Technical Forensic Speaker Identification from a Bayesian Linguist's Perspective

Phil Rose
Phonetics Laboratory, School of Language Studies, Australian National University
Joseph Bell Centre for Forensic Statistics and Legal Reasoning, University of Edinburgh
philip.rose@anu.edu.au

Abstract

Important methodological aspects of Technical Forensic Speaker Identification are discussed and exemplified. The centrality of the Likelihood Ratio of Bayes' Theorem as the proper way of forensically evaluating speech evidence is emphasised, as well as the many different types of evidence that are of use in discriminating same-speaker from different-speaker speech samples.

1. Introduction

The most typical scenario in speaker recognition for forensic purposes involves the comparison of one or more samples of a known voice with samples of unknown origin. Often the unknown samples are claimed to be of the individual alleged to have committed an offence, and the known voice belongs to the suspect. The court is then concerned with being able to say on the basis of the evidence whether the two samples have come from the same person, and thus be able either to identify the suspect as the offender, or eliminate them from suspicion. In other words, the court needs to know the probability of the hypothesis of same-speaker origin, given the evidence: $p(H_{SS}|E)$.

When the decision is theoretically informed, the terms Technical Forensic Speaker Identification, or FSI by Expert are often used. In contrast to this, so-called Naive Speaker Recognition refers to the unreflected everyday abilities of people to recognise voices - apparently not very well, unless the voices are familiar, and even then still not 100%.

This paper gives an overview of important aspects of Technical Forensic Speaker Identification (or Recognition - the terms are used synonymously) from the perspective of a practitioner and researcher. As will be seen, the focus is slightly more on the 'forensic', than the 'speech'. Interested readers are referred to [1] and [2] for further details.

2. Bayes' Theorem

Given that probability is the best standard of the uncertainty pertaining to a claim like "these two samples were spoken by the same speaker"[3], the logically correct way of estimating the probability of a hypothesis given the evidence brought in its support must be by Bayes' theorem (BT), and its centrality is the one non-negotiable thing in FSI. The well-known odds form of BT, which is the most convenient way to consider it, is shown at (1) in a form applicable to FSI (H_{ss} = prosecution hypothesis that the samples were spoken by the same speaker; H_a = alternative (defence) hypothesis; and E_{fsi}= forensic-speaker-identification evidence adduced in support of H_{ss} ; $p(E_{fsi} \mid H_{ss})$ etc.= probability of getting the evidence conditional upon the truth of the same-speaker claim.)

$$\underbrace{\frac{p(H_{ss} \mid E_{fsi})}{p(H_a \mid E_{fsi})}}_{\text{Posterior Odds}} = \underbrace{\frac{p(H_{ss})}{p(H_a)}}_{\text{Prior Odds}} * \underbrace{\frac{p(E_{fsi} \mid H_{ss})}{p(E_{fsi} \mid H_a)}}_{\text{Likelihood Ratio}} \quad (1)$$

As can be seen, (1) states that the posterior odds in favour of a hypothesis given the evidence adduced in its support are the product of the prior odds in favour of the hypothesis and the likelihood ratio (LR) for the evidence. The LR - the central notion in FSI - is the ratio of the probability of getting the evidence assuming the hypothesis is true to the probability of the evidence assuming an alternative hypothesis (one cannot estimate the probability of a hypothesis without comparing it to some alternative). Since voices are heavily multidimensional, it is possible to calculate LRs for each separate feature and then combine them into an overall LR. The easy combination of LRs (at least it is easy if the evidence is independent) is one of the beauties of the Bayesian approach. The conditions upon $p(H)$ are actually more complicated [4], and involve, for example, assumptions of how well the data are statistically modelled, and other background knowledge, in FSI for example whether a suspect is known to be bilingual. The prosecution claim H_{ss} is straightforward, but there are important aspects of the other terms that warrant discussion.

3. Prior odds

The prior odds are the odds in favour of the hypothesis before the evidence is adduced. Suppose the suspect is one of a group of five males known to be in a house at the time of an incriminating phone intercept. The prior odds are then 4 to 1 *against* them being the owner of the intercepted voice. Suppose further from comparison of known and unknown phone intercepts the evidence is estimated as 100 times more likely if the same speaker is involved (LR = 100). The posterior odds on the suspect being the speaker now shift to (100 * 1/4 =) 25 to 1 in favour.

It is clear from BT that, unless the FSI expert knows the prior odds, they logically cannot estimate the probability of the hypothesis. Since the FSI expert is usually not privy to information that informs the prior odds - and in fact there are good reasons why they shouldn't be [1] - they cannot logically state the probability of the hypothesis. Since this, in the author's experience, is precisely what is usually expected of the FSI expert by the court and police, this can be a big problem [1], [5].

It also needs to be acknowledged that this point is sometimes not appreciated even by the FSI experts themselves [6].

Most textbooks on the evaluation of forensic evidence/forensic statistics, e.g. [7], [8], stress that it is the role of the identification expert to estimate the strength of the evidence by estimating its LR - the probabilities of the evidence under competing prosecution and defence hypotheses. It would clearly be difficult to argue why FSI practitioners should be exempt from this, and thus a correct format for a FSI conclusion might sound something like this. "*There are always differences between speech samples, even from the same speaker. In this particular case, I estimate that you would be about 1000 times more likely to get the difference between the offender and suspect speech samples had they come from the same speaker than from different speakers. This, prior odds pending, gives moderately strong support to the prosecution hypothesis that the suspect said both samples*".

4. Likelihood Ratio

The likelihood ratio (LR) is by far the most important construct in FSI, since it quantifies the strength of the evidence in support of the hypothesis. Its numerator estimates the probability of getting the evidence assuming that the prosecution hypothesis is true; its denominator estimates the probability of the evidence under the alternative, defence, hypothesis. The relative strength of the evidence in support of the hypothesis is thus reflected in the magnitude of the likelihood ratio. The more the likelihood ratio approaches unity, the more the evidence is likely under both prosecution and defence hypotheses, and thus the more useless. The more the likelihood ratio deviates from one, the greater support for either prosecution (for LR > 1), or defence (for LR < 1).

Verbal equivalents for LRs exist. For example, for 100 < LR < 1000, evidence is described as giving "moderately strong" support for the prosecution hypothesis [5]. However, neither the verbal equivalents nor their use is universal and they can be criticised as circular [2]. Another, related, problem is that it is difficult to come to terms with the idea that, for example, 'strong support' is being claimed for a hypothesis which can be overturned when the prior odds are taken into account (although it is in fact all too common for the prior odds to be ignored by the court - whether by commission or omission is not clear).

In FSI, the LR numerator quantifies the degree of *similarity* between the offender and suspect samples, and its denominator quantifies the degree of *typicality* of the offender and suspect samples in the relevant population. Then the more similar the two samples are, the more likely they are to have come from the same speaker and the higher the ratio. But this must be balanced by their typicality: the more typical the samples, the more likely they are to have been taken at random from the population under consideration, and the lower the ratio. The value of the LR is thus an interplay between the two factors of similarity and typicality. BT makes it clear that both these factors are needed to evaluate identification evidence: it is a very common fallacy to assume that similarity is enough: that if two speech samples are similar that indicates common origin.

5. Alternative hypothesis and background data

The similarity between the forensic samples has to be evaluated for typicality against background (also called reference) data. The background data depends on the alternative hypothesis H_a, which needs careful consideration. Quite often H_a will simply be that the voice of the unknown speaker does not belong to the accused, but to another same-sex speaker of the language. In this case, a representative distribution of the parameter for appropriately sexed speakers of that language is needed. This is often a default assumption, because under many jurisdictions there is no disclosure to a prosecution expert of H_a before trial. If H_a is that the offender voice is of someone who sounds like the accused, then a distribution of the parameter in pairs of similar-sounding voices needs to be used. H_a might also be that the unknown speech is not from the accused but their brother, in which case the evaluation is considerably simplified into a closed-set comparison between the probability of observing the parameter assuming the accused, and the probability of observing it assuming his brother. H_a can on occasion get quite complicated. In a recent case, for example, it has been claimed, sensibly, that the questioned voice was not that of the female accused, but of a male speaker who sounds similar to the accused because her voice sounds like a male.

Proper implementation of BT requires that an adequate background distribution exists. In many cases, it does not, and has to be estimated by various means. The lack of adequate background data is one of main factors that makes the accurate estimation of LRs problematic.

6. LR formulae

As stated in the *locus classicus* for LR derivation [9] "There can be no general recipe [for a LR formula], only the principle of calculating the [Bayes'] factor to assess the evidence is universal". One of the factors influencing choice of formula for speech is whether the variances of the parameter in the forensic samples can be considered equal. Up to now, work on LRs has involved mostly refractive indices of glass, and has assumed equal variances, at least to derive analytic solutions. Several equal-variance LR formulae can be found in [9]. However, it is well-known that there is both between- and within-speaker variation in variance, and this will therefore make any LR estimate assuming equal variances less accurate.

A second factor is whether the background distribution for a parameter can be considered normal (appropriate, presumably, for cepstral coefficients, and for some formants [11]). It is conceded that this is probably an unrealistic default assumption [9], and this will be true for many traditional speech parameters. For non-normality, various formulae with simple numerical integration can be used [9], or a kernel density / GMM estimation [7]. A LR formula incorporating gaussian kernel density estimation, from [7] is given at (2). It can be seen that it still assumes uniform within-speaker variance, which is therefore likely to affect accuracy of LR estimates.

7. Evidence and forensic speaker identification features

It is necessary to distinguish three different things when discussing the notion of strength of forensic evidence as quantified by a LR. Firstly, there is the raw data: for example a fingerprint, blood spatter, or digitised speech sample on a CD. Next there is information that the court receives from the expert witness concerning their qualifications, experience, methods of analysis, and findings: this is evidence in the legal sense: relevant information that the court has then to weigh. Finally, there is the evidence in the Bayesian sense - information that the expert witness extracts from the raw data and quantifies. In FSI, this kind of evidence is then the ensemble of differences between the forensic speech samples when extracted and quantified with some analytic technique, such as LP-estimated MCC's or formant centre-frequencies. It is important to note these distinctions, because, firstly, typically there will be information in the raw data that is not exploited. This will be due, trivially, to time constraints, but much more importantly also to analytic approach: a local, perhaps formant-based approach will be unable to make use of much of the individual-specific information in the samples that can be extracted automatically; a global automatic approach is by definition unlikely to pick up important between-sample differences in the realisation of a single phoneme. It is also important to remember that, as with other areas of forensic science, different methods can result in different strengths of evidence, even on the same raw data.

7.1 Types of features

There are four main types of Bayesian evidence in FSI, usefully (but not crucially) characterised as the intersection of two binary features: *Auditory/Acoustic* and *Linguistic/Non-linguistic* [1]. Auditory features are those that can be extracted by trained, theoretically-informed listening [2]. The theory comes from all aspects of linguistic structure, not just phonetics, and the training is the kind provided by courses which teach (1) how to reliably transcribe and productionally interpret any speech-sound (and ideally any human vocalisation), and (2) how to analyse linguistic structure and the way it varies. An auditory analysis is just that - analytic - and not a holistic, undifferentiated "these two samples sound as if they have come from the same speaker" (although it is in principle possible to assign a LR to natural gut feelings like this [2]).

Acoustic features are self-explanatory, and can be subcategorised into *traditional* and *automatic*. Traditional features relate in a direct way to aspects of speech production, like formant centre-frequencies, F0, or jitter. Automatic features are those like cepstral coefficients. The distinction is important, since it reflects a tension between interpretability and discriminant power: traditional features have much greater interpretability, which is a bonus for explanations and justifying methodology in court. Automatic features are very much more powerful as evidence: they will, on average, yield LRs that deviate much more from unity [10]. Although the particular background of an expert will tend to influence their choice between automatic and traditional features, there is no reason why both types of features should not be combined in case-work [12]. Since different types of evidence are generally tapped by the two approaches, this would result in potentially even more powerful LRs.

$$LR = \frac{K \exp\left\{-\frac{(\bar{x}-\bar{y})^2}{2a^2\sigma^2}\right\}\sum_{i=1}^{k}\exp\left\{-\frac{(m+n)(w-z_i)^2}{2\left[\sigma^2+(m+n)s^2\lambda^2\right]}\right\}}{\sum_{i=1}^{k}\exp\left\{-\frac{m(\bar{x}-z_i)^2}{2(\sigma^2+ms^2\lambda^2)}\right\}\sum_{i=1}^{k}\exp\left\{-\frac{n(\bar{y}-z_i)^2}{2(\sigma^2+ns^2\lambda^2)}\right\}} \quad (2)$$

where

$$K = \frac{k\sqrt{(m+n)}\sqrt{(\sigma^2+ms^2\lambda^2)}\sqrt{(\sigma^2+ns^2\lambda^2)}}{a\sigma\sqrt{(mn)}\sqrt{\left[\sigma^2+(m+n)s^2\lambda^2\right]}}$$

and

$\bar{x}, \bar{y}$ = means of offender, suspect samples

m, n = number of observations in offender, suspect samples

s^2 = variance in reference population (between - speaker variance)

σ^2 = within - speaker variance

λ = smoothing factor for kernel density estimate

$a = \sqrt{(1/m)+(1/n)}$ $\qquad$ $w = \left(m\bar{x}+n\bar{y}\right)/(m+n)$

k = number of kernel functions

z_i = value at which probability density is evaluated for the ith kernel

Since there is evidence that the exclusive use of auditory or acoustic features is associated with considerable shortcomings, the consensus among practitioners is that both are necessary to evaluate differences between samples. An auditory approach on its own is problematic because it is possible, due to aspects of the resolution of the perceptual mechanism, for two speech samples to sound similar even though there are significant differences in acoustics [13]. By the same token, two forensic samples can have very similar acoustics and yet crucially differ in a single auditory feature. For example, one sample may uniformly have a labio-velar approximant [ʊ] for the English rhotic phoneme /r/, while the other is uniformly post-alveolar [ɹ] [14]. There is often an enormous amount of potentially useful - even crucial - information available from the auditory features, although the evidentiary value of a feature is often language-dependent (creaky phonation is a normal speech sound in Standard Vietnamese, and therefore of no forensic use; it can be a marker of individuality in varieties of English, although even there its forensic use is restricted because it can function paralinguistically to signal temporary boredom, and linguistically to signal end of turn at talk). Trivially, a prior auditory analysis is necessary to decide whether the samples are comparable in the first place, and if they are, what is to be compared (this also includes deciding how many speakers are involved, and partitioning the speech into putative speakers, since forensic samples are unlikely to be monologues).

Linguistic features have to do with how the units of Language - the supremely human code that links speech sound to meaning - are organised and realised. Linguistic features can be broadly grouped into: phonological (having to do with speech sounds - e.g. the choice of /rʊm/ or /rʊm/ for *room*); morphological (having to do with the structure of words - e.g. the choice of /juθs/ or /juðz/ for the plural of *youth*) and syntactic (the ways

words are strung together to form larger units like phrases or sentences - e.g. *these clothes need washed* vs. *these clothes need washing*).

Speakers of the same language can and do differ in linguistic features, although this depends on the language. Samples in languages with a strong norm, and less dialectal variation, like Australian English, generally contain less such features. Samples in languages with less established norms, and great dialectal variation, like Chinese, generally contain more.

8. Examples of forensic application

8.1 Acoustic-linguistic features

One of the commonest acoustic-linguistic features used in forensic comparison is vocalic formant centre-frequencies. F1 (except possibly for low vowels) and F4 (except for rhotics) are counter-indicated because of differential effects of the telephone transmission [15], [16], but F2 and F3 are usually reliably and usefully quantifiable for some vowels in even average quality recordings [2]. A further advantage is the availability of useful

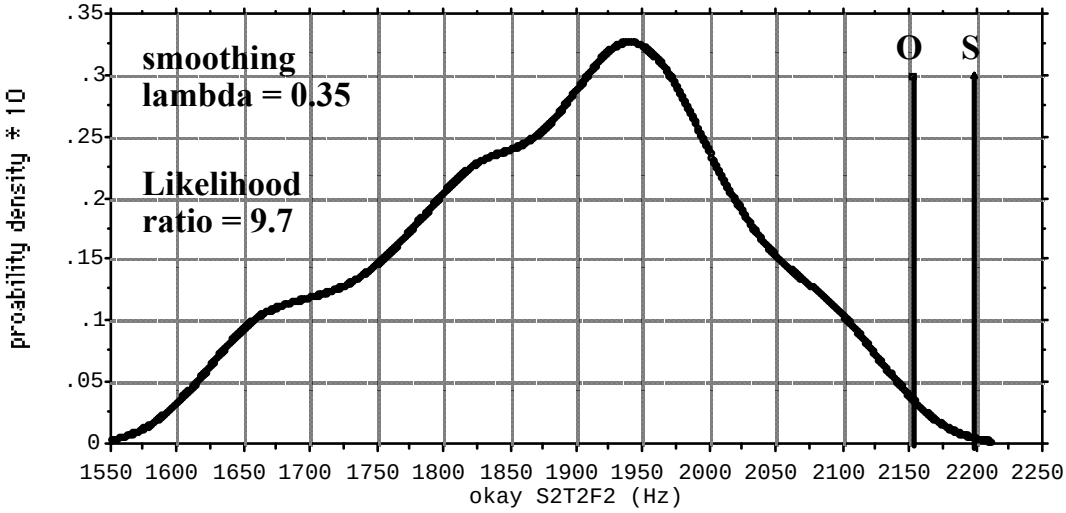

Figure 1: Forensic kernel density estimation of an acoustic-linguistic feature in *okay*. O = location of mean of offender samples, S = location of grand mean of suspect samples

Non-linguistic features can be defined negatively as what's left when you remove the linguistic ones. These may be habitual articulatory or phonatory settings like the use of nasalised or breathy or creaky voice; fast or slow speech rates; lower than average pitch etc. They may also be pathological features.

reference distributions for F-pattern centre frequencies. The feature is linguistic because, due to the long-known relationship between the lower formants and auditory vowel quality (height, backness, rounding), the lower formants relate clearly to the linguistic unit being signalled.

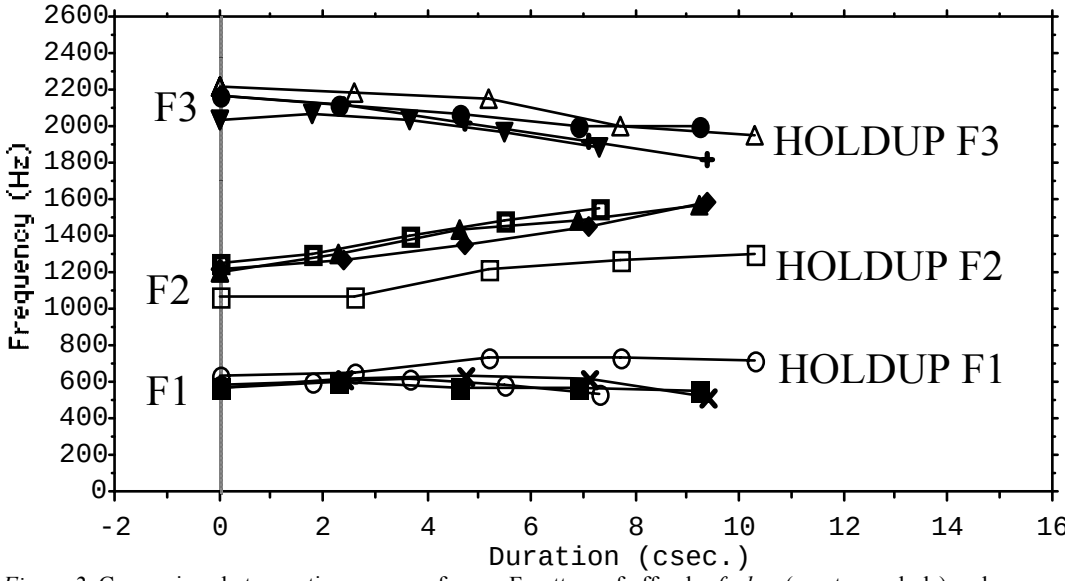

Figure 2: Comparison between time course of mean F-pattern of offender *fucken* (empty symbols) and mean F-patterns of *fucken* from three intercepted suspect phone calls (solid symbols)

Table 1: Data for LR comparison of mid-nucleus F-pattern in suspect (S) and Offender (O) samples of *fucken*. C1-C3 = suspect conversations 1 - 3. R = reference data for Broad (B) and combined Broad and General (B+G) Australian male /ɐ/ F-pattern. x = mean (Hz), sd = standard deviation (Hz), n = number in sample.

		F1			F2			F3		
		x	sd	n	x	sd	n	x	sd	n
O		**734**	92.1	7	**1215**	99.8	6	**2153**	59.6	4
S	C1	**574**	28.0	3	**1426**	43.3	3	**2072**	24.5	3
	C2	**621**	38.4	5	**1346**	67.3	5	**2021**	97.2	5
	C3	**611**	57.1	14	**1399**	74.4	13	**2029**	159.0	11
R	B	**737**	69.4	56	**1416**	93.1	56	**2526**	146	56
	B+G	**744**	68.5	117	**1414**	84.4	117	**2513**	151.2	118

Figure 1 represents the evaluation of evidence in a specific fragment of case-work based on the F2 centre frequency of the second diphthongal target in /ɛɪ/ in the Australian English word *okay* [2]. *Okay* is a very common word in conversations, and yields several forensically useful features. This particular frequency reflects how high and how front the speaker locates their tongue body at the end of the diphthong, as well, of course, as the overall dimensions of their tract. This is a value that is known to show a large ratio of between- to within-speaker variance [16], and in this particular case both suspect and offender samples were perceived to have a very close, very front offset to the /ɛɪ/ diphthong in this word. A comparison is shown between the mean value of 2151 Hz from four offender *okay*s in a single conversation, and a grand mean value of 2199 Hz from the means of several *okay*s in seven different conversations of the suspect. The difference between the suspect and offender values is evaluated using the kernel density estimation formula at (2) against the reference distribution of the same feature in the conversational speech of 10 male speakers of Australian English derived from [17]. As can be seen from the LR value, one would be about 10 times more likely to observe this difference had the samples come from the same rather than different speakers: weak support for the prosecution. Note that the LR magnitude in this example is still not very big, even though the offender and suspect values are fairly similar and atypical.

Another common word in forensic samples of probably many varieties of English is *fuck* or *fucken*. Figure 2 shows details from another acoustic-linguistic comparison between the F-pattern of the short open /ɐ/ vowel in a set of seven *fucken*s recorded during a hold-up and three sets of *fucken*s intercepted from separate telephone calls involving the suspect. The F-pattern was sampled at 25% points of the duration of the nucleus. The vowels in the criminal sample sounded backer than those in the suspect samples, and this difference corresponds to the clear difference in relative position of F1 and F2. Table 1 gives the numerical data (means, standard deviations, number in sample) for the first three formant centre-frequencies at the mid-point of the vowel, both for offender sample, criminal samples and reference distribution. The reference distribution against which the differences between the samples were compared consists of formant data from a relatively large number of male Australian English speakers [18]. Two sets of reference distribution values are given, corresponding to the two alternative hypotheses entertained: the offender is a broad-speaking male other than the suspect (denote by B); and the offender is someone other than the suspect with a non-cultivated accent (denote by B+G). (Australian accents are customarily classified on the basis of the quality of some vowels into three types, called Broad, General and Cultivated. In the case of the /ɐ/ vowel being tested, it can be seen that there is little difference between Broad and General values, and the results are therefore very similar for both alternative hypotheses.)

Table 3: Likelihood ratios for /ɐ/ F-pattern comparisons between suspect and offender *fucken* (S vs. O) and within-suspect *fucken.* C1 = suspect conversation 1 etc. *n* SS/DS = *n* times more likely to observe difference between samples if from same speaker/different speaker. B, B+G = LRs for different alternative hypotheses (see text). Shading indicates LRs counter to known reality.

	F1		F2		F3		Combined LR	
Within-suspect	B	B+G	B	B+G	B	B+G	B	B+G
C1 vs. C2	6.0 SS	7.4 SS	1.9 DS	2.1 DS	312 SS	176 SS	985 SS	620 SS
C1 vs. C3	14.4 SS	18.2 SS	1.7 SS	1.5 SS	204 SS	117 SS	4994 SS	3194 SS
C2 vs. C3	13.0 SS	11.7 SS	1.1 SS	1.1 SS	660 SS	350 SS	9438 SS	4505 SS
O vs. S	4.3 DS	3.7 DS	14.7 DS	15.5 DS	11.2 SS	6.8 SS	6 DS	8 DS

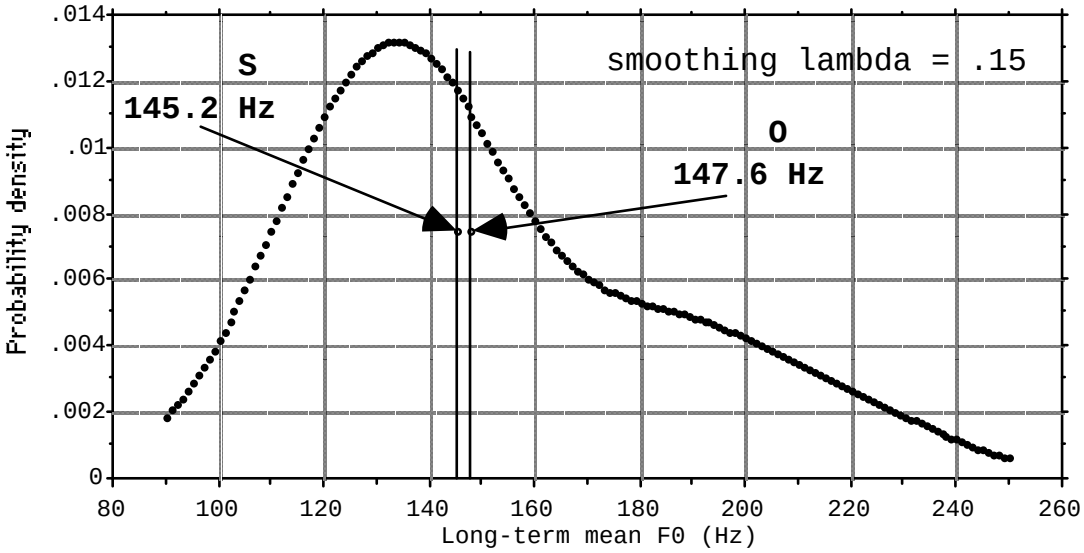

Figure 3: Mean Suspect and Offender LTF0 samples compared against a reference distribution of LTF0 in Cantonese

LRs were estimated for each of the formants, not only for the important offender-suspect comparison, but also for the within-suspect comparisons. A pooled-variance version of a LR formula was used, which assumes normality and equal variances [10]. Results are in table 2, which shows, for example, that when comparing the /ɐ/ F1 means in the suspect's two conversations C1 and C2, the difference between their values would be about six times more likely were they from the same than different speakers, assuming H_a = B, and about seven times more likely, assuming H_a = B + G. Since it is known that the data are in fact from the same speaker, this is an encouraging result. Note, however, that this is not the case with the F2 results for C1 vs C2, where the difference between the values is in fact marginally more typical for different speakers. When the values for all three formants are combined, in the right-most column, the difference is clearly considerably more likely assuming same-speaker provenance, consistent with the known facts. (The combined LR is the product of the individual LRs assuming independent evidence; the LR DS values must be converted back to their proper, reciprocal form.) As far as the overall known-suspect comparisons are concerned, table 3 shows that they are all more likely assuming that the data have come from the same speaker. The result is the opposite with the comparison between the offender and suspect samples, where the combined LRs of 6 to 8 indicate weak support for the defence hypothesis that they have come from different speakers. Note again that the differences between the F3 values are more likely to have been observed assuming same-speaker provenance.

It is necessary to point out that, for several reasons, this is actually a very crude estimate of the LR for this small piece of evidence. Firstly, the samples have been compared with respect to F-pattern at only one point in the vowel! (Comparison at other points is difficult because of lack of reference data.) Figure 2 shows, however, that there are differences in F-pattern throughout the formants' time course, so the LRs would probably show greater support for the defence hypothesis. Next must be reiterated the shortcomings in the assumptions underlying the choice of the LR formula. Thirdly, the reference data are not totally comparable to the forensic data: the reference data are for stressed /ɐ/ vowels before a final alveolar consonant as in *hut*, whereas the /ɐ/ vowel in the samples occurs before a velar. I suspect that the large difference between the samples' F3 measurements and those of the reference data are partly referable to this. Finally, in simply taking the product of the LRs to estimate a combined LR, no account has been taken of possible correlations between different formant measurements.

8.2 An acoustic-non-linguistic feature

An acoustic-non-linguistic feature often used in forensic comparison is long term average F0 (LTF0). This is non-linguistic because it reflects both Intrinsic Indexical features like length and mass of the cords, and state of health, as well as non-linguistic aspects of Communicative Intent like Affect and Self-presentation (the capitalised terms are part of an explicit model for the information content in a voice [19], [1] - a third conceptual framework which, together with BT and Linguistics, underlies TFSI). Figure 3 represents a forensic comparison between suspect and offender in LTF0, again using kernel density estimation. The language is Cantonese. The suspect's LTF0 is the mean of 14 phone calls; the offender's value is from one phone call adjudged long enough to provide a good estimate of their LTF0 [20]. The reference distribution is from means of 17 Cantonese males speaking over the phone. The 2.3 Hz difference between the offender and suspect LTF0 is extremely small - it represents only about 2% of a male Cantonese speaker's typical range (2 * LTF0sd). However, the values also lie near the mode and are thus fairly typical. One would only be about twice as likely (LR = 2.3) to observe this difference were the samples from the same speaker. This is a good example of why similarity between samples is only half the story in forensic comparison.

8.3 Auditory features

There is effectively a limitless number of potential auditory features that can be used in the forensic comparison of speech samples. Table 3 contains some typical examples of differences observed between offender and suspect samples in a case involving Chinese [2]. (It is worth noting that the voice in both

Table 3: Example of auditory-linguistic comparison of forensic voice samples in Putonghua (Standard Chinese).

	Suspect's samples			Offender's samples	
	[utterance]	*gloss* (Pinyin)		[utterance]	*gloss* (Pinyin)
1	ni ɕɛn tɕʰy ba	*better you go first* (nǐ xiān qù ba)		tɐŋ iɕa ɕɛn a	*wait a bit first* (děng yíxià xiān a)
2	ʂwɔ	*speak* (shuō)		swɔ	*speak* (shuō)
3	ʂʐwu	*fifteen* (shíwǔ)		szxɔʊ	*time* (shíhou)
4.	na	*In that case* (nà)		lali	*where?* (nǎlǐ)

samples sounded very similar in general phonetic features like overall pitch and phonation type - similarities that one would perhaps be more likely to observe were they from the same speaker.)

The first example in table 3 is of differential placement of the time adverb xiān *first*: pre-verbal in suspect sample; post-verbal in offender. Examples 2 and 3 are of a consistent difference between a word-initial retroflex fricative (suspect) and an alveolar fricative (offender). This reflects a more general phonological situation where the offender's sample lacks a whole set of phonemic contrasts between retroflex and alveolar initials that is present in the suspect sample. Example 4 shows a correspondence between alveolar nasal (suspect) and lateral (offender). The features in the offender sample are in fact typical of a Cantonese speaker speaking Standard Chinese (Cantonese does not have retroflex consonants, typically lacks [n], and puts time adverbs after the verb). It is known that the suspect was born and grew up in Peking, where they do have retroflexes and [n] and prepose time adverbs, and his speech reflects that.

It is difficult, though not impossible, to conceive of a situation where the same speaker might show these features in two different conversations. However, it is clear that these auditory-linguistic differences would be far more likely under the assumption that the samples had come from different speakers. This would be enough to balance the "same-speaker" LR that might come from consideration of the abovementioned similarity between the samples in voice quality features.

9. Testing

The well-known USA Supreme Court *Daubert* rulings on admissibility of scientific evidence [21] include, as one criterion, whether the theory or technique can be, and has been, tested. In Federal and State Australian courts the practice notes requiring reliability, replicability and transparency on the part of expert testimony are *de facto* adoptions of *Daubert*. It is a natural question, therefore, to ask to what extent the Bayesian approach to FSI outlined above has been tested.

The idea of testing a theorem is not coherent since it does not possess the property of being wrong, and its truth is guaranteed [8]. Rather, it is that part of the analytical approach which has to do with the extraction and quantification of the differences between the samples that can be tested. Given that the LR of BT is predicted to be greater than unity for same-subject data, but less than one for different-subjects, it can be used as a discriminant distance around a threshold of 1, and the evidence consisting of known same-speaker and different-speaker pairs tested to see to what extent they are correctly resolved - a relatively straightforward discrimination between same-speaker and different-speaker pairs. An experiment of this kind was recently carried out with non-contemporaneous phone recordings from 60 Japanese males, using only three phonetic segments - a vowel [ɔː], a voiceless fricative [ɸ] and a nasal [ŋ]. LRs were estimated for two kinds of analysis commonly found in FSI - F-pattern and cepstrum. Both analyses yielded useful strengths of evidence, but the automatic approach, not surprisingly, was stronger on average by a factor of 18. With formants, a LR bigger than unity was on average about 50 times more likely if the samples were from the same speaker; with the cepstrum, LR > 1 was about 900 times more likely. The strength of evidence associated with LR values of this magnitude is characterised as "moderate" and "moderately strong" respectively [10]. Similar strength of evidence results for formants have been obtained in another similar LR-based experiment with Japanese, using fewer (10) male speakers, but only six formants. In this experiment, LR > 1 was about 30 times more likely with same-speaker pairs [22]. A LR>1 was found to be 50 times more likely with same-speaker pairs in another experiment with 11 Australian males, discriminated only with respect to their F2 in the five long phonemic monophthongs [11].

Auditory-linguistic and non-linguistic features in the word *okay* in Australian English have also been tested in this way, and found to yield low but useable LRs [23]. The features, which are categorical, were observed in the spontaneous speech of ten young Australian males in two separate conversations separated by at least several months, and include: palatalisation of /k/ to fronted velar [c]; frication of /k/ to velar fricative [x]; voicing of /k/ to [g]; realisation of first diphthong /oʊ/ as [ɛɪ]; nasalisation of diphthongs; use of creaky voice. The LRs associated with these features are such that a certain incidence of, say, palatalisation in both offender and suspect samples would be nine times more likely if they were from the same speaker. Since the same average LR was found for the creaky voice feature, two samples both with creaky palatalised *okay*s would be 81 times more likely assuming same speaker.

The experiments just mentioned primarily investigate the use of linguistic features in LR-based discrimination. There is of course ample demonstration from a long line of ever diminishing EERs in automatic verification experiments - for example the NIST evaluations - that same-speaker pairs can be discriminated from different-speaker pairs with considerable reliability, even under fairly tough conditions. Some of these experiments, e.g. [24], [25] have used explicitly Bayesian methods.

One aspect of the LR that *is* open to testing is the nature of the formula used, and it is of obvious interest and importance to see to what extent the various formulae cope with speech.

10. Summary

This paper has discussed some important aspects of Technical Forensic Speaker Identification, focusing on both the necessary logical framework for evaluation of forensic speaker identification evidence, and how non-automatic methods, using higher-level linguistic knowledge, can be of forensic use. The main message, I think, given the excellent performance of automated systems, is nevertheless that not all evidence is being exploited in estimating Likelihood Ratios. Some fruitful collaboration is in order, envisaged perhaps by the recent name change from the *International Association for Forensic Phonetics* to the *International Association for Forensic Phonetics and Acoustics.*

11. Acknowledgments

I thank Dr. James Robertson, Head of the Forensic Service of the Australian Federal Police, and Hugh Selby, Reader in Law at the Australian National University, for useful discussions about evidence. The views expressed in this paper are of course my own.

12. References

1. Rose, Philip, *Forensic Speaker Identification*, Taylor & Francis *Forensic Science* Series, London & New York, 2002.
2. Rose, Phil, *The Technical Comparison of Forensic Voice Samples,* Issue 99, *Expert Evidence*, (series eds. Freckelton, I. & Selby, H.), Thomson Lawbook Company, Sydney, 2003.
3. Lindley, D. V., "Probability", in Aitken & Stoney (eds.) *The Use of Statistics in Forensic Science*, Ellis Horwood, Chichester: 27-50, 1991.
4. Bernado, José M., "Bayesian Statistics", in *Encyclopedia of Life Support Systems* UNESCO, 2001.
5. Champod, C. & Evett, I., Commentary on Broeders (1999), *Forensic Linguistics* 7/2: 238-43, 2000.
6. Broeders, A.P.A., "Some observations on the use of probability scales in forensic identification", *Forensic Linguistics* 6/2: 228-41, 1999.
7. Aitken, C.G.G., *Statistics and the Evaluation of Evidence for Forensic Science*, Wiley, Chichester, 1995.
8. Robertson, B. & Vignaux, T., *Interpreting Evidence*, Wiley, Chicester, 1995.
9. Lindley, D.V., "A problem in forensic science", *Biometrika* 64/2:207-13, 1977.
10. Rose P., Osanai, T. & Kinoshita, Y., "Strength of Forensic Speaker Identification Evidence - Multispeaker formant and cepstrum based segmental discrimination with a Bayesian Likelihood ratio as threshold", *Forensic Linguistics* 10/2: 179-202, 2003.
11. Alderman, T., *Refining the Likelihood Ratio Approach to Forensic Speaker Identification - Effects of Non-Normality in the Background Distribution as Modelled with the Bernard Data for Australian English*, Unpublished First-Class Honours Thesis, Australian National University, 2004.
12. Künzel, H.J. & Gonzalez-Rodriguez, J., "Combining Automatic and Phonetic-Acoustic Speaker Recognition Techniques for Forensic Applications", *Proc. 15th ICPhS*: 1619 - 22, 2003.
13. Nolan, F., "The Limitations of Auditory-Phonetic Speaker Identification", in Kniffka H. (ed.) *Texte zur Theorie und Praxis forensischer Linguistik*, Max Niemayer Verlag, Tübingen, 1990.
14. Nolan, F. & Oh, T., "Identical Twins, Different Voices", *Forensic Linguistics* 39: 39-49, 1996.
15. Rose, P. & Simmons, A., "F-pattern Variability in Disguise and over the Telephone: Comparisons for Forensic Speaker Identification", in McCormack P. & Russell A. (eds.) *Proc. 6th Australian Intl. Conf. on Speech Science and Technology*, Australian Speech Science & Technology Association, Canberra, 121-6, 1996.
16. Künzel, H. J., "Beware the telephone effect: the influence of transmission on the measurement of formant frequencies", *Forensic Linguistics* 8/1: 80-99, 2001.
17. Elliott, J., "Auditory and F-pattern variations in Australian *okay*: a forensic-phonetic investigation", *Acoustics Australia* 29/1: 37-41, 2001.
18. Bernard, J.R.L., *Some measurements of some sounds of Australian English*, unpublished Ph.D. thesis, Sydney, 1967.
19. Nolan, F., *The Phonetic Bases of Speaker Recognition*, CUP, Cambridge, 1983.
20. Rose, Phil, "How effective are long-term mean and standard deviation as normalisation parameters for tonal fundamental frequency?", *Speech Communication* 10: 229-24, 1991.
21. Daubert, *Daubert v. Merrell Dow Pharmaceuticals, Inc.* 113 S Ct 2786, 1993.
22. Kinoshita, Y., "Use of likelihood ratio and Bayesian approach in forensic speaker identification", in C.Bow (ed*.) Proc. 9th Australian Intl. Conf. Speech Science and Technology*, Australian Speech Science & Technology Association: 297-302, 2002.
23. Elliott, J., *Okay, what are the odds?* unpublished M.Phil. Thesis, Australian National University, 2002.
24. Meuwly, D, & Drygajlo, A. "Forensic speaker recognition based on a Bayesian framework and Gaussian Mixture Modelling (GMM)", *Proc. 2001 Speaker Odyssey - Speaker Recognition Workshop*: 145-50, 2001.
25. Gonzalez-Rodriguez, J., Ortega-Garcia, J.,& Lucena-Molina, J.J., "On the application of the Bayesian framework to real forensic conditions with GMM-based systems", *Proc. 2001 Speaker Odyssey - Speaker Recognition Workshop*: 2001.