

RESEARCH ON SPOKEN LANGUAGE PROCESSING

Progress Report No. 27 (2005)

Indiana University

**When and Why Feedback Matters in the Perceptual Learning of
Visual Properties of Speech¹**

Stephen J. Winters, Susannah V. Levi and David B. Pisoni

*Speech Research Laboratory
Department of Psychological and Brain Sciences
Indiana University
Bloomington, Indiana 47405*

¹ This work was supported by grants from the National Institutes of Health to Indiana University (NIH-NIDCD T32 Training Grant DC-00012 and NIH-NIDCD Research Grant R01 DC-00111). We would like to thank Christina Fonte, Jen Karpicke, Sara Phillips, and Melissa Troyer for their help in running subjects, constructing stimuli, and analyzing data.

When and Why Feedback Matters in the Perceptual Learning of Visual Displays of Speech

Abstract. This study investigated how feedback can be used to improve the perception of speech from visual-only displays. Participants saw English words spoken in two different, visual-only displays: full-face displays, in which a speaker's whole face is seen under normal lighting conditions, and point-light displays, which preserve only some of the dynamic information that is visible in speech. The participants attempted to identify words in one of these display formats, and then received feedback information about the identity of the words after each trial. Six different forms of feedback were provided to the participants. Participants who saw point-light displays improved most at the visual-only word identification task when they received feedback which re-presented the stimulus in its original, visual form; these participants also improved more rapidly when they received feedback in audio, rather than in orthographic form. However, the form in which feedback was presented to participants who saw full-face displays of speech did not have as strong an effect on their rate of perceptual improvement. In both display conditions, feedback only improved identification accuracy on stimuli which participants had seen before, without facilitating generalization to novel stimuli. These results suggest that the information contained in visual displays of speech is retained in memory in a highly-detailed, modality-specific format, and that observers draw upon this detailed, episodic information in memory in the process of perceptual learning.

Introduction

Formal linguistic theory (e.g., Chomsky & Halle, 1968) typically represents the phonological structure of language in highly abstract terms. These formal structures are assumed to represent the knowledge that an “ideal” speaker-hearer has of the sound structure of his or her language, and are thus independent of the particular physical system in which they may be manifested (Chomsky, 1965). Whenever we perceive speech in the course of everyday life, however, we perceive it in a particular set of idiosyncratic circumstances, as produced by a particular speaker, through a particular medium. The phonological structures underlying the speech that we hear therefore never come to us in an ideal form. Moreover, they are shaped in a wide variety of ways by the unique characteristics, or “indexical properties” of the speakers that we hear (Abercrombie, 1968). For this reason, it is often possible to identify certain personal characteristics of the speakers we hear—such as their age, their gender, their socio-economic or geographic background—as we interpret the linguistic content of what they are saying. Similarly, it is often possible to identify specific characteristics of the medium through which speech is transmitted to us—for example, over the telephone, over the radio, in a noisy (reverberant) room, etc. None of these “indexical properties” or medium-specific characteristics of the speech that we hear, though, are ever encoded into a formal description of phonological structure. Since these properties of speech cannot affect the meaning or content of a linguistic message, they have all traditionally been considered “extra-linguistic” properties of a speech signal (cf. Laver, 1994; Kreiman, VanLancker-Sidtis, & Gerratt, 2005).

It has often been assumed that these extra-linguistic details are discarded in the perception of speech, as listeners pare down what they hear to an essential linguistic core. For instance, Halle (1985) claimed that:

“...when we learn a new word we practically never remember most of the salient acoustic properties that must have been present in the signal that struck our ears; for example, we do not remember the voice quality, speed of utterance, and other properties directly linked to the unique circumstances surrounding every utterance.” (p. 101)

The process by which extra-linguistic details are filtered or removed from the speech signal in order to yield a formal, abstract and sparsely detailed linguistic representation in memory is known as “perceptual normalization”. (see Pisoni, 1997, for a review). A growing body of evidence suggests, however, that the perception of speech does not “normalize” away extra-linguistic details, but actually yields representations that include far more talker-dependent and medium-specific detail than is typically contained in a formal description of language (Pisoni & Levi, 2005). It has been shown, for instance, that listeners retain highly detailed information about the voice of the speaker in recognition memory experiments using spoken words (Palmeri, Goldinger, & Pisoni, 1993). Palmeri et al. showed that listeners can recognize words more quickly and accurately when they are re-presented to them in the same voice, rather than in a different voice. This talker repetition effect is robust across varying numbers of speakers and occurs even when listeners are not asked to attend to the voices of the speakers who are producing the words they hear. Similarly, Nygaard, Sommers, and Pisoni (1994) and Nygaard and Pisoni (1998) have shown that listeners can better identify words and sentences in noise if they are spoken by familiar talkers than if they are spoken by unfamiliar talkers. This finding indicates that listeners not only encode and store talker-specific information in memory when listening to speech, but also that they actively use this information when processing the semantic content of spoken messages. These findings are similar to earlier studies showing that readers store information in memory about the orientation and font face of typewritten words that they have read (Kolers, 1973).

On the basis of such evidence, some theoreticians (Johnson, 1997; Pierrehumbert, 2001) have proposed that specific experiences of speech and language are stored in a highly-detailed, medium-specific format in memory. This view of speech perception holds that normalization is unnecessary; instead, linguistic generalizations emerge during perception from the process of extracting meaningful information out of a wide variety of similar category “exemplars” in memory. Further evidence in support of these exemplar models of speech perception comes from Goldinger (1997), who studied the role of lexical frequency in a phenomenon he dubbed “spontaneous vocal imitation.” Goldinger reported that listeners who are asked to repeat words produced by other speakers will reflexively mimic the low-level acoustic characteristics of the words that they hear. That is, listeners’ repetitions of spoken words will more closely match the acoustic characteristics of the originals if the words are low in frequency. Goldinger hypothesized that this frequency effect occurs because the acoustic structure of specific word repetitions is based on a combination of what the listener hears and the aggregate average of the acoustic details of the listener’s experiences of that word in memory. The repetition of low frequency words, which have fewer exemplars in memory, will thus be more heavily influenced by the acoustic structure of the input signal than high frequency words.

While studies such as Goldinger (1997), Palmeri, Goldinger, and Pisoni (1993) and Nygaard, Sommers, and Pisoni (1994) have shown that the auditory processing of speech preserves highly detailed, “extra-linguistic” information in memory, much less is known about the extent to which human observers preserve extra-linguistic or episodic information in the processing of speech in the visual domain. It has been established in a variety of studies that normal-hearing listeners can extract some meaningful linguistic information from visual recordings of speech which completely lack acoustic information (Breeuwer & Plomp, 1986; Demorest & Bernstein, 1992). Sumbly and Pollack (1954) found that visual speech signals significantly augment the intelligibility of speech in adverse listening conditions, while McGurk and McDonald (1976) demonstrated that the visual cues to certain speech sounds may override

the audio cues to different speech sounds in the perception of audio-visually mismatched stimuli. The visual perception of speech thus appears to be highly robust and pervasive (Rosenblum, 2005).

Evidence has begun to emerge from recent work that talker- and token-specific details from particular productions of speech are preserved in the process of visual speech perception. Rosenblum, Yakel, Basser, Panchal, Nodarse and Niehus (2002) showed that observers can match talkers in visual-only speech stimuli across point-light and full-face display formats. Lachs and Pisoni (2004b,c) have reported that observers can match individual tokens of words across modalities and various acoustic transformations. The available evidence thus suggests that observers retain “extra-linguistic” characteristics of the visual signal in memory, just as they preserve the fine-grained acoustic-phonetic details of speech in memory during the process of perception.

Pilot Study

In an earlier pilot study (Winters & Pisoni, 2004), we investigated whether the extra-linguistic, modality-specific details of visual experiences of speech could be used to facilitate the perceptual learning of the visual properties of speech. We asked participants to identify isolated, monosyllabic English words from visual-only displays of speech and then provided them with feedback. This feedback information was presented to different groups of participants in one of three different formats. One group received audio-visual feedback, in which they saw the original, visual stimulus again, while simultaneously hearing the word that was spoken in the video. Another group of participants received audio-only feedback, in which they only heard the audio track from the original stimulus video. The third group of participants received orthographic-only feedback, in which an alphabetic display of the word that had been spoken in the original stimulus video was presented to them on the video screen. A fourth group of participants, in a control condition, received no feedback on their responses.

Prior to this study, we expected feedback to improve the participants’ ability to identify the words in each silent video, since feedback information would provide the participants with a linguistic interpretation of the speech events they had seen in the silent, visual displays. Furthermore, we hypothesized that improvement would be proportional to the amount of information provided in feedback to the participants about the speech events they saw in each video. In particular, we expected audio-visual feedback to improve participants’ identification accuracy more than audio-only or orthographic-only feedback because it re-presented the stimulus in a visual form that exactly matched the participants’ memory of their initial experience of that stimulus. We also expected audio-only feedback to improve identification accuracy more than orthographic-only feedback because the audio-only signal would more closely match the idiosyncratic, dynamic structure of the speech events in the original, visual-only stimulus. Orthographic-only feedback, on the other hand, would only provide the participants with a static, symbolic linguistic representation of the word which had been spoken, which would not provide the participants with any detailed information about the dynamics of the speech events in the original visual signal.

We quantified the amount of perceptual learning the participants made by comparing the participants’ accuracy in identifying whole words and sub-lexical units (such as phonemes) between the first and the second halves of the experiment. We found that accuracy did improve over the duration of the experiment, but that the amount of improvement in identification accuracy was almost always unrelated to the type of feedback the participants received. Statistically equivalent gains in whole word identification accuracy were made by all groups of participants—even those who received no feedback at all. The only significant effect of feedback type on perceptual improvement emerged in the identification of word-initial phonemes. However, the observed effects of feedback on identification accuracy in this

context did not match what had been predicted. Participants who received orthographic-only and audio-only feedback identified a higher percentage of word-initial phonemes correctly in the second half of the experiment than they did in the first half. The participants who received audio-visual feedback and no feedback, however, made no comparable gains in perceptual improvement.

We speculated that the specific type of feedback might not have affected whole word identification accuracy because none of the word stimuli were ever re-presented to participants after they had received feedback on them. Therefore, the participants could not apply what they had learned through feedback to the process of identifying the same words on subsequent experimental trials. The effect of feedback type on phoneme identification accuracy, on the other hand, may have emerged because certain phonemes were presented in more than one word in the experiment. Participants could therefore apply what they had learned through feedback about the visual properties of phonemes to the process of identifying those same phonemes on subsequent experimental trials.

The metric that was used to assess perceptual improvement in Winters and Pisoni (2004) may have actually obscured gains in improvement made during each half of the experiment itself. The audio-visual feedback group had a small, but not significant advantage in identification accuracy for word-initial phonemes over the other feedback groups. This advantage may have been the result of rapid perceptual learning during the first half of the experiment by the audio-visual feedback group. The audio-visual feedback group may also have been better at identifying word-initial phonemes at the beginning of the experiment than the other groups, independent of their ability to improve in word identification accuracy throughout the experiment.

Current Study

For the present study, we modified the experimental paradigm used in the pilot study in order to determine whether the expected effects of feedback on perceptual learning and the visual perception of speech would emerge under more relevant testing conditions. The visual-only word identification task remained the same in this study, but the number of experimental trials was expanded and split into three separate phases: pre-test, training, and post-test. In the pre-test, participants saw 16 video stimuli without receiving feedback. Performance in this pre-test thus provided a baseline measure for the inherent ability of each group of participants to do the visual-only word identification task. In the training phase, participants saw 64 videos and received feedback after each trial. Most of the videos they saw during training were also presented to them again, during training, after they had already received feedback on those videos. By re-presenting stimuli in this way, we enabled participants to apply what they had learned through feedback directly to the identification of the stimuli they had received feedback on. Recently, Pashler, Cepeda, Wixted, and Rohrer (2005) have shown that re-presenting test stimuli, after participants have received feedback on them, is an effective way of improving the identification of lexical items in a unfamiliar language; hence, we also expected observer identification accuracy to improve after repeated viewings of the same visual stimuli in training. Comparing identification accuracy after successive presentations of each stimulus in training also provided a more objective means by which to gauge the effects of feedback on perceptual improvement in the task than did the arbitrary first half/second half split that had been used in the pilot study. Finally, in the post-test phase, participants saw 16 new videos without receiving feedback on any of them. The structure of the post-test was thus identical that of the pre-test. Participant performance in the post-test could thus be directly compared to their performance in the pre-test to gauge how well the participants had improved in lip-reading accuracy over the course of the experiment. Comparing identification accuracy between pre-test and post-test phases also provided a direct measure of whether any gains in identification accuracy which had been made during training would generalize to novel video stimuli.

Along with the three forms of feedback which were used in the pilot study—audio-visual (AV), audio-only (A) and orthographic-only (O) feedback—participants in this investigation also received feedback in three new forms which combined a simultaneous or sequential presentation of the visual signal with either audio or orthographic information. These new forms of feedback were included in order to provide an equitable means of testing the effects of combining audio and orthographic feedback with visual information on perceptual learning. In the orthographic-visual (OV) feedback condition, orthographic and visual information were simultaneously presented to the observers by superimposing an orthographic representation of the spoken word on the silent visual stimulus. In the sequential feedback conditions, observers first received information about the identity of the spoken word through either an acoustic-only signal (A-then-V feedback) or an orthographic-only presentation of the word on the computer screen (O-then-V feedback) prior to viewing the silent video stimulus again. With these six forms of feedback, the effects of dynamic audio feedback on perceptual learning could be directly compared to the effects of static orthographic feedback along three separate dimensions, two of which involved a re-presentation of the original video stimulus. A summary of these feedback conditions is provided in Table 1.

	Audio	Orthographic
Simultaneous feedback	AV	OV
Sequential feedback	A-then-V	O-then-V
Non-visual feedback	A	O

Table 1. Summary of feedback types.

As in Winters and Pisoni (2004), we expected that feedback would not only improve observers' identification accuracy for visual stimuli on repeated presentations during training, but that certain types of feedback would improve identification accuracy more than others. When observers see a stimulus that they have seen before on a previous trial, they can use what they have learned about the linguistic properties of that stimulus through feedback to help them identify its linguistic content on the repeated presentation. The ability of observers to do this, however, will depend on how much feedback information they encode and store in memory. If observers store all the modality-specific details that they see during feedback (e.g., the visual properties of the spoken word that they have seen in audio-visual feedback, or the dynamic spectral properties of speech that they have heard in audio-only feedback), then feedback which shares more features in common with the visual-only stimuli should improve identification accuracy more than feedback which does not. However, if such modality-specific detail is discarded in the perceptual analysis of the visual or auditory properties of speech, then the type of feedback the observers receive should not affect how much perceptual improvement observers make in the visual-only word identification task. Performance should only improve if they receive some kind of feedback, regardless of the form in which it is presented to them.

By hypothesizing that observers do not discard modality-specific, "extra-linguistic" details of visible speech tokens from memory, we expected that visual feedback would improve identification accuracy more than non-visual feedback. We also expected that audio feedback would facilitate perceptual learning better than orthographic feedback, since audio feedback matches the dynamic information in the visual speech signal while orthographic feedback does not. It was unknown whether sequential feedback would facilitate perceptual learning better than simultaneous feedback. However, we had a priori reasons for expecting that OV feedback would not facilitate perceptual learning as well as AV feedback, because observers must divide their visual attention in attempting to perceive both an

orthographic and a visual representation of a word at the same time. Observers do not need to divide their attention between modalities when either audio or orthographic feedback is presented in sequence with a repetition of the visual speech signal. The sequential feedback conditions were therefore expected to provide a clearer test of the effects of presenting dynamic (audio) vs. non-dynamic (orthographic) feedback, in conjunction with the original visual signal, on observer accuracy in the visual-only word recognition task.

This study investigated the perceptual learning of the visual properties of speech by using two different kinds of visual displays: full-face displays and point-light displays. Full-face displays of speech present a speaker's face under normal, visible lighting conditions. It has been known since Sumbly and Pollack (1954) that normal-hearing observers can readily extract meaningful linguistic information from full-face displays of speech. It has also been shown in a wide variety of studies that the ability of observers to perceive speech in visual-only full-face displays improves over the course of a short training experiment, especially if the participants receive feedback (Bernstein, Auer, & Tucker, 2001; Black, O'Reilly, & Peck, 1963; Gesi, Massaro, & Cohen, 1992; Massaro, Cohen, & Gesi, 1993; Massaro & Light, 2004; Walden, Erdman, Montgomery, Schwartz, & Prosek, 1981; Walden, Prosek, Montgomery, Scher, & Jones, 1977).

Point-light displays are animated sequences of illuminated dot patterns (Johansson, 1973). Figure 1 shows two example frames from a point-light display of a person executing a placekick.



Figure 1. Example frames from a point-light display of a person executing a placekick.

Such point-light displays may be constructed by attaching luminescent points to the major joints on a person's body (e.g., shoulders, elbows, wrists, hips, knees, ankles) and then filming that person executing some motion under darkened lighting conditions. Johansson found that observers could identify human motions in point-light videos made in this way, even though the point-light videos contained much less information than fully illuminated videos of the same motions. Point-light displays of speech were first constructed by Summerfield (1979), who investigated whether they contained enough information to support word identification in adverse listening conditions. Summerfield's point-light displays consisted of only four luminescent points which had been attached to a talker's lips in video-recordings of speech made under darkened lighting conditions. Summerfield presented these point-light displays to observers in conjunction with acoustic speech signals in noise, but found that they did not significantly increase the intelligibility of words over a control condition in which listeners saw no visual information whatsoever. Rosenblum, Johnson, and Saldana (1996), however, found that point-light displays that were made with more than four point-lights in the configuration did improve the intelligibility of speech in noise.

Furthermore, Rosenblum and Saldana (1996) demonstrated that point-light displays also induce McGurk-like effects in audio-visually mismatched tokens of speech. Both Rosenblum, Johnson, and Saldana (1996) and Winters and Pisoni (2004) have also shown that the perception of speech in point-light displays improves rapidly over the course of a short experiment. The visual perception of speech in point-light displays thus exhibits the same basic properties as the visual perception of speech in full-face displays, despite the fact that point-light displays contain much less visual information than fully illuminated displays of speech. It is unknown, however, whether feedback can facilitate the perceptual learning of the visual properties of speech in point-light displays, as it does for full-face displays of speech. Experiment 1 reports the results of using the proposed experimental paradigm to test the effects of feedback on the perceptual learning of the visual properties of speech in point-light displays, while Experiment 2 reports the results of using the same paradigm with full-face displays of speech.

Experiment 1: Perceptual Learning of Point-light Displays of Speech

Methods

Participants. Participants were introductory psychology students at Indiana University in Bloomington, Indiana. A total of 147 subjects participated in the study; seven were removed from analysis (one because of computer failure, one because of self-reported hearing impairment, one who was bilingual, and four because they did not provide responses), resulting in twenty participants in each of the six feedback conditions and twenty in the control condition. All participants were between the ages of 18 and 25, native speakers of English, with normal or corrected-to-normal vision and no reported hearing or language deficits at the time of testing. None of the participants had any previous experience with the audio-visual speech stimuli used in this experiment. All participants received partial course credit for participation in the experiment.

Materials. The point-light displays of speech that were used in this experiment were selected from a digital database originally created by Lachs and Pisoni (2004a). A single talker produced all stimuli. In each video, the talker read one of 96 English words of the form consonant-vowel-consonant (CVC) (e.g., “base”). The talker was video-recorded with glow-in-the-dark dots attached to her face, under black light illumination. The dots were each approximately 3 mm in diameter and were attached to the talker’s face in the pattern shown in Figure 2. There were five dots on each cheek, one on the nose, two on the chin, four on the lower edges of the lips, four on the outer edges of the lips, two on the corners of the lips, one on the tip of the tongue, and two dots each on the lower and upper rows of teeth. Figure 3 shows an example from one of the finished point-light videos.



Figure 2. Configuration of point-lights

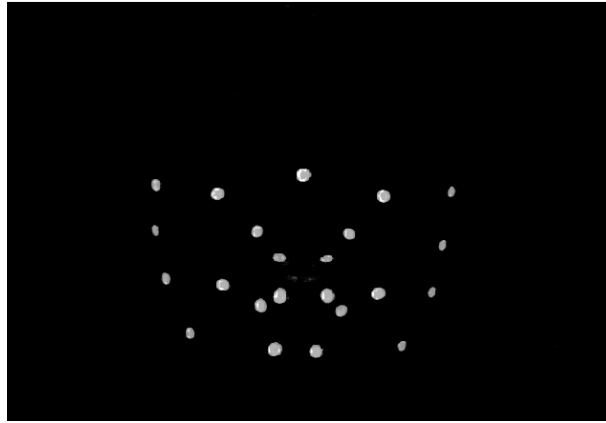


Figure 3. Example frame from point-light video

During pilot-testing, several observers complained that the motion of these point-light displays began before they could orient themselves to the pattern of lights they saw on the screen. The difficulty of interpreting the point-light displays was alleviated by extending the first frame of each video by 500 milliseconds.

For the audio-only feedback condition, the audio track from each video was saved as an .AIFF file using QuickTime software. For the simultaneous orthographic-visual feedback condition, a duplicate set of point-light videos were constructed, using FinalCut Pro Software, in which the spoken word appeared, in lucida grande font face, size 36, centered just beneath the speaker's chin, for the duration of the audio signal in the original video.

Procedure. The experiment was implemented on a customized SuperCard (version 4.1.1) stack, running on a Macintosh G3 computer in a quiet testing room. Participants sat in front of a computer while wearing Beyer Dynamic DT-100 headphones. Their primary task was to watch individual, silent point-light videos and then identify the word that was spoken in each video. After the presentation of each video, the participants answered the on-screen question, "What did the speaker say?" by typing in a response on the computer keyboard.

A brief description of how the point-light stimuli were created was given to the participants prior to the experiment. The participants were told that it might be difficult to perceive speech in the point-light videos, but that they should always provide a response to each video on every trial, even if they had to guess.

The experiment consisted of three phases: pre-test, training and post-test. In the pre-test, the participants saw 16 different point-light videos and responded to each. The participants received no feedback during this portion of the experiment, and none of these videos were shown again during the study.

During training, the participants saw 64 different point-light stimuli. 16 of these stimuli were presented eight times, 16 were presented four times, 16 were presented twice, and 16 were presented only once. The participants received feedback after they responded to each video during training. After the

presentation of feedback, participants clicked on an on-screen button to move on to the next trial. None of the videos which were presented during the training session had been presented during the pre-test.

The experiment used a between-subjects design, so that each group of participants received only one form of feedback each during training. Details of the method used to provide each form of feedback to the various groups of participants are summarized below in Table 2.

Feedback Form	Method of Presentation
AV	Participants saw the original video stimulus again and also heard the original audio track from the test stimulus.
A-then-V	Participants first heard the audio track from the original stimulus and then, after a 500 millisecond pause, they saw the silent, video test stimulus again.
A	Participants heard the audio track from the test stimulus, without seeing the visual stimulus again.
OV	Participants saw a simultaneous presentation of the original, silent visual stimulus and an orthographic representation of the spoken word.
O-then-V	Participants saw the word which had been spoken in text on the computer screen for 1000 milliseconds, and then, after a 500 millisecond pause, saw the silent, visual stimulus again.
O	Participants saw the word which had been spoken in text on the computer screen for 1000 milliseconds, without seeing the visual stimulus again.
N	Participants received no feedback.

Table 2. Method of presenting each type of feedback during training.

The form of the post-test was identical to that of the pre-test. The participants saw 16 different point-light videos and responded to each. The participants received no feedback during this portion of the experiment, and none of the videos which were presented during the post-test had been presented during either training or in the pre-test.

Videos were presented to the participants in random order, with the restriction that no video was ever shown on two consecutive trials during training. The groups of videos which were selected for the pre-test, post-test, and the four different presentation groups in training were also selected at random for each participant. Most participants completed the experiment within 45 minutes.

Analysis. The participants' responses were scored using three levels of analysis: whole word, phoneme, and viseme. A "viseme" denotes a visually equivalent class of sounds (Walden, Prosek, Montgomery, Scher, & Jones, 1977); for instance, the bilabials /b/, /p/ and /m/ belong to the same category. All stimuli and responses were converted into phonetic transcriptions by matching them with entries in the Carnegie Mellon pronouncing dictionary, which lists transcriptions for English words in ARPA notation. Every word in the original video stimuli had a consonant-vowel-consonant form, so the phonetic transcriptions in the dictionary for each match were segmented into an "onset", a "nucleus" and

a “coda.” The vowel in each stimulus word formed its “nucleus”, while the initial consonant was its “onset” and the final consonant was its “coda.”

Even though all participants were informed, prior to the experiment, that they would only see monosyllabic words in each video, many of their responses contained more than one syllable. The “nucleus” of all participant responses—no matter how many syllables they contained—was therefore taken to be the vowel with the highest stress level in the response word or phrase. All segments—including any consonants or vowels—which preceded this response “nucleus” were then taken to be the “onset” of the response, while all segments which followed 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/. The /ae1/ vowel has the highest stress level in the word, so it formed the “nucleus” of the response. The /k/ which preceded it then formed the response “onset,” while the final /m ax0 r ax0/ sequence formed the “coda.”

A response was scored correct at the whole word level if the phonetic transcription of its onset, nucleus and coda matched the corresponding transcriptions of the stimulus onset, nucleus and coda. Homonyms (e.g., “wear” and ‘ware’’) were thus considered to be correct identifications of whole stimulus words. At the phoneme level of analysis, response onsets, nuclei, and codas 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.

The onset and coda of all stimuli and responses were also classified by viseme. The different viseme types included bilabials (/p/, /b/, /m/), labio-dentals (/f/, /v/), interdental (/θ/, /ð/), dorso-linguals (/t/, /d/, /n/, /k/, /g/, /ŋ/), palato-alveolars (/ʃ/, /ʒ/), and separate categories for /s/, /r/, /h/, /l/, and /w/ (Walden, Prosek, Montgomery, Scher, & Jones, 1977). (Corresponding viseme categories for vowels have not been defined.) Those response onsets and codas which contained more than one segment were classified as having a “mixed” viseme type—unless all of the segments in those onsets and codas happened to agree in viseme type. 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 viseme type of the response onsets and codas were only counted as “correct” identifications if they exactly matched the corresponding viseme features of the stimulus. One participant, for instance, gave the response “damp” to the “dame” stimulus. In “damp,” the coda /mp/ was classified as having a bilabial viseme type, since both /m/ and /p/ are bilabial consonants. This was scored as a correct identification of the stimulus coda viseme, since the coda /m/ in “dame” is also a bilabial. Another participant, however, identified the same “dame” stimulus as “table.” Since the coda of “table” includes both /b/ and /l/ segments, which have a bilabial and a lateral viseme classification, respectively, the coda was categorized as having a “mixed” viseme type. This response was therefore scored as an incorrect identification of the bilabial viseme type in the stimulus coda /m/.

Many of the participants’ responses could not be matched to any entry in the CMU pronouncing dictionary. Responses that were obvious misspellings (e.g., “cheif”) were corrected in the original data file and then matched with the corresponding dictionary entry, while responses that were not obviously

English words (e.g., “rith”) were given onset-nucleus-coda transcriptions by hand and then scored accordingly.

Results

Analyses of variance (ANOVAs) were run on the percentages of whole words, phonemes and visemes correctly identified by the participants in order to determine the effects that testing session, feedback type and repetition number had on participants’ response accuracy. Three separate ANOVAs were run for all three levels of analysis (words, phonemes, visemes): one comparing participant performance in pre- vs. post-test, another analyzing participant performance on the initial presentation of each group of stimuli during training, and another analyzing participant performance on the final presentation of each group of stimuli during training. Essentially the same pattern of effects on identification accuracy emerged from the separate ANOVAs at the three different levels of analysis, so only the results of the whole word ANOVAs will be reported here, since this is the linguistic level at which participants entered their responses and at which feedback was given.

Pre- vs. Post-test. A repeated measures ANOVA with testing session (pre-test vs. post-test) as a within-subjects factor and feedback condition (AV, A-then-V, A, OV, O-then-V, O, N) as a between-subjects factor revealed a significant main effect of testing session ($F(1,132) = 22.634; p < .001$) but no main effect of feedback. The percentage of words correctly identified in the post-test (3.9%) was significantly higher than the percentage of words correctly identified in the pre-test (1.4%). There was no significant interaction between feedback condition and test session.

Training: Initial Presentation. In order to establish a baseline to measure the effects of stimulus repetition during training on participant response accuracy, a two-way repeated measures ANOVA was run using the percentages of words correctly identified on the initial presentation of each point-light stimulus in training as a dependent variable, feedback condition (AV, A-then-V, A, OV, O-then-V, O, N) as a between-subjects factor and presentation group (one, two, four or eight) as a within-subjects factor. Presentation group number was included as a factor in the ANOVA in order to establish that there were no pre-existing differences in ease of identifiability between the words in each group of stimuli prior to their repetition during training. This ANOVA failed to reveal any significant main effects for feedback or presentation group on whole word identification accuracy. There was also no significant interaction between these two factors. None of the presentation groups thus contained words with inherent differences in intelligibility.

Training: Final Presentation. A repeated measures ANOVA was run using the percentages of words correctly identified on the final presentation of each stimulus during the training phase as a dependent measure in order to determine what effects feedback type and stimulus repetition had on participants’ response accuracy. The independent factors in this ANOVA included presentation number (one, two, four, eight) as a within-subjects factor and feedback type (AV, A-then-V, A, OV, O-then-V, O, N) as a between-subjects factor. This analysis revealed significant main effects of both presentation number ($F(3,130) = 127.193; p < .001$) and feedback type ($F(6,132) = 10.248; p < .001$).

Table 3 shows the percentage of words correctly identified during training on the final presentation of words in each presentation group. Paired samples t-tests on the main effect of presentation number showed that participants identified words more accurately on the eighth presentation than they did on the fourth, second and first presentations (all $p < .001$). Likewise, they identified more words correctly on the fourth presentation than they did on the second and first presentations (both $p <$

.001), and they identified a significantly higher percentage of words correctly on the second presentation than they did on the first ($p < .001$).

Presentation	% Correct
1	3.2%
2	8.2%
4	15.0%
8	25.6%

Table 3. Percentage of words correctly identified, by presentation group.

Table 4 lists the percentages of words correctly identified by participants in each feedback condition during training. Post-hoc Tukey tests on the main effect of feedback condition indicated that participants in the N feedback group identified fewer words correctly than participants in all of the other feedback groups (AV, A-then-V, O-then-V: $p < .001$; O: $p = .009$; OV: $p = .01$; A: $p = .022$). The AV and A-then-V feedback groups also correctly identified a significantly higher percentage of words than participants in the A feedback group ($p = .037$ in both cases). Comparisons between all other groups yielded no significant differences in word identification accuracy.

Video Presentation	A	O	N
Simultaneous	18.1%	11.8%	---
Sequential	18.1%	16.4%	---
None	11.1%	11.7%	3.7%

Table 4. Percentage of words correctly identified, by feedback condition.

The repeated measures ANOVA also yielded a significant interaction between presentation number and feedback condition ($F(18,396) = 3.252$; $p < .001$). Figure 4 shows the percentages of words correctly identified by participants in each feedback condition, on the final presentation of each word during training. Post-hoc Tukey tests on the interaction between feedback group and presentation number revealed that, on the second presentation of each stimulus, the AV and A-then-V groups were significantly more accurate than the N feedback group ($p = .004$ and $p = .012$, respectively). On the fourth presentation of each stimulus, the AV, A-then-V and O-then-V groups were significantly more accurate than the N feedback group ($p < .001$, $p < .001$ and $p = .001$, respectively). All groups receiving feedback were significantly more accurate than the N feedback group on the eighth presentation of each stimulus (AV, A-then-V, O-then-V: $p < .001$; A: $p = .002$; OV: $p = .003$; O: $p = .001$). Comparisons between all other feedback groups, for all presentation numbers, yielded no significant differences.

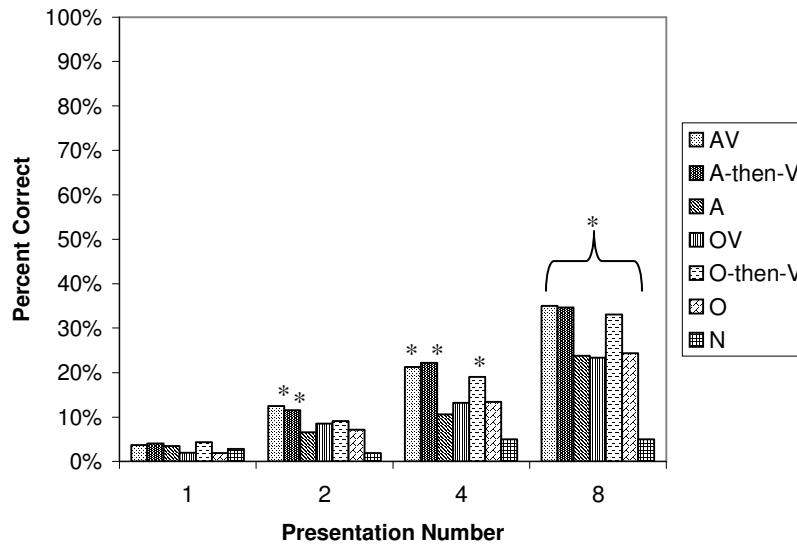


Figure 4. Percentage of whole words correctly identified, by feedback condition and presentation number, on final presentation in training. (* denotes that identification accuracy in that feedback condition was significantly higher at the $p=.05$ level than identification accuracy in the N feedback condition, after an equivalent number of stimulus presentations.)

Discussion

Two kinds of perceptual learning emerged from the results of this study—one which relied on the type of feedback that participants received, and one which relied on practice. The perceptual learning that relied on practice did not depend on feedback type and took place between pre-test and post-test. All groups of participants, regardless of whether or not they received feedback during training, improved in their ability to identify words from visual-only speech stimuli between pre-test and post-test. As even the group of participants which did not receive any feedback improved in identification accuracy between pre- and post-test, this form of perceptual learning seems to be just a generalized practice effect—the result of acclimating to the experimental procedure and task.

The form of perceptual learning which did rely on feedback type emerged during the training session. The results from training consistently showed two broad trends in identification accuracy: performance improved with successive re-presentations of the same stimuli, and performance was also higher for the groups which received feedback than it was for the group which received no feedback. These two trends also interacted in that several feedback groups performed significantly better than the group which received no feedback after fewer repetitions of the same stimuli.

Taken together, these results indicate that participants were able to use what they had learned through feedback to improve their ability to identify stimuli they had seen before on previous training trials. In other words, simply re-presenting stimuli to participants after they have received feedback on them enabled that feedback information to improve identification accuracy. This effect of feedback did not emerge in the earlier pilot study by Winters and Pisoni (2004) because none of the stimuli were ever presented more than once to participants. Within the confines of a short-term learning experiment, it is clear that feedback only improves participants' ability to identify the stimuli for which they have received feedback. For this reason, the differential effects of feedback on identification accuracy which

emerged in the training session did not generalize to any of the novel stimuli the participants saw in the post-test.

Even when participants attempt to identify stimuli they have seen before and have already received feedback on, they show greater improvement in identification accuracy with some types of feedback than others. In general, the participants who received feedback that re-presented the original point-light stimulus in visual form improved more rapidly in word identification accuracy than did the groups of participants who received only audio or orthographic feedback during training. For instance, the AV and A-then-V feedback groups displayed significantly better word identification accuracy than the N feedback group after the second presentation of the same word during training. The O-then-V feedback group attained the same level of performance after the fourth presentation of the same word in training. The A and O feedback groups, however, did not identify a significantly higher percentage of words than the N feedback group until the eighth presentation of repeated words in training. The group of participants who received OV feedback did not improve in word identification more rapidly than the A and O feedback groups because of the expected difficulties in dividing visual attention between the simultaneous orthographic and visual presentation of the target word.

The pattern of results observed in this study suggests that participants preserved in memory the fine-grained visual details of what they saw during feedback, and they were able to use this information to facilitate their perception of identical visual stimuli on subsequent training trials. When participants do not receive any visual feedback—as in the orthographic-only and audio-only feedback conditions—they cannot directly apply what they learn during feedback to the perception of identical stimuli on subsequent training trials. In order for participants to improve their identification accuracy in these conditions, they must learn to interpret the visual-only stimuli they see on each training trial in terms of the audio or orthographic representations of feedback they have in memory. Since this is more difficult than simply using visual representations in memory to interpret visually identical test stimuli, participants in the A or O feedback conditions did not improve as quickly in identification accuracy during training as the participants who received visual feedback.

We suggested earlier that audio feedback would improve identification accuracy more than orthographic feedback, because audio and visual representations of spoken words share dynamic features which static, orthographic representations do not preserve. The results of the present study only confirmed this hypothesis in the visual feedback conditions. The A-then-V and AV feedback groups both identified a significantly higher percentage of words after fewer stimulus repetitions in training than did the O-then-V and OV feedback groups. There was, however, no difference in the time-course of perceptual improvement between the A and O feedback groups. The fact that audio and visual representations of spoken words share dynamic properties thus only affected perceptual learning when the spoken word was presented to participants in both modalities during feedback. Observers, that is, are apparently only able to use the dynamic information in audio feedback to improve their perception of visual-only stimuli when the dynamic connection between audio and visual representations of speech is explicitly shown to them in feedback. Dynamic information may not affect perceptual learning in the non-visual feedback conditions because it is more difficult for observers to notice the shared dynamic properties of audio and visual representations of speech when there is a longer lag between the presentation of audio feedback and the re-presentation of the original visual stimulus.

The results of this study also showed that, in certain circumstances, sequential feedback facilitated perceptual learning better than simultaneous feedback. The sequential presentation of orthographic and visual (O-then-V) feedback consistently improved identification accuracy more rapidly than the simultaneous presentation of orthographic and visual (OV) feedback. This result is confounded,

however, by the aforementioned difficulties that the simultaneous presentation of two different types of visual information present in OV feedback. No differences emerged in the rate of perceptual improvement between the sequential A-then-V and simultaneous AV feedback groups in whole word identification accuracy, although the A-then-V feedback group did improve more quickly than the AV feedback group in both phoneme and viseme identification accuracy.² These results thus provide only limited evidence confirming the efficacy of providing feedback in sequential form on perceptual learning.

Comparing the rate of improvement across different levels of analysis—whole word identification accuracy, phoneme identification accuracy and viseme identification accuracy—revealed very few differences between the various feedback groups. For example, the AV feedback group correctly identified a significantly higher percentage of words than the N feedback group after only two presentations of words in training, but only reached the same level of performance in phoneme and viseme identification accuracy after four presentations of words in training. Other than minor differences such as these, most feedback groups progressed in identification accuracy at comparable rates, regardless of whether their responses were scored in terms of whole word, phoneme or viseme accuracy. The fact that there are no substantial differences in improvement between the whole-word and sub-lexical levels suggests that participants are processing stimuli and making use of feedback on a relatively holistic level, rather than building up their knowledge of the visual properties of speech from smaller perceptual units at the segmental or featural levels. The fact that the participants received more feedback and experience with the various phoneme and viseme categories, in a wider variety of phonological environments, than they did with whole words during training may make the absence of stronger learning effects at the sub-lexical level seem surprising. However, a pattern of learning suggesting that participants might have perceived the visual-only speech stimuli in a holistic fashion echoes earlier findings that the visual perception of speech is largely a holistic, top-down process (Heider & Heider, 1940). The top-down nature of the perceptual learning of the visually degraded point-light stimuli in this experiment provides converging evidence for Davis, Johnsruide, Hervais-Adelman, Taylor, & McGettigan's (2005) finding that the perceptual learning of noise vocoded speech depends on access to top-down lexical information.

In summary, the results of Experiment 1 indicate that participants preserve more than just formal, abstract and symbolic linguistic structures in memory when they perceive speech in the visual domain. The observers of visual-only speech stimuli in this study preserved in memory modality-specific details of the information they received in feedback, and used that information to help identify the linguistic content of the same stimuli when they were re-presented on subsequent trials in training. Observers do not discard these modality-specific details in the visual perception of speech through some sort of perceptual normalization process. It is for this reason that feedback which re-presents stimuli in their original, visual form facilitates perceptual learning better than other forms of feedback. The results of Experiment 1 also showed, however, that this form of perceptual learning does not generalize to novel stimuli; the observers in this study improved their perception of novel visual-only speech stimuli through practice alone.

Experiment 2: Perceptual Learning of Full-Face Displays of Speech

Experiment 1 showed that feedback facilitates the perceptual learning of the visual properties of speech in point-light displays, which are unusual, highly degraded visual displays of speech. Experiment 2 investigated whether the same forms of feedback would affect the perceptual learning of the visual

² A-then-V participants identified a significantly higher percentage of phonemes and visemes than the N feedback group after only two presentations of the same stimuli in training ($p = .014$ and $p = .044$, respectively), but the AV feedback group did not reach the same level of performance until the fourth presentation of the same stimuli in training ($p = .012$ and $p = .008$, respectively).

properties of speech in full-face displays in the same way. Full-face displays of speech differ from point-light displays in two important ways: first, they present more information about speech, and second, observers have more experience perceiving full-face displays of speech than they do perceiving point-light displays of speech. For both of these reasons, we expected participants' ability to perceive speech in full-face displays to be significantly better than their ability to perceive speech in point-light displays.

Based on the results of previous research, we also expected that feedback would improve observers' perception of visual-only, full-face displays of speech. Black, O'Reilly, and Peck (1963), Massaro, Cohen, and Gesi (1993) and Bernstein, Auer, and Tucker (2001) have all shown that orthographic-only feedback improves the visual-only perception of speech. Gesi, Massaro, and Cohen (1992) have shown that audio-visual feedback also improves the visual-only perception of speech. Other studies, such as Walden, Prosek, Montgomery, Scher, and Jones (1977), Walden, Erdman, Montgomery, Schwartz, and Prosek (1981) and Massaro and Light (2004), have provided feedback by informing their participants whether or not their responses in a visual-only speech perception task were correct or incorrect—and then repeated the presentation of those same stimuli until the participants responded correctly. This form of feedback also significantly improved observers' perception of visual-only speech stimuli. While all of these forms of feedback improve the visual-only perception of speech in full-face displays, it is unknown whether some forms of feedback might improve visual-only speech perception more than others, as was shown for point-light displays of speech in Experiment 1.

Methods

Participants. Participants were drawn from the same pool of subjects as in Experiment 1 and met the same criteria for inclusion. A total of 143 subjects participated in the study. Three were not included in the analysis of the response data (two for self-reported hearing impairment, one for a bilingual language background), thus resulting in twenty participants in each of the seven feedback conditions.

Materials. Full-face visual stimuli were selected from the Hoosier Audiovisual Multi-Talker Database (Lachs & Hernandez, 1996). This database consists of digitized videos of ten different talkers (five males and five females) producing 300 CVC English words under normal lighting conditions. Only stimuli produced by one (female) talker were included in Experiment 2; the words produced in those videos were identical to the list of 96 CVC words produced in the point-light videos in Experiment 1. An example frame from one of these videos is shown in Figure 5.



Figure 5. Example frame from a full-face video.

To parallel the presentation of the point-light videos in Experiment 1, the first frame of each full-face video was extended by 500 milliseconds at the beginning of the clip. For the OV feedback condition, a duplicate set of full-face videos were constructed using FinalCut Pro Software in which the spoken word appeared in lucida grande font face, size 36, centered just beneath the speaker's chin, for the duration of the audio signal in the original full-face video. For the audio-only feedback condition, audio files were constructed by simply saving the audio track of each full-face video to an .aiff file using QuickTime software.

Procedure. The procedures for Experiment 2 were identical to those used for Experiment 1. Participants were encouraged to provide a response to each stimulus, even if they were not sure what word had been spoken in the full face video.

Analysis. The analysis of participant responses to the fully illuminated videos in Experiment 2 was identical to the analysis of responses in Experiment 1. This process thus yielded correct identification percentages for whole words, phonemes and visemes, in each of the three parts of the experiment.

Results

Analyses of variance (ANOVAs) were run on the percentages of whole words, phonemes and visemes correctly identified by the participants in order to determine what effects testing session, feedback type and repetition number had on the participants' response accuracy. Three separate ANOVAs were run for all three sets (words, phonemes, visemes) of response accuracy data: one comparing participant performance in pre- vs. post-test, another analyzing participant performance on the initial presentation of each group of stimuli during training, and another analyzing participant performance on the final presentation of each group of stimuli during training. Once again, the same pattern of results emerged from the separate ANOVAs at the different levels of analysis. Therefore, only the results of the whole word ANOVAs are reported here.

Pre- vs. Post-test. A repeated measures ANOVA with testing session (pre-test vs. post-test) as a within-subjects factor and feedback condition (AV, A-then-V, A, OV, O-then-V, O, N) as a between-subjects factor revealed a significant main effect of testing session ($F(1,133) = 4.567; p = .034$) but no main effect of feedback. The percentage of words correctly identified in the post-test (24.1%) was significantly higher than the percentage of words correctly identified in the pre-test (21.3%). There was no significant interaction between feedback condition and test session.

Training: Initial Presentation. A two-way repeated measures ANOVA was run using the percentage of words correctly identified on the initial presentation of each point-light stimulus in training as a dependent variable and both feedback condition (AV, A-then-V, A, OV, O-then-V, O, N) and presentation group (one, two, four or eight) as independent factors. The results of this ANOVA did not reveal any significant main effects for feedback type or presentation group. There were also no significant interactions between these two factors. None of the presentation groups in Experiment 2 thus contained words with inherent differences in intelligibility.

Training: Final Presentation. A repeated measures ANOVA was run using the percentage of words correctly identified on the final presentation of each stimulus as a dependent measure in order to determine what effects feedback type and stimulus repetition had on participants' response accuracy. The independent factors in this ANOVA were presentation group number (one, two, four, eight), a within-

subjects factor, and feedback type (AV, A-then-V, A, OV, O-then-V, O, N), a between-subjects factor. This ANOVA revealed significant main effects of both presentation number ($F(3,131) = 248.461; p < .001$) and feedback type ($F(6,133) = 12.496; p < .001$).

Table 5 shows the percentages of words correctly identified during training on the final presentation of words in each presentation group. Paired samples t-tests on the main effect of presentation number showed that participants identified words more accurately on the eighth presentation than they did on the fourth, second and first presentations (all $p < .001$). Likewise, they identified more words correctly on the fourth presentation than they did on the second and first presentations (both $p < .001$), and they identified a significantly higher percentage of words correctly on the second presentation than they did on the first ($p < .001$).

Presentation	% Correct
1	21.9%
2	34.5%
4	49.3%
8	59.6%

Table 5. Percentage of whole words correctly identified, by presentation number.

Table 6 lists the percentages of words correctly identified during training by participants in the different feedback conditions. Post-hoc Tukey tests on the main effect of feedback condition indicated that participants in the N feedback group identified fewer words correctly than participants in all of the other feedback groups ($p < .001$ in all cases). Comparisons between all other groups yielded no significant differences in word identification accuracy.

Video Presentation	A	O	N
Simultaneous	45.7%	45.5%	---
Sequential	41.7%	44.2%	---
None	42.2%	46.4%	23.5%

Table 6. Percentage of whole words correctly identified, by feedback condition.

The repeated measures ANOVA also revealed a significant interaction between presentation number and feedback condition ($F(18,399) = 5.194; p < .001$). Figure 6 shows the percentages of words correctly identified by participants in each feedback condition, on the final presentation of each word in each presentation group, during training. Post-hoc Tukey tests on the interaction between feedback group and presentation number revealed that, on the second presentation of each stimulus, all groups receiving feedback except for the A group correctly identified a significantly higher percentage of words than the N feedback group (O: $p < .001$; AV: $p = .001$; O-then-V: $p = .007$; OV: $p = .017$; A-then-V: $p = .038$). On the fourth and eighth presentations of words, all feedback groups identified a significantly higher percentage of words than the N feedback group ($p < .001$ in all cases). No comparisons between any other feedback groups, for all presentation numbers, yielded significant differences in performance.

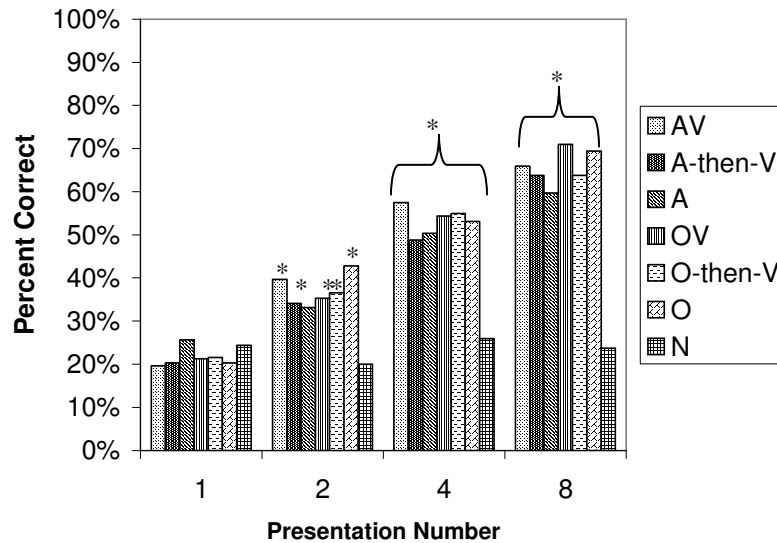


Figure 6. Percentage of whole words correctly identified, by feedback condition and presentation number, on final presentation in training. (* denotes that identification accuracy in that feedback condition was significantly higher at the $p=.05$ level than identification accuracy in the N feedback condition, after an equivalent number of stimulus presentations.)

Discussion

Taken together, the results obtained in Experiment 2 were quite similar to the findings obtained in Experiment 1. Participants consistently improved in their ability to identify the words, phonemes and visemes in the full-face stimuli between pre- and post-test, regardless of which type of feedback they received. Participants were also able to improve their perception of novel full-face stimuli through practice alone, even when they received no feedback during training at all. Those participants who did receive feedback during training, however, accurately identified a significantly higher percentage of words in repeated presentations of the same stimuli than did participants who received no feedback. The participants in Experiment 2 were thus able to use the information they received in feedback to improve their perception of stimuli they had seen before on previous trials. As in Experiment 1, however, feedback only improved the ability of participants to identify stimuli they had seen before and did not generalize to the novel stimuli presented during the post-test.

As expected, the level of identification accuracy was much higher for the full-face stimuli in Experiment 2 than it was for the point-light stimuli in Experiment 1. A variety of participants who received feedback in Experiment 2 also improved in identification accuracy more rapidly with respect to the participants who did not receive feedback. On only the second repetition of words in Experiment 2, for instance, all groups of participants who received feedback (except for the A feedback group) identified a significantly higher percentage of whole words than the N feedback group. In contrast, only the AV and A-then-V feedback groups reached the same level of performance in whole word identification accuracy on the second presentation of training stimuli in Experiment 1.

Although the effect of feedback on perceptual learning which emerged in training in Experiment 2 did not differ as much between feedback types as it did in Experiment 1, the A feedback group lagged

behind the other groups in its rate of perceptual improvement. The A feedback group did not identify a significantly higher percentage of words than the N feedback group until the fourth presentation of stimuli in training. In this respect, the A feedback group was actually outperformed by the O feedback group, even though the opposite was true for participants in Experiment 1. This result is surprising, given our expectation that audio feedback should facilitate perceptual learning better than orthographic feedback. Aside from this difference, though, all other feedback groups—including those who received AV, OV, O-then-V and O feedback—improved in identification accuracy at the same rate in comparison to the N feedback group.

In Experiment 1, we found that the rate of perceptual improvement which emerged in training depended on the specific type of feedback the participants received. This result provided evidence for the hypothesis that observers store specific instances of visual speech in a highly detailed, modality-specific format in memory. Since the feedback-based improvement which emerged in Experiment 2 did not depend as strongly or as consistently on the type of feedback the participants received, it is difficult to draw similar conclusions about the representation of feedback information in memory for full-face displays of speech. What appears to be the case, instead, is that it is simply easier for observers to make use of feedback—in a variety of different forms—to improve their perception of full-face, visual-only speech stimuli that they have seen before. This may be the case for the same reasons why it is easier to perceive speech in full-face displays than it is to perceive speech in point-light displays: full-face stimuli not only contain more visual information, but observers also have more experience viewing speech in full-face form in everyday life. Participants in an experiment such as this one thus have much more information and knowledge to draw from—as well as more practice perceiving speech in full-face displays—to help them interpret the visual-only stimuli.

Previous studies have shown that the perception of speech in full-face displays improves when participants are provided with either AV or O feedback during a short training experiment. The results of this study corroborate those earlier findings, but also show that, for full-face displays of speech, neither AV nor O feedback improves the visual-only perception of repeated stimuli better than the other. Interestingly, however, one form of feedback that was not used in any of the previous studies—A feedback—did not improve the perception of whole words quite as rapidly as the other forms of feedback that were used in this study. It is possible that participants in the A feedback condition lagged behind the other feedback groups in their rate of perceptual improvement because the structure of the experiment required participants to identify words spoken in visual-only stimuli by typing them into a computer. In order to do this, all participants had to access the orthographic representations of each word. Audio-only feedback is the only form of feedback which did not present the spoken word to the participants in either visual or orthographic form. Without receiving information in either of these forms, it may have been more difficult for the participants in the A feedback condition to use the feedback information they received to interpret visual-only stimuli in orthographic terms than it was for participants in the other feedback groups. This hypothesis could be tested by investigating whether A feedback would facilitate the perceptual learning of the visual properties of speech more rapidly in an experimental paradigm where participants speak their responses, rather than type them. In this paradigm, the output of the participants' spoken responses would be in the same modality as the feedback they receive in the A condition. Participants would not have to generalize across modalities when interpreting their spoken responses in terms of the feedback information they receive under these conditions. It might therefore be easier for them to use A feedback to modify their responses to more closely match what they see in the visual-only speech stimuli.

In summary, the results of Experiment 2 confirmed that the perception of visual-only full-face displays of speech also improved when stimuli were re-presented to participants and feedback was

provided after each trial in a short training experiment. As with the point-light displays of speech in Experiment 1, this perceptual learning effect was only observed in stimuli the participants had seen before and did not, therefore, generalize to novel stimuli. The form in which participants received feedback did not affect the rate of perceptual learning for full-face displays as much as it did for point-light displays, however. The perception of speech in visual-only full-face displays improved rapidly when observers received several different forms of feedback, regardless of whether or not feedback re-presented the stimulus in its original, visual form.

General Discussion

This study investigated whether modality-specific information is preserved in memory when observers are asked to identify spoken words from visual displays of speech. Evidence from the perceptual learning of visual-only point-light displays of speech indicated that highly detailed, modality-specific information was preserved in the visual perception of speech. Feedback that re-presented point-light speech stimuli in their original, visual form to participants improved the perception of point-light displays of speech better than feedback which did not. This result indicates that the fine-grained visual details of point-light stimuli are encoded and retained in memory, and are used to facilitate the perception of previously seen speech stimuli, rather than being discarded in favor of an abstract, linguistic representation resulting from perceptual normalization processes at the time of initial encoding. The visual perception of speech thus preserves “extra-linguistic” visual details in memory just as the auditory perception of speech preserves speaker-specific information (Goldinger, 1997; Nygaard, Sommers, & Pisoni, 1994; Nygaard & Pisoni, 1998; Palmeri, Goldinger, & Pisoni, 1993) and reading preserves information about the font face and orientation of written material (Kolers, 1973).

The effect of visual vs. non-visual feedback on perceptual learning in this study was, however, limited to the perceptual learning of point-light displays of speech. The rate of perceptual learning of full-face displays of speech was, in contrast, largely independent of the form in which feedback was provided to the participants. The perceptual learning of full-face displays of speech may have been insulated from the form in which feedback was provided to participants for at least two reasons. First, compared to point-light displays, full-face displays provide more visual information to observers about the speech events they are trying to perceive, and second, observers have far more experience perceiving full-face speech displays outside the laboratory setting. In future research, it may be possible to test how much each of these two factors interacts with feedback in the perceptual learning of the visual properties of full-face displays of speech by varying them independently in a study on the perception of point-light displays of speech. For instance, more visual information could be provided in point-light displays of speech by simply adding more points of light to the articulators, or even by completely illuminating some articulators, such as the lips, without showing the speaker’s whole face. A range of point-light displays along a continuum of informativeness could be created and then presented to observers in a perceptual learning paradigm such as this one. The perceptual learning of point-light displays which are highly visually informative should, presumably, be less susceptible to differences in feedback form than those point-light displays which are less visually informative. Similarly, participants’ experience with point-light displays could also be increased through a passive viewing task, in which they simply watch speech in point-light displays (with sound) without responding to what they see. Participants with varying amounts of exposure to point-light displays in such a task could then be tested in a perceptual learning experiment. The gains in perceptual accuracy made by those participants with greater amounts of exposure to point-light displays of speech should be less sensitive to the form of feedback they receive than the gains in perceptual accuracy made by those participants made with less experience to point-light displays. This line of inquiry might, however, prove impractical because enormous amounts of exposure

to point-light displays might be required before the amount of experience observers have with point-light displays would begin to approximate their level of experience with full-face displays of speech.

It is important to note that the feedback-based gains in perceptual accuracy made by the participants in this study were limited to stimuli they had seen before on previous trials. No group of participants in either experiment displayed generalization of what they had learned through feedback to improve their perception of novel visual-only stimuli. However, participants in all feedback conditions, in both experiments, were able to improve their identification accuracy between pre-test and post-test through practice alone. This effect of practice on perceptual learning suggests that there is more to the process of visual speech perception than the mere retention of episodic details in memory. Through practice, observers can apparently “tune in” to the properties of visual-only displays of speech which may be relevant for the identification of linguistic information. In other words, practice enables observers to improve in their ability to pick up information from the visual signal per se, independently of how well they can match up those visual stimuli with feedback information in memory. Since perceptual learning due to practice generalizes to novel stimuli, it likely reflects some form of higher-order knowledge of the articulation of speech sounds in a variety of different phonetic environments.

An important question for future research to answer is what—if any—role the “extra-linguistic” details stored in memory from previous experiences with speech play in the perception of novel speech stimuli. It may be possible to answer this question by modifying the training paradigm that was used in this study in some way so that feedback improves the perceptual identification of novel speech stimuli, as well as repeated stimuli. One modification which may make such generalization possible is to incorporate more variability into the training stimuli, in order to force observers to abstract away from arbitrary, idiosyncratic, instance-specific details which are specific to particular training stimuli. For instance, generalization of feedback-based knowledge may be facilitated by training participants to perceive visual-only speech tokens produced by a wide variety of talkers, rather than just one, as used in the present set of experiments. Similarly, generalization might also be facilitated by training participants on sentence-length stimuli, rather than on individual words. Variations of this “High Variability Training Paradigm” have been used with success in previous work on training Japanese listeners to identify the English /r-/l/ distinction (Lively, Pisoni, Yamada, Tohkura, & Yamada, 1994), and normal-hearing listeners to both understand synthetic speech (Greenspan, Nusbaum, & Pisoni, 1988) and identify dialects of American English (Clopper & Pisoni, 2004). It is also possible that training observers with nonsense words or semantically anomalous sentences may facilitate generalization, because observers would be forced to extract linguistic information solely from what they perceive in the visual-only speech signals, without relying on higher-order knowledge to facilitate processing (Burkholder, 2005).

Developing a training paradigm which can improve the visual perception of speech is, of course, important for practical as well as theoretical reasons. The ability to perceive speech through the visual domain can dramatically improve the intelligibility of speech in adverse listening conditions for normal-hearing listeners, as well as improve the ability of the hearing-impaired to communicate (Bergeson & Pisoni, 2004). The results of this study, along with that of previous research, have shown that incorporating feedback into a training paradigm is an effective way of improving the visual perception of speech. The results of this study also suggest, however, that future researchers should consider the form in which they provide feedback to participants in any given training paradigm. With respect to the perceptual learning of the visual properties of speech, not all forms of feedback are created equal.

References

- Abercrombie, D. (1967). *Elements of General Phonetics*. Edinburgh: Edinburgh University.
- Bergeson, T.R. & Pisoni, D.B. (2004). Audiovisual speech perception in deaf adults and children following cochlear implantation. In G. Calvert, C. Spence, & B.E. Stein (Eds.), *Handbook of multisensory processes* (pp. 749-772). Cambridge, MA: MIT Press.
- Bernstein, L.E., Auer, E.T., & Tucker, P.E. (2001). Enhanced speechreading in deaf adults: can short-term training/practice close the gap for hearing adults? *Journal of Speech, Language and Hearing Research, 44*, 5-18.
- Black, J.W., O'Reilly, P.P., & Peck, K. (1963). Self-administered training in lipreading. *Journal of Speech and Hearing Disorders, 28*, 183-186.
- Breeuwer, M. & Plomp, R. (1986). Speechreading supplemented with auditorily presented speech parameters. *Journal of the Acoustical Society of America, 79*, 481-499.
- Burkholder, R.A. (2005). *Perceptual learning of speech processed through an acoustic simulation of a cochlear implant* (Research on Spoken Language Processing Technical Report No. 13). Bloomington, IN: Speech Research Laboratory, Indiana University.
- Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge: MIT Press.
- Chomsky, N. & Halle, M. (1968). *The sound pattern of English*. New York: Harper & Row.
- Clopper, C.G. & Pisoni, D.B. (2004). Effects of talker variability on perceptual learning of dialects. *Language and Speech, 47*, 207-239.
- CMU pronouncing dictionary, version 0.6 Available at <http://www.speech.cs.cmu.edu/cgi-gbin/cmudict>.
- Davis, M.H., Johnsrude, I.S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (May, 2005). Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General, 134*, 222-241.
- Demorest, M.E. & Bernstein, L.E. (1992). Sources of variability in speechreading sentences: a generalizability analysis. *Journal of Speech and Hearing Research, 35*, 876-891.
- Gesi, A.T., Massaro, D.W., & Cohen, M.M. (1992). Discovery and expository methods in teaching visual consonant and word identification. *Journal of Speech and Hearing Research, 35*, 1180-1188.
- Goldinger, S.D. (1997). Words and voices: perception and production in an episodic lexicon. In K.A. Johnson & J.W. Mullennix (Eds.), *Talker variability in speech processing*. (pp. 33-66). Academic Press: San Diego, CA.
- Greenspan, S.L., Nusbaum, H.C., & Pisoni, D.B. (1988). Perceptual learning of synthetic speech produced by rule. *Journal of Experimental Psychology: Human Learning, Memory and Cognition, 14*, 421-433.
- Halle, M. (1985). Speculations about the representations of words in memory. In V. Fromkin (Ed.), *Phonetic linguistics*. (pp. 101-114). Academic Press: Orlando.
- Heider, F. & Heider, G.M. (1940). An experimental investigation of lipreading. *Psychological Monographs, 52*, 124-153.
- Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics, 14*, 201-211.
- Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K.A. Johnson & J.W. Mullennix (Eds.), *Talker variability in speech processing*. (pp. 145-166). Academic Press: San Diego, CA.
- Kolers, P.A. (1973). Remembering operations. *Memory and Cognition, 1*, 347-355.
- Kreiman, J., VanLancker-Sidtis, D., & Gerratt, B.R. (2005). Perception of voice quality. In D.B. Pisoni & R.E. Remez (Eds.), *The handbook of speech perception*. (pp. 338-362). Blackwell Publishing: Malden, MA.

- 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.
- Lachs, L. & Pisoni, D.B. (2004a). Specification of crossmodal source information in isolated kinematic displays of speech. *Journal of the Acoustical Society of America*, *116*, 507-518.
- Lachs, L. & Pisoni, D.B. (2004b). Cross-modal source information and spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *30*, 378-396.
- Lachs, L. & Pisoni, D.B. (2004c). Crossmodal source identification in speech perception. *Ecological Psychology*, *16* (3), 159-187.
- Laver, J. (1994). Principles of phonetics. New York: Cambridge University Press.
- Lively, S.E., Pisoni, D.B., Yamada, R.A., Tohkura, Y., & Yamada, T. (1994). Training Japanese listeners to identify English /r/ and /l/. III. Long-term retention of new phonetic categories. *Journal of the Acoustical Society of America*, *96*, 2076-2087.
- Massaro, D.W., Cohen, M.M., & Gesi, A.T. (1993). Long-term training, transfer and retention in learning to lipread. *Perception & Psychophysics*, *53*, 549-562.
- Massaro, D.W. & Light, J. (2004). Using visible speech to train perception and production of speech for individuals with hearing loss. *Journal of Speech, Language and Hearing Research*, *47*, 304-320.
- McGurk, H. & McDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746-748.
- Nygaard, L.C., Sommers, M.S., & Pisoni, D.B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, *5*, 42-46.
- Nygaard, L.C. & Pisoni, D.B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, *60*, 355-376.
- Palmeri, T.J., Goldinger, S.D., & Pisoni, D.B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *19*, 309-328.
- Pashler, H., Cepeda, N., Wixted, J., & Rohrer, D. (2005). When does feedback facilitate learning of words and facts? *Journal of Experimental Psychology: Learning, Memory and Cognition*, *31*, 3-8.
- Pierrehumbert, J.B. (2001). Exemplar dynamics: word frequency, lenition and contrast. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure*. (pp. 137-158). John Benjamins: Amsterdam.
- Pisoni, D.B. (1997). Some thoughts on "normalization" in speech perception. In K.A. Johnson & J.W. Mullennix (Eds.), *Talker variability in speech processing*. (pp. 9-32). Academic Press: San Diego.
- Pisoni, D.B. & Levi, S.V. (2005). Some observations on representations and representational specificity in speech perception and spoken word recognition. In *Research on Spoken Language Processing Progress Report No. 27*. Bloomington, IN: Speech Research Laboratory, Indiana University.
- Rosenblum, L.D. (2005). Primacy of multimodal speech perception. In D.B. Pisoni & R.E. Remez (Eds.), *The handbook of speech perception*. (pp. 51-78). Blackwell Publishing: Malden, MA.
- Rosenblum, L.D., Johnson, J.A., & Saldana, H.M. (1996). Point-light facial displays enhance comprehension of speech in noise. *Journal of Speech and Hearing Research*, *32*, 921-929.
- Rosenblum, L.D. & Saldana, H.M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 318-331.
- Rosenblum, L.D., Yakel, D.A., Baseer, N., Panchal, A., Nodarse, B.C., & Niehus, R.P. (2002). Visual speech information for face recognition. *Perception & Psychophysics*, *64*, 220-229.
- Sumby, W.H. & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, *26*, 212-215.
- Summerfield, A.Q. (1979). Use of visual information for phonetic perception. *Phonetica*, *36*, 314-331.

- 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.
- Walden, B.E., Erdman, S.A., Montgomery, A.A., Schwartz, D.M., & Prosek, R.A. (1981). Some effects of training on speech recognition by hearing-impaired adults. *Journal of Speech and Hearing Research*, 24, 207-216.
- Winters, S.J. & Pisoni, D.B. (2004). Some effects of feedback on the perception of point-light and full-face visual displays of speech: a preliminary report. In *Research on Spoken Language Processing Progress Report No. 26* (pp. 139-164). Bloomington, IN: Speech Research Laboratory, Indiana University.