AN AUDITORY SPEAKER-RECOGNITION EXPERIMENT

Marion Shirt

Department of Linguistics & Phonetics, Leeds University

This is an account of a pilot study conducted as an attempt to
answer a motion passed at the Colloquium of British Academic
Phoneticians in 1980 'that phoneticians should not consider them-
selves expert in speaker identification until they have demonstra-
ted themselves to be so'. This in turn was a response to the
increasing disquiet about phoneticians acting as expert witnesses
in court cases involving speaker identification. The study was
supported by a research grant from Leeds University.

Experimental set-up
A tape was compiled consisting of spontaneous speech samples,
courtesy of the Home Office Police Scientific Development Branch.
They kindly made available to us a number (37) of tapes of 74
different speakers. From these tapes 4-second samples were
selected to use in three different kinds of listening test:-

1) Section A - 6 closed Identification Tests
    In each test there were seven speech samples and the listeners
    were asked to match the seventh voice with one of the
    preceding six.

2) Section B - 2 closed and 2 open Pairing Tests
    In each test there were ten speech samples and the task was
    to pair matching voices. Test 1 had 5 matching pairs, Test 2
    had 2 matching pairs. In Tests 3 and 4 the listeners were not
    told how many (if any) matching pairs there were.

3) Section C - 3 Discrimination Tests using material from a
    different source. The idea was to use two speakers from the
    same family who might be expected to have similar voices. The
    speakers read a prepared text. Listeners were asked to indi-
    cate whether the two speech samples in each test were the same
    speaker or different. Test 1 was a direct recording, Test 2
    band-limited and Test 3 used whispered speech. Judging by
    results this section was (too?) easy, though some listeners
    were fooled by the whisper.

The other sections were considered by phoneticians and non-
phoneticians alike to be very difficult, especially the pairing
tests.

Short samples of speech were deliberately chosen primarily to
avoid having too long a tape and thereby discouraging listeners
from the start. But the brevity of the samples also served

AN AUDITORY SPEAKER-RECOGNITION EXPERIMENT

another purpose; that of preventing subjects from relying too heavily on segmental and linguistic information at the expense of the actual voices. All tests but one used groups of speakers with similar accents as judged by the author. Some information on the speakers' linguistic backgrounds was provided with the tapes.

The only test using speakers with different accents was solved by 90% of the listeners (Phoneticians 95%; Non-phoneticians 85%).

The subjects
Test-tapes were sent out to 40 volunteer phoneticians in July 1981. Subjects were told they could listen to the material as often as they liked using any listening technique they wished. Space was provided on the answer sheet to give details of both. 26 replies were received but only 20 were used. The least accurate answers were discarded. There appears to be no correlation between time spent on the test and accuracy. Those people who re-recorded the material and then played the two tapes side by side on similar machines were among the highest scorers.

It was decided, therefore, to use this technique for the phonetically naive subjects. Twenty non-phoneticians, 15 undergraduates and 5 older subjects originating from various parts of Britain listened to the tapes during the Summer Term 1982, in controlled conditions in the Department of Linguistics and Phonetics, Leeds. The average time spent on the tapes by these listeners was $2\frac{1}{2}$ hours with a short rest period. Phoneticians' time varied between 14 hours and 1 hour.

Results
Results suggest that phoneticians perform only marginally better than non-phoneticians.

Overall % accurate matching (All Sections)

P(honeticians)	53	range (76 - 38)
N(on)P(honeticians)	46	range (76 - 19)

Overall % accurate matching (excluding Section C)

P	46	range (72 - 33)
NP	40	range (72 - 6)

AN AUDITORY SPEAKER-RECOGNITION EXPERIMENT

Spread of scores over the two groups

% accuracy	P	NP	All
76	1	1	2
71	2	-	2
67	1	-	1
62	2	3	5
57	1	3	4
52	3	3	6
48	4	1	5
43	3	1	4
38	3	1	4
33	-	4	4
30	-	1	1
24	-	1	1
19	-	1	1

% accuracy for each test-type

Section A	P	70	(83 - 50)
	NP	50	(83 - 0)
Section B	P	35	(75 - 8)
	NP	35	(67 - 8)
Section C	P	92	(100 - 66)
	NP	85	(100 - 66)

AN AUDITORY SPEAKER-RECOGNITION EXPERIMENT

% accuracy for each test

Section A	P	NP
Test 1	95	85
2	85	70
3	15	0
4	70	40
5	80	50
6	75	55

Section B	P	NP
Test 1	55	57
2	28	28
3	11	15
4	40	15

Section C	P	NP
Test 1	95	95
2	95	90
3	85	70

It proved interesting to see how the subjects agreed with each
other. For example, in Test 3 of Section A, 29 of the 40 subjects
(including the best performers) chose the same incorrect match.

Conclusions
Various statistical tests have been tried on the data. For
Section A the Mann-Whitney U test (2 tailed) reveals that there is
a significant difference between the two groups of subjects at the
5% level. For Section B, Tests 1 - 3, and Section C, no signifi-
cant difference in performance is found. Test 4 in Section B
presents rather a problem in analysis. There are no matching pairs
in this test and it is felt that those listeners who found several
pairs ought to be considered more wrong than those who found only
one.

The Mann-Whitney test is perhaps not the most suitable for this
data. Probably the z-test is more appropriate. This test can be
used when comparing two groups and testing for agreement. Accord-
ing to the z-test there is no significant difference in the
performance of the two groups.

We hope to conduct a more extensive investigation in the near
future.