(1951) J. Exp Psy #41 PP 329-335

THE INTELLIGIBILITY OF SPEECH AS A FUNCTION OF THE CONTEXT OF THE TEST MATERIALS [1]

BY GEORGE A. MILLER, GEORGE A. HEISE, AND WILLIAM LICHTEN

Harvard University

For many years communication engineers have used a psychophysical method called the "articulation test" (2, 3). An announcer reads lists of syllables, words, or sentences to a group of listeners who report what they hear. The articulation score is the percentage of discrete test units reported correctly by the listeners. This method gives a quantitative evaluation of the performance of a speech communication system.

There are three classes of variables involved in an articulation test: the *personnel*, talkers and listeners; the *test materials*, syllables, words, sentences, or continuous discourse; and the communication *equipment*, rooms, microphones, amplifiers, radios, earphones, etc. The present paper is directed toward the second of these three classes of variables, the test materials. The central concern can be stated as follows: Why is a stimulus configuration, a word, heard correctly in one context and incorrectly in another?

Three kinds of contexts are explored: (*a*) context supplied by the knowledge that the test item is one of a small vocabulary of items, (*b*) context supplied by the items that precede or follow a given item in a word or sentence, and (*c*) context supplied by the knowledge that the item is a repetition of the immediately preceding item. All three kinds of context enable the listener to limit the range of alternatives from which he selects his response. A word selected from a small vocabulary must be one of the few words agreed upon in advance. A word in a sentence must be one of the relatively few words that make a reasonable continuation of the sentence according to grammatical rules agreed upon in advance. A repeated word must be one of the few words similar to the word just heard. Not anything can happen, and the listener can set himself to make the required discrimination. Context, in the sense the word assumes here, is the *S's* knowledge of the conditions of stimulation. The experimental problem is to vary the nature and amount of this contextual knowledge in order to study its influence upon perceptual accuracy.

Equipment and Procedure

The apparatus consisted of components from military communication equipment used during the recent war. The output voltage of a carbon microphone was amplified and delivered to the listener's dynamic earphones. The talker monitored his speaking level with a volume indicator (VU meter) that responded to the voltage generated at the output of the amplifier. A random noise voltage, with a spectrum that was relatively uniform from 100 to 7,000 cps, was introduced at the listener's earphones. The signal-to-noise ratio (S/N) was varied by holding the average voice level constant and changing the level of the noise. The S/N was measured by a vacuum tube voltmeter across the terminals of the earphones, and the measurements reported in

[1] This research was carried out under Contract N5ori-76 between Harvard University and the Office of Naval Research, U.S. Navy (Project NR147-201, Report PNR-74). Reproduction for any purpose of the U.S. Government is permitted.

the following pages represent the ratio in decibels of the average peak deflection of the meter for the words (in the absence of noise) to the level of the noise in the 7,000-cycle band. A S/N of zero db means, therefore, that the electrical measurements indicated the two voltages, speech and noise, were equal in magnitude. Since the earphones transduce frequencies only up to about 4,500 cps, however, the acoustic level of the noise was about 2 db lower than these electrical measurements indicate. The over-all acoustic level of the voice at the listener's ears was approximately 90 db re .0002 dyne/cm^2.

The speech channel was not a high quality system. Only the speech frequencies between 200 and 3,000 cps were passed along to the listener.[2]

Only two *Ss* were used throughout the experiments. Both had normal auditory acuity, and both were familiar with the design and theory of the experiments. The *Ss* were located in different rooms, connected only by the communication channel described above, and they alternated as talker and listener. Some particular S/N was set up in the channel, and the talker proceeded to read a list of test items. These items were pronounced after a carrier sentence, "You will write. . . ." During this carrier sentence the talker adjusted his voice level to give the proper deflection of the monitoring VU meter, and then the test item was delivered with the same degree of effort. This procedure preserves the inherent variability of English words—the word "peep" has much less acoustic energy than the word "raw" when both words are pronounced with equal emphasis by a normal talker. By monitoring the carrier sentence rather than the test item, the relative intensities of the speech sounds are preserved in a natural fashion. The listener then recorded the item on a test blank, and these test sheets were later graded and the scores converted to percentages.

Importance of Test Materials

The kind of speech materials used to test communication systems is an important variable determining the results of the tests. Figure 1 illustrates how much difference the test materials can make. These three functions were obtained for the communication channel and the personnel described above. The test materials used for these three functions were the following.

(*a*) The *digits* were pronounced *zero, wun, too, thuh-ree, four, f-i-i-v, six, seven, ate, niner.* (*b*) The *sentences* were those constructed at the Psycho-Acoustic Laboratory (1). A sentence consists of five major words connected by auxiliary "of's," "the's," etc. The score shown in Fig. 2 represents the percentage of these major words heard correctly. (*c*) The *nonsense syllables* used were also those published by Egan (1). To standardize the pronounciation and recording of the nonsense syllables, an abbreviated phonetic symbolism was used.

The values of S/N necessary for 50 per cent correct responses are approximately -14 db for digits, -4 db for the individual words in a sentence, and $+3$ db for nonsense syllables. At a S/N where practically no nonsense syllables were recorded correctly, nearly all the digits were correctly communicated. Differences of this magnitude require explanation. What differences among these spoken stimuli make some easy to hear and others quite difficult? A list of perceptual aspects—rhythm, accent, grouping, meaning, or phonetic composition—can be suggested.

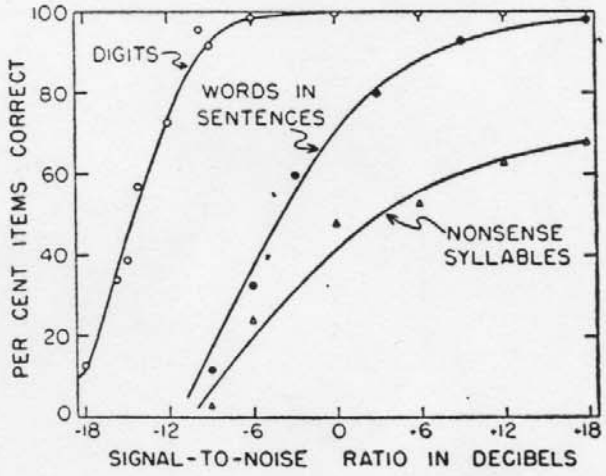

Fig. 1. Relative intelligibility of different test materials

[2] For the convenience of those who may wish to apply one of the several schemes for predicting articulation scores, the frequency response of the system may be obtained by ordering Document 3250 from American Documentation Institute, 1719 N Street, N. W., Washington 6, D.C., remitting $1.00 for microfilm (images 1 in. high on standard 35 mm. motion picture film) or $1.00 for photocopies (6 × 8 in.) readable without optical aid.

Our experiments indicate, however, that these various characteristics of the stimulus that *did* occur are less important than the characteristics of the stimuli that *could* have occurred but didn't. The most important variable producing the differences is the range of possible alternatives from which a test item is selected. A listener's expectation (or, more precisely, his freedom of choice) is determined by the context in which the particular phonetic pattern occurs. When digits are used, the listener can respond correctly with a marginal impression of the relatively intense vowel sound alone, because all the digits, with the exception of *five* and *nine*, have different vowels. Since the alternatives are thus limited, the digit is interpreted correctly, although the same acoustic stimulus is quite ambiguous when the alternatives are not so limited. With nonsense syllables, however, this limitation of possibilities is far less helpful; the listener feels that anything can happen. To record the nonsense syllable correctly, a listener must perceive each phoneme correctly, and the perception of one phoneme in a syllable does not give a clue to the other phonemes in the same syllable. Not only must the listener hear the vowels correctly, but the less intense consonant sounds must also be distinguished.

Size of Test Vocabulary

An articulation test is a rather unusual combination of the familiar psychophysical procedures. The experiment requires the listener to select, not one out of two or three, but one out of several thousand alternative responses. Thus the number of alternatives involved becomes an interesting variable.

Suppose we try to adapt spoken stimuli as closely as possible to the traditional method of constant stimuli. To this end we might use a single speech sound or a single syllable as the stimulus, present this speech unit at various intensities, and ask *S* to report whether or not he heard each presentation. This procedure determines a threshold of audibility for the particular speech unit. The practical value of this isolated datum is negligible. The experiment must be repeated for all the forty or fifty different speech sounds or the thousands of different syllables of English. And then we know only about audibility, not intelligibility.

Consider this distinction between audible speech and intelligible speech. It is intuitively clear that the words *audible* and *intelligible* are not synonyms, and listeners give reliably higher thresholds when asked to make continuous discourse "just understandable" instead of "just audible" (2). The crux of the difference is that intelligibility involves a complex discrimination and identification, whereas audibility is simply a discrimination of presence or absence.

It seems reasonable, therefore, to call a speech unit intelligible when *it is possible for an average listener with normal hearing to distinguish it from a set of alternative units*. By a speech unit is meant any combination of vocal noises—phonemes, syllables, words, phrases, sentences. The act of distinguishing can take various forms—repeating the unit, writing it down, pointing to it, behaving in accordance with its content, etc. The critical part of this definition concerns the set of alternative speech units from which the particular unit is selected. This part of the definition reduces intelligibility to discriminability, and avoids the questions of semantic rules and meaning. Discriminability is a function of the number of alternatives and

the similarities among them. The word "loot" is easily discriminated if all the alternatives are trisyllabic, but difficult to distinguish, other things being equal, in a set of alternatives that includes "boot," "loop," "jute," "lewd," "mute," "loose," etc.

An articulation test is analogous to a test of visual acuity where the percentage of correct judgments of a fixed set of test figures is plotted as a function of the level of illumination. A differential judgment is required under various favorable and unfavorable conditions. In such an experiment we determine the most unfavorable conditions under which the discrimination can be made, rather than the most unfavorable conditions under which the presence of the stimuli can be detected. These are clearly different thresholds and correspond to what we have called the thresholds of intelligibility and of audibility.

A difficult discrimination quickly becomes impossible as the conditions are made unfavorable, whereas an easy and obvious difference remains noticeable almost as long as the stimuli can be detected. The discrimination of a difference of 3 cycles in frequency, for example, is fairly accurate under favorable conditions—at 1,000 cps and 100 db. If the intensity is progressively reduced, however, such a small difference becomes imperceptible. For a simpler discrimination, say 30 cycles difference in frequency, the listener can respond accurately at all intensities down to 5 or 10 db above the threshold of audibility.

The situation is manageable so long as we have some index of the difficulty of the discrimination. Thus, in the tonal example, the difficulty can be gauged by the size of the difference in frequency. With the articulation test, however, such an index is not available. We could utilize known differences in the spectra of the sounds to construct an index of the distance between speech sounds, but this index is not yet available. For the present we must approach the problem in a simpler way.

Imagine a many-dimensional space with a separate coordinate for each one of the different frequencies involved in human speech sounds. Along each coordinate plot the relative amplitude of the component at that frequency. In this hyperspace each unique speech sound is represented by a single point. Each point in the hyperspace represents a single acoustic spectrum. The group of similar sounds comprising a phoneme is represented by a cluster of points in the hyperspace. If a language utilized only two different phonemes, the hyperspace could be split into two parts, one for each phoneme. The distance between the two phonemes could be made as large as the vocal mechanism permits, and discrimination would be relatively easy. But suppose the number of different phonemes in the language is increased from two to ten. With ten different phonemes the hyperspace must be divided into at least ten subspaces, and the average distance between phonemes must be smaller with ten phonemes than it is with two. The discriminations involved must be correspondingly more precise. If the number of alternative phonemes is increased to a thousand, then the listener is required to make even more precise discriminations.

In other words, the ease with which a discrimination of speech sounds can be made is limited according to the number of different speech sounds that must be discriminated. From this line of reasoning it follows that the number of alternatives can be used to gauge the difficulty of discrimination. This argument has been developed by Shannon (**5**) to give a measure of the amount of information in a message. The interesting aspect of this index of difficulty, or of amount of information, is that it does not depend upon the characteristics of the particular item,

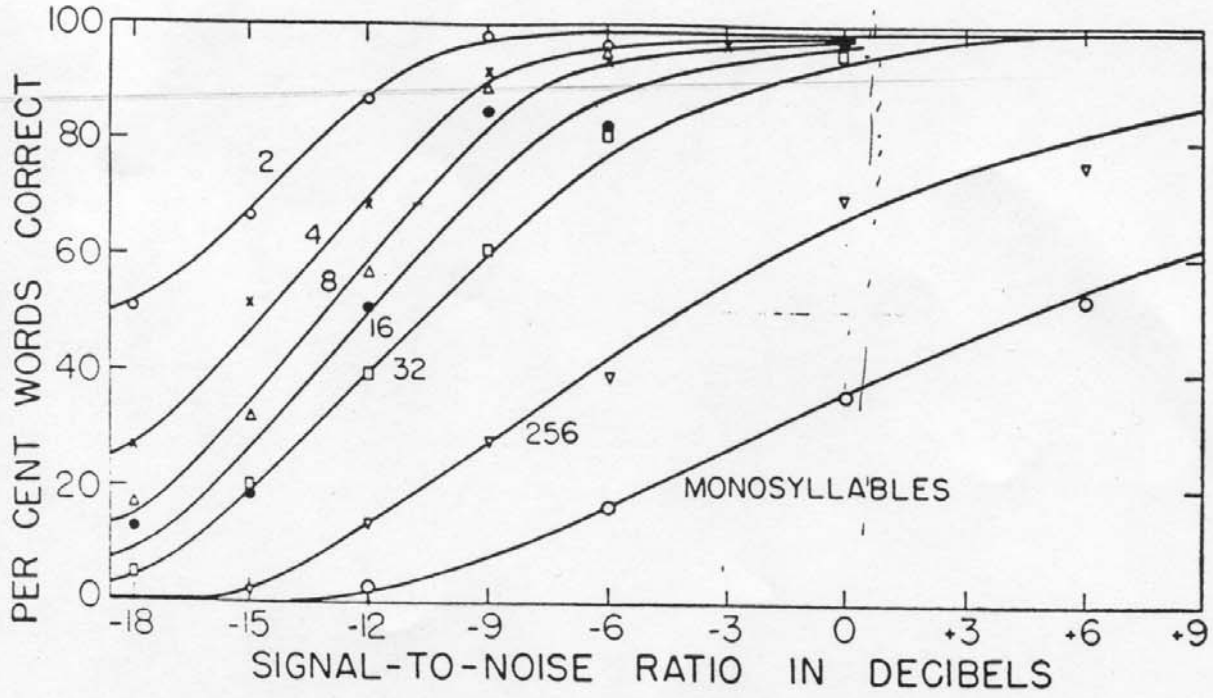

FIG. 2. Intelligibility of monosyllables as a function of the size of the text vocabulary. (Data are not corrected for effects of chance.)

but upon the range of items that *could* occur.

The range of alternatives was used as the experimental variable in the following way. The listener was informed that each test word would be one of the items from a given restricted vocabulary. The size of the test vocabulary was alternatively 2, 4, 8, 16, 32, or 256 words. The talker always spoke one of the words from the prearranged list.

The words used in the restricted vocabularies were chosen at random from the list of phonetically balanced monosyllables published by Egan (2). For the two-alternative vocabulary, different pairs of words were chosen and typed on the listener's answer sheet. The talker read one of the pair, and the listener checked the item he heard. A similar procedure was used for the four- and eight-word vocabularies. For the 16-, 32-, and 256-word vocabularies the listeners had a list of all the words before them, and studied this list until they made their choice. The choice was recorded and a signal given to the talker to proceed to the next item. The *S*s studied carefully the particular list used in any test and arranged the words according to the vowel sounds before the tests began.

The results are summarized in Table I and in Fig. 2. Included with the data for restricted vocabularies are data for words from the original list of 1,000 monosyllables, obtained with no list of choices available to the listener. When these data are corrected for chance, the two-word vocabulary gives a threshold (50 per cent of the words correct) at −14 db, the 256-word vocabulary gives a threshold at −4 db, and the unrestricted list of

TABLE I

PER CENT WORDS CORRECT IN ARTICULATION TESTS WITH VOCABULARIES OF VARIOUS SIZES

S/N in db	Size of Vocabulary							
	2	4	8	Digits	16	32	256	Monosyllables
−21	49							
−18	51	27	17	13	13	5		
−15	67	52	32	38	19	20	2	
−12	87	69	57	73	51	39	14	3
− 9	98	92	89	92	85	61	28	
− 6		94	95	99	82	81	39	17
− 3		96						
0				100	97	95	70	37
+ 6				100			76	53
+12				100			90	70
+18								82

monosyllables gives a threshold at +4 db. With the same test words the threshold is changed 18 db by varying the number of alternatives. This result supports the argument that it is not so much the particular item as the context in which the item occurs that determines its intelligibility.

Context of the Sentence

A word is harder to understand if it is heard in isolation than if it is heard in a sentence. This fact is illustrated by Fig. 3. Sentences containing five key words were read, and the listener's responses were scored as the percentage of these key words that were heard correctly. These data are shown by the filled circles in Figs. 1 and 3. For comparison, the key words were extracted from the sentences, scrambled, and read in isolation. The scores obtained under these conditions are shown by the open circles of Fig. 3. The removal of sentential context changes the threshold 6 db.

The effect of the sentence is comparable to the effect of a restricted vocabulary, although the degree of restriction is harder to estimate. When the talker begins a sentence, "Apples grow on ———," the range of possible continuations is sharply restricted. This restriction makes the discrimination easier and lowers the threshold of intelligibility. A detailed statistical discussion of the restrictions imposed by English sentence structure is given by Shannon (**5**), and is used in a simple recall experiment by Miller and Selfridge (**4**).

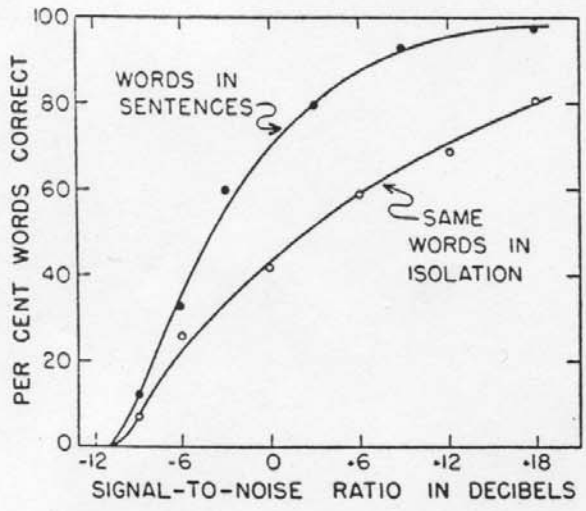

Fig. 3. Effect of sentence context on the intelligibility of words

Effects of Repetition

When an error occurs in vocal communication, the listener can ask for a repetition of the message. The repeated message is then heard in the context provided by the original message. If the original message enabled the listener to narrow the range of alternatives, his perception of the repeated message should be more accurate. A series of tests were run with various kinds of test materials to evaluate the importance of the context of repetition. These tests were run with automatic repetition of every item and, also, with repetitions only when the listener thought he had not received the test item correctly.

The improvement in the articulation scores obtained with automatic and with requested repetitions was found to be about the same. A slight but insignificant difference was found in favor of the requested repetition, and if we add to this the savings in time achieved by omitting the unnecessary repetitions, the requested repetition is clearly superior.

The advantage gained by repetition is small for all types of test materials. In Fig. 4 data are given for the effects of repeating automatically the monosyllabic words. The difference in threshold between one presentation and three successive presentations is only 2.5 db. Similar data for words heard in sentences show a shift of 2 db, and for digits, 1.5 db.

These results indicate that the improvement that can be achieved by the simple repetition of a message is

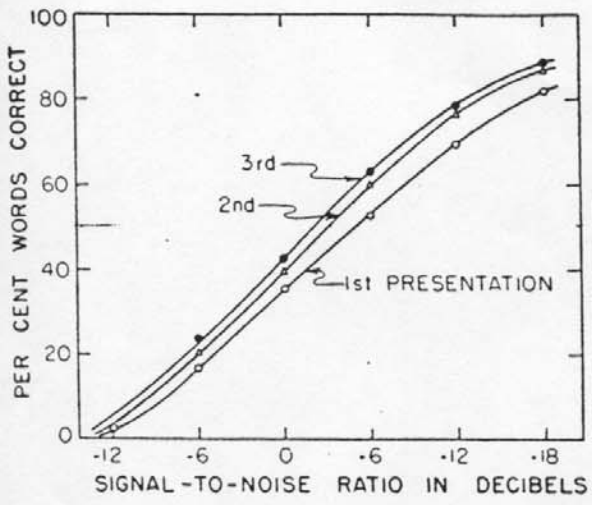

FIG. 4. Effects of repetition of test words on the articulation score

slight. The repeated message contains approximately the same information, and the same omissions, that the original message contained. If the listener thinks he heard the word correctly, he persists in his original response, whether it is right or wrong. If he thinks he heard the word incorrectly, he does not use this presumably incorrect impression to narrow the range of possibilities when the item recurs. In any case, no strong factor is at work to improve the accuracy on repeated presentations, and so we obtain only the slight improvement indicated in Fig. 4.

The results indicate that far more improvement in communication is possible by standardizing procedures and vocabulary than by merely repeating all messages one to two times.

In general, therefore, the results are in qualitative agreement with the mathematical theory of communication presented by Shannon (**5**). A precise quantitative comparison of the data with the theory cannot be made in the absence of trustworthy information about the distributions of errors. Seemingly reasonable assumptions about the error distributions give results consistent with theoretical predictions, but a more thorough study would be rewarding. For a given signal-to-noise ratio the listener receives a given amount of information per second (according to Shannon's definition), and articulation scores can be predicted for different types of test materials on the basis of the average amount of information needed to receive each type of test item correctly.

SUMMARY

Articulation tests showed the effects of limiting the number of alternative test items upon the threshold of intelligibility for speech in noise. The number of alternative test items was limited by providing three kinds of context: (*a*) restricting the size of the test vocabulary, (*b*) using the words in sentences, and (*c*) repeating the test words. Differences among test materials with respect to their intelligibility are due principally to the fact that some materials require more information than others for their correct perception. The relative amount of information necessary for a given type of item is a function of the range of alternative possibilities. As the range of alternatives increases, the amount of information necessary per item also increases, and so the noise level must be decreased to permit more accurate discrimination.

(Manuscript received April 24, 1950)

REFERENCES

1. EGAN, J. P. *Articulation testing methods, II.* OSRD Report No. 3802, February, 1942. (Available through Office of Technical Services, U.S. Department of Commerce, Washington, D.C., as PB 22848.)
2. EGAN, J. P. Articulation testing methods. *Laryngoscope*, 1948, **58**, 955–991.
3. FLETCHER, H., AND STEINBERG, J. C. Articulation testing methods. *Bell Syst. Tech. J.*, 1929, **8**, 806–854.
4. MILLER, G. A., AND SELFRIDGE, J. Verbal context and the recall of meaningful material. *Amer. J. Psychol.*, 1950, **63**, 176–185.
5. SHANNON, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.*, 1948, **27**, 379–423, 623–656.