Perception & Psychophysics
1993, 53 (5), 549-562

Long-term training, transfer, and retention in learning to lipread

DOMINIC W. MASSARO, MICHAEL M. COHEN, and ANTOINETTE T. GESI
University of California, Santa Cruz, California

A long-term training paradigm in lipreading was used to test the fuzzy logical model of perception (FLMP). This model has been used successfully to describe the joint contribution of audible and visible speech in bimodal speech perception. Tests of the model were extended in the present experiment to include the prediction of confusion matrices, as well as performance at several different levels of skill. The predictions of the FLMP were contrasted with the predictions of a prelabeling integration model (PRLM). Subjects were taught to lipread 22 initial consonants in three different vowel contexts. Training involved a variety of discrimination and identification lessons with the consonant-vowel syllables. Repeated testing was given on syllables, words, and sentences. The test items were presented visually, auditorily, and bimodally, at normal rate or three times normal rate. The subjects improved in their lipreading ability across all three types of test items. Replicating previous results, the present study illustrates that substantial gains in lipreading performance are possible. Relative to the PRLM, the FLMP gave a better description of the confusion matrices at both the beginning and the end of practice. One new finding from the present study is that the FLMP can account for the gains in bimodal speech perception as subjects improve their lipreading and listening abilities.

In face-to-face communication, visible speech contributes to speech perception. As the signal-to-noise ratio of the speech signal decreases, the benefits of viewing the talker increase (Dodd, 1977; Erber, 1969; Hutton, 1959; Neely, 1956; O'Neill, 1954; Sumby & Pollack, 1954). Even when auditory speech is intelligible, visual information from the talker's face can influence speech perception (Massaro & Cohen, 1983; McGurk & MacDonald, 1976). Bimodal speech perception can be characterized as a process in which the auditory and visual sources each provide continuous information that is combined or integrated to achieve an overall goodness of match with each possible alternative. The perceptual judgment is determined by the relative goodness of match of each of the relevant alternatives (Massaro, 1987; Summerfield, 1979). The percept that emerges from this processing reflects the contribution of both sources of information. Given an auditory /da/ and a visual /ba/, for example, the perceiver often categorizes the event as /bda/. This experience is a reasonable outcome, given the use of both sources, because an auditory /da/ is similar to auditory /bda/, and visual /ba/ is similar to visual /bda/. A relatively close match on both sources is a more optimal decision than a good match on one source and a mismatch on the other, as would be the case for either /ba/ or /da/ responses.

Massaro, Thompson, Barron, and Laren (1986) asked normal-hearing preschool children and normal-hearing adults to identify bimodal syllables and to lipread visible syllables without sound. The visible syllables /ba/ and /da/ were crossed with a five-step auditory /ba/-/da/ continuum. The auditory speech was also presented while the talker did not move the lips, and the visible speech was presented without auditory speech. Adults were more accurate in lipreading and showed a larger contribution of visible speech in the identification of the bimodal syllables relative to the preschoolers. This result could not be explained by the possibility that children were less likely to attend to the visible speech (Massaro, 1984). Across both groups of subjects, there was a positive correlation between lipreading accuracy and the contribution of visible speech in bimodal speech perception. MacLeod and Summerfield (1990) evaluated the correlation between lipreading ability and the contribution of visible speech to bimodal speech perception. Twenty undergraduates lipread sentences, with a performance range between 0% and 70% correct. The same subjects also processed a different set of sentences given auditory and bimodal speech. The dependent measure was the difference in the signal-to-noise ratio required to report the content words correctly in auditory and bimodal speech. There was a strong correlation of .89 between lipreading ability and the benefit gained from a view of the talker in bimodal speech perception. Like the research in our laboratory, these results imply that improvements in lipreading ability should

The research reported in this paper and the writing of the paper were supported, in part, by grants from the Public Health Service (PHS R01 NS 20314), the National Science Foundation (BNS 8812728), a James McKeen Cattell Fellowship, and the graduate division of the University of California, Santa Cruz. The authors would like to thank Lester Krueger and three anonymous reviewers, whose comments were very helpful in the revision of this paper. Correspondence should be addressed to D. W. Massaro, Program in Experimental Psychology, University of California, Santa Cruz, CA 95064 (e-mail: massaro@fuzzy.ucsc.edu).

necessarily increase the contribution of visible speech in bimodal speech perception.

The positive correlation between lipreading accuracy and the influence of visible speech in bimodal speech perception can justify the effort required for improving lipreading skills. A good lipreader should more accurately perceive bimodal speech if he or she has a hearing loss or if there is a degraded auditory input. In the present study, we ask how the improvement in lipreading ability with training will facilitate the perception of bimodal speech. As far as we know, this question has not been answered in previous research because quantitative models of performance were not available. Without quantitative models, it is not possible to address the question of how the improvement in lipreading facilitates bimodal speech perception. For example, the improvement in bimodal speech perception might result from simply a gain in the information available from visible speech or some modification in how the auditory and visual information is combined.

The primary goals of the present study are (1) to expand the venue for tests of extant models of speech perception, and (2) to test the models at different levels of skill. Most quantitative tests among models have involved judgments of synthetic and/or natural speech that has been modified (Massaro, 1987). Furthermore, stimulus-response confusions are not usually analyzed. Even Braida's (1991) tests of these models were limited to predictions of overall accuracy rather than the stimulus-response confusions. In the present study, extant models are tested against the stimulus-response confusion matrices generated at several levels of performance skill. We now articulate several accounts of bimodal speech perception.

Fuzzy Logical Model of Perception

The benefits of lipreading and learning to lipread can be rationalized within the context of a fuzzy logical model of speech perception (FLMP). The assumptions central to the FLMP are as follows: (1) There are multiple sources of information supporting speech perception, (2) each source of information is evaluated to give the degree to which that source specifies various alternatives, (3) the sources of information are evaluated independently of one another, (4) the sources are integrated to provide an overall degree of support for each alternative, and (5) perceptual identification and interpretation follow the relative degree of support among the relevant alternatives.

In a speech perception situation, all prototypes corresponding to the perceptual units of the spoken language are activated. Consider a speech signal /ba/, spoken in face-to-face communication. The sensory systems transduce the physical event and make available various sources of information called *features*. The syllable /ba/ might have visible featural information related to the moving contours of the optic display and audible information corresponding to the second and third formant transitions. These two features must share a common metric if they eventually are going to be related to one another. To serve this purpose, fuzzy truth values (Zadeh, 1965) are used because they provide a natural representation of the degree of match (Massaro, 1987). Fuzzy truth values lie between 0 and 1, corresponding to a proposition's being completely false or completely true. The truth value .5 corresponds to a completely ambiguous outcome, whereas .7 would be more true than false, and so on.

Figure 1 illustrates the three operations assumed by the FLMP. Feature evaluation provides the degree to which each feature in the syllable matches the corresponding feature in each prototype in memory. During the second operation of the model, called *feature integration*, the features (actually the degrees of matches) corresponding to each prototype are combined (or conjoined in logical terms). The outcome of feature integration consists of the degree to which each prototype matches the syllable. During the decision stage, the merit of each relevant prototype is evaluated in relation to the sum of the merits of the other relevant prototypes. This decision operation is modeled after Luce's (1959) choice rule, called a relative goodness rule (RGR) by Massaro and Friedman (1990). This RGR predicts the proportion of times the syllable is identified as an instance of the prototype. The RGR can predict response probabilities between 0 and 1. If absolute goodness were used, only the probabilities 0 and 1 could be predicted—an unreasonable prediction (Massaro & Friedman, 1990).

According to the FLMP, independent sources of information come together to influence speech perception. Most importantly, all sources of information contribute to perception solely as a function of their information value. One modality does not necessarily have any dominance over another. If visible speech perception is very good, it will also make a strong contribution to bimodal speech perception. If training in lipreading improves visible speech perception, the FLMP predicts that the visible contribution to bimodal speech perception should also increase. Thus, the model can be used to justify training in lipreading because of the predicted benefits in bimodal (or even multimodal) speech perception. Because of the specific predictions of the FLMP, training studies offer another venue for empirical tests.

It has been found in earlier research that lipreading ability improves with training (e.g., Dodd, Plant, & Gregory, 1989; Gesi, Massaro, & Cohen, 1992). Although improvement over training was measured, the primary

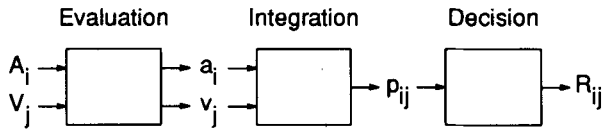

Figure 1. Schematic representation of the three operations of the fuzzy logical model of perception that are involved in perceptual recognition. The sources of information are represented by uppercase letters. The psychological representations are given by lowercase letters. The evaluation process transforms these into psychological values (indicated by lowercase letters), which are then integrated to give an overall value. The decision operation maps this value into some response, such as a discrete decision or a rating.

purpose of this study was to test the FLMP's strong prediction of performance in a long-term lipreading experiment. Of additional interest is the nature of integration of audible and visible speech as a function of experience in the task (familiarity). It is possible that the integration becomes more efficient with increasing familiarity (Braida, 1991). Testing the FLMP at different levels of skill provides a test of this hypothesis. If this hypothesis is correct, the FLMP should give a better description of performance after rather than before training, because the FLMP is an optimal model for integrating information (Massaro & Friedman, 1990).

The present analyses of confusion matrices offer two novel domains for testing the FLMP. First, the model has usually been tested in factorial designs with synthetic speech (e.g., Massaro & Cohen, 1990). One distinguishing property of these tests is that there is no correct answer for each test stimulus (and subjects are *not* given feedback during the experiment). The observed and predicted dependent measures are the proportions of identifications for each stimulus. In the present study, there is a normatively correct answer for each stimulus and feedback is given during the training sessions. Thus, the FLMP will be extended to predict the accuracy of identification and the proportion of times one stimulus is confused with another. Second, previous studies have measured and predicted performance at more or less a single level of skill. The present study offers tests of the FLMP across a range of skill levels, as subjects improve in the lipreading task.

Prelabeling Integration Model

Braida (1991) developed and tested a prelabeling integration model (PRLM). In the taxonomy of Massaro and Friedman (1990) and Cohen and Massaro (1992), the PRLM is a multidimensional version of the theory of signal detectability (TSD). A presentation of a stimulus in a given modality locates that stimulus in a multidimensional space. Given that the process is noisy (Gaussian), the location may be displaced from the stimulus center. There is also a *response center* (prototype) in the multidimensional space. The multidimensional space for a bimodal presentation is simply the combination of the spaces for the two unimodal presentations. For example, if the auditory and visual sources are each represented in three-dimensional space, the bimodal information is represented in six-dimensional space. In all cases, the subject chooses the response alternative whose response center (or prototype) is closest to the location of the stimulus in the multidimensional space.

Braida (1991) tested the FLMP and PRLM against confusion matrices from multimodal speech identification experiments. In these experiments, subjects were tested under three presentation conditions: two unimodal and one bimodal. In most cases, subjects were given auditory speech, visual speech, or auditory-visual (bimodal) speech, although tactile or electrical speech were sometimes used rather than auditory speech. Subjects were asked to identify the stimuli with a fixed set of response alternatives.

In his tests of the PRLM, Braida (1991) used a multidimensional scaling (MDS) technique to find the optimal locations of stimulus centers in order to minimize the errors in prediction of each unimodal condition. The response prototypes were assumed to be equal to their respective stimulus centers. The bimodal judgments were predicted from the combined spaces of the unimodal judgments. For his fits of the FLMP, Braida simply used the unimodal data to directly predict the bimodal points. Neither of these two tests is optimal, because only the unimodal results are used. In Braida's test of the PRLM, the bimodal results cannot influence the location of the stimulus centers in the multidimensional space. In the test of the FLMP, he assumed that the unimodal results are an error-free measure of the parameters of the FLMP. In the present paper, however, minimization model-fitting techniques are applied to both the unimodal and the bimodal results for the tests of both the PRLM and the FLMP. Thus, we should have a direct comparison between these two models when both models are performing as optimally as possible.

Previous Research on Training Lipreading

Walden, Prosek, Montgomery, Scherr, and Jones (1977) investigated whether training could improve lipreading of consonants, and if so, how much training was required. They trained 31 hearing-impaired adults to identify consonant-vowel (CV) syllables by using a set of 38 exercises that increased in difficulty (i.e., fewer to more CV syllables; syllables from different viseme categories to syllables from the same viseme categories). Although the subjects were hearing impaired, they wore their hearing aids throughout the study. A viseme is a visible speech category in which several different phonemes have roughly the same visible articulation. For example, /b/, /p/, and /m/ are in one viseme category because their visible articulations cannot be distinguished. The subjects were given 14 sessions of intensive, individualized lipreading instruction, administered by three different clinicians. One training task required subjects to make *same-different* judgments between CV syllables. After this task was completed, a second training task required the subjects to identify CV syllables. Feedback about response accuracy (i.e., correct or incorrect) was given on each trial for both training tasks. In addition, items that were judged or identified incorrectly were repeated until correct responses were made.

The effects of training were measured by lipreading differences between a pretest and a posttest. The pretest and posttest, which were identical, consisted of 400 items (20 consonants paired with the vowel /*a*/, each randomly presented 20 times) by a male talker. The male talker was not one of the clinicians used during training. Using a criterion of 70% correct, the number of visemes that could be recognized increased from six to nine. Before training, the subjects could distinguish /fv/, /θð/, /wr/,

/pbm/, /šž/, and /tdszjkngl/. Although learning occurred for all viseme classes, there were some dramatic increases in accuracy from pretraining to posttraining. After training, the subjects could distinguish /fv/, /θð/, /w/, /r/, /pbm/, /sz/, /šž/, /tdjkng/, and /l/. The viseme /sz/, for example, increased from 34.9% to 79.2% correct on the posttraining. Duration appears to be a cue used to distinguish /sz/ from the other alveolar and velar consonants. The correct identification of /r/ increased from 36.1% to 88.6%. An additional result revealed that the majority of the learning of the viseme classifications occurred within the first 5–6 h of training. A smaller but gradual and consistent increase in performance occurred throughout the rest of training. These findings suggest that training can improve the lipreading abilities of hearing-impaired adults.

Walden, Erdman, Montgomery, Schwartz, and Prosek (1981) extended the Walden et al. (1977) study to assess the transfer of training on individual segments to the recognition of bimodal sentences and to include training of some subjects on the auditory modality. The subjects had high-frequency hearing loss and were new hearing-aid users enrolled in a 2-week inpatient aural rehabilitation program. They wore their hearing aids throughout the experiment. The testing consisted of 22 English consonants in an /a/-C-/a/ context. The training materials were similar to those used in Walden et al.'s study. Using their same criterion, six viseme groups emerged in the posttest. For the visually trained group, correct recognition on these visemes improved about 10% between the pretest and posttest. For the auditorily trained group, correct recognition on these visemes improved about 7% between the pretest and posttest. A control group given just the posttest on the syllables was not included. This group would have measured any improvement that might have occurred, given repeated testing. On the bimodal sentence test, the improvement for subjects given the auditory and visual training was 28% and 23%, respectively. These values are larger than the 10% gain for subjects enrolled in their standard program and given no specific training on syllables. These results, therefore, show some positive transfer from training on syllable segments to bimodal sentences.

Present Study

Although Walden et al. (1977) and Walden et al. (1981) found gains in lipreading ability with a relatively short training period, several questions remain. First, only a single vowel was used; the relative difficulty of the consonants and the confusions among them might not occur in other vowel contexts. We tested consonants in three different vowel contexts. Second, the subjects in the Walden et al. (1981) study were pretested and posttested on unimodal syllables. Auditory training and visual training were a between-subjects variable. In our experiment, the pretest and posttest had auditory, visual, and bimodal items. Third, although Walden et al. (1981) tested transfer to bimodal sentences, we assessed transfer to auditory, visual, and bimodal sentences. Finally, no long-term retention was measured in Walden et al.'s studies. To determine the value of training, its influence should be assessed for both short-term and long-term retention. We measured retention after a long-term 7.5-week delay.

In a recent study (Gesi et al., 1992), subjects with normal hearing were trained to lipread CV syllables for 3 days. In addition, we compared discovery and expository methods of learning to lipread. Subjects learned with training, but there was no difference between the two learning methods. As a retention measure, subjects returned 4 weeks later and repeated the training. There were significant savings of the original learning. Thus, there is some evidence for improvement in visible speech perception with experience, and for a retention of this improvement for 1 month.

The present study extended the studies of Walden et al. (1977), Walden et al. (1981), and Gesi et al. (1992) to address important issues in learning to lipread. We measured not only the expected improvement in lipreading but also its contribution to bimodal speech perception at different levels of lipreading skill. The accuracy of visible consonant recognition in different vowel environments (/a/, /i/, and /u/) was also assessed. There is some evidence that visible consonants are easier to recognize in some vowel environments than in others (Owens & Blazek, 1985). We also evaluated not only the degree to which training on CV syllables improved recognition of these syllables, but also to what extent training enhanced the contribution of visible speech in words and sentences. Finally, we assessed whether there is long-term retention or savings of the training. We extended the training period to several months and assessed the transfer of lipreading training on syllables to lipreading words and sentences.

METHOD

Subjects

Six female work-study students served as subjects and were paid $5.56/h. Their ages were 18, 19, 20, 21, 21, and 32. None of the subjects was hearing-impaired. One of the subjects reported a "25% hearing loss" in the left ear. However, this subject showed no decrement on the auditory task in relation to the other subjects. The subjects appeared to be highly motivated throughout the experiment. Three subjects were tested and trained simultaneously in separate sound-attenuated rooms.

Test Method Overview

The subjects were given six sets of lipreading tests, with five intervening training tasks. Each lipreading test consisted of three parts in the following order: (1) single-syllable words (word), (2) CV-syllable tests (syllable), and (3) full-sentence tests (sentence). We will first describe the methods for the three types of tests. All of the test stimuli were spoken by a white adult male on a laser videodisk recorded by Bernstein and Eberhardt (1986). The exact word list and sentences used in the tests and the words used in the training sessions are available from the senior author by anonymous FTP to fuzzy.ucsc.edu under file pub/train.dat, or on floppy disk.

Word Method

Stimuli. The stimuli were 420 single-syllable English words, given in the Appendix. They included words from the Modified

Rhyme Test (House, Williams, Hecker, & Kryter, 1965), as well as additional words. The duration of the frames on the laser disk containing the words averaged 1,345 msec. However, there were several frames at the beginning and end of each word that had no visual or auditory speech. The average duration of visual movement during the words was 1,072 msec, and the duration of the audible signal was 552 msec. The words were also presented three times as fast for a *fast* rate. For the fast presentation rate, only every third visual and/or auditory frame was presented. This means that for each 100 msec in the original word, only the first 33 msec were presented. At this speed, the segments played averaged 448 msec, with 357 msec of visual movement and 184 msec of audible signal.

Apparatus. All experimental events were controlled by a DEC PDP-11/34A computer. The stimuli were presented by a computer-controlled SONY LDP-1500 laser disk player on NEC C12-202A 12-in. color monitors. Four sound-attenuated subject rooms were used, each illuminated by two 60-W incandescent bulbs in a frosted-glass ceiling fixture. Each room contained a chair that faced a table with two displays—a TVI950 terminal and a color monitor. The audio feedback portion of the experimental tape was presented to the subjects over the built-in speakers of the monitors at a comfortable listening level of about 61 dB-A (measured at the approximate position of the observer's head using a B&K 2123 sound level meter with a 4134 microphone on the fast setting). A trial number was displayed on a Televideo TVI-950 terminal, and subjects used a button on the terminal keyboard to indicate that they were ready for the next trial.

Procedure. There were two sessions of 210 trials, each taking about 20 min with a 5-min break in between. The subjects were reminded of the trial number by a display on a terminal screen adjacent to the stimulus display. Each trial started with a blank 1,000-msec interval. Next, a test word was presented, which the subjects identified by writing on a test sheet in a numbered box. No feedback was given. After the subjects wrote their answers, they indicated that they were ready for the next trial by pressing a button. The next trial began 5 sec after each subject pressed the button.

Words were presented under one of six conditions combining speed—either at normal or triple normal (fast) rate—and modality, using auditory-alone (A), visual-alone (V), or bimodal (B) channels. The 420 stimulus words were divided into six groups of 70, as shown in the Appendix, for presentation under one of these conditions. The presentation condition was reassigned to a different word group during each of the first six sessions so that every word in a group was presented under one of the six presentation conditions. During each session, the subjects were presented all of the 420 words once, sampled randomly without replacement. The seventh session was equivalent to the first. For all subjects, the order of presentation of the six conditions was the same.

A 400-msec warning tone produced by the terminal keyboard started each trial. During each word presentation, the laser disk was set for the required modality and speed. The laser disk was then moved to the starting frame of the word, where it stayed for 400 msec. Next, all frames of the word were played and the disk was left at the final frame until all subjects indicated that they were ready for the next trial. On auditory-alone trials, subjects saw a blank screen.

The subjects' written responses were each typed into a computer for scoring. To carry out this scoring, a phonetic version of Webster's 7th edition dictionary was used. The consonants were also converted from phonemes to the nine visemes in order to carry out the viseme analyses. Table 1 gives the conversions used from phonemes to visemes. These translations from consonants to visemes are based on the results of Walden et al. (1977). The three vowels correspond to different visemes (Montgomery & Jackson, 1983). These data (9 visemes presented × 9 viseme responses × 2 presentation rates) were examined with an analysis of variance.

Table 1
The Consonant Phoneme to Viseme Translations Used in the Present Experiments

Phoneme	Viseme
p, b, m	b
f, v	v
θ, ð	ð
t, d, n, k, g, j, y, h, x	d
s, z	z
l	l
r	r
š, ž, č, ǰ	ž
w	w

Syllable Method

Stimuli. The stimuli were 132 CV syllables combining the 3 vowels /i/, /a/, and /u/ with the 22 initial consonants used by Walden et al. (1981): /b/, /č/, /d/, /ð/, /f/, /g/, /h/, /j/, /k/, /l/, /m/, /n/, /p/, /r/, /s/, /š/, /t/, /θ/, /v/, /w/, /z/, and /ž/. Two different exemplars of each of these 66 syllable types were used. The duration of the frames on the laser disk containing the syllables averaged 1,047 msec, with the average duration of visual movement during the syllables at 723 msec and the duration of the audible signal at 389 msec. As with the words, the syllables were also presented three times as fast for a *fast* rate. At this speed, the segments played averaged 348 msec, with 240 msec of visual movement and 130 msec audible signal.

Procedure. The apparatus and procedure were as in the word test, except as noted below. There were two sessions of 132 trials, each taking about 17 min with a 5-min break in between. The subjects wrote their answers.

As in the word test, the syllables were presented under one of six conditions: at either normal or fast rate, and along A, V, or B channels. The 132 stimulus syllables were randomized as in the word condition.

Sentence Method

Stimuli. The stimuli were 96 sentences selected from the CID-100 sentence list. They varied in length from 2 to 15 syllables.

Procedure. The apparatus and procedure were as in the word test, except as noted below. There were two sessions of 48 trials, each taking about 17 min with a 5-min break in between.

As in the word and syllable tests, the sentences were presented under one of six conditions: at either normal or fast rate, and along A, V, or B channels. The 96 stimulus sentences were divided into six groups of 16. For a particular subject, each of the 16 sentences in a group was presented according to one of the six presentation methods. During each test, the subjects were presented each of the 96 sentences once, with one sixth of the sentences in each condition. Because of scheduling problems, the sentence test was given only during the first four sessions of testing. Over the four administrations of the test, the first four stimulus groups were used in each of the six conditions. For all subjects, the order of presentation of the four conditions was the same.

Training Method

The syllable, word, and sentence tests were given at the beginning of the experiment, after each of the five sets of training tasks, and after a retention period of 7.5 weeks. The subjects were given five sets of lipreading training tasks, with the previously described test tasks (syllable, word, and sentence test) intervening. The training tasks were: (1) two-word two-choice identification (2W2I), (2) one-word two-choice identification (1W2I), (3) one-word nine-choice identification (1W9I), (4) four-talker CV-syllable nine-choice identification (4CV9I), and (5) one-talker CV-syllable nine-choice

identification (1CV9I). The training tasks occurred in the order presented below.

2W2I task. The stimuli were eight sets of 18 CVC words spoken at the normal rate. Within each set, there were 9 words that started with one viseme and 9 words that started with another viseme. The second and third visemes of each word were balanced among the two subsets of 9 words. The two initial visemes of the eight sets were /d-l, d-z, z-ž, r-w, b-d, b-r, d-ž/ and /b-l/ (see Table 1). The subjects were tested on each of these distinctions until the average proportion correct exceeded .9. A group average was used as the criterion, because 3 subjects were tested simultaneously.

There were one or two sessions of 162 trials, each taking about 20 min with a 5-min break in between sessions. The video output from the laser disk player was fed through an Amiga 1300 genlock incorporated in an Amiga 1000 computer. The Amiga was used to overlay feedback about the correct response on each trial onto the video image in the lower right-hand corner of the screen. The Amiga received the information to be displayed via an asynchronous line from the PDP-11/34A computer.

Each trial started with a 1,000-msec intertrial interval that contained a 400-msec keyboard beep. Two words were then presented, one from each initial viseme subset, separated by a 500-msec interval. Each word presentation consisted of 200 msec of the initial frame, the moving frames of the word itself, and 200 additional milliseconds of the final word frame. The subjects made their responses by pressing one of two buttons (e.g., /d-l/ or /l-d/) to indicate the order of the initial visemes in the two words that they had been presented. After a 500-msec blank interval, feedback was given by displaying the correct viseme response highlighted on the screen for 1,500 msec, followed by another 500-msec blank interval.

1W2I task. Six of the eight stimulus sets from the 2W2I task were used: /d-l, d-z, z-ž, r-w, b-d/ and /d-ž/.

The procedure was the same as that used for 2W2I, except that a single word was presented from one of the two viseme subsets on each trial. The subjects identified the initial viseme (e.g., /d/ or /l/) by pressing one of two buttons. As in the 2W2I training, feedback was given by displaying the correct viseme.

1W9I task. In this task, a set of 81 CVC words from the laser disk was used. There were nine subsets, each with 9 words starting with one of the nine visemes in Table 1.

The procedure was the same as that for 1W2I, except that on each trial, any of the 81 stimuli could be presented and the subject had to identify the initial viseme of the word by pressing one of nine buttons, which were labeled with the identities of the nine possible response alternatives. The keyboard of the TVI-950 terminal was used by the subjects to make their responses and was labeled with the names of each CV syllable. As in the 2W2I training, feedback consisted of the correct response displayed for 1,500 msec.

4CV9I task. The training stimuli, which were taken from Gesi et al. (1992), consisted of 27 unique CV syllables spoken by each of four different talkers (two males and two females). These 27 syllables were constructed by combining nine initial voiced consonants and three vowels. One phoneme was taken from each of the nine viseme classes in Table 1 (i.e., /ð/, /v/, /ž/, /z/, /b/, /d/, /w/, /r/, /l/) combined with the vowels /i/, /a/, and /u/. In a given block of trials, all 108 possible syllables (27 CVs × 4 talkers) were presented. Eighteen such blocks were recorded in different random orders.

An experimental tape was made from the master tape and was played on a Panasonic NV-8200 VHS video recorder. Each trial consisted of a speech stimulus without sound and a feedback stimulus consisting of both visible and audible speech. The feedback presentation occurred after a 5-sec delay, so that enough time would be allowed for each subject to make a response before its presentation.

The keyboard of the TVI-950 terminal was used by the subjects to make their responses and was labeled with the names of each CV syllable.

All subjects completed 3 training blocks per day for 6 days, thus completing 18 blocks of the 108 CV syllables.

1CV9I task. The stimuli were the 132 CV syllables used in the syllable test.

For each session, there were two presentations of the 132 CVs for a total of 264 trials. The procedure and the feedback on each trial were as those in the 1W9I task.

RESULTS

Word Test

Although 23 consonants were tested, they can be classified into one of the nine viseme categories in Table 1. Because some of the visemes were not tested under all of the presentation conditions, the visual identification results were calculated from identification performance averaged over visemes. Figure 2 gives the proportion of correct lipreading responses, pooled over rate of presentation, as a function of sessions. The six curves correspond to the 6 subjects. The overall mean proportion of correct initial viseme responses was .777. Lipreading performance was better (.813) for the normal than for the fast (.740) presentation rate [$F(1,5) = 28.396$, $p = .004$]. As can be seen in Figure 2, lipreading performance improved over sessions [$F(6,30) = 4.828$, $p = .002$], but this absolute improvement did not differ significantly for the two presentation rates [$F(6,30) = 1.240$]. As can be seen in the figure, most of the subjects showed some improvement across the seven test sessions.

Figure 3 shows the average proportion of correct initial phoneme identifications of the words for auditory, visual, and bimodal conditions as a function of test sessions for normal and fast presentation rates. Overall, the proportion of correct identifications increased over test sessions, from .741 to .808 [$F(6,30) = 13.063$, $p < .001$]. The performance on the first and last test sessions went from .884 to .938 for the auditory, from .382 to .509 for the visual, and from .958 to .977 for the bimodal condition. The results differed significantly as a func-

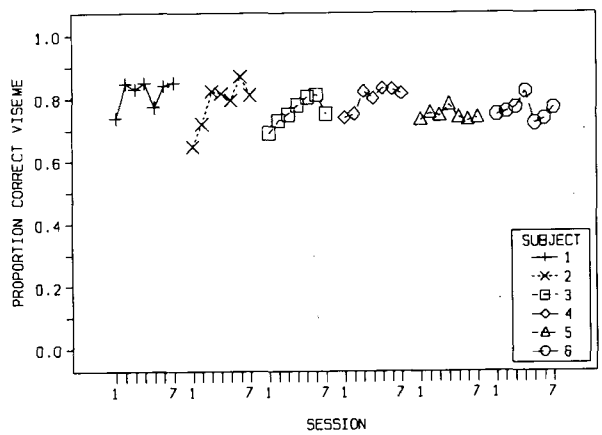

Figure 2. Average proportion correct of lipreading the initial consonant viseme in the visual presentation of the words as a function of the seven sessions of testing for each of the 6 subjects.

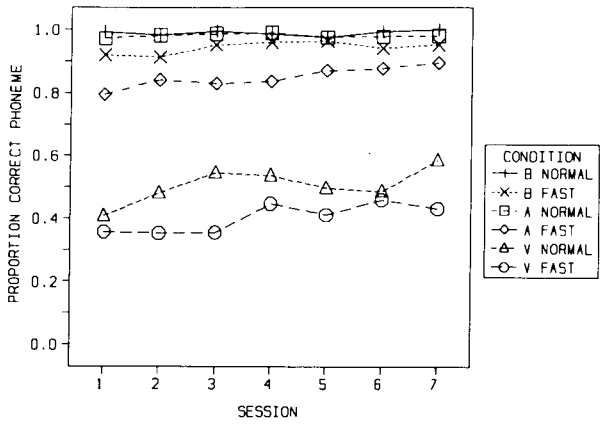

Figure 3. Average proportion correct of initial phoneme identifications on words for bimodal, auditory, and visual conditions, as a function of the seven sessions of testing for normal and fast presentation rates.

tion of modality [$F(2,10) = 2362.227, p < .001$], and there was also a significant effect of rate, and significant interactions of test session × modality, session × rate, modality × rate, and session × modality × rate (all $ps < .001$).

Syllable Test

Figure 4 gives the proportion of correct initial consonant viseme responses for the visual-alone condition, for the 6 subjects as a function of test sessions. As can be seen in the figure, performance improved over test sessions for each of the 6 subjects. The individual differences were larger at the beginning of training than at the end. Performance ranged between .262 and .785 for Session 1 and .708 and .898 for Session 6.

Figure 5 shows the average proportion of correct initial phoneme identifications on CV syllables for auditory, visual, and bimodal conditions, as a function of test sessions for normal and fast presentation rates. Overall, the proportion of correct identifications over test sessions increased from .592 to .699 [$F(6,30) = 2.726, p = .031$]. Accuracy increased from .721 to .850 for the auditory, from .266 to .352 for the visual, and from .788 to .896 for the bimodal condition. Performance differed significantly as a function of modality [$F(2,10) = 4429.740$, $p < .001$], and there was also a significant interaction of rate and modality [$F(2,10) = 5.577, p = .023$].

Figure 6 gives the overall proportion of correct initial consonant viseme responses for the nine visemes, pooled over rate, as a function of test sessions. Performance improved over test sessions [$F(6,30) = 8.439, p < .001$], and this improvement did not differ significantly for the two presentation rates [$F(6,30) = 1.403, p = .245$]. The overall mean proportion of correct initial viseme responses was .750, with .747 for normal and .753 for fast presentation rates. This difference was not significant [$F(1,5) = .324$]. Performance differed as a function of the different visemes [$F(8,40) = 36.698, p < .001$], and this effect interacted with test session [$F(48,240) = 3.037, p < .001$] and rate of presentation [$F(8,40) = 4.184, p = .001$]. The triple interaction of rate, session, and viseme was not significant [$F(48,240) = .795, p = .828$].

Vowel context had only a small overall influence on consonant identification [$F(2,10) = 11.88, p < .003$]. Identification of the initial viseme was best for /a/ (.69), poorest for /i/ (.64), and intermediate for /u/ (.67). There was also a significant interaction of vowel and presentation condition [$F(4,20) = 7.56, p < .001$]. Table 2 gives the average results for each vowel for the three presentation conditions. For visible speech, consonant identification was best in the context /a/. For audible speech, consonant identification was poorest in the context /i/.

The present result showing no large differences in lipreading as a function of vowel context appears to contradict findings of Owens and Blazek (1985). For visual phoneme recognition, their results indicated that overall

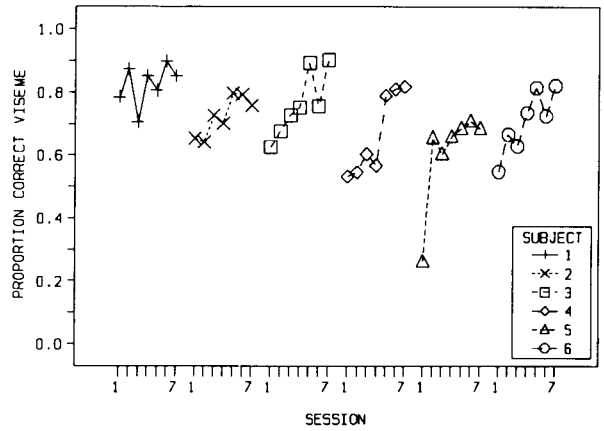

Figure 4. Average proportion correct of lipreading the initial consonant viseme in the visual presentation of the consonant-vowel syllables, as a function of the seven sessions of testing for each of the 6 subjects.

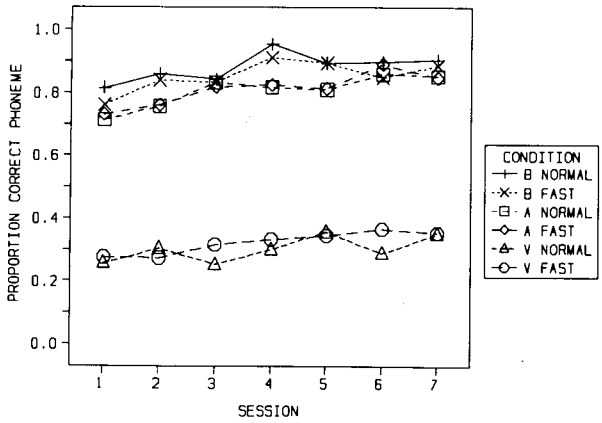

Figure 5. Average proportion correct of initial phoneme identifications on consonant-vowel syllables for bimodal, auditory, and visual conditions as a function of the seven sessions of testing for normal and fast presentation rates.

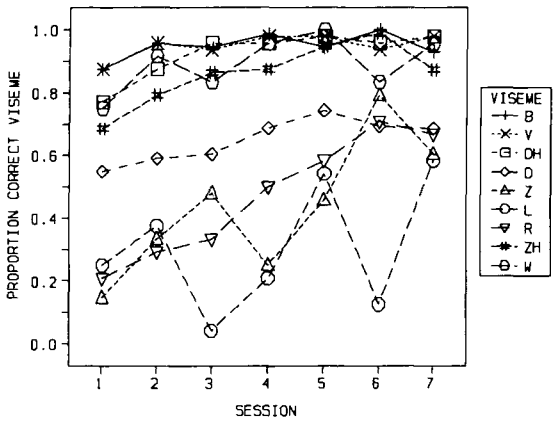

Figure 6. Average proportion correct identifications of the initial visemes of consonant-vowel syllables for the nine visemes as a function of the seven sessions of testing (pooled over normal and fast presentation rates). The visemes DH and ZH correspond to /ð/ and ž, respectively.

Table 2
The Average Proportion of Correct Phoneme Recognitions in the Syllable Test at Each of the Three Vowel Contexts for the Three Presentation Conditions

	Vowel		
Modality	/i/	/a/	/u/
Auditory	.755	.826	.843
Visual	.305	.344	.292
Bimodal	.850	.891	.861

identification was best for /a/ (.40), intermediate for /i/ (.33), and poorest for /u/ (.24). On the other hand, Greene, Kuhl, and Meltzoff (1988) found a larger influence of visible speech with the vowel /i/ than the vowel /a/. The vowel /u/ gave a minimal effect of visible speech. In summary, very little can be said about the effect of visible speech as a function of vowel environment.

Sentence Test

Figure 7 shows the proportion of words identified correctly as a function of test sessions for auditory, visual, and bimodal conditions at normal and fast rates. The overall proportion of correct identifications was .686, with .851 auditory, .306 visual, and .902 bimodal. Accuracy increased significantly from .570 to .751 over the first four test sessions [$F(3,15) = 67.92, p < .001$]. Performance was significantly better for normal rate (.828) versus fast rate (.544), and there were also significant interactions of session × rate, rate × modality, and session × rate × modality (all $ps < .001$).

To assess performance as a function of sentence length, the results were analyzed as a function of short and long sentences. The proportion of words reported correctly was larger for short (.74) than for long (.66) lengths [$F(1,5) = 62.26, p < .001$].

Training Results

Table 3 shows the performance on the eight discrimination pairs in the 2W2I training. Only two pairs, /d-l/ and /d-s/, were repeated once because of initial group scores below .900. Table 4 shows the performance on the six discrimination pairs in the 1W2I training.

Table 5 gives the average accuracy of performance in lipreading the visemes in the 1CV9I training. As can be seen, performance was very good. For the 4CV9I training, performance increased fairly rapidly from .570 in the first block to .756 in the last block. Table 5, which gives the mean performance in the last half of training, shows that performance was similar for the 1CV9I and 4CV9I training except that a somewhat greater number of confusions occurred when there were four possible talkers rather than just one. The mean proportions correct for the two tasks were .830 (1CV9I) versus .745 (4CV9I) [$F(1,5) = 23.57, p = .005$].

Retention

For the word and syllable tests, the seventh test session was run 7.5 weeks after the sixth test session. No additional training was given between these two sessions, to keep the retention test as pure as possible. As can be seen in Figures 2–6, there was no dramatic change in performance between the sixth and seventh test sessions. Thus, the learning that did occur appears to have been maintained for at least 7.5 weeks. For the word test, the overall proportion correct visemes did not change significantly between the sixth and seventh test sessions [$F(1,5) = .759$]. For initial phonemes, the proportion of correct identifications actually increased slightly from .789 to .808 [$F(1,5) = 10.946, p = .021$]. For the syllable test, the overall proportion of correct initial visemes for the visual condition showed some tendency for slightly poorer performance between Test Sessions 6 and 7 for the fast rate [$F(1,5) = 7.494, p = .040$]. This result is surprising, given that rate had no overall effect across the test sessions.

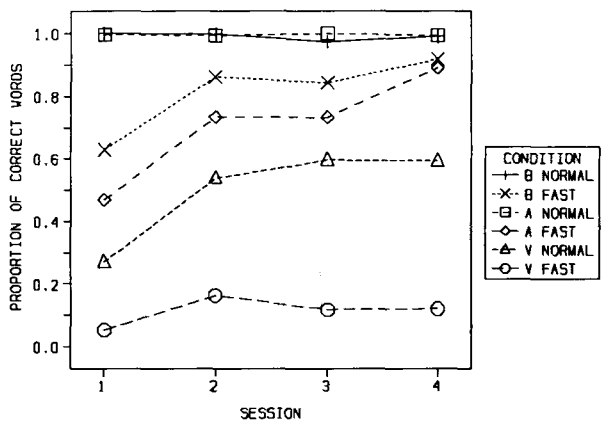

Figure 7. Proportion of correct word identifications in the test sentences, as a function of session, for bimodal, auditory, and visual conditions at normal and fast presentation rates.

Table 3
The Proportion of Correct Judgments as a Function of the Training Pair of Visemes and the Test Session in the 2W2I Training Task

		Subject					
Pair	Session	1	2	3	4	5	6
/d-l/	1	.907	.852	.920	.815	.932	.673
	2	.914	.907	.914	.883	.914	.747
/d-z/	1	1.000	.944	.914	.772	.870	.772
	2	1.000	.963	.914	.790	.994	.796
/z-ž/	1	1.000	.994	1.000	1.000	1.000	.988
/r-w/	1	1.000	.975	.963	.938	.938	.870
/b-d/	1	.994	1.000	.981	1.000	1.000	.975
/b-r/	1	1.000	1.000	1.000	1.000	1.000	.957
/d-ž/	1	.988	1.000	.939	1.000	.982	.944
/b-l/	1	.994	.994	.957	1.000	1.000	.988

Table 4
The Proportion of Correct Judgments as a Function of the Training Pair of Visemes and the Test Session in the 1W2I Training Task

		Subject					
Pair	Session	1	2	3	4	5	6
/d-l/	1	.969	.802	.728	.753	.821	.698
	2	.981	.877	.907	.821	.871	.642
/d-z/	1	1.000	.932	.901	.759	.969	.778
	2			.963	.796		
/z-ž/	1	.994	1.000	1.000	1.000	1.000	.975
/r-w/	1	1.000	.944	.895	.944	.944	.907
/b-d/	1	1.000	.981	1.000	1.000	.994	.951
/d-ž/	1	.988	.975	.951	1.000	.981	.895

Table 5
The Proportion of Correct Viseme Judgments in the 1CV9I and 4CV9I Training Tasks

	Subject					
Task	1	2	3	4	5	6
1CV9I	.890	.834	.794	.847	.802	.810
4CV9I	.830	.732	.782	.751	.701	.676

Unimodal and Bimodal Confusions: Data and Theory

We now consider in some detail the phoneme and viseme responses made in the syllable and word tests for the auditory, visual, and bimodal conditions. These results allow us to test extant models of speech perception. For the syllable identifications, the solid line circles in Figures 8 and 9 show typical phoneme and viseme responses, combined across rate, for the sixth word session and first syllable session, respectively. The lines in Figure 9 partition the phonemes into the nine different viseme categories.

The FLMP was tested against the confusion matrices combined over presentation rate, separately for the first and sixth test sessions. The theoretical assumptions were discussed in the introduction and the mathematical form of the model is presented in Massaro (1987) and Massaro and Cohen (1990). When the FLMP is applied to the auditory, visual, and bimodal syllable identification results, each of the unimodal sources is assumed to provide continuous and independent evidence for each of the response alternatives. We denote these feature values by A_{ji} and V_{ji} for the support of response alternative j given stimulus i, for the auditory and visual sources of information, respectively. In each unimodal condition, each stimulus s_i leads to a particular response r_j with the probability $P(r_j|s_i)$. For the auditory condition, we will call these probabilities $P(Ar_j|As_i)$ and similarly $P(Vr_j|Vs_i)$ for the visual. For the bimodal condition, we have $P(Br_j|As_iVs_i)$. These bimodal response probabilities can be predicted from fuzzy feature values, which give the degree to which the auditory and visual modalities support each alternative. The predicted probability of response j given stimulus i for the auditory presentation is

$$P(Ar_j|As_i) = \frac{A_{ji}}{\sum_j A_{ji}}, \tag{1}$$

and similarly for the visual presentation:

$$P(Vr_j|Vs_i) = \frac{V_{ji}}{\sum_j V_{ji}}. \tag{2}$$

The form of Equations 1 and 2 reflects the relative goodness rule (RGR) given by the decision operation in the FLMP. The probability of response j given stimulus i is equal to the support given j divided by the sum of the support given all relevant alternatives in the task (Massaro & Friedman, 1990). Note that the denominator of Equations 1 and 2 need not sum to one, because all of the response alternatives can receive varying degrees of support (feature values).

For the bimodal case, the multiplicative integration of the auditory and visual sources of information determines the support for alternative j. The predicted probability of a response, given the RGR, is thus equal to

$$P(Br_j|As_iVs_i) = \frac{A_{ji} \times V_{ji}}{\sum_j A_{ji} \times V_{ji}}. \tag{3}$$

Equation 3 predicts that the support for a bimodal alternative is the multiplicative combination of the two unimodal degrees of support for that alternative divided by the sum of the support for all of the relevant alternatives.

The quantitative predictions of the model are determined by using the program STEPIT (Chandler, 1969). A model is represented to the program in terms of a set of prediction equations and a set of unknown parameters. By iteratively adjusting the parameters of the model, the program minimizes the squared deviations between the observed and predicted points. The outcome of the program STEPIT is a set of parameter values that, when put into the model, come closest to predicting the observed results. Thus, STEPIT maximizes the accuracy of the description of a given model. We report the goodness-of-fit of a model by the root mean square deviation

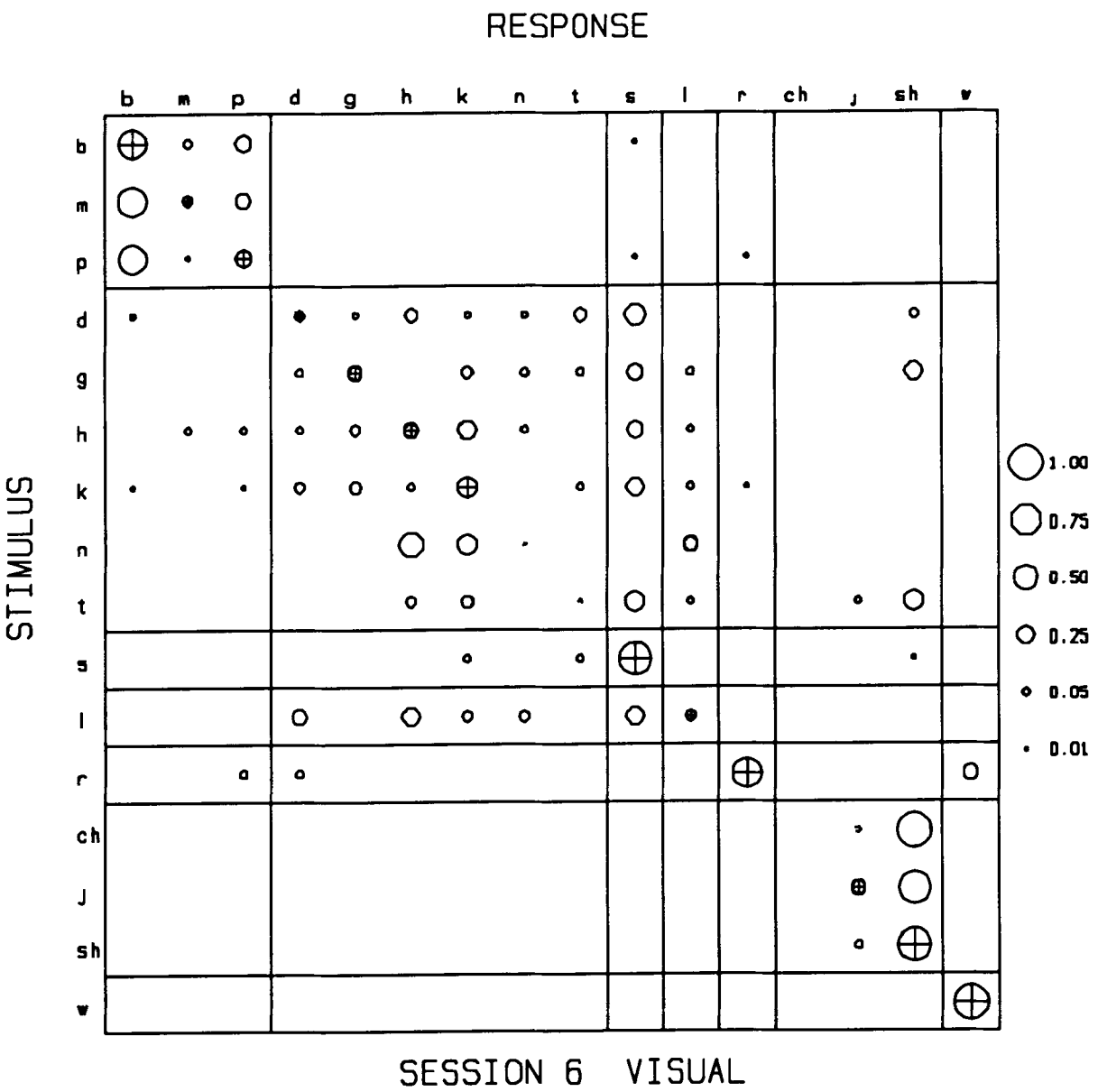

Figure 8. Observed (solid circles) and fuzzy logical model of perception predicted (dashed circles) initial phoneme responses for visual words in the sixth test session.

(RMSD)—the square root of the average squared deviation between the predicted and observed values.

Word test. The models were tested against the proportion of initial phoneme identifications for the word test stimuli. Because of some missing data cells, the data we examined for the word test have been condensed from 23×23 to 16×16 consonant responses. The solid line circles in Figure 8 give the observed initial phoneme responses for the visual words in the sixth test session. Given that dashed circles are rarely seen in the figures, the predictions of the FLMP usually fall on the observations given by the solid circles.

For the test of the FLMP, Equations 1-3 were used in conjunction with STEPIT to predict the 768 data points on the basis of 16×16 auditory and 16×16 visual parameters. The dashed circles in Figure 8 give the predictions of the FLMP. Although we would normally have 512 free parameters, the parameter space is fairly sparse with many near-zero feature support values, so we reduced this number by setting some of the feature values to 0 if they gave very little (less than .02 for Session 1 and less than .002 for Session 6) support for a given alternative. These constraints gave 157 free parameters for Session 1 and 153 for Session 6, respectively. The model provides a good

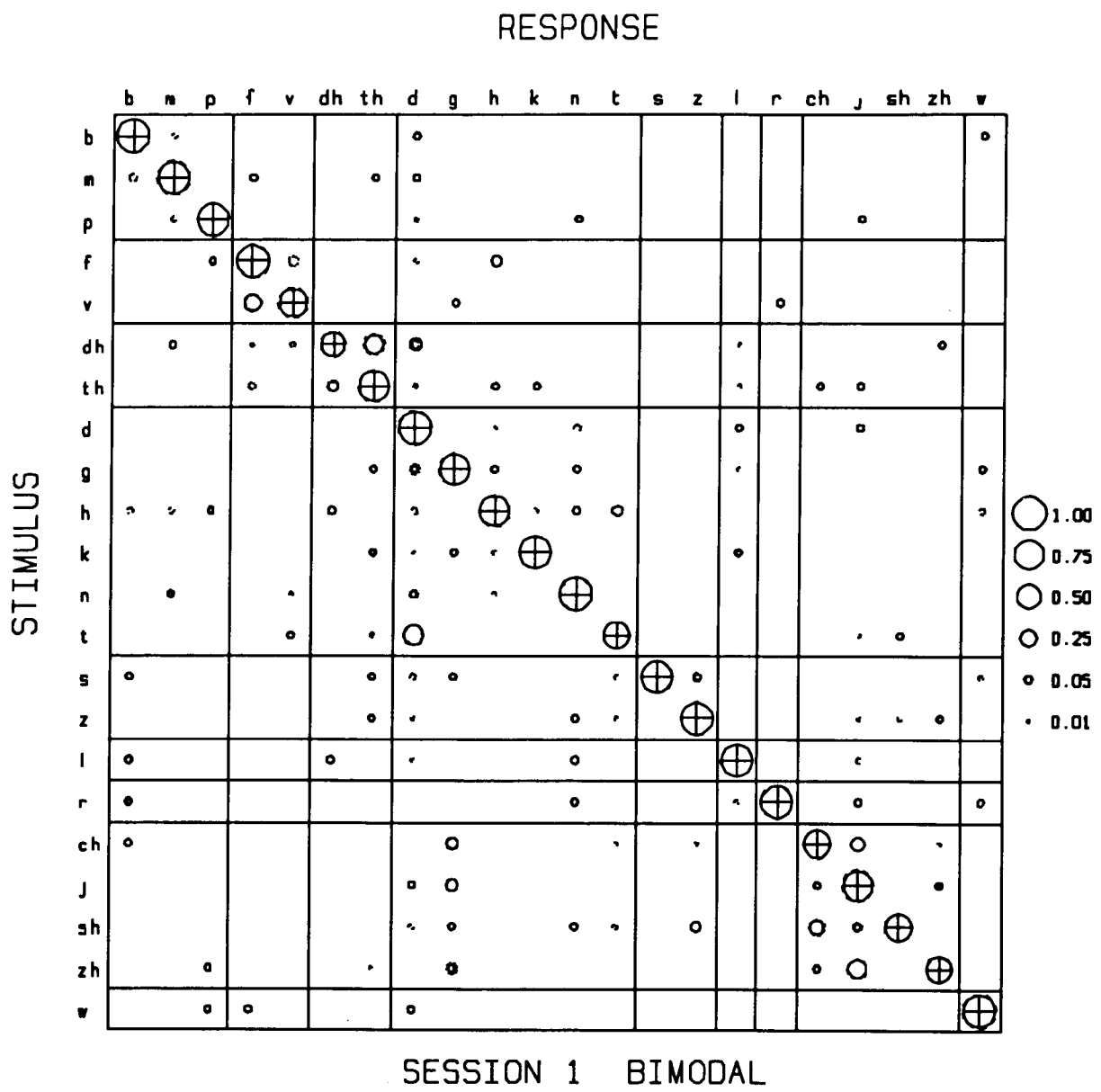

Figure 9. Observed (solid circles) and fuzzy logical model of perception predicted (dashed circles) initial phoneme responses for bimodal syllables in the first test session.

description of the identifications of both the unimodal and the bimodal syllables, with an RMSD of .0167 for the first session and .0167 for the sixth session. Given the equivalence in these two RMSDs, there is no evidence for Braida's (1991) conjecture that more familiar items will be processed more optimally.

The PRLM was also tested against these results. For each of the stimulus modalities (auditory and visual), each stimulus and response was represented as a three-dimensional point in space, thus having a six-dimensional bimodal representation. Thus, 3 (dimensions) × 32 centers (16 stimuli + 16 responses) × 2 modalities (auditory and visual) requires a total of 192 free parameters for the model, a number somewhat higher than that required for the FLMP. Because the PRLM involves the addition of noise to the stimuli, a closed solution is not possible in computing the predictions. Rather, numerical integration by a Monte Carlo technique is employed. To compute the confusion matrices, we first set the 768 response probabilities (256 for each modality condition) to 0 and reset the random number generator. Then, for each of the modalities and for each of the 16 stimuli, 1,000 simulated trials occurred. On each simulated trial, random deviates from a normal (Gaussian) distribution computed

by the *Box-Muller* method (Press, Flannery, Teukolsky, & Vetterling, 1988) with a standard deviation of 1 were added to the appropriate stimulus location, and the closest response center was computed, which then received 1/1,000 probability. This entire process was repeated until the RMSD was minimized. These fits resulted in RMSDs of .0291 and .0267 for Sessions 1 and 6, respectively, about 1.7 times worse than the FLMP.

Syllable test. The parameters of the current model consist of the 22×22 auditory and 22×22 visual feature values, used to predict the 1,452 observed data points. As with the word data, we were able to reduce the number of free parameters from the normal 968 by setting some of the feature values to 0 if they gave very little (less than .02) support for a given alternative. This constraint gave 323 free parameters for Session 1 and 174 for Session 6. The model provides a good description of the identifications of both the unimodal and bimodal syllables with an RMSD of .0171 for the first session and .0139 for the sixth session. Given that dashed circles are rarely seen in the figures, the predictions of the FLMP usually fall on the observation given by the solid circles. As with the words, the near equivalence in these two RMSDs provides no support for Braida's (1991) conjecture that more familiar items will be processed more optimally. Because the FLMP is an optimal model, the low RMSDs show that subjects were equally optimal before and after training. Figure 9 shows the observed and predicted responses to the bimodal syllables during the first test session. To give an index of the relationship of area to proportion in Figure 9, looking at $P(d|g)$ we have an observed value of .042 versus a prediction of .096. Similarly for $P(g|g)$ we have an observed value of .792 versus a prediction of .842. To give some measure of the deviations between predicted and observed values that would be noticeable in Figures 8 and 9, the difference between the predicted and observed values for $P(d|g)$ in Figure 8 is .054. The deviations between the predicted and observed responses in these two examples are higher than the average deviation. As can be seen in the figures, however, there are many cells in which the differences between observed and predicted values are close to 0. Also, many cells have both observed and predicted values near 0. Both of these results bring down the overall RMSD.

For the test of the PRLM, we once again employed a Monte Carlo approach. Using 3 (dimensions) × 44 centers (22 stimuli + 22 responses) × 2 modalities (auditory and visual) requires a total of 264 free parameters for the model, a number roughly equivalent to the average 249 parameters required for the FLMP. Once again, the obtained RMSDs (.0300 for Session 1 and .0228 for Session 6) are worse than the FLMP by a factor of about 1.7.

Interactive activation models have not been developed to account for stimulus-response confusions and would probably be unwieldy to test in this way. The good fit of the FLMP, however, is simultaneously evidence against interactive activation. One central assumption of the FLMP is that the auditory and visual sources of information are independent. Generic IAMs (interactive activation models), on the other hand, assume crosstalk between sources of information so that the activation of one modifies the activation of the other. It follows that the activation of the auditory representation should differ as a function of whether or not the visual source is present. If this were the case, the FLMP should not have been capable of describing the stimulus-response confusions, because no crosstalk occurs in this model.

In summary, both the prelabeling model and interactive activation models cannot account for the stimulus-response confusions. The FLMP, on the other hand, gives a good description of identification accuracy and responses among the consonants across improvements in lipreading skill.

DISCUSSION

The findings from the present study indicate that some segments of speech are easier to lipread than others. These segments can be grouped into viseme groups in which the segments within a group are not easily discriminable from one another (Owens & Blazek, 1985). A finding that replicated the Gesi et al. (1992) study is that there is a significant difference among the different visemes. Some viseme classes are more easily perceived by eye than are other classes. Segments produced at the front of the mouth are more easily lipread than sounds produced at the back.

The observed improvements in lipreading in the present study have roughly the same magnitude as in previous studies (Dodd et al., 1989; Walden et al., 1981; Walden et al., 1977). The wide range of designs used to study the teaching of lipreading and the consistent finding of improved performance across training for each of these different designs suggest that lipreading can be taught to some degree (e.g., Gesi et al., 1992; Walden et al., 1977). However, it is not clear from these studies, including the present study, what aspects of the training and experience account for this improved performance. The repeated testing experience could have been as beneficial as the training procedures because it has been shown that tests can be potent learning events (Schmidt & Bjork, 1992). The design of these studies precludes isolating the specific aspects of training and experience that are responsible for the improvements in lipreading skill. It appears, however, that explicit instruction might not be as beneficial as one would expect. Gesi et al. (1992) found that an expository method of teaching in which subjects were told where to look and what to look for had no advantage over a discovery method in which subjects were given no explicit instructions. In Gesi et al.'s study, both groups of subjects were simply given auditory feedback paired with the visible speech segment presented on each trial. In the present study, it was not necessary to train lipreading with bimodal syllables. The subjects practiced on visible speech without sound. Given the substantial support

for the FLMP assumption of independent evaluation of the A and V sources, it should not be critical whether training is carried out with visual or bimodal speech.

It should be noted too that subjects also improved in their auditory identification as well as in their lipreading. Each modality (auditory, visual, bimodal) showed similar improvements in performance across training. This result should not be surprising, if it is accepted that processing visible speech is as common and natural as processing sound. If gains are made in one modality, it should not be surprising if gains should also be made in the other.

The rate of presentation of the test items was important for the word but not the syllable test. Presenting the items at a fast rate interfered with performance only in the word test. This result might be explained in terms of the perceptual processing time required for speech perception (Massaro, 1972). There is evidence that backward masking of one segment can occur when a second segment is presented before the first is recognized. With CV syllables, the CV segment can be processed during the blank period following its presentation—even when the segment is presented at a rapid rate (Massaro, 1974). For the words, however, the final VC segment can interfere with processing of the initial CV syllable, and this backward masking would be particularly damaging at the fast rate of presentation. Thus, the effect of rate of presentation for words but not for syllables is consistent with the importance of perceptual processing time in speech perception and extends previous findings in auditory speech perception to visual and bimodal speech perception.

Finally, the accounting for improvements in lipreading and auditory speech perception, and their concomitant contribution to bimodal speech perception, provided a new and unique means of testing the FLMP. The FLMP is based on the assumption that the unimodal sources provide continuous and independent evidence for each response alternative. Bimodal speech perception involves the integration of the unimodal sources. Insofar as perceivers learn more about the unimodal sources, their bimodal speech perception should also improve. Confusion matrices were used to measure unimodal and bimodal speech perception and any improvement across training. The FLMP not only provided an excellent account of the confusion matrices, but also described the joint improvement in unimodal and bimodal speech perception. With regard to practice, the present research was successful in showing that improved lipreading ability facilitates bimodal speech perception. Therefore, training in lipreading is worthwhile even if the perceiver normally perceives speech bimodally.

Schmidt and Bjork's (1992) review of practice effects reveals that there are some general principles of practice and learning that generalize across specific domains, as, for example, commonalities in motor and verbal learning. For this reason, we expect that our findings in speech perception are not limited and should be of general interest to perceptual psychologists. The FLMP has previously been shown to describe performance across a wide variety of domains (Massaro, 1987, 1992), and the present research demonstrates its predictive power at several levels of skill. This demonstration was not possible in previous studies, because investigators did not measure stimulus-response confusions during learning.

REFERENCES

BERNSTEIN, L. E., & EBERHARDT, S. P. (1986). *Johns Hopkins lipreading corpus I-II: Disc I* [Videodisk]. Baltimore, MD: The Johns Hopkins University.

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

CHANDLER, J. P. (1969). Subroutine STEPIT-Finds local minima of a smooth function of several parameters. *Behavioral Science*, **14**, 81-82.

COHEN, M. M., & MASSARO, D. W. (1992). On the similarity of categorization models. In F. G. Ashby (Ed.), *Probabilistic multidimensional models of perception and cognition* (pp. 395-447). Hillsdale, NJ: Erlbaum.

DODD, B. (1977). The role of vision in the perception of speech. *Perception*, **6**, 31-40.

DODD, B., PLANT, G., & GREGORY, M. (1989). Teaching lip-reading: The efficacy of lessons on video. *British Journal of Audiology*, **3**, 229-238.

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

GESI, A. T., MASSARO, D. W., & COHEN, M. M. (1992). Discovery and expository methods in teaching visual consonant and word identification. *Journal of Speech & Hearing Research*, **35**, 1180-1188.

GREENE, K. P., KUHL, P. K., & MELTZOFF, A. N. (1988). Factors affecting the integration of auditory and visual information in speech: The effect of vowel environment. *Journal of the Acoustical Society of America*, **84**, S155.

HOUSE, A. S., WILLIAMS, C. E., HECKER, M. H. L., & KRYTER, K. (1965). Articulation-testing methods: Consonantal differentiation with a closed response set. *Journal of the Acoustical Society of America*, **37**, 158-166.

HUTTON, C. (1959). Combining auditory and visual stimuli in aural rehabilitation. *Volta Review*, **61**, 316-319.

LUCE, R. D. (1959). *Individual choice behavior*. New York: Wiley.

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

MASSARO, D. W. (1972). Preperceptual images, processing time, and perceptual units in auditory perception. *Psychological Review*, **79**, 124-145.

MASSARO, D. W. (1974). Perceptual units in speech recognition. *Journal of Experimental Psychology*, **102**, 199-208.

MASSARO, D. W. (1984). Children's perception of auditory and visual speech. *Child Development*, **55**, 1777-1788.

MASSARO, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Hillsdale, NJ: Erlbaum.

MASSARO, D. W. (1992). Broadening the domain of the fuzzy logical model of perception. In H. L. Pick, Jr., P. Van den Broek, & D. C. Knill (Eds.), *Cognition: Conceptual and methodological issues* (pp. 51-84). Washington, DC: American Psychological Association.

MASSARO, D. W., & COHEN, M. M. (1983). Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception & Performance*, **9**, 753-771.

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

MASSARO, D. W., & FRIEDMAN, D. (1990). Models of integration given multiple sources of information. *Psychological Review*, **97**, 225-252.

MASSARO, D. W., THOMPSON, L. A., BARRON, B., & LAREN, E. (1986). Developmental changes in visual and auditory contributions to speech perception. *Journal of Experimental Child Psychology*, **41**, 93-113.

MCGURK, H., & MACDONALD, J. (1976). Hearing lips and seeing voices. *Nature*, **264**, 746-748.

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

NEELY, K. K. (1956). Effects of visual factors on intelligibility of speech. *Journal of the Acoustical Society of America*, **28**, 1276-1277.

O'NEILL, J. J. (1954). Contributions of the visual components of oral symbols to speech comprehension. *Journal of Speech & Hearing Disorders*, **19**, 429-439.

OWENS, E., & BLAZEK, B. (1985). Visemes observed by hearing-impaired and normal-hearing adult viewers. *Journal of Speech & Hearing Research*, **28**, 381-393.

PRESS, W. H., FLANNERY, B. P., TEUKOLSKY, S. A., & VETTERLING, W. T. (1988). *Numerical recipes: The art of scientific computing*. Cambridge: Cambridge University Press.

SCHMIDT, R. A., & BJORK, R. A. (1992). New conceptualization of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, **3**, 207-217.

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

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

WALDEN, B. E., ERDMAN, S. A., MONTGOMERY, A., SCHWARTZ, D. M., & PROSEK, R. A. (1981). *Journal of Speech & Hearing Research*, **24**, 207-216.

WALDEN, B. E., PROSEK, R., MONTGOMERY, A., SCHERR, C. K., & JONES, C. J. (1977). Effects of training on the visual recognition of consonants. *Journal of Speech & Hearing Research*, **20**, 130-145.

ZADEH, L. A. (1965). Fuzzy sets. *Information & Control*, **8**, 338-353.

APPENDIX
The Six Groups of Words Used in Word Test

1		2		3		4		5		6	
sing	thaw	sin	law	pin	rake	din	Jake	bin	shake	dung	ray
pack	rave	kin	raze	pus	race	name	rate	win	came	tin	sake
thin	gale	shin	male	gin	pale	chin	shale	lin	jail	sip	bale
tip	tale	sit	park	hit	mark	sick	shark	lick	lark	thick	dark
nick	bark	hick	hark	chick	rent	Rick	dent	wick	gent	kick	tent
fit	went	pit	sent	tig	bent	fib	toil	fizz	foil	bit	coil
bet	boil	bait	soil	pip	oil	wit	hold	kit	fold	kith	told
kid	cold	lid	gold	guide	sold	quid	sag	rid	sat	mid	sap
hid	zap	bid	sack	king	sass	kiss	sad	rig	pan	wig	pad
jig	pat	pig	pang	big	pass	fig	tab	tick	tan	dip	tack
rip	tang	lip	tam	hip	tap	zip	badge	chip	bat	ship	bought
gyp	bass	dim	back	ditch	bad	dish	ban	fin	sang	dig	gang
did	rang	dill	hang	sill	bang	pill	fang	shill	mass	bill	math
hill	madge	fill	mad	will	map	kill	man	till	mat	look	game
cook	dame	nook	shame	shook	same	took	fame	hook	tame	book	sane
mop	safe	top	sale	lop	save	bop	cane	shop	cave	chop	shave
cop	cape	hop	cake	pop	case	rust	gay	lust	say	just	way
dust	may	must	day	gust	lane	bust	laze	gun	lace	run	lake
Hun	lame	shun	lay	nun	late	sun	pace	fun	page	sub	pave
sung	vest	sop	they	suck	nest	such	best	sud	rest	sup	zest
bum	test	sum	west	putt	den	pick	hen	puck	ten	pub	zen
pup	pen	puff	men	pun	then	pug	fed	dub	shed	dun	bed
dud	red	dull	wed	duck	led	dug	said	bus	blue	buff	clue
bun	Jew	buck	shoe	but	glue	buzz	due	boot	rue	Bert	true
bug	brew	cuff	screw	cud	crew	cup	grew	cub	drew	cut	shrew
cuss	woo	hot	loo	tot	zoo	lot	moo	not	boo	got	too
pot	couch	sot	crouch	shot	ouch	jot	pouch	yacht	grouch	health	vouch
heave	ground	heal	bound	heath	found	hear	hound	heap	round	jeep	sound
cheap	cite	sheep	quite	peace	mite	peach	kite	peal	rite	peak	bight
peat	blight	peas	bright	team	fight	tease	flight	teach	fright	teak	pay
teal	pane	beach	bow	bean	flow	bead	blow	beak	mow	beam	row
tear	throw	keel	bite	reel	crow	feel	glow	peel	know	eel	low
heel	slow	meat	stow	feat	sew	beat	show	heat	Joe	neat	tow
sheet	boat	seep	coat	seem	goat	seen	throat	seat	moat	seed	batch
seethe	bash	seek	bath	raw	match	paw	catch	jaw	cache	saw	latch

(Manuscript received June 15, 1992;
revision accepted for publication September 22, 1992.)