SPEAKER VERIFICATION UNDER REALISTIC FORENSIC CONDITIONS

Phil Rose

Department of Linguistics (Arts)
Australian National University

ABSRACT - A forensic phonetic experiment is described which investigates the nature of within- and between-speaker variation in demonstrably similar sounding voices. The centre frequencies of F1 - F4 in the naturally produced single word utterance *hello* are compared for 6 adult males. ANOVA results show that even similar sounding voices differ in F-pattern, but some of these differences are not realistically demonstrable. The magnitude of smallest significant difference between similar speakers is proposed as a way of estimating the involvement of more than one speaker forensically.

INTRODUCTION

An extensive literature now exists on speaker recognition using acoustic features (see, e.g. the bibliography in Hollien (1990: 350-357). However, the results are not generalisable in any straightforward way to the typical verificational situation in forensic speaker identification (FSI). Speaker recognition experiments are typically forensically inadequate in several important aspects which have mostly to do with unrealistic test material and closed-set testing (Rose 1995). Since the recognition literature concentrates on success rates, it is also difficult to find actual measurements of magnitude of difference between speakers. This paper looks at the problem of speaker verification under two forensically more realistic, and hence more stringent conditions. It also provides some data on magnitude of differences.

The primary inadequacy redressed in this paper is lack of control for prior similarity in voices. Under normal forensic circumstances, two voice samples to be compared will sound similar: otherwise it is unlikely they would be compared in the first case. In recognition exeriments, no attempt is usually made to control for similarity, other than chosing speakers with roughly the same accent. The consequent degree of between-speaker variation of course facilitates the task of identification. It still therefore needs to be asked what the nature of variation is between speakers who sound similar. To this end, the voices of six male subjects with demonstrably similar voices were used . In contrast to Rose (1996) -- which looks at how different the same speaker can be -- this paper asks how acoustically similar speakers can get.

Speaker recognition experiments have also been criticised (Nolan 1983, 12) for their tendency to underrate the importance of within-speaker variation, and realistic FSI has to be able to take this variation into account. This applies especially to the kind of linguistic variation which characterises situations with different degrees of formality, as for example when a suspect is being interviewed by the police, contrasted with when they are chatting over the phone to a friend, or when perchance they are committing an armed robbery. An attempt to redress this shortcoming is made in the present paper by eliciting utterances with realistic uncontrolled within-speaker variation. Subjects were asked to say the word *hello* as they thought they might say it under different conditions: questioning if someone was there; meeting a long-lost friend in the corridor; answering the phone; announcing their arrival home; reading it off the page. *Hello* also has the advantages, because of its pragmatic function, of, firstly, being able to be said naturally, thus avoiding the 'yellow lion roar' effect (Nolan 1983, 75). Secondly, it is capable of taking naturally a wide range of contrasting intonational nucleii, thus providing a potentially greater range of within speaker variation. The Australian *hello* also permits F1 and F2 to be examined over a fairly wide range.

PROCEDURE

Six adult male speakers of general to slightly broad Australian English were used. These speakers had been chosen initially on the basis of anecdotally reported similarity and were shown in experiments reported in Rose and Duncan (1995) to indeed have voices similar enough to be confused even by closest family members ('familiar' listeners). Four of the speakers are closely related: **JM** (49 y.o.), his two sons **DM** (23 y.o.) and **EM** (16 y.o.), and his nephew **MD** (24 y.o.). **RS** (50 y.o.) and **PS** (29 y.o.) are father and son. The confusions occurred in two types of experiments -- open class identification, and discrimination -- on three types of spoken material differing in length from one word (*hello*) through a short utterance to a 45 seconds text. Discrimination tests were also carried out with 21 unfamiliar listeners on the *hello* and longer utterance.

Details of the confusions reported in Rose and Duncan (1995) are summarised in table 1. Misidentification rate is shown in square brackets for the open identification tests by familiar listeners. Bold is used for familiar listener response, and italic for *hello* responses. The table is to be read in two different ways. Firstly, for open identification by familiar listeners (figures in bold), as "utterance from speaker on left was misidentified as spoken by speaker along the top". Thus the table shows that in one occurrence out of ten one of PS' *hello*s was recognised as spoken by EM by a familiar listener in an open identification test "***[1/10]***". Otherwise (non-bold figures) the table is to be read as "utterances of both speakers were heard as coming from the same speaker". Thus the table shows that in 11 occurrences out of 21 unfamiliar listeners identified *hello*s spoken by RS and MD as coming from the same speaker "*11/21*".

Table 1. Confusion data for the subjects' *hello*s. See text for details.

	JM	EM	DM	MD	PS	RS
JM	▨			***[1/10]*** 3/21		
EM	***[1/20]*** **1/30**	▨	5/21		***1/10*** *4/21* **1/10** 5/21	
DM		***[5/30]*** **[1/30]**	▨	***[1/30]***		
MD		***[1/10]*** **3/10** 10/21	***[1/10]*** **[1/30]**	▨		
PS		***[1/10]*** **[1/20]** **2/10** 7/21		[8/50]** ***4/10*** *8/21*	▨	
RS				***1/10*** *11/21*	*21/31**	▨

* this figure represents the sum of 6/10 familiar and 21/31 unfamiliar responses.

** this represents the sum of 7/20 for 45 sec text and 1/30 for longer utterance.

In the open identification test using the single word *hello*, every one of the 5 speakers used was misidentified as another one of the group at least once. In the longer utterance, three speakers were misidentified. Even the 45 seconds utterance was not free from judicially fatal errors, with 2 speakers heard as other members in the group. Generally, the same misidentifications occurred in the discrimination tests, where two different speakes' utterances were identified as coming from the same speaker. The same patterns characterised discrimination by unfamiliar listeners also, but more drastically so. Table 1 shows that every speaker had their voice misidentified as another in the group^1. Although Rose and Duncan (1995) point out that expectation effect also undoubtedly contributed, it seems justified to assume that these patterns of confusion, especially by close family members, are mostly to be ascribed to the auditory similarity of the stimuli. In particular, there is interaction between DM, EM, PS and MD, with DM's voice appearing to be especially similar to EM's, and PS's to MD. JM and especially RS do not show much interaction with the others.

Speakers were each recorded at a single sitting in the phonetics laboratory recording studio of the Linguistics department at the Australian National University. A Nakamichi 500 stereo cassette deck was used with a Nakamichi CM 300 microphone and wind shield. Forty-nine tokens of *hello* were elicited from the six speakers. An auditory analysis of the hellos revealed that /l/ was always velarised, and /ou/ varied between speakers in the fronting and rounding of its offglide. There was considerable between- and within-speaker variation in the realisation of the /h/ and the first vowel, and measurements from these segments were not used. MD produced two very different vowels in the second syllable: 6 with an unrounded off-glide; 4 with a rounded vowel, and one had a combination, with a glide towards a fronter position (high F2) in the first part of the rhyme, and then a rounded offglide (lower F2). The first two had to be treated as separate samples; the last was discarded. Between- and within-speaker differences in intonation and phonation type were also noted.

The *hellos* were digitised at 10 KHz and analysed with the ILS API routine which uses linear prediction spectral modelling with cepstrally based pitch period extraction. A filter order of 14, with hamming window and 100% preemphasis were used. The boundaries of the /l/, the offset of modal phonation in /ou/, and the onset of the first vowel were determined from inspection of the wave form produced by the ILS SGM command, in conjunction with conventional analog wide band spectrograms. Centre frequencies and bandwidth of the first four formants, as well as fundamental frequency, were sampled at the middle of the /l/ (labelled "/l/" below); at 25 percent intervals of the duration of the /ou/ ("0% /ou/, 25% /ou/, ... 100% /ou/"); and at the middle of the first vowel ("V") if present. Analog wide-band (350 Hz) contour spectrograms, which have a greater dynamic range than nornal bar spectrograms, were also made to assist in checking and interpreting the F-pattern extracted by the API analysis.

In addition to the expected acoustic characteristics for velarised lateral (Fant 1960, 162-168), several speakers showed sporadic additional poles from the middle of the lateral lasting sometimes well into the middle of the /ou/. Speakers also differed markedly with respect to higher frequency acoustic structure. JM showed for example many resonances above F2 that could only with difficulty be aligned with the other speakers' F3 and F4. Because of this, I was confident of comparability between his and the others' F4 only at the last two sampling points in /ou/. MD had clear formant structure up to F5.

Tokens were pooled for statistical purposes only if they sounded to have the same segmental target. (Thus MD's two different *hellos* were not pooled). Means and standard deviations were calculated for all parameters at all sampling points. These, together with number in sample, are given in table 2, which shows for example that JM's 5 *hellos* had a mean F1 for /l/ of 418 Hz, with a standard deviation of 44 Hz. It is readily apparent from table 2 that there are some very similar measurements for some parameters. For example, four speakers' mean F1 values in /l/ lie within a range of 20 Hz, and three speakers have mean F2 values at /ou/ 75% that lie within 8 Hz. The F-patterns of the most similar pair -- DM and MD -- are compared in figure 1 (at the end of the paper). There are also some obvious between-speaker differences, with a range of 332 Hz separating the lowest and highest F2, for example. The set of measurements for MD's F1 and F2 in his rounded vowel *tokens* is clearly different from those with the unrounded off-glide, but interestingly his F3 and F4 are very similar for both types.

RESULTS

Table 2. Mean and standard deviation values for F1 - F4 in six speakers' *hellos* Shading refers to excluded material -- see text.

F1	/V/	/l/			ou		
			0%	25%	%50%	75%	100%
JM	**582**	**418**	**548**	**607**	**550**	**496**	**380**
5	42	44	4	39	49	62	94
DM	**509**	**410**	**479**	**554**	**518**	**430**	**273**
17	100	78	83	45	65	79	45 *15*
EM	**582**	**412**	**499**	**556**	**498**	**352**	**270**
3	30	38	63	33	36	47	65 *2*
MD	**551**	**425**	**538**	**607**	**515**	**388**	**295**
6	79	42	52	63	86	78	64
/o/	**624**	**484**	**544**	**581**	**579**	**497**	**349**
4	6	53	35	34	37	41	74
RS	**678**	**499**	**654**	**673**	**615**	**499**	**489**
	65 *6*	17	40	29	54	92	135
PS	**548**	**405**	**585**	**622**	**574**	**454**	**368**
4	90	30	33	5	33	75	66
F2	/V/	/l/			ou		
			0%	25%	%50%	75%	100%
JM	**1055**	**760**	**998**	**1240**	**1366**	**1494**	**1491**
	124	40	43	55	44	41	86
DM	**1076**	**977**	**1061**	**1204**	**1377**	**1605**	**1686**
	103	72 *15*	93 *16*	99	95	65	84
EM	**1138**	**1011**	**1084**	**1216**	**1290**	**1436**	**1681**
	86	81	82	40	54	102	47
MD	**1023**	**851**	**995**	**1149**	**1485**	**1694**	**1701**
	54	45	50	60	74	87	139
	1103	**837**	**910**	**993**	**975**	**807**	**729**
/o/	46	64	41	70	44	45	98
RS	**1093**	**1011**	**1035**	**1131**	**1225**	**1293**	**1383**
	113 *6*	76	64 *4*	47 *6*	79 *6*	104	204
PS	**1092**	**948**	**1041**	**1165**	**1317**	**1388**	**1369**
	41	65	43	125	79	80	198

A single factor ANOVA, with 95% confidence limit, was carried out on the centre frequency data for F1 - F4. Results are shown in table 3, which lists the following information for each parameter: the F ratio (column F), asssociated significance (p), and the range in hertz between the lowest and highest values observed (R). Thus it can be seen that for F1 in /l/ the range between the lowest and highest mean values was 93 Hz, and that the F ratio of 2.56 gave a marginally significant probability of just under .05 that there are F1 values in /l/ that differ. Column "D" lists the number of significant differences in the corpus for the given parameter. Of more importance is the magnitude of the smallest significant difference between two similar sounding speakers ("SSDSSS Scheffé"). The SSDSSS values quoted are based on the (for the forensic context appropriately conservative) Scheffé post-hoc significance test for unequal sized samples (Elzey 1987,155). Thus for F1 in /l/ it can be seen that no significant differences between speakers can be demonstrated, but that for F2 in /l/ there are 6 pairs of speakers that differ, and the smallest significant difference is 126 Hz, between MD and DM. The magnitudes can be used as a kind of prior probability threshold to give some indication of provenance from different vocal tracts. If, for example, a significant difference is established between two recordings in the F2 of /l/ in comparable environments, the difference can be compared with the magnitude of SSDSSS found here. If it exceeds the SSDSSS of 126 Hz, the probability of the t test can be quoted that two different voices are involved.

Table 2 (con't)

F3	/V/	/l/			ou		
			0%	25%	%50%	75%	100%
JM	**2503**	**2529**	**2502**	**2677**	**2440**	**2426**	**2256**
	220 *3*	201 *3*	265 *2*	*1*	100 *4*	69 *4*	114
DM	**2460**	**2573**	**2617**	**2555**	**2476**	**2418**	**2395**
	155 *15*	88 *14*	100 *15*	95 *16*	129	120	168
EM	**2742**	**2778**	**2791**	**2694**	**2236**	**2150**	**2226**
	167	36	73	95	73	59	3 *2*
MD	**2418**	**2355**	**2376**	**2428**	**2452**	**2330**	**2390**
	44	75	69	45	72	144 *5*	214
MD	**2460**	**2423**	**2446**	**2369**	**2056**	**2107**	**1889**
/o/	68	45	47	123	112	154	70 *2*
							2576
							187
RS	**2529**	**2561**	**2577**	**2516**	**2365**	**2199**	**2225**
	57	38 *6*	108	126	166	119	185
PS	**2602**	**2651**	**2728**	**2732**	**2550**	**2424**	**2397**
	37	49	73	98	27	51	62 *3*

F4	/V1/	/l/	ou	ou	ou	ou	ou -
			0%	25%	50%	75%	100%
JM	**3761**	**3921***	-	**3792***	**3945***	**3633**	**3441**
	180 *3*	267 *2*		224 *3*	166	99	158
DM	**3513**	**3492**	**3525**	**3426**	**3389**	**3392**	**3446**
	161 *16*	108	90 *15*	74 *14*	122 *14*	157 *16*	257 *15*
EM	**3885**	**3796**	**3671**	**3320**	**3255**	**3166**	**3092**
	42	76 *2*	- *1*	125	87	63	42 *2*
MD	**3620**	**3610**	**3502**	**3534**	**3342**	**3358**	**3443**
	71	68	112	156	105	59	145 *5*
MD	**3614**	**3657**	**3587**	**3537**	**3472**	**3388**	**3349**
/o/	38	24	62	41	81	185	159
RS	**3700**	**3572**	**3589**	**3344**	**3065**	**3095**	**3018**
	203	154 *5*	135 *6*	198	112 *5*	120 *5*	180 *4*
PS	**3789**	**3707**	**3698**	**3640**	**3530**	**3473**	**3506**
	180	113	73	20	92 *3*	145	50 *3*

*It is not clear whether these resonances in JM are comparable with the other speakers' F4.

The six similar speakers were compared pairwise with respect to differences between them that had been shown by the Scheffé test to be significant. Distribution of significant differences between pairs is given in table 4. Three conditions of comparison are shown: ideal (column "I"), with all available centre frequency information, and nominal telephone ("T_n"), with effective information between 350 Hz and 3.5 KHz. In table 2, shading is used to highlight material excluded under telephone conditions. The third condition is realistic ("R"), which represents a situation with medium to poor quality non telephone speech with information restricted effectively to F1 and F2. The number of significant differences between speakers is given in column " D ".

Table 4 shows that there are no pairs that do not differ significantly at at least one of the 20 or 24 measuring points, providing information from all the first four formants is used (the figure of 20 must be quoted for JM, since four of his F4 measurements were excluded). This suggests that given a tightly enough circumscribed environment -- as for example the F-pattern in *hello* -- every speaker does inhabit their own region of variation in multidimensional acoustic space. If information is restricted to typical telephone bandwidths, however, there is one pair of speakers (PS/EM) for whom no significant differences can be demonstrated. If only the first two formants are considered, three pairs (PS/EM, PS/RS, DM/EM) cannot be shown to differ significantly.

Table 5 shows the 15 pairs of speakers ranked according to the distance between them measured in number of significant differences ("SD"). The significant differences are also broken down by formant. It can be seen that RS and DM differ the most, with 10 out of a possible 24 significant differences. But

Table 3 ANOVA results (95% confidence). See text for explanation

	p	F	R	D	SSDSSS (Scheffé)
F 1					
/i/	.0439	2.56	93	0	-
0%	.0001	8.35	175	2	155 EM/RS
25%	.0001	8.79	119	2	117 EM/RS
50%	.0155	3.27	117	1	98 DM/RS
75%	.0354	2.71	147	0	-
100%	.0001	7.98	219	2	195 MD/RS
F 2					
/i/	.0001	11.9	251	6	126 MD/DM
0%	.3337	1.2		0	-
25%	.2289	1.45	109	0	-
50%	.0002	6.79	260	2	152 DM/RS
75%	.0001	25.71	401	7	201 MD/JM JM/RS
100%	.0001	9.4	331	4	303 DM/RS
F 3					
/i/	.0001	10.08	423	5	202 MD/RS
0%	.0001	9.37	415	3	241 MD/DM
25%	.0003	6.53	304	3	216 PS/RS
50%	.0154	3.29	314	0	-
75%	.0003	6.41	276	2	219 DM/RS
100%	.1106	1.96	220	0	-
F 4					
/i/ *	.0001	8.58	424	2	306 MD/EM
0% *	.0002	7.64	347	3	259 EM/RS
25% *	.0031	5.09	320	2	296 PS/RS
50% *	.0001	9.05	465	3	277 MD/RS
75%	.0001	10.63	538	5	241 DM/JM
100%	.0193	3.23	488	0	-

* = excluding JM

Table 4. Distribution of significant differences between pairs of similar sounding speakers. Details in text

D	I	T_n	R
0	-	1	3
1	3	-4	5
2	4	3	4
3	2	3	1
4	2	1	-
5	1	1	1
6	1	1	-
7	1	-	-
8	-	-	-
9	-	1	-
10	1	-	-

Table 5. Significant differences between speakers

	Pair	SD	F1	F2	F3	F4
1	RS - DM	10	4	3	1	2
2	RS - MD	7	1	4	1	1
3	MD - EM	6		1	3	2
4	PS - MD	5		2	3	
5	PS - RS	4			1	3
6	DM - EM	4			2	2
7	RS - EM	3	2			1
8	MD - DM	3		1	2	
9	PS - DM	2		2		
10	RS -.JM	2		2		
11	DM - JM	2		1	1	
12	JM - EM	2		1	1	
13	MD - JM	1		1		
14	PS - JM	1		1		
15	PS - EM	1				1

there is a high proportion of pairs -- ca. 50% -- that only differ by 3 or less out of 20 or 24 possible differences . This lack of differentiation reflects corpus-internal within-speaker variance as well as similarity in mean F-pattern. Note that the pair MD - EM shown in figure 1 as having the most similar mean F-pattern is not the pair with the least significant differences. Whatever the source of the similarity, it will be appreciated that it might not be easy to distinguish some of these speakers under realsitic forensic conditions.

CONCLUSION

The results of this experiment indicate that under optimum conditions, with highly comparable segmental material, it should be possible to demonstrate a difference in the acoustics of two similar sounding voices. The magnitudes of least differences in F-pattern that will discriminate between similar sounding voices (SSDSSS) have been given above for selected segments. However, under more realistic circumstances, it has been shown that there is between 7% (1/15) and 20% (3/15) chance that an actual difference in speakers will not be reflected in a significant difference in acoustics.

It is clear that these exstimates will have to be revised, however, because of two conditions of the experiment which bias the results towards differentiation of speakers: control for contemporaneity and control for segmental differences. There is a need to explore the extent to which these results will be affected by relaxing these constraints towards even more realistic conditions. It is well known that recognition rates decrease dramatically if non-contemporaneous speech sample are used (Nolan 1983, 12). Therefore, the variation found in the voices of similar sounding speakers across time needs to be taken into account. The same six speakers were in fact re-recorded one year later, and the results are now being processed. It is to be expected that greater within-speaker variation exists across both corpora than demonstrated here, and that consequently it will be less easy to demonstrate differences between speakers.

The second factor which will have contributed to greater speaker differentiation in this experiment is the forced segmental comparability, both between- and within-speaker, from using the same word. Notwithstanding its desirability in this context, the degree to which within-speaker variance was minimised thereby is of course unrealistic. We as yet know little about the index of comparability that can be associated with most natural classes, although it is clear for example that unstressed vowels are not comparable with those in stressed environments (Ingram M.s.). It is planned to explore this issue using the large amount of longer utterances and connected text also recorded by our six similar speakers.

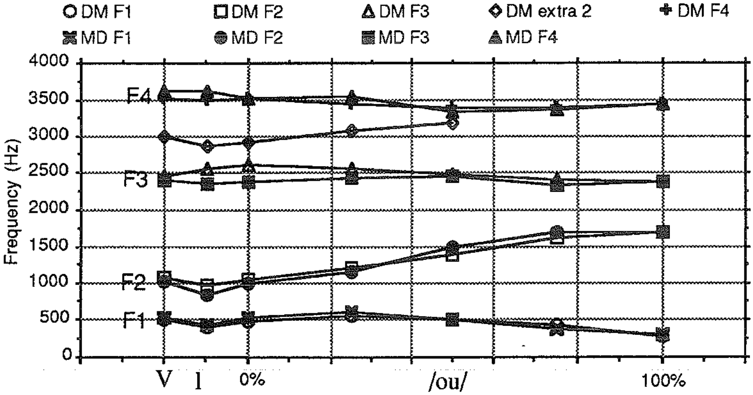

Figure 1. Mean F-patterns of the acoustically most similar pair of speakers DM & MD

NOTES

(1) Not all pairs were tested, so absence of data in a cell does not imply that speakers were not confused, only that the two were not compared in a discrimination test. The results in Table 1 are exhaustive, as far as the discrimination results are concerned. For the open identification test, of course, it is as if all speakers were being tested against each other.

REFERENCES

Elzey, Freeman F. (1987) *Introductory Statistics: A Microcomputer Approach* , Brooks/Cole Publishing Co.

Fant, G. (1960) *Acoustic Theory of Speech Production* , Mouton.

Hollien, H. (1990) *The Acoustics of Crime* , Plenum.

Ingram, J. (M.s.) "Formant Trajectories as Indices of Phonetic Variation For Speaker Identification" to appear in *Journal of Forensic Linguistics*

Nolan, Francis (1983) *The phonetic bases of speaker recognition*, Cambridge U.P.

Rose, Phil (1995) "On the acoustics of similar voices" Paper given at the Intl. Conf. on Forensic Linguistics, U.N.E.

Rose, P. & Duncan; S. (1995) "Naive Auditory Identification and Discrimination of Similar Voices by Familiar Listeners" *Journal of Forensic Linguistics* 2/1, 1-17.