CATCHING CRIMINALS BY THEIR VOICE – COMBINING AUTOMATIC AND TRADITIONAL METHODS FOR OPTIMUM PERFORMANCE IN FORENSIC SPEAKER IDENTIFICATION

Australian Research Council Discovery Project DP0774115

Dr. Philip Rose, Principal Investigator

School of Language Studies, Australian National University

1 AIM

Consider the following scenario. A fraud involving millions of dollars is perpetrated over the phone. Because such large transactions are monitored by the financial institution, the police have a recording of the fraudster's voice. They also have a suspect, and several intercepts of his voice on the phone. They want to know whether the suspect's and offender's voice come from the same person.

This is a typical example of Forensic Speaker Identification (FSI), taken from my case-work. One or more speech samples of a known voice are compared with samples of unknown origin. The unknown samples are usually of the individual alleged to have committed an offence, and the known voice belongs to the suspect, defendant or accused. The interested parties (police, court, legal counsel) are 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 defendant as the offender, or exonerate them. At the moment, two very different, but complementary, approaches – Traditional and Automatic – are used to do this.

- **The aim of this project is to improve substantially on the methods used to identify criminals by their voice by combining these two approaches.**

2 BACKGROUND

The beginning of Forensic Speaker Identification (or Recognition – the terms are used synonymously) can be reasonably dated from its first institutional use, by Germany's Bundeskriminalamt in 1980. It has emerged as an applied discipline in the last 15 or so years in response to an ever-increasing demand from the legal profession and security agencies as the speech of more and more offenders, and terrorists, is recorded. During this time, FSI has experienced two major revolutions: in methodology, and in evaluation of evidence. Crucial to the understanding of the project, these two things must now be explained in some detail.

Evaluation of Forensic Identification Evidence: the Likelihood Ratio Over the last twenty or so years considerable attention has been focused on the proper, rationalist evaluation of forensic evidence. This is the result of the post-1968 "new evidence scholarship" debate and the increased incidence, from 1985 onwards, of DNA profiling and its subsequent statistical evaluation. Some spectacular miscarriages of justice due to incorrect statistical reasoning have also helped to bring about a revolution in the approach to evaluating forensic evidence. As a result, the Likelihood Ratio of Bayes' Theorem now plays a central role in quantifying evidential strength.

The idea behind the Likelihood Ratio (LR) is intuitive and easy to understand. It is the ratio of two conditional probabilities. One is the probability of getting the forensic evidence assuming the prosecution hypothesis is true – e.g. the defendant is guilty. The other is the probability of the evidence assuming the defence hypothesis. If you are more likely to get the forensic evidence assuming that the prosecution hypothesis is true than if the defence hypothesis is true – i.e. if the LR is greater than one – this counts as support for the prosecution. Values of the LR less than unity indicate support for the defence, and of course an LR of one means that you are just as likely to get the evidence under both hypotheses and that it is therefore useless.

The actual strength of the forensic evidence is reflected in the magnitude of the LR. An LR of 100 means you are 100 times more likely to get the evidence under the prosecution hypothesis than under the defence hypothesis. An LR of 10,000 means that the evidence is one hundred times stronger than that. Corresponding LRs for the defence would be 0.01 and 0.0001.

The LR is what the forensic expert, whether working on DNA, glass fragments, clothing fibres or speech, must try to estimate. This view continues to be stressed in the main textbooks on the evaluation of forensic evidence, e.g. Robertson & Vignaux (1995), or forensic statistics, e.g. Aitken & Stoney (1991), Aitken & Taroni (2004), Lucy (2005). It will also be found endorsed by the judiciary, e.g. Doheny (1996).

Estimating Likelihood Ratios for speech In FSI, the evidence is the ensemble of differences (or similarities, for similarities are just small differences) between the questioned and known speech samples (Rose 2002). Denote this evidence E_{fsi} (for *forensic-speaker-identification evidence*). Denote the prosecution hypothesis that the questioned and known speech samples were said by the *same speaker* H_{ss}; denote by H_{ds} the alternative, defence, hypothesis that the samples were spoken by *different speakers*. The LR for the forensic speaker identification evidence is then as given at (1) (p = probability, *"|"* = conditional upon).

$$\boxed{LR = \frac{p\,(E_{fsi} \mid H_{ss})}{p\,(E_{fsi} \mid H_{ds})} \quad (1)}$$

In order to estimate a LR it is necessary to have not only the questioned and known samples, but also a reference, or background, sample. This is because a LR is a ratio of *similarity* to *typicality*: it quantifies how similar the two samples are, and then evaluates that similarity with respect to *typicality*, i.e. how likely we would be to observe the samples in randomly selected pairs of different speakers from the relevant population (e.g. young male speakers of Hong Kong Cantonese). The more similar the samples are than typical, the greater the LR will deviate above unity, and the greater the support for the claim that they come from the same speaker. The more typical they are than similar, the greater the deviation below unity, and the greater the support for different speaker provenance.

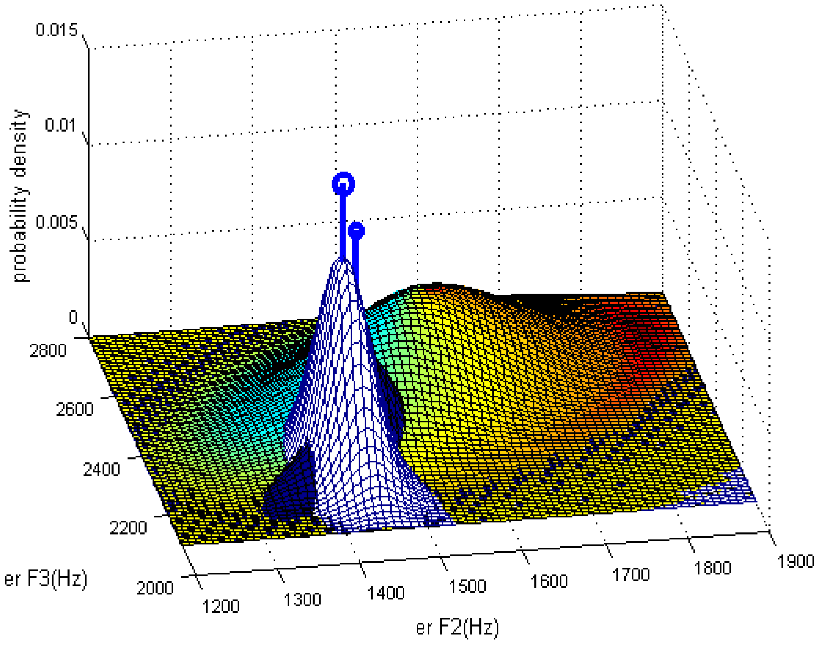

Figure 1. Joint bivariate normal probability density distributions for *er F2* and *er F3* in reference population, and in known (suspect) and questioned (offender) samples. Suspect = large ball on top; offender = small ball.

This is illustrated in figure 1 (from Rose 2006b), which shows the estimation of a LR from two speech samples. The two samples are being compared with respect to two acoustic features, called formants, from an *er* vowel (as in *bird*). These features are frequencies at which the air in the speaker's vocal tract was vibrating during the articulation of the vowel. The two features (labeled *erF2* and *erF3*) are plotted along the horizontal axes. The vertical axis can be thought of as probability. The flat distribution, more or less in the middle, is the reference sample, based on data from 57 speakers. The reference sample has to be chosen with respect to the alternative (defence) hypothesis. In figure 1 the known and questioned samples were typical of a male speaker with a so-called Broad Australian accent, so the reference sample had to be from Broad Australian males (this would correspond to a defence hypothesis: "the incriminating speech is not from my client but from another male speaker with a Broad Australian accent").

The two peaked distributions, to the bottom left, are the known and questioned samples, consisting of about 15 *er* vowel tokens each. It can be seen that the distributions of the questioned and known samples intersect – so they are quite *similar*. Moreover, their separate location relative to the reference sample shows that they are *atypical*: you would not expect them to be frequently sampled at random from this reference distribution. The ratio of *similarity* to *typicality* for this comparison is thus going to be above unity, and count as support for the prosecution hypothesis that the samples were spoken by the same speaker. The calculations of LRs for speech acoustics are very complicated and computationally intense, because they have to take into account the complicated variation of speech, for example, as here, that the two features are correlated. The LR for this comparison is 32.8, indicating that you would be about 30 times more likely to get the difference between known and questioned samples assuming they had come from the same speaker (as indeed they did).

Forensic Speaker Identification Methodology

There are currently two very different approaches used to estimate Likelihood Ratios (i.e. strength of evidence) in FSI: "Traditional" and "Automatic" (Rose 2003: 4069ff).

In a Traditional forensic comparison, the expert treats the speech signal analytically as the output of a vocal tract that is executing all the complex gestures that are required to make speech sounds, and convey a linguistic message. Thus samples are compared using features that relate in a relatively straightforward way to aspects of speech production, like what the speaker is doing with their tongue, or vocal cords, or lips, to produce a particular speech sound. It is common in Traditional approaches for samples to be compared with respect to the quality of particular vowels. For example, the first vowel in the word *fucken'* might sound as if it has been produced further back in the mouth in the offender than in the suspect samples (Rose 2006a). The acoustic frequencies corresponding to this difference would be ascertained, and a LR for this difference estimated. A comparison of this type was illustrated in figure 1.

The *comparanda* in a Traditional FSI approach are not necessarily exclusively speech sounds: they can include any feature of linguistic structure: whether or not the samples distinguish between *you* (singular) and *youse* (plural) for example. Traditional approaches rely heavily on expertise in Linguistics (the science of Language), and especially its sub-parts: Phonetics (how speech sounds are produced, transmitted acoustically, and perceived); Phonology (how speech sounds are organized in Language); Dialectology and Sociolinguistics (how Language varies as a function of geography and social structure respectively). The *locus classicus* for LR-based approaches in Traditional FSI is my textbook (Rose 2002). At the time of writing, the most up-to-date account of Traditional FSI is Rose (2006a).

In Automatic forensic speaker identification, speech is treated purely statistically, as a time-varying signal, with no special attention to any particular linguistically meaningful sub-part or deliberate production thereof. The features used to compare samples do not relate in any meaningful way to individual sounds, or words or structures, but are mathematical abstractions that best model, or account for, the fluctuations in signal amplitude as a function of time. The state-of the art approach involves the Gaussian Mixture Modeling of a set of cepstral coefficients, usually mel-weighted. These signal processing techniques are intimidatingly complex and cannot be explained here. The underlying theory is taken from signal detection. The most up-to-date account of Automatic forensic speaker identification at the time of writing is Gonzalez-Rodriguez et al. (2006).

The origins of the Automatic approach lie in attempts to get computers to perform speech and speaker identification automatically. An Automatic approach is often seen as desirable because it is assumed that it avoids the so-called subjectivity associated with a hermeneutic approach like Linguistics, Phonetics or Phonology. Furthermore, it theoretically makes it possible to do FSI without the extensive training in Phonetics, Phonology, Dialectology and Sociolinguistics required for Traditional approaches.

The two approaches differ in many other respects, discussed in detail in Rose (2006a). Here are the ones that are relevant for this proposal.

Discriminant Power Automatic features are very much more powerful as evidence: they will, on average, yield likelihood ratios that deviate much more from unity. I was the first to show this in Rose et al. (2003), where it was found that analyses with both Traditional and Automatic types of feature yielded useful strengths of evidence, but the Automatic approach was stronger on average by a factor of 18. With Traditional features (vowel acoustics), an LR bigger than unity was on average about 50 times more likely if the samples were from the same speaker; with the Automatic approach, LR > 1 was about 900 times more likely.

Globality Automatic approaches are *global*. They take into account the speech as a whole. Traditional approaches are *local*: they work with a few features extracted from the samples, for example the acoustics of a subset of the vowels. From the point of view of time, being able to zap the whole of the available speech material is a big advantage, since time is not used in listening to the samples for suitable segments to quantify and compare. One disadvantage of the global approach is that it requires a minimum of data. One of the currently competitive FASR systems must have for example at least 60 seconds of net speech from each of the suspect, offender and reference speakers. This is actually a lot: quite often the speech available in real case-work (usually involving phone calls between males) is considerably less. There is no such minimum requirement in Traditional approaches (although, of course, the more speech available, the more likely will be the chance of finding comparable material). For example, the samples to be compared may be much shorter than 60 seconds, but nevertheless contain several tokens of an abnormal *r* sound, the word *fucken*', or pause particles like *er….* or *ummm.* , which could enable a useful estimate of a LR.

Channel Sensitivity The main drawback of Automatic approaches is that they are extremely channel-sensitive. The handsets, the transmission pathway and characteristics, the mobile vs. landline connection, the type of data compression: all these have a substantial effect on the acoustics. The effect is great enough to seriously compromise the approach, were it not for channel normalizing techniques like cepstral subtraction. Even then, it would be very complicated, for example, to compare a police interview recording of the suspect made directly onto a cassette tape with an intercept of the offender's voice from their mobile. Questioned, suspect and reference data have to been recorded under exactly the same conditions for the comparison to work properly, and often it is not possible to even find out from the police what the conditions of the recordings were – for example what kind of data compression, if any, was used. In such cases, complicated compensatory techniques have to be used (Gonzalez-Rodriguez et al. 2006). Traditional features are more robust, and not so severely compromised by telephone transmission, although all the acoustics are always affected to some extent (Rose 2003: 5101-5113). Some Traditional features not dependent on acoustics, like what kind of an *r* is being used, are not affected at all.

As can be seen from the above, both approaches have their strengths and weaknesses, and are in many respects complementary. The main message, given the excellent performance, yet extreme channel sensitivity, of automated systems, is nevertheless that *not all evidence is being exploited in estimating Likelihood Ratios* (Rose 2006a). It is clear that the Traditional approaches, which lack the globality of the Automatic approaches, are not extracting all information relevant to the estimation of a LR. It is equally clear that Automatic methods, which by definition do not take true higher level linguistic or paralinguistic information into account, will be missing information of evidentiary value, since it has been shown that this higher level information can furnish on its own strong LRs in support of either defence or prosecution. Consider, for example, a case where two samples are from different speakers who have very similar global acoustics, like some identical twins, but where one twin consistently uses a funny *r* sound (technically a labio-dental approximant). It is likely that a global Automatic approach, which cannot focus on single speech sounds, will evaluate the difference between the two samples as more probable assuming they have come from the same speaker. A Traditional approach would not make this mistake. If it is conceded that the aim of FSI is to estimate the strength of evidence with a LR, then a LR must be estimated for *all possible information in a FSI case*. A complete and proper integration of both Traditional and Automatic approaches to FSI is clearly the way to go. The result will be potentially even more powerful and more accurate LRs, consequently more reliable FSI.

Showing that it works

In 1993 the USA Supreme Court ruled in *Daubert* that for scientific evidence to be admitted, the theory or technique in question must be testable, and has been tested (Daubert 1993). 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*. Thus the testing of approaches with the appropriate LR-based methodology is crucial in the real world. It will determine whether the results of a FSI case actually make it into court, or, if they have, whether they are subject to appeal.

The LR-based approach to FSI is tested in the following way. Given that the LR 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 the appropriate threshold (i.e. LR =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 pairs and different-speaker pairs (Rose 2002). In order to do this, speech samples are recorded from many speakers on two different occasions. Denote these samples as Sp1.1 Sp1.2, Sp2.1 Sp2.2 … Sp*n*.1, Sp*n*.2, where Sp = speaker, and 1 & 2 = the two occasions. (So Sp1.1 refers to the speech sample collected from speaker 1 on the first occasion, and Sp1.2 means their sample on the second occasion.) Comparisons are then made, using a LR as a discriminant function, for all **same-speaker pairs** – e.g. Sp1.1 and Sp.1.2; and for all **different-speaker pairs**, e.g. Sp.1.1 and Sp2.1; Sp.1.1 and Sp.2.2 etc. If a particular pair is evaluated with a LR greater than unity and it is a same-speaker pair, the pair is correctly discriminated. If it is evaluated with an LR lower than unity it is incorrectly evaluated. The same applies *mutatis mutandis* for different speaker pairs.

The extent to which the same-speaker pairs can be correctly discriminated from different-speaker pairs shows how well the approach works. For example, one might find for a particular method that all same-speaker pairs were correctly classified as same-speaker pairs, and 89% of different-speaker pairs correctly classified as different-speaker. Such a situation is shown in figure 2, in order to make the approach easier to understand. Figure 2 shows a so-called Tippett, or reliability, plot (the current way of displaying the results of a FSI discrimination). It is taken from a FSI experiment to investigate how well same-speaker pairs could be discriminated from different speaker pairs using Traditional vowel acoustics (Rose 2006c). The dashed line shows results for 110 different-speaker trials; the solid line shows results for eleven same-speaker trials. The horizontal axis shows logLR values (so the threshold is 0, not unity); the vertical axis shows cumulative probability. The graph shows for what proportion of same- or different-speaker trials one observes a Likelihood Ratio bigger than a given abscissa LR value. It can be seen that about 10% of different-speaker trials were incorrectly evaluated, with LRs greater than threshold, so about 90% of different-speaker trials were correctly discriminated. All same-speaker trials were correct, with LRs bigger than threshold.

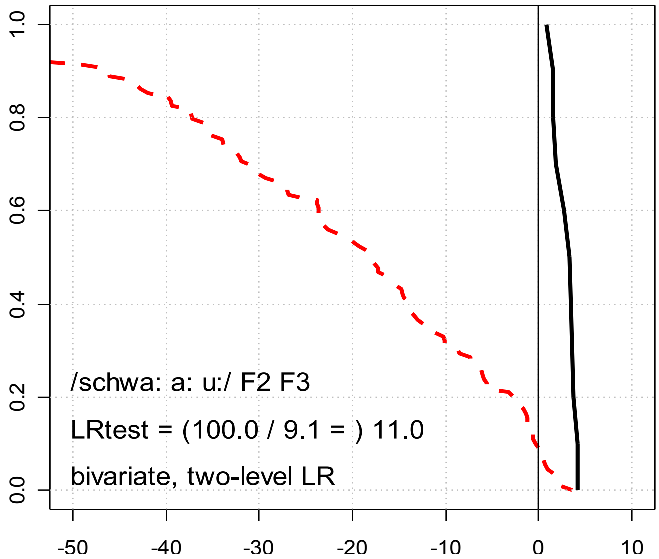

Figure 2. Example of Tippett plot for LR-based discrimination (after Rose 2006c).

The results thus reflect the reliability of the approach (here, comparing speech samples with respect to LRs derived from the acoustics of some of their vowels). Plots like these are used as follows. Suppose that in a given case a logLR of, say, 3.5 was obtained using the vowel acoustics (i.e. the evidence is 3,162 times more likely were the samples from the same speaker). One could then report to the court that, on the basis of this experiment, you would expect a logLR value of 3.5 in about 98% same-speaker comparisons and in about 2% of different-speaker comparisons. Since one would be about 50 times more likely to get the evidence assuming that the sample had come from the same speaker than from different speakers, the result offers moderate support for a prosecution hypothesis that the samples had in fact come from the same speaker. The court may want to pay particular attention to the proportion of so-called 'judicially fatal' *incorrectly evaluated different-speaker trials* – for the case of LogLR = 3.5, that would mean that there is still a 2% chance that a different speaker is involved where the test claims the opposite.

This approach has already been used in forensic discrimination experiments with forensically common materials, e.g. DNA (Evett et al. 1993), refractive indices for glass (Aitken & Lucy 2004), and is well-established. It has been tested on speech in a suite of experiments with both Automatic and Traditional approaches, e.g. Kinoshita (2000); Meuwly & Drygajlo (2001); Rose (2003); Rose Osanai & Kinoshita (2003); Leeuwen and Bouten (2004), Alderman (2005); Gonzalez-Rodriguez et al. (2006), where is has been shown to be successful, discriminating different-speaker pairs from same-speaker pairs with acceptable and specifiable levels of reliability. I have also been using it in real-world forensic case-work for some time. Being able to combine information from both Traditional and Automatic approaches can be expected to result in a considerable improvement over these results in discriminatory power, both for same-speaker and different-speaker pairs.

3 SIGNIFICANCE and INNOVATION

I don't think the significance of this project needs belabouring. It is important for the successful functioning of our Criminal Justice System, and, nowadays, for our national security. You simply need to be able to correctly convict the offenders and exonerate the innocent. FSI is also still far from an optimum solution: this experiment will improve the situation. The approach will also be able to be used in FSI with other languages (the CI is often asked to do case-work on Chinese dialects), and other dialects of Australian English. For example, so-called 'woglish' – the accent of first generation Australians of parents from the Mediterranean area, especially Greece, Italy, and the Middle East - is becoming more frequent in FSI cases.

As far as innovation is concerned, the idea of combining Traditional and Automatic approaches is not new (although this does not mean it is not a very good idea!). It has already been suggested (Künzel & Gonzalez-Rodriguez 2003: 1619). Suggested, but not demonstrated: Künzel & Gonzalez-Rodriguez simply *compared* the *outcomes* of two cases using Traditional and Automatic approaches, and showed that they corroborated each other. They did not *combine* them. It is also the case that Automatic approaches are beginning to find an improvement in performance if they incorporate some very low-level Traditional features, like fundamental frequency parameters (Reynolds et al. 2003).

That no-one has yet undertaken the proper combination of both approaches is firstly because there are very few practitioners in the world who are properly trained in Traditional FSI. Secondly, it is because, having been responsible for showing how to properly implement the LR-based approach to traditional features in FSI, only the CI (and his forensic-phonetic research students) know how to do it properly (for published results in two books and several papers, see references immediately above).Thirdly, we have been lucky in Australia in having an extensive data-base of formant frequencies from about 170 speakers to use as a reference population for testing.

4 APPROACH and METHODOLOGY

General The project will involve obtaining speech samples from a large number of Australian males on several different occasions and under forensically realistic conditions. Three forensic discrimination tasks will then be carried out between same-speaker and different-speaker pairs of samples, using Likelihood Ratios as a discriminant distance. The first two tasks will implement Traditional and Automatic approaches. The last will combine both Automatic and Traditional approaches. This can be done by simply taking the product of the LRs from both analyses. The easy combination of evidence from different sources is one of the beauties of the LR-based approach (Rose 2002). At least, it is easy if the evidence is independent: it will be necessary to remove any speech used in the Traditional analysis from the speech used in the Automatic analysis to avoid problems in LR estimation from correlation of features (Rose 2006c, Rose et al. 2004). It is expected that there will be a considerable improvement in discrimination over the performance from the separate approaches. The components of this experimental design are described and justified below.

Speakers *Sex* Although it would be desirable to sample both sexes, the vast majority of crimes committed where FSI is required (armed robbery, blackmail threats, bomb threats, murder threats, drug offences) are by males. Male data will therefore be acquired and tested. *Accent & age* It is customary, after Mitchell and Delbridge (1965), to classify Australian accents into three groups - Broad, General and Cultivated - mostly on the basis of vowel, especially diphthongal, quality. A recent study on the vowel acoustics of groups of designated Broad General and Cultivated speakers (Harrington et al. 1997) showed that this auditory classification does have some basis in acoustic fact, but it is probably the case that for many vowel acoustics these are positions on a continuum (Horvath 1985: 68). The commonest accents encountered forensically are in the Broad to General range, but it is not forensically realistic to test such a wide range, because many of the different-speaker pairings in it will be of different-sounding speakers who should be easy to discriminate. For this reason, only speakers will be used who can be characterised as having a General accent, and, as accent can vary with age, aged between 20 and 40. This will increase the proportion of similar-sounding speakers to be discriminated, and make the experiment both more realistic and more demanding. *Number* Estimating sample-size is important in forensic statistics, where cost has to be balanced against accuracy (Lucy 2005: 179-189). The number of subjects used determines the number of same-speaker and different-speaker comparisons, and this in turn determines the accuracy of the estimate of the reliability of the method (as in the example in figure 2 above, where there was a 2% probability of getting a LogLR greater than 3.5 with different-speaker comparisons). The confidence limits for a parameter decrease as a function of number in sample, so it is obviously important to have as many speakers as possible.

50 speakers is a reasonable lowest limit, taking into consideration the three most important factors, viz: the number of same-speaker pairings; the number of different-speaker pairings; and the number of speakers constituting the reference sample. Modeling the data with the appropriate beta(α, β) distribution, as recommended by the *European Network of Forensic Institutes*, shows that, if you were to get a 95% correct discrimination with 50 speakers, you could be 95% sure that the actual discrimination was **at least 89.2%**. Increasing the number of speakers to 60 would result in an improvement of less than 1% in the lower confidence limit (to 89.6%), whereas lowering to 40 would lower the lower confidence limit by just less than 1% (to 88.4%). Using 50 speakers also of course means a precision of only +/- 2% in the same-speaker results. It should not be allowed to get any less than this by having a smaller number of speakers: it would otherwise be difficult to justify to a court.

From the point of view of protecting the innocent, the number of different-speaker pairings is the most important, since that will determine the accuracy of the estimate of the proportion of incorrectly evaluated different-speaker pairs. 50 speakers means (n*[(n-1)/2]) = 490 different-speaker comparisons. With this number, there is a relatively high lower confidence limit. For beta(465.5, 24.5), you can be 95% sure that, if you got 95% correct discrimination of different-speaker pairs, the actual proportion of correct decisions would be at least 93%.

The number of speakers sampled for the reference data must be large enough to adequately reflect the background population, to ensure that the LR estimates are as accurate as possible. It is usually assumed that the population mean and standard deviation will be closely approximated by a so-called ‘large sample’, with 30 or more observations. However, this is for a normal distribution, and it is known that many Traditional features are not normally distributed. A 50 speaker data-base should be large enough, just, for accurate kernel-density modeling of the distributions of the reference sample, with LR comparisons then being done on a ‘leave-one-out’ basis. As an additional check, however, the data from the 60 or so General speakers in the Bernard data-base can also be used as a reference population for the Traditional analysis.

The necessary steps will be taken to preserve subjects' anonymity, commensurate with ethical guidelines: the project will be thoroughly vetted by the ANU's *human research ethics committee*.

Number of recordings It is a phonetic truism that an individual's speech is never invariant, but changes from occasion to occasion. Generally, the greater the time separation between two speech samples, the greater the differences between them (I demonstrated this in a forensically motivated study of long- and short-term variation in Traditional features (Rose 1999)). This means that generally it will easier to forensically discriminate between two speech samples recorded with only a short separation in time than with a longer time separation. In the real world, it is usually known fairly precisely when a forensic speech sample was spoken, and thus the time separation between samples can be quantified. Forensic speech samples can be effectively contemporaneous, as for example when an offender makes two phone calls, one timed immediately after the other. The time between samples can also of course be much longer – of the order of months or years. It is clear that, in order for this project to be forensically realistic, it must take into account within-speaker variation over time. Discrimination of speech samples must therefore be attempted under three degrees of time separation: **contemporaneous, short-term** and **long-term**. Thus three sets of speech samples need to be collected from each speaker. There is no non-arbitrary division of the time separation continuum. It is proposed that the first two recordings can be separated by about two weeks to give short-term comparisons; in order to allow long-term comparisons within the time-frame of the experiment, recordings 2 and 3 must be separated by at least several months. Sufficient data will be collected in one of the recordings to enable contemporaneous comparison. It is expected that the same-speaker discriminant performance will decrease, the greater the time separation of the recordings.

Elicitation In order to conform to forensic realism, it is vital to obtain natural speech, and enough speech to ensure that sufficiently representative mean values for desired parameters can be extracted. It is also necessary to control for content, so that, say, vowels of a particular type can be elicited. Not all vowels have the same individual identifying potential, so it is important to use those that do. Elliott (2001) has shown how the Edinburgh 'map' task can be well adapted to eliciting natural, yet controlled, data for Traditional forensic testing. This involves giving the subject a map and asking them to explain how to get from a to b. The important references are names containing the desired vowels, so for example if one wanted to elicit the *er* vowel in a stressed syllable, *Sherbrooke* street might appear on the map; or a B.P. service station will give two nice examples of the "ee" vowel. In addition, a suspect is often asked in a police interview to spell their name, or the name of the place where they live. This can also be required of the experimental subject in order to elicit the desired number of vowel tokens. The vowels that will be used in this experiment are the five long monophthongal phonemes of Australian English /iː aː oː əː ʉː/, the diphthongs /aɪ aʊ ɛɪ aʉ ɔɪ ɪə jʉ/; the short vowels /ɪ ɛ æ a/, and the sonorants /m/ /n/ and /l/. The individual-identifying potential of diphthongs and sonorant consonants has not yet been ascertained for LR-based FSI, and constitutes a novel, and important aspect of the experiment (since these sounds are common in forensic speech samples). All tokens will be in stressed position. Ten tokens per type will be collected. The analyst will be free to make use of whatever other linguistic or paralinguistic information is present in the recordings.

As far as data for Automatic extraction is concerned, it will be sufficient to engage the subject in conversation for a long enough period of time to obtain several minutes of net speech.

Recording, digitisation & processing The studio of the A.N.U's School of Language Studies will be used for recording speakers. Data will be digitized straight onto computer. Automatic systems require fairly low sampling frequencies e.g. 8 kHz, However, a Traditional analysis might require access to more than a nyquist of 4 KHz of information, so a sampling frequency of 12 kHz will be used. In case-work, there is effectively no control over forensic recording conditions. This, however, is the one aspect where experimental conditions must deviate from realism, in order to exercise adequate control. Once the data are digitized, they can, if desired, be put through various randomly generated filters, with or without noise, to simulate the effect of uncontrollable telephone transmission. It will then be possible to compare the relative performance of Traditional and Automatic approaches under differing degrees of noise. This in turn will make it easier to decide in noisy case-work whether it is prudent to proceed with an Automatic analysis in additional to the Traditional.

Further analysis can be done on P.C. For the Traditional approach, standard speech acoustic extraction software like *Praat* will be used. For the Automatic approach, the Spanish BATVOX Forensic Automatic Speaker Identification software will be used. The consistently good performance of this software has been documented in the USA's National Institute of Standards and Technology evaluations, both for normal and forensic speaker identification (Gonzalez-Rodriguez et al. 2006). The CI is collaborating in the development of this software. Kernel density Likelihood Ratios will be estimated for Traditional features using LR estimation software developed at Edinburgh University's *Joseph Bell Centre for Forensic Statistics and Legal Reasoning*. The CI has close and continuing connections with this centre, having been a British Academy Visiting Professor there in 2004, and having collaborated in publications. The LR programs will be tailored to forensic comparison of speech sample acoustics, being able to take into account the four features of speech which make computing a LR extremely complex, namely: correlation between variables, three levels of variance (between-speaker, within-speaker, within-speaker~between occasions), unequal variance, and non-normal distribution (Rose 2006a). Earlier versions of the programs have already been successfully trialed (Rose et al. 2004).

Schedule Data acquisition and preparation can begin immediately, will take three months and will be completed by the end of the first year. Traditional analysis can start as soon as the data becomes available. It will take a year to process each of the recordings and do the discrimination, thus three years in all. The Automatic analysis can be done in the second year. Combining the results, and final write-up, will be done in the fourth year.

5 NATIONAL BENEFIT

Because the project will show how better to forensically discriminate between voices, it will have the following significant social and economic outcomes for Australia. It will …

- *improve the equity of the Criminal Justice System* in voice identification cases, which are at present of necessity biased against conviction.
- *reduce the room for potential divergence in opinion* between prosecution and defence forensic experts by reducing the current amount of necessary 'guess work' informing those opinions.
- result in *substantial* savings in Court time and expenditure and thereby increase the efficiency of the Criminal Justice System.
- result in *substantial savings in Legal Aid funding* will result from the decrease in time required for an expert to carry out forensic speaker identification analysis.

It is also clear, of course, that the project has direct relevance to counter-terrorism, and national security. Was that again the voice of Bin Laden in January 2006? The CIA said yes, but their decision was based, incredibly, on an outdated and discredited method of speaker identification using 'voiceprints', incalculably inferior to the methods proposed here. Another recent case in point is the 2005 "terrorist video clip" of a male speaker with an Australian accent promising retribution in the name of Islam. If a suspect were identified, a forensic speaker identification could be carried out, and its results would be far more reliable than previously attainable.

6 COMMUNICATION OF RESULTS

Results need to be disseminated in several different scholarly fora. Submissions will be to *International Journal of Speech Language and the Law*; *Australian Journal of Forensic Sciences*;

Speech Communication; *Journal of the International Phonetic Association.* The CI will write papers to be given at four international conferences: in 2007 (International Association of Forensic Phonetics & Acoustics), 2008 (International conference on Spoken Language Processing), 2009 (International Association of Forensic Phonetics & Acoustics) and 2010 (International Australian Conference on Speech Science and Technology). As chair of the *Forensic Speaker Identification Standards Committee* of the Australian Speech Science and Technology Association, I will arrange for media releases at the appropriate times.

7 ROLE of PERSONNEL

This project breaks down neatly into five tasks. (1) data collection and preparation; (2) discrimination with Traditional methods; (3) discrimination with Automatic methods; (4) discrimination with combined Automatic and Traditional. (5) write-up & dissemination. The CI will be responsible for the last two tasks, and any training of personnel in the others. The first three tasks will be the responsibility of three separate individuals. This will ensure that the discriminations are completely independent.

8 REFERENCES

Aitken CGG & Stoney DA (1991) *The Use of Statistics in Forensic Science*, Chichester: Ellis Horwood.

Aitken, CGG & Taroni F (2004) *Statistics and the Evaluation of Evidence for Forensic Scientists*, Chichester: Wiley.

Aitken CGG & Lucy D (2004) 'Evaluation of trace evidence in the form of multivariate data', *Applied Statistics*, 53(4):109-122.

Aitken CGG, Lucy D, Zadora G. & Curran J.M (in press) 'Evaluation of transfer evidence for three-level multivariate data with the use of graphical models', *Computational Statistics and Data Analysis*, in press.

Alderman T (2005) *Forensic Speaker Identification: A Likelihood Ratio-based Approach Using Vowel Formants,* LINCOM Studies in Phonetics 01, Munich: Lincom Europa.

Daubert (1993) Daubert vs Merrell Dow Pharmaceuticals, Inc. 113 S Ct 2786.

Doheny (1996) R v Doheny. Court of Appeal Criminal Division. No. 95/5297/Y2.

Elliott J (2001) 'Auditory and F-pattern variations in Australian *okay*: a forensic-phonetic investigation', *Acoustics Australia*; 29/1: 37-41.

Evett IW, Scrange J & Pinchin R (1993) 'An Illustration of the Advantages of Efficient Statistical Methods for RFLP Analysis in Forensic Science', *American Journal of Human Genetics* 52: 498-505.

Gonzalez-Rodriguez J, Drygajlo A, Ramos-Castro D, Garcia-Gomar M, & Ortega-Garcia J (2006) 'Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition', *Computer Speech and Language* Special Issue, 20(2-3):331-355.

Harrington J, Cox F & Evans Z (1977) 'An Acoustic Phonetic Study of Broad, General, and Cultivated Australian English Vowels', *Journal of Australian Linguistics*: 155-184.

Horvath B (1985) *Variation in Australian English*. Cambridge: Cambridge University Press.

Kinoshita Y (2002) 'Use of likelihood ratio and Bayesian approach in forensic speaker identification'. In Bow C, ed. *Proc. 9th Australian Intl. Conf. Speech Science and Technology*, Melbourne: Australian Speech Science & Technology Association: 297-302.

Künzel H & Gonzalez-Rodriguez J (2003) 'Combining Automatic and Phonetic-Acoustic Speaker Recognition Techniques for Forensic Applications', *Proc. 15th Int. Congr. Phonetic Sciences*, Barcelona: 1619 – 1622.

Leeuwen DA & Bouten JS (2004) 'Results of the 2003 NFI-TNO Forensic Speaker Recognition Evaluation'. In Ortega-García J, González-Rodríguez J, Bimbot F, Bonastre J-F, Campbell J, Magrin-Chagnolleau I, Mason J, Peres R, Reynolds D, eds. *Proc. Odyssey-04, The Speaker and Language Recognition Workshop*: 81-82.

Lucy D (2005) *Introduction to Statistics for Forensic Scientists*, Chichester: Wiley.

Meuwly D & Drygajlo A (2001) 'Forensic Speaker Recognition Based on a Bayesian Framework and Gaussian Mixture Modelling (GMM)', *Proc. of the 2001 Speaker Odyssey - Speaker Recognition Workshop*: 145-50.

Mitchell AG, & Delbridge A (1965) *The speech of Australian adolescents*. Sydney: Angus & Robertson.

Reynolds D, Andrews W, Campbell J, Navratil J, Peskin B, Adami A, Jin Q, Klusacek D, Abramson J, Mihaescu R, Godfrey J, Jones D & Xiang B (2003) 'The SuperSID Project: exploiting high-level information for high accuracy speaker recognition', Paper at Intl. Conf. on Acoustics Speech & Signal Processing.

Robertson B & Vignaux GA (1995) *Interpreting Evidence*, Chichester: Wiley.

Rose P (1999) 'Long- and Short-term within-speaker differences in the formants of Australian *hello*', *Journal of the International Phonetics Association* 29/1: 1-31.

Rose P (2002) *Forensic Speaker Identification,* London & New York: Taylor and Francis.

Rose P (2003) *The Technical Comparison of Forensic Voice Samples*, Issue 99, Expert Evidence, Freckelton & Selby (series eds.), Sydney: Thomson Lawbook Company.

Rose P (2006a) 'Technical Forensic Speaker Recognition: Evaluation, Types and Testing of Evidence', *Computer Speech and Language* Special Issue, 20(2-3):159-191.

Rose P (2006b) 'Forensic Speaker Recognition at the Beginning of the Twenty-first Century – An Overview and a Demonstration', *Australian Journal of Forensic Sciences* 37/2: 4-30.

Rose P (2006c) 'Accounting for Correlation in Linguistic-Acoustic Likelihood Ratio-based Forensic Speaker Discrimination'. Paper accepted for *Proc. Odyssey-06, The Speaker and Language Recognition Workshop*.

Rose P, Lucy D & Osanai T (2004) 'Linguistic-acoustic Forensic Speaker Identification with Likelihood Ratios from a Multivariate Hierarchical Random Effects Model: A 'Non-Idiot's Bayes' Approach'. In: Cassidy S. ed. *Proc. 10th Australian Intl. Conf. on Speech Science and Technology*, Sydney: Australian Speech Science & Technology Association:492-497.

Rose P, Osanai T & Kinoshita Y (2003), 'Strength of Forensic Speaker Identification Evidence - Multispeaker formant and cepstrum based segmental discrimination with a Bayesian Likelihood ratio as threshold', *Speech Language and the Law*, 10(2):179-202.