More is better: likelihood ratio-based forensic voice comparison with vocalic segmental cepstra frontends

Phil Rose

Abstract
The suitability of vowel cepstral spectra for forensic voice comparison is explored within a likelihood ratio-based framework, and non-technical explanations provided for some basic concepts of cepstral analysis and forensic voice comparison. Non-contemporaneous landline telephone recordings of 297 male Japanese speakers are compared using only two replicates per recording of each of their five read-out vowels. 14 cepstrally-mean-subtracted LPC cepstral coefficients modelling the spectral shape to 5 kHz are used as features. When evaluated intrinsically with kernel density multivariate likelihood ratios, all 297 same-speaker comparisons are correctly discriminated as coming from the same speaker, and only 173 of the 43,956 different-speaker comparisons (0.4%) are incorrectly evaluated as coming from the same speaker. The log-likelihood ratio cost for this comparison is very low at 0.013. Fusion with a speaker’s long-term spectral data marginally improves the different-speaker error rate to 0.27% and the log-likelihood ratio cost to 0.009. It is concluded that the approach warrants further examination.

KEYWORDS FORENSIC VOICE COMPARISON, LIKELIHOOD RATIO, VOWEL SPECTRA, CEPSTRUM

Affiliation
School of Culture, History and Language, College of Asia and the Pacific, Australian National University
email: philjohn.rose@gmail.com
Introduction

In forensic voice comparison (FVC) the expert typically compares suspect and offender speech samples to help the trier of fact decide whether the suspect said the incriminating speech. The evidence in FVC is the ensemble of observed differences between the suspect and offender speech samples, and the crucial concept is its strength, when quantified as a likelihood ratio (LR) (Rose 2006; Morrison 2010). This is the ratio of the probabilities of observing the evidence under the competing hypotheses – usually the prosecution hypothesis that the suspect said both speech samples, and the defence hypothesis that they were said by different speakers. The LR is crucial because it provides the logical link between the evidence and the prosecution and defence hypotheses, and the usefulness of an expert is considerably restricted if they are unable to say how likely the evidence is under both hypotheses. The LR has also been shown to be a powerful function in the essential testing, as is done in this article, of the discriminability of various forensic media, e.g. DNA (Evett, Scrange and Pinchin 1993), fingerprints (Neuman, Evett and Skerrett 2012; Morrison 2012), handwriting (Hepler, Saunders, Davis and Buscaglia 2012), SMS texts (Ishihara 2012) and speech (Gonzalez-Rodriguez, Rose, Tamos, Torre and Ortega-García 2007).

The strength of the evidence in a particular case, or equivalently the magnitude of the LR, depends on many factors, but ceteris paribus on the features used to compare the samples. Cepstral coefficients have long been the feature of choice in automatic speaker and speech recognition, and automatic forensic speaker recognition, where they are applied globally (Gonzalez-Rodriguez, Drygajlo, Ramos-Castro, Garcia-Gomar and Ortega-Garcia 2006). The aim of this article is to demonstrate how some basic signal-processing methods used in automatic speaker recognition can be used to enhance FVC, cf. Morrison (2011a). Specifically, it is to explore the potential of cepstral coefficients when used locally, to characterise segmental phones such as vowels: this is what is meant by the term segmental cepstrum. Previous work on the use of the segmental cepstrum in LR-based forensic voice comparison focused on the voiceless palatal fricative [ɕ] and nasal consonants [m n N] (Rose, Osanai and Kinoshita 2003; Rose 2011a; Yim and Rose 2012), and on vowels in the fore-runner to this article (Rose 2011b).

Cepstral coefficients

Cepstral coefficients represent a way of optimally parametrising a spectrum for speech and speaker recognition. They also have computationally useful mathematical properties.
Their nature can be demonstrated with the following example. The left-hand panel of Figure 1 shows a spectrogram of a stressed vowel spoken by a young Australian female during a map task designed to elicit unedited but controlled speech data. She was saying ‘along Erskinville (lane)’ and the vowel occurred at the beginning of the word Erskinville. It has a mid front rounded quality. Formant centre-frequencies have been extracted and superimposed on the vowel spectrogram (the Burg option in Praat was used, with a setting of five formants below 6 kHz). A spectrogram of the utterance in which the vowel occurred is shown on the right, where it can be seen from about csec. 35 to csec. 50.

The initial portion of the vowel, from c. csec. 0.5 to csec. 5.5, shows non-modal phonation. Although it looks like creak (which has a deliberate indexical value for some young Australian females), it is not audible as such and is probably best interpreted as the realisation of a glottal stop associated with the juncture between the word final /ŋ/ in along and the word-initial vowel in Erskinville. The F-pattern is effectively steady-state for most of the duration of the modally phonated portion of the vowel, between c. csec. 4.5 and csec. 10.5, with F1, F2 and F3 at c. 600 Hz, 1.7 kHz and 2.75 kHz respectively. The F-pattern over the last couple of centiseconds of vowel duration shows expected perturbations induced by the alveolar place of the following consonant /s/.

Assuming that we would want to quantify the acoustics of this speaker’s vowel in a forensic voice comparison, it would be normal to do this by estimating its F-pattern centre-frequencies, and the best way of doing this, given its quasi steady-state nature, would be to estimate its mean F-pattern centre-frequencies over the putative steady-state portion. This is easily done: estimates using Praat for example are: 560 Hz (F1), 1.73 kHz (F2) and 2.617
kHz (F3). The value of 3.787 kHz for F4 would not normally be included, as it would be above the telephone bandpass encountered in many forensic voice comparison cases.

It is known, however, that formant centre-frequencies constitute only a part of the acoustic information associated with the vowel relevant for speech recognition (Zahorian and Jagharghi 1993) and speaker recognition (Rose and Clermont 2001; Rose et al. 2003). This can be appreciated if we look at the vowel’s mean spectrum, in Figure 2. Figure 2 shows the mean LPC spectrum calculated over the c. 6 csec. of modally phonated steady-state, together with the corresponding FFT spectrum, to 5 kHz. The non-modally phonated part was excluded, as the spectral slope is expected to be less for creak, and combining them would therefore produce a meaningless result. A 12th order LPC was applied, using the rule-of-thumb of ‘two plus twice the expected formant number’, here, five formants below 5 kHz. The first four formant centre-frequencies are very clear, but there are also other features contributing to the spectral shape, which contain speaker-specific information – the formant bandwidths, for example (Gonzalez-Rodriguez 2011), or the spectral slope. It would clearly be better to be able to make use of more information from spectral shape parameters than just the formant centre-frequencies, and automatic speaker-recognition experimentation, the locus classicus for which is Furui (1981), has shown that one way of doing this is to parametrise the
spectral shape with a set of cepstral coefficients – now the algorithmic mainstay of the automatic approach.

Cepstral coefficients allow the parametrising of a spectrum with a greater degree of smoothing than with linear prediction, and therein lies their power, since the smoothing generally exhibits ‘strong immunity to non-information variabilities in the speech spectrum’ (Rabiner and Juang 1993) and thus turns out to be optimum for speech and speaker recognition. The LP spectrum requires smoothing because its peakyness, especially in the lower formants signalling linguistic rather than individual information, means that small differences in frequency between two otherwise similar tokens can lead to large overall spectral differences (Furui and Akagi 1985: 8). The smoothing can be adjusted to be just enough to preserve the differences between spectra that are important for distinguishing individuals, or speech sounds, but just enough also to eliminate the differences that will tend to hamper such differentiation (Furui and Akagi 1985: 7). Figure 3 illustrates the vowel’s cepstrally smoothed linear prediction spectrum, referred to below as a cepstrally smoothed spectrum (CS spectrum). As its name implies, this spectrum is the result of a cepstral signal-processing operation on a linear prediction spectrum of the type illustrated in Figure 2 (effectively it is like taking the spectrum of a spectrum). Figure 3 shows both LP and CS spectra for the female speaker’s vowel, where the smoothing of the CS spectrum relative to the LP spectrum is clear.

Figure 3: LPC and cepstrally smoothed spectra of the steady-state portion of the vowel in Figure 1.
The CS spectrum is built up from more simple spectral components. However, unlike radiated/LPC vowel spectra, these components are not the resonant curves of individual formants. Rather, the CS spectrum is constituted from a set of much more simple individual spectral components, each determined by a separate cepstral coefficient (Rose 2003: 408ff.). The number of individual spectra used (equivalently, the number of cepstral coefficients) depends on several factors, one of which is how well one wants to approximate the details of the actual spectrum. A typical number of CCs is between 12 and 14. The CS spectrum in Figure 3 was derived from 12 CCs. They are shown in Table 1.

**Table 1: First 12 LP cepstral coefficient values (to 3 significant figures) for the CS spectrum in the vowel in Figure 1. k = coefficient index, C = coefficient value.**

<table>
<thead>
<tr>
<th>k</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>1.308</td>
<td>-0.194</td>
<td>0.295</td>
<td>0.073</td>
<td>0.331</td>
<td>-0.116</td>
<td>-0.090</td>
<td>-0.054</td>
<td>-0.595</td>
<td>-0.208</td>
<td>0.133</td>
<td>0.011</td>
</tr>
</tbody>
</table>

Figure 4 shows how a CS spectrum is built up from its individual cepstral spectral components. The top left panel shows the full CS spectrum, i.e. from all 12 CCs. To its right is the cepstral spectrum of just the first cepstral coefficient (CC1). As can be seen, its spectral amplitude is a very simple sinusoidal curve, symmetrical around 0 dB. This is because the curves are based on the trigonometrical cosine function: the spectrum of CC1, for example, represents a half of a cosine curve. The cepstral spectral amplitude \( H \) as a function of frequency \( f \) is given by:

\[
H(f) = C \cdot \cos\left(\frac{\pi}{R} \cdot f \cdot k\right) \cdot 10
\]

where \( C \) is the cepstral coefficient value, \( k \) its index, and \( R \) the frequency range of the spectrum (here, 5 kHz). Thus, for example, the cepstral spectral amplitude of CC1 at 1000 Hz is:

\[
= 1.308 \text{ (the value of the first cepstral coefficient)} \cdot \cos\left(\frac{3.142}{5000} \cdot 1000 \cdot 1\right) \cdot 10 \\
= 1.308 \cdot \cos\left(6.284 \cdot 0.004 \cdot 1000 \cdot 1\right) \cdot 10 \\
= 1.308 \cdot \cos\left(0.6284\right) \cdot 10 \\
= 1.308 \cdot 0.809 \cdot 10 \\
= 10.6\ldots \text{ dB}
\]

It can be seen in Figure 4 that the amplitude of the CC1 spectrum is indeed about 10 dB at 1 kHz. It is also easy to see from the shape of the CC1 spectrum...
why it is sometimes taken to represent spectral slope (Furui and Akagi 1985: 8).

The bottom left panel of Figure 4 shows the cepstral spectrum for the ninth cepstral coefficient (CC9), the value of which was \(-0.5953\). As it is the ninth coefficient it shows 4.5 repeats of a cosine, and as the CC’s value is negative its initial value is negative. It can be seen that CC9 contributes considerably to the four spectral peaks of the full spectrum. This contribution is reflected in the magnitude of CC9 in Table 1, which in this token is the second greatest of all 12 CCs. The amplitude of CC9 at 1 kHz is \((-0.5953 \times \cos(0.628319 \times 9) \times 10 = \) \(-4.81\) dB, and Figure 4 shows that the amplitude of CC9 is indeed about \(-5\) dB at 1 kHz.

The bottom right panel of Figure 4 shows how the spectra of CCs are combined to obtain a closer approximation of the full spectral shape. The combination is easy: simply the sum of the spectra, or equivalently the spectrum from the summed individual CCs. In Figure 4 the spectrum of CC1 is shown combined with that of CC9, and it can be seen that the result, with the CC9 spectrum appearing to sit on top of the CC1 spectrum, is actually quite close to the full spectrum. The amplitude of this spectral combination at 1 kHz is simply the sum of the amplitudes of the individual components. These were calculated above to be 10.58 dB and \(-4.81\) dB for CC1 and CC9 respectively, thus the value of the spectrum of CC1 and CC9 at 1 kHz is \((10.58 + \(-4.81\) =) 5.77 dB.

Figure 4: Cepstral spectra components of the vowel in Figure 1. Dotted line shows spectrum from all 12 CCs.
The idea behind cepstral spectra is thus that the complicated shape of a radiated speech spectrum can be represented by a set of cosinusoidal components, each defined by their cepstral coefficients, which when added together reconstitute the complicated spectral shape to any desired degree. It is the set of cepstral coefficients that can be used as features for comparing forensic speech samples.

The main forensic advantage of the cepstrum is its power. For example, Rose et al. (2003: 192) showed that for the same data it was capable of delivering about five times stronger evidence, i.e. LRs of greater magnitude, than formant centre-frequencies alone. This greater power is probably simply because, in capturing the whole of the spectral envelope, there is potentially more information in a cepstral than a formant comparison. There is therefore a greater chance of picking up important similarities or differences between samples with the cepstrum. For example, the cepstral-spectral envelope of a vowel can be expected to reflect not only the dimensions of the tract that produced it (in its F-pattern centre-frequencies – to the extent they are resolved), but also aspects of the phonatory activity of the source (in its spectral slope), and the tract compliance (in its formant bandwidths). (Since the CS spectrum is derived from an all-pole LPC, it will not model the effect of zeros from nasalisation or subglottal coupling.)

A second advantage over formants is that cepstral coefficients are so much easier to extract: all one has to do is determine the start and end points of the portion to be analysed. Formants, including forensically important sub-glottal and nasal resonances, can be notoriously difficult to identify and extract in the often degraded speech encountered in real casework. The vowel data analysed in this article are a case in point. They are in telephone recordings, each said on a relatively constant pitch typical for reading out, and the group is said on a higher than normal pitch associated with utterance-initial position, both factors thus making it difficult for the extraction algorithm to distinguish harmonic from formant frequencies. Only an optimistic masochist would attempt to extract them. This one gave up after trying just 10 of the 297 speakers who sounded non-nasal enough to be tractable.

The major drawback of the cepstrum is its sensitivity to transmission effects (compared, say, with vocalic F2). It will be appreciated from the demonstration above that each cepstral coefficient contributes to the spectrum of the whole of the frequency range. When the shape of the spectrum is perturbed, as for example in telephone transmission, all of the cepstral coefficients will change. Another often-cited drawback is its general lack of interpretability in terms of speech production. Most of the cosinusoidal components which are added together to constitute the final smoothed spectral shape do not correspond to
anything real in speech acoustics or production. They are simply abstract mathematical constructs that, when combined, give the best approximation to the shape. Exceptions are the first cepstral coefficient, with its relationship to spectral slope, and the second and third coefficients, which relate to vowel height and backness respectively (Clermont and Itahashi 2000). It is in this sense that they are automatic parameters. This is unlike the traditional features of formant centre-frequencies, the lower of which of course can be related to aspects of speech production, acoustics and perception.

A final, but important point to be made in justifying CCs is that it is not a case of supplanting formants, but of selecting the features that are likely to give the best strength of evidence under the real-world circumstances of the case (for there is little point in spending time quantifying a feature that is likely to return a LR close to 1, i.e. that is equally likely under either hypothesis). The LR-based approach, where combination of heterogeneous features is very easy, means that one can mix and match: formants for this easy vowel, perhaps, and CCs for this difficult nasal segment. One might even use both formants and CCs for the same segment (cf. Holmes, Holmes and Garner 1997), or make use of band-limited cepstra (Clermont and Mokhtari 1994; Khodai-Joopari, Clermont and Barlow 2004) to partition the frequency range to comprise, say, the first two formants of a vowel, and CCs for the rest of its spectrum. As explained below, any within-segment correlation in cases like these is easily handled by a multivariate likelihood ratio.

**Procedure**

**Database, speakers, corpus**

The recordings used in this experiment are part of a database collected some time ago by the Japanese National Research Institute of Police Science (NRIPS) for forensic speaker recognition tests (Osanai, Tanimoto, Kido and Suzuki 1995). Its main benefit is its size, which allows testing with a reasonably large number of speakers. It comprises recordings, digitised at 10 kHz with 12 bit quantization, of 297 adult male Japanese from 11 different prefectures around Japan. This means nearly 300 same-speaker comparisons and 44,000 different-speaker comparisons can be made. All speakers were members of the Japanese police force and were uncontrolled for age, which ranged from c. 20 to 50 years. The recordings were made centrally, on the same NRIPS equipment, of incoming landline telephone calls phoned-in from all over Japan. Most importantly for realistic forensic testing, two non-contemporaneous recordings were made for each speaker, separated by a rather large period of three to four months. Each recording for each speaker contains about 70–80 seconds net speech
comprising single- and many-word utterances, and a set of five Japanese vowels read out from hiragana いえあおう representing the five Standard Japanese vowel phonemes /i e a o ɯ/. It is these vowel readings that were used in the experiment. In each recording the speaker repeated all the data, giving just two replicates of each vowel per non-contemporaneous recording session. The small replicate number is emphasised because, since the strength of the evidence/LR magnitude is a function of the number of replicates in each sample, normally one would want to be able to compare more than just two replicates in suspect and offender samples.

Each token of each vowel was quantified with 14 LPC CCs after 98% preemphasis. Table 2 shows the raw input data for the first two speakers’ /i/ vowels (two per recording) to give an idea of the nature of the information that was used in the experiment. It can be seen that each speaker’s /i/ vowel is represented by four sets of 14 cepstral coefficients (C1–C14): two sets are from one recording and two from the second, non-contemporaneous recording. The organisation of data for the four other vowels /e a o ɯ/ was the same. Eyeballing the data row-wise – CC1 for example, or CC3 – it can be seen that generally the CC values are similar within speakers and different between speakers.

The CCs used in this experiment had already been extracted from the
speakers’ vowels in a forensically motivated PhD thesis (Khodai-Joopari 2006). Mel-frequency cepstral coefficients, which model a speech spectrum based on a mel frequency instead of a linear frequency scale and are now the preferred automatic feature, were not used, as it was found that for all vowels except /a/ they did not perform as well as the LPC CCs. In order to extract representative CCs for the vowels’ spectra, Khodai-Joopari first identified a given vowel token’s best continuous interval. This contained all the speech waveform within 10% of the maximum amplitude of the vowel, and, by visual examination, no extraneous material (e.g. non-speech transients). From the 14th order autocorrelation LP of the vowel’s best continuous interval, four consecutive single 25.6 ms frames were then identified such that a single frame could be chosen from them which simultaneously 1) best represented the whole of the best continuous interval, and 2) minimised the variance across a speaker’s four vowel tokens. The CCs from this single frame were then used to characterise the vowel token in question. In this way the within-speaker spectral variability was claimed to be minimised. This final part of the procedure was of course forensically unrealistic, because it forced the two non-contemporaneous recordings for each speaker to be as similar as possible. In reality, one does not of course know whether suspect and offender speech samples are from the same speaker or not, and one can imagine what defence counsel would have to say if it transpired that the forensic voice comparison expert had tried to make them as similar as possible.

Cepstral mean subtraction

As mentioned above, the speech spectrum can be considerably affected by telephone transmission, and it is customary in automatic speaker and speech recognition to implement some means of compensating for the distortion. One way is with cepstral mean subtraction (also called cepstral mean normalisation). The idea behind cepstral mean subtraction is explained by Garcia and Mammone (1999) as follows. When someone talks over the phone the acoustics of what they say are filtered by the telephone transmission, so that what emerges – what we have to analyse in forensic voice comparison – is a corrupted version of the original. If we think of the speech signals and the filter as represented spectrally by a temporal sequence of cepstral vectors $\text{CEP}$ (one such vector was shown in Table 1), the filtering effect of the phone transmission $\text{CEP}_{\text{transmission}}$ on the original signal $\text{CEP}_{\text{original}}$ which results in the phone signal $\text{CEP}_{\text{phone}}$ is very simple. In the cepstral domain, the set of cepstral vectors of the corrupted signal $\text{CEP}_{\text{phone}}$ is the sum of the cepstral vectors of the original signal $\text{CEP}_{\text{original}}$ and the cepstral vectors of the transmission filter $\text{CEP}_{\text{transmission}}$. 
where CEP = a sequence of cepstral vectors.

Thinking of the spectrum of the signals and the filter as represented by the mean of the sequence of cepstral coefficient vectors ĆEP, this becomes:

\[
ČEP_{\text{phone}} = ĆEP_{\text{original}} + ĆEP_{\text{transmission}}
\]

Now, if the distribution and variety of [speech] sounds in the clean signal [i.e. \(ČEP_{\text{original}}\)] is such that the average spectrum of the utterance is relatively flat, the cepstral mean vector of the original signal \(ČEP_{\text{original}}\) will approximate 0, and the mean cepstral spectrum of the phone signal \(ČEP_{\text{phone}}\) will be the same as the mean cepstral spectrum of the transmission \(ČEP_{\text{transmission}}\) (under the second major assumption that the transmission channel remains invariant over the duration of the signal). Thus:

\[
ČEP_{\text{phone}} = 0 + ĆEP_{\text{transmission}} = ĆEP_{\text{transmission}}
\]

Now think of a segment in the phone recording, such as [i], represented by its cepstral vector \(ČEP_{\text{segment}}\). If the mean cepstrum of the corrupted signal \(ČEP_{\text{phone}}\) is subtracted from the mean cepstral vector representing the segment \(ČEP_{\text{segment}}\) to get a mean cepstrally subtracted representation \(ČEP_{\text{cms \ segment}}\) for the segment, that is equivalent to subtracting the effect of the channel \(ČEP_{\text{transmission}}\), thus removing the effect of the phone transmission on the segment:

\[
ČEP_{\text{cms \ segment}} = ĆEP_{\text{segment}} - ĆEP_{\text{phone}}
\]

Of course, the major assumption of a flat spectrum for the original speech signal \(ČEP_{\text{original}}\) is unrealistic – one expects, for example, a nominal spectral drop-off of \(-12\) dB per octave due to radiation and spectral slope, although this can be offset by pre-emphasis. Moreover, the channel invariance assumption is probably only justified for landline transmission; in mobile phones the bit rate can and does change very quickly to respond to the current network load, and measures to counter lost and corrupted packets can also have an effect (Guillemin and Watson 2008). Nevertheless, as will be shown in this article, cepstral mean subtraction can be very effective.

To illustrate cepstral mean subtraction, we take a clean signal, pass it through
a landline telephone, perform cepstral mean subtraction and compare the cepstral mean subtracted signal with the original. As original signal we take the mean spectrum of [ʃ] in the word *shape* in the well-known *rainbow* passage, recorded in a sound-proof phonetics recording studio by a young Australian male in a suite of experiments investigating the effect of telephone transmission on formants (Rose 2003: 1501ff.). From the wide-band spectrograms of *the shape* in Figure 5 a typical landline telephone bandlimiting effect can be clearly seen, at least in the upper frequencies, where there is little or no visible energy above c. 3.5 kHz: gone are F4 in /ɛɪ/ and the palatal channel resonance in the [ʃ]; and the extraction would have to perform well to get a coherent centre-frequency out of the true /ɛɪ/ F3. The low frequency bandlimiting can be seen in the higher values for the diphthongal offglide F1.

The telephone effect on the spectral properties of the /ʃ/ and the degree to which it is compensated by cepstral mean subtraction are shown in Figure 6. It can be seen in the left panel of Figure 6 that there is a large difference between the CS spectra of the [ʃ] in the original recording and after it has been transmitted over the phone. The insert shows how the cepstrally mean subtracted spectrum of [ʃ] is obtained by subtracting the channel CS spectrum estimate from the phone spectrum. For example, at 3 kHz the spectral amplitude of the phone [ʃ] and the mean spectrum are about 4.5 and 0.5 units respectively. Subtracting the channel from the phone amplitude gives an amplitude for the cepstrally mean subtracted spectrum of about 4 units. In the right panel of Figure 6 the cepstrally mean subtracted CS spectrum is compared with the original spectrum, and it can be seen that the fit is better than between the phone and the original in the left panel.

The result with individual segments is not always as good as this. Inspection of individual cases shows that sometimes it makes no difference, and in some cases it can actually *increase* the difference between the original and the cms CS spectrum. Usually, however – and this will be shown below – cms compensates for the channel enough to considerably improve on the results achieved without the compensation.

A set of cepstrally mean subtracted CCs (cms CCs) was obtained for each vowel by subtracting each speaker’s mean cepstral vector obtained from all their speech in a repeat from the vowel’s CCs. Figure 7 shows an example from speaker 52’s [i] vowels. A Praat wideband spectrogram of his first [i] token, with superimposed formant centre-frequencies (*Burg*, 5 formants below 5 kHz), shows the typical diffuse-acute F-pattern for [i]. F1 is at c. 350 Hz, F2 at just over 2 kHz, a weak F3 appears to meander around 3.3 kHz, and F4 is at about 3.8 kHz. A weak pole is present at just over 1 kHz over the first few centiseconds, without the extraction of which F2 will be estimated too low. The
Figure 5: Wide-band spectrograms of the shape before (left) and after (right) transmission over a landline phone. The [ʃ] occurs between c. csec.6 and csec.19.

Figure 6: Cepstral mean subtraction illustrated with [ʃ]. CS spectrum of original [ʃ] (solid line) compared in left panel to CS spectrum of phone [ʃ] (dotted line), and in right panel to cepstrally mean subtracted CS spectrum of [ʃ] (dashed line). Insert shows derivation of cepstrally mean subtracted CS spectrum (dashed line) by subtracting mean cepstral spectrum for the whole recording (thick line) from the phone [ʃ] (dotted line).
peaks of the raw cepstral spectra for speaker 52’s non-contemporaneous [i] means (shown with solid lines in the bottom left panel) correspond fairly well with the spectrographic F-pattern, but it must be remembered that their spectral shape will be due in part to the transmission effect, which will be contained in the speaker’s mean cepstral spectra over their whole recordings shown in the dotted lines in the bottom left panel. The most obvious feature is of course the upper passband appearing to cut in at just below 4 kHz and truncating the mean cepstral spectrum and of course also the vowel spectrum.

Figure 7: Illustration of cepstral mean subtraction in [i]. Top: wideband spectrogram of [i]. Bottom left panel: solid lines = 12th order LPC cepstral spectra of phone-recorded [i]; dashed lines = speaker’s mean cepstral spectra. Right panel: cepstrally mean subtracted spectra for [i]. The paired lines in each panel show means for both non-contemporaneous recordings of [i].
In subtracting the mean cepstrum from the [i] CCs one hopes at least partially to deconvolve the original signal from the channel: the result is shown in the cms CS spectra for the two non-contemporaneous means in the bottom right panel. Some changes in the vowels’ overall spectral shape can be seen, notably in the decreased spectral slope (because of the falling slope of the mean spectrum). Note that the difference between the cms spectra appears to be actually a little bigger than between the raw spectra, but it must be remembered that the LR depends not just on the absolute difference between the spectra, but on the difference when evaluated against the between- and within-speaker variances. For the sake of interest, the log_{10}LR for this same-speaker comparison after cepstral mean subtraction was 2.76, showing a modest increase on the raw log_{10}LR value of 2.47: the difference between the two mean cms CS cepstra in the bottom right panel would be c. 575 times more likely assuming same rather than different speaker provenance, compared to c. 295 times with the raw spectra in the left panel.

**Likelihood ratio estimation**

LRs for separate vowels were estimated from both raw and cms CCs with the generative multivariate likelihood ratio formula (MVLR) developed at Edinburgh University’s Joseph Bell Centre for Forensic Statistics and Legal Reasoning (Aitken and Lucy 2004). This formula estimates multivariate LRs (MVLRs) taking into account any correlation between variables within a segment. Although the variables in this case are CCs, which are orthogonal by definition, there is always the chance that correlation will arise by virtue of the spectral shape of the actual sound being modelled, as was shown for [ɕ] in Rose et al. (2003: 194). The distribution of the reference sample used to estimate the probability of observing the difference between the suspect and offender assuming they have come from different speakers is modelled with a kernel density (the formula then sometimes being referred to as MVKD). The distributions of the suspect and offender are modelled normally.

The numerator and denominator of the MVKD formula are shown at (6) and (7). The multivariate likelihood ratio is their ratio. Several parts of the formula are worth pointing out. Firstly, there are the mean vectors of the suspect and offender ($\bar{y}_1, \bar{y}_2$). These contain the mean values for all of the CCs for a suspect or offender’s vowel. Table 3 shows an example using speaker 1’s /i/ CCs data from Table 2. In this example the suspect’s mean vector is the vector of the means of the 14 /i/ CCs for speaker 1’s first recording, and the offender’s mean vector is the vector of the means of the 14 /i/ CCs for his second recording. The evidence is the difference between these two vectors: if one is more
likely to get this set of differences assuming the suspect and offender samples have come from the same speaker (which they have, namely speaker 1) the log_{10} MVLR will be bigger than 0 (i.e. positive). If one is more likely to get this set of differences assuming the suspect and offender samples have come from different speakers, the log_{10} MVLR will be smaller than 0 (i.e. negative).

In order to estimate the probability of getting the difference between the mean vectors under the prosecution hypothesis that they have come from the same speaker, one needs to know the sampling distribution of the difference between means, which can be estimated from the variation found within speakers. Since multivariate data are used, this estimate actually takes the shape of a variance-covariance matrix (this is the term $U$ in the formula). Table 4 shows what it looks like using a portion of the within-speaker variance-covariance matrix $U$ for /i/ cms CCs (only the first 6 CCs are shown). The diagonal elements, in bold, are the variances for the individual CCs; the off-diagonal elements are the covariances between the different CCs: these denote how much one CC varies together with another, and are used to take the correlation between the individual CCs into account. It can be seen that the matrix is symmetrical around its diagonal, since covariance is commutative. It can also be seen that generally the covariance magnitudes are very small, indicating little correlation between the individual CCs.

Table 3: Example of mean vectors of suspect and offender CCs (speaker 1 /i/)

<table>
<thead>
<tr>
<th>CC</th>
<th>suspect</th>
<th>offender</th>
<th>difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>0.302</td>
<td>0.464</td>
<td>-0.162</td>
</tr>
<tr>
<td>C2</td>
<td>-0.741</td>
<td>-0.496</td>
<td>-0.245</td>
</tr>
<tr>
<td>C3</td>
<td>0.814</td>
<td>0.871</td>
<td>-0.057</td>
</tr>
<tr>
<td>C4</td>
<td>0.326</td>
<td>0.198</td>
<td>0.128</td>
</tr>
<tr>
<td>C5</td>
<td>0.390</td>
<td>0.567</td>
<td>-0.177</td>
</tr>
<tr>
<td>C6</td>
<td>0.295</td>
<td>0.156</td>
<td>0.139</td>
</tr>
<tr>
<td>C7</td>
<td>-0.269</td>
<td>-0.275</td>
<td>0.006</td>
</tr>
<tr>
<td>C8</td>
<td>-0.393</td>
<td>-0.091</td>
<td>-0.302</td>
</tr>
<tr>
<td>C9</td>
<td>0.297</td>
<td>0.152</td>
<td>0.145</td>
</tr>
<tr>
<td>C10</td>
<td>-0.010</td>
<td>0.015</td>
<td>-0.025</td>
</tr>
<tr>
<td>C11</td>
<td>-0.200</td>
<td>-0.146</td>
<td>-0.054</td>
</tr>
<tr>
<td>C12</td>
<td>-0.143</td>
<td>-0.189</td>
<td>0.045</td>
</tr>
<tr>
<td>C13</td>
<td>-0.095</td>
<td>-0.049</td>
<td>-0.045</td>
</tr>
<tr>
<td>C14</td>
<td>0.037</td>
<td>0.054</td>
<td>-0.017</td>
</tr>
</tbody>
</table>
Table 4: Portion of within-speaker variance-covariance matrix $U$ for /i/. Bold = diagonal elements = variance.

<table>
<thead>
<tr>
<th></th>
<th>CC1</th>
<th>CC2</th>
<th>CC3</th>
<th>CC4</th>
<th>CC5</th>
<th>CC6</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC1</td>
<td>0.03777</td>
<td>0.00639</td>
<td>-0.00880</td>
<td>0.00378</td>
<td>-0.00225</td>
<td>-0.00120</td>
</tr>
<tr>
<td>CC2</td>
<td>0.00639</td>
<td>0.02973</td>
<td>-0.00163</td>
<td>-0.00247</td>
<td>0.00169</td>
<td>-0.00290</td>
</tr>
<tr>
<td>CC3</td>
<td>-0.00880</td>
<td>-0.00163</td>
<td>0.01949</td>
<td>-0.00192</td>
<td>-0.00242</td>
<td>0.00008</td>
</tr>
<tr>
<td>CC4</td>
<td>0.00378</td>
<td>-0.00247</td>
<td>-0.00192</td>
<td>0.01345</td>
<td>-0.00106</td>
<td>-0.00165</td>
</tr>
<tr>
<td>CC5</td>
<td>-0.00225</td>
<td>0.00169</td>
<td>-0.00242</td>
<td>-0.00106</td>
<td>0.00926</td>
<td>-0.00041</td>
</tr>
<tr>
<td>CC6</td>
<td>-0.00120</td>
<td>-0.00290</td>
<td>0.00008</td>
<td>-0.00165</td>
<td>-0.00041</td>
<td>0.00917</td>
</tr>
</tbody>
</table>

\[(6)\]

numerator of kernel density MVLR =

\[
(2\pi)^{\frac{p}{2}}|D_1|^{1/2}|D_2|^{1/2}|C|^{1/2}(mh^p)^{1/2}D_1^{-1} + D_2^{-1} + (h^2C)^{-1}\]^{1/2} \\
\times \exp\left\{\frac{1}{2}\left(\mathbf{y}_1 - \mathbf{y}_2\right)^T(D_1 + D_2)^{-1}\left(\mathbf{y}_1 - \mathbf{y}_2\right)\right\} \\
\times \sum_{i=1}^{m} \exp\left\{-\frac{1}{2}\left(y_i^* - \bar{y}_i\right)^T\left(D_1^{-1} + D_2^{-1}\right)^{-1}\left(h^2C\right)^{-1}\left(y_i^* - \bar{y}_i\right)\right\}
\]

\[(7)\]

denominator of kernel density MVLR =

\[
(2\pi)^{p/2}|C|^{1/2}(mh^p)^{1/2}\prod_{i=1}^{m} |D_i|^{1/2}|D_1 + C^{-1}D_1^{-1} + h^2C|^{1/2} \\
\times \sum_{i=1}^{m} \exp\left\{-\frac{1}{2}\left(y_i - \bar{y}_i\right)^T\left(D_1 + h^2C\right)^{-1}\left(y_i - \bar{y}_i\right)\right\}
\]

where $U, C =$ within-, between-speaker variance/covariance matrices

$n_1, n_2 =$ number of replicates per speaker

$m =$ number of speakers in reference population

$p =$ number of assumed correlated variables per speaker

$D_1 = D_1, D_2 =$ offender, suspect variance /covariance matrices = $n_1^{-1}U, n_2^{-1}U$

$h =$ optimal smoothing parameter for kernel density = $\frac{4}{(2p + 1)}^{1/(p+4)} m^{-1/(p + 4)}$

$\bar{y}_i = \bar{y}_1, \bar{y}_2 =$ offender, suspect mean vectors

$y_i^* = \left(D_1^{-1} + D_2^{-1}\right)^{-1}\left(D_1^{-1}\bar{y}_1 + D_2^{-1}\bar{y}_2\right)$

$\bar{y}_i =$ within-speaker means of reference population.

In order to estimate the probability of getting the difference between the mean vectors under the defence hypothesis that they have come from different speakers, the MVLR formula uses an estimate of the between-speaker variation denoted $C$. Table 5 shows the portion of $C$ for /i/ for the first 6 CCs. It can be seen that the between-speaker variances for the individual CCs in $C$ in Table 5
Table 5: Portion of between-speaker variance-covariance matrix $C$ for /i/. Bold = diagonal elements = variance.

<table>
<thead>
<tr>
<th></th>
<th>CC1</th>
<th>CC2</th>
<th>CC3</th>
<th>CC4</th>
<th>CC5</th>
<th>CC6</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC1</td>
<td>0.05447</td>
<td>0.01134</td>
<td>-0.01487</td>
<td>0.00868</td>
<td>-0.00268</td>
<td>0.00004</td>
</tr>
<tr>
<td>CC2</td>
<td>0.01134</td>
<td>0.03823</td>
<td>-0.00358</td>
<td>-0.00552</td>
<td>0.00684</td>
<td>-0.00769</td>
</tr>
<tr>
<td>CC3</td>
<td>-0.01487</td>
<td>-0.00358</td>
<td>0.04187</td>
<td>-0.00574</td>
<td>-0.00873</td>
<td>0.00614</td>
</tr>
<tr>
<td>CC4</td>
<td>0.00868</td>
<td>-0.00552</td>
<td>-0.00574</td>
<td>0.02558</td>
<td>-0.00601</td>
<td>-0.00792</td>
</tr>
<tr>
<td>CC5</td>
<td>-0.00268</td>
<td>0.00684</td>
<td>-0.00873</td>
<td>-0.00601</td>
<td>0.01684</td>
<td>-0.00117</td>
</tr>
<tr>
<td>CC6</td>
<td>0.00004</td>
<td>-0.00769</td>
<td>0.00614</td>
<td>-0.00792</td>
<td>-0.00117</td>
<td>0.02202</td>
</tr>
</tbody>
</table>

are higher than for their corresponding within-speaker values in $U$ in Table 4, and also that the covariance magnitude is again small.

The difference between the suspect and offender data is evaluated using a Mahalanobis distance. An example can be seen in the numerator product $(\bar{y}_1 - \bar{y}_2)^T (D_1 + D_2)^{-1}(\bar{y}_1 - \bar{y}_2)$. As can be seen, in addition to the suspect and offender mean vectors $\bar{y}_1$, $\bar{y}_2$, it also includes the (inverted) suspect and offender variance-covariance matrices $D_1$, $D_2$. As pointed out by Khodai-Joopari (2006: 145) these variance-covariance matrices contribute both to the decorrelation of the individual features (the CCs) and the equalisation of their contribution. The remaining complexities of the formula are to handle kernel-density modelling of the between-speaker probability density (for example the smoothing parameter $h$), and the variance ratio. This is the ratio of between- to within-speaker variances $(C/U)$ and acts as a major scaling factor of the LR: other things being equal, the greater the magnitude of between-speaker variance $C$ to within-speaker variance $U$, the greater the magnitude of the LR.

Thus the process involves evaluating the Mahalanobis difference between the suspect and offender mean vectors against measures relating to the same-speaker variances and the different-speaker variances to determine the LR.

The crucial question in forensic voice comparison is the extent to which same-speaker speech samples can be discriminated from different-speaker speech samples. The LR-based approach allows us to test this. In this experiment, each speaker’s first session recording was compared to their second, non-contemporaneous, recording to get a set of 297 known non-contemporaneous same-speaker MVLRs. To get different-speaker MVLRs, each speaker’s first recording was compared with the first recording of all the other speakers. This gave MVLRs from 43,956 known different-speaker comparisons. The other independent different-speaker comparison, between a speaker’s second recording and the others’ second recording, was not evaluated, as the results are usually quite similar. It is important to note, however, that both sets are important for one kind of estimate of reliability of the system (Morrison 2011b).
To the extent that the cms CCs for the five vowels can discriminate between same-speaker pairs and different-speaker pairs, their log₁₀MVLRs will be greater than 0 for same-speaker comparisons, and less than 0 for different-speaker comparisons. It is then just a case of seeing to what extent this is true. LRs were also estimated with a GMM–UBM approach (Gaussian mixture model–universal background model) (Morrison 2011a), but this, surprisingly, gave markedly poorer performance and was not further pursued.

Results: Tippett plots and calibration

Figure 8: Tippett plots from MVLR-based comparison using /i/ showing shifting and scaling effect of logistic regression calibration. Left = uncalibrated scores, right = calibrated Likelihood Ratios. Clr = Log likelihood-ratio cost. SS = same-speaker comparisons, DS = different-speaker comparisons (see next section).

The conventional way of showing the results of a LR-based comparison is with so-called Reliability, or Tippett plots. The left panel of Figure 8 shows the Tippett plot for the output from the MVLR formula for the LR-based comparison for the single vowel /i/. This involves MVLRs for 297 comparisons between the mean CS spectrum, represented by 14 cms CCs, of a speaker’s /i/ vowel on one occasion and their /i/ vowel spectrum on a second occasion several months later. The cumulative log₁₀LR curves for these same-speaker comparisons are shown rising to the right. There are also log₁₀LRs from 43,956 different-speaker comparisons, the curve for which rises towards the left. Log₁₀LR values, rather than linear LR values are used, as this results in a symmetrical scale around 0, it being easier to appreciate that a log₁₀LR of −2 indicates the same strength of evidence as a log₁₀LR of 2, the only difference being in the hypothesis they support (cf. the same comparison with linear LRs of 0.01 and 100). Examples of how the curves are read are: 40% of different-
speaker comparisons had $\log_{10} \text{LRs}$ greater than $-10$, or 85% of same-speaker comparisons had $\log_{10} \text{LRs}$ greater than 0.

There are two ways of interpreting the kind of results shown in Figure 8. The first is as a graphical demonstration of the degree of discriminability of a forensic feature – in this case the CS spectrum of /i/. As already said, the most basic question one must ask of a forensic feature, and of a forensic voice comparison system as a whole, is to what extent it can be used to distinguish between comparisons involving the same subject and comparisons involving different subjects. In reality all one has is speech samples: the trier of fact wants to know whether they are more likely to be same-speaker speech samples or different-speaker speech samples. To the extent that the feature can tell whether they are speech samples from the same speaker or from different speakers, it has forensic use. The likelihood ratio is the ratio of the probability of getting the difference between the speech samples assuming they have come from the same speaker relative to the probability of getting the difference between the samples assuming they have come from different speakers. Thus it is expected to be bigger than unity for same-subject comparisons and less than unity for different-subject comparisons, and can be used as a discrimination function to test the forensic discriminatory power of a feature. Many experiments over the past 15 years have done this. Since the common log of the LR is used, the question is whether the $\log_{10} \text{LR}$ for a comparison is greater or less than 0. If it is less than 0, the comparison is evaluated as different-subject, if greater than 0 then as a same-subject. We then test many known same-subject and different-subject comparisons to see to what extent they are correctly resolved.

If one were treating these results as a discrimination between same-speaker comparisons and different-speaker comparisons, then, a common evaluation metric of the system’s performance would be its equal error rate (EER). This is the value where the error rate for same-speaker comparisons is the same as that for different-speaker comparisons. The EER can be easily seen in a Tippett plot if plotted in the way shown in Figure 8: it is where the curves cross, at $\log_{10} -2$. This shows that the best overall performance of the system would be obtained were the decision threshold between same-speaker and different-speaker speech samples set at $\log_{10} -2$, it and would be about 3%: c. 3% of both same-speaker and different-speaker comparisons would be incorrectly evaluated. This is of course quite a respectable result, and so the method shows good discrimination. However, for two reasons that is strictly speaking not the correct interpretation of these curves. The first reason (the second is discussed later) is that in LR-based comparison the threshold is predicted by theory to be at $\log_{10} 0$. This is clearly not the case in the left panel of Figure 8. There is not only some overlap between the two sets of same-speaker and different-speaker values, but,
as already mentioned, their crossover point is displaced to the left, at about \(\log_{10} -2\). To be sure, only a few of the different-speaker comparisons – about 1% of 43,956 – had counterfactual values above \(\log_{10} 0\). On the other hand, about 15% of the 297 same-speaker comparisons had counterfactual values below \(\log_{10} 0\).

This situation, where either same-speaker comparisons or different-speaker comparisons, or both, are associated with large counterfactual LRs, is described as poor calibration. It is typical of MVLR-based comparisons where a large number of features (here, 14) is used, and usually it is the same-speaker comparisons that are badly evaluated (i.e. the uncalibrated Tippett plot is typically shifted to the left). This bad calibration may be due to several factors. It may relate to the inadequacy of estimating multidimensional probability density functions with too few subjects: for a large number of variables, as here, one needs a very large number of subjects for a good estimate, and in fact the originators of the MVLR explicitly suggest limiting testing with it to low-dimensional data (i.e. two or three variables) because of this (Aitken and Lucy 2004). The bad calibration may also have to do with inadequacies of the model related to the fact that it is not designed for speech acoustics, where, for example, within-speaker variance cannot automatically be assumed to be invariant. It is interesting to note that the bias in fingerprint LRs (Morrison 2012) and SMS texting LRs (Ishihara 2012) is exactly the opposite, with uncalibrated data migrating to the right. In any case, either the same-speaker LRs are being estimated too low or the different-speaker LRs too high, or a mixture of both, to give LRs that are overall too low.

Because of this, it is customary to refer to these MVLRs by the technical term scores, that is, as difference measures which take not only similarity into account, but also typicality, and to use additional processing to convert them into true LRs (Campbell, Brady, Campbell, Granville and Reynolds 2006; Morrison 2012).

To do this conversion it is normal to calibrate the scores with logistic regression (Ramos-Castro, Gonzalez-Rodriguez and Ortega-Garcia 2006). Logistic regression is used when regressing a binary on a continuous variable (here, regressing whether the comparison is same- or different-speaker on the value of the score, and is another common automatic FVR procedure (Pigeon, Druyts and Verlinde 2000)). It involves firstly regressing the log-odds of the hypotheses for a score \(s\), i.e. \(\log \left[ \frac{p(H_{ss}|s)}{p(H_{ds}|s)} \right] \) on the score, and then using the intercept and slope of the linear regression to transform the scores to true LRs. Morrison (2012) gives an excellent explanation of how this works. The output of such a calibration is a set of true LRs, which have been shifted and scaled by the intercept and slope of the logistic regression. For example, the logistic regression coefficients for the /i/ data scores in Figure 8 were 0.5836 (slope) and 3.0533 (intercept). These are natural logs, so to calibrate a score of, say, \(\log_{10} -2\)
it would first have to be converted to its natural log: $-4.605$. This can then be scaled and shifted with the coefficients (also called weights): $-4.605 \times 0.5836 + 3.0533 = 0.3657$, and the result can then be converted back to its $\log_{10}$ LR of 0.1588.

The effect of calibrating the /i/ MV scores in the left panel of Figure 8 is shown in the right. It can be seen, firstly, that the crossover point has been shifted to the right so that it is in fact very close to the $\log_{10}0$ threshold. The bad counterfactual same-speaker scores have been reined in, so that only 2.5% of the 297 same-speaker comparisons now have counterfactual LRs, and they are moreover not very big – lying higher than $\log_{10} -1$. These improvements cannot be without cost, however. The different-speaker comparisons now no longer show such strength of evidence as their scores: now about 96% of different-speaker LRs lie below $\log_{10}0$, compared to about 99% of scores. Note, too, the very slight worsening of the counterfactual different-speaker LRs: they now extend to $\log_{10}3$, compared to just below $\log_{10}3$. This is also typical. The error rate for /i/ around the $\log_{10}0$ threshold is now 2.4% for same-speaker comparisons and 3.6% for different, but because the EER is located so close to $\log_{10}0$ we can use it (3%) as an estimation of the probability of being wrong. If you were to claim same-speaker provenance on the basis of a MVLR for /i/ being greater than $\log_{10}0$, or different-speaker provenance on the basis of an /i/ MVLR being less, the probability you would be wrong would be about 3%.

A useful and simple measure of this discriminatory performance, taken from medical diagnostic practice, is to treat the MVLR comparison as a kind of test for same-speaker data, and compare its sensitivity to 1 – its specificity. The sensitivity is the probability of getting a LR greater than threshold for a same-speaker comparison: this measures how sensitive the test is to the presence of a same-speaker comparison. The specificity is the probability of getting a LR less than threshold for a different-speaker comparison: this measures how specific the test is to same-speaker comparisons – i.e. how well it manages to reliably reject different-speaker comparisons. Since this statistic is also a likelihood ratio, it was called $L_{test}$ by Rose et al. (2003). Its formula is given at (8). Obviously, the higher the $L_{test}$ the better. The $L_{test}$ for the /i/ results is $(97.6 / 3.7) = 26.4$. After calibration you are about 26 times more likely to get a $\log_{10}$LR greater than 0 if it is from a same-speaker comparison than a different-speaker comparison.

(8)

\[
L_{test} = \frac{\text{sensitivity}}{1 - \text{specificity}} = \frac{p(\log_{10}LR > 0 \mid \text{same-speaker comparison})}{1 - p(\log_{10}LR < 0 \mid \text{different-speaker comparison})}
\]
The results in the right panel of Figure 8 show, then, that, assuming a threshold of $\log_{10} 0$, non-contemporaneous same-speaker speech samples can be discriminated from different-speaker speech samples reasonably well on the basis of their /i/ vowel spectra as represented by 14 cms CCs. In real terms 294 of the 297 same-speaker comparisons and 42,374 of the 43,956 different-speaker comparisons were correctly discriminated. The test obviously shows fairly high sensitivity and specificity.

**Log likelihood ratio cost $C_{\text{llr}}$**

Although intuitively appealing, especially perhaps to the legal profession, interpreting and evaluating the performance of likelihood ratio detection systems in terms of error rates is strictly speaking incorrect, and this brings in the second way of interpreting Tippett plots mentioned above: as indicators of strength of evidence. Deriving error rates from LRs in this way is conceptually incorrect, because error rates imply posterior probabilities: an error-rate of 3% refers to the probability the hypothesis is not true, given the evidence: $p(\neg H|E)$. Likelihood ratios, on the other hand, are the other way round. They are ratios of probabilities of evidence under competing hypotheses: $p(E|H)/p(E|\neg H)$. They require prior odds to convert them to posteriors. Thus the set of LR values in Figure 8 can mean totally different things depending on the prior odds in the case. If we obtained a $\log_{10} \text{LR}$ of 2 for a comparison between suspect and offender /i/ vowel cms CCs, and there were 11 other people including the suspect who could have said the incriminating speech, the posterior odds that they did indeed say it would be (100:1*1:10 = ) 10 to 1 in favour of them saying it – a posterior probability of (10/11 = ) 91%. But if there were 1001 other people, including the suspect, who could have said it, the posterior probability that they said it falls to 9% – i.e. they very probably didn’t. Another way of appreciating LRs in their proper guise – as strength of evidence indicators – is with the maximum strength of evidence for same-speaker comparisons. In Figure 8 it is just over $\log_{10} 3$. This means that the trier of fact, or prosecuting body, can be informed that, with this feature, in order to get a posterior probability of c. 95% there would need to be other evidence indicating no more than about 50 other people including the suspect who could have said the incriminating speech. Any more possible perpetrators and the posterior probability falls – and that is the best possible scenario, with the strongest LR possible! With a medium strength $\log_{10} \text{LR}$ of about 2.3, the number of possible perpetrators would need to be narrowed down to about 10 for a 95% posterior probability. (Calculations can also be made assuming any ‘reasonable doubt’ equivalent, or a balance of probabilities appropriate to a civil case.) Tippett plots
viewed in this way, then, reflect the kind of data one would have to collect and test in a real case. If one wants to talk about error rates with LRs it is good to bear in mind that a prior of 0.5 is assumed.

$$C_{llr} = \frac{1}{2} \left( \frac{1}{N_{hp}} \sum_{j} \log_2 \left( 1 + \frac{1}{LR_{j}} \right) \right) + \frac{1}{N_{hd}} \sum_{j} \log_2 \left( 1 + LR_{j} \right)$$

where $Hp =$ the set of LRs for the prosecution hypothesis = LRs from known same-subject comparisons, and $Hd =$ the set of LRs for the defence hypothesis = LRs from known different-subject comparisons.

Currently the evaluation metric of choice for the performance of LR-based detection systems, like the one in this article, is the log likelihood ratio cost ($Cllr$) (Brümmer and du Preez 2006). The formula of this simple information-theoretic cost function is given at (9), where it can be seen to consist of the mean of two hypothesis-dependent logarithmic functions. The left-hand term evaluates the performance of all same-speaker LRs; the right-hand one the performance of all different-speaker LRs. The purpose of $Cllr$ is to severely penalise highly misleading LRs. For example, a different-speaker LR of 1000 (misleadingly and dangerously indicating that one would be 1000 times more likely to get the difference between the suspect and offender samples had they come from the same speaker) attracts a value of $\log_2 (1 + 1000) \approx 6.909$. Since $Cllr$ effectively does not reward correct cases, even if they involve high LRs, this value of 6.909 then has a high contribution to the mean of all the different-speaker LRs and the overall $Cllr$ value. A $Cllr$ value below unity means that the system is delivering some information, and the lower the $Cllr$ the better. The $Cllr$ values in Figure 8 are decomposed into same-speaker and different-speaker values to show the contribution of the two different sources to the overall $Cllr$ value. Note how the high uncalibrated same-speaker contribution (0.72) reduces to 0.09 after calibration.

$Cllrs$ are relative; they can only be used to compare the performance of systems on the same data. The $Cllr$ value for the uncalibrated MV scores and the calibrated MVLRs in Figure 8 show that calibration results in quite a large improvement from about 0.4 for the uncalibrated scores to about 0.1 for the likelihood ratios. Often with high-dimensional data the uncalibrated scores have $Cllrs$ much higher than unity.

The results for the remaining four vowels’ cms CCs likelihood ratios (i.e. after calibration) are shown in Figure 9, and the numerical results for all vowels gathered in Table 6. Although all vowels perform well, with $Cllrs$ all below 0.25, /i/ has the greatest individual-identifying potential and remains capable of
delivering the strongest evidence. Interestingly (because individual CCs generally lack phonetic content), there is a clear relationship between vowel phonetics and evidential strength, which correlates primarily with height: open vowels are less strong than mid vowels which are less strong than high vowels.

Table 6: Numerical results for individual vowel LR-based comparison

<table>
<thead>
<tr>
<th>Vowel</th>
<th>SS error rate (%)</th>
<th>DS error rate (%)</th>
<th>LRtest</th>
<th>Cllr</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>2.36</td>
<td>3.7</td>
<td>26.40</td>
<td>0.11</td>
</tr>
<tr>
<td>e</td>
<td>4.38</td>
<td>7.27</td>
<td>13.15</td>
<td>0.22</td>
</tr>
<tr>
<td>a</td>
<td>7.74</td>
<td>8.08</td>
<td>11.42</td>
<td>0.24</td>
</tr>
<tr>
<td>o</td>
<td>4.71</td>
<td>7.23</td>
<td>13.18</td>
<td>0.22</td>
</tr>
<tr>
<td>ü</td>
<td>3.03</td>
<td>4.77</td>
<td>20.32</td>
<td>0.14</td>
</tr>
</tbody>
</table>

Figure 9: Tippett plots for individual vowels /e a o ü ('w')/ cms CC MVLRs.
Logistic regression fusion

Although the MVLR takes into account any correlation between features within a segment – here between the cms CCs within a vowel – any between-segment correlation must be also be handled. This problem arises commonly when comparing suspect and offender speech samples with respect to the acoustics of several different vowels, the features of which may be correlated (Rose 2010a), but which do not constitute multivariate data (for example, there may be five replicates of an /a/ vowel in the suspect data and seven of an /i/ vowel). Between-segment correlation can also be handled by logistic regression, which in this case combines the LRs from the different vowels according to the correlation between the vowels’ scores. It is then called logistic regression fusion. The good news is that one does not need to calibrate and fuse separately. Since logistic regression does both simultaneously, the multivariate scores from the separate vowel comparisons can be input into the logistic regression, and the output will be a set of calibrated LRs for all five vowels combined. Thus the calibration step described in the previous section is not necessary unless one wants for some reason, as here, to inspect how the individual vowels compare.

Figure 10 shows the Tippett plot of the results for the MVLR-based discrimination/strength of evidence estimates using all five vowels combined with fusion. To show the effect of cepstral mean subtraction, results using all 14 raw CCs have also been included and plotted with dotted lines. The insert, which zooms in on details around the log_{10} 0 threshold, shows that all of the 297 same-speaker comparisons were correctly evaluated with a LR that would be more likely had they come from same speakers; and out of the 43,956 different-speaker comparisons, less than 0.5%, or 173, had counterfactual LRs. The LR_{test} for these data is 254: one would be about 250 times more likely to get a log_{10} LR greater than 0 if the suspect and offender samples were from the same rather than different speakers. The Cllr is also very low at 0.013. The effect of fusing the scores from all five vowels thus improves considerably on the individual vowels’ performance. This echoes earlier results from speaker-recognition experiments using individual vowels that ‘speaker-characterizing vowels differ from speaker to speaker’ (Furui and Matsui 1994: 1463) and is one of the senses in which ‘more is better’ (another sense is that the performance improves with more CCs).

The beneficial effect of cepstral mean subtraction is also clear from the plot: it improves on the discrimination and strength of evidence with the raw CCs, more than halving the Cllr, and doubling the LR_{test} value. Furui (1981: 264) demonstrated the same dramatic improvement from mean cepstral subtraction. Figure 11 gives a clearer idea of the effect of cepstral mean subtraction by
Figure 10: Tippett plot for multivariate LRs derived from comparisons using 14 LPC CCs from all five vowels combined. Solid lines = LRs from cepstrally-mean-subtracted-CCs; dotted lines = LRs from raw CCs. Insert shows detail around log_{10} = 0 threshold.

Figure 11: Cepstrally mean subtracted vs raw MVLRs for 297 same-speaker comparisons (left) and 43,956 different-speaker comparisons (right). The line for which cms LR = raw LR is also plotted.
plotting the cms vs raw log₁₀ LRs separately for each of the 297 same-speaker and 43,956 different-speaker comparisons (a relationship that is not clear in the cumulative Tippett plots). It can be seen how in the majority of cases the cepstral mean subtraction increases the raw same-speaker LRs so that most cms comparisons lie above the line of equality between cms and raw values, and decreases the raw different-speaker LRs so that most cms LRs lie below it.

The effect of cepstral mean subtraction on the different-speaker error rate is also clear: the number of counterfactual cases with raw LRs (where raw LR > 0) is visually greater than the number of counterfactual cases with cms LRs (where cms LR > 0) (there is no corresponding effect for same-speaker comparisons as they are all correctly evaluated). This cepstral mean subtraction effect is achieved by increasing the ratio of between- to within-speaker variance (C/U).

As mentioned above, this variance ratio acts as the main scalar of the LR formula, forcing raw different-speaker log₁₀ LRs towards $-\infty$, and mutatis mutandis for raw same-speaker LRs. It has no effect on the absolute difference between spectra, which remain the same for raw and cms comparisons.

The fact that the performance improves considerably with an attempt to compensate for the effect of telephone transmission is important. It appears to discount one further explanation for the good results, namely that the good speaker discrimination was due at least in part to transmission effects. (I suspected this from inspection of some same-speaker comparisons of the type shown in Figure 7, where I would have expected greater differences between the non-contemporaneous mean cepstra.) If it were the case, for example, that each speaker used different, yet invariant telephone channels (each phoning in from their own office, perhaps, with the same routing), this would enhance the discrimination. A speaker’s individual vocalic spectra characteristics would be reinforced by the individual spectral characteristics of their transmission, both of which would be inextricably encoded by the cepstrum into a single spectrum. However, if there were differences in channel transmission contributing to discrimination in this way, it is difficult to see how, when an attempt is made, as in this experiment, to compensate for them with cepstral mean subtraction, the discrimination is actually enhanced. It can also be shown with a bandlimited cepstral analysis (Clermont and Mokhtari 1994; Khodai-Joopari et al. 2004) that removal of the top 1 kHz of the spectrum does not affect the results. This also suggests minimal contribution from the telephone transmission.

These results are still capable of improvement. It will be recalled that, in order to perform cepstral mean subtraction, it was necessary to estimate the mean cepstrum for the remaining, much larger, part of a speaker’s utterances which was then taken as an estimate of the phone transmission. However,
because the assumption regarding spectral flatness of the original signal is not realistic, the long-term mean cepstral spectrum will also contain information on a speaker’s long-term average spectrum, which has been shown to be effective in speaker recognition (Furui, Itakura and Saito 1972; Doherty 1976; Hollien and Majewski 1977). Just fusing a speaker’s scores from their long-term mean CCs with scores from their vocalic cms-CCs results in a slight improvement over the vowels alone: the $\text{Cllr}$ drops a little to 0.01, the different-speaker error drops to 0.3%, and the same-speaker error remains at 0%, giving a $LR_{\text{test}}$ of 326. Still more improvement can be obtained by first performing a cepstral mean subtraction on the long-term mean cepstral data, by subtracting from them the overall mean of all speakers. This is like removing the effect of telephone transmission common to all calls. If this is done, and long-term scores fused with those from the individual vowels, the $\text{Cllr}$ reduces to 0.009, the different-speaker error rate to 0.027% and the $LR_{\text{test}}$ increases to 369. Figure 12 shows the Tippett. The improvement can be best appreciated visually by comparison with the plot for vowels alone in Figure 10. It can be seen, for example, that c. 65% of different-speaker comparisons have $\log_{10}$ LRs below $-10$ with vowels alone, compared to 75% with fused vowels and long-term spectra. With same-speaker comparisons, 50% have $\log_{10}$ LRs greater than c. 5.7 with vowels alone, compared to $\log_{10}$ LRs greater than 6.2 with fused data.

![Figure 12: Tippett plot for fused vocalic and cepstrally-mean-subtracted long-term mean cepstral CCs](equinoxonline)
In view of these good results it is also worth bearing in mind that of the 119 incorrectly evaluated different-speaker comparisons in Figure 12, although most (82%) had LR values less than 100, 10 had LRs bigger than 1000, and the worst was 522,093 times more likely given same-speaker provenance. This is the price one pays for including higher-order CCs: the magnitude of the worst different-speaker LRs drops considerably with lower-order analyses (although of course the Cllr and the LR_{text} worsen).

This is a salutary caution that different speakers can still occupy exceedingly similar positions in heavily multidimensional space. With (5 vowels + 1 cepstrally mean subtracted long-term spectrum × 14 CCs =) 84 dimensions there is a 1/44,000 chance of finding two speakers similar enough in their five-vowel cepstral spectra to attract a counterfactual LR of c. 522,000! Rose (2010b) shows a similar thing with Australian English. Even using formant data from all 18 Australian vowel phonemes (c. 50 dimensions) of 171 males it was not possible to correctly discriminate all c. 14,500 different-speaker pairs, and there was one different-speaker pair with a counterfactual LR of 10,000. It would be interesting to compare the longer utterances of the two speakers in the worse different-speaker comparison – speaker 145 from Kyuushuu and speaker 255 from Niigata – to see if their longer utterances also show the same very high degree of similarity.

Discussion

The results just presented indicate that vocalic segmental cms CCs might be of use within a likelihood-ratio based approach to discriminate between same-speaker and different-speaker speech samples. One should always be sceptical of exceptionally good results when forensically testing speech data, however, and this is especially true in this experiment, as several factors will have contributed to them. Firstly, the good results undoubtedly reflect the very favourable conditions of the comparison. As described above, considerable care was taken to ensure a priori minimal within-speaker variation in CCs. This will favour same-speaker comparisons, and I suspect it is the reason why all same-speaker comparisons are correctly evaluated. Isolated vowels, with their spectra unperturbed by consonantal effects, are also highly comparable. Isolated vowels might also be expected to have a longer duration than in normal speech (for example, the duration of the Australian female’s normally spoken vowel in Figure 1 is about half that of Japanese speaker 52’s read-out vowel in Figure 5). Normally, this will have enhanced the frequency resolution of the cepstral spectrum, but in this experiment it will be recalled that only a single frame was used, so the duration difference is unlikely to have been a factor. In any case,
the reduction in variance associated with these factors will presumably have
offset to a certain extent the effects associated with non-contemporaneity,
telephone transmission and small number of replicates. (Lest it be thought,
however, that isolated vowels are totally unrealistic – i.e. do not occur in reality
– the author has more than once had to deal with forensic speech samples
containing them, when speakers spell out names, for example.)

Database heterogeneity

There will also be some contribution to the good results from the Japan-wide
sampling. The 11 different prefectures from which the 297 database speakers
were selected were chosen to give a reasonable geographical coverage.
Consulting a map of Japan will show that the sampling is fairly even – only
Hokkaido is not represented. Given this wide coverage, and the aims of this
article, it is sensible to ask about the degree of pan-Japanese vocalic homo-
geneity. This is because if there are sources of heterogeneity in addition to non-
linguistic between-speaker differences – for example dialectal differences – this
will contribute to advantageous different-speaker LRs and a better overall
discrimination. Since the NRIPS data were collected from prefectures over
most of mainland Japan, one possible a-priori source of heterogeneity in the
vowel acoustics might be differences in the /high back vowel/ allophones, i.e.
dialects can be classified into two or three groups: either East, West and
Kyuushuu, or Eastern and Western, with the Kyuushuu varieties belonging to
the latter (Shibatani 1990: 196). The /high back/ vowel diaphone is generally
rounded in the West and Kyuushuu group of dialects, but unrounded in the
East dialects (Shibatani 1990: 198). Osanai et al. (1995) were careful to sub-
classify the NRIPS data by prefecture, and this allows us to group speakers by
major dialect area. Table 7 lists the prefectures, the speaker indices and the
dialect area to which they belong. It can be seen that the East and West dialect
areas are represented by five prefectures each, with one prefecture for the
Kyuushuu area, and that each prefecture except Osaka is represented by 30
speakers (data from three of the Osaka speakers were corrupted and omitted).
This information allows us to do a crude check for vocalic differences associ-
ated with dialect area. Figure 13 shows the mean cms CC spectrum for the
/high back vowel/ in the three dialect areas of West, East and Kyuushuu. It can
easily be seen that the speakers from the East dialect prefectures have a spectral
prominence at 1.2 kHz, which presumably corresponds to the higher F2
expected from an unrounded high back vowel. Now, the subjects were sup-
posed to be speaking Standard Japanese, but it could be heard from casual
listening that sometimes they were retaining aspects of their local pitch accents. It appears then that at least some were also retaining dialectal vowel quality differences, at least with respect to their /high back vowel/.

Table 7: Breakdown of database by prefecture and dialect area. E W Ky = East West Kyushuu

<table>
<thead>
<tr>
<th>Prefecture name</th>
<th>Dialect area</th>
<th>Speaker numbers</th>
<th>Prefecture name</th>
<th>Dialect area</th>
<th>Speaker numbers</th>
<th>Prefecture name</th>
<th>Dialect area</th>
<th>Speaker numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hyogo</td>
<td>W</td>
<td>1–30</td>
<td>Saga</td>
<td>Ky</td>
<td>121–150</td>
<td>Niigata</td>
<td>E</td>
<td>241–260</td>
</tr>
<tr>
<td>Aichi</td>
<td>E</td>
<td>31–60</td>
<td>Fukushima</td>
<td>E</td>
<td>151–180</td>
<td>Toyama</td>
<td>W(?)</td>
<td>261–280</td>
</tr>
<tr>
<td>Osaka</td>
<td>W</td>
<td>61–78</td>
<td>Wakayama</td>
<td>W</td>
<td>181–210</td>
<td>Iwate</td>
<td>E</td>
<td>281–300</td>
</tr>
<tr>
<td></td>
<td></td>
<td>82–90</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kanagawa</td>
<td>E</td>
<td>91–120</td>
<td>Ehime</td>
<td>W</td>
<td>211–240</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 13: Mean cepstral spectra to 4 kHz for /high back vowel/ in the three major dialect groups represented in the data. Arrow indicates putative F2 in East dialects.

There are not only dialect-related differences between speakers, however. The pairwise differences between prefectures estimated using the quefrency-weighted cepstral distance (a very simple distance measure appropriate for the cepstrum) showed that there were also clear differences between some prefectures. Interestingly, these differed for different vowels, so, for example, Wakayama shows the greatest distance from nearly all prefectures in the /low
vowel/ cepstral spectrum. The prefecture spectra for the /low vowel/ are shown in Figure 14. (Each spectrum is the mean of 30 speakers, except Osaka, with 27). It can be seen that the Wakayama spectrum appears to differ from the others in having a lower F2 which might indicate a backer articulation. Its F3 is also more prominent. Toyama is furthest from most other prefectures in the /mid front vowel/, whereas Osaka differs the most from all other prefectures – even those adjacent to it in the same dialect area – in its /high front vowel/. These findings, while also showing that the cepstral spectrum might be useful for comparative phonetic purposes, thus confirm that there are sources of heterogeneity other than non-linguistic between-speaker differences in the data – some /vowels/ are expected to be phonetically different across different prefectures and dialect areas – and this will have contributed to its overall discriminability; another reason not to get too excited with this performance, therefore.

![Figure 14: Mean cepstral spectra to 4 kHz for /low vowel/ in the 11 prefectures represented in the NRIPS database showing different spectrum for Wakayama.](image)

**Extrinsic testing with independent MVLR and logistic regression parameters**

Another factor contributing to the discriminatory performance is the *intrinsic* nature of the comparison.

The LR-based comparisons described in this article were done intrinsically, that is, by using the same set of data for testing as training (i.e. getting the parameters for the testing). In a LR-based comparison of the type just per-
formed there are two sets of parameters to be estimated: the between- and within-speaker variance–covariance matrices used to obtain the multivariate scores, and the coefficients used in the logistic regression calibration/fusion to convert the scores to LRs. In the intrinsic comparison just described, these were obtained from the same set of data used for testing: this will guarantee optimum performance.

To get a better idea of how well the approach really performs, results should be obtained extrinsically, in the time-honoured engineering way, using separate data for training and testing. One way of doing this is by cross-validation with a leave-one-out method. But since a lot of speakers are available, it would be simpler to separate them into three independent groups: one to obtain estimates of the between- and within-speaker variance–covariance matrices, one to estimate the logistic regression coefficients, and one for testing. Space prevents describing the results of such experiments, where it would be expected that the $Cllr$ and $LR_{test}$ values would be worse than the intrinsic values, but they are important because they relate to another very important consideration in properly testing LR-based approaches.

It is customary in LR-based FVC testing, and engineering approaches in general, to use what has been called a ‘plug-in’ approach (Brümmer 2011). This is where the model parameter estimates, for example the logistic regression coefficients, are obtained from a set of data, and the independent test data and parameter estimates then ‘plugged-in’ to the appropriate formulae to get a result. Brümmer (2011) correctly points out that this is misleading, as the parameter values are of course only one of a set of estimates – a different set of results would almost certainly be obtained from a different sample. In reality, there is thus uncertainty about the model parameter values, and this uncertainty should be incorporated into the analysis. Brümmer (2011) shows how; he calls the approach ‘fully-Bayesian’. You cannot aspire to greater strength of evidence if you are less certain of the input parameter values, and one reasonable outcome of the fully-Bayesian approach is that the greater the uncertainty accompanying the model parameters, the more the LR will tend to $\log_{10} 0$.

**Summary, conclusion, way forward**

This article was motivated by the aim of quantifying the whole of a speaker’s vocalic spectrum in forensic voice comparison rather than just two or three of their formants. It has used a large database of speakers to show how same-speaker vowel samples can be well discriminated under conditions of non-contemporaneity and telephone transmission from different speakers’ vowels using multivariate likelihood ratios derived from their cepstrally subtracted
segmental cepstrum. The experiment has shown that the segmental cepstrum as an addition to the forensic voice comparison arsenal deserves some further study – perhaps with non-vocalic sonorants; definitely with mobile phones. It has also demonstrated once again the power of the likelihood ratio as a forensic discriminant function.

The forensic applications of the cepstrum are many, and this article has barely scratched the surface. It has only described the use of the real cepstrum; the complex cepstrum, which includes the phase relationships within the original signal, may also contain speaker-individualising information of use in estimating LR.s. Likewise, the potential of delta-cepstra (for diphthongs?) and mel-frequency segmental CCs remains to be investigated within the LR paradigm. There are also many other basic variations on the theme: the use of a normally modelled (as opposed to a kernel density) reference sample; various solutions to MVLR instability like Nair, Alzqhoul and Guillemin’s (2012) principal component model and Khodai-Joopari’s (2006) inverse-weighted cepstrum, as well as band-limiting (Clermont and Mokhtari 1994; Khodai-Joopari et al. 2004). The latter would be especially appropriate for phone data, as it enables one to ignore upper and lower frequency regions presumed to have been most compromised by transmission. Better ways of estimating the crucial between- and within-speaker variance for MVLR estimation are needed. There is also the enticing proposal to do away with independent between- and within-speaker variance estimates altogether by deriving them from the quefrency-weighted cepstrum (Khodai-Joopari 2006). It is also perhaps naïve to suppose that all vowels should be processed in the same way. Most important of all, I think, is exploration of Brümmer’s fully-Bayesian approach. It is hoped that some of these possibilities will be addressed in a later article. It will be interesting to see if any of them can improve on the intrinsic results described above.

About the author

Phil Rose was Reader at the Australian National University, where he taught Phonetics and Chinese Linguistics for 30 years. He has been a British Academy visiting professor at the University of Edinburgh’s Joseph Bell Centre for Forensic Statistics and Legal Reasoning, is chairman of the Forensic Speech Science Committee of the Australasian Speech Science and Technology Association and is a member of the Australian Academy of Forensic Sciences. His 2002 book pioneered the application of the likelihood ratio of Bayes’ Theorem to traditional forensic voice comparison, thus bringing it in line with forensic DNA profiling. He has done forensic voice comparison casework on Australian
English and varieties of Chinese for nearly 20 years and was last seen teaching courses on Chinese Phonetics and Forensic Voice Comparison in Cantonese at the Hong Kong University of Science and Technology.

Acknowledgments

This is a much expanded version of my paper at the 2011 International Congress of Phonetic Sciences which was written as part of Australian Research Council Discovery Grant DP0774115. Many, many thanks to Drs Takashi Osanai and Mehrdad Khodai-Joopari: they did all the really hard work – and it really is hard work – to collect the data and write the code to automatically process and extract CCs. I want especially to acknowledge the massive contribution of Dr Frantz Clermont, the main supervisor of Mehrdad’s PhD, the source of so many good ideas and data. Frantz has always been unstintingly generous in trying to impart to me his knowledge of the cepstrum, including now even some of what he calls its ‘dirty bits’. I doubt whether these three scholars would disagree much with what I have written, but it is best to let the usual disclaimers apply. Many thanks also to my two anonymous reviewers for their careful reading and helpful comments.

References


