Auditory-Visual Speech Processing
(AVSP'98)
Terrigal - Sydney, Australia
December 4-6, 1998

DOES TRAINING ENHANCE VISUAL SPEECH PERCEPTION?

L. E. Bernstein, E. T. Auer, Jr., & P. E. Tucker

House Ear Institute, Los Angeles, California 90057, USA

ABSTRACT

This study investigated whether visual speech perception can be improved with short-term training. This paper first reviews the literature on lipreading training. Then a study is reported that involved 6-7 hours of lipreading training within at most three weeks. In Exp. I, subjects were adults with impaired hearing (IH) or with normal hearing (NH). They alternated training and testing on sets of prerecorded isolated sentences. Exp. II replicated Exp. I, except that subjects also received vibrotactile speech signals during training. It was hypothesized, based on previous studies, that this might promote learning. Evidence was obtained for learning in both experiments during initial test periods, primarily for the NH groups in both experiments. However, the IH group was overall more accurate. The training effects and the differential performance of IH versus NH groups imply the need to take perceptual learning and experience into account in evaluating practical applications involving visible speech.

1. INTRODUCTION

A fundamental question about lipreading (speechreading) is whether performance can be improved through training. If so, what is the time course of learning? Does the subjects' audiological status and history affect performance? Is there transfer of learning across talkers? Answers to these questions have implications for theories of visual speech perception (and speech perception generally) and for practical and clinical applications.

1.1. Previous Lipreading Training Studies

During much of this century, whether people could be effectively trained to lipread, and whether other abilities predict lipreading performance have been important issues for clinicians and educators of deaf children in the United States [1]. In 1940, Heider and Heider [2] reported on the lipreading performance of large groups of children from the Clarke School for the Deaf. Across a five-year study, they found that the children's relative proficiency levels did not change. Although lipreading was the main avenue for deaf children to acquire language and general knowledge, many children failed to develop excellent lipreading despite intensive training [3, 4]. As a result, it has frequently been concluded that good lipreaders are born and not made. However, there have also been some experimental studies suggesting that training can improve performance.

Walden et al. [5] investigated identification of 20 prerecorded CV syllables, before and after training on CV syllable discrimination and identification tasks presented live voice by several clinicians. Adult males with sensorineural hearing impairments and a similar group of controls participated. Before training, responses clustered into five groups of mutually confused phonemes (i.e., visemes), and after training, into nine groups. An analysis of the results obtained after 14 hours of training showed that most of the change took place during the first few hours. Controls showed essentially no change.

Walden et al. [6] studied adult males with mostly high-frequency sensorineural hearing impairments, enrolled in a 50-hour program of aural rehabilitation. Men were assigned randomly to either the rehabilitation program only or rehabilitation with seven hours of consonant training. Training involved listening to or lipreading nonsense syllables. Pre- and post-training tests employed pre-recorded sentence and VCV materials. The lipreading training resulted in a small but significant improvement in identifying consonants within viseme groupings. However, there was not a control for the lipreading training on consonants, so improvement could be attributed to simply retesting the materials. Both auditory and visual experimental training protocols resulted in significant improvements (approximately 26 percentage points) in identifying words in audiovisual sentences, an improvement significantly beyond that obtained by the control group. Subjects were not tested on sentence lipreading, so it is unknown whether the CV training generalized to sentences.

Gesi et al. [7] compared two training methods (each required three hours of participation), one in which the subjects (undergraduates with normal hearing) were told information about visible speech characteristics and the other in which they were told only how to perform the task, which was forced choice identification of CV syllables with feedback. After a four-week break, subjects returned for another three-day training set. Results showed an approximately 20-percentage point improvement in CV identification (see figures in [7]) that was not dependent on training method. Three weeks later, subjects received a transfer task in which monosyllabic words were presented for forced choice identification. Transfer was not demonstrated.

Massaro et al. [8] conducted a study over several months. Training involved word and syllable stimuli. Testing involved monosyllabic words, CVs, and sentences [9], repeated seven times across the experiment: once at the beginning, once after each of five courses of training, and once after a retention period of 7.5 weeks. Conditions were audio, video, and audiovisual. Words were scored in terms of initial viseme correct and initial phoneme correct. Small improvements across lipreading test sessions were observed. Some of this improvement might be attributable to remembering the words from the highly intelligible audiovisual conditions. The syllable test also showed improvements across test sessions but this could have been due to learning stimulus tokens. The sentence tests showed approximately 20 percentage point increases between the first and second sessions and little change after that. However, the same 96 sentences were presented under all conditions and were perfectly intelligible under the audiovisual condition. Improved scores could have been due to remembering the sentences from the high intelligibility conditions.

1.2. Training with Vibrotactile Speech Aids

Studies on vibrotactile devices to enhance lipreading by deaf people provide additional information on lipreading training. These studies frequently employ sentence stimuli across a series of aided and unaided conditions. Researchers assume that learning the vibrotactile speech stimuli will take a long time, and as a result, these studies provide information about lipreading across many observations. Almost inevitably, lipreading *without* the aid is shown to improve [e.g., 10, 11, 12, 13, 14, 15].

Boothroyd et al. [12] reported on one study in which eight IH (greater than 90 dB hearing thresholds) adults participated. Each received three months of training, four hours a week. Training employed sentences and phonemes under conditions of lipreading alone and with a multichannel vibrotactile aid, which presented voice fundamental frequency (F0). The results for lipreading of isolated sentences showed reliable improvements across time in both lipreading alone and lipreading with the vibrotactile aid. In another study, three of the IH adults returned for a replication of the first study, but with a different type of F0 vibrotactile stimuli. Learning occurred for both aided and unaided tests.

Kishon-Rabin et al. [14] studied four IH adults who received training on vibrotactile stimuli, including identification of frequency contours, identification of stressed words, vibrotactile feedback with the subject's own voice, identification of words in video-recorded sentences, and training with live voice connected speech. Training for three weeks with the aid and for one week by lipreading alone was repeated two or three times, depending on the subject. The results showed learning for both conditions by two subjects, learning for only the aided lipreading by one subject, and no learning for the fourth subject.

Eberhardt et al. [13] studied lipreading with four different F0 vibrotactile aids. Groups of NH adults were assigned to each of the F0 conditions or to a lipreading-only control condition. Training and testing with the vibrotactile aid involved five alternations of lipreading alone and aided lipreading, except for the controls who received no vibrotactile stimuli. Stimuli were isolated sentences [9] presented once only. Feedback was given throughout the experiment. Learning took place in unaided conditions across all subjects and was statistically confirmed in 12 out of the 15 individual subjects. Different CID Everyday Sentences were tested at pre- and post-test. A significant learning effect was obtained, with pre-test mean of 24.9 percent words correct and post-test mean of 35.3 percent words correct across all subjects. One set of CV nonsense syllables was identified at pre- and post-test, and identification improved significantly by 3.7 percentage points. One set of sentences was repeated five times in the lipreading only conditions. At the first testing, the mean was 25.4 percent words correct. At the fifth presentation, the mean was 64.4 percent words correct. For the sentences that were not repeated during lipreading only, the comparable means were 21.8 percent and 37.5 percent words correct. That is, repetition of the same sentence, although across a long period of training (36 sessions across four to five months), resulted in elevated scores. Repetition of materials increased scores beyond the learning that generalized across materials.

Bernstein et al. [10] studied lipreading with three different multichannel vibrotactile aids that provided spectral speech information. Adult subjects were pre-screened to be average-or-better lipreaders. Five had profound hearing impairments acquired by age four years. Eight had normal hearing. Training and testing required 65-70 hours of one-hour sessions per subject. Training employed identification of prerecorded words in sentences and consonants in nonsense syllables, and identification of words spoken live in connected text. Testing employed the same tasks, excluding consonant identification. Two control subjects received no vibrotactile stimuli. Learning occurred in testing and training for both aided and control subjects. Different CID Sentences tested at pre- and post-test times resulted in significant improvements of approximately 10 percentage points across subjects. CV syllable identification across pre- and post-tests improved a statistically significant 3.2 percentage points.

1.3. Summary: Training Studies

The literature on lipreading training provides evidence that CV nonsense syllable identification can be improved with short-term training [5,6,7,8]. However, the evidence is equivocal in regard to transfer of CV learning to word and sentence-length stimuli. The vibrotactile speech aid literature provides evidence that extensive experience with sentence-length stimuli results in improved performance, with and without the use of a vibrotactile aid [e.g., 10, 11, 12, 13, 14, 15]. These studies do not inform us about short-term learning effects for sentence-length stimuli. Such effects are potentially important for assessing practical applications involving visible speech. The evidence from the vibrotactile aid studies suggests the possibility that vibrotactile speech stimuli might enhance or accelerate learning to lipread. Finally, comparisons across IH and NH groups suggest that these groups might differentially benefit from training.

2. THE CURRENT STUDY

Several questions were investigated: 1) Does short-term training on sentence stimuli result in learning? 2) Does training (in Exp. II) with a vibrotactile speech aid accelerate or enhance learning to lipread? 3) Do learning effects differ across NH vs IH adults?

2.1. Experiments I and II: Method

IH subjects. Eight IH subjects for each experiment were screened for the following characteristics: (a) aged 18 to 40 years; (b) Gallaudet University student; (c) sensorineural hearing impairments greater than 80 dB HL average in the better ear across the frequencies 500, 1000, and 2000 Hz;[1] (d) no self-report or university record of disability other than hearing impairment; (e) self-reported use of spoken English as the primary language of the subject's family; (f) self-report of English (including a manually coded form) as the subject's native language; (g) education in a mainstream and/or oral program; (h) self-report of good lipreading skills; and (i) vision at least 20/30 in each eye. Subjects were paid by the hour.

NH subjects. Eight NH subjects for each experiment were recruited from the Gallaudet University community. They had English as a first language, self-report of good lipreading skill, and normal or corrected-to-normal vision. Subjects were paid by the hour.

Design and materials. A multiple single-subject design was used in which training alternated with testing. Sentence stimuli were produced by a male and a female who spoke General American English [9]. Table 1 shows the schedule of training and testing. The B-E Sentences (female talker), B-E Sentences (male talker), and CID Sentences were sorted into lists with equal expected means based on previous results [16]. Training sentences were randomly assigned to lists. Subjects never saw the same stimulus twice. Pre-test and post-test sentence sets were counterbalanced across subjects.

Session	Sentence Sets
1	**Practice:** 20 Sentences **Pre-Test:** a) 50 B-E Sentences, female talker b) 10 CID Sentences, male talker c) 5 B-E Sentences, male talker
2-5	**Training:** 66 B-E Sentences **Test:** a) 10 CID Sentences, male talker b) 5 B-E Sentences, male talker **Training:** 66 B-E Sentences, male talker **Test:** a) 10 CID Sentences, male talker b) 5 B-E Sentences, male talker
6	**Training:** 66 B-E Sentences, male talker **Post-Test:** a) 50 B-E Sentences, female talker b) 10 CID Sentences, male talker c) 5 B-E Sentences, male talker

Table 1. Training and testing schedule.

Procedures. The subject sat at a small table in a darkened, sound attenuating room. A videodisc player, controlled by a personal computer, was used to instruct the subject prior to testing, to present stimuli, and to record responses. Subjects pressed a key to see the first stimulus and pressed the return key following each subsequent stimulus and response. Subjects were instructed to type exactly what they thought the talker had said. Partial responses were encouraged. During training, the correct sentence was printed on the computer screen following each response. During testing, no feedback was given, and subjects gave confidence ratings on each response.

Experiment I versus II procedures. Experiments were identical except that in Exp. II, during training, the subjects also received vibrotactile

[1] In reviewing audiological records after the experiment, it was found that two subjects had less severe impairments, one had 77-dB HL and the other 78-dB HL better pure tone averages.

vocoder speech stimuli. The vocoder is a bank of bandpass filters whose output is the energy level passed by each of the bands. The vibrotactile array was 16 solenoids that vibrated at an amplitude proportional to the energy passed by the corresponding filter band. Use of this vocoder has been shown [10] to enhance lipreading of sentences.

Measures. Several measures were obtained: 1) Mean proportion words correct; 2) Mean proportion phonemes correct; 3) Mean proportion phonemes incorrect; and 4) Mean incorrect phoneme distance. Proportion words correct was calculated across sentence sets for each subject. Arcsine transformed scores were used for statistical analyses of proportion words correct. The other three measures were the result of phoneme-to-phoneme sequence comparison that produced alignments and distance measures [17] for each response. Phonemes correct was the exact matches between stimulus and response phonemes. Phonemes incorrect was the alignments of incorrect response phonemes with stimulus phonemes. Phonemes correct and phonemes incorrect were normalized on the number of stimulus phonemes in each sentence. Incorrect phoneme distance was the estimated perceptual distance between stimulus phonemes and incorrect phonemes, and was normalized on the number of response phonemes plus stimulus phoneme deletions.

Results: Experiments I and II

Due to space limitations, we focus on CID Sentence and B-E Sentence (female talker) results. Omnibus repeated measures analyses of variance showed that there were not significant effects of Exp. I versus II. Results were combined across experiments.

Figs. 1-4 show the mean CID Sentence scores across groups for each of the four measures across the ten test periods, taking into account Test-1 versus Test-10 counterbalancing. Repeated measures analyses of variance, with test as the within subjects factor and group (IH versus NH) as the between subjects factor, showed that group was significant for all the measures (range of p = .011 to p = .000): The IH group was overall more accurate in terms of words and phonemes correct; They had fewer phonemes incorrect and; Their errors overall had smaller mean perceptual phoneme-to-phoneme distances. Test was significant for all measures (p = .000). Except for proportion words correct, the interaction between test and group was also significant [phonemes correct, [F(9, 252) = 2.72, p = .022; phonemes incorrect F(9,252) = 3.15, p = .001; and incorrect phoneme distance, F(9,252) = 2.12, p = .028].

Simple contrasts for the main effect of test for the *mean proportion words correct* measure showed that Test 1 was significantly lower than each of Tests 2, 3, 4, 7, 8 (range of p = .000 to p = .045). For *mean proportion phonemes correct*, the simple main effect of the first test versus the others showed that Test 1 was significantly lower than each of Tests 2, 3, 5, 6, 7, and 8 (range of p = .000 to p = .006). There was a significant Test 1 to Test 2 increase for the IH group. Increases from Test 1 to Tests 2, 3, and 4 were each significant for the NH group. For *mean proportion phonemes incorrect*, only Test 9 differed from Test 1, and this varied across groups (p = .045). It is not known why Test 9 caused so many errors for the NH group. For *mean incorrect phoneme distance*, for the main effect, every test (except Test 4) had a significantly smaller distance than Test 1 (range of p = . 036 to p = .000). The interaction of group with test was due to no difference between Test 1 and either Test 3 or Test 9 for the IH group, in contrast with reliable decreases for the NH group. The analyses of the interaction effects suggest that the NH group learned more than did the IH group.

Table 2 shows the Pearson correlations among the phoneme measures for the CID Sentence tests. The correlations between phoneme distance and phonemes incorrect, although significant, account for little variance. This result, in combination with Figs. 3 and 4, suggests that the drop in perceptual distance over time is relatively independent of the number of incorrect phonemes in responses: Reduced perceptual distance implies improved phonetic perception even though the phoneme error rate remained the same.

IH Group	Distance	Phonemes Correct
Phonemes Correct	-.667**	
Phonemes Incorrect	.135**	.-494**
NH Group		
Phonemes Correct	-.682**	
Phonemes Incorrect	.073**	-.483**

Table 2. Correlations among CID Sentence phoneme measures. **= p < .01.

Regression analyses were conducted on the data from individual subjects. The number of individuals with significant upward trends in Exp. I for the proportion words correct scores was: one IH (p = .016) and two NH (marginally, p < .093). In the same experiment, the IH subject had a significant upward linear trend for phoneme scores. In Exp. II, two NH individuals had significant upward trends for the proportion words correct scores (p < .040).

No significant trends for phonemes scores were obtained in Exp. II.

The results from the pre- and post-test B-E Sentences (female talker) were examined with analysis of variance. Group (IH versus NH) and test were significant main effects [respectively, F(3,84) = 6.83, F(1,28) = 8.17, p <.008]. Simple comparisons across test times showed a significant improvement of approximately 4.5 percentage points between the first pre-test and the second post-test [F(1,28) = 18.04, p = .000] (see Fig. 5). Regression analyses on individuals showed that one IH subject and four NH subjects had significant upward linear trends across tests (p < .05).

3. DISCUSSION

Summarizing the results: Analyses of variance for mean proportion words correct and mean proportion phonemes correct suggest that most learning took place during the initial days of training. However, regression analyses on individuals revealed that learning continued throughout the experiment for a few subjects. Greater learning was observed among the NH subjects, however IH subjects were significantly more accurate overall. There was no evidence that the vibrotactile speech signals accelerated or enhanced test performance. Testing using the female talker's speech, viewed only at pre- and post-test showed reliable improvements across time, suggestive of task learning and/or carryover from training with the male talker's sentences.

Learning in the early part of the experiments is probably indicative of learning to perform the task and learning something general about the sentences presented. Learning probably also reflects perceptually extracting additional phonetic information from the stimuli. This possibility is supported by the mean incorrect phoneme distance measures. Reductions in estimated perceptual distance between the stimulus and aligned incorrect response phonemes suggests that subjects' errors became based on more fine-grained phonetic information. The continued learning throughout the experiment by a few individuals suggests that there are individual differences in the ability to benefit from training.

The results demonstrate that for individuals who do not rely on vision for speech communication, the NH group, several hours of experience may be required before performance asymptotes. Nevertheless, the asymptotic levels achieved are generally below those of proficient lipreaders in the IH group. Group differences in perceptual experience need to be considered for experimental designs and for practical applications.

(This research was supported by the NIH/NIDCD DC02107.)

Fig. 1: Mean Proportion Words Correct

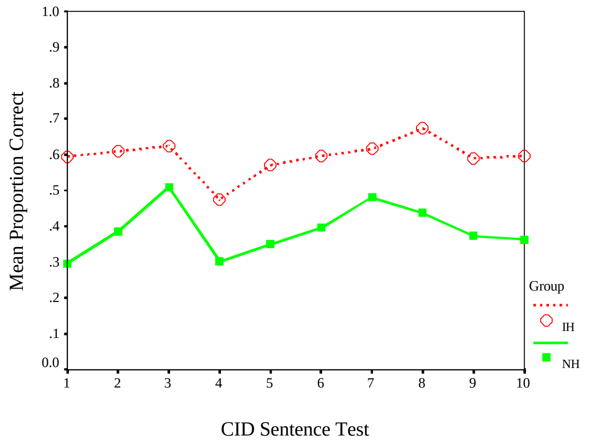


Fig. 2: Mean Proportion Phonemes Correct

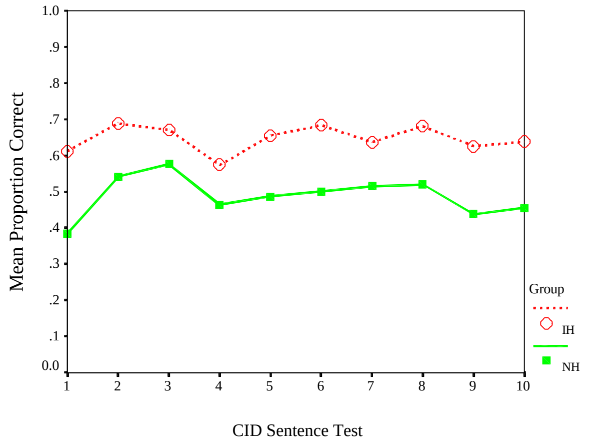


Fig. 3: Mean Proportion Phonemes Incorrect

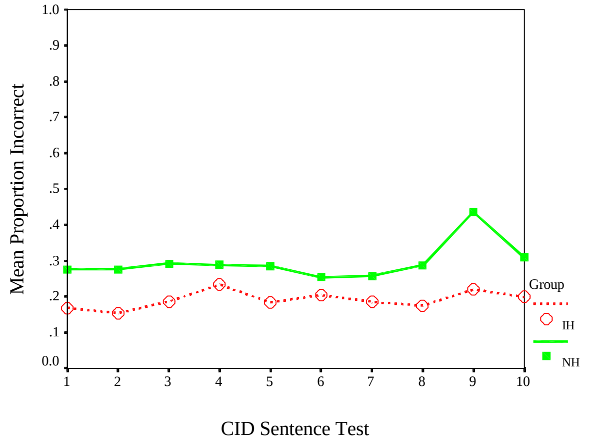


Fig. 4: Mean Incorrect Phoneme Distance

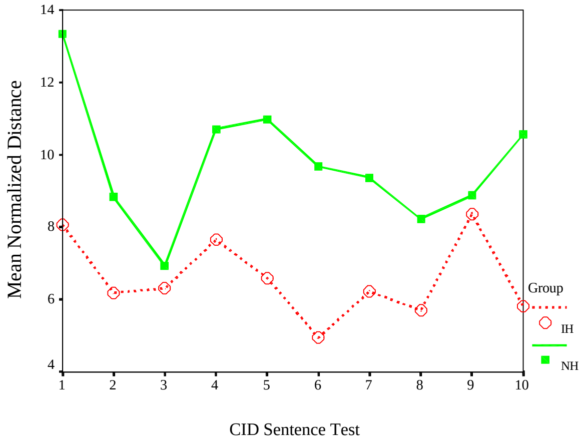
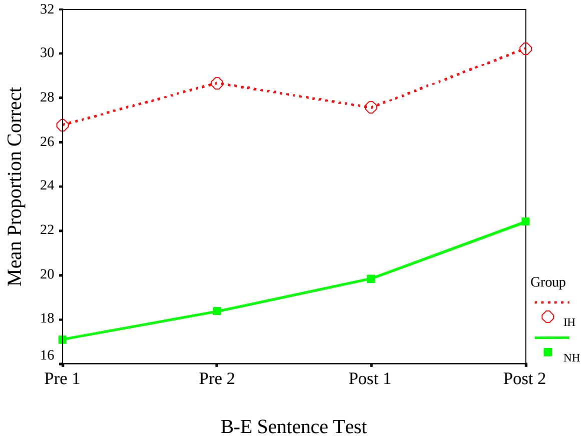


Fig. 5: Proportion Words Correct (Female Talker)



4. REFERENCES

1. Jeffers, J. and Barley, M. Speechreading (Lipreading), Charles C. Thomas, Springfield, IL, 1971.

2. Heider, F. and Heider, G. M., “An experimental investigation of lipreading”, Psychol. Monogr., Vol. 52, 1940, pp. 124-153.

3. Rönnberg, J. “Perceptual compensation in the deaf and blind: Myth or reality?”, In R. A. Dixon and L. Bäckman (Eds.), Compensating for Psychological Deficits and Declines, Erlbaum, Mahwah, NJ, 1995, pp. 251-274.

4. Summerfield, Q. “Visual perception of phonetic gestures”, In I. G. Mattingly and M. Studdert-Kennedy (Eds.), Modularity and the Motor Theory of Speech Perception, Erlbaum, Hillsdale, NJ, 1991, pp. 117-137.

5. Walden, B. E., Prosek, R. A., Montgomery, A. A., Scherr, C. K., and Jones, C. J., “Effects of training on the visual recognition of consonants”, J. Sp. Hear. Res., Vol. 20, 1977, pp. 130-145.

6. Walden, B. E., Erdman, S. A., Montgomery, A. A., Schwartz, D. M., and Prosek, R. A., “Some effects of training on speech recognition by hearing-impaired adults”, J. Sp. Hear. Res., Vol. 24, 1981, pp. 207-216.

7. Gesi, A. T., Massaro, D. W., and Cohen, M. M., “Discovery and expository methods in teaching visual consonant and word identification”, J. Sp. Hear. Res., Vol. 35, 1992, pp. 1180-1188.

8. Massaro, D. W., Cohen, M. M., and Gesi, A. T. “Long-term training, transfer, and retention in learning to lipread”, Percept. Psychophys., Vol. 53, 1993, pp. 549-562.

9. Bernstein, L. E. and Eberhardt, S. P. Johns Hopkins Lipreading Corpus I-II: Disc 1, Corpus III-IV: Disc 2. Johns Hopkins University, Baltimore, MD, 1986.

10. Bernstein, L. E., Demorest, M. E., Coulter, D. C., and O'Connell, M. P., “Lipreading sentences with vibrotactile vocoders: Performance of normal-hearing and hearing-impaired subjects”, J. Acoust. Soc. Am., Vol. 90, 1991, pp. 2971-2984.

11. Boothroyd, A., and Hnath-Chisolm, T., “Spatial, tactile presentation of voice fundamental frequency as a supplement to lipreading: Results of extended training with a single subject”, J. Rehabil. Res. Dev., Vol. 25, 1988, pp. 51-56.

12. Boothroyd, A., Kishon-Rabin, L., and Waldstein, R., “Studies of tactile speechreading enhancement in deaf adults”, Semin. Hear, Vol. 16, 1995, pp. 328-342.

13. Eberhardt, S. P., Bernstein, L. E., Demorest, M. E., and Goldstein, M. H., “Speechreading sentences with single-channel vibrotactile presentations of voice fundamental frequency”, J. Acoust. Soc. Am., Vol. 88, 1990, pp. 1274-1285.

14. Kishon-Rabin, L., Boothroyd, A., and Hanin, L., “Speechreading enhancement: A comparison of spatial-tactile display of voice fundamental frequency (F0) with auditory F0”, J. Acoust. Soc. Am., Vol. 100, 1996, pp. 593-602.

15. Weisenberger, J. M., Broadstone, S. M., and Saunders, F. A., “Evaluation of two multichannel tactile aids for the hearing impaired”, J. Acoust. Soc. Am., Vol. 86, 1989, pp. 1764-1775.

16. Bernstein, L. E., Demorest, M. E., and Tucker, P. E., “Speech perception without hearing”, (submitted for publication, 1998).

17. Bernstein, L. E., Demorest, M. E., and Eberhardt, S. P., “A computational approach to analyzing sentential speech perception: Phoneme-to-phoneme stimulus-response alignment”, J. Acoust. Soc. Am., Vol. 95, 1994, pp. 3617-3622.