Visual Contribution to Speech Intelligibility in Noise*

W. H. SUMBY† AND IRWIN POLLACK
Human Factors Operations Research Laboratories, Washington 25, D. C.
(Received November 5, 1953)

Oral speech intelligibility tests were conducted with, and without, supplementary visual observation of the speaker's facial and lip movements. The difference between these two conditions was examined as a function of the speech-to-noise ratio and of the size of the vocabulary under test. The visual contribution to oral speech intelligibility (relative to its possible contribution) is, to a first approximation, independent of the speech-to-noise ratio under test. However, since there is a much greater opportunity for the visual contribution at low speech-to-noise ratios, its absolute contribution can be exploited most profitably under these conditions.

INTRODUCTION

IN many practical work situations, the standard criteria for speech interference levels, based upon laboratory articulation test procedures, may be misleading. This condition is the result of several factors, two of which are the subject matter of the present study: the information associated with the class of possible messages and the contribution of visual factors to speech intelligibility.

First, if only a small number of possible messages may be communicated, we can tolerate higher noise interference levels than if the class of possible messages is large.[1] And, second, if visual factors supplementary to oral speech are utilized, we can tolerate higher noise interference levels than if visual factors are not utilized.[2]

This study considers the interaction of these two factors. Specifically, we shall examine the contribution of visual factors to oral speech intelligibility as a function of the speech-to-noise ratio and the size of the possible vocabulary.

APPARATUS AND PROCEDURE

1. Experimental Variables

The experimental variables manipulated were: the absence or presence of supplementary visual observation of a speaker's lips and facial movements, the speech-to-noise ratio under test, and the size of the vocabulary under examination.

2. Speech Materials

The speech materials employed were 256 bisyllabic words of the spondaic stress pattern, e.g., cupcake, baseball. These words were chosen because they were less subject to inter-speaker variation than other classes of words examined.

Vocabularies of 8, 16, 32, 64, and 128 words were randomly selected from the entire group of 256 spondees. Test lists of 25 and 50 items were then constructed from each restricted vocabulary-source. A different ordered series was assembled for each group of subjects and for each testing-session.

In a series of supplementary tests, words of three different lengths were considered—monosyllables, spondees, and trisyllablic phrases. This series was designed to test the generality of the findings to other speech materials. The trisyllabic phrases were constructed by combining a spondee and a monosyllable into a meaningful pair with equal speech stress on each syllable, e.g., "hardware store."

3. Speech Signal

Trained speakers read the lists of spondaic words into a suspended microphone (RCA 88-A). A high quality auditory system (±1 db between 25 and 20 000 cps) was employed between the microphone and earphones (Permoflux PDR-8 mounted in doughnut cushions). The over-all speech level was measured in terms of the average peak deflection of a Daven VU meter. The signal level was monitored at a constant level by a test supervisor.

4. Noise

Noise, derived from a gas-tube source, was mixed electrically with the speech signal. It was uniform in level per cycle in the frequency band of 20–10 000 cps. The level at the listener's ears was db S.P.L., based upon an overall reading of a Daven VU meter. A S/N ratio of 0 db was defined in terms of an equal overall reading of each of the two signals upon the VU meter. The speech-to-noise ratio was varied by holding the noise level constant and varying the speech level.

5. Test Procedure

Before each test list was presented, the speaker recited the test vocabulary in order to define the words under test. A reference list, alphabetically arranged, of the test vocabulary was furnished to the subject. The speed of reading was determined by the subjects' response rate. If a word was not clearly received, the subjects were instructed to select a word from the restricted vocabulary on the basis of any marginally available cues. The order of presentation of the various tests conditions was varied at random. No carrier sentence was used. In its place, a warning light was turned on approximately one second before each word was read. Immediately after each test list, each subject corrected his own test responses.

6. Subjects

Six subjects were seated about a table in a group. Their average distance from the speaker was five feet. Each subject wore a tight fitting headset. Each subject handheld the cushion nearest the speaker in order to insure negligible direct air transmission over the noise background.

 The writers wish to thank Mr. John Schjelderup for his assistance with the experimental equipment and A/3C Paul Baringer for his assistance with tabulation of the experimental data. This report is a condensation of an HFORL report written by the first author with the guidance of the second author.

† Now at the University of Virginia, Charlottesville, Virginia.

[1] Miller, Heise, and Lichten, J. Exptl. Psychol. 41, 329 (1951) especially Fig. 2, p. 333.

[2] J. J. O'Neill, unpublished doctoral dissertation, The Ohio State University, 1951.

Half of the subjects watched the speaker's facial movements as he spoke (auditory and visual presentation); the other half faced away from the speaker (auditory presentation alone). Each subject alternated between the two listening conditions. A total of 129 subjects—enlisted military and civilian laboratory personnel and undergraduate university students—participated. No special practice in lip-reading was given and all had normal auditory and visual acuity. In the supplementary test series, nine university undergraduate students were employed.

RESULTS

Speech intelligibility scores, under conditions of auditory presentation alone, are presented in Fig. 1 as a function of the speech-to-noise ratio. The parameter is the size of the vocabulary under test. In general, speech intelligibility decreases as the speech-to-noise ratio is decreased and as the size of the vocabulary is increased. However, little further change in speech

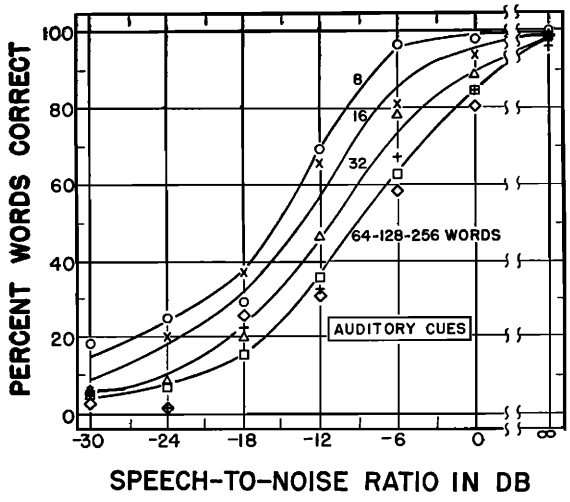

FIG. 1. Speech intelligibility under conditions of auditory presentation alone as a function of the speech-to-noise ratio under test. The parameter on the curves is the size of the vocabulary (spondee words) under examination. Each point in Figs. 1 and 2 represents the average results for 450 determinations pooled over subjects.

intelligibility is observed as the size of the vocabulary under test is increased beyond 64 words.[3]

A parallel examination of the results associated with combined auditory presentation and visual observation of the speaker is presented in Fig. 2. The major relationships of Fig. 1 are again obtained. The outstanding difference, however, is the higher resistance to noise for bisensory presentation. This finding is illustrated by the gentler slopes of the empirical functions of Fig. 2.

For each experimental condition, the *difference*

[3] This latter finding is contrary to that obtained by Miller, Heise and Lichten. They found continued decrements in performance as the size of the vocabulary was extended from 32 to 1000 words. The discrepancy is due, we suspect, to the fact that our subjects' check lists were arranged in an arbitrary (alphabetic) fashion, whereas, Miller's were arranged in logical groupings of common vowel sounds. Thus, for larger vocabularies, our lists were probably of little value to the subject. See: Miller, Heise, and Lichten, Reference 1.

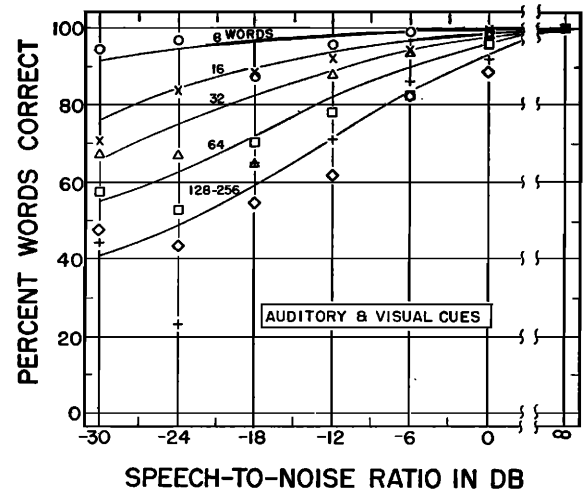

FIG. 2. Speech intelligibility under conditions of simultaneous auditory presentation and visual observation of a speaker's facial movements as a function of the speech-to-noise ratio under test. The parameter on the curves is the size of the vocabulary (spondee words) under examination.

between the average intelligibility associated with auditory presentation alone and that associated with bisensory presentation is presented in Fig. 3. The major relationship presented in Fig. 3 is that this difference-score increases as the speech-to-noise ratio is decreased. Specifically, when the speech signal is inaudible and where only visual factors operate (S/N ratio of -30 db), the differences between the intelligibility scores associated with the two experimental conditions range from 40 percent for the 256-word vocabulary to 80 percent for the 8-word vocabulary. In contrast, under noise-free conditions, there is little difference in the intelligibility scores associated with the two test conditions.

Each difference-score may be regarded, alternatively, as the contribution of visual observation of the speaker's

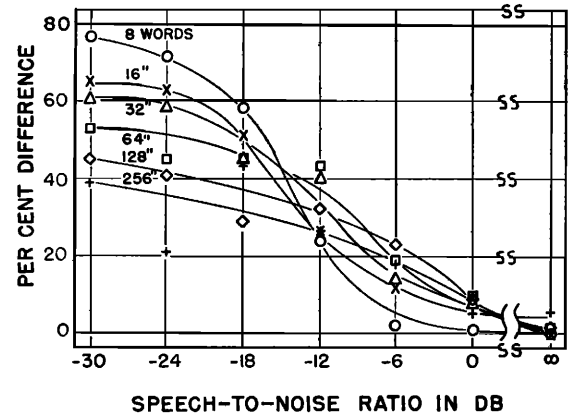

FIG. 3. The difference between the speech intelligibility scores under conditions of bisensory presentation (Fig. 2) and auditory presentation alone (Fig. 1) as a function of the speech-to-noise ratio under test. The parameter on the curves is the size of the vocabulary (spondee words) under test.

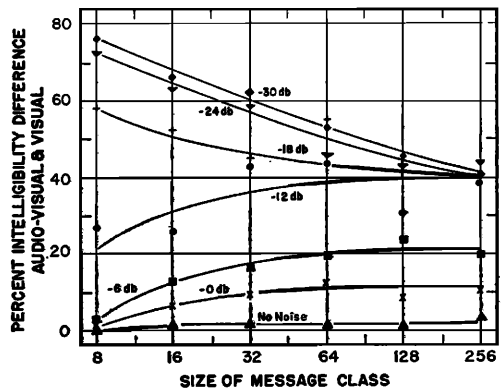

FIG. 4. The difference between the speech intelligibility scores under conditions of bisensory presentation (Fig. 2) and auditory presentation alone (Fig. 1) as a function of the size of the vocabulary. The parameter on the curves is the speech-to-noise ratio under test.

facial and lip movements to oral speech intelligibility in noise. In these terms, the main conclusion to be drawn from Fig. 3 is that this visual contribution becomes more important as the speech-to-noise ratio is decreased.

An alternative description of this visual contribution to oral speech intelligibility is presented in Fig. 4. The parameter and abscissa of Fig. 3 have been interchanged in Fig. 4 in order to examine the role of vocabulary-size. In general, under conditions of low speech-to-noise ratios, the visual contribution (in terms of the difference-score) is decreased with an increase in vocabulary size. Under conditions of high speech-to-noise ratios, however, the visual contribution is increased with an increase in vocabulary-size. This latter finding is due, in part, to the upper ceiling placed upon the intelligibility scores which leaves no room for improvement for highly restricted vocabularies under high speech-to-noise ratios.

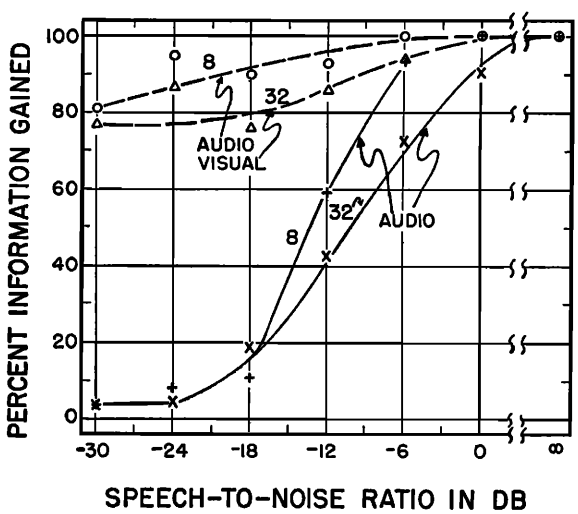

FIG. 5. The information transmitted, relative to the information presented, as a function of the speech-to-noise ratio under test. The parameter on the curves is the size of the vocabulary (spondee words) under test. Upper curves are for auditory and visual presentation. Lower curves are for auditory presentation alone. Each point in Figs. 5 and 6 represents the results for 50 observations, pooled over subjects, for each of the words.

An alternative examination of the results, in terms of measures of transmitted information, is presented in Fig. 5. The information transmitted, relative to the information presented, is plotted as a function of the speech-to-noise ratio. The results associated with bisensory presentation are presented as the upper curves and with auditory presentation alone as the lower curves. The parameter is the size of the vocabulary under test.

In general, the major finding of the intelligibility analysis is verified: the visual contribution to speech intelligibility (in terms of the difference in transmitted

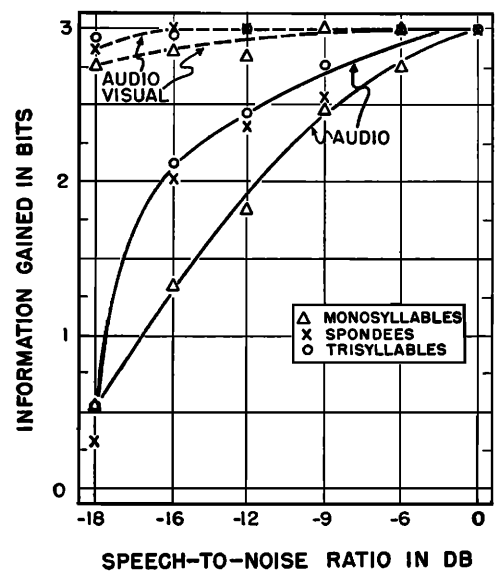

FIG. 6. The information transmitted, relative to the information presented, as function of the speech-to-noise ratio under test. The parameter is the type of words employed. Data for vocabularies of eight words. The upper curves are associated with bisensory presentation; the lower curves are associated with auditory presentation alone.

information associated with bisensory presentation and with auditory presentation alone) increases as the speech-to-noise ratio is decreased.

The results of the supplementary tests are presented in Fig. 6 in terms of the relative measure of transmitted information. There is little difference in favor of polysyllabic words under conditions of bisensory presentation. Under conditions of auditory presentation alone, higher intelligibility scores are associated with the polysyllabic words with little difference between the scores with the bi- and trisyllabic words.

DISCUSSION

The information analysis, presented in Fig. 5, may be extended by an analysis of the individual components

of transmitted information to achieve a somewhat more general result. In the following analysis, we are essentially duplicating a similar analysis presented by McGill[4] in a somewhat different situation.

It may be noted that the absolute visual contribution (as defined by the difference-scores between auditory presentation alone and bisensory presentation) must, necessarily, be small at high speech-to-noise ratios. The reason is that, under these conditions, there is little room for improvement with bisensory presentation because intelligibility is high under conditions of auditory presentation alone. Conversely, there is a much greater opportunity for a visual contribution at low speech-to-noise ratios because, under these conditions, intelligibility is low under the condition of auditory presentation alone.

The more meaningful question, perhaps, is "What is the visual informational contribution *relative* to the *possible available* contribution in the absence of visual cues?" The actual contribution, as shown schematically in Fig. 7 as A, is the difference between the scores associated with auditory-visual presentation and auditory presentation alone. The possible available contribution, as shown schematically in Fig. 7 as B, is the difference between the total response information[5] and the transmission under auditory presentation alone. The ratio of these two scores, A/B answers the question.

This ratio is approximately constant over a wide range of speech-to-noise ratios. Specifically, for the 8-word vocabulary, the ratio increases from about 0.81 at a S/N ratio of -30 db to about 0.95 at a S/N ratio of -6 db. For the 32-word vocabulary, the ratio increases from 0.77 to 0.81 over the same range. Thus, to a first approximation, the relative visual informational contribution supplied by observing a speaker's facial and lip movements is *independent* of the speech-to-noise ratio under test.

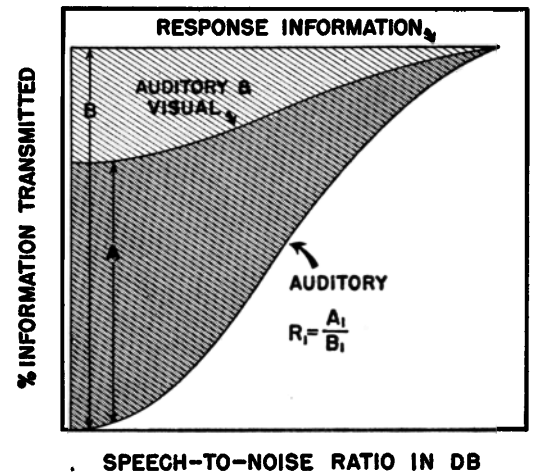

FIG. 7. Schematic illustration of the visual contribution to oral speech intelligibility relative to its maximum possible contribution.

Practically speaking, however, since there is a much greater opportunity for the visual contribution at low speech-to-noise ratios, its absolute contribution can be exploited most profitably under these conditions. And, since these conditions are the rule, rather than the exception, in many military and industrial situations, the results suggest that oral speech intelligibility may be appreciably improved in many practical situations by arrangement for supplementary visual observation of the speaker.

[4] W. J. McGill, Paper delivered before the American Psychological Association, September, 1953.

[5] The total response information in our tests is almost indistinguishable from the stimulus information. This is a value of 100 percent or 1.0 in Fig. 5.