A COMPARISON OF TWO ACOUSTIC METHODS FOR FORENSIC SPEAKER DISCRIMINATION

Phil Rose* and Frantz Clermont**

***Phonetics Laboratory, Linguistics Program, Australian National University**

****School of Computer Science, University of New South Wales (ADFA)**

ABSTRACT A pilot forensic-phonetic experiment is described which compares the performance of formant- and cepstrally-based analyses on forensically realistic speech: intonationally varying tokens of the word hello said by six demonstrably similar-sounding speakers in recording sessions separated by at least a year. The two approaches are compared with respect to F-ratios and overall discrimination performance utilising a novel band-selective cepstral analysis. It is shown that at the second diphthongal target in hello the cepstrum-based analysis outperforms the formant analysis by about 5%, compared to its 10% superiority for same-session data.

1. INTRODUCTION

A recurrent topic in Forensic Phonetics, where speaker verification under very much less than optimum conditions is a major concern, is its relationship to Automatic Speaker Recognition (ASR). Leading forensic phoneticians (e.g. Künzel 1995: 79) emphasise the difference in the real-world conditions between Automatic and Forensic speaker recognition, especially in the lack of control over operational conditions in forensic speaker identification, and point out that fully automated forensic speaker identification is not a possibility.

This should not imply, however, that some of the analytical techniques common in ASR are of no forensic use. Although forensic speaker identification must usually rely, *inter alia*, on comparison of individual formants (e.g. Nolan 1990, Labov & Harris 1990: 287ff.), it is generally assumed that in ASR cepstrally-based methods are superior. This is because the cepstrum tends to exhibit strong immunity to "noninformation-bearing variabilities" (Rabiner & Juang, 1993: 169) and, hence, greater sensitivity to distinctive features of speech spectra. Aside from actual performance, the cepstrum is more easily extracted than the F-pattern, with its inevitable problems of identification and tracking of the higher formants. Although the nature of the cepstrum *qua* smoothed spectral shape is far from *unanschaulich*, arguments against the use of the cepstrum in forensic phonetics centre on the abstract nature of its mathematical basis (van der Giet 1987: 125), and include its indirect relationship to auditory and articulatory phonetic features - the latter of considerable importance in forensics - and the difficulty of explaining it to the jury (Rose 1999b: 7). It is of both interest and importance, therefore, to examine the performance of that algorithmic mainstay of ASR - the cepstrum - on forensically realistic data. That is the aim of this paper. It has only recently become practical due to mathematical developments (Clermont & Mokhtari 1994), which make it possible to specify the upper and the lower bound of any frequency band directly in the computation of the cepstral distance.

The speech data we use are forensically realistic in four important ways. Firstly, they are from speakers that sound similar. This is an obvious requirement on any forensically realistic speaker discrimination experiment, since if two speech samples do not sound similar, *ceteris paribus*, it makes little sense to claim that they come from the same speaker. Secondly, the data are from different sessions. If the data were from the same session, the criminal would be known. Thirdly, the data is not controlled intonationally. This is because, even if the same word occurs in criminal and suspect samples, it is unlikely that it will occur in exactly the same prosodic environment in criminal and suspect material. Lastly, we use a word very common in telephone intercepts - *hello*. We therefore use the most controlled data that can be realistically expected, namely variation within a segment that occurs in the same position in intonationally varying repeats of the same word. The word *hello* is capable of taking naturally a wide range of contrasting intonational nuclei, thus providing a potentially greater range of within-speaker variations.

2. PROCEDURE

Subjects

Six demonstrably 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. Similar-sounding means similar sounding to naive listeners, and presumably rests on similarities in auditory *voice* quality rather than *phonetic* quality (for this important distinction, see the collection of papers in Laver (1991)). The speakers had been chosen initially on the basis of anecdotally reported similarity (it was claimed for example that a father and son were commonly confused by their wife and mother over the telephone). The six speakers were shown in subsequent experiments reported in Rose and Duncan (1995) to indeed have voices similar enough to be confused in open identification and discrimination tasks even by closest family members. It is not surprising that perceptual discrimination tests with naive unfamiliar listeners also showed the six voices to be highly confusable.

Recordings and Cepstral Processing

Use was made of two sets of recordings to furnish genuine long-term data for comparison. These were separated by a period of four years (DM) and one year (the others), and are referred to as R(ecording) 1 and R(ecording) 2. Details of the within- and between-speaker variation in the two sets of recordings can be found in Rose (1999a) for R1, and Rose (1999b) for R2. Two sets of data were obtained in the second recording, and data from the second set were used. In order to elicit a selection of realistically varying intonational patterns, speakers were asked to say the word *hello* as they imagined they might say it under six different situations. (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, (6) reading it off the page. In the second recording session these were expanded to: (7) meeting the Prime Minister, (8) admiring someone's appearance, and (9) trying to attract someone's attention. Some speakers, especially EM, preferred utterances other than *hello* (e.g. *Hi, Hey, G'day*) for some situations, and so had less hello tokens than the others.

Table 1. Numbers of tokens recorded

	DM	JM	EM	PS	RS	MD
R1	17	6	3	4	7	6
R2	9	9	7	10	9	9

The *hellos* were recorded using professional equipment in the A.N.U. phonetics laboratory recording studio. The resulting analogue signals were then sampled at 10 kHz, and analysed (ILS API routine) by linear prediction (LP-order 14) of 20msec Hamming-windowed frames with 100% pre-emphasis and a frame advance of 6.4msec. The boundaries of the /l/, the offset of modal phonation in /oʊ/, and the onset of the first vowel were determined from inspection of the wave-form produced by the ILS SGM command (yielding a quasi-spectrogram plot), in conjunction with conventional analog wide-band spectrograms. The following seven temporal landmarks were defined: the middle of the /l/; 25% intervals of the duration of the /oʊ/; and the middle of the first vowel if present. The ILS analysis frames corresponding to the landmarks were then identified, the centre- frequency of the first four formants identified, and transferred to a spreadsheet for statistical analysis. In addition, the set of 14 LP-derived cepstral coefficients corresponding to each landmark were retained for further processing.

Cepstral distances were calculated both for the entire Nyquist interval, and also for sub-bands of this interval. This was done in order to obtain cepstral analogues of the formant-based measures of variance and distance that are commonly sought in forensic speaker identification. To this end we used Clermont & Mokhtari's (1994) parametric formulation of the cepstral distance, which permits a posteriori specification of the upper and the lower bound of any frequency sub-band between 0Hz and the Nyquist frequency. The sub-bands corresponded to the spectral regions straddling the frequency range of each of the four observed formants. The upper bound for each sub-band was set at the frequency of the highest mean-formant's centre-frequency observed plus one standard deviation, and the lower bound at the lowest mean minus one standard deviation. For example, the highest mean centre-frequency (499 Hz) for F1 in /l/ was produced by RS, with a standard deviation of 17 Hz; and the lowest mean centre-frequency (405 Hz) for F1 in /l/ was produced by PS, with a standard deviation of 30 Hz. The sub-band constrained to the F1-range was thus specified in terms of an upper bound of 499+17 = 516 Hz and a lower bound of 405-30 = 375 Hz.

3. RESULTS

Intonation

As intended, the different situations did elicit a forensically realistic variety of different intonational patterns. Thirteen different patterns occurred, which were formally classifiable according to their nuclear pitch into five types: *Fall, Rise, Downstep, Fall-Rise and Rise-fall* (Rose 1999b: 10). With the exception of JM, who produced proportionately more downsteps, the between-speaker intonational variety was largely comparable.

Auditory phonetic quality

Although the speakers were largely comparable in the suprasegmental aspects of their phonetic quality, they showed both between- and within-speaker segmental variation in the backness and rounding of the diphthongal offglide in /oʊ/. (Realisations of Australian /oʊ/ typically show a wide range in the backness of the diphthongal offglide). The /oʊ/ diphthongs in the data collected here have an offglide ranging between [y-] and [ʉ-/ʊ+] (and a fairly open central initial target [ɐ]). They are thus representative of a major part of the typical range. Two speakers (PS and RS) consistently had what sounded like a backer/rounder off-glide: [ʊ+]; DM's offglide was consistently fronter: [ʉ], and JM's offglide sounded slightly fronter and lower: [ʉ˕]. The other two speakers showed within-speaker variation. Some of EM's /oʊ/ tokens sounded the same as DM's, and some sounded backer/more rounded, although not as much as PS and RS. MD was notable for his wide range of off-glide realisations, from [ʊ+] through [ʉ] to [y-]. Also noticeable were differences in the secondary articulation of /l/ (pharyngealised vs velarised), and incidental differences in the first vowel phoneme /ʌ/ vs. /ɛ/ vs. /æ/. An important point is that, as a result of these auditory linguistic differences, it was possible to discriminate rather easily some pairs of speakers who had similar voice quality but different phonetic quality.

Formant analysis

As might be expected from the similarity in their auditory voice quality, some pairs of speakers had very similar mean F-patterns. Within-session Euclidean distances were calculated for all between-speaker pairs for all four formants both combined, and individually for both recordings. Figure 1 shows the mean F-patterns of the two most similar speakers in R2 (PS, DM), according to overall Euclidean distance. The

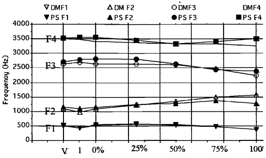

Figure 1. Mean F-patterns compared for PS and DM.

mean Euclidean distance for this pair over all four formants was 109 Hz, with individual formants, from F1 through F4, as follows: 15 Hz, 138 Hz, 118 Hz, 120 Hz. (In R1 the most similar pair was DM and MD, who were separated overall by 101 Hz, with individual formant differences of 39, 81, 160, and 85 Hz (Rose 1999a: 17)). It can be seen from figure 1 that PS and DM display a fairly high level of congruence in all formants except F3 at the onset of the diphthong, and F2, F3 and F4 at offset. Notably, the difference in F2 over the last two landmarks in /ou/ corresponds to an audible difference in the acuteness of the second diphthongal target.

In spite of the problems in formant identification alluded to above, it was generally easy to identify these six speakers' formants—for some even up to F5. There were two exceptions. In both recordings, JM appeared to have two close resonances in the area of F4, neither of which was unambiguously continuous. The higher of the two had to be identified as "true" F4 and the other as a singer's formant (Rose 1999a: 12-13). In RS's second recording, his F3 and F4 were not reliably extracted. These two speakers offer the possibility for cepstral analysis to demonstrate its superiority.

4. ANOVA COMPARISON

Preliminary to the discrimination, in order to find out where the points of greatest within- to between-speaker variation in *hello* lay, a single factor ANOVA was carried out for both the formant and cepstral data, on the data in both R1 and R2, at each of the sampling points. The resulting F-ratios are given in table 3. Very few comparable points exist between the cepstral and formant F-ratios (one of the reasons for this is because the F-ratios for the formants across both recordings are significantly correlated, whereas those for the cepstral analysis are not). However, both cepstral and formant analyses do agree in the status of the 75% landmark. This is the point at which the highest within- to between-speaker values occur both across recordings and across analyses. (For formants this is in terms of the sum of the F-ratios of the individual formants at a landmark; for the cepstra in terms of the highest whole-range value.) It is thus possible to say that the greatest between- to within-speaker long-term variation occurs at the same landmark (75%) in both the F-pattern and the C-pattern. Their discrimination performance was accordingly tested at the 75% landmark.

Table 2. F-ratios for cepstral and formant analysis.

	V	I			ou%		
R1 cep			0	25	50	75	100
F1-range	2	2	7	5	3	5	2
F2-range	7	5	4	3	5	16	7
F3-range	4	7	6	5	3	6	2
F4-range	4	6	6	6	4	4	2
Full-range	3	5	5	5	4	6	3
R2 cep							
F1-range	13	5	13	13	3	11	10
F2-range	10	14	12	9	9	13	10
F3-range	12	12	15	18	13	6	5
F4-range	5	4	5	7	15	15	7
Full-range	8	7	9	9	10	11	7
R1 form							
F1	4	3	8	9	3	3	8
F2	0	11	1	2	7	25	9
F3	4	6	7	5	4	6	2
F4	7	7	14	18	24	13	4
R2 form							
F1	5	2	7	9	2	7	9
F2	2	23	6	10	19	23	12
F3	10	15	13	11	13	11	10
F4	9	14	10	10	43	46	17

5. DISCRIMINATION ANALYSES

In forensic phonetic case-work, the emphasis is on discrimination between same-voice samples and different-voice samples. This differs somewhat from the conventional (identification) sense of discriminant analysis, which is concerned with assigning to a set of pre-established classes (here speakers) an unlabelled token observed in addition to those used to determine the classes (Woods, Fletcher & Hughes 1986: 266). Forensically, identification is the secondary result of a process of discrimination. If it is decided that two samples come from the same voice, the suspect is identified as the criminal. If not, no identification results. In this experiment, therefore, discrimination does not mean being able to identify individuals, but being able to say, given any pair of *hellos* from our data set, whether or not they come from the same speaker. In this experiment, we wanted to find out how much better a cepstral analysis can do this than a formant analysis.

Both cepstral and formant analyses in the preceding section showed that hello has the most individual-identifying information at the 75% landmark. Two tests were accordingly performed to compare the discriminant power of formant- and cepstrally-based analyses at the 75% landmark on same-speaker and different speaker pairs of hellos. The first test was carried out with the same-session data of the second recording. In this test, all possible within- and between-speaker pairs of *hellos* were tested. Thus, for example, DM's first hello token in his second recording was compared with all his other tokens in his second recording, and all other tokens from the second recording of all other speakers. In all, then, 210 within-speaker pairs of *hellos* were compared in the first test, and 1168 between-speaker pairs.

Although it quantifies the relative performance of the cepstral and formant analyses, this test is forensically unrealistic because it uses single-session data (Rose 1999b:1,2). Therefore a second, forensically more realistic, test was performed with the long-term different session data provided by recordings 1 and 2. In this test, the within-speaker comparison was, of course, across the two recording sessions. Thus, for example, all DM's hello tokens in his first recording were tested against all his hello tokens in his second recording, and against all the *hellos* of all other speakers in both recordings. The second test involved 376 within-speaker and 3688 between-speaker comparisons. The second test thus simulates a situation where a criminal and a suspect sample, separated by a long stretch of time, are being compared using one hello token in each sample. (In reality, of course, much more material in each sample would be compared, and usually the samples would be separated by a much shorter stretch of time.)

The tests are crude, and make use of nothing but unweighted distances between samples as thresholds. First, the mean between-speaker and within-speaker distances, and the mean standard deviation of the between-speaker and within-speaker standard deviations, were calculated for values at the 75% point. The discriminant threshold was then set at halfway between the between- and within-speaker mean values. Given the similar standard deviations observed with this procedure, this should ensure that values close to an EER should be obtained, assuming distributional normality. The EER was then found as the mean of the discriminant performances for the between- and within-speaker comparisons. Because the F-ratio values for F1 and F3 at 75% were not so high as for F2 and F4, performance was evaluated only for F2 and F4 in the formant analysis. We did not know what to expect for the cepstrum, so we evaluated the cepstral performance at all four formant ranges, as well as over the whole range.

6. RESULTS

Results are shown, as equal error percent correct performance, in table 3. Table 3 shows firstly that, as expected, performance decreases with the different session data. The best performance (79%) is clearly obtained by (whole-range) cepstral analysis for the same session data, but both analyses perform equally well, as far as best performances are concerned, for the different session data: the value for F4 (64%) is effectively the same as the 63% for the whole-range cepstrum. It is, of course, highly unlikely that F4 will be available for use in real forensic case-work, barring comparison with rhotics, so perhaps it is more realistic to draw comparisons with F2. Here the results are clearer. The cepstral analysis is 10% better than the formant in the same session data (79% vs. 69%), and 5% better (63% vs. 58%) for the different session data. While F4 is not fully admissible on grounds of availability or measurement unreliability, cepstral analyses spanning the entire Nyquist interval are not thus hampered, and can therefore be justifiably exploited to implicate the higher-formant range. It can be noted, moreover, that the F2 range cepstrum performance (62%) is still 4% better than the formant analysis with F2.

Table 3. Equal discrimination performance (%) of Cepstrum (C) and Formant (F) analyses at 75% of /oʊ/ in hello.

Formant/ Cepstral range	same session (R2)		different session (R1 & R2)	
	C	F	C	F
F1	59		56	
F2	70	69	62	58
F3	64		56	
F4	72	71	58	64
Full	79		63	

The fairly good agreement observable between the performance for the individual formants and that for the cepstral formant ranges is presumably because the former are the primary determinants of the spectral shape. However, it is also a nice indication that the cepstral sub-band analysis works. It is important to note that the fact that no sub-band analysis outperformed the full-range analysis does not automatically indicate the superiority of the latter. This is because we deliberately constrained the sub-bands to correspond to formant ranges. In the different-session comparison, the F2 sub-band (62%) contains effectively as much discriminating information as the whole spectral range (63%). It is therefore entirely possible that better performance might occur with unconstrained sub-bands than with the full range. If this is the case, an unconstrained sub-band cepstral discrimination might have the potential to outperform a formant discrimination by more than the demonstrated 5%.

7. SUMMARY, CONCLUSION AND WAY AHEAD

This paper has shown that, in the very restricted task of comparing samples at a single landmark in hello, the cepstrum does discriminate forensically realistic data better than formants, by at least 5%. Moreover, the overall good discrimination performance of the cepstrum at some other landmarks in hello (not demonstrated in this paper) indicates that it is less sensitive to different landmarks than the formant analysis. In forensic science, performance must outweigh understandability for juries. In addition, the practicality of the cepstrum in avoiding measurement problems that are inherent to formant-frequency estimation is also a criterion in its favour. Thus we conclude that spectral shape parameters like the LP-cepstral coefficients do have a more important role to play in forensic speaker identification than has been demonstrated to date. Whether this role is as an adjunct or as an alternative to the formants remains to be seen.

Further research is required into the effects of different recording conditions (e.g. telephone); the pre-treatment of cepstral coefficients; of sample size; and the use of more sophisticated discrimination strategies, including weighting; the involvement of more than one landmark; and unconstrained sub-bands. It will also be interesting to see whether the cepstrum produces a more homogeneous set of results with respect to individual speakers and speaker pairs. With formants, different-session within-speaker discrimination of RS is particularly bad, for example, and only offset by good performance with other speakers. Ideally, of course, all same-speaker hello pairs must be discriminable from different-speaker pairs. Ultimately, however, we will need to move away from the average probabilities of overall discrimination rates to the calculation of likelihood ratios for cepstral distances, so that the latter can be used within the appropriate Bayesian approach for forensic science (Champod & Meuwly 2000).

REFERENCES

Champod, C. & Meuwly, D. (2000) "The inference of identity in forensic speaker recognition." *Speech Communication* **31**, 193-203.

Clermont, F. & Mokhtari, P. (1994) "Frequency-band specification in cepstral distance computation." In Roberto Togneri (ed.) *Proc. Fifth Australian Internat. Conf. on Speech Science and Technology.* Canberra: ASSTA, 354-359.

van der Giet (1987) "Der Einsatz des Computers in der Sprechererkennung." In Künzel, H.J. *Sprechererkennung: Grundzüge forensischer Sprachverarbeitung.* Heidelberg: Kriminalistik Verlag, 121-132.

Künzel, H.J. (1995) "Field Procedures in forensic speaker recognition." In J.W. Lewis (ed.) *Studies in general and English phonetics. Essays in honour of J.D. O'Connor.* London: Routledge, 68-84.

Labov, William & Harris, Wendell A. (1990) "Addressing social issues through linguistic evidence." In John Gibbons (ed.) *Language and the Law.* New York: Longman, 265-305.

Laver, J. (1991) *The Gift of Speech.* Edinburgh: EUP.

Nolan, Francis (1990) "The limitations of auditory-phonetic speaker identification." In H.Kniffka (ed.) *Texte zur Theorie und Praxis forensischer Linguistik.* Tübingen: Niemayer.

Rabiner, L. & Juang, B-H.J. (1993), *Fundamentals of Speech Recognition.* Englewood Cliffs, N.J. Prentice-Hall.

Rose, P. (1999a) "Differences and distinguishability in the acoustic characteristics of *hello* in voices of similar-sounding speakers: a forensic phonetic investigation." *Australian Review of Applied Linguistics* **22**(1), 1-42.

Rose, P. (1999b) "Long- and short-term within-speaker differences in the formants of Australian *hello.*" *Journal of the International Phonetic Association* **29**(1), 1-31.

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

Woods, A. Fletcher, P. & Hughes, A. (1986) *Statistics in language studies.* Cambridge: CUP.