FINAL REPORT ON THE PRONUNCIATION TEST SYSTEM

Julie B. Lovins

1. Introduction

This report concerns the final series of experiments with the main program in a system for computer-assisted pronunciation-hearing tests that has been developed and studied since 1968.[1] Some general conclusions about the system will be presented, with particular reference to the data obtained in this latest experiment.[2] The results, comprising extensive information on the English pronunciation of seven female junior high school students (S1-S7), will also be compared with those reported on previously[3] in relation to the Level Improvement Program and Component Additivity Test Program.

2. Error-evaluation capability of the test program

It is possible to get some sense of how successfully the test program operates in judging[4] the performance of a student on individual segments or clusters, given only a binary judgment on a whole word by the native-speaker evaluator, by comparing students' s/c scores on WF's with the evaluator's off-line impressions of pronunciation problems. This comparison will now be made for several of the errors observed (by the evaluator) to be most prevalent among S1-S7.[5]

---

1) For an explanation of the design of the system, and discussion of previous results, see especially the papers listed in References. Each of these also lists other reports on the project.

2) This experiment was conducted in the summer of 1973 by Miss Fumiko Mitsuya, whose skill and perseverance are solely responsible for its completion. The author gratefully acknowledges also the helpful comments made by Dr. Osamu Fujimura on the results, and the supplementary information kindly provided by Miss Yoshiko Umezu, the students' English teacher at the time of the experiment.

3) Cf. Mitsuya (1973).

4) The algorithm used in this procedure is described in Harada (1971).

5) During the test sessions, the evaluator attempted to judge students' performance purely as a native English speaker. Off-line comments were however based also on linguistic training and experience in teaching English in Japan.

A more complete listing of such pronunciation problems would include the following:

lowering of [I ɛ] to [ɛ æ] respectively
unrounding of /ow uw/
various distortions of /oy/ (sometimes lowered to /ay/)
/z-/ pronounced as /dz-/ and /-dz/ as /-z/
confusion of /b/ and /v/
difficulty in pronouncing /θ ð/, or confusion with other fricatives
difficulty with final clusters (especially omission of /s/ after /-(C)t/)
trouble in perception of /p/ and /b/
pronunciation of /aɹ/ tending toward /ɹ/
difficulty in pronouncing /Cw-/

Table 1. s/c scores for selected errors prevalent among S1-S7.

WF	Level assignments[6]			Usual error	s/c scores for students						
	I	II	III		S1	S2	S3	S4	S5	S6	S7
I	5	4	5	I → ɛ	-1	-1	.25	-1	-1	.58	-1
OJ	2	3	3	(varies)	.80[7]	.77	.61	.45	1	.77	.65
Z-	5	5	5	z- → dz-	-1	-1	1	-1	-1	.20	.40
-DZ	4	5	5	-dz → -z	.04	-1	-1	.33	1[8]	-1	.50
B-	1	3	3	(varies)	.68	.76	.75	.61	.83	.91	.88
BL-	1	4	4	(varies)	.38	.33	.20	.45	.64	.82	.82
-B	3	3	3	-b → -v	-1	.33	.33	1	1	.40	.67

In Table 1, the scores for those items in which significant difficulty was noted off-line by the evaluator are underlined. The non-underlined scores are given to provide some idea of the results for items not judged to give particular trouble to that student. It can be seen that the underlined numbers do tend to be lower than the others, which implies that an 'N' response by the evaluator did lead to an appropriately low evaluation by the program of the student's performance on the WF the 'N'

6) In this report three different level assignments are discussed. The first two are mentioned in Mitsuya (1973); they are based on test results for the EJJ and NEJJ students. III-assignments resulted from the final experiments, with S1-S7.
7) The evaluator noted off-line that OJ /oy/ sounded somewhat like AJ /ay/, but did not in general give an 'N' response to this pronunciation, meaning it was within an acceptable range of variation.
8) This student was noted to have trouble with -DZ, but the printout shows that 'N' responses were actually given only to some -RDZ words (tabulated separately).

response was based on; and conversely, that less difficult WF's (for a given student) were not disproportionately downgraded by the program.

Although this general statement is borne out by additional examination of the data (not presented here), it must be interpreted with some caution. The most obvious problem is the unavoidably large number of cases, throughout the experiment, in which a WF is presented a very small number of times (say one to three). [9] One may well question the reliability of conclusions based on this small a sample, particularly in view of the 'context' problem to be discussed below. As a matter of fact, the (correct) important observation, by the program, that the students in the latest experiment (III) almost all had trouble with I [i], may be regarded as fortuitous, since in five of the seven cases the program presented just two instances of I, which the students failed on; and on the basis of this sample 'gave up on' this WF for the rest of the experiment. In the other two cases, additional words were presented after the student had correctly pronounced at least one of the first two examples. This rather striking example suggests that we might not have been so lucky in cases in which the first few words presented with a given WF happened to be 'bad' choices for one reason or another--e. g. words completely unfamiliar to the student, or containing a sequence of WF's particularly difficult for reasons not taken into account in constructing the program. [10] (Or for that matter, a WF might be given a grade of 100%, on the basis of just one or two occurrences, although it might cause some difficulty in different contexts.) What comes to mind first in this connection is the prevalence of E [ɛ] and IJ [i:] in the tests of initial consonantal WF's, since these two vowels were assigned to Level 1 at the I-stage of the experiments. There are several prominent processes in Japanese phonology connected with specific vowels, expecially high vowels. While a comparatively low score for the consonants affected by the processes, relative to others, would correctly reflect special pronunciation problems in English for Japanese speakers, such a low score would also represent an unintentional bias in the test if it were based on the occurrence of these consonants solely in the environment in which they were most difficult. That this might happen was shown in relation to I above; and indeed something of the sort may have occurred for S1 and S2 with Y-, which in Japanese phonology disappears before both front vowels. These two students failed on both YELP and YES and were then given no more Y- words. The other students got an 'OK' on at least one of these first two words, and were then allowed to

---

9) As described in Harada (1971, q. v.), the program was designed to shorten the test session by avoiding presenting a word once it has been demonstrated to be 'too difficult or too easy' for the student in terms of the WF's it contains. This is done by setting lower and upper limits on the 'state' values for WF's, at the beginning of each session. Since the limits were -2 and +2 in the latest experiment, two consecutive identical responses frequently determined the final judgment on that WF.

10) At an earlier stage of the project a plan was made for studying interactions between sequential WF's, but this option was not implemented in the version of the program reported on here. (Cf. Smith et al., 1970.)

proceed to others (containing various other vowels), on which they scored generally very high.

To return to the larger question of whether the prevalence of small samples (small c's) significantly decreases the reliability of the program, let us look at the scores for WF's having c=1 or 2. Table 2 shows the frequencies of various s/c scores for three of the seven students (ranking first, third, and fifth in overall performance as measured by total average s/c). These scores are divided into three categories which in effect cover the following possibilities: s/c = -1 or 0; s/c = .50; and s/c = 1. There are few scores in the second category, and to include them in the first group would not change the result. The table gives the number of WF's for which each (category of) score occurred with c = 1 or 2; and some additional information about each student's test.

Table 2. Frequencies of various s/c scores, for three students.

Student rank	#1	#3	#5
Average of s/c scores	.86	.64	.51
s/c < .50	6	20	30
s/c = .50	2	10	3
s/c = 1.	137	90	86
Total WF's tested	236	234	236
Total word count	453	513	564
Proportion of 'OK' words	.92	.75	.68

We can conclude that these frequency figures, for high and low s/c scores, accurately reflect the overall s/c performance of individual students--that small-c scores are not distributed very differently from large-c ones. However, one should still note the considerable role the small-c scores play in creating the averages they mimic: in the extreme case above, that in which only 453 words were presented to the student, 145/236 = 60% of the WF's tested were presented only once or twice (16 once, 129 twice).

Whatever validity small-c scores may have in gauging a student's overall performance--and just how much it is, is not really clear--it is quite easy to show that when used in the calculation of level assignments for individual WF's, they can be the cause of intuitively strange results. (This matter will be returned to below, in connection with the Level Improvement Program.) Consider the most recent experiment, in which seven students were tested on each WF. For a large number of these WF's s/c was either -1 or 1 (by far the most frequent results when c=1 or 2). Suppose that for various WF's, the scores ran like this:

Table 3. Possible results of $|s/c| = 1$ for all students.

WF_a	$s/c = 1$ for 7 students	average=1 level = 1
WF_b	$s/c = 1$ for 6 students, -1 for 1 student	average=.71 level = 3
WF_c	$s/c = 1$ for 5 students, -1 for 2 students	average=.43 level = 5
WF_d	$s/c = 1$ for 4 students, -1 for 3 students (etc.)	average=.14 level = 5

Then the difference between WF_a and WF_b, and WF_b and WF_c, is considered to be two levels, based on the performance of one student in each case, possibly in just one test word.

To make this example less theoretical, we can observe that even in the case of the third (the 5th-ranking) student mentioned above, 66 difficult-looking WF's originally assigned to levels 3-5 were given $s/c = 1$ on the basis of one or two test items. The full listing for $c = 1$ or 2, for this student, is as follows:

Table 4. WF's on which one student attained various s/c scores, for $c = 1$ or 2, compared with level of WF's at stage I.

	Level 1	Level 2	Level 3	Level 4	Level 5
$s/c < .50$		Y- -RV -LT -RK -LFTH -RLD	PL- -LTH SL- -LSJT -B -NS -SJ -NGKTH -RD	SKL- -TJT -LSJ -DTH	I -THS -LTHS Z- -DHZ -NTHS -ZJ -KSTS -RTHS -MD -MPTS
$s/c = .50$				-LKT -LFS	THW-
$s/c = 1$	BY- FY- -LB	HY- -NTH -G -NST -NG -LDJ -SP -LST -MP -LPS -KS -LVZ -DHD -MPS -DJD -NTJT -NGK	PY- -NGZ -V -NTJ -VZ -RSJ -FT -RND -FS -RDJ -BD -LMZ -ZD -LDZ -NZ -SPS -LK -SPT -LS -MPT -RM -RTJT -NGKS	SF- -PTS -RMZ DW- -DST -RLZ VY- -SKS -LPT -TH -RTJ -LBD -VD -RPT -LVD -RF -RBD -LBZ -RS -RMD -NGKT -MZ -RKS -LDJD -RZ -RPS -RDJD -RL -RBZ	-RN -PTH -RB -TTH -RTS -LMD -RST -RMTH -RFT -RKT -RVD -RFS -RNZ -RVZ

Clearly many more WF's are assigned questionably high s/c scores than questionably low ones, on the basis of the small-count scores alone. As an example of just two small-c scores out of seven disproportionately influencing the seven-student s/c average, however, we have the results of the Y-problem mentioned above. The individual s/c scores were computed as

$$\frac{-2}{2} = -1, \quad \frac{-2}{2} = -1, \quad \frac{6}{10} = .60, \quad \frac{7}{8} = .88, \quad \frac{6}{8} = .75, \quad \frac{8}{9} = .89, \quad \frac{8}{9} = .89$$

yielding an average of .29, which is clearly not representative of the performance of the group as a whole.

It is important to emphasize that any biases that small-c scores may have introduced into the data gathered in these pronunciation tests are due to well-defined and easily correctable programming problems or to a particular choice of test parameters, as in the low settings for 'state' threshold values. For example, it would not be difficult to 'weight' scores somewhat in terms of their likelihood of ultimate validity, based on the size of c; to take into account the median as well as the mean when computing averages; and so on. It would also be very easy to decrease the number of small-c scores, even if (for the sake of not appreciably lengthening the test time) it were necessary to omit some WF's in the process. With several such minor alterations in the program, the test results would undoubtedly be as sound as the general conception of the test program.

3. Level Improvement Program

An attempt was made to determine the effectiveness of the Level Improvement Program, comparing the results of its application to three different sets of data (see footnote 6). There are two main questions to be answered: are the level assignments made by the program reasonable, in terms of actual student performance; and is there any obvious trend in differences between assignments made for the respective three groups?

The apparent tendency toward randomness in level reassignment noted in Mitsuya (1973) for the stage II data was borne out by the results in III. Especially notable is the preponderance in both II and III of WF's reassigned to Level 1 from low-level assignments in I: and the assignments in I seem to have a firmer experimental and linguistic basis. The most probable cause of this misleading result has already been mentioned: the small-count problem.

Assuming that the I level assignments were 'better', however, we still need an overall measure of the discrepancies between I, II, and III level assignments. To obtain some rough approximation to this new measure, a tabulation was made of which two (of the three) level assignments were closer to each other, for each WF, along with the minimal difference between one of these two (often identical) values, and the third. (Cases in which the three level assignments were equal, or no two of the three were closer, were also tabulated.) This was done separately for initial clusters final clusters beginning with -L-, -R-, or a non-liquid respectively; and vowels. For instance, the WF DR- was assigned to Levels 4, 3, and 4. Then (assignment) I was closer to (actually equal to) III, and the minimum difference between the various level assignments was 4-3=1. The reason

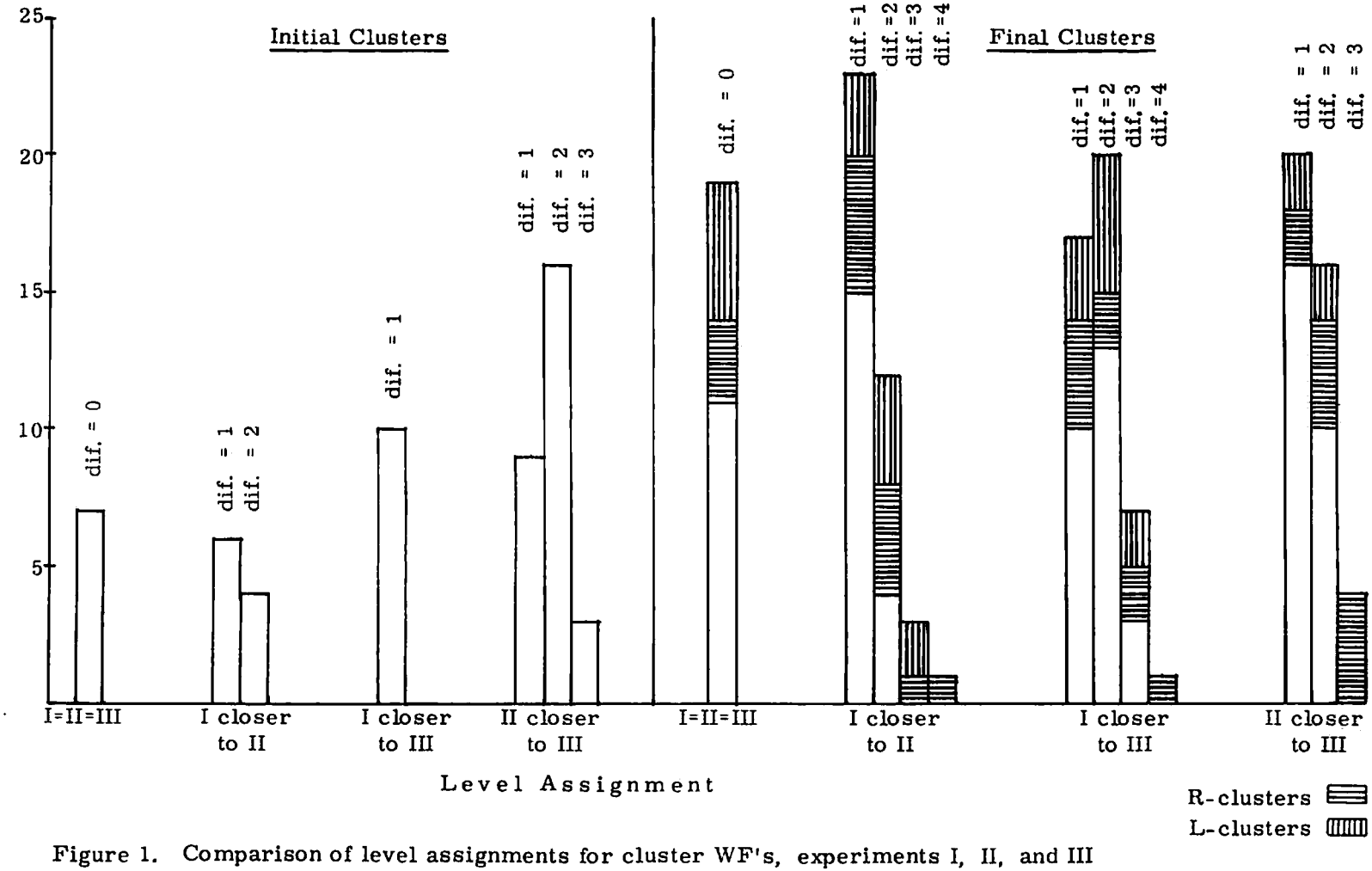

Figure 1. Comparison of level assignments for cluster WF's, experiments I, II, and III

for noting this minimal difference is that a difference of 1 in level assignments is much less likely to be significant than a difference of 2 or more, given the arbitrary partitioning of percentages into the five 'level' categories.

The results [11] of this tabulation are shown in Figure 1. Any two of I, II, or III (level assignments) may be considered to be more consistent with each other when the total length of the bars, for a given comparison, is greatest. Thus it appears quite clearly that for initial clusters, II and III are appreciably closer to each other than either is to I. The results for final clusters are not nearly so clear: overall, III appears closest to I, but not by a wide margin. For -L- final clusters, neither II nor III seems appreciably closer to I; for -R- ones, it appears that II and III are again closest, if we take into account the size of the 'higher' differential (the bar furthest to the right).

It is difficult to draw any firm conclusions from these results, other than the negative one that if I assignments are 'better', the program did not do as well as we might have hoped in the two subsequent experiments, judging from the lack of any definite overall correlation between I and either II or III or both. It is quite possible, though, that much better results would be obtained if the small-count problem could be eliminated.[12]

4. Component Additivity Test Program

This program yielded results comparable to those described in Mitsuya (1973) for experiment II: the values of W-A and W-M [13] were positive more frequently than negative, implying that students found clusters easier to pronounce than any of their components. Table 5 [14] below is analogous to Mitsuya's Table 4 (1973, 103):

---

11) The data on vowels are omitted for reasons related to the content of footnote 12.

12) Quite aside from this, the following caveat must be added: after these calculations were made, it was discovered that a number of level assignments in III (at least) may have been made incorrectly because of an apparent (untraceable) error somewhere in the computer system. While this probably does not affect the general results drastically, since the direction of the errors varied randomly, it does make it necessary to allow a fairly wide margin of error when considering the figures on level assignment.

13) As noted in Mitsuya (1973, 102), W is the s/c score of a cluster; A the average of the s/c scores for the components of that cluster; M the minimal s/c score among those for the components (all in percent). Then when W-A is negative, the average difficulty of the components of a cluster is greater than that of the cluster; when it is positive, the cluster is easier than the average component; and so on.

14) S3 is not included in the table because sufficient data were not available

Table 5. Frequencies of various values of W-A and W-M for 'cluster' WF's.

	W - A			W - M		
	-	0	+	-	0	+
S1	80	1	104	65	1	119
S2	86	4	97	74	5	108
S4	89	7	89	81	6	98
S5	37	33	115	37	33	115
S6	56	4	113	46	5	122
S7	83	2	100	64	2	119

The most obvious phonetic explanation for such data is the feasibility of 'cluster reduction' in English pronunciation. No serious attempt to justify this hypothesis in terms of the content of the clusters tested was made, however, because the data are not dependable enough for this special type of analysis (see footnote 12). (The small-count problem must also be considered: it is relevant to the s/c scores listed for the clusters, but less so in relation to s/c scores given for the WF components (single consonants), since these were generally presented with much higher frequency).

5. Other results

One factor in these experiments that is difficult to know how to evaluate is the relative performance of students on words they are familiar with and those they have never heard before. The latter are in effect 'nonsense' words; and, when testing junior high school students, there are a large number of them to be considered. [15] Some of the students found no major difficulty in pronouncing unfamiliar words; but others had perception problems clearly aggravated by this factor, despite the use of a visual display [16] of the test words. It is well known that the results of trying to repeat unfamiliar words (particularly in a foreign language) are considerably more varied than when the words have been encountered previously.

For example, several students pronounced 'boss' as 'woss'; /b-/ was also confused with other voiced bilabials. The aspiration of /p-/ was presumably responsible for some renderings of this phoneme as /f-/ or

15) The students tested in III were near the beginning of their second year in junior high school, i. e. most had been studying English a little over one year. Also, many of the words on the test list are quite infrequent in standard usage.

16) See Harada (1971). The same test procedures were followed in stages I, II, and III of the experiments, although different recordings of the data were used in each case. The only notable difference between II and III was that a male speaker of American English recorded the test words for II and a female speaker did so for III, because of possible acoustic phonetic problems due to the students' all being female.

/h-/. Final /-p/ was sometimes pronounced as a fricative, too. The universal tendency to confuse /θð/ with /f v/ respectively also turned up. It is quite probable that such errors would have been less frequent within a more limited vocabulary. It would of course be contrary to the spirit of these experiments--and of the test per se--to limit the vocabulary severely; the point is that if we want to correct the errors, we should consider their varying origins.

Another interesting aspect of the data in the final experiment was the appearance of apparent hypercorrection phenomena. This may well be the reason for the frequent pronunciation of /-b/ as /-v/. One student pronounced 'sharp' as 'sharped', perhaps over-conscious of the difficulty of perceiving final unreleased stops. As in the case of the other matter just mentioned (unfamiliar vocabulary), one conclusion that can be drawn from this is that the reliability of results derived from these experiments is then all the more likely to be proportional to the amount of data available on each WF being tested.

References

Harada, K. I. "Experiments on the T-T system," Annual Bulletin (Research Institute of Logopedics and Phoniatrics, University of Tokyo) No. 5, 51-68 (1971).

Mitsuya, F. "Status report on the pronunciation test system," Annual Bulletin (Research Institute of Logopedics and Phoniatrics, University of Tokyo) No. 7, 93-111 (1973).

Smith, D., K. Ito, C. Sato, H. Ishida and O. Fujimura. "The Pronunciation-Hearing Test Using a Hybrid Magnetic Tape System," Annual Bulletin (Research Institute of Logopedics and Phoniatrics, University of Tokyo) No. 4, 111-113 (1970).