct identifications. This improvement was probably due in large part to the transitional cues afforded the viewers by the visual offset of the vocalic portion of the CVC stimuli, rather than any particular configuration or dynamic transformation of the fluorescent dots on the interdental articulators.

The point-light groups also had particular difficulty identifying the retroflex consonant /r/ in both onset and coda positions. They correctly identified less than 5% of the /r/ tokens in onset position, while the full-face groups correctly identified between 40% and 60% of the /r/ tokens in the onset of their stimuli. For the /r/s in coda position, the point-light groups' percentages correct increased modestly to between 5% and 10%, while the full-face groups' performance decreased slightly to around 40% correct, in general. The reason behind the participants' inability to identify /r/ in the point-light stimuli may, in essence, be the same as that proposed for the difficulty they had in identifying interdentals—i.e., a lack of fluorescent dots at the appropriate places on the retroflex articulators. However, the two groups' perception of /r/ differed in more than just their ability to identify this segment correctly; the point-light groups consistently misidentified /r/ tokens in a different way than the full-face groups misidentified them. Tables 11 and 12 show confusion matrices for the retroflex stimuli, broken down by stimulus type, feedback condition and experiment half. Table 11 shows this data for the /r/s in onset position while Table 12 provides the same data for /r/s in coda position. Table 12 reveals that participants most often misidentified /r/ in the onset of the full-face stimuli as labio-velar /w/. For example, one participant misidentified “rang” as “ways.” For the point-light stimuli, however, the participants primarily misidentified onset /r/ as a bilabial (e.g., /b/, /p/ and /m/). One participant, for example, misperceived “rang” as “bounce.” That /r/ might be misperceived as /w/ in the full-face stimuli is not surprising, since both consonants have similar lip-rounding gestures (Johnson, 2002). Children often substitute /w/ for /r/, in fact, before they have learned to produce the (invisible) tongue-curling and pharynx-constricting gestures necessary to make an adult-like /r/ sound in English. Why the /r/ tokens in the point-light stimuli were so often misidentified as bilabials and not labio-velars, however, is not clear. This may reflect a bias in the participants to identify any sound with a labial gesture as bilabial, since this is the place of articulation they are most likely to identify correctly and therefore receive positive feedback on.

Whatever the reason behind the point-light groups' misidentification of /r/ may be, however, it is interesting to note that they were able to identify labio-velars in onset position quite well—especially those groups who received feedback. Those three groups correctly identified between 60% and 80% of labio-velars in onset position, whereas the no feedback group scored 35% to 5% for the same stimuli across both halves of the experiment. The scores for the point-light groups were, in fact, nearly as high as those for the full-face groups, which topped out near ceiling between 95% and 100%. These results suggest that the participants could readily identify the cues to labio-velar place of articulation that were preserved in the point-light stimuli, and that feedback (of any kind) was also particularly helpful in attuning their perceptual systems to these cues. However, this pattern of results suggests that whatever

aspects of articulation appear to be visually similar between retroflex /r/ and labio-velar /w/ in the full-face stimuli were not preserved in the PLDs used in this experiment, since the point-light groups did not have difficulty identifying these cues in the labio-velar stimuli themselves. In a broader sense, this pattern of misidentifications also indicates that the point-light stimuli do not just transmit a simplified representation of the dynamic cues in the full-face displays. Instead, there is some degree of dissociation between the two visual representations of the same set of articulatory gestures.

Point-Light													
AVF	bi	ld	lv	id	al	la	re	pa	ve	gl	None	Other	total
1 st	67	5	2	1	8	5	2	4	1	0	5	8	108
2 nd	75	3	11	0	4	4	9	3	3	0	2	9	123
AF	bi	ld	lv	id	al	la	re	pa	ve	gl	None	Other	total
1st	96	4	2	0	10	0	1	4	4	2	8	8	139
2nd	83	5	9	0	13	0	4	5	2	2	2	11	136
OF	bi	ld	lv	id	al	la	re	pa	ve	gl	None	Other	total
1st	78	4	5	1	8	3	5	3	2	5	5	10	129
2nd	85	5	8	0	7	1	7	1	3	4	7	7	135
NF	bi	ld	lv	id	al	la	re	pa	ve	gl	None	Other	total
1st	65	5	4	0	13	2	3	1	3	3	9	17	125
2nd	33	0	4	0	3	0	2	0	1	1	3	4	51
Full-Face													
AVF	bi	ld	lv	id	al	la	re	pa	ve	gl	None	Other	total
1st	2	0	38	0	0	0	66	0	0	0	2	3	111
2nd	2	0	32	0	1	1	84	0	0	0	0	0	120
AF	bi	ld	lv	id	al	la	re	pa	ve	gl	None	Other	total
1st	3	0	57	0	0	1	58	0	0	0	0	0	119
2nd	3	0	48	0	3	0	68	0	0	1	0	0	123
OF	bi	ld	lv	id	al	la	re	pa	ve	gl	None	Other	total
1st	5	2	47	0	0	0	48	0	0	1	1	2	106
2nd	2	1	46	0	1	0	62	0	1	0	0	1	114
NF	bi	ld	lv	id	al	la	re	pa	ve	gl	None	Other	total
1st	4	1	48	0	0	0	41	0	1	1	2	3	101
2nd	2	0	54	0	1	0	61	1	0	0	0	0	119

Table 11. Response totals for /r/ phonemes in the onset of the stimulus. (Response place key: bi = bilabial, ld = labio-dental, lv = labio-velar, id = interdental, al = alveolar, la = lateral, re = retroflex, pa = palato-alveolar, ve = velar, gl = glottal).

The two groups of participants also showed a different set of response biases when they misidentified retroflex /r/ in coda position. Table 12 provides more details about these biases; note that the point-light group often misidentified coda /r/s as either alveolars or bilabials. Over the course of the experiment, however, this bias appears to shift away from the bilabials and towards more anterior places of articulation, such as lateral and velar. For the full-face stimuli, on the other hand, there are very few alveolar or bilabial misidentifications in either half of the experiment. There are, however, many misidentifications of retroflex /r/ as lateral /l/, which seem to increase in frequency—across most feedback conditions—between the first and the second halves of the experiment. One participant in the

Point-light											
AVF	bi	ld	id	al	la	re	pa	ve	None	Other	Total
1st	7	0	2	13	5	4	1	7	7	4	50
2nd	3	1	0	16	7	5	1	8	3	11	55
AF	bi	ld	id	al	la	re	pa	ve	None	Other	Total
1st	12	1	2	12	1	5	3	4	9	8	57
2nd	5	1	1	23	9	7	0	11	5	6	68
OF	bi	ld	id	al	la	re	pa	ve	None	Other	Total
1st	11	0	1	13	7	7	0	2	7	9	57
2nd	6	4	0	7	9	8	1	7	7	14	63
NF	bi	ld	id	al	la	re	pa	ve	None	Other	Total
1st	10	0	2	12	0	0	0	1	5	6	36
2nd	8	0	3	12	1	2	1	3	6	8	44
Full-face											
AVF	bi	ld	id	al	la	re	pa	ve	None	Other	Total
1st	0	1	0	5	7	22	0	1	5	12	53
2nd	1	1	0	0	14	21	0	1	5	9	52
AF	bi	ld	id	al	la	re	pa	ve	None	Other	Total
1st	1	0	0	1	2	26	0	1	0	13	44
2nd	0	1	0	5	11	24	0	0	5	20	66
OF	bi	ld	id	al	la	re	pa	ve	None	Other	Total
1st	0	2	0	5	8	14	0	2	6	9	46
2nd	1	0	0	6	6	21	1	3	5	11	54
NF	bi	ld	id	al	la	re	pa	ve	None	Other	Total
1st	3	1	0	3	9	22	1	1	3	6	49
2nd	0	1	0	0	12	25	0	1	4	8	51

Table 12. Response totals for /r/ phonemes in the coda of the stimulus. (Response place key: bi = bilabial, ld = labio-dental, lv = labio-velar, id = interdental, al = alveolar, la = lateral, re = retroflex, pa = palato-alveolar, ve = velar, gl = glottal).

full-face condition, for example, misidentified “beer” as “bowl” in the full-face condition, while a different participant misidentified the corresponding point-light stimulus as “put.” It is not clear why the participants tended to misidentify coda /r/s as /l/s in the full-face stimuli; it is possible that they may have picked up on a shared element of semi-syllabicity in the codas ending in these two liquids. The tendency for the participants in the point-light groups to misidentify coda /r/s as alveolars, on the other hand, may simply reflect the participants’ broader tendency to respond with words which ended in alveolars, especially when there was insufficient evidence in the stimulus to specify an alternative place of articulation in the coda consonant.

Conclusions

This study was designed to investigate the potential differences in perception between the visual-only PLDs and full-face displays of speech. We explored the effects that different kinds of feedback might have on the participants’ ability to perceive speech in either kind of visual-only display. The results of this investigation confirmed the prediction that it should be easier to perceive speech in full-face displays than in PLDs. The participants in the full-face conditions correctly identified the words, phonemes and visemes at consistently higher levels than the participants in the point-light conditions. These results are likely due to the fact that full-face displays provide more visual information about articulatory gestures to perceivers than PLDs do. The full-face displays not only show all of a speaker’s face but also potentially provide complementary static cues to perceivers instead of just the dynamic, time-varying cues found in PLDs. The participants in this study also had no experience perceiving speech in PLDs prior to this experiment, and may therefore, have found them perceptually difficult simply because they were unfamiliar with them. The results of this study do confirm, however, that—despite the impoverished visual signals in the point-light stimuli and the fact that none of the participants had seen them before—observers can, to a limited extent, accurately perceive speech in visual-only PLDs (cf. Rosenblum et al., 1996, Rosenblum & Saldaña, 1996; Lachs & Pisoni, submitted). This finding indicates that observers can perceive speech using only dynamic cues—i.e., moving points of light—without necessarily being able to recognize the pattern of the talker’s face or their individual articulators. It is presumably possible for observers to do this because the pattern of point-light movements is highly constrained and lawfully determined by the articulatory events in speech that the observers perceive as meaningful.

Investigating the differences in the participants’ ability to perceive particular places of articulation in the two different types of stimuli revealed that the PLDs are not just simplified or impoverished versions of the full-face displays. Certain places of articulation—e.g., interdental and retroflex—were nearly impossible for participants to perceive in the PLDs despite their relative ease of perceptibility in the full-face stimuli. Furthermore, certain systematic misidentifications emerged in the perception of point-light stimuli (e.g., bilabials for retroflexes) which differed from more common patterns of misidentification found in full-face displays (e.g., labio-velars for retroflexes). These results indicate that the particular patterns of dot placement used in the production of the point-light stimuli in this experiment were probably less than ideal; they failed to capture certain aspects of the dynamics of articulation which are clearly perceptible in the full-face displays and also gave viewers unusually misleading information about certain places of articulation. For this reason, the point-light stimuli developed a peculiar form of perceptual independence from their full-face counterparts; the participants perceived articulatory events in them which they would not have perceived in full-face displays of the same events. This finding serves as an important caution against concluding too much about the visual perception of speech under normal viewing conditions based only on the results of studies which have shown that observers can successfully perceive speech in PLDs (e.g., Rosenblum & Saldaña, 1996; Lachs & Pisoni, submitted); the perception of speech in one type of display does not necessarily reflect the

perception of speech in the other. It is also not known at this time whether different patterns of dot placement might eliminate this perceptual independence between the two visual representations of speech. However, determining a pattern of dots to use in point-light speech stimuli which can more faithfully represent the dynamics of articulation in fully-illuminated displays may provide a fruitful line of future research.

The results of this preliminary study on the effects of feedback on participants' improvement in the perceptual task were unfortunately somewhat inconclusive. No main effects of feedback were found in any of the experimental conditions, and the one significant feedback by experimental half interaction which did emerge yielded a pattern of improvement across the four feedback conditions which was not consistent with any of the effects that feedback was predicted to have on the participants' perceptual improvement. Thus, while it may seem rational to suggest that dynamic, event-based feedback—in audio or audio-visual form—may improve viewer performance in a visual-only speech perception task more than either static, symbolic feedback or no feedback at all would, the results of this experiment do not provide any solid empirical support for that hypothesis. This null result leaves open the question of just how important dynamic, articulatory event-based feedback really is to viewers in improving their skills in the visual-only perception of speech. It also remains unclear, for that matter, whether feedback is of any particular use to participants in a visual-only perception task, when they can evidently improve just as much without feedback as they can when they receive either orthographic or audio-visual feedback on the stimuli they have just seen. The lack of statistically significant differences in perceptual improvement between the various feedback groups suggests, in fact, that participants may have been able to improve by simply becoming more familiar with the experimental task. Practice with the visual-only perception task may have helped the participants fine-tune their perceptual systems into what they already knew about the acoustic consequences of particular articulations. For example, one participant remarked after the experiment that she had often tried to articulate the response she had in mind in order to determine if it matched the articulatory gestures she had seen the speaker make on the computer screen. Tapping into such tacit knowledge of articulation may thus provide a better guide to correctly identifying visual-only stimuli than being informed—after the fact—what a speaker in a silent video has just said.

However, the results of this preliminary study do not necessarily close the theoretical door on the possible efficacy of feedback on improvement in a visual-only speech perception experiment. There are a number of ways in which the paradigms used in this experiment might be modified in order to provide more optimal conditions for the emergence of the anticipated effects of feedback on perceptual improvement. For example, even though the various amounts of improvement made by the participants in the different feedback groups were not significantly different from one another, some groups did show a trend at improving more than other groups in the identification of certain aspects of the stimulus words. The audio and audio-visual feedback groups, for instance, improved more on vowel identification than either the orthographic or no feedback groups. Such a trend might become statistically significant if the experiment contained more than just 96 visual-only stimuli, thereby giving the participants more time (and practice) to improve their perceptual skills.

Feedback may also become more efficacious if participants get second chances to identify stimuli they have already seen before and received feedback on. In this experiment, participants only received feedback on individual words, all of which they saw only once. In this vein, it may also be more informative to analyze the efficacy of feedback in terms of how the ability of the participants to identify a particular stimulus (or part thereof) improves with each successive presentation of that stimulus in the experiment. The analyses carried out in this report looked at the effects of feedback over the comparatively broad scopes of the first and second halves of the experiment. Small-scale improvements within each half may have thus been lost in the analysis. Analyzing feedback effects in terms of the number of times a particular stimulus has been seen might also be made easier by balancing the

proportions of places of articulation in coda and onset consonants across all stimuli. The coda consonants in this experiment's stimuli included a large proportion of alveolars, while the onset consonants had—to a lesser extent—disproportionate amounts of bilabials and alveolars. It was therefore not possible to make equitable assessments of the participants' improvement in identifying particular places of articulation after a certain number of presentations of that place of articulation in the experiment—the participants saw too many tokens of some places of articulation and too little of others. Furthermore, the predominance of alveolars—which do not have strong visual cues to their place of articulation—in the coda position of the experiment's stimuli also facilitated deceptively positive effects of response bias on the percentage of correct coda identifications. This disproportionate representation of alveolars in coda position may also have inhibited the improvement of participants' abilities to identify more visually salient—but less well-represented—places of articulation in the coda consonants. Balancing the stimuli for place of articulation in future studies may help eliminate some of the confounding influences that the variability of ease of place identifiability may have on the participants' ability to improve in a visual-only speech perception task.

Although such possibilities for improving on the experimental paradigm in this study remain, its results did nonetheless demonstrate that participants could accurately identify individual words and phonemes from PLDs, and that their ability to perform this perceptual task improved over the course of a one-hour experiment. That human observers can perceive speech from only the impoverished, dynamic cues represented in these displays is quite remarkable; studying further just how much this ability may improve—through either feedback or practice—may shed further light on the importance of dynamic cues in visual-only speech perception, as well as the role that event-based information—in any sensory modality—might play in helping people perceive the significant articulatory events which produce the dynamic features of speech.

References

- CMU Pronouncing Dictionary, version 0.6. Available at <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>.
- Fowler, C.A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, *14*, 3-28.
- Fowler, C.A. & Dekle, D.J. (1991). Listening with eye and hand: cross-modal contributions to speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *17*, 816-828.
- Fowler, C.A. & Rosenblum, L.D. (1991). Perception of the phonetic gesture. In I.G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the Motor Theory*. Pp. 33-59. Hillsdale, NJ: Lawrence Erlbaum.
- Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, *14*, 201-211.
- Johnson, K. (2002). *Acoustic and Auditory Phonetics*. 2nd ed. Oxford: Blackwell.
- Lachs, L. (2002). *Vocal tract kinematics and crossmodal speech information*. (Research on Speech Perception Technical Report No. 10). Bloomington, IN: Speech Research Laboratory, Indiana University.
- Lachs, L., & Pisoni, D.B. (2004). Crossmodal source information and spoken word recognition. *Journal of Experimental Psychology: Human Perception & Performance*, *30*, 378-396.
- Lachs, L., & Pisoni, D.B. (submitted). Specification of crossmodal source information in isolated kinematic displays of speech. *Journal of the Acoustical Society of America*.
- Lachs, L. & Hernandez, L.R. (1998). Update: the Hoosier Audiovisual Multitalker Database. In *Research on Spoken Language Processing Progress Report No. 22* (pp. 377-388). Bloomington, IN: Speech Research Laboratory, Indiana University.
- McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746-748.

- Pashler, H., Cepeda, N., Wixted, J., & Rohrer, D. (In press). When does feedback facilitate learning of words and facts? *Journal of Experimental Psychology: Learning, Memory and Cognition*.
- Reed C.M., Durlach N.I., Braida L.D., & Schultz M.C. (1989) Analytic study of the Tadoma method: effects of hand position on segmental speech perception. *Journal of Speech and Hearing Research, 32*, 921-929.
- Rosenblum, L.D., Johnson, J.A. & Saldaña, H.M. (1996). Point-light facial displays enhance comprehension of speech in noise. *Journal of Speech and Hearing Research, 39*, 1159-1170.
- Rosenblum, L.D. & Saldaña, H.M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 22*, 318-331.
- Sheffert, S.M., Lachs, L. & Hernandez, L.R. (1996). The Hoosier Audiovisual Multitalker Database. In *Research on Spoken Language Processing Progress Report No. 21* (pp. 578-583). Bloomington, IN: Speech Research Laboratory, Indiana University.
- Sumby, W.H. & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26*, 212-215.
- Walden, B.E., Prosek, R.A., Montgomery, A.A., Scher, C.K. & Jones, C.J. (1977). Effects of training on the visual recognition of consonants. *Journal of Speech and Hearing Research, 20*, 130-145.

