F-PATTERN VARIABILITY IN DISGUISE AND OVER THE TELEPHONE - COMPARISONS FOR FORENSIC SPEAKER IDENTIFICATION

Phil Rose and Alison Simmons

Department of Linguistics (Arts)
Australian National University

ABSTRACT - A pilot experiment was carried out which investigates the nature of variability in the F-pattern of 3 speakers under conditions of disguise and telephone speech. Evidence is presented which may point to effective exclusion of F1 and F4 in comparisons involving the latter. The acoustic and forensic consequences of three different types of disguise are outlined.

INTRODUCTION

One major problem confronting forensic speaker identification (FSI) is lack of invariance between speaker and acoustic signal. This invariance is a property of natural speech, and can be seen for example when a speaker's mean fundamental frequency changes with emotional state. In FSI however, there are often additional, external, sources of invariance masking the identity of a speaker, as when he deliberately adopts a disguise, or when the signal is band-limited by the telephone, or indeed both. Disguise is known to be a devastatingly confounding variable in FSI (Hollien and Majewski 1977; Doherty and Hollien 1978), but little appears to be reported on the magnitude and nature of the acoustic distortion involved. The limited bandwidth of the telephone is a commonplace in the literature on FSI. One of its forensic consequences is that restricting acoustic information to that passed by a typical telephone increases the probability that two different speakers cannot be distinguished by their F-pattern (Rose 1996).

This paper reports a pilot study which examines the nature of distortion in the F-pattern occurring as the result of the above two factors -- disguise and telephone transmission. The two obvious forensically relevant questions to be addressed are to what extent an individual's F-pattern remains the same over the telephone, and whether there are any properties of the F-pattern which remain constant under disguise.

PROCEDURE

Three male speakers of general to broad Australian English from one family (J C S) were recorded under four conditions: two undisguised, and two disguised. The undisguised conditions were direct, and over the telephone. Speakers were first recorded in the A.N.U. phonetics laboratory, and, ten days later, over the telephone. The calls were made from different phones, but recordings were made over the same receiver using a high quality hand held cassette recorder.

Speakers were also asked to disguise their voices in any two ways. Disguises were recorded in the laboratory, rather than over the phone, to permit observation of methods of disguise, and to avoid confounding disguise with telephone effects. The subjects adopted a variety of disguise strategies. They used different phonation type, imitation of different regional or individual accent, and external distortion (speaking through a handkerchief; holding nose). Two of the disguises had to be discarded. One -- with a falsetto phonation type -- could not be adequately analysed by the computer without further processing because the formant tracker tended to resolve harmonics rather than formants. The other disguise -- speaking whilst holding the nose (but obviously with the soft palate down) -- resulted in relatively unintelligible speech, which would be forensically unrealsitic, there being little point for the criminal in incomprehensible annonymity.

The corpus comprised 10 /CVC/ words (*bead, deed; boot, coot; bought, board; cart, tart; bird, dirt;*) where V was one of the 5 Australian phonologically non-diphthongal long vowel phonemes: /ᵊi/ , /(ᵊ)ʉ/, /o/ , /a/ , /ə/. These vowels were chosen to give a good idea of an Australian speaker's F1/F2 acoustic space: the first 4 vowels can be expected to delimit the space, with /ə/ in the centre. As can be seen, the vowels were preceded by a voiced or voiceless stop /p t k, b d g/ and followed by an alveolar stop /d/ or /t/. In order to restrict the prosodic environment to sites of intonational prominence, the words were embedded once at the end of a short frame sentence where they could be expected to carry the tonic accent, e.g. *Where is the 'deed?*, and once at the end of a short sentence or clause immediately after the tonic accent e.g. *It was a 'good deed, not 'bad.* Four tokens of each vowel were recorded, except for /ə/, which by mistake was only recorded twice.

Table 1. Means and standard deviations for F-pattern in 3 speakers. See text for explanation.

Speaker S		F1		F2		F3		F4	
Vowel	Con	X	sd	X	sd	X	sd	X	sd
/əi/ T1	U_d	392	28	1577	95	2531	44	3293	129
	U_f	492	7	1588	99	2575	82	3582	131
	D_{ef}	384	48	1604	61	2575	125	3414	138
/əi/ T2	U_d	281	31	1939	96	2668	137	3223	138
	U_f	412	14	1992	43	2641	53	-	-
	D_{ef}	316	46	1973	90	2676	81	3380	74
	D_{ph}	240	26	2146	51 *3*	2668	0 *1*	3014	64
/əʉ/ T1	U_d	384	33	1528	103	2342	65	2982	60
	U_f	479	29	1585	148	2355	96	3477	325
	D_{ef}	364	55	1630	144	2514	106	3346	128
/əʉ/ T2	U_d	317	21	1703	48	2335	84	2889	108
	U_f	431	19	1750	93	2249	101	-	-
	D_{ef}	323	49	1734	15	2423	94	3138	35
	D_{ph}	247	37	905	86	2067	95	2960	56
/ə/	U_d	424	0	1444	32	2504	78	3206	57
	U_f	464	2	1515	25	2483	25	3353	94
	D_{ef}	499	58	1503	40	2531	38	3763	0 *1*
	D_{ph}	372	23	1112	28	1396	23	2855	100
/o/	U_d	364	27	750	25	2214	157	2758	113
	U_f	411	23	827	29	2257	84	3055	260
	D_{ef}	456	94	923	146	2492	123	3288	0 *1*
	D_{ph}	382	24	690	49	2253	77	2625	92
/a/	U_d	620	27	1343	36	2237	133	3216	63
	U_f	580	38	1300	39	2193	33	3354	370
	D_{ef}	658	32	1343	22	2129	49	3537	0 *1*
	D_{ph}	611	65	1110	71	1993	110	2801	0 *1*

Speaker C		F1		F2		F3		F4	
Vowel	Con	X	sd	X	sd	X	sd	X	sd
/əi/ T1	U_d	382	21	1498	76	2740	63	3346	79
	U_f	513	12	1514	36	2680	57	3311	165
/əi/ T2	U_d	246	12	2286	115	2820	139	3312	64
	U_f	353	36	2273	65	2713	61	3240	37
[i]	D_{sc}	232	19	2261	126	2839	166	3341	50
/əʉ/ T1	U_d	317	44	1522	159	2428	121	3006	119
	U_f	470	51	1449	176	2406	66	3011	99
/əʉ/ T2	U_d	276	17	1682	78	2259	30	2921	72
	U_f	378	75	1711	113	2208	50	2786	52
[y]	D_{sc}	243	10	1848	229	2535	190	3148	113
/ə/	U_d	390	18	1434	80	2726	23	3350	19
	U_f	451	0	1452	51	2695	13	3278	0
[ɩ] /_r	D_{sc}	568	0 *1*	1544	0 *1*	2688	0 *1*	3235	0 *1*
[œ] /_r	U_{sc}	256	0 *1*	2212	0 *1*	2778	0 *1*	3427	0 *1*
/o/	U_d	359	46	659	28	2805	178	2946	107
	U_f	374	66	737	52	2599	34	3012	67
	D_{sc}	432	73	833	54	2651	209	3078	117
/a/	U_d	765	32	1398	15	2517	269	3130	155
	U_f	754	52	1304	58	2060	148	2530	114
[a a̠ ɐ ɑ]	D_{sc}	536	86	1169	146	2682	71	3157	111

Utterances were first transcribed, and then processed with the Kay CSL equipment. After digitisation at 16 KHz, the time-aligned wave form and wideband spectrogram of the word containing the vowel to be analysed were produced. Formant tracks were then superimposed on the spectrogram with the 'LPC formant history' command. A 20 order model and frame size of 25 ms. was used. This was found to constitute the lowest acceptable order for adequate performance of the formant tracking. Spectral slices with FFT and LPC envelopes were made at sampling points chosen by visual inspection of the spectrogram. Where possible, centre frequencies of the first four formants were measured from the peaks of the LPC envelope. The F-pattern was sampled at points taken to represent vowel targets, i.e. points where the F2 time course appeared least affected by transitions to and from adjacent consonants. For /o/, where the F2 transition to a following alveolar starts early, a point before mid vowel duration was chosen. This was also necessary for some rhotacized forms produced under disguise, where both F2 and F3 were afffected. For /ə/ and /a/, and additional monophthongal forms produced under disguise, mid vowel duration was taken to represent the vocalic target. For /əi/, the typically dynamic F-pattern of which indicates two vocalic targets, the first target was sampled immediately after obvious onset transitions (usually about the third or fourth glottal pulse), and the second at a point of stabilisation in F-pattern before offset transition. As the transcription for /(ə)ʉ/ indicates, some tokens did not show an extrinsic vocalic onglide. However, it was expedient for purposes of comparison to sample all /(ə)ʉ/ in the same way as /əi/ with two assumed targets.

Table 1. (con't)

Speaker J Vowel	Con	F1 X	F1 sd	F2 X	F2 sd	F3 X	F3 sd	F4 X	F4 sd
/əi/ T1	U_d	477	43	1535	109	2634	88	3687	41
	U_t	465	33	1485	154	2478	116	3445	61
	D_h	419	70	1448	194	2636	104	3739	142
/əi/ T2	U_d	293	26	2172	19	2580	53	3495	175
	U_t	342	36	2192	59	2609	25	3422	65
	D_h	336	31	2242	137	-	-	33633	53
/əʉ/ T1	U_d	426	35	1569	99	2370	106	3524	100
	U_t	513	63	1557	178	2376	121	3388	54
	D_h	408	55	1494	245	2531	31	3460	103
/əʉ/ T2	U_d	313	35	1842	37	2268	65	3469	191
	U_t	361	6	1806	20	2227	60	3388	60
	D_h	314	8	1952	138	-	-	3482	69
/ə/	U_d	488	28	1428	52	2598	40	3815	133
	U_t	528	85	1446	93	2514	8	3403	26
	D_h	461	33	1591	37	2607	66	3702	132
/o/	U_d	424	7	735	34	2621	56	3409	37
	U_t	441	52	792	64	2529	15	3318	21
	D_h	412	66	780	109	2587	36	3325	42
/a/	U_d	796	18	1279	20	2683	48	3466	54
	U_t	756	30	1284	44	2739	111	3106	281
	D_h	645	27	1630	42	2769	75	3573	96

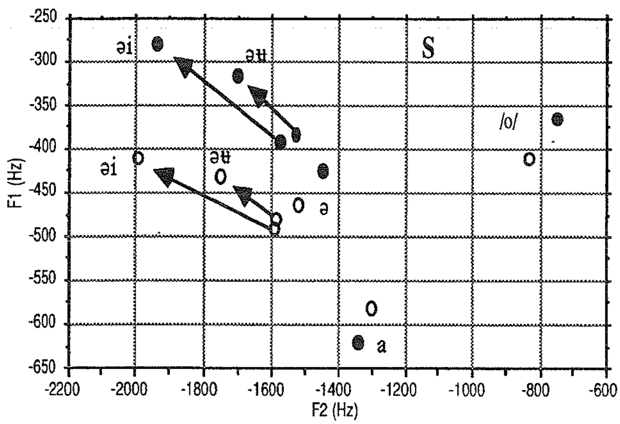

Figure 1. F1/F2 plot of speaker S's telephone (open circles) and direct vowels.

Means and standard deviations were calculated for the three speakers' formant centre frequencies in their undisguised direct and telephone speech, and disguised speech, and are presented in table 1. The vowel is shown on the extreme left (T1,2 = first, second vocalic target), then the condition ('Con'): U_d = undisguised direct; U_t = undisguised telephone; D_{ef} = disguised effeminate voice; D_{sc}= disguised Scottish accent; D_{ph} = voice disguised as Peter Harvey; D_h = voice disguised by handkerchief over mouth. Unless shown by an integer in italics after the standard deviation, number of items in sample is 2 for /ə/ and equivalent vowels, 4 otherwise.

TELEPHONE VS. DIRECT SPEECH

The acoustic relationship between telephone and direct speech vowels was very similar for all speakers. As far as the direct speech is concerned, all speakers showed the typical configuration for mildly broad Australian English. /a/ occupies a low central position, with mid /ə/ lying slightly forward of it. /o/ has a back, half-close position slightly higher than /ə/. /^{ə}i/ and /$^{(ə)}$ʉ/ are glides to high from an initial target position near /ə/: the second target is fully front for /^{ə}i/, and half-way between front and central for /$^{(ə)}$ʉ/. This configuration is clearly shown in figure 1, a conventional F1/F2 plot of speaker S's data. It is also evident from figure 1 that all vowels except the low /a/ show higher F1 positions in their telephone recordings. These differences were also found in speakers J and C, although J's were smaller. For speakers S and C, telephone vowels are higher by 92 Hz (S) and 77 Hz (C). For both these speakers, a two-way ANOVA shows telephone F1 in all vowels except /a/ to be significantly higher at at least the .0001 level. (All ANOVA reported in this paper have 95% confidence limits.) For J, F1 is higher in telephone vowels by 38 Hz, with ANOVA significant at p= .004.

It is unlikely that these observed differences in F1 reflect real articulatory differences. F1 might perhaps be expected to shift down, rather than up, as a result of decreased mouth opening from restricted jaw movement when speaking into the phone. Something of this nature may be occurring with the telephone /a/ of S and J . (T test shows J's telephone /a/ F1 is almost significantly lower (p= .06) than in his normal voice.) It is more likely that the differences are artefactual: due to the interaction between the real centre frequency of the (non-low vowel) formants and the low frequency cut-off characteristic of either the telephone, or , less likely, the microphone (or indeed a combination of the two). Experiments are in progress to determine the source of the F1 raising effect. If the telephone turns out to be responsible, however, F1 will be excluded from consideration in forensic comparisons using telephone speech for anything except low vowels, that is anything with male F1 lower than ca. 550 Hz.

Although some differences in F2 can be seen in figure 1, two-way ANOVA showed no significant differences between the telephone and direct speech for the three speakers. F2 would therefore seem to be a safe candidate for forensic comparison involving telephone speech. ANOVA on F3, however, showed speaker C to have significantly higher F3 values in his direct speech (p=.001). Post-hoc t tests showed F3 in his telephone /a/ and /o/ to be 457 Hz and 206 Hz significantly lower (p= .025, .026 respectively). These are fairly large differences, even for F3, and it needs to be checked whether another resonance was being registered here. Until this is clarified, it seems that comparisons involving F3 must be treated with caution.

Speakers J and C also showed significant differences in F4 (both p=.0002), with in all cases the direct recordings being higher than the telephone: J's by a mean of121 Hz, C's by 125 Hz. Speaker S's telephone F4 values could not be compared because their rather large variances pointed to the involvement of more than one resonance. The effect observed here is the opposite of that for F1, and it seems reasonable to assume that, again, artefactual differences are being registered. If the telephone is responsible, F4 cannot be used in forensi comparisons either.

DISGUISE

One of the disguises chosen by speaker J was to speak with a handkerchief held over his mouth (but not his nose). This resulted in some auditory distortion, but generally the message was clear. Acoustically, the disguise resulted in increased damping, and an abrupt drop in spectral energy above ca. 4 KHz. However, it was generally possible to identify formant peaks in the spectrum and measure most of the formants present in the speaker's laboratory recording (one exception was F3 at /əʉ/ offset). An unambiguous nasal resonance at 1134 Hz was clearly visible in J's disguised /ə/. In /a/, there were two candidate resonances for F2: one at 1165 Hz and one at 1630 Hz. Despite the similarity between the lower resonance and J's undisguised F2 of 1279 Hz, it was decided that this also represented the same nasal resonance as in /ə/. The higher value for F2 in /a/ would then be consistent with overall higher F2 values in J's disguised voice. Lowering of the soft palate may be a reaction to external obstruction of the vocal tract, since speaker C, who held his nose for one of his disguises, also did this.

In the F1/F2 plane, all J's disguised vowels, especially "/a/" and "/ə/", lie further forward. Two-way ANOVA shows the overall difference in F2 of 76 Hz to be significant (p=.019). ANOVA also shows a significant diference in F1, with disguised vowels having a 32 Hz overall lower F1 (p= .008). A significant interaction effect exists in both cases, probably reflecting much greater differences for the disguised "/a/", which lies 151 Hz closer and 351 Hz further forwards, and the disguised "/ə/", which lies 163 Hz further forward than undisguised counterparts. J's vocal tract's response of fronting and raising is fairly plausible, given the restriction on jaw opening caused by holding the handkerchief in place. Two-way ANOVA showed no significant differences between J's disguised and undisguised voices in F3 and F4.

Speaker J's data shows that under external distortion significant differences can be expected of the magnitude of 350 Hz in F2 and 150 Hz in F1 for open vowels, but no significant differences in F3 and F4. Where distortion of this kind is suspected therefore, large F1 and F2 differences are not sufficient for positive elimination, but significant differences in F3 and F4 might provide grounds for suspecting involvement of different speakers.

Speaker S affected an effeminate voice as one of his disguises. This voice was characterised mainly by dental articulation of /t/ /d/ and /s/ and /z/, as for example in [ˈwɒt̪ɪz̪ə ˈd̪ᵊid̪] *What is a deed?* The dentality was not totally consistent, however. For example, *Where is the boot?* was [ˈwɛɪz̪ə ˈbʉ̃t] with an apical alveolar in *boot*. As seen in the previous example, nasality was also audible, especially in /ʉ/ and /ə/. The speaker also tended to use a higher fundamental frequency, and preferred a rise-fall nucleus.

The content of all the utterances was clearly audible. Acoustically, the nasal setting showed up predominantly around 1 KHz in the region between F1 and F2 for [ɔ̃] and [ɐ̃], and also /əi/. Generally it was possible to identify formant peaks in the spectrum and measure the formants present without difficulty. However, it was noted from comparison of the LPC envelope with the FFT profile that often the F1 peak in the LPC appeared somewhat lower than the corresponding local peak in the FFT. This usually occurred when there was a significant amount of spectral energy present below F1, presumably from intrinsic or extrinsic nasalisation. It may be the case, then, that F1 has been estimated too low in the disguised voice.

Except for /a/, speaker S's effeminate disguised vowels lie further forward, and all except the onsets of /əi/ and /ʉ/ lie lower than the non-disguised vowels. Two-way ANOVA shows that disguised vowels differ significantly from undisguised in both higher F1 (with a mean difference of 30 Hz, $p= .024$) and higher F2 (with a mean difference of 58 Hz, $p= .018$). There is no significant interaction effect. Of note are the significant F1 differences of 75 Hz for /ə/, and 92 Hz for /o/. The higher F2 values are consistent with those reported for two speakers' dentalised, as opposed to neutral, voice in Nolan (1983: 169, 178). The overall higher F1 values for S's effeminate voice, however, show the opposite trend. As noted above, this may be a rather undesirable artefact of LPC extraction with nasalised vowels. Two-way ANOVA on the F3 values shows the disguised vowels to have on average 86 Hz significantly higher F3 ($p= .024$), with a significant interaction effect. A higher F3, by a very similar amount, is also typical of dentalised voice (Nolan 1983: 169,179). Of note here is a difference of 278 Hz in /o/ F3. F4 values in the disgusied voice are also significantly higher by a mean of 275 Hz (two-way ANOVA, $p= .0001$, with significant interaction). Of note here are large differences of 557 Hz for /ə/, 530 Hz for /o/, 364 Hz for /ʉ/ onset, and 321 Hz for /a/. These results demonstrate that the higher formants do not necessarily remain relatively constant under disguise, and where a change in articulatory setting is suspected fairly large changes in formant values cannot eliminate suspects.

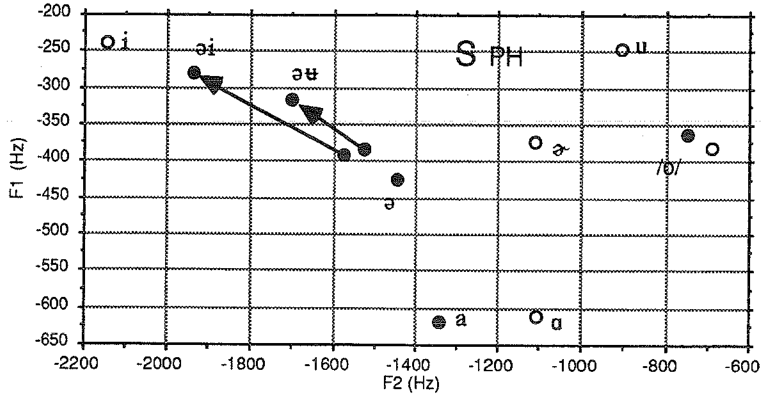

Figure 2. F1/F2 plot of speaker S's normal vs. Peter Harvey (open circles) vowels

The last two disguises involve imitation of different accents, one regional -- Scottish, by speaker C -- and one individual, or perhaps sociolectal -- Peter Harvey (a political commentator) by speaker S. Both these imitations had in common that different extrinsic vowel targets were involved, and not just vowels perturbed thorough an articulatory setting, or by external deformation. There is, however, always the possible complicating factor that the bogus accent incorporates a shift in articulatory setting as well.

Speaker S's Peter Harvey voice had considerably lower overall pitch than his undisguised voice, with the impression of a concomitantly lowered larynx. Open vowels had very salient creak. Vowel quality also differed considerably from the undisguised voice. /əi/ corresponded to monophthongal [i]; /əʉ/ to a very much backer monophthongal [u]. The accent was rhotic, but not entirely consistently so. Undisguised /ə/ corresponded to a backer rhotacized [ɚ], and /a/ corresponded to a backer [ɑ] followed by a post-alveolar continuant. One token of *board* was not rhotacized; in the other, the vowel was followed by a post-alveolar continuant. One other noticeable trait concerned the word-final voiceless alveolar stop, which in disguise was fairly regularly a strongly aspirated long apico-post alveolar, thus [ʃi boʈː ʰ ɪʈː ʰ nɒ ʈːʰ hɪ̈m] *'she bought it not 'him.* The retroflexion was also noticable on the voiced counterparts.

Figure 2 shows an F1/F2 acoustic plot of the vowels of speaker S's Peter Harvey voice compared to his undisguised vowels. It can be seen that the relative positions correspond fairly well to the auditory impressions, with [u], [ɑ], and [ɚ] having much lower F2 than their undisguised counterparts. [i] does not differ significantly from the offglide to /əi / in height, but is 207 Hz significantly fronter (t-test, p = .02). There is no significant difference between disguised and undisguised /o/, which means that disguised [u] and [ɑ] lie well outside the speaker's undisguised acoustic space. Two-way ANOVA showed that F3 values were overall significantly lower in the disguised voice (p= .0001). The overall rather large mean difference of 332 Hz is probably due primarily to the lowering effect of rhotization on F3 in conjunction with a backwards shift in tongue body. A lowered larynx may also be partly implicated. The F3 difference for /ə/ ~ [ɚ] was an enormous 1108 Hz (p= .004), and for [a] ~ [ɑʳ] 244 Hz (p= .031). [u] and /əʉ/ also differed in F3 by a significant 268 Hz (p= .005). The reason for the significant interaction effect in the ANOVA can be seen in the fact that the other two vowel pairs -- /o/ ~ [o], and /əi/ ~ [i] had very small differences in F3. Two-way ANOVA also revealed a significant overall difference in F4 of 192 Hz (p= .0001), with an interaction effect. Three pairs of vowels had small differences in F4: [u] ~ /əʉ/, /o/, and [a] ~ [ɑʳ]. Of the remaining three pairs, /əi/ ~ [i] differed significantly by 209 Hz (p= .05), /ə/ ~ [ɚ] by 351 Hz (p= .07), and [a] ~ [ɑʳ] by 415 Hz (p= .03). The lower F4 in this disguise may relate to a lowered larynx.

In Speaker C's Scottish accent, /ʉ/ was substitued by a fronter [y], /əi/ by a monophthongal [i]. /a/ and /ə/ showed a little variation in their substitutions. /a/ appeared as [a], [a̠], [ɐ] and [ɑ], and /ə/ was realised as [œ] in *bird*, but as [ɪ] in *dirt*. The accent was rhotic, but not consistently so. Post-alveolar approximants were produced in conjunction with a preceding shorter vowel in words having an orthographic *r*. However, one occurrence of *tart* [ta̠t] lacked a rhotic, and one had an off-glide [aᵊ]. Only one token of *board* had a rhotic. Other verisimilitudinous segmental substitutions were a voiceless labial-velar approximant and glottal-stop for *w* and *t* in *what*, and preglottalised alveolar stops in e.g. *bought.*

Acoustically, speaker C's two high front vowels /əi/ and [i] share the same target (they do not differ significantly in F1 or F2), but his [y] and /ʉ/ occupy significantly different positions corresponding to their auditory descriptions, as do the front rounded vowel [œ] and the corresponding /ə /, and the two /o/ and [o] vowels. The vowels corresponding to /a/ have about 200 Hz lower F1. A two-way ANOVA showed no overall significant difference between his disguised and undisguised voice in F3 or F4.

SUMMARY

This paper has highlighted some of the distortions in F-pattern that may occur in telephone speech and under disguise. In telephone speech, it may be the case that F1 and F4 distortion precludes their use in forensic comparison. In disguise, the comparisons above have demonstrated that the first two formants nearly always show significant differences irrespective of the type of disguise adopted. Neither does there seem to be a vowel type which is more refractory to disguise. F3 and F4, however, are not always significantly different under disguise. When a different articulatory basis is adopted, as with the effeminate voice, or probably with the Peter Harvey imitation, F3 and F4 show significant and large differences. However, with external distortion, or the imitation of a foreign accent without change in articulatory basis, F3 and F4 do not differ significantly. Obviously, before these findings can be applied forensically, they must be checked with more speakers. The carrier sentences should also not be neglected: it may be that there are less differences between a speaker's disguised and undisgusied voice further away from the word carrying the nuclear accent.

REFERENCES

Doherty, E.T. & Hollien, H. (1978) "Multiple factor speaker identification of normal and distorted speech" *Journal of Phonetics* 6, 1-8.

Hollien, H. & Majewski, W. (1977) "Speaker Identification by long-term spectra under normal and distorted speech conditions." *JASA* 62, 975-980.

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

Rose (1996) "Speaker Verification under realistic forensic conditions" This volume.