A FORENSIC PHONETIC INVESTIGATION INTO NON-CONTEMPORANEOUS VARIATION IN THE F-PATTERN OF SIMILAR-SOUNDING SPEAKERS

Phil Rose

Phonetics Laboratory, Department of Linguistics (Arts), Australian National University

ABSTRACT

A Forensic Phonetic experiment is described which investigates the nature of non-contemporaneous within-speaker variation in the F-pattern of intonationally different repeats of the same word hello said by 6 speakers in recordings separated by at least a year. Within-speaker variation is quantified by ANOVA on differences and Scheffé's F for centre frequencies of the first 4 formants at 7 well-defined points in the word.

1. INTRODUCTION

What differences will there be between the same word said by the same speaker on different occasions? This is a question of obvious importance in Forensic Phonetics, where a typical task involves comparison of 2 or more speech samples in order to either give an opinion on whether the samples come from the same speaker or different speakers, or (in a Bayesian approach) to quote probabilities of observing the magnitude of the difference between samples under competing hypotheses. Such comparison must rely on knowledge of both within- and between-speaker variation for the parameters being compared. Of crucial importance, however, is that the within-speaker variation be non-contemporaneous. This is because greater within-speaker variation is known to often characterise non-contemporaneous samples than contemporaneous. Many speaker-recognition experiments have shown that drastic drops in recognition performance occur when non-contemporaneous samples are used (Nolan 1983:12), and it is to be assumed that this is because of the concomitant increase in within-speaker variation relative to between-speaker variation. In addition, of course, use of non-contemporaneous samples reflects the reality of the forensic situation: if samples were contemporaneous, the identity of the criminal would be known.

This paper describes an experiment to find out the nature of non-contemporaneous within-speaker variation in the F-pattern of the same word hello said by 6 speakers in recordings separated by at least a year. Tokens were elicited with different intonations to reflect the reality of the forensic situation where intonation cannot, but tonic stress can be controlled. An additional way in which the real-world situation dictates procedure in Forensic Phonetic experiments is that comparison should be made between speakers who sound similar. Although this does not affect the present analysis, where within-speaker comparisons are being made, a set of 6 similar-sounding speakers is nevertheless used. These are speakers who had voices similar enough to be confused even by close family members in open identification and discrimination tests (Rose & Duncan 1995), and who were used in Rose (1996) to investigate the nature of between-speaker variation in similar-sounding speakers. It is intended to combine the NCV data described in this paper with data on the between-speaker differences in order to determine the limits of discrimination under realistic forensic conditions.

2. PROCEDURE

Six similar-sounding adult male native speakers of general to slightly broad Australian English were recorded. Four of the speakers are closely related: JM, his two sons DM and EM, and his nephew MD. RS and PS are father and son. Strictly speaking, of course, all speakers should have been recorded with different equipment, since suspect and criminal will not be recorded under identical conditions in reality. It was felt, however, that this represented an unacceptable loss of control over experimental conditions.

In order to obtain truly non-contemporaneous data, 2 recordings of the speakers were made separated by more than a year: the first (Rec(ording) 1) in 1994, and the second (Rec. 2) a little more than 4 years later (DM) or 1 year later (others). In order to elicit a range of realistically varying intonational patterns, speakers were asked to say the word hello as they imagined they might say it under different situations. In Rec. 1, 6 situations were stipulated: (1) answering the phone, (2) announcing their arrival home, (3) questioning if someone was there, (4) greeting a long-lost friend, (5) passing someone in the corridor, and (6) reading it off the page. Two speakers were asked to produce more than 1 repeat of the 6 situational hellos in order to provide more detailed information on within-speaker variation: 3 consecutive repeats were elicited from DM and 2 from MD. Sometimes a speaker produced an utterance other than hello. This happened particularly for "passing someone in the corridor", where DM and EM both said Hi!, and PS g'day. Evidently, hello is not the preferred lexical item for casual greeting in Australian English. EM also said Hi! for "announcing arrival home", and Hey! Buddy ! for "seeing a long lost friend". This reduced the number of his hellos to 3 in all. PS also had a different response (Hey! How yer doin'.) for "seeing a long lost friend". RS and DM cited the word hello once in conversation before formal elicitation, and these additional tokens were also used. In all, 49 tokens of hello were elicited from the 6 speakers in Rec. 1: DM 17; MD 12; JM 6; EM 3; PS 4; RS 7.

In order to elicit a still wider range of intonational patterns the number of situations was expanded in Rec. 2 to include: (7) meeting the Prime Minister, (8) admiring someone's appearance, and (9) trying to attract someone's attention. A larger number of tokens was also elicited by

incorporating 2 repeats for each speaker except DM, who produced 3. The repeats were separated by a ca. 60 sec. long reading of the 'rainbow passage' and are designated Rec. 2.1, 2.2, and (for DM) 2.3. Recording 2 yielded less alternative utterances to hello than Rec. 1, except for EM, who still preferred utterances other than hello (Hi!, Hey!)) for some situations. In all, 115 hello tokens were elicited from the 6 speakers in Rec. 2: RS and MD produced 18 each; DM 27; PS produced 2 and JM 1 extra token for 20 and 19 tokens respectively, and the recalcitrant but consistent EM produced 13.

The word hello was chosen because it can be said naturally on its own, thus avoiding the 'yellow lion roar' effect (Nolan (1983:75). It is capable of taking naturally a

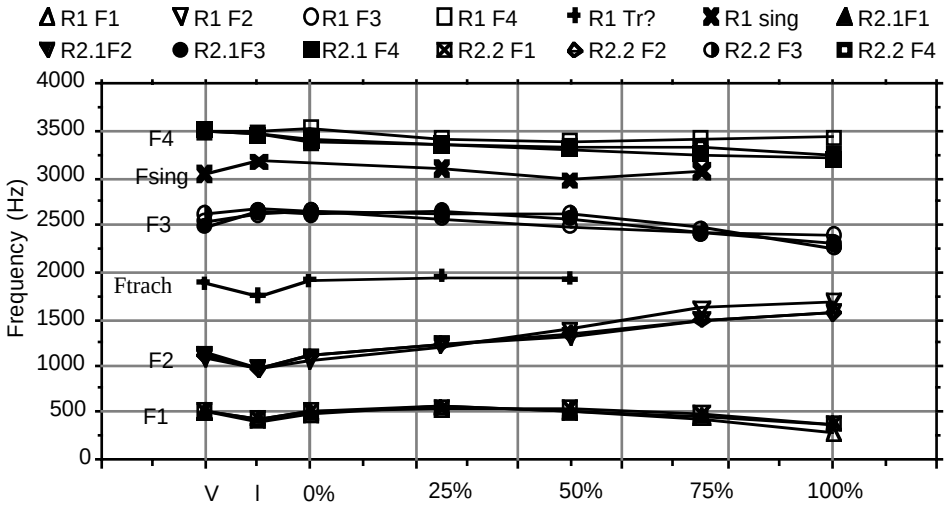

Figure 1: Non-contemporaneous F-pattern variation in DM's hello.

fairly wide range of contrasting intonational nuclei, thus providing a potentially greater range of within-speaker variation. Conversational analysis also shows hello to be an exceptionally high frequency lexical item in domestic telephone conversations, where it constitutes the answer in the summons-answer sequence. Since intercepted telephone conversations are common material for forensic analysis, the forensic value of knowing about between- and within-speaker discriminability in hello is obvious. The quality of the two vocalic targets in the second syllable diphthong of the Australian hello, and its typically velarised lateral, also permit the lower formants to be examined over a wide, though not ideal range. Australian /ou/ typically has a fairly front offglide ranging between [y] and [ ]. It is thus near in quality to high front segments that tend to have the most individual-identifying potential.

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 maximum number of peaks to be extracted was set at 5. 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. The following 7 sampling points were defined with respect to these events: the middle of the /l/ (labelled 'l' below); 25 percent intervals of the duration of the /ou/ ('0%', '25%', etc.) and the middle of the first vowel ('V') if present. The ILS analysis frames corresponding to the sampling points were then printed out, and centre frequencies of the resonances transferred to a spread sheet for statistical analysis (F0 and bandwidth were also extracted). Analog wide-band (350 Hz) spectrograms were also made to assist in checking and interpreting the F-pattern extracted by the API analysis.

Formants were identified on the basis of expectation and continuity. The former criterion, as is well known, involves a degree of circularity (Nolan, 1983: 86,87). A formant is identified because one knows from previous studies and the acoustic theory of speech production where in the frequency range to expect it for a given segment. According to this criterion, it was assumed that the first 4 resonances shared by all tokens in the latter part of the /ou/ diphthong represented F1 to F4. Resonances continuous with these in the first part of the diphthong, the lateral, and first vowel were then also identified as F1-F4. It was generally easy to identify F1 and F2 in this way, but there were the expected problems with some speakers' higher formants. In particular, RS's F3 and F4 in his 2nd recording were difficult to identify, and were not further analysed in this paper. JM had 2 fairly consistent resonances in the F4 region, the lower of which (labelled 'Fs' below) was probably a singer's format. This resonance has been included in the NCV comparison.

To show a typical F-pattern for hello and how it varies over [redacted] es DM's mean F-pattern in his 17 [redacted] ellos with recordings 2.1 and 2.2 of his mean F-pattern 4 years later. Note the relatively acute value of the diphthongal offset. It can be seen that there is very little obvious variation in his first 3 formants and only slightly more in F4. As with several other speakers, DM's first recording gave evidence of a tracheal (or possibly lateral F3) resonance between F2 and F3. He also had a singer's formant around 3 KHz. Neither of these was consistently resolved in his 2nd set of recordings.

Variation in F-pattern was quantified by ANOVA. Significance was set at the 95% confidence level and the (for the forensic context appropriately conservative) Scheffé post-hoc significances for unequal sized samples. (Scheffé's F ($F = (\ _i - \ _j)^2 / MS_w (1/N_i + 1/N_j$

)) shows the size of the squared difference between the two means being compared (e.g. Recs.1 and 2.1) relative to the average within-group variance (i.e. for all 3 Recs. 1, 2.1, 2.2). The within-group mean square is adjusted by a term correcting for different sample sizes. Before evaluating the non-contemporaneous variation across the recordings separated by more than a year, the short-term variation was examined between the 2 (in DM's case 3) repeats of the second recording.

3. RESULTS

The procedure was intended to elicit a variety of different intonational patterns in order to introduce forensically realistic within-speaker variation, and the hellos in both 1st and 2nd recordings were indeed characterised by a variety of intonations. In the 1st recording, 4 of the 6 speakers produced up to 5 different intonation patterns on hello. Typically, the 2nd recordings included 1 or 2 new patterns (low prehead followed by rise-fall nucleus [L.LHL] was common for 'admiring someone's appearance'), but lacked some of the patterns in the 1st recording. This gave a substantial overlap between 1st and 2nd recordings and about the same variety across recordings. For exa[redacted]H], Downstep = [L.H!H], [H!H], [redacted]H]. Fall-Rise = [L.HLH], [L.LHLH]. Rise-Fall = [L.LHL].

	Fall	Rise	Down step	Fall Rise	Rise Fall
DM1	12	59	24	6	-
DM2	26	44	15	4	11
MD1	58	17	25	-	-
MD2	44	17	28	-	11
JM1	17	50	17	17	-
JM2	32	26	37	5	-
PS1	25	50	25	-	-
PS2	15	50	20	5	10
RS1	43	43	-	14	-
RS2	21	42	21	16	-
EM1	33	66	-	-	-
EM2	27	45	9	18	-

Table 1: Percent distribution of intonation types in corpus. Fall = [L.HL], [H.HL]. Rise = [L.LH], [H.LH],

had 5 different intonation patterns in both his 1st and 2nd recordings. The 1st recording contained: rising nucleus on second syllable with both low and hi prehead on first ([L.LH], [H.LH]); stylised downstepped high on second syllable with low prehead ([L.HH!]); fall-rise spread over both syllables ([HL.LH]); low prehead with fall ([L.HL]). The 2nd recordings lacked [H.LH] and [HL.LH], but added low prehead with fall-rise ([L.HLH]), and stylised downstepped high across both syllables ([H.H!]). The distribution of the basic intonation types (Fall, Rise, Downstep, Fall-Rise, Rise-Fall) in the corpus is shown in table 1.

	JM	DM	EM	MD	RS	PS	
F1	29	42	31	23	17	30	29
F2/F3	69	62	84	79	71	75	73
F4	178	86	83	98	-	171	123
F1/F3	.51	1.21	.98	.67	.25	1.01	.77
F2/F4	2.57	2.06	1.15	1.59	1.04	2.74	1.86

Table 2: Mean non-contemporaneous variation in formant (Hz, above), and Scheffé's F (below).

Results of the contemporaneous comparison show firstly that significant within-speaker differences in mean formant values do occur in the same word spoken in samples separated by as little as a couple of minutes. However, they are rare -- only 15 out of 138 pairs -- and are confined mostly (10/15 occurrences) to F4. Significant differences are not distributed evenly with respect to speaker. In all his data DM has no significant differences, and PS only 1, whereas JM has 8 and MD 4. Generally the magnitude of the Scheffé F is very small for contemporaneous within-speaker variation, as might be expected: ca. 0.5 for F1, 1.0 (F2/F3), 3.0 (F4). Scheffé's F values greater than ca. 3.5 indicated significant differences between means in this corpus.

Since the contemporaneous variation showed some, albeit few, significant differences, it was decided to quantify the non-contemporaneous variation separately with respect to second recording sessions. Thus differences between Rec. 1 and Recs. 2.1 and 2.2 (and 2.3 for DM) were analysed separately by ANOVA. The ANOVA results showed a wide range of differences, both between- and within-speaker, between the means of the two recordings. Thus, taking differences at the 75% sampling point in F1 as an example, PS had very similar small differences of 12 Hz and 7 Hz between his 2 sets of recordings, EM had much larger, but still similar, indeed identical differences of 77 Hz across recordings, whereas JM's 2 recordings differed considerably: a difference of 18 Hz for Rec. 1 vs Rec. 2.1, and 86 Hz for Rec. 1 vs. Rec. 2.2. Another example of a large difference between recordings was JM's F4 at the 0% sampling point. Differences were 43 Hz (Rec. 1 vs. Rec. 2.1), compared to 397 Hz (Rec. 1 vs. Rec. 2.2). Out of the 298 paired comparisons, there were 36 instances of non-contemporaneous differences significant at 95%. These appeared to be distributed non-randomly with respect to Formant (F2 and F4 showed many more significant differences than F1 and F3), and Speaker, (JM and PS showed more significant differences than the others).

A 3-way Generalised Factorial ANOVA (Speaker x Formant x Sampling point) was carried out on the non-contemporaneous differences, and associated Scheffé's F values, using a General Linear Model. For the non-contemporaneous differences, this showed no significant overall difference for Sampling Points ($F = 1.067$, $p = .38$); or Speakers ($F = 1.552$, $p = .176$), but a very highly significant difference for Formants ($F = 29.98$, $p < .0001$). Conservative post-hoc tests (Tamhane's T) showed significant differences between all except F2 and F3. Significant interaction at 95% was noted between Speaker and Formant ($F = 2.54$, $p = .003$), and between Speaker and Sampling Point ($F = 1.665$, $p = .032$). Significant interactions were noted for Formant/Speaker, and Sampling Point/Speaker. For Scheffé's F, there were significant differences for Formants ($F = 16.19$, $p < .0001$), and for Speakers ($F = 3.01$, $p = .013$), but not for Sampling Point. There were also significant interactions between all 3 factors. Post-hoc tests showed no significant differences between F1 and F3, and between F2 and F4; and no significant differences between any pairs of speakers (probably due to a difference in the conservatism between the ANOVA and the post-hoc test).

These results are summarised in table 2, where NC differences and Scheffé F values are pooled for all sampling points, but speakers are kept separate. Difference values

		F1			F2			F3			F4		
			d	n			n			n		d	n
JM	1-2.1	-21	72	345	-1	102	340	70	293	330	-30	233	135
											-61	247	157
	1-2.2	6	88	315	4	137	305	13	291	275	131	247	107
											170	289	162
DM	1-2.1	-31	92	1053	28	124	1044	-9	202	1035	92	205	968
	1-2.2	-29	92	1053	31	115	1044	-43	197	1035	75	189	1001
	1-2.3	-42	92	1038	28	130	1029	-70	196	1035	45	189	969
EM	1-2.1	3	70	114	-50	96	123	-39	153	126	27	199	114
	1-2.2	-9	82	127	-44	128	147	1	158	144	24	135	140
MD	1-2.1	-8	95	366	50	132	369	42	245	337	-30	173	353
	1-2.2	-29	98	378	34	142	369	73	142	360	134	150	351
RS	1-2.1	5	101	419	-10	151	348	-	-	-	-	-	-
	1-2.2	-11	82	400	-23	154	345	-	-	-	-	-	-
PS	1-2.1	-11	72	272	-73	123	280	-86	159	266	222	163	232
	1-2.2	-22	76	264	47	168	273	66	124	262	-145	180	240
All signed		-14	85		-1	132		6	196		48	190	
Allunsigned		16			33			47			92		

Table 3: Means (Hz) & standard deviations (Hz) for between-token NCV differences. n = number of pairs.

for F2 and F3, and Scheffé values for F1 and F3, and F2 and F4 are pooled. Table 2 shows for example that JM's mean difference in F1 between non-contemporaneous samples of the same word was 29 Hz. Mean differences for all speakers are ca. 30 Hz for F1, 80 Hz for F2 and F3, and 130 Hz for F4. (Mean standard deviation values for the differences (not shown in table 2) are ca 20 Hz (F1), 50 Hz (F2, F3), and 80 Hz (F4)).

The results above have quantified the non-contemporaneous variation that obtains between mean F-pattern values. This knowledge is of use if there are several repeats of the same word in both samples from which mean values can be calculated. However, suppose that the samples to be compared forensically contained just one token each. In order to evaluate this situation, the distribution of non-contemporaneous differences between individual tokens must be known. A program was written to calculate the signed mean and standard deviation of the differences between each non-contemporaneous pair in a speaker's data. (The mean difference will of course be the same as the difference between a speaker's sample means, previously discussed.) Signed, rather than unsigned or Euclidean differences were calculated, because the latter's distribution, truncated at zero, has properties which make it statistically intractable. Results are given in table 3, which also includes values for JM's singer's formant as the upper of the 2 sets of figures in his F4 column. Table 3 shows, for example, that of DM's 1053 pairs of tokens from Rec. 1 (17 tokens) and Rec. 2.1 (9 tokens) the mean difference in his F1 was -31 Hz -- i.e. F1 in Rec 2.1 was on average 31 Hz higher than Rec. 1 -- and the standard deviation of the differences was 92 Hz. A 2-way ANOVA shows significant differences ($p < .0001$) between the standard deviation values for all formants except F3 and F4. There is also a significant difference between speakers ($p < .0001$), which concerns differences (reflected in a significant interaction effect ($p = .004$)) between JM and some other speakers (EM, PS, MD) in F3 and F4 standard deviation.

Considered across all 6 speakers, the signed differences for F1 to F3 can be seen to cancel out fairly well. However, in forensic discrimination we do not know the sign, but only the magnitude of the difference. The mean magnitudes of the differences for the 6 speakers, also given in table 3, are ca. 20 Hz (F1), 40 Hz (F2), 50 Hz (F3) and 100 Hz (F4). Together with the standard deviations, these values can be used (assuming distributional normality) to estimate the probability of one of the terms which determine the Likelihood Ratio in a Bayesian approach: the probability of observing a given difference between samples assuming they were spoken by the same speaker. For example, comparing two of DM's non-contemporaneous tokens (1.1 & 2.1), the mean difference in F3 was 44 Hz. The probability of this observation assuming the tokens are from the same speaker (i.e. using within-speaker non-contemporaneous mean and sd values of 50 and 200 Hz) is 0.038. The probability of observing this difference assuming different speakers (using for illustration between-speaker mean and standard deviation values for F3 of 161 and 215 Hz from a comparison between the two most similar speakers DM and MD) is 0.0034. This means that the observation is 11 times more likely if the samples were from the same speaker.

5. REFERENCES

1. Nolan, Francis, The Phonetic Bases of Speaker Recognition, Cambridge University Press, Cambridge, 1983.

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

3. Rose, Phil, "Speaker Verification under Realistic Forensic Conditions", in Paul McCormak and Alison Russell (eds.), Proceedings of the Sixth Australian International Conference on Speech Science and Technology, Australian Speech Science and Technology Association, Canberra: 109-114, 1996.