Visual Vowel and Diphthong Perception across Speakers

Sharon A. Lesner and Patricia B. Kricos
The University of Akron

The purpose of this study was to investigate whether different speakers, all within a normal range of articulation and intelligibility, present the same visual information during a lipreading task. Four speakers who were found to vary in the ease with which they could be lipread were videotaped while they presented 15 vowels and diphthongs in an /h/-V-/g/ context. Twelve normal-hearing subjects phonetically recorded the vowel or diphthong they perceived. Viseme categories were determined through the use of a hierarchical clustering analysis. Results indicated that the number and nature of the viseme categories varied across speakers and were related to the ease with which the speakers could be lipread. A comparison of the perception of the speakers' production of sentences from the Utley Test of Lipreading Ability and the speakers' production of the experimental vowels and diphthongs was also made. A significant correlation ($r = 0.96$) between diphthong perception and Utley score was found. Visual perception of diphthongs was also found to be significantly better than visual perception of vowels ($F = 22.37, p < .05$).

Various classification systems have been proposed for describing the visually distinctive signals which are associated with articulatory movements. In order to lipread, these visually distinctive movements or visemes (Fisher, 1968) must be decoded. Nitchie (1950) and Bruhn (1949) have noted that vowels, unlike consonants, have unique articulatory movements associated with their production so theoretically they can all be visually differentiated. Yet, investigators have been unsuccessful in categorizing vowels into viseme categories, and it has been consistently found that vowels are not identified with complete correctness (Berger, 1970; Woodward & Lowell, 1964; Wozniak & Jackson, 1979).

Of the classification systems that have been proposed, Nitchie (1950) advanced a system which is similar to the traditional vowel triangle used by phoneticians. That is, vowels are categorized according to the width of the

Sharon A. Lesner, Ph.D., and Patricia B. Kricos, Ph.D., are Assistant Professors of Audiology, Department of Speech Pathology and Audiology, The University of Akron, Akron, Ohio.

mouth opening (narrow, medium, wide) and lip shape (puckered, relaxed, extended). The implication is that all vowel sounds can be visually differentiated since they each have unique articulatory positions associated with their production. Nitchie warned, however, that some vowels can only be distinguished through context.

In experimental investigations, Jeffers and Barley (1971) described vowels in terms of "lipreading movements," and they implied that there are seven viseme groupings under ideal viewing conditions and four viseme groupings of vowels under usual viewing conditions. Fisher (1968) concluded that there are only four visually contrastive vowel visemes. Based on results using multidimensional scaling techniques, Jackson, Montgomery, and Binnie (1976) proposed five dimensions which underlie vowel lipreading performance. Several authors have suggested that some of the variations in viseme categories that have been obtained by various investigators could be accounted for, at least in part, by speaker differences (Berger, 1972; Jeffers & Barley, 1971; Kricos & Lesner, 1980; Wozniak & Jackson, 1979).

The purpose of this study was to investigate whether different speakers, all within a normal range of articulation and intelligibility, present the same information during a lipreading task. Secondary purposes were to determine if a speaker's production of vowels is related to perception of the speaker's production of sentences in a lipreading task and to determine if diphthongs are easier to discriminate visually than vowels.

METHOD

Speakers

Four speakers, who had been presenters in a previous study of visual consonant recognition (Kricos & Lesner, 1980) and who were found to vary in the ease with which they could be lipread, were videotaped while they presented the test stimuli. The four speakers were female graduate students in the Department of Speech Pathology and Audiology at The University of Akron. They all used a general American dialect and were identified as having normal articulation and intelligibility through administration of the Weiss Comprehensive Articulation Test (1978).

Stimulus Items

Stimuli for the experimental task consisted of the following 15 English vowels and diphthongs: /i/, /ɪ/, /e/, /æ/, /a/, /ɔ/, /ʊ/, /u/, /ʌ/, /ɝ/, /eɪ/, /ou/, /aɪ/, /au/, and /ɔɪ/. Each of these phonemes was produced in an /h/-V-/g/ context to form fifteen consonant-vowel-consonant monosyllabic nonsense syllables. These syllables were randomized into four lists, one for each speaker, with each list consisting of four presentations of each stimulus item for a total of 60 syllables.

Test stimuli also included sentences from the Utley Test of Lipreading Ability (Utley, 1946). Since the Utley Test consists of two forms which have been equated for difficulty as a visual task, the sentences on each form were randomized into two lists, providing a different list for each speaker.

The speakers were recorded on black and white videotape (JVC 6060U videocassette recorder and Sony AVC 3260 camera) while they presented the experimental stimuli. The recorded image was a front view that included each speaker's head and shoulders. Two reflector hoods with 150 watt incandescent bulbs were used at a 45-degree angle approximately four feet from the speaker's mouth to supplement the recording room's normal illumination. The speakers were encouraged to use a natural speaking manner without exaggeration during the videotaping. A neutral background was used to avoid the presence of distracting stimuli.

Subjects

Subjects for the experimental task were 12 normal-hearing female college students with no prior training in phonetics and no prior experience with lipreading tasks. They ranged in age from 18 to 23 years and all had normal or corrected-to-normal vision (defined as 20/25 or better distance acuity as measured using the Titmus Vision Tester).

Presentation of Stimulus Items

The presentation of the stimulus items consisted of only the video portion of the recordings. An 18-inch monitor (Shibaden 19 UL) was utilized in a quiet, well-illuminated room. During the viewing sessions, subjects were tested in pairs, with each subject seated five feet from the monitor at an approximate zero degree angle.

Subjects were asked to view each speaker as they presented the Utley Test (Utley, 1946) items, followed by the CVC syllables. A sufficient interval of time was allowed so that the subjects could record what they had observed following each presentation of a stimulus item. A key that phonetically listed the fifteen test vowels and diphthongs was provided for each subject. The order of speaker presentation was counterbalanced.

RESULTS

Individual and group 15 × 15 confusion matrices were constructed for each speaker from the response sheets. From these, the percentage of correct vowel and diphthong identifications was determined for each speaker as well as an overall vowel/diphthong recognition score. A mean score on the Utley Test of Lipreading Ability (Utley, 1946) was also calculated for each speaker. This was accomplished by giving a point for each sentence for which the subject had preserved the main idea. These results are shown in Table 1.

By averaging the rank orders of the sets of scores for each speaker, Speaker

Table 1
Absolute Percent Correct Recognition Scores by Speaker

	Stimuli[a]			
Speakers	Utley	Diphthongs	Vowels	Vowel/Diphthong
1	35.5 (2)	70.1 (2)	34.6 (3)	44.5 (3)
2	34.2 (3)	62.9 (3)	46.8 (1)	50.6 (1)
3	39.3 (1)	75.5 (1)	39.5 (2)	49.1 (2)
4	11.6 (4)	48.5 (4)	31.3 (4)	36.2 (4)

[a] Numbers in parentheses indicate the speaker's rank.

3 was found to be the easiest to lipread (mean Utley Score = 39.3%; mean diphthong score = 75.5%; mean vowel score = 39.5%), while Speaker 4 was the most difficult to lipread (mean Utley score = 11.6%; mean diphthong score = 48.5%; mean vowel score = 31.3%).

The percentage of correct recognitions of the test stimuli were calculated for each speaker, and these results are shown in Table 2. Inspection of the table reveals that the test stimuli differed in the degree to which they could be correctly identified. For all speakers, though, the /ɔɪ/, /aʊ/, and /σu/ were identified correctly more than 50% of the total number of times they were presented, while the /ʊ/, /ɛ/, /ɔ/, /ʌ/, /ɝ/, and /æ/ were identified correctly less than 50% of the total number of times they were presented. There were differences in the percent correct rankings of the phonemes for each speaker. Thus, although /ɔɪ/ was correctly identified more than 90% of the total number of times when it was presented by Speaker 1, for example, it was identified correctly only between 60 to 70% of the total number of times it was presented by Speaker 2. As a group, diphthong perception was significantly better than vowel perception across the speakers ($F = 22.37$; df 1, 3; $p < 0.05$).

A Pearson Product Moment Correlation analysis was conducted to determine the relationship between a speaker's diphthong recognition score and the Utley scores. A positive .96 coefficient of correlation was obtained, indicating a marked relationship between the two sets of scores ($t = 4.8$; $p < .05$). A statistically significant relationship was not found between a speaker's vowel recognition score and the Utley score ($r = 0.61$).

An analysis of the specific confusions that occurred in terms of vowel and diphthong perception for each speaker was done. Johnson's (1967) hierarchical clustering analysis was performed in order to determine each subject's viseme categories. This technique was used by Walden, Prosek, Montgomery, Scherr, and Jones (1977) to determine pre- and post-training visemes. The clustering analysis enables the investigator to determine the arrangement of stimuli into clusters based on the similarity of the stimuli. Clusters were

Table 2
Percentages of Correct Identification for Each Vowel and Diphthong Across Speakers

Percent Correct Scores	Speaker 1	Speaker 2	Speaker 3	Speaker 4
91-100	ɔɪ	i	aʊ	
81-90		aʊ, o	ɔɪ	
71-80	aɪ		aɪ, o	
61-70	aʊ	ɑ, ɔɪ	ɑ	u, ɔɪ
51-60	e, ɑ, o	e, a	u	ɪ, au, o
41-50	ɪ, æ, u	ɪ, u, ɝ	ɪ, ʌ, æ	i, æ
31-40	i	æ, ɔ	i, e	
21-30	ɝ, ʌ	ʌ	ɔ	e, aɪ, ɑ, ɔ, ʊ, ɝ
11-20	ɛ, ʊ, ɔ	ɛ	ɛ, ɝ	ɛ
0-10		ʊ	ʊ	ʌ

Table 3
Vowel Viseme Categories for Each Speaker

Speaker 1	Speaker 2	Speaker 3	Speaker 4
aɪ	aʊ	aɪ	i, ɪ, ʌ
ɜɪ	o	o	
i, ɪ	i	i, ɪ	
e, ɛ, æ	e, ɛ, æ	aʊ	
	ɑ, ɔ	ɔɪ	

accepted as visemes when the within-cluster responses constituted 75% or more of the responses. Using this criterion, the visemes shown in Table 3 were obtained for each of the four speakers. A comparison of the viseme categories indicates that they not only differ in number across speakers but also that the nature of the viseme categories varies across speakers.

DISCUSSION

The results of this study suggest several important implications for the assessment and training of lipreading skills. It is apparent that not all speakers, even with normal articulation and a general American dialect, will reveal the same visemes to individuals who are lipreading them. Since a viseme is "any individual and contrastive visually perceived unit" (Fisher, 1968, p. 800), it is reasonable to expect that speakers who present more visemes will be easier to lipread than speakers who present fewer.

A comparison of Speaker 4 (determined to be relatively difficult to lipread) and Speaker 3 (determined to be relatively easy to lipread) suggests that both the number and the nature of a particular speaker's viseme categories are related to the ease with which the speaker can be lipread. Speaker 3 had five viseme categories while Speaker 4 had only one viseme category. In addition, of the visemes presented by Speaker 3, four out of the five were individual phonemes. For the remaining viseme, this speaker's production of /ɪ/ was confused with the speaker's production of /i/ almost as often as it was correctly identified. Speaker 4's single viseme category consisted of /i/, /ɪ/, and /ʌ/. Not only did this speaker reveal fewer visually contrastive units, but the viseme that the speaker did reveal contained three homophenous sounds as opposed to a single phoneme. That is, the /i/, /ɪ/, and /ʌ/ were mutually confused. Therefore, persons lipreading Speaker 4 would be able to identify if these vowels were presented versus the other 12 vowels, yet they would not be able to identify which one of the three vowels within the viseme group was being presented.

It would appear to be critical that speakers who administer lipreading tests, who conduct lipreading training, or who educate hearing-impaired children attempt to determine the particular nature of the visual information that they are conveying. They should not merely assume that they are providing as much information as is suggested by proponents of traditional lipreading methods or by results of experimental investigations.

Finally, the degree of visibility of the phonemes varied across speakers, with diphthongs being relatively more visible than vowels. Wozniak and Jackson (1979) have reported that the movement involved in the production of diphthongs appears to be perceivable and that this movement aids a lipreader in discriminating them. The high correlation between diphthong perception and sentence perception and the lack of a significant correlation between vowel perception and sentence perception may be related to the latter and the fact that normal-hearing subjects who had not had lipreading training were employed in this study. Heider and Heider (1940) indicated that differences between vowels are very minimal; therefore, they suggest that it is possible with training to learn to make progressively finer differentiations. Perhaps, if a group of sophisticated lipreaders were tested, a higher correlation between vowel recognition scores and sentence scores might be the result.

ACKNOWLEDGMENT

The authors would like to thank L. Christine Shields for her assistance in the collection and analysis of the data.

REFERENCES

Berger, K.W. Vowel confusions in speechreading. *Ohio Journal of Speech and Hearing,* 1970, *5,* 123-128.

Berger, K.W. *Speechreading: Principles and methods.* Baltimore, Md.: National Educational Press, 1972.

Bruhn, M.E. *Mueller-Walle Method of Lipreading.* Washington, D.C.: The Volta Bureau, 1949.

Fisher, C.G. Confusions within six types of phonemes in an oral-visual system of communication. Unpublished doctoral dissertation, Ohio State University, 1963.

Fisher, C.G. Confusions among visually perceived consonants. *Journal of Speech and Hearing Research,* 1968, *11,* 796-804.

Heider, F., & Heider, G.M. An experimental investigation of lipreading. *Psychological Monographs,* 1940, *52,* 124-133.

Jackson, P.L., Montgomery, A.A., & Binnie, C.A. Perceptual dimensions underlying vowel lipreading performance. *Journal of Speech and Hearing Research,* 1976, *19,* 796-812.

Jeffers, J., & Barley, M. *Speechreading (Lipreading).* Springfield, Ill.: Charles C. Thomas, 1971.

Johnson, S.C. Hierarchical clustering schemes. *Psychometrika,* 1967, *32,* 241-254.

Kricos, P.B., & Lesner, S.A. Differences in visual distinctive features across speakers. Paper presented at the Annual Convention of The American-Speech-Language-Hearing Association, Detroit, Michigan, 1980.

Nitchie, E.H. *New lessons in lipreading.* Philadelphia, Pa.: J.B. Lippincott, 1950.

Utley, J. A test of lipreading ability. *Journal of Speech and Hearing Disorders,* 1946, *11,* 109-116.

Walden, B.E., Prosek, R.A., Montgomery, A.A., Scherr, C.K., & Jones, C.J. Effects of training on the visual recognition of consonants. *Journal of Speech and Hearing Research,* 1977, *20,* 130-145.

Weiss, C.E. Weiss Comprehensive Articulation Test. Boston, Ma.: Teaching Resources Corporation, 1978.

Woodward, M.F., & Lowell, E.E. A linguistic approach to the education of aurally-handicapped children. United States Department of Health, Education, and Welfare. Project #907, 1964.

Wozniak, V.D., & Jackson, P.L. Visual vowel and diphthong perception from two horizontal viewing angles. *Journal of Speech and Hearing Research,* 1979, *22,* 354-365.