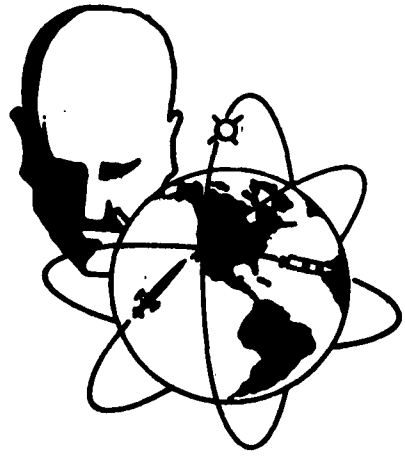
SPEECH-INTELLIGIBILITY AND TALKER-RECOGNITION TESTS OF AIR FORCE VOICE COMMUNICATION SYSTEMS

TECHNICAL DOCUMENTARY REPORT NO. ESD-TDR-63-224
FEBRUARY 1963

Stephen E. Stuntz

OPERATIONAL APPLICATIONS LABORATORY
DEPUTY FOR TECHNOLOGY
ELECTRONIC SYSTEMS DIVISION
AIR FORCE SYSTEMS COMMAND
L. G. Hanscom Field, Bedford, Massachusetts



PROJECT 7684

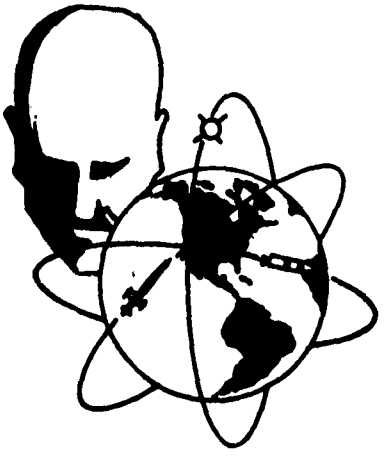
ESD-TDR-63-224

SPEECH-INTELLIGIBILITY AND TALKER-RECOGNITION TESTS OF AIR FORCE VOICE COMMUNICATION SYSTEMS

TECHNICAL DOCUMENTARY REPORT NO. ESD-TDR-63-224
FEBRUARY 1963

Stephen E. Stuntz

OPERATIONAL APPLICATIONS LABORATORY
DEPUTY FOR TECHNOLOGY
ELECTRONIC SYSTEMS DIVISION
AIR FORCE SYSTEMS COMMAND
L. G. Hanscom Field, Bedford, Massachusetts



PROJECT 7684

ABSTRACT

Word-intelligibility (psycho-acoustic), Articulation Index (electrical) and voice recognition tests were made on the Thule, Greenland, to Cape Dyer, Baffin Island, link of the DEWDROP tropospheric-scatter communication system, and on the Gander, Newfoundland, to Harmon Air Force Base, Newfoundland, link of the POLEVAULT tropospheric-scatter communication system, and compared against similar tests taken on a laboratory reference system. Harvard phonetically-balanced (PB) word-lists were used in the psycho-acoustic tests; an abbreviated octave-band form of the Articulation Index procedure (after Kryter) was the basis of the electrical tests. These two measures agreed in ranking the three systems in order of intelligibility from highest to lowest: laboratory, POLEVAULT and DEWDROP. It is concluded that the Articulation Index technique suitably modified, is feasible and useful for voice-system performance evaluation and quality control testing. It is also concluded, tentatively, that system characteristics affecting intelligibility do not necessarily affect listeners' ability to recognize individual talkers.

Reviewed and approved for publication.

HERBERT RUBENSTEIN, Chief
Information Presentation Division
Operational Applications Laboratory

WILLIAM V. HAGIN, Col, USAF
Director
Operational Applications Laboratory

CONTENTS

Project 7684

SPEECH-INTELLIGIBILITY AND TALKER-RECOGNITION TESTS OF AIR FORCE VOICE COMMUNICATION SYSTEMS

Stephen E. Stuntz

INTRODUCTION

To evaluate the quality of service provided by voice-communication systems, it has long been the engineering practice to express system performance in terms of steady-state signal parameters each as frequency-response, attenuation of a test-tone ("transmission loss") over the path of interest, level of inherent noise, and harmonic distortion of a test-tone. Rigorous standards of acceptability are based on data of this sort. Just how they relate to the transmission of intelligible speech is not clearly apparent, either from published technical manuals or from the literature of research and development. That a relationship is inferred by somebody is evident from a "Merit Rating" scheme which classifies ranges of signal-to-noise (presumably speech-to-noise) ratio figures along a five-point quality scale, with notations of the grade of service to be expected at each merit-rating level. The source of this information is AACSM 100-5 dated 1 February 1959, Figure 13-7, page 13-13. A number of critical comments might be offered regarding this particular mode of system evaluation, but would probably be of little significance, since it has fallen into almost total disuse. As far as can be determined at this time, the concept of speech intelligibility does not figure in present engineering practices.

During the last fifteen years there have emerged two methods for more or less directly measuring the intelligibility of speech transmitted between talker and listener by electrical techniques. Both of them proceed from known characteristics of spoken language. One of them, the psychoacoustic method, measures communication in terms of the accuracy with which listeners receive standardized messages spoken by live talkers over the system being tested. This approach has been very widely used, mostly in the laboratory. Its sensitivity to system variables has been well established, with particular regard to the effects of talker and listener language-background, background noise, frequency-response, interruptions, and various types of amplitude distortion such as limiting and peak-clipping. Compared with run-of-the-mill engineering tests it is cumbersome and time-consuming, but does give valid and highly reliable measurements when properly conducted. Its final datum is a "word articulation score" (WAS) expressing intelligibility in percent of correctly received words.

The other method, known as the Articulation Index, samples signal-to-noise conditions at from five to twenty critical points in the spectrum of speech sounds by purely electrical measurement. While it does not allow for talker or listener language peculiarities, it quite adequately covers the effects of the system variables mentioned above. It gives a simple index running from zero to one (0 to 1) showing the speech-handling capability of the system under test. The procedure is much less involved than the psychoacoustic test, since the data can be taken with available signal-generators, filters and transmission measuring sets, and requires only simple multiplication and addition to get the final answer.

Rather surprisingly, considering the sophistication of the present engineering arts, neither of these tests has been widely used in system evaluation. The first one (word articulation) has become almost a laboratory standard for communication research; most of our theoretical knowledge about voice reception stems from applications of it in various forms. The articulation index technique has very seldom been used as a laboratory-research tool, and only sparingly in field evaluation of communication systems, despite considerable refinement and simplification (4) since it first appeared fifteen years ago (3). Recently its validity has been reaffirmed by new research investigations (7), on the basis of which it was strongly recommended "to evaluate the performance of many speech communication systems when speech intelligibility testing of the system is not practical."

Users of reduced-bandwidth systems have been concerned about preserving the features of talkers' voices which enable listeners to tell with certainty who is talking. It has been recognized that voice messages contain other information besides the meaning of the words themselves, not only as to the identity of the person speaking, but also as to his emotional and physiological status. Experimental research on the matter has not as yet revealed any fundamental invariants which can be described in simple form, although the conviction prevails that "there is something there". Up to the present, most of the effort has been spent in trying to measure the effects of the usual transmission variables on talkers' voices, to see how well listeners can identify them over systems containing various amounts of noise, frequency-distortion and amplitude nonlinearity. So far the findings have been piecemeal and highly specific to the conditions imposed; few generalizations have come to light which might be built into theoretical explanations as to how people recognize other people by their voices.

PURPOSES OF THIS EXPERIMENT

The present study was designed for three purposes:

1. To determine the relative feasibility of using talker-listener and Articulation Index techniques for evaluating voice systems in the field.

2. To confirm, from data taken on working real-life systems, the correlation between Psychoacoustic and Articulation Index methods of estimating intelligibility.

3. To examine the relationship between intelligibility and talker-recognition with respect to the transmission characteristics of selected Air Force systems.

A fourth purpose, incidental but expected, was a comparison of the several systems' capacity to transmit intelligible speech and to preserve the individual identity of speakers.

PROCEDURE

Three tape-recorded voice communication tests were transmitted over two Air Force point-to-point systems and one experimental laboratory system. The tests were: (1) phonetically-balanced word-lists (for talker intelligibility); (2) lists of short declarative sentences (for talker identification); (3) 10 second bursts of white noise alternating with 10-second no-signal intervals. The systems were: (a) DEWDROP tropospheric-scatter link between Thule, Greenland and Cape Dyer, Baffin Island; (b) POLEVAULT tropospheric-scatter link between Gander Air Station and Ernest Harmon Air Force Base, Newfoundland; (c) a laboratory tape playback-loudspeaker setup. Tests and systems are described in detail in the section to follow.

A. TESTS

(1) Intelligibility word-lists. Under laboratory conditions, four talkers (two male, two female) each recorded four equivalent but not identical 50-item lists of phonetically-balanced monosyllabic English words, taken from published sources (Beranek (1) pp 770-772). Talkers were instructed to maintain uniform loudness, and monitored their output by means of a volume indicator. All tests were transmitted through all systems.

(2) Talker-identification sentences. Five teams of four male talkers each recorded under laboratory conditions three lists of twenty short declarative sentences per team, in the following manner: each talker in turn read from two to four sentences at the beginning and end of which he identified himself with a number (e.g. "This is talker 1", "This is talker 2", and so on), to provide preliminary

listener-training in associating a particular voice with a particular identity. Following this training sequence, each talker read one sentence followed by a pause and his identification number. All talkers read a total of five sentences each; talkers appeared in random order throughout the 20-item test sequences. Sentences were taken from published sources (Beranek, (1) pp 774-777). All sentences were transmitted through all systems.

The tests described above, after transmission and re-recording in the manner discussed under Systems below, were presented by means of a tape playback-loudspeaker system (Ampex 600 and 620) to groups of listeners in a low-reverberation, isolated laboratory room. In the word-intelligibility tests, they were required to write down each test word as they heard it. Their answers were then checked against master-lists used by the talkers during original recording; the numbers of right answers given by each listener on each test for each system were the quantitative data which were analyzed as described later. In the talker-identification tests, listeners wrote the number assigned to a given talker, immediately after he uttered each test-sentence; after a brief pause on the tape, the talker's correct identity-number was announced and written by the listener. Where the listener's response-number and the announced identity-number agreed, the listener was credited with a correct answer. Thus for each test the numbers of correct answers given by each listener for each system became the data by which talker-identification was measured.

(3) Noise-bursts. The output of a noise generator (Grason-Stadler type 455-A) was connected to the input of an Ampex type 600 portable tape-recorder. With recording gain set at maximum undistorted level (indicated by -3 VU on the recorder meter), a series of ten-seconds-on, ten-seconds-off cycles was recorded.

In addition to (3) above, a sine-wave tone of 800 cycles was recorded at a constant level for three minutes for use as a gain-setting reference during play-back transmissions in the field.

B. SYSTEMS

At the terminals of both tropospheric-scatter systems tested, tape recording-playback equipment (Ampex type 600) was connected directly to transmitter-modulator input and receiver-demodulator output, eliminating local-service drops between transmitter-receiver sites and communication centers. Transmission over both systems was one-direction only; thus a tape playback at the transmitting end (e.g. "P" Mountain, Thule, and Gander) fed signals through the system to a tape recorder at the receiving end (Cape Dyer and

Lookout Mountain, Stephenville, Newfoundland).

(1) DEWDROP link, (System C) Transmissions from Thule to Cape Dyer were made on multiplex channel 13 (Group 2, Channel 1); data were taken with Talk-Level Regulator (TLR) in service and with Compandor equipment bypassed, i.e., out of service.

(2) POLEVAULT link, (System B) Transmissions from the U. S. Air Force tropo-scatter site, Gander Air Station, to Harmon Air Force Base, Stephenville, Newfoundland, were made on a central multiplex channel; circuitry equivalent to Talking-Level Regulator and Compandor were operative during the tests.

(3) Laboratory system, (System A) Original recordings used for transmission of intelligibility word-lists and talker-identification sentences were played via Ampex 600 playback and associated Ampex 620 amplifier-loudspeaker directly to listeners in acoustically-isolated, low-reverberation laboratory room. No electrical compression, filtering or other processing devices were interposed between playback and speaker-amplifier.

Figure 1 shows in simplified block diagram the equipment configuration of both field and laboratory systems.

C. SUBJECTS

Listeners for all psychoacoustic tests (intelligibility and talker-identification) were female college students with clinically normal hearing. In the intelligibility series, a crew of five listeners was trained on one half of the word-lists until they achieved group average of 99% correct responses on two successive presentations of equivalent but not identical 50-word lists. The crew was then tested on the remaining half of the word-lists. In the talker-identification tests, a crew of four listeners was trained on five of the twelve tests recorded, one 20-item test per talker team. Thereafter the listening crew was tested on the remaining ten identification tests.

TREATMENT OF DATA

For both word-articulation and talker-identification tests, mean percent correct responses were computed for each of the three systems. In addition, data from both tests were subjected to analysis of variance.

From the noise-sample recordings, sound spectrograms were prepared by means of a Kay Electric Company Sonagraph, utilizing

the section-display feature of the instrument. This affords a frequency-versus-amplitude picture of the contents of a 5-millisecond segment drawn from a 2.4 second sample of the recorded material, over a dynamic amplitude range of 35 db and along a frequency scale extending from 80 to 8000 cycles per second. Measurement with suitably calibrated rules along the frequency and amplitude axes yielded signal-plus-noise and noise-only data in the frequency-bands 150-300, 300-600, 600-1200, 1200-2400 and 2400-4800 cycles per second. For each frequency band a signal-plus-noise/noise ratio was determined by subtracting the noise amplitude figure (in db) from the signal-plus-noise figure (also in db). For each of the three systems, the ratios so derived were weighted according to the recommendations of Kryter (2), and added to give the articulation index (AI). By reference to a published nomograph,(2, Fig. 15), the obtained AI for each system was converted to a word-articulation percentage score.

Each of the three systems, then,is described by a percentage-correct word articulation score obtained by averaging the scores of five listeners on two word-list tests; by a percentage-correct talker identification score averaged over four listeners and five tests; by an articulation index computed from signal and noise measurements; and by a percentage-correct word articulation score derived from previously-determined relationships between AI and word articulation.

RESULTS

Table I shows average obtained percent-correct word scores, obtained articulation indices, and predicted percent-correct scores for all three systems.

TABLE I

	Obt. WA%	SE%	Obtained AI	Predicted WA% (from Fig. 2)
System A (Lab)	99.2	0.45	0.989	99.0
System B (POLEVAULT)	94.8	1.11	0.877	97.0
System C (DEWDROP)	78.0	6.55	0.392	60.0

Table II summarizes results of analysis of variance in the word-articulation data, based on raw scores (number of words correctly understood).

TABLE II

	Sum of Squares	DF	Mean Square	F	Significance
T(alkers)	732.60	3	244.20	58.70	.001
L(isteners)	7.56	4	1.89	...	....
S(ystems)	1245.40	2	622.70	149.69	.001
T x L	15.78	6	2.63	...	....
T x S	707.40	12	58.95	14.17	.001
S x L	16.42	8	2.05	...	....
T x L x S (error)	99.75	24	4.16	...	....
Total	2824.91				

Articulation Index Data. Employing the approach described by Kryter, the signal-plus-noise to noise ratios for five octave-wide frequency bands were obtained from the Sonagraph sections of noise samples through each system. These data, the weighting factors, and resulting Articulation Indices (AI's) are shown in Table III.

TABLE III

		(A) LAB SYSTEM		POLEVAULT		DEWDROP	
OCTAVE BAND	WEIGHT	S+N/N	WTD. S+N/N	S+N/N	WTD. S+N/N	S+N/N	WTD. S+N/N
150-300 cps	.0013	30 db	.039	30 db	.0390	10 db	.0130
300-600	.0042	30	.126	29	.1218	7	.0294
600-1200	.0067	30	.191	27	.1674	5	.0335
1200-2400	.0105	30	.315	27	.2835	15	.1575
2400-4800	.0106	30	.318	25	.2650	15	.1590
	SUM = A.I. =		.989		.877		.392

Typical Sonagraph sections, taken from recorded noise-

transmissions, are shown in Fig. 3, which also indicates the octave-band limits listed in Table III.

Results of the talker-recognition tests are shown in Table IV.

TABLE IV

PERCENT CORRECT RECOGNITION OF TALKERS,
for all five teams and three systems

SYSTEM	TALKER TEAMS I	II	III	IV	V	Av. for Systems
A (Lab)	78.8	70.0	59.4	78.8	76.3	72.7
B (Polevault)	91.3	71.9	63.1	78.8	75.6	76.1
C (Dewdrop)	88.8	69.4	48.1	65.0	65.6	67.4
Av. for Teams	86.3	70.4	56.9	74.2	72.5	- - -

Table V summarizes results of analysis of variance in the talker-identification data, based on raw scores (number of correct identifications).

TABLE V

	Sum of Squares	DF	Mean Square	F	Significance
T(alker-Teams)	210.85	6	38.78	16.91	.001
L(isteners)	39.31	8	3.88	1.87	.10
S(ystems)	31.03	12	3.28	1.58	.10
T x L	58.59	2	32.22	14.04	.001
T x S	34.55	3	11.52	5.54	.10
S x L	22.11	4	5.53	2.66	.10
T x S x L (error)	49.81	24	2.08	.....	....
Total	446.25				

DISCUSSION

Intelligibility. From results displayed in Table I, it appears that the psychoacoustic (word-articulation) and electroacoustic (articulation index) methods substantially agree in rank-ordering the three systems with respect to intelligibility. This suggests that they may in fact be measuring the same thing - the effect of system characteristics on speech transmission. While the label "system characteristics" applies to a complex of hardware and transmission-medium parameters (e.g., noise, distortion, frequency-response and the like), its significance is attested by its contribution of over two-thirds of the total variance in the word-articulation data (Table II). While the present study does not afford a finer definition of "system characteristics", the fact that the octave-band unweighted signal-to-noise measures of Table III clearly differentiate the three systems suggests that system noise may be most influential in delimiting the voice-handling capabilities of these systems. Frequency-response is, of course, another potential limitation. System A, being essentially flat to at least 7,000 cps, might be expected to yield higher word-articulation scores and articulation indices than the other two systems, which in fact it did. However, it was also the most noise-free of the three as well. Systems B and C, having practically identical frequency-response (see Figure 3) characteristics, show word-articulation differences which are highly significant in the statistical sense; in view of the signal-to-noise data on these two systems, it appears that the effect of system noise outweighs that of frequency-response. Thus we may conclude that the high proportion (about two-thirds) of total variance uncovered in the analysis of word-articulation data (Table II) attributed to "Systems" is due primarily to noise-levels in the three systems.

Of further interest in assessing the capabilities of the three systems and the transmission-characteristics they exhibit is the effect of differences among talkers, upon listeners' reception-accuracy. As shown by the first line in Table II, talker characteristics account for about one-sixteenth of the total variance, a highly significant fraction. Since two of the talkers were male and the other two female, it might be predicted that the systems tested favor one sex over the other. This prediction is partially borne out by a further analysis of the word-articulation data, which shows a moderately-significant advantage in favor of the male talkers in this experiment (CR = 4.90, significant at the .05 level of confidence). However, sex-difference alone does not explain all the findings. While Figure 4 suggests that female voices become unintelligible more rapidly than male voices as system conditions deteriorate (that is, as S/N gets lower), it is worth noting that under the worst

noise tested (System C), individual differences among talkers, male and female alike, become much more noticeable. That is, their word-intelligibility scores spread farther apart. While not otherwise reflected in the data, it was noted that the two female talkers had rather strong Northeastern-United States ("New England") accents; the two males were pretty good examples of the General American ("Midwestern") pattern. There were also some subtler articulatory differences between the males and between the females, which might have further influenced their individual intelligibility. One might generalize that freedom from strong regional accent is a prerequisite for intelligible transmission over noisy voice circuits. Whether this consideration overrides sex-difference effects cannot be estimated from the present findings.

Articulation Index. From Table III it can be seen that the purely electro-acoustic measures (articulation indices, calculated from weighted octave-band signal-to-noise ratios) substantially agree with the obtained psychoacoustic (word-articulation percentage) scores in ranking the three systems from highest to lowest in speech-transmission capability. That is to say, they appear to be highly and positively correlated, to the end that they both appear to measure the same thing despite their basic difference of technique. This confirms the considerably more rigorous findings of French and Steinberg (3) and Kryter (2). Of course, the Articulation Index is not inherently sensitive to the linguistic, phonemic and personal variables which affect listeners' comprehension of talkers. However, it is much less laborious and time-consuming to determine than is the word-articulation measure, and has been shown (3) to accurately reflect the influence of such electrical variables as microphone and receiver response, and by Kryter (2) the effects of peak-clipping and frequency distortion. Kryter, Flanagan and Williams (4), conclude that "..... the octave band method for the calculation of AI can be used in place of the more detailed 20 band method without any appreciable loss in the accuracy with which speech intelligibility test scores are predicted." Within the limits of the present data, this is confirmed: the articulation indices calculated from measurements on the three systems were projected onto the 1,000 PB-word curve of Kryter (2), reproduced here as Figure 2, to get the predicted word-articulation scores shown in the extreme right-hand column of Table I. While the predicted percent-correct scores so obtained do not agree in exact magnitude with the scores obtained by measurement of the three systems, both sets of scores arrange the systems in the same rank-order, viz: highest, system A (laboratory); second, system B (POLEVAULT); lowest, system C (DEWDROP). On the basis of rank agreement among the three sets

of measures (obtained word-scores, obtained articulation indices, and word-scores predicted from obtained articulation indices) it seems reasonable to accept the highly simplified articulation-index method used here as a valid way of assessing the speech-handling capacities of communication systems.

Talker identification. Analysis of the variance contained in the talker-identification data is summarized in Table V. It reveals two highly significant effects: first, that the behavior of the talkers themselves profoundly influenced listeners' judgment of talker identity, and second, that a talker-listener interaction occurred. Examination of the raw data shows that discriminating the talkers of some teams was consistently harder for some of the listeners than for others. The explanation of this is not clear; (talkers had not been previously equated or scaled for inherent discriminability), nor was each talker systematically teamed with all other talkers. There is a suggestion that the voices of some teams sounded much more alike to all listeners than did the voices of other teams; Table IV illustrates this point, and also shows how the different teams were affected by the three systems. Comparing the "Total" entries for the five teams of talkers, the talkers of Team I are seen to be most easily distinguished from eachother, while those of Team III are hardest to tell apart. This difference is statistically significant (CR = 7.88, $p = .001$), and says in effect that recognition of talkers depends not only on the individual characteristics of each talker, but also - and very heavily - on the characteristics of the other talkers with whom he is being compared. In a word, the listeners' discrimination task changed with each change of talker-team; listeners had to adjust their discrimination standards from team to team.

The analysis of variance indicates that the transmission characteristics of the three systems may have had some effect on the listeners' ability to distinguish voices in the various teams. Table IV gives a detailed account as to how each team fared. The recognizable differences among talkers in Teams I, II and III seem to have been enhanced somewhat by the characteristics of System B; Teams IV and V were practically unaffected. In the case of Team I, a statistically reliable improvement was wrought by System B; something about that system rendered four highly distinguishable voices even more so (CR = 2.25, $p = .03$). This effect persisted in System C, although not to a significant extent. Systems B and C differed from System A in one aspect: they incorporated sharp-cutoff lowpass filters (nominal value, 3000 cycles; see the spectrograms of Figure 3). Despite the confounding effects of the much lower signal-to-noise levels in System C, as indicated in Table III, it is possible that lowpass filtering does

something to preserve and perhaps even exaggerate those cues which listeners need to identify different talkers' voices. This effect is not mentioned in the research literature specifically, although Pollack, Pickett and Sumby (5) indicate that no serious impairment of talker-recognition is produced by lowpass filtering until cutoff frequencies below about 2000 cycles are imposed. Peters (6) tended to confirm this with data on four voices, and also found that when octave-wide passbands were used in the range 150 to 4800 cycles per second, the band 1200-2400 cycles permitted transmission of significantly more talker-identity information than any of the others.

It is appropriate to observe here that the task of trying to identify the features of speech and voice which lead listeners to single out particular individuals from a whole ensemble of talkers is beset by at least two grave difficulties. First, each sample of talkers delimits the range of comparisons immediately available to the listener, laying open to challenge the generality of any conclusions one may draw from a highly specific experiment. Second, each listener brings to a given talker-identification experiment certain by-products of previous experience which affect his criteria of judgment, and so render this type of investigation very difficult to control with scientific rigor. Strictly speaking, there probably is no such thing as a naive observer where talker-identification is concerned. These two comments omit any consideration of the multi-level complexity of the stimulus itself.

Comparison of Word Intelligibility and Articulation Index Tests. It has already been shown that the three systems' speech-handling capacity can be pretty sharply delineated by either the talker-listener method or the articulation index analysis alone, and that the results of both agree in ranking the systems from highest to lowest as to intelligibility. (In the language of statistics, there is a very high positive rank-order correlation.) Figure 2 details the general relationships between the Articulation Index and intelligibility of various kinds of voice messages. The data from which these curves were drawn result from systematic experiments conducted in several laboratories. The curves represent averages of many tests taken with a wide variety of talkers and listeners, under controlled laboratory conditions. Thus they may be taken as reliable predictors of speech transmission over systems whose characteristics can be described in terms of AI.

One of the aims of this investigation is to assess the two techniques for estimating intelligibility for use in the field as technical management tools. From the preceding discussion, it can be

seen that a great deal more time and technical effort went into collecting and analyzing the word-intelligibility data than was required for the articulation index measurements. In general, this was due to the fact that human performance was the basis of measurement in the former; since people vary, both within themselves from time to time and also from individual to individual on any task, time-consuming counterbalances had to be employed as in any experiment on human perception. Thus several talkers read several lists, which were run several times over each system, then played several times to several listeners, so that the average effects of system variables could be ascertained, while taking into account the circumstances which caused the people to vary in their behavior. As a matter of fact, the significance of the data was tested mainly in terms of the people's variability in response to system conditions. This required employing a standard, well developed technique designed for this purpose - the statistical analysis of variance - preceded by reduction of the data to a suitable form. The amount of time and specialized technical skill needed to prepare the tests and to obtain and analyze the data exceeded actual system-transmission time by several orders of magnitude. Figuring about five minutes each for 16 wordlists, this means that the minimum down-time per tested circuit ran about an hour and a half - and this did not give the final answer as to circuit quality. It is conservatively estimated that another 60 to 90 minutes per wordlist was required to select and record the words, test the listener-groups, reduce the listener scores, and finally analyze the data to get a final answer. (It is emphasized that the classic laboratory-research method of word-articulation testing was followed throughout, with no corner-cutting to see how far the procedure could be simplified without jeopardizing results.)

By contrast, the articulation index approach was much more economical in several ways, although it too employed laboratory-research methods. In the first place, circuit down-time ran about two minutes per test transmission, equally divided between periods of signal and no-signal. After each test-transmission there was immediately available the raw data ready for analysis. The costliest part of the procedure was the analysis, in both instrumentation and working time: it took about 30 minutes to write the spectrogram for a 50-millisecond slice from each of the recorded two-minute transmissions, mark this spectrogram off in octave intervals, measure sound-levels in the bands, figure signal-to-noise ratios, and calculate the Articulation Index for any two 50-millisecond subsamples (one of transmitted noise, one of system residual noise). While this represents a very high-ratio gain with respect to the Word Intelligibility method, and yields data of equivalent validity, the Articulation Index technique as applied in

this study is probably still too cumbersome and involved for routine use outside a well-equipped laboratory where work can proceed at a deliberate pace.

Relative Effects of Systems on Intelligibility and Identification of Talkers. In view of the marginal significance of system-connected variance in the talker-identification phase of these experiments, indicated both in Table V and in the "Av. for Systems" column of Table IV, one might surmise that the transmission variables which degrade intelligibility do not necessarily mask the cues a listener uses to identify individual speakers. In terms of the results emanating from the present experiment, however, this notion can only be proposed very tentatively.

Relative Intelligibility of the Three Systems. While not designed as a check on current engineering-management procedures, the tests reported here clearly confirm the word-of-mouth opinions expressed by operating personnel on the basis of hard experience: the Thule-Cape Dyer DEWDROP link was significantly inferior to the Gander-Harmon POLEVAULT link, as regards intelligibility. Admittedly, the data taken in this study do not pinpoint causes other than that the average signal-to-noise conditions of the DEWDROP path were significantly poorer than those of the POLEVAULT link - within the sampling limitations and at the particular time (September and October, 1961) of the tests. It is no surprise that the reference laboratory system surpassed both field systems, since all relevant variables were controlled specifically to enable the most rigorous comparisons. It is believed that significant modifications have now been made in the DEWDROP equipment which may have changed the picture; however, no evidence on this point is now available in terms comparable with the results of this experiment.

From the results presented in Table I, the Gander-Harmon POLEVAULT link is seen to give almost as high intelligibility as the reference laboratory system. Actually the difference in word-articulation percentage (WA %) between the two is highly significant in the purely statistical sense (CR = 3.7, $p = .01$); however the practical importance of this finding is diminished by the observation that both systems are capable of supplying highly intelligible communication. That is especially pointed up by reference to the curve of Figure 2 showing the relation between AI (articulation index) and sentence intelligibility, which probably represents real-life voice messages more reasonably than do test-transmissions of single, unassociated words. It will be seen that the laboratory systems AI of .989 and the POLEVAULT link's AI of .877 both translate into approximately the same

value of sentence intelligibility - 99%.

SUMMARY AND CONCLUSIONS

1. The results of this study have confirmed the relationship between AI and word-articulation techniques for estimating the effects of transmission characteristics on speech intelligibility. Additionally, they have demonstrated that this relationship may be extended to apply to operational communication systems in the field, where precise laboratory controls do not prevail.

2. It has been shown that the AI method requires very significantly less circuit down-time, off-line analysis of data, and prior technical knowledge to secure indicative results, in comparison with the word-articulation test procedure.

3. Tentatively, it appears that system characteristics which depreciate intelligibility do not necessarily have a like effect on listeners' identification of talkers.

4. As an incidental finding, it has been shown that the DEWDROP tropospheric-scatter link from Thule, Greenland, to Cape Dyer, Baffin Island, as it existed in the autumn of 1961, was significantly less capable of transmitting speech intelligibly than was the Gander, Newfoundland-Harmon Air Force Base POLEVAULT tropo link, as of the same time. Further, it appeared that the latter was probably as good, for practical communication purposes, as a laboratory-reference system in which adventitious variables had been carefully controlled.

From the foregoing, then, it is concluded that:

a. The AI test, with suitable modifications to shorten and simplify its procedure, would yield a valid and useful evaluation of voice systems in the field, for both quality-control and service prediction.

b. The AI procedure, being entirely electrical, is susceptible of simplification by use of electronic instruments selected or designed especially for the purpose. The minimum requirements might be - a signal source (for example, random-noise generator) at the transmitting end of the circuit under test, and a suitable transmission measuring set at the receiving end - for instance, a bank of one-third-octave, half-octave or octave-band filters, plus a calibrated indicator such as a VU-meter.

c. It would appear that system-performance factors which

degrade intelligibility do not necessarily adversely affect voice recognition: it is not clear whether the reverse holds.

It may further be concluded that -

d. From the limited experience of this study, it seems that female talkers may be significantly less intelligible over noisy, band-limited circuits than male talkers.

e. Tentatively, it looks as if speech which is free of strong regional accent would be much more intelligible on noisy, band-limited circuits than speech which is so accented.

ACKNOWLEDGMENT

This study was conducted with co-operation and assistance from many sources, identified below and acknowledged with gratitude:

Headquarters, Air Force Communication Service - specifically Maj. Stanley E. Golon, Chief, System Standards Branch, DCS/Comm, for administrative support during the planning phase.

Headquarters, North Atlantic Region, AFCS - specifically Col. William E. Geyser, commanding officer, for guidance in selecting system-links for test, and in providing technical data about the characteristics of DEWDROP and POLEVAULT systems.

Thule DEWDROP Project Office, Federal Electric Corp., Thule AB, Greenland, Mr. Clifton F. Foss, Project Engineer, for technical assistance.

Federal Electric Corp. DEW communications office, Cape Dyer, Baffin Island, Mr. Larry E. Strople, Sector Chief C&E, for technical assistance.

Personnel of the 1933 Communications Squadron, Harmon AFB, and Gander Air Force Station, Newfoundland, for technical assistance.

First Lt. Albert Whitehurst, USAF, Operational Applications Laboratory, ESD, for technical assistance in collecting data on both DEWDROP and POLEVAULT systems.

REFERENCES

1. Beranek, L. L. Acoustic Measurements. New York: McGraw-Hill (1949)

2. Kryter, K. D. Proposed Methods for the calculation of the Articulation Index. Contract AF 19(604)-4061. ESD TDR-62-35 (October 1961).

3. French, N. R. and J. C. Steinberg. "Factors governing the intelligibility of speech sounds", Journal of the Acoustical Society of America 19 1949, pp 90-119.

4. Kryter, K. D., G. Flanagan and C. Williams. A test of the 20 Band and Octave-Band Methods of Computing the Articulation Index. Contract AF 19(604)-4061. ESD-TDR-62-4 (October 1961).

5. Pollack, I., J. M. Pickett and W. H. Sumby. "On the identification of speakers by voice", Journal of the Acoustical Society of America, 26, 1954, pp. 403-406.

6. Peters, Robert W. Studies in Extra-Messages: Listener Identification of Speakers' Voices Under Conditions of Certain Restrictions Imposed Upon the Voice Signal. Contract N6onr 22525. U. S. Naval School of Aviation Medicine Joint Project Report NM oo1 064.01.30 (October 1954).

7. Kryter, K. D. The Validity of the Articulation Index. Contract AF 19(604)-4061 ESD-TDR-62-3 (October 1961).

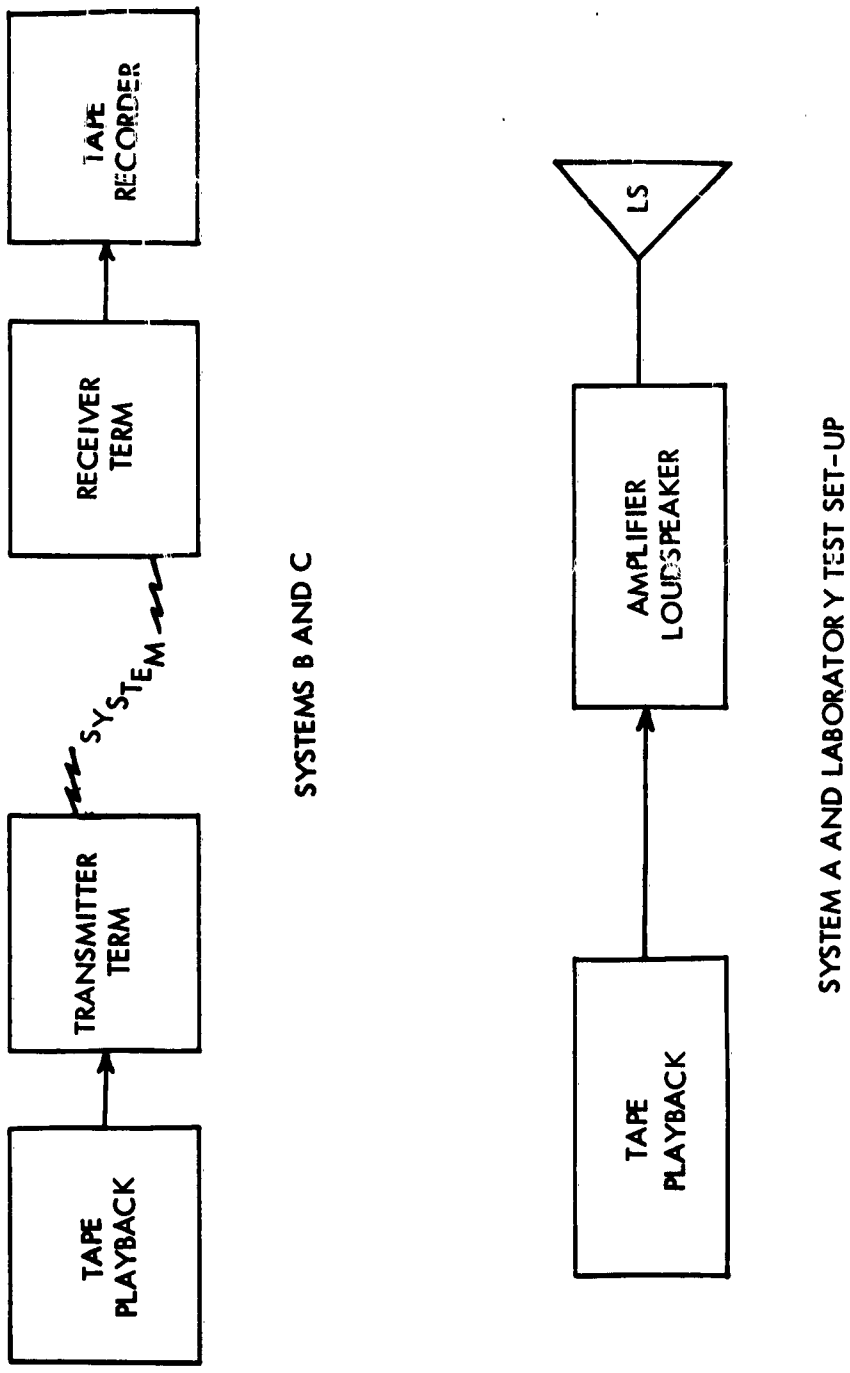


FIGURE I. BLOCK DIAGRAM OF SYSTEMS

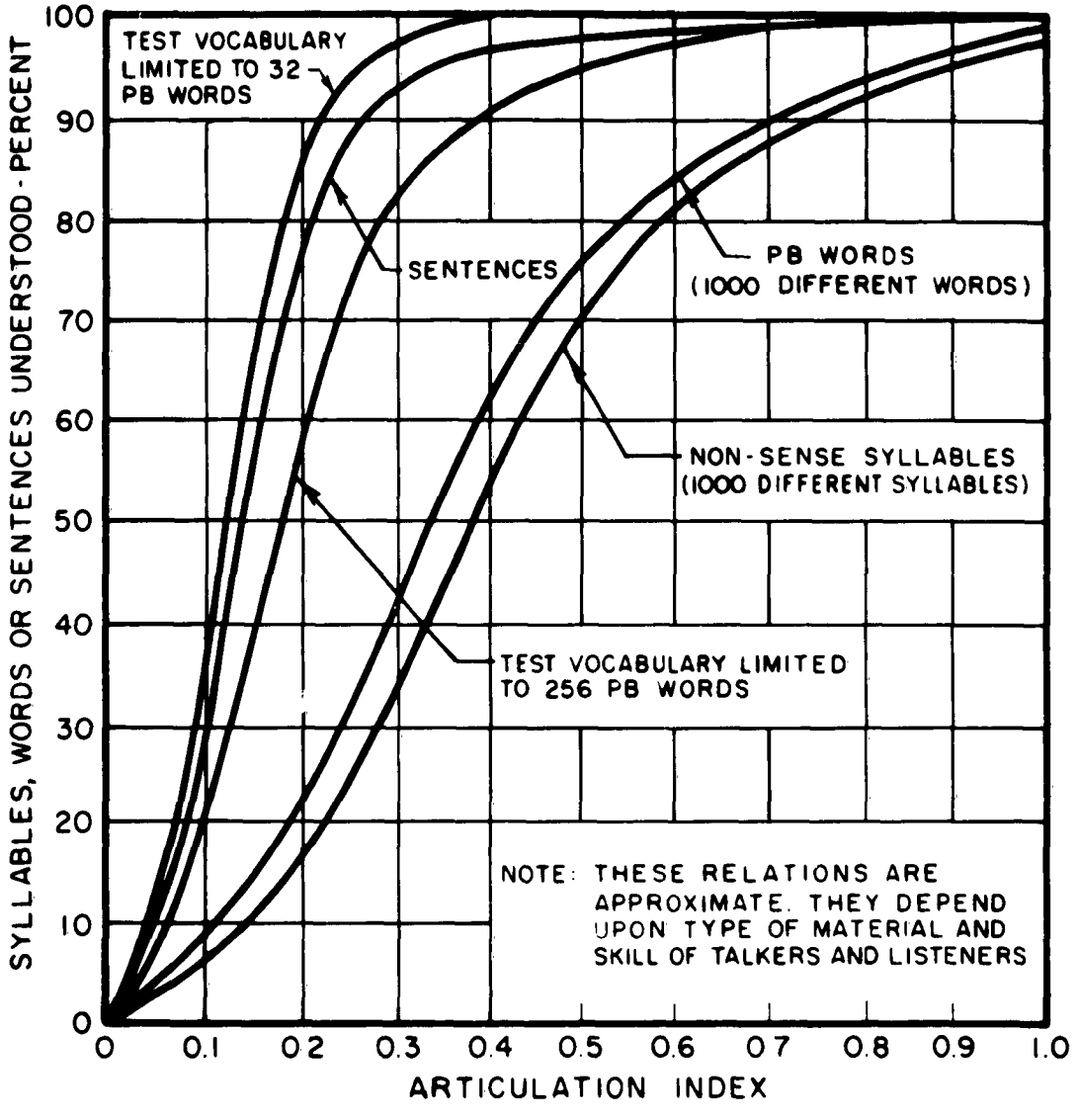


FIGURE 2. RELATIONSHIPS BETWEEN THE ARTICULATION INDEX (AI) AND VARIOUS MEASURES OF SPEECH INTELLIGIBILITY

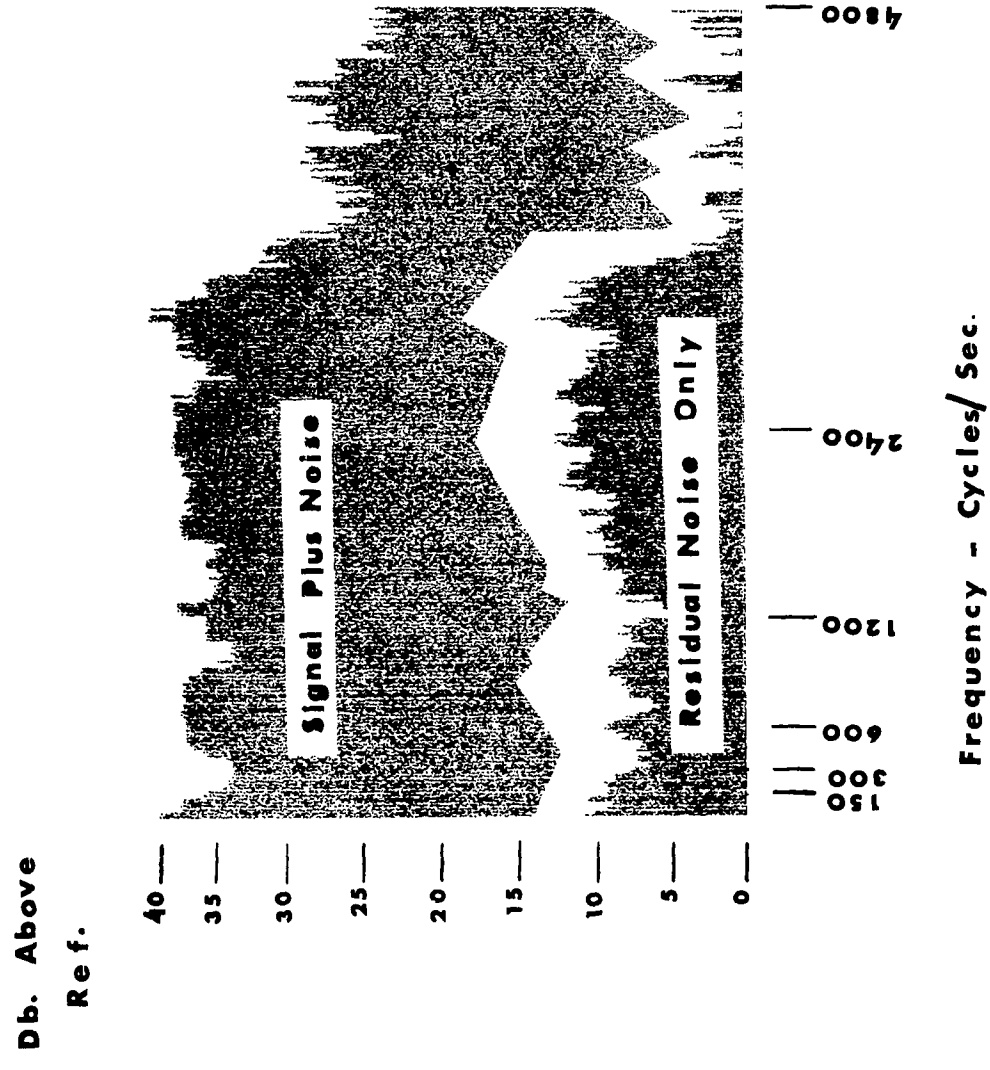


FIGURE 3. TYPICAL SONAGRAPH SECTION, SYSTEMS B AND C

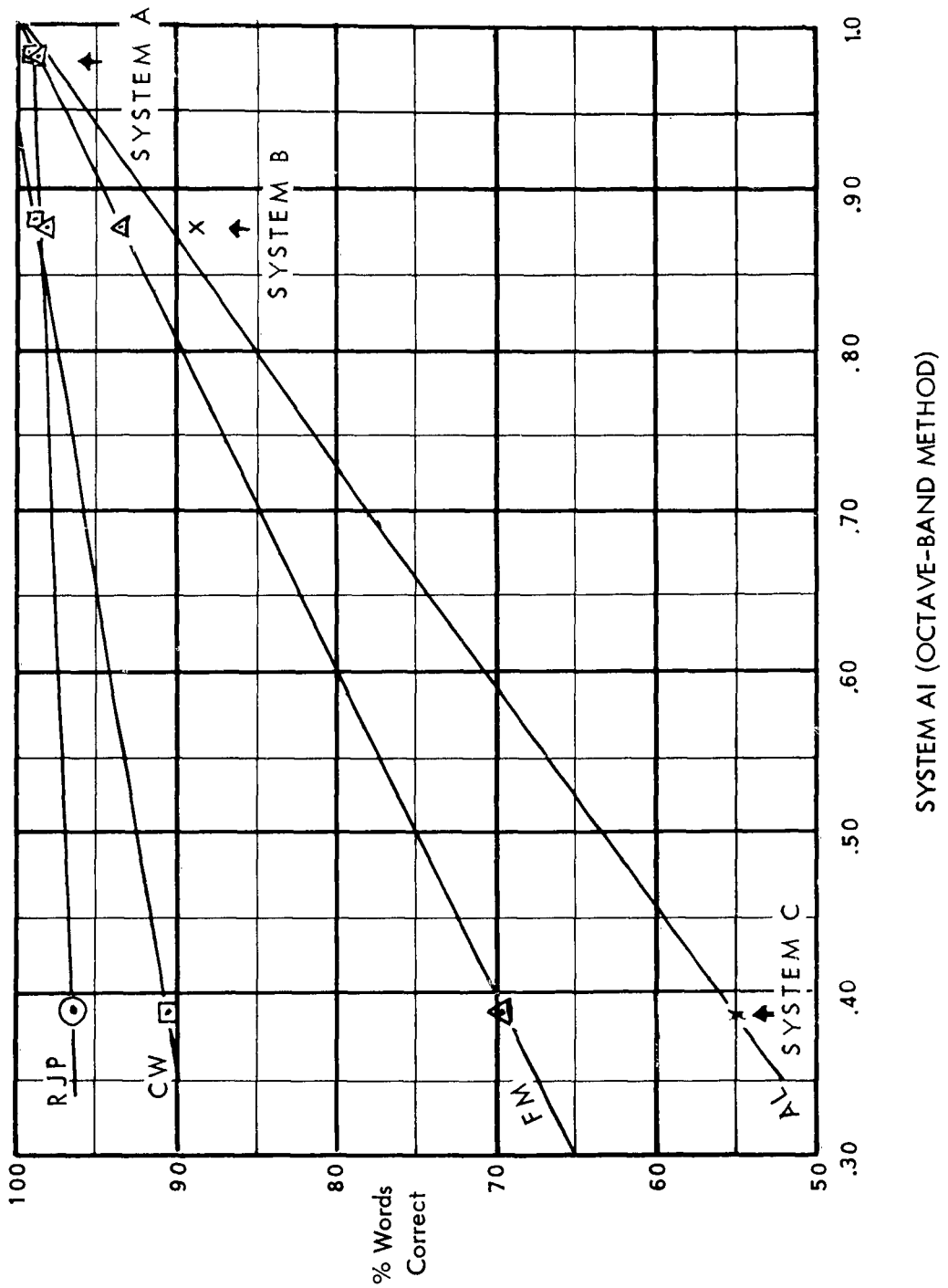


FIGURE 4. Showing how intelligibility of individual talkers is affected by system characteristics (reflected in Articulation Indices). Talkers RJP and CW are male, FM and AI female.