ABSTRACT
A pilot forensic-phonetic experiment is described which compares the performance of formant- and cepstrum-based analyses on forensically realistic speech. This novel approach was utilised by a novel formant selective cepstral analysis. It is shown that at the performance utilising a novel band-selective cepstral analysis outperforms the second diphthongal target in h ello the cepstrum forward analysis by about 5%, compared to its 10% superiority for same-session data.

INTRODUCTION
A recurrent topic in Forensic Phonetics, where speaker verification under very much less than full control is a major concern, is its relationship to Automatic Speaker Recognition (ASR). Leading forensic phoneticians (e.g., Künzel 1995: 79) in fact, the difference is 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 this is not a possibility.

This should not imply, however, that some of the analytical techniques common in ASR are of no use. Although forensic speaker identification must usually rely, inter alia, on no forensic use. Although forensic speaker identification does not sound in the comparison of individual formants (e.g., Nolan 1993; Labov & Harris 1990: 287ff.), it is required to use the cepstrum in forensic phonetics. This is because the analysis extracts ear immunity to "noninformation-bearing variables" (Rabiner & Lejung, 1993: 169) and, hence, greater sensitivity to distinctive features of speech spectra. This is achieved in the sense of the cepstrum as discussed above and practical to mathematical developments (Clermont & Mokhtari 1994), which make it possible to compute directly to the upper and lower bound of any frequency band directly in the computational distance of the cepstral.

The speech data we use are forensically realistic in many important ways. Firstly, they are from speakers that sound similar. This is an obvious requirement on any forensic analysis of presence of the first four formants identified, and transferred to the upper and lower bound of any frequency band directly in the possible to specify the upper and lower bound of any frequency band directly in the computation of the cepstral distance.

The speech data we use are forensically realistic in many important ways. Firstly, they are from speakers that sound similar. This is an obvious requirement on any forensic analysis of presence of the first four formants identified, and transferred to the upper and lower bound of any frequency band directly in the possible to specify the upper and lower bound of any frequency band directly in the computation of the cepstral distance.

The speech data we use are forensically realistic in many important ways. Firstly, they are from speakers that sound similar. This is an obvious requirement on any forensic analysis of presence of the first four formants identified, and transferred to the upper and lower bound of any frequency band directly in the possible to specify the upper and lower bound of any frequency band directly in the computation of the cepstral distance.

The speech data we use are forensically realistic in many important ways. Firstly, they are from speakers that sound similar. This is an obvious requirement on any forensic analysis of presence of the first four formants identified, and transferred to the upper and lower bound of any frequency band directly in the possible to specify the upper and lower bound of any frequency band directly in the computation of the cepstral distance.

The speech data we use are forensically realistic in many important ways. Firstly, they are from speakers that sound similar. This is an obvious requirement on any forensic analysis of presence of the first four formants identified, and transferred to the upper and lower bound of any frequency band directly in the possible to specify the upper and lower bound of any frequency band directly in the computation of the cepstral distance.

The speech data we use are forensically realistic in many important ways. Firstly, they are from speakers that sound similar. This is an obvious requirement on any forensic analysis of presence of the first four formants identified, and transferred to the upper and lower bound of any frequency band directly in the possible to specify the upper and lower bound of any frequency band directly in the computation of the cepstral distance.

The speech data we use are forensically realistic in many important ways. Firstly, they are from speakers that sound similar. This is an obvious requirement on any forensic analysis of presence of the first four formants identified, and transferred to the upper and lower bound of any frequency band directly in the possible to specify the upper and lower bound of any frequency band directly in the computation of the cepstral distance.

The speech data we use are forensically realistic in many important ways. Firstly, they are from speakers that sound similar. This is an obvious requirement on any forensic analysis of presence of the first four formants identified, and transferred to the upper and lower bound of any frequency band directly in the possible to specify the upper and lower bound of any frequency band directly in the computation of the cepstral distance.

The speech data we use are forensically realistic in many important ways. Firstly, they are from speakers that sound similar. This is an obvious requirement on any forensic analysis of presence of the first four formants identified, and transferred to the upper and lower bound of any frequency band directly in the possible to specify the upper and lower bound of any frequency band directly in the computation of the cepstral distance.

The speech data we use are forensically realistic in many important ways. Firstly, they are from speakers that sound similar. This is an obvious requirement on any forensic analysis of presence of the first four formants identified, and transferred to the upper and lower bound of any frequency band directly in the possible to specify the upper and lower bound of any frequency band directly in the computation of the cepstral distance.

The speech data we use are forensically realistic in many important ways. Firstly, they are from speakers that sound similar. This is an obvious requirement on any forensic analysis of presence of the first four formants identified, and transferred to the upper and lower bound of any frequency band directly in the possible to specify the upper and lower bound of any frequency band directly in the computation of the cepstral distance.

The speech data we use are forensically realistic in many important ways. Firstly, they are from speakers that sound similar. This is an obvious requirement on any forensic analysis of presence of the first four formants identified, and transferred to the upper and lower bound of any frequency band directly in the possible to specify the upper and lower bound of any frequency band directly in the computation of the cepstral distance.

The speech data we use are forensically realistic in many important ways. Firstly, they are from speakers that sound similar. This is an obvious requirement on any forensic analysis of presence of the first four formants identified, and transferred to the upper and lower bound of any frequency band directly in the possible to specify the upper and lower bound of any frequency band directly in the computation of the cepstral distance.

The speech data we use are forensically realistic in many important ways. Firstly, they are from speakers that sound similar. This is an obvious requirement on any forensic analysis of presence of the first four formants identified, and transferred to the upper and lower bound of any frequency band directly in the possible to specify the upper and lower bound of any frequency band directly in the computation of the cepstral distance.

The speech data we use are forensically realistic in many important ways. Firstly, they are from speakers that sound similar. This is an obvious requirement on any forensic analysis of presence of the first four formants identified, and transferred to the upper and lower bound of any frequency band directly in the possible to specify the upper and lower bound of any frequency band directly in the computation of the cepstral distance.

The speech data we use are forensically realistic in many important ways. Firstly, they are from speakers that sound similar. This is an obvious requirement on any forensic analysis of presence of the first four formants identified, and transferred to the upper and lower bound of any frequency band directly in the possible to specify the upper and lower bound of any frequency band directly in the computation of the cepstral distance.

The speech data we use are forensically realistic in many important ways. Firstly, they are from speakers that sound similar. This is an obvious requirement on any forensic analysis of presence of the first four formants identified, and transferred to the upper and lower bound of any frequency band directly in the possible to specify the upper and lower bound of any frequency band directly in the computation of the cepstral distance.

The speech data we use are forensically realistic in many important ways. Firstly, they are from speakers that sound similar. This is an obvious requirement on any forensic analysis of presence of the first four formants identified, and transferred to the upper and lower bound of any frequency band directly in the possible to specify the upper and lower bound of any frequency band directly in the computation of the cepstral distance.

The speech data we use are forensically realistic in many important ways. Firstly, they are from speakers that sound similar. This is an obvious requirement on any forensic analysis of presence of the first four formants identified, and transferred to the upper and lower bound of any frequency band directly in the possible to specify the upper and lower bound of any frequency band directly in the computation of the cepstral distance.
regions straddling the frequency range of each of the four observed formants were processed separately using input parameters to the BSD defined as follows. The upper bound for each region 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 spectral region thus processed by the BSD was specified in terms of an upper bound of (499 +17 n) 516 Hz and a lower bound of (405+30 n) 575 Hz.

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, Downtone, Fall-Rise and Rise-fall (Rose 1999b: 10). With the exception of JM, who produced proportionately more downtones, the between-speaker intonational variety was comparably 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 backrounding and rounding of the diphthongal offglide in /ou/. (Realisations of Australian /ou/ typically show a wide range in the backness of the diphthongal offglide). The /ou/ diphthongs in the data collected here have an offglide ranging between [y-] and [e- -u+] (and a fairly open central initial target [e]). They are thus representative of a major part of the typical range. Two speakers (PS and RS) consistently had what sounded like a backerrounder offglide: [u+]; DM's offglide was consistently fronter: [a], and JM's offglide sounded slightly fronter and lower: [e]. The other two speakers showed somewhat comparable variation. Some of EM's /ou/ tokens sounded the backer/more rounded, although not as much as PS and RS. MD was notable for his wide range of off-glide realisations, from [u+] through (t) to [y-]. Also noticeable were differences in the secondary articulation of /l/ (pharyngealised vs velarised), and incidental differences in the first vowel phomene /l/ vs. /l/ vs. /l/. 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.

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

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 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.

ANOVA COMPARISON

As a preliminary to the discrimination, in order to find out where the points of greatest within- to between-speaker variation in /helios/ lay, a single second analysis was undertaken. Table 2 gives the results for /helios/. 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 discrimination occurs 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.

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 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 helios 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.

Table 2: F-ratios for cepstral and formant analyses.

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 3: F-ratio analysis.

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>72</td>
<td>90</td>
<td>95</td>
<td>90</td>
</tr>
<tr>
<td>V2</td>
<td>80</td>
<td>90</td>
<td>95</td>
<td>90</td>
</tr>
<tr>
<td>I</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

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 test was conducted, of course, across the two recording sessions. Thus, for example, DM's first *hello* token 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 distances, 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 same 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 want to expect for the cepstrum, so we evaluated the cepstral performance at all four formant ranges, as well as over the whole range.

**RESULTS**

Results are shown, as equal error percent correct performance, in table 3. Table 3 shows first that, as expected, performance decreases with the different session data. The best performance (75%) 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 work, barring comparison with mockos, so perhaps it is more realistic to draw conclusions with F2. Here the results are closer. 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 cepstra. 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 justly 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>
<thead>
<tr>
<th>Formant/Cepstral range</th>
<th>same session (R2)</th>
<th>different session (F1 &amp; R2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>59</td>
<td>56</td>
</tr>
<tr>
<td>F2</td>
<td>70</td>
<td>62</td>
</tr>
<tr>
<td>F3</td>
<td>84</td>
<td>76</td>
</tr>
<tr>
<td>F4</td>
<td>72</td>
<td>51</td>
</tr>
<tr>
<td>Full</td>
<td>79</td>
<td>63</td>
</tr>
</tbody>
</table>

Finally, the good agreement observable between the performance for the individual formants and that for the cepstral range is presumably because the formant are the primary determinants of the spectral shape. However it is also a nice indication that the cepstral sub-band analysis works. Note also that the sub-band analysis enables us to see that in the different session comparison the F2 sub-band contains effectively as much discriminating information as the whole spectral range.

**DISCUSSION**

The results reported in this paper do indicate that spectral shape parameters, such as the LP-cephstral coefficients, may have a more important role to play in forensic speaker identification than has been demonstrated to date. In addition to the fact that such parameters do not pose the measurement problems that are inherent to formant-frequency estimation, the band-selected formulation of the cepstral distance has rendered more viable the task of searching for spectro-temporal regions that hold the potential of yielding more reliable discrimination. Moreover, the overall good performance of the cepstrum at some other landmarks (not demonstrated in this paper) suggests that it is less sensitive to different landmarks than the formant analysis. However, an unequivocal prescription for the use of the cepstrum, either as an add-on or as an alternative to the formants, must await further research into the effects of varying recording conditions (e.g. on telephone data); the pre-treatment of cepstral coefficients; the effects of sample size; and the use of more advanced discrimination strategies, including weighting, and the involvement of more than one landmark.

The overall discrimination performance is not the only relevant parameter for comparison between the two approaches, of course. It is also necessary to compare performance on 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. It will be interesting to see whether the cepstrum produces a more homogeneous set of results.

**REFERENCES**


