AVSP 2003 - International Conference
on Audio-Visual Speech Processing
St. Jorioz, France
September 4-7, 2003

Perception of Point Light Displays of Speech by Normal-Hearing Adults and Deaf Adults with Cochlear Implants

Tonya R. Bergeson[†], David B. Pisoni[†‡] and Jeffrey T. Reynolds[‡]

† Indiana University School of Medicine, Department of Otolaryngology, 699 West Street – RR044,
Indianapolis, Indiana, USA 46202

Bloomington, Indiana, USA 47405-1301
E-mail: tbergeso@iupui.edu, pisoni@indiana.edu

Abstract

Normal-hearing (NH) adults display audiovisual enhancement when degraded auditory input (e.g., words, sentences) is paired with point-light displays (PLDs) of speech, which isolate the kinematic properties of a speaker's face [10]. Do deaf adults who use cochlear implants (CIs) benefit in the same way? Does feedback influence NH adults' performance? In the present study, we investigated audiovisual (AV) word recognition using PLDs of speech in postlingually deaf adults with CIs and NH adults. Both groups displayed evidence of AV enhancement with PLDs. Moreover, NH participants' Visual-alone performance improved over time with Auditory-alone and AV feedback. These results suggest that NH and CI adults were sensitive to the kinematic properties in speech represented in the PLDs, and they were able to use kinematics to improve their word recognition performance even with highly degraded visual displays of speech. In addition, NH adults were able to use temporal cues from Auditory-alone and AV feedback to improve their word recognition performance with point-light visual displays of speech.

1. Introduction

Visual information about speech articulation obtained from lipreading has been shown to improve speech perception in adults with normal hearing [1, 2], hearing loss [3], and deaf adults with cochlear implants (CIs) [4, 5, 6]. In fact, many audiologists and speech and hearing scientists have assumed that the primary modality of speech perception is vision for hearing-impaired people [1, 7, 8].

In this connection, it has been proposed recently that hearing-impaired adults who are highly successful lipreaders exhibit larger audiovisual benefit than those who are poor lipreaders [4]. What are the cues that hearing-impaired adults attend to while lipreading? It is possible that good lipreaders are more sensitive to the articulatory changes over time, or kinematics, common to both auditory and visual speech patterns. One method of assessing sensitivity to time-varying visible speech information is to use point-light displays (PLDs). This method involves placing small point-lights on target locations of a darkened actor's face, videotaping the actor articulating a list of words, and then playing back the videotapes to individuals so that only the movement of the lights can be seen. When presented statically, PLDs cannot be recognized as a human face. However, once the point-lights begin to move and there is change over time, observers are able to recognize the displays as a human face articulating words. Thus, PLDs can be used to isolate the kinematic properties of the visual speech signal [9, 10].

2. Experiment 1

Several studies have shown that normal-hearing (NH) adults display AV enhancement when degraded auditory input (e.g., words, sentences) is paired with PLDs of speech [10, 11, 12]. Do deaf adults who use CIs benefit in the same way? In the present experiment, we investigated audiovisual (AV) word recognition using PLDs of speech in a small group of postlingually deaf adults with CIs.

2.1 Method

2.1.1 Participants

Five CI patients were deafened after the age of five, had used their CIs for at least one year, and were between the ages of 36 and 74 years (M = 56.2 years) (see Table 1). Six NH participants were also included as a control group, ranging from 22 to 24 years of age (M = 23.0 years).

2.1.2 Stimuli and Procedure

The PLDs of speech were constructed by darkening the face of the talker and by placing 10 dots symmetrically around the lips and mouth of the talker, 2 dots on the chin, 8 dots along the jaw line, 2 dots on the cheeks, 1 dot on the tip of the nose, 2 dots on the upper teeth, 2 dots on the lower teeth, and 1 dot on the talker's tongue (28 points total) (see Figure 1) [10]. The talker was videotaped while reading isolated English words under an infrared light so only the reflective disks surrounding the talker's articulators could be seen. The talkers in the full-face displays (FFDs) and PLDs were two different women. The FFD and PLD conditions each contained a different set of 96 monosyllabic English words, equally divided into words with high and low visual intelligibility.

	Age at Test (years)	Age at Implantation (years)	Duration of Implant Use (years)
CI 18	39	30	8
CI 50	74	73	1
CI 80	62	53	9
CI 94	36	35	1
CI 95	70	65	4

Table 1: Characteristics of CI participants.

Both groups of participants were instructed to repeat aloud what they thought the talker said under three presentation conditions: Auditory-alone (A-alone), in which the words were presented via a loudspeaker while the computer screen remained blank, Visual-alone (V-alone), in which visual displays of the talker articulating words were presented on the computer screen while the loudspeaker was off, and Auditory-visual (AV), in which the words were presented via the loudspeaker and the computer screen. A constant background noise (55 dB SPL) was present in the testing room. We presented the auditory stimuli at 75 dB SPL for CI users, but at 60 dB SPL for NH listeners so that gains due to visual input could be observed.

There were three phases of the present study: First, both groups of participants were initially given a practice session using a FFD. Following the practice session, participants were then given the word recognition tests first with the FFD, and finally with the PLD.

2.2 Results and Discussion

The data were scored by the number of whole words, phonemes, and visemes correctly identified. A viseme is a visual category of speech consisting of speech sounds that look similar when articulated [13]. For example, one viseme group might contain “va” and “fa” while another viseme group comprises “ma,” “ba,” and “pa”.

Figures 2 and 3 show the percent correct for word, phoneme, and viseme recognition in CI and NH participants across the two visual displays (full-face and point-light) and three presentation conditions (A-alone, V-alone, and AV). When CI participants’ responses were scored by words correctly identified, we found statistically significant main effects of visual display ($F(1, 4) = 24.79$, $p < .01$) and presentation format ($F(1, 4) = 69.60$, $p = .001$). The interaction between visual display and presentation format was not statistically significant. As shown in Figure 2, CI participants’ performance was better in the full-face condition than in the point-light condition. Performance was best in the AV presentation condition, followed by the A-alone presentation condition, and then the V-alone presentation condition.

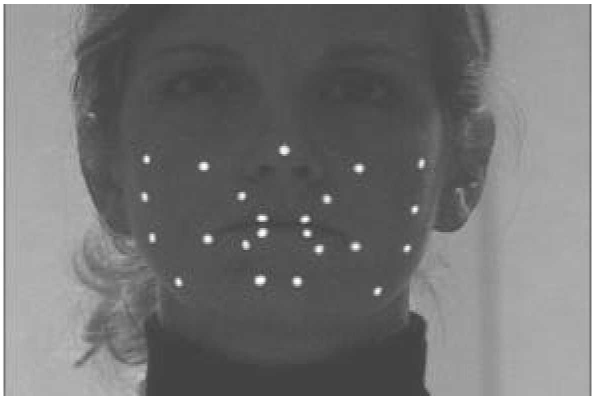

Figure 1: Dot configuration for point-light display. Five dots are not visible due to occlusion by the lips.

When NH participants’ responses were scored by words correctly identified, we found a statistically significant main effect of presentation mode ($F(1, 5) = 9.76$, $p < .05$), but no significant main effect of visual display, and no significant interaction between visual display and presentation format. As shown in Figure 3, performance for NH participants was best in the AV presentation condition, followed by the A-alone presentation condition, and then the V-alone presentation condition, similar to the pattern of performance for CI participants. Performance was not better in the full-face condition compared to the point-light display condition.

When CI participants’ responses were scored by the percentage of phonemes correctly identified, we found a statistically significant main effect of visual display ($F(1, 4) = 88.85$, $p = .001$) and presentation format ($F(1, 4) = 54.62$, $p < .01$). The interaction was not statistically significant. Again, as shown in Figure 2, CI participants’ performance was better in the full-face condition than in the point-light condition. Although performance was best in the AV presentation condition, there was no clear advantage for performance in the A-alone presentation condition compared to the V-alone presentation condition.

When NH participants’ responses were scored by the percentage of phonemes correctly identified, we also found a statistically significant main effect of visual display ($F(1, 5) = 48.34$, $p = .001$) and a marginally significant main effect of presentation format ($F(1, 5) = 5.18$, $p = .072$). The interaction was not statistically significant. As shown in Figure 3, NH participants’ performance was best in the AV presentation condition, followed by the A-alone presentation condition, and then the V-alone presentation condition. Performance was slightly better in the full-face condition than in the point-light display condition.

Finally, when CI participants’ responses were scored by percent of visemes correctly identified, we found statistically significant main effects of visual display

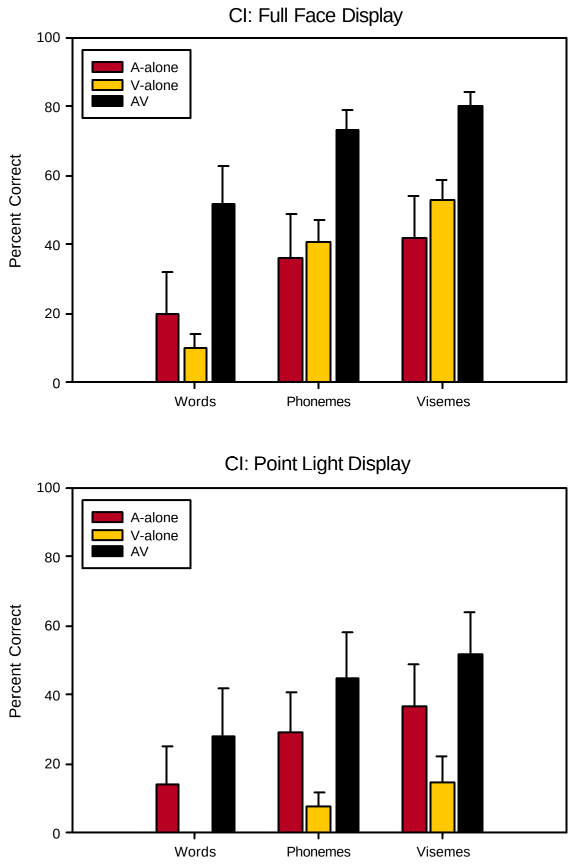

Figure 2. Word, phoneme, and viseme recognition performance for CI adults in A-alone, V-alone, and AV conditions for full-face versus point-light visual display.

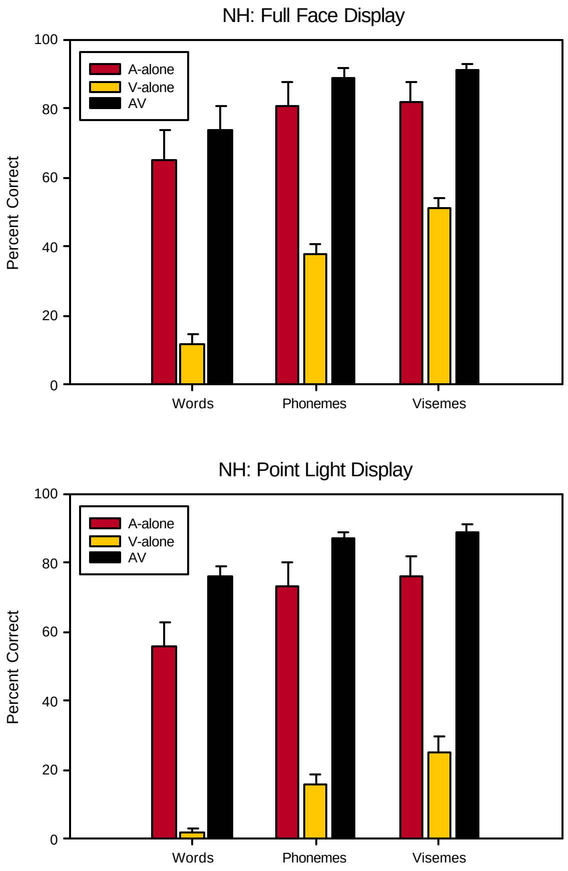

Figure 3. Word, phoneme, and viseme recognition performance for NH adults in A-alone, V-alone, and AV conditions for full-face versus point-light visual display.

($F(1, 4) = 50.41$, $p < .01$) and presentation format ($F(1, 4) = 37.77$, $p < .01$), as well as a statistically significant interaction between visual display and presentation format ($F(1, 4) = 7.79$, $p < .05$). Once again, CI participants' performance was better in the full-face condition than in the point-light condition. Performance was best in the AV presentation condition. In the point-light display condition, performance was better in the A-alone presentation condition than in the V-alone presentation condition. However, in the full-face display condition, performance was better in the V-alone presentation condition than in the A-alone presentation condition.

When NH participants' responses were scored by percent of visemes correctly identified, we found a statistically significant main effect of visual display ($F(1, 5) = 26.96$, $p < .01$) and a marginally significant main effect of presentation format ($F(1, 5) = 5.61$, $p = .064$). The interaction was not statistically significant. As shown in Figure 3, NH participants' performance was best in the AV presentation condition, followed by the A-alone presentation condition, and then the V-alone presentation condition. Performance was slightly better in the full-face condition compared to the point-light display condition.

Note that V-alone word, phoneme, and viseme recognition performance appears to be similar for NH and CI listeners across full-face and point-light displays. In fact, a two-tailed t-test revealed no statistically significant differences between NH and CI listeners in each condition. Vision has been assumed to be the primary modality of speech perception for hearing-impaired people [1, 7, 8]. The present results, however, show that the lipreading skills of this small group of postlingually deaf adult CI users are not better than the lipreading skills of NH adults.

3. Experiment 2

Although both CI and NH adults recognized some words in the V-alone PLD condition, their scores were quite low. Previous studies have shown small but significant improvements in full-face lipreading ability with training [14, 15]. In the second experiment, we investigated the effects of orthographic, A-alone, and AV feedback on NH adults' V-alone word recognition using PLDs of speech.

3.1 Method

3.1.1 Participants

Seventy-eight Indiana University undergraduate students between the ages of 18 and 24 years (M = 19.2 years) participated in this study for partial course credit in Introductory Psychology. All participants (49 women, 29 men) were native speakers of English, had no history of speech or hearing disorders, and had normal or corrected-to-normal vision at the time of testing.

3.1.2 Stimuli and Procedure

Only the V-alone set of PLDs of speech, consisting of 96 monosyllabic English words, was used in Experiment 2. Participants were instructed to type on a computer keyboard what they thought the talker said. In contrast to Experiment 1, participants in this experiment were told that each target word should be a one-syllable English word. In addition, participants were assigned to one of four feedback conditions: No Feedback (NFB), in which participants simply completed the experiment as they did in Experiment 1 (n = 16), Orthographic Feedback (OFB), in which participants saw the correct word displayed on the computer monitor for 500 ms following their response (n = 22), A-alone Feedback (AFB), in which participants heard the correct word following their response (n = 20), and AV Feedback (AVFB), in which participants were shown the same PLD along with the correct auditory word (n = 20).

3.2 Results and Discussion

The data were scored by the number of whole words, phonemes, and visemes correctly identified. Figure 4 shows the percent correct for word, phoneme, and viseme recognition across the four feedback conditions (NFB, OFB, AFB, AVFB) and two experiment conditions (first half, second half). When responses were scored by words correctly identified, we found no significant main effects of feedback or experiment half and no significant interaction. This is likely due to the fact that participants performed very near floor level using this scoring method.

However, when responses were scored by phonemes and visemes correctly identified, we found a significant main effect of experiment half (phonemes: *F*(1, 74) = 19.41, *p* < .0001; visemes: *F*(1, 74) = 13.77, *p* < .0001). The main effect of feedback and the interaction between feedback and experiment half were not significant. Thus, in contrast to the scores based on correct word recognition, participants' phoneme and viseme recognition performance improved across the experimental session.

Post-hoc t-tests were carried out to determine the effect of experiment half on phoneme and viseme recognition in each of the feedback conditions. Phoneme and viseme recognition performance was similar across the two experiment halves when participants received no feedback. Viseme recognition performance also did not significantly differ across the two halves of the experiment in the OFB condition, but phoneme recognition performance slightly

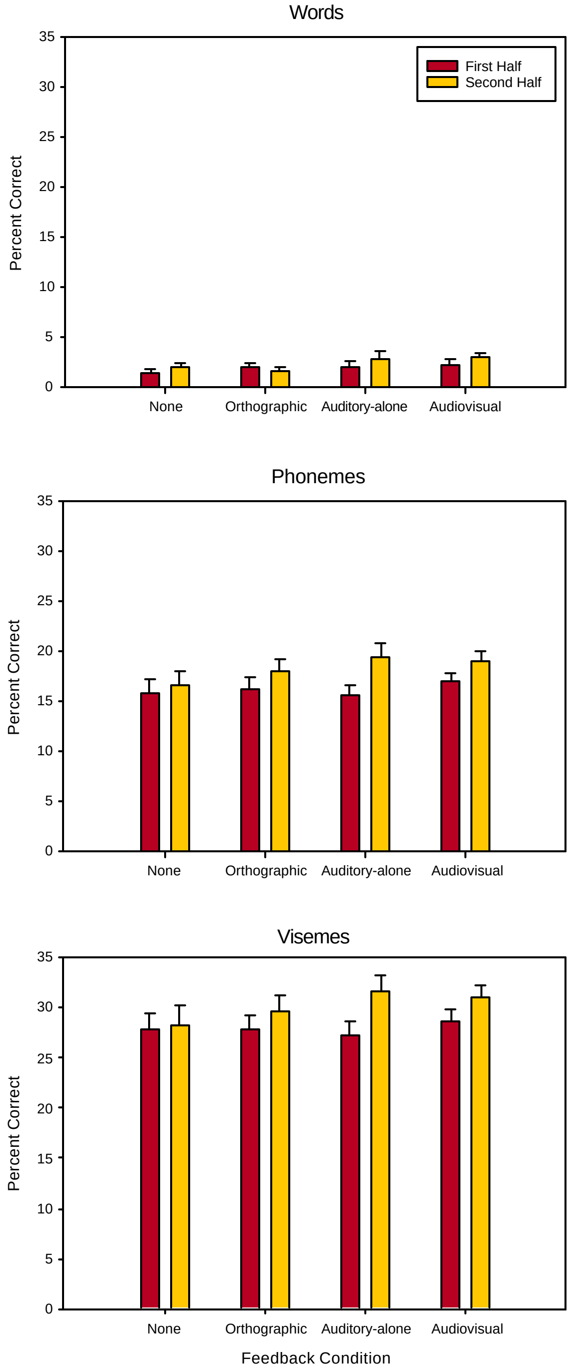

Figure 4. V-alone word, phoneme, and viseme recognition performance for NH adults in the four feedback conditions across the two halves of the experimental session.

increased in the second half of the experiment compared to the first half ($t(21) = 2.07$, $p = .051$). The largest improvements over the course of the experiment were found in the A-alone and Audiovisual feedback conditions. Both phoneme and viseme recognition performance increased in the second half of the experiment compared to the first half with A-alone Feedback (phonemes: $t(19) = 3.41$, $p < .01$; visemes: $t(19) = 3.29$, $p < .01$) and with AV Feedback (phonemes: $t(19) = 2.79$, $p < .05$; visemes: $t(19) = 3.03$, $p < .01$). Surprisingly, AV feedback did not improve phoneme and viseme recognition performance more than A-alone feedback.

Although overall performance did not differ whether participants received feedback (OFB, AFB, AVFB conditions) or no feedback (NFB condition), participants did receive a significant gain in phoneme and viseme recognition performance from A-alone and AV feedback, as well as a marginally significant gain in phoneme recognition performance from Orthographic feedback over the course of the experimental session. The present results are consistent with Gesi et al. [14], who used two different training methods over three days to teach full-face visual CVC syllable identification to NH adults. Participants were given auditory feedback after each trial in both training methods. Gesi et al. [14] found that visual syllable identification performance improved over time regardless of training method, presumably due to the auditory feedback.

Several researchers have suggested that both auditory and orthographic presentation of the target word would result in similar internal phonological/lexical representations of that word [e.g., 16, 17]. If this were the case, we would expect similar gains in the orthographic feedback condition as in the A-alone and AV feedback conditions. However, we found only a marginally significant gain in phoneme recognition performance, and no gain in viseme recognition performance in the orthographic feedback condition. The present results also contrast with those reported by Bernstein et al. [15], who found that lipreading performance improved with orthographic feedback training. However, the participants in the Bernstein et al. [15] study received lipreading training in six sessions across several weeks, compared to the participants in the present study who received feedback training only across 96 trials in one experimental session lasting about one hour. It is possible with increases in orthographic feedback training time NH adults could improve their lipreading abilities using PLDs of speech.

Nevertheless, the largest gains in performance over the course of the present experiment were found in the A-alone and AV feedback conditions. Note that the feedback cues in each of these conditions are temporal, i.e., they unfold over time. Although the previous studies of lipreading training used full-face visual displays, the PLDs in the present study isolated the kinematic properties of the visual speech signal. Thus, it is likely that the temporal cues in the A-alone and AV feedback conditions provided more benefit than the static cues in the Orthographic feedback condition to the participants in the present study.

4. CONCLUSIONS

The results of Experiment 1 showed that both CI and NH listeners displayed evidence of audiovisual enhancement with PLDs of speech. Overall, participants performed most accurately in the AV condition, followed by the A-alone condition, and then the V-alone condition. These results suggest that adult CI users, like NH adults, were sensitive to the kinematic properties in speech represented by the dynamic changes in the point-light displays, and they were able to use kinematics to improve their word recognition performance even with highly degraded visual displays of speech.

The results of Experiment 2 showed that NH adults were also able to use temporal cues encoded in A-alone and AV feedback to improve their word recognition performance with visual PLDs of speech. These results suggest that NH adults were sensitive both to the kinematic properties in speech represented in the PLDs as well as the associated dynamic changes in the auditory speech signal that represent and code the talker's underlying articulatory gestures.

5. ACKNOWLEDGEMENTS

This work was supported by NIH-NIDCD Training Grant T32DC00012 to Indiana University and NIH-NIDCD Research Grant R01DC00111 to the Indiana University School of Medicine.

REFERENCES

[1] Erber, N. P. "Interaction of audition and vision in the recognition of oral speech stimuli", *Journal of Speech and Hearing Research*, **12**, pp. 423-425, 1969.

[2] Sumby, W. H., and Pollack, I. "Visual contribution to speech intelligibility in noise", *Journal of the Acoustical Society of America*, **26**, pp. 212-215, 1954.

[3] Erber, N. P. "Auditory-visual perception of speech", *Journal of Speech and Hearing Disorders*, **40**, pp. 481-492, 1975.

[4] Grant, K. W., Walden, B. E., and Seitz, P. F. "Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration", *Journal of the Acoustical Society of America*, **103**, pp. 2677-2690, 1998.

[5] Kaiser, A. R., Kirk, K. I., Lachs, L., and Pisoni, D. B. "Talker and lexical effects on audiovisual word recognition by adults with cochlear implants", *Journal of Speech, Language, and Hearing Research*, in press.

[6] Tyler, R. F., Parkinson, A. J., Woodworth, G. G., Lowder, M. W., and Gantz, B. J. “Performance over time of adult patients using the Ineraid or Nucleus cochlear implant”, *Journal of the Acoustical Society of America*, **102**, pp. 508-522, 1997.

[7] Gagné, J.-P. “Visual and audiovisual speech perception -P. Gagné & N. Tye-Murray (Eds.), *Research in Audiological Rehabilitation: Current Trends and Future Directions (Monograph). Journal of the Academy of Rehabilitative Audiology*, **27**, pp. 133-159, 1994.

[8] Seewald, R. C., Ross, M., Giolas, T. G., and Yonovitz, A. “Primary modality for speech perception in children with normal and impaired hearing”, *Journal of Speech and Hearing Research*, **28**, pp. 36-46, 1985.

[9] Bingham, G. P., Rosenblum, L. D., and Schmidt, R. C. “Dynamics and the orientation of kinematic forms in visual event recognition”, *Journal of Experimental Psychology: Human Perception and Performance*, **21**, pp. 1473-1493, 1995.

[10] Rosenblum, L. D., and Saldaña, H. M. “An audiovisual test of kinematic primitives for visual speech perception”, *Journal of Experimental Psychology: Human Perception and Performance*, **22**, pp. 318-331, 1996.

[11] Lachs, L., Pisoni, D. B., and Kirk, K. I. “Use of audiovisual information in speech perception by prelingually deaf children with cochlear implants: A first report”, *Ear & Hearing*, **22**, pp. 236-251, 2001.

[12] Rosenblum, L. D., Johnson, J. A., and Saldaña, H. M. -light facial displays enhance comprehension of speech in noise”, *Journal of speech and Hearing Research*, **39**, pp. 1159-1170, 1996.

[13] 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”, *Journal of Speech and Hearing Research*, **20**, pp. 130-145, 1977.

[14] Gesi, A. T., Massaro, D. W., and Cohen, M. M. “Discovery and expository methods teaching visual consonant and word identification”, *Journal of Speech and Hearing Research*, **35**, pp. 1180-1188, 1992.

[15] Bernstein, L. E., Auer, E. T. Jr., and Tucker, P. E. “Enhanced speechreading in deaf adults: Can short-term training/practice close the gap for hearing adults?”, *Journal of Speech, Language, and Hearing Research*, **44**, pp. 5-18, 2001.

[16] Seidenberg, M. S., and McClelland, J. L. “A distributed, developmental model of word recognition and naming”, *Psychological Review*, **96**, pp. 523-568, 1989.

[17] Slowiaczek, L. M., Soltano, E. G., Wieting, S. J., and Bishop, K. L. “An investigation of phonology and orthography in spoken-word recognition”, *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, **56A**, pp. 233-262, 2003.