STANDARDIZATION OF SPEECH MATERIALS FOR UNDERWATER RESEARCH 1: COMPARATIVE INTELLIGIBILITY OF MONOSYLLABIC WORD LISTS

Robert F. Coleman, et al

Florida University

Prepared for:

Office of Naval Research

1 September 1972

TECHNICAL REPORT

Office of Naval Research
Engineering Psychology Programs
Grant # N00014-68-A-0173-0003
Project # 196-114

UNDERWATER SPEECH COMMUNICATION

Principal Investigator — Harry Hollien, Ph. D.
Director

Co-principal Investigators — Howard B. Rothman, Ph. D.
Assistant Professor

Stephen H. Feinstein, Ph. D.
Assistant Professor

Report Title: Standardization of Speech Materials for Underwater Research I: Comparative Intelligibility of Monosyllabic Word Lists

Report Number: CSL/ONR Technical Report #25

Date of Report: September 1, 1972

Communication Sciences Laboratory
Department of Speech
University of Florida
Gainesville, Florida
32601

DOCUMENT CONTROL DATA - R & D

1 ORIGINATING ACTIVITY (Corporate author)

Communication Sciences Laboratory
University of Florida
Gainesville, Florida 32601

2a. REPORT SECURITY CLASSIFICATION: UNCLASSIFIED

2b. GROUP:

3 REPORT TITLE

Standardization of Speech Materials for Underwater Research I: Comparative Intelligibility of Monosyllabic Word Lists

4 DESCRIPTIVE NOTES (Type of report and inclusive dates)

Technical Report

5 AUTHOR(S) (First name, middle initial, last name)

Robert F. Coleman
Harry Hollien

6 REPORT DATE	7a. TOTAL NO OF PAGES	7b NO OF REFS
September 1, 1972	~~14~~ 15	5

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
N00014-68-A-0173-0008	CSL/ONR Technical Report #25
8b. PROJECT NO 196-114	
c.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d.	

10 DISTRIBUTION STATEMENT: Approved for public release; distribution unlimited. [redacted] tion

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Engineering Psychology Programs Office of Naval Research Arlington, Va. 22217

13. ABSTRACT

Two experiments are reported in which the underwater intelligibility scores of four standard word lists were compared. As would be expected, closed-set word lists produced higher scores than did open-ended tests. Alternate forms of two tests were compared and the results indicated that lists which are "equated for difficulty" in normal environments also are reasonably equated underwater. A preliminary analysis of phoneme type distortion was derived from one closed-set test. The results appeared to demonstrate that, for this type of experiment, the most common phoneme error was in place of production and that fricatives were most affected.

1

DD FORM 1 NOV 68 1473 (PAGE 1)

S/N 0101-807-6801

UNCLASSIFIED
Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	W
Divers Speech						
Underwater Communication						
Speech Materials						
Phonetic Analysis						
Speech Intelligibility						

STANDARDIZATION OF SPEECH MATERIALS FOR UNDERWATER RESEARCH I: COMPARATIVE INTELLIGIBILITY OF MONOSYLLABIC WORD LISTS

Robert F. Coleman

Harry Hollien

Conducting research on underwater speech intelligibility presents several unique problems; one major difficulty has been the lack of uniformity in speech materials used for testing. For example, speech samples used in past experiments have varied from simple counting procedures to the use of elaborate multiple choice word lists and/or sustained reading passages.

The standardization of speech materials for underwater communication research appears to be important in at least two respects. First, due to the wide differences in the type of tests used, comparisons among such variables as talkers, listeners, communication systems, depth, range, gas mixtures, etc. are inherently difficult and should not be further complicated by problems with the speech materials; secondly, materials which have been "equated for difficulty" in air may not necessarily be so equated underwater. However, such problems may be due (in large part) to the restrictions placed upon the diver's speech mechanism by the underwater milieu. Included among these restricting elements are straps, mouthpiece bits, muzzle mobility, face masks and other essential but cumbersome life support equipment -- as well as such varying factors as regulator back-pressure and gas mixtures. Each of these variables may be expected to affect certain types of phonemes more than others and, therefore, may affect speech materials differently. Hence, standarized , equated speech materials are necessary for underwater investigations of this nature whether the research protocols call for gross intelligibility data or phonemic analysis of the most detailed type.

Despite this need, no data previously have been reported which have compared the intelligibility of specific word lists and sentences spoken underwater. The purpose of this paper is to report results of the first two in a series of experiments designed to evaluate various types of speech materials to be considered for use in underwater intelligibility testing. Specifically, these experiments were designed to compare certain monosyllabic speech discrimination tests used in an underwater environment, and to test the intelligibility of alternate lists which previously had been equated for difficulty in air.

METHOD

PROCEDURE FOR EXPERIMENT I

Test Environment: The procedure reported was carried out at the Underwater Sound Reference Division of NRL, Leesburg, Florida. The test site is a spring with a depth of 175 feet and a diameter of about 400 feet. The temperature of the water is a constant 72° F., and the noise level in the water is negligible. Situated over the deepest portion of

the spring is a large barge, secured by cables from the shore. A well in the center of the barge enables equipment and personnel to be lowered into position for underwater experiments.

For these investigations, a Diver Communication Research System, DICORS, was employed (1). Briefly, this system consists of a frame of plastic tubing arranged so that divers may be comfortably but reliably positioned and calibrated at specific depths. Underwater television monitors and surface communication devices are also provided; DICORS was positioned at a depth of 35 feet for these experiments.

Communication System: For this experiment it was necessary to utilize a diver communication system; the unit employed was the Aquasonics UO-42 coupled to a Nautilus muzzle and double hose regulator. This unit utilizes a 42 KHz. amplitude modulation carrier signal which is fed into the water by a transducing coil. A receiving pick-up coil is coupled to a demodulator and the final output is heard as normal speech. In the experiment, the diver/talkers wore an appropriate unit, the receiving transducer (coupled to the surface unit) placed approximately twenty feet in front and ten feet above the diver. The output from this surface unit was recorded on an Ampex 602 recorder.

Word Lists: The word lists evaluated were two multiple choice discrimination tests: the four-item intelligibility test developed by Black and Haagen (2) and the five-item phonemically balanced rhyme test developed by Clarke (3). The Clarke test consists of monosyllabic words differing by a single phoneme or phoneme cluster in initial, medial and final positions. The Black and Haagen test consists of test-words which are often confused with each other. In addition, a third test -- the Campbell PB_{25} lists (4) was used also. This test is composed of twenty-five monosyllabic words drawn from the PB_{50} lists; they are equated for difficulty in air. The Campbell lists involve an open-ended task rather than a closed or multiple choice set as do the Clarke and Black lists.

Subjects and Recording Procedure: Six males and four females were diver/talkers for this experiment; all were experienced in using underwater communication systems as well as being experienced SCUBA divers. Each diver, after familiarizing himself with the assigned word list, descended to DICORS and positioned himself in front of a TV monitor upon which the test words were presented. The diver read each word - preceded by the phrase "You will say...," as it appeared on the TV monitor. A pause of five seconds was inserted between each word in order to allow for adequate spacing on the recordings. Each diver read the same Clarke and Black lists; for the Campbell lists, each diver read all versions. In actual practice, most divers read two or three words on each breath but each was allowed to use the breathing pattern he or she found most comfortable.

Listening Task: The words on the resulting tape recordings were randomized within a list and were then presented to listeners. No listener heard any talker or list more than twice; all list-talker combinations were heard by at least 11 listeners who were college students paid for their work. Prior to participating in these tasks, volunteers were given a screening test consisting of 1) 50 words from CID Auditory Word

List A-3 (Hirsh recording) recorded in white noise at a S/N ratio of +10dB, 2) 25 words recorded in a helium environment, 3) 25 words spoken over diver communication systems, and 4) 50 words from CID Auditory Word List 4-A. On the latter test, listeners were required to score a minimum of 92% correct. After passing the screening procedure, listeners were presented the lists in random order -- for both lists and talkers. They circled the word they heard for each of the groups of five-words-per-set seen on the Clarke response sheets and four on the Black response sheets. In response to the Campbell lists, listeners wrote the word they heard. The resulting data then were the intelligibility scores (in percent) based on these listener evaluations.

RESULTS AND DISCUSSION FOR EXPERIMENT I

The results of this experiment are summarized in Table I. The Black list attained highest intelligibility scores with an average percent correct of 81.0%, followed by the Clarke list with 73.1%, and the Campbell lists with 57.5% correct. As would be expected, the closed set lists produced scores considerably higher than did the open-ended Campbell lists.

The wide dispersion of scores for the Campbell lists may have been produced by several factors: First, the use of more than one form of the test might have introduced errors due to possible lack of equality between the lists in the demanding underwater environment (for data on this issue, see experiment 2); second, talker differences may have a more profound effect on open-ended listening tasks, wherein many of the cues intrinsic to the test are not furnished.

Table II presents the rank of each diver/talker for each test from highest to lowest for both individual and overall ranks. In general, the civilian male Experimental Phoneticians scored higher than did either military male or female civilian divers; however, this trend was not consistent over all three tests. For example, diver 3.1 (a female civilian) was ranked near the group mean for both the Black and Clarke scores, yet scored highest on the Campbell lists. The three lowest overall ranked divers, however, were female civilians. Their low scores reflect either a lack of experience in producing intelligible speech material underwater in comparison to the male divers - or design inadequacies of the mask and communication system with respect to female facial structure and/or speech acoustics. Obviously, investigations of diver/talker performance are needed as, for example, the relationships between the speech productions of males and females for air may not hold in a fluid environment.

PROCEDURE FOR EXPERIMENT II

One of the desirable features for standardized speech materials is that the word or sentence lists have alternate forms which are equated for difficulty. With such alternates, the test is much more viable as, for example, the number of listeners required for a valid evaluation may be significantly reduced -- a critical factor when the research protocols specify the use of underwater listeners. In this regard, several tests are available with alternate forms which are either "phonemically balanced" and/or "equated for difficulty" in normal atmospheres. Such balance among the alternate forms, however, may not be maintained in an underwater speaking

environment.

The purpose of this experiment, therefore, was to investigate the practicality of using alternate forms of two intelligibility tests, the Campbell PB_{25} lists (4) and the Griffiths' Rhyming Minimal Contrasts test (5) for underwater intelligibility studies.

Test Site and Recording Procedure: The location for this experiment was Rainbow Springs, near Dunnellon, Florida. The test site was within a large basin adjacent to the headwater springs; water depth was approximately fifteen feet, the temperature a constant 72°. Despite a substantial hourly flow from the springs, the ambient noise level is near sea state zero. As in Experiment I, DICORS was employed in the study. In this instance, it was placed on the bottom of the lagoon; diver depth, hence, was approximately twelve feet. The transducing equipment (i.e., Aquasonics UO-42 with Nautilus muzzle and double-hose regulator) and recording procedures otherwise were identical to those employed in Experiment I.

Subjects: Diver/talkers for this experiment were four male Experimental Phoneticians and two female divers; all personnel were experienced in reading speech intelligibility materials underwater and were accomplished SCUBA divers.

Word List Presentation: The word lists compared by this experiment were the eight alternate forms of the Campbell PB_{25} lists and five alternate forms of the 50-word Griffiths' Rhyming Minimal Contrasts test. Each diver read every list, preceding each test word by the phrase "You will say..." The lists were quasi-randomized with respect to order of alternate forms -- and between the Campbell and Griffiths lists.

Listening Task: Tape recordings of the six divers reading all forms of the Griffiths and Campbell lists were randomized and presented to groups of eleven listeners, with the restriction that no listener heard the same talker more than twice. A total of 73 listeners (paid college students) were used; all had previously passed a screening test described in Experiment I (but had not been used in that experiment).

RESULTS AND DISCUSSION FOR EXPERIMENT II

The number of words correctly identified for each list was determined and expressed as a percentage. A summary of these results for the Campbell lists is presented in Table III. With respect to the Campbell data, the wide range of scores obtained by different talkers on each list (see Column 4: "Range") makes comment on the "equality" of these lists hazardous. However, within the limitations of this procedure and in view of the wide standard deviation for each of the alternative forms, there is no reason to suspect that consistently higher scores would be expected for any particular one of the alternative lists. Figure I presents the data for the Campbell lists in graphic form and further demonstrates both talker variability and that the rank for each diver/talker on the different lists changed frequently (although generally the male divers scored higher than did the females).

The results of the Griffiths tests are presented in Table IV. The

similarity of the mean list scores for all divers, coupled with small standard deviations, lead to the conclusion that the alternate forms of this test may be considered to be "equated for difficulty" underwater as well as in air. The differences in the variances of the alternate forms of the Campbell and Griffiths lists may arise in part from the open versus closed response sets: the Campbell list responses would be expected to vary more due to listener "guesses" on such an open-ended task.

The experiment involving the Griffiths test permitted an error analysis to be carried out on listener responses; a confusion matrix was constructed listing the cumulative errors for each test word in the five lists. For example, when the test word in item 1 of the Griffiths test was "bat," 68 listeners responded correctly while 4 marked "batch," 1 marked "bass," and none marked "bash" or "badge." When "bash" was the test word, however, only 36 listeners responded correctly while 22 marked "bass," 8 marked "bat," 4 marked "batch," and 3 "badge." Chi-square tests of significance were conducted on this matrix under two assumptions: (1) that listeners were responding randomly (20% probability for each word) and (2) that the test word would be expected to be chosen the same number of times as the overall mean for the test list used. Sixty-three of the 250 test corpus were significantly different (.01 level of confidence) from the scores expected under the random response hypothesis, while 29 received significantly fewer correct responses than might be expected under the more rigorous assumption that listeners would respond to each word set as the average of the list.

Finally, the 29 words which were significantly less intelligible than the other 221 words were plotted with respect to phoneme distortion on an articulation chart presented in Figure 2. The phonemes are arranged in a manner-place display, and arrows from the various phonemes denote substitutions for that phoneme. For example, the phoneme /n/ was often heard as /b/ or /d/, and /h/ was heard as /k/, and so on. Examination of Figure 2 will show that, for this experiment, the most obvious relationship is for listeners to make "place" errors, and that phoneme confusions were most common in the "fricative" class. Moreover, there seemed to be a moderate trend for nasals to shift toward plosives, an event which would not be unexpected in view of the fact that a diver's mask separates the nasal port from the oral cavity and possibly functions as a closed resonator blocking the release of air from the nasal cavities. A more extensive analysis of phoneme substitution patterns currently is being carried out, but this preliminary evidence appears to indicate that a major effect of attempting to speak in the underwater environment in SCUBA rig is the distortion of "place" cues in intelligibility test materials.

SUMMARY and CONCLUSIONS

The results of the two experiments described above demonstrate the need for standardization of speech materials for underwater intelligibility research. The wide dispersion of scores obtained from the different types of tests and the variability of the open-ended alternate lists indicate the need for close attention both to the choice of test and consideration of all alternate forms to be used as well. That is, the choice of a test vehicle for underwater speech research involves all of the standard considerations for its use in air plus special ones related to the uniqueness of

the fluid environment.

While the wide dispersion in scores among the Campbell lists is discouraging, it is interesting to note that the overall scores for these lists were quite similar in both the first and second experiments; that is, for the first experiment, the range of scores obtained was from 34 to 80 percent correct with a mean of 57 percent while for the second experiment these values were 32 to 80 percent and 46 percent respectively. Such a relationship is encouraging since it is to be remembered that listeners, some diver/talkers, water depth, and location changed between the two projects.

Finally, it appears that the closed-set type of intelligibility test may furnish a practical tool for making informed judgements on the type of phoneme distortions which may be introduced by the underwater environment, transmitting equipment, and/or divers themselves. From the results of the second experiment, a tentative conclusion may be drawn that a major effect of present underwater speaking environments is the distortion of "place" cues for individual phonemes.

References

1. Hollien, H. and Thompson, C.L., A Diver Communication Research System (DICORS), CSL/ONR Progress Report No. 2, Office of Naval Research, Physiological Psychology Branch, Grant NONR 580 (20), January 15, 1967, 1-8, (AD 648-935).

2. Black, J.W. and Haagen, C.H., Multiple-Choice Intelligibility Tests, Forms A and B, Journal of Speech and Hearing Disorders, 28, 1963, 77-86.

3. Clarke, F.R., Technique for Evaluation of Speech Systems, Stanford Research Institute's Final Report 5090, U.S. Army Electronics Laboratory, Contract DA 28-043 AMC-00227 (E), August, 1965, 1-65, (AD-473 995).

4. Campbell, R.A., Discrimination Test Word Difficulty, Journal of Speech and Hearing Research, 8, 1965, 13-22.

5. Griffiths, John D., Rhyming Minimal Contrasts: A Simplified Diagnostic Articulation Test, Journal of Acoustical Society of America, 42, 1967, 236-241.

Table I. Summary of results of intelligibility tests using the Black, Clarke, and Campbell lists. Scores are expressed as a percentage of possible words correct.

List	Black	Clarke	Campbell
Range	70.4-88.8	63.2-81.2	34.0-80.8
Mean	81.0	73.1	57.5
Median	82.9	73.2	60.4
Standard Deviation	6.8	6.3	15.0

Table 11. Rank for talkers reading each list (Black, Clarke, Campbell). Talkers arranged in order of highest overall rank to lowest. Class specifications are: (1) Male Civilian, experimental phoneticians; (2) Male Navy Divers; (3) Female Civilian Divers.

Talker/Diver	Black	Clarke	Campbell	Mean
1.1	3	1	4	2.7
1.2	1	2	5	2.7
2.1	2	3	8	4.3
2.2	4	4	6	4.7
1.3	5	6.5	3	4.8
3.1	6	9	1	5.3
1.4	8	6.5	2	5.5
3.2	7	8	7	7.7
3.3	10	10	11	10.3
3.4	11	11	10	10.7

Table III. Summary of scores obtained for alternate forms of Campbell PB_{25} lists. Scores are based on six diver/talkers and a total of 73 listeners.

List	Mean Score %	Standard Deviation	Range
M	50.8	13.09	32.8-68.4
N	42.6	6.64	33.6-50.3
O	39.7	12.77	23.7-57.7
P	56.6	14.65	32.4-80.7
Q	37.4	13.74	15.3-58.0
R	45.7	15.18	20.0-68.8
S	49.4	17.56	24.0-72.4
T	45.6	11.89	28.3-61.3

Table IV. Summary of scores obtained for alternate forms of Griffiths test of rhyming minimal contrasts. Scores are based on six diver/talkers and a total of 73 listeners.

List	Mean Score %	Standard Deviation	Range
A	71.2	6.4	60.2-77.2
B	68.1	4.2	63.8-75.0
C	68.0	5.5	61.4-76.9
D	73.4	3.2	69.8-79.0
E	69.1	5.0	62.2-75.2

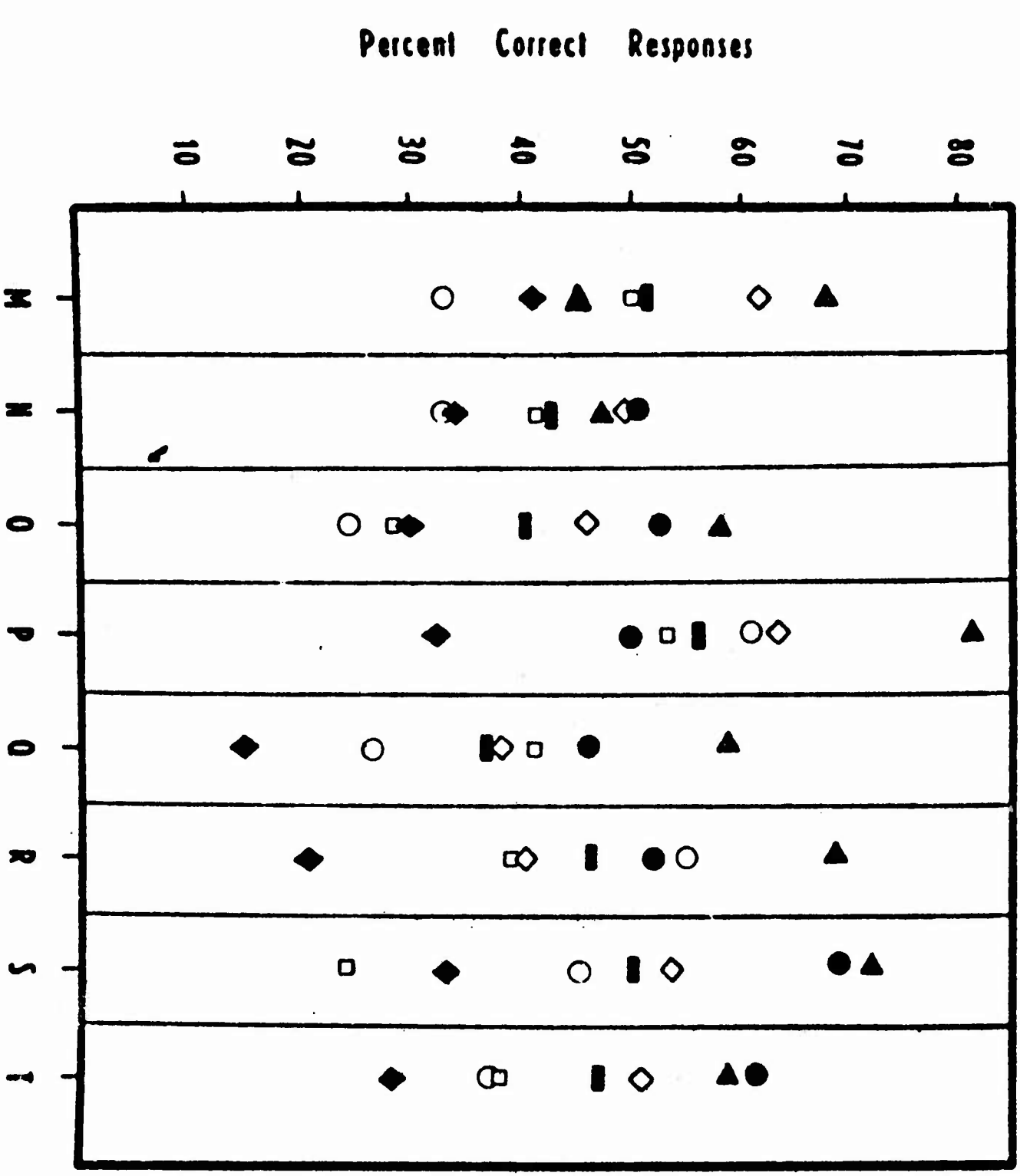

Figure 1: Scores obtained by six diver/talkers reading the eight alternate Campbell PB_{25} lists. Each symbol in a column represents a single diver's score; solid bar represents median score obtained for six divers and a total of 73 listeners.

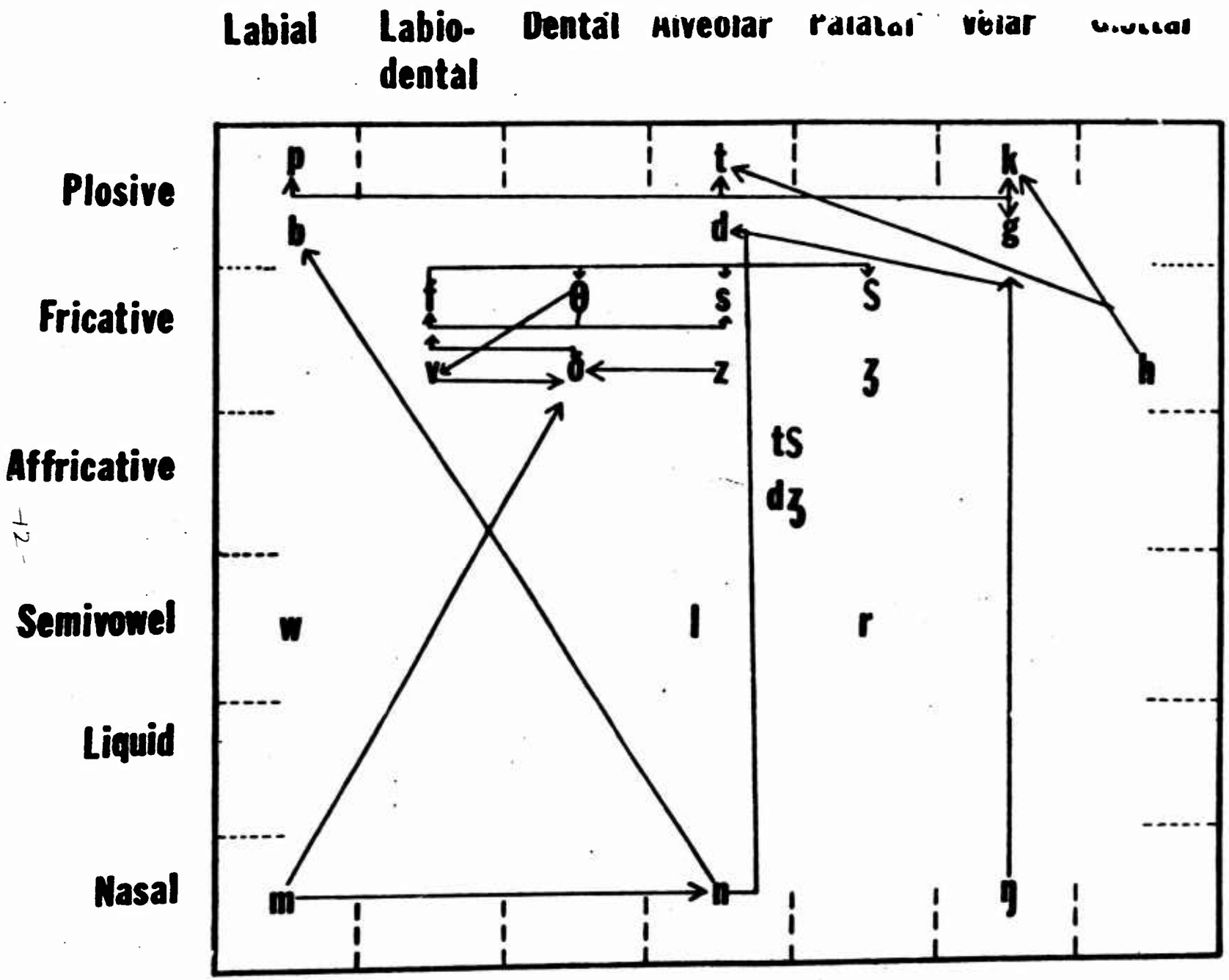

Figure 2: Phoneme error analysis of Griffiths lists. Arrows point to phonemes substituted for intended phoneme.