Viseme Weighting Systems and Lipreading: A Comparative Reappraisal

William K. Ickes, Ph.D.
Texas Tech University
Lubbock, Texas

ABSTRACT

This series of experiments is concerned with the application of viseme weighting systems to research in lipreading. Such systems are reported in the literature but not widely employed. Four such systems are compared for validity (Experiment 1) and reliability (Experiment 2) with the AHS system appearing to be valid and the most reliable. In Experiment 3, the AHS system was used to determine if lipreading difficulty is more related to sentence length or to degree of visibility. Forty test items were constructed consisting of ten short, visually easy sentences; ten short, visually difficult sentences; ten long, visually easy sentences; and ten long, visually difficult sentences. The sentences were presented without auditory cues to senventeen native subjects. An analysis of variance suggests that sentence length, at least up to ten words, is not a significant factor in lipreading difficulty; whereas visibility is a significant factor. An alternative interpretation of previous research which indicates that sentence length is a factor in lipreading difficulty is presented along with suggestions for future research.

The art of lipreading[1] among persons with acquired hearing loss is something of an enigma which has puzzled scholars in the hearing sciences for years. The processes involved have

[1] The author is aware that in scholarly circles the term speechreading is preferred since visual communication involves more than just lip movements. However, the term lipreading is used here for two reasons. First, the research reported here is concerned primarily with lip movements only, indeed lipreading. Second, the primary purpose of words is to communicate. Generally speaking, the term lipreading is better understood outside of the hearing sciences and has come to signify the visual communication process in its totality. People seem to know what is meant by the term lipreading but are confused by the term speechreading.

remained so obscure that little meaningful research is currently being conducted to add additional knowledge to the meager amount of information that we already have. What we do know if fraught with conjecture and conflicting information which appears to make the problem even more complex. As anyone who has taught lipreading can attest, there appears to be little doubt that superior lipreaders exist. But how did they acquire this skill and what personal characteristics are necessary to become a good lipreader? Conversely, poor lipreaders may remain poor lipreaders even after months of training. What personal characteristics are lacking in a poor lipreader which retards his advancement? We simply do not have any clear-cut answers to these questions. However, it is not the purpose of this paper to explore all the ramifications of the problems associated with lipreading. There are other sources which do this quite adequately (Berger, 1972).

Specifically, the present research is concerned only with the visual signal employed in communication through lipreading. It is an attempt to investigate further the usefulness of viseme weighting as a functional tool in lipreading research and lipreading training. The term viseme was suggested by Fisher (1963) to differentiate the visual characteristics of the speech signal from the purely auditory aspects of the speech signal as characterized by the phoneme. The basis of the present study resides in a viseme weighting system originally devised by the Board of Education of the City of New York and published by the American Hearing Society (1943). In essence this weighting system places a numerical value of either .25, .50, .75, or 1.00 on visemes according to their assumed difficulty in being seen. For any given sentence, one sums the viseme values contained therein and divides the total by the number of visemes to obtain an average difficulty score, which is then multiplied by a constant of 100. Unfortunately, the available literature does not make clear how individual visemes were initially assigned to their point values, but it appears that this was done on a **priori** basis. In any case, the author(s) of the AHS scale provide research data which show correlations between computed visibility scores and actual visibility scores for untrained lipreaders ranging from .23 to .70 depending on whether or not a clue word was included. This data appears in Table 1.

Doubt has been cast on the efficacy of the original AHS viseme weighting systems by O'Neill (1954), who reported no statistical significance attributable to visibility for either vowels or consonants when the system was included in a discrimination

TABLE 1. Four Viseme Weighting Systems

VISEMES	AHS SYSTEM	O'NEILL SYSTEM	WOODWARD & LOWELL	BERGER SYSTEM
k & g	.25	.50	.25	.50
h	.25	.50	.75	.50
ŋ	.25	.50	.25	.50
p & b	1.00	1.00	1.00	1.00
t & d (I)	.50	.50	.75	.75
t & d (F)	.50	.50	.75	.50
f & v	1.00	1.00	1.00	1.00
θ & ð	1.00	1.00	.75	.50
s & z	.50	.50	.75	.50
ʃ & ʒ	1.00	1.00	1.00	1.00
tʃ & dʒ	1.00	1.00	1.00	1.00
m	1.00	1.00	1.00	1.00
n (I)	.50	.50	.75	.75
n (F)	.50	.50	.75	.50
r	.50	.50	1.00	.75
w	1.00	1.00	1.00	1.00
j	1.00	1.00	.50	.25
l	.75	1.00	.75	.50
i	.75	.50	.75	.75
I	.50	.25	.25	.50
e	.75	.75	.50	.50
ɛ	.50	.25	.25	.25
æ	1.00	1.00	1.00	.75
a	1.00	1.00	.75	.75
ɔ	1.00	1.00	.75	.75
o	1.00	.75	.75	1.00
U	.50	.50	.75	.50
u	1.00	.75	.50	.50
ʌ	.50	.50	.50	.50
ə	.50	.50	.50	.50
ɝ	.50	.50	.75	.75

task at -20dB speech-to-noise ratio. However, O'Neill's data were based on his own **a priori** modification which clearly may have affected his results. Table 1 shows the O'Neill viseme weighting scale as nearly as I have been able to reconstruct it from the information he provides. O'Neill mentions specific consonants and vowels which he shifted among categories, but for vowels and consonants which he did not mention, the relative weighting shown in Table 1 remain as indicated in the AHS scale. Since the information in Table 1 may not be an accurate representation of what O'Neill actually did, perhaps it is a misnomer to call it the O'Neill scale. Nevertheless, for the sake of a label for a second viseme weighting system based on **a priori** categorization, the data in Table 1 are referred to as the O'Neill scale.

I have been able to construct two additional viseme weighting scales based on empirical data. The first is attributed to Woodward and Lowell (1964) and the second to Berger (1970, 1972b). Table 1 also presents these two additional weighting scales, and the data contained therein were taken from data presented by Berger (1972a, pp. 82-95). In tabular form Berger has presented his own data which are compared to Woodward and Lowell's data and which show percent of correct responses and percent of confusions for vowels and consonants. The percent scores were collected from subjects involved in an experimental lipreading task. The rationale used to construct these two additional scales is simple enough. Visemes which show a high percent correct score and a low percent confusion score are assumed to be easier to lipread than visemes which show a low percent correct score and a high percent confusion score. Therefore, visemes with a percent correct score of .76% or better were placed in Category I and given a weighting of 1.00. Visemes with a percent correct score between 51% and 75% were placed in Category II and given a weighting of 0.75. Visemes with a percent correct score between 26% and 50% were placed in Category III and given a weighting of 0.50; and visemes with a percent correct score between 1 to 25% were placed in Category IV and given a weighting of 0.25.

A major question, and the essence of Experiment 1, pertains to the validity of weighting systems. If the procedure of weighting visemes according to degree of visibility is valid, then there ought to be a substantially significant correlation between actual subject lipreading scores and the degree of visibility assigned to the various visemes. Therefore, which of the four weighting systems described above offers the highest correlation with actual lipreading scores.

In an early attempt to provide some validation data for this type, Taaffe and Wong (1957) employed the AHS system in an experiment designed to study variables in lipreading stimulus material. Visibility scores were computed for 60 sentences contained in **The Film Test of Lip Reading** (1957). A coefficient of correlation between visibility scores and **P** scores was obtained and no statistical significance was found (**P** scores represent sentence difficulty based on the number of words correctly identified by the experimental subjects). Recognizing that visibility scores are based on "letter determinations" [visemes], Taaffe and Wong attempted to better equate visibility scores with **P** scores by dividing the total number of visibility units found in a sentence by the total number of words contained in a sentence. The coefficient of correlation between this average and **P** scores also was not statistically significant. The problem with the scoring method employed by Taaffe and Wong is that of trying to compare qualitatively different units of analysis (e.g. apples and oranges). Visibility scores are based on visual units whereas the basic component of words is the phoneme which is auditory. Further, dividing the total number of visemes in a sentence by the number of words in a sentence is not likely to yield an interpretable metric since the number of visemes in a word may vary.

In order to improve the scoring process, in the present series of experiments only the visual units, or visemes, were employed. The sentences constructed for this research were assigned visibility scores based on the four weighting systems described previously. Subject scores were determined by counting the number of visemes per sentence correctly identified by each subject. Group totals per sentence were divided by the number of subjects to obtain a mean group score per sentence. Since sentences could vary in length, and hence vary in the number of visemes contained therein, the visual difficulty of the sentence was assessed by dividing the mean group score by the number of visemes within the sentence and this average was employed in computing the coefficient of correlations to be reported. The computational equation is as follows:

$$Ss = \frac{Xr}{NsNv} \text{ X } 100$$

Where: S_s is equal to subject score per sentence,

X_r is equal to the group sum of correct responses,

N_s is equal to the number of subjects,

N_v is equal to the number of visemes in the sentence, and

100 is a constant

It should be noted that this scoring procedure is identical to that employed by Taaffe and Wong except visemes, not words, were mployed in the computation. Multiplying subject scores by a constant of 100 removed fractions and is consistent with the process used to obtain visibility scores from the viseme weighting systems.

Experiment 2 is concerned with reliability of the visibility score measures and seeks to determine if the results obtained in Experiment 1 are replicable with a second group of subjects. Further, which of the four weighting systems can be expected to give consistent results when employed with other stimulus material used to measure lipreading ability?

Experiment 3 makes use of the information obtained in Experiments 1 and 2 applying a weighting system to a fundamental problem in lipreading research. It has been claimed that lipreading difficulty increases as sentence stimuli increase in number of words or syllables (Morris, 1944; Taaffe and Wong, 1957). However, can this claim still be made if the stimuli are equated in terms of visibility difficulty? In other words, is sentence length really a factor in lipreading difficulty or is sentence length merely an artifact of increased visual difficulty? These last questions are the ones of greatest theoretical importance which the present series of experiments attempt to clarify.

EXPERIMENT I

Method

Subjects. The subjects are 14 undergraduate students with normal hearing who were all untrained, naive lipreaders. The age range was from 18 to 26 years.

Stimulus Material. Twenty-five sentences were constructed and weighted in visual difficulty according to each of the four weighting systems described earlier. The sentences were analyzed only in terms of visemes, not phonemes. Therefore, consonant blends were given the weighting value of the greater of the two components. Similarly, vowel dipthongs were given the weighting value of the greater of the two components. The order in which the sentences were presented to subjects as randomly determined.

Procedure. The sentences were presented by an adult male speaker, without voice[2], in a well-lit room. There was full light on the speaker's face. The subjects were positioned in such a way so

[2] Arguments for and against the use of live vs. filmed or taped tests of lipreading are presented by Berger (1972, Chapter VIII). The major criticism of a live presentation appears to be possible fluctuation of visibility on repeated presentations. Since in this experiment the sentences were presented only once to the subjects, the question of reliability need not be of major concern.

that no subject was outside a 45° sight angle. No subject was closer than 1.5 meters to the speaker nor further away than 5.0 meters. Sentences were read only once after the subjects were instructed as follows:

> This is a test of your lipreading ability. It is being conducted for research purposes. There are 25 sentences. Each sentence will be read only once; and I will not use voice. Record what you see. You may record whole sentences, whole words, syllables, or even isolated phonemes [visemes]. If you record only syllables or phonemes, place them on your answer sheet in the approximate position in which you think they occurred. Are there any questions?

The scoring of subject responses was identical to the procedure described earlier. Since the idea was to see how closely subject scores correlated with what was actually transmitted by the speaker, homophenous representations were not included in the scoring.

Results. Subject scores per sentence were correlated with visibility scores for each of the four weighting systems employed in this research withe the following results:

1. AHS weighting systems	r = .67
2. O'Neill weighting system	r = .69
3. Woodward and Lowell weighting system	r = .63
4. Berger weighting system	r = .31

With 23 degrees of freedom, an r of .505 was needed for statistical significance at the .01 level of confidence. The high correlations, with the Berger weighting system excepted, indicates validity for employing viseme weighting systems to lipreading research when subject responses are scored as indicated in this study.

EXPERIMENT 2

The next consideration was establishing the reliability of the visibility scores based on the various weighting systems. In other words, are the results obtained in Experiment 1 replicable with other subject groups or with different stimulus material? Experiment 2 was designed to answer this question.

Method

Subjects. The subjects were 17 undergraduate students with normal hearing who were all untrained, naive lipreaders. The age range was from 19 to 25 years. None of the subjects who participated in Experiment 1 were allowed to participate in Experiment 2.

Stimulus Material. The stimulus material consisted of 26 sentences[3] which were weighted in difficulty according to each of the four weighting systems employed in this research. Thirteen (that is, one half) of the sentences were common to both Experiments 1 and 2. The remaining 13 sentences were newly constructed. As with Experiment 1, the attempt was to analyze difficulty according to the number of visemes contained in each sentence. The order in which the sentences were presented was randomly determined.

Procedure. The procedure and instructions presented to the subjects prior to testing were identical to those of Experiment 1. Only the number of sentences presented was different. The scoring of subject responses was also identical to the method employed in Experiment 1.

Results. Subject scores per sentence were correlated with visibility scores for each of the four weighting systems employed in the research with the following results:

1. AHS weighting system	r = .60
2. O'Neill weighting system	r = .52
3. Woodward and Lowell	r = .34
4. Berger weighting system	r = .51

With 24 degrees of freedom, an r of .496 was needed for statistical significance at the .01 level of confidence. With the exception of the Woodward and Lowell system, all of the other systems were significant at the .01 level of confidence. However, since a primary goal of Experiment 2 was to test the comparitave reliability of the four systems, it is important to note that the greatest consistency was obtained with the AHS system. The correlational results between subject groups used for Experiment 1 and Experiment 2 were almost identical for the AHS system, and suggest that the AHS system is not only valid but perhaps the most reliable of the four systems.

As a test of the reliability of the scoring procedure, a coefficient of correlation was obtained between subject scores for the 13 sentences common to both subject groups. With 11 degrees of freedom, an r of .684 was needed to demonstrate statistical significance at the .01 level of confidence. In this instance, r was found to be .88. This very high correlation indicates high consistency in the scoring procedure and further suggests consistent subject responses from one subject sample to the next.

[3] There were actually 40 sentences presented to this group of subjects. However, only the first 26 were employed in this reliability study. The data obtained from all 40 sentences is to be reported in Experiment 3.

EXPERIMENT 3

Since the AHS viseme weighting system appears to be valid and a reliable tool in lipreading research, attention was directed toward using this tool in gaining information about sentence length as it relates to lipreading difficulty. Previous research (Morris, 1944; Taaffe and Wong, 1957) has suggested that lipreading difficulty increases as sentence length increases. Experiment 3 was designed to determine whether lipreading difficulty is indeed related to sentence length or whether the degree of visibility of the stimulus material is the factor which best accounts for lipreading difficulty?

Method

Subjects. The subjects were 17 undergraduate students who were naive and untrained in lipreading skills. They were the same subjects as those employed in Experiment 2.

Stimulus Material. Forty sentences were constructed so as to provide 10 sentences which were short with low visibility, 10 sentences which were short with high visibility, 10 sentences which were long with low visibility, and 10 sentences which were long with high visibility. Sentence length was determined by number of words, with 5 words or less constituting a short sentence and more than 5 words constituting a long sentence. No sentence was less than 3 words in length nor longer than 10 words in length. The mean number of visemes for short sentences was 20.0, or almost twice as many. Using the AHS viseme weighting system, a dichotomy was formed for sentence visual difficulty. Difficult sentences ranged in visual difficulty from a visibility score of 50 to a visibility score fo 59. Easy sentences ranged in visual difficulty from a visibility score of 70 to 92. Sentence difficulty was adjusted by means of a t test until the long and short, difficult sentences failed to show a significant difference in visibility scores. A statistically significant difference was expected for visibility scores between the short, visually easy sentences and the short, visually difficult sentences; and between the long, visually easy sentences and the long, visually difficult sentences; and t tests confirmed this expectation.

The sentences were presented to the subjects only once with identical instructions as provided in Experiments 1 and 2. The order of sentence presentation was randomly determined and the procedure and scoring technique were those employed in Experiments 1 and 2.

Results. The subject scores were subjected to an analysis of variance. A highly significant (P◄.001) F ratio was found for visibility difficulty, indicating that, with sentence length held constant, subjects were much less accurate in reading visibly difficult sentences than in reading visibly easy ones. Neither sentence length nor the interaction of sentence length with visibility difficulty yielded statistically significant effects. These data were summarized in Table 2. The mean scores for long sentences and short sentences were 22.1 and 28.2, respectively; but since the analysis of variance failed to show statistical significance for this factor, this mean difference must be attributed to chance.

TABLE 2. Analysis of variance for visual difficulty and sentence length.

Source	SS	df	m/s	F
SS_A (visibility)	8,352.10	1	8,352.10	26.64*
SS_B (sentence length)	372.10	1	372.10	1.19
Interaction A X B	184.90	1	184.90	.59
Residual	11,286.00	36	313.50	
Total	20,195.10	39		

* Significant at .01 level or beyond

DISCUSSION

It would seem that too little attention has been paid to the purely visual aspects of stimulus material used in lipreading research. Without taking into account the visual difficulty of stimulus material, it is possible to draw erroneous conclusions concerning the relationship between sentence length and lipreading accuracy. Thus, Taaffe and Wong (1957) concluded that as sentence length increases, so does lipreading difficulty. Yet they present discrepant data which suggests eight word sentences are easier to lipread than are five, six, or seven word sentences. In fact their eight word sentences are about as easy to lipread as are four word sentences. They admit that this discrepancy is difficult to explain, but they suggest that eight word sentences may be easier because eight word sentences contain more "contextual cues." From their data they further suggest that nine, ten, and eleven word sentences are the most difficult and the explanation is offered that the advantage provided by contextual cues for sentences longer than eight words are offset by the greater number of words. Taaffe and Wong's data is displayed in Figure 1.

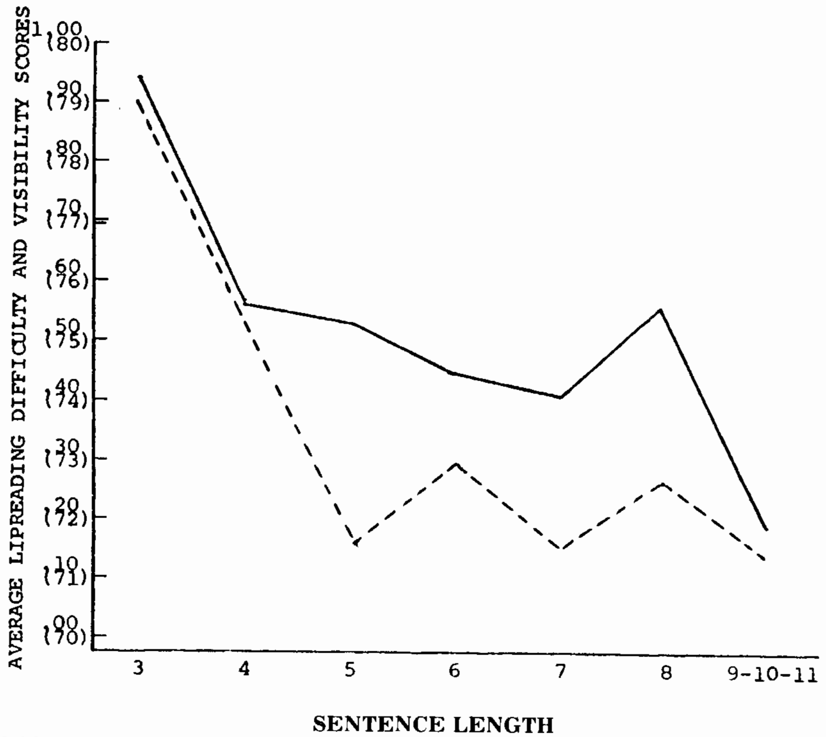

FIGURE 1. Average sentence lipreading difficulty (Taaffe and Wong) and visibility scores (Ickes) by length of sentence. Visibility scores are shown in parenthesis along the ordinate and are represented in the graph by the broken line.

I used the AHS viseme weighting scale to assign visibility difficulty scores to the sentences contained in the **Film Test of Lip Reading** which Taaffe and Wong (1957) used as stimulus material for their research, and then obtained mean visibility scores for the three, four, five, six, seven, and eight word sentences. The nine, ten, and eleven word sentences were pooled just as Taafe and Wong had done, and a mean for these sentences was also obtained. When graphically plotted (Figure 1) these mean visibility scores appear very similar to the sentence difficulty curve presented by Taaffe and Wong. Eight word sentences are easier than seven word sentences and nine, ten, and eleven word sentences (pooled) are the most difficult of all. This suggests to me that relative visibility, not sentence length, is responsible for the date presented by Taaffe and Wong. The agreement is not perfect but the difference in scoring (words correct as opposed to visemes correct) may account for the graphic differences.

In summary, the use of viseme weighting in lipreading research appears to be a valid and reliable procedure. Subject lipreading scores, at least for untrained subjects as employed in this research, appears to be systematically related to the degree of visual difficulty the stimulus material presents. I would suggest that a number of previous research findings in the lipreading literature ought also to be re-examined. For instance, are declarative sentences really more difficult to lipread than are interrogative sentences as has previously been claimed (Taaffe and Wong, 1957)? Do lipreaders really have more difficulty with negatively constructed sentences than they do with passively constructed sentences (Schwartz and Black, 1967)? Is the length of words or the number of syllables (Berger, 1972, p. 104) really related to lipreading difficulty? Or, alternatively, are all of the above factors more related to the degree of visual difficulty contained in the stimulus material? The research possibilities are numerous.

REFERENCES

Berger, K.W., Vowel confusions in speech reading. **Ohio J. Speech and Hearing** 5, 123-128 (1970).

Berger, K.W., **Speechreading Principles and Methods.** Baltimore: National Education Press (1972a).

Berger, K.W., Consonant confusions in speechreading. **Ohio J. Speech and Hearing** (1972b). (In Press)

Fisher, C.G., Confusions among visually perceived consonants. **J. Speech Hearing Res.**, 11, 796-804 (1963).

Morris, D.M., A study of some of the factors involved in lipreading. M.A. Thesis, Smith College (1944).

New Aids and Materials for Teaching Lip-Reading. **Washington: American Society For the Hard of Hearing.** 22, 27 (1943).

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

Schwartz, J.R., and Black, J.W., Some effects of sentence structure on speechreading, **Central States Speech J.**, 18, 86-90 (1967).

Taaffe, G., **A Film Test of Lip Reading.** Studies in Visual Communication II, Los Angeles: John Tracy Clinic (1957).

Taaffe, G., and Wong, W., Studies of variables in lip reading stimulus materials, **John Tracy Clinic Research Papers,** III, pp. 21 (1957).

Woodward, M.E., and Lowell, E.E. A linguistic approach to the education of aurally-handicapped children. **U.S. Dept. of Education, Education and Welfare.** Project No. 907 (1964).