LINKAGE PROJECT APPLICATION

PROJECT ID: LP100200142

A2 Proposal Title
Making demonstrably reliable forensic voice comparison a practical everyday reality in Australia

A4 Summary of Proposal (750 characters inc spaces)
To make forensic voice comparison a practical everyday reality in Australia, for use in police investigations and for presentation of evidence in court, forensic scientists must have a practical and demonstrably reliable forensic-voice-comparison system including a large representative database of Australian English voices. In collaboration with Australian police forensic laboratories and other partners we will develop and test such a system, improving on existing systems by combining the strengths of acoustic-phonetic and automatic approaches. The system will comply with the strictest international standards for the admissibility of scientific evidence in court, using the same evaluative framework as is applied to DNA.

[729]

A5 Summary of National/Community Benefit (For Public Release) (750 characters ins spaces)
The police obtain audio recordings of a criminal and a suspect, but what next? To assist Australian law-enforcement agencies and courts in the process of the conviction of the guilty and the exoneration of the innocent, we will develop and test a practical and demonstrably reliable forensic-voice-comparison system for use with Australian English voices. This will allow forensic scientists to produce reliable strength-of-evidence statements for presentation in court using the same evaluative framework as used with DNA. In addition, application of the system during criminal investigations may lead to the refocussing of investigations on other suspects, or may help leverage guilty pleas, thus saving substantial expenditure of time and money.

[748]

PROJECT TEAM (not part of application in this format)

Researchers:

– Geoffrey Stewart Morrison	Research Associate, School of Language Studies, Australian National Univeristy (ANU); Visiting Researcher, School of Electrical Engineering and Telecommunications (EE&T), University of New South Wales (UNSW). Will move to UNSW full time to lead project.
– Julien Epps	Senior Lecturer, EE&T, UNSW.
– Eliathamby Ambikairajah	Head of School, EE&T, UNSW.
– Gary Edmond	Director, Program on Expertise, Evidence and Law, Centre for Interdisciplinary Studies of Law, UNSW.
– Joaquín González-Rodríguez	Co-Director, Biometric Recognition Group (ATVS), Autonomous University of Madrid (UAM).
– Daniel Ramos	Assistant Professor, ATVS, UAM.
– Cuiling Zhang	Associate Professor, China Criminal Police University (CCPU).

Partners:

– Australian Federal Police (AFP), Forensic and Data Centres

Dr K Paul Kirkbride, Chief Scientist

Mr Graeme Kinraid, Team Leader, Forensic Imaging

– Victoria Police (VicPol), Forensic Services Department, Digital & Document Evidence Branch

Mr Dean Catoggio, Branch Manager

– Western Australian Police (WAPol), Forensic Division

Superintendent Hadyn Green, Officer in Charge

– National Institute of Forensic Science (NIFS)

Mr Alastair Ross, Director

– Universidad Autónoma de Madrid (UAM), Escuela Politécnica Superior

– Australasian Speech Science and Technology Association (AASTA)

Prof Denis Burnham, President

– Guardia Civil, Departamento de Ingeniería - Área de Acústica

Lieutenant Coronel José Juan Lucena-Molina

Consultants:

– Dr Philip Rose	Internationally recognised expert in acoustic-phonetic forensic voice comparison.
– Prof Claude Roux	Director, Centre for Forensic Science, University of Technology Sydney; Chair, Australian and New Zealand Association of Forensic Science Educators.

AIMS AND BACKGROUND

Goal (distal aim)

- The goal of the project is to make forensic voice comparison of high demonstrable reliability a practical everyday reality in Australia.

Falsifiable criteria for success (proximal aims)

- Build a forensic voice comparison system which outperforms current systems on reliability and human-labour expense (for typical cases we aim to be able to provide casework reports in a time frame of approximately 6 weeks at a price of approximately $10k).
- Be judged to have submitted the most reliable system at the 2012 US National Institute of Standards and Technology (NIST) Human-Assisted Speaker Recognition Evaluation (HASR).
- Publish key research findings in at least four major journal articles.
- Train 2 PhD students to be highly competent forensic scientists, capable of conducting research and performing casework.
- Develop a business plan (working with UNSW Global's Expert Opinion Services) and begin performing casework.

What is forensic voice comparison?

Forensic voice comparison (FVC) is the comparison of one or more audio recordings of the voice of a known speaker with an audio recording of the voice of a speaker of questioned identity for the purpose of presenting expert testimony in court or during pre-trial investigation. Typically the known voice is that of a suspect/defendant and the questioned voice is that of an offender. Here is a representative scenario: In a major fraud case involving hundreds of millions of dollars an audio recording of a telephone call made by the offender to the bank is available. An audio recording of a telephone call made by a suspect, a former bank employee, is also available (the defence do not contest the identity of the speaker on this recording). A forensic scientist conducts a forensic comparison of the two voice recordings. In court, the forensic scientist testifies that one would be 2000 times more likely to observe the acoustic differences between the voice recordings under the prosecution's proposition that the recordings are of the same speaker than under the defence's proposition that they are of different speakers. This, along with other evidence, leads to a conviction.

Historically, there have been four basic approaches to FVC: *auditory*, *spectrographic (voice-printing)*, *acoustic-phonetic*, and *automatic* (see summaries in Rose, 2002; Jessen 2008; and Morrison, 2009b). Of these, only that latter two are based on objective measurements of the acoustic properties of the voices, and hence only the latter two are good candidates for the development of demonstrably reliable FVC systems. The ***acoustic-phonetic*** approach was developed by phoneticians. Typically, in this approach comparable phonetic units are located in both known and questioned speech samples and then the acoustic properties of these units are measured. An example of a phonetic unit is the vowel /aɪ/ (the vowel sound in the words "I", "hi", "buy", etc.). A phonetic unit could be a phoneme (a basic building block of speech in phonological theory), but could also cover shorter or longer stretches of speech. Examples of acoustic properties are the resonances of the vocal tract (formants) which in phonetic theory are primary acoustic correlates of vowel category (phoneme) identity, i.e., they are the primary cues used by listeners to determine whether a speaker said /aɪ/, /aʊ/, /æ/, etc.. The ***automatic*** approach was developed by engineers as an application of signal processing with little reference to phonetic theory. Typical features in an automatic system are spectra from 20–30 ms windows extracted over the entire speech-active portion of the recording, and quantified using cepstral coefficients. Typically no explicit attempt is made to exploit information relating to phonetic units.

An advantage of an automatic system is that large amounts of information can be rapidly analysed with little human-labour involved. A disadvantage is that linguistic information has traditionally been treated as noise (unwanted variability).The noise problem is overcome by analysing massive amounts of data, but system performance could potentially be improved if the linguistic information were exploited. An advantage of an acoustic-phonetic system is that linguistic information is exploited. A disadvantage is that data analysis is human-labour intensive, and therefore only a relatively small amount of data can be analysed. System performance could potentially be improved if larger amounts of data could be analysed without dramatically increasing the human-labour costs. Another advantage of a typical acoustic-phonetic system is that it exploits acoustic properties of the speech signal which

are more robust to channel effects than those typically used on automatic systems. (A channel effect is, for example, the changes to a signal as the result of passing it through a telephone system; dealing with channel effects, and session variability in general, has been a major research focus in automatic speaker recognition, Kinnunen & Li, 2010.)

A realisation of the relative strengths and weaknesses of each approach has lead to research aimed at combining the two. This includes taking calibration and fusion techniques originally developed for use with automatic systems, and applying them in acoustic-phonetic systems (González-Rodríguez *et al*, 2007; Morrison, 2009a). Although not developed for forensic applications, some research on automatic speaker recognition has explored exploiting linguistic information via techniques such as using automatic speech recognisers to identify phoneme-like units (Shriberg & Stolke, 2008).

How should forensic evidence be evaluated?

Today we are in the midst of what Saks & Koehler (2005) have called a *paradigm shift* in the evaluation and presentation of evidence in the forensic sciences which deal with the comparison of the quantifiable properties of samples of known and questioned origin, e.g., DNA profiles, fingermarks, glass fragments, handwriting, and speech recordings. The new paradigm and the history of its application to FVC are described in detail in Morrison (2009b), a brief description of the new paradigm is provided here. The new paradigm can be characterised as quantitative implementation of the likelihood-ratio framework with quantitative evaluation of the reliability of results.

Seminal legal events in this paradigm shift occurred in the 1990s: The US Supreme Court decision in *Daubert v Merrell Dow Pharmaceuticals* [(92-102), 509 US 579 (1993)] ruled that when considering the admissibility of scientific expert evidence, the judge must, among other things, consider whether the scientific methodology has been empirically tested and found to be reliable. The Appeal Court of England and Wales' decision in *R v Doheny & Adams* [(1996) EWCA Crim 728] ruled that a forensic expert must present the probability of observing the evidence given the hypotheses of same versus different origin, and must not present the probability of the same-origin hypothesis given the evidence.

Also in the 1990s the relatively new field of forensic DNA profile comparison was rapidly developing and assuming a central role in many criminal cases. Arguably the newness of forensic DNA profile comparison and the strong scientific training of those developing it made this branch of forensic science ideal for the quantitative implementation of a framework which would allow it to meet the requirements of the *Daubert* and *Doheny & Adams* rulings. This framework is the *likelihood-ratio framework*, which is now standard for the evaluation of DNA comparison evidence (Foreman *et al*, 2003) and is widely recognised as the logically correct framework for the evaluation of forensic evidence (Aitken & Taroni, 2004; Association of Forensic Science Providers, 2009; Balding, 2005; Buckleton, 2005; Champod & Meuwly, 2000; Evett, 1998, 2009; Jessen, 2008; Lucy, 2005; Rose, 2002, 2006). The applicants and their collaborators have been on the vanguard of the adoption of quantitative implementations of the likelihood-ratio framework for FVC, e.g., González-Rodríguez *et al* (2006), Morrison (2009a), Thiruvaran, Ambikairajah, & Epps (2008).

In the likelihood-ratio framework, the task of the FVC expert is to provide the court with a *strength-of-evidence* statement in answer to the question:

- How much more likely are the differences/similarities between the voice samples to arise under the hypothesis that they were both produced by the same speaker than under the hypothesis that they were produced by different speakers?

The answer to this question is quantitatively expressed as a likelihood ratio, calculated using the following formula:

$$\text{likelihood ratio} = \frac{p(\text{ observed difference between samples} \mid \text{same origin hypothesis})}{p(\text{ observed difference between samples} \mid \text{different origin hypothesis})}$$

Where the numerator of the likelihood ratio can be considered a *similarity* term, and the denominator a *typicality* term. In calculating the strength of evidence, the forensic expert must consider both the degree of similarity between the two voice samples, and also the degree of typicality of the samples with respect to the potential population of offenders. For example, even if two voice samples are very similar on some acoustic measure, this does not lead to a high strength of evidence in favour of the same-speaker hypothesis if these acoustic properties are also very typical and two voice samples

selected at random from any two speakers in the population are likely to be equally similar. In contrast, if two voice samples are very similar in terms of acoustic properties which are very atypical in the population, this would result in a high strength of evidence in favour of the same-speaker hypothesis.

Since typicality must be considered when calculating quantitative likelihood ratios, it is essential to have a database of voice samples from members of the relevant population. Such a database is also necessary to measure the reliability of the FVC system.

Note that FVC requires the evaluation of the probability of the evidence given competing hypotheses, and is not the same as speaker recognition, speaker identification, or speaker verification which require the evaluation of the probability of hypotheses given evidence.

The current situation

Although internationally many researchers are working on automatic speaker recognition and on auditory and acoustic-phonetic approaches to FVC, the number of research groups working on *FVC in the new paradigm* is small. There are three research groups (Lausanne, Madrid, and Canberra) with a track record of around 10 years, and two groups (Sydney and EU) with a 2-year track record.

In February and April of this year the US National Research Council (NRC) Report to Congress (NRC, 2009) and the Law Commission of England and Wales Consultation Paper (Law Commission of England and Wales, 2009) both called for greater emphasis on reliability in forensic science. See Edmond (2008) on the need for greater emphasis on reliability by Australian courts.

- "[S]ome forensic disciplines are supported by little rigorous systematic research to validate the discipline's basic premises and techniques. There is no evident reason why such research cannot be conducted" (NRC, 2009, S-16).
- "The development of scientific research, training, technology, and databases associated with DNA analysis have resulted from substantial and steady federal support for both academic research and programs employing techniques for DNA analysis. Similar support must be given to all credible forensic science disciplines if they are to achieve the degrees of reliability needed to serve the goals of justice." (NRC, 2009, S-9)

In 1997 the Guardia Civil in Spain began funding research to develop a likelihood-ratio automatic FVC system, and in 2004 they began creating a large database of Spanish voices. By 2005 they were satisfied with the reliability of their system and database, and began using them for investigative purposes and for presentation of evidence in court. The number of FVC reports submitted to the courts by the Guardia Civil's forensic audio laboratory has steadily increased from 30 in 2005 to 98 in 2008.

In contrast, in Australia the ability of forensic scientists to present FVC evidence has been hampered by the lack of investment in a practical and reliable FVC system including a database of Australian English voices. Each case must be handled on an ad hoc basis, and the cost and time needed to handle casework is such that it usually exceeds the budget available to either prosecution or defence, and cannot be completed in a legally acceptable time frame. True costs can exceed $50k and time to completion can exceed 6 months. Demonstrably reliably likelihood-ratio FVC is therefore almost never performed in Australia. To date, it has only been presented in two trials in Australia, one in Victoria in 2007, and one in New South Wales in 2008. In both cases the forensic expert was Rose.

Police forensic laboratories in Australia receive a number of requests to perform forensic voice comparison each year (Victoria Police Forensic Services Department reports receiving approximately 50 requests per year), but none of the police laboratories have the capability to provide this service. FVC is performed in Australia by five or six university-based phoneticians (including Rose and Morrison), but the time and resources they are able to devote to casework is limited. Rose and Morrison receive around 15 requests per year from Australia and overseas, but are only able to deal with a small portion of these. There is therefore demand in excess of the availability of FVC in Australia given present capabilities and costs. Greater availability of FVC at affordable prices and within reasonable time frames would also likely increase demand for service.

In Australia, prosecutors often make use of audio recordings from telephone intercepts and rely on the testimony of police officers, laypersons with respect to FVC, who listen to the recordings and assert that they believe the voice on the recording to be that of the defendant. Defence counsel are beginning to realise that they can call on forensic scientists to challenge the reliability of such testimony (Morrison was involved in one such case earlier this year). If this trend continues, it is likely to result

in much higher demand for demonstrably reliable FVC performed by experts.

In some quarters there is considerable resistance to the adoption of the new paradigm for FVC, most notably in the United Kingdom (French & Harrison, 2007; but see counterarguments in Morrison, 2009b, and Rose & Morrison, 2009). The principle objection appears to be the lack of existing voice databases which are representative of the potential population of offenders and sufficiently large to allow for a statistically reliable quantitative implementation of the likelihood-ratio framework.

SIGNIFICANCE AND INNOVATION

Expected scientific outcomes

Our work on combining acoustic-phonetic and automatic approaches will be cutting-edge, and we expect it to lead to a substantial improvement of performance compared to present incremental improvement when only one approach is used.

We expect to make advances in the understanding of the extraction of acoustic information from voice recordings for the purposes of FVC. The use of formant trajectories has so far only been tested on small databases of laboratory quality speech. We will investigate its effectiveness using larger databases of voices under more forensically realistic conditions, including testing its robustness to transmission channel effects. We will also thoroughly test other acoustic information, including nasal spectra, which theoretically may be highly effective, but which have not previously been empirically investigated. We expect to produce at least two major journal article on the basis of this acoustic-phonetic research.

We also expect to make important advances in the understanding of how the size of the database and the number of tokens in known and questioned samples affect the accuracy and precision of forensic likelihood ratios. This will provide answers to what is the most pressing issue in FVC research at the present time: *demonstrating reliability in terms of accuracy and precision*. We expect to produce at least one major journal article on the basis of this forensic-evaluation research.

We also expect to make advances in determining which procedures for the calculation of forensic likelihood ratios are most effective for acoustic-phonetic data and whether additional improvement can be achieved by combining these with automatic data. We expect to produce at least one major journal article on the basis of this speech-processing research.

We expect that our research will clearly demonstrate that data-based likelihood-ratio FVC can be practically implemented, an important step in convincing sceptics in FVC and other branches of forensic science to adopt the new paradigm.

The large database of Australian English voices suitable for FVC research and casework which we create will be unique and will become an important asset for conducting future research in FVC, and also potentially for other research on acoustic phonetics and speech processing.

Expected practical outcomes

By the end of the project, we expect to have developed an FVC system which can be demonstrated to be highly reliable and which meets the most stringent international requirements for the admissibility of forensic evidence. The database of Australian English voices will form an integral part of the system and allow trained forensic speech scientists to perform demonstrably reliable FVC in cases where the known and questioned voices are of Australian English speakers. Such analyses would be conducted within an appropriate time frame and at a price which is affordable for prosecution and defence. This would allow a shift from the current situation where FVC is essentially not performed because it is either of undemonstrated reliability or is too expensive and takes too long, to a situation where it can become a practical everyday reality. The outcome of our project would therefore assist the courts in the process of the conviction of the guilty and the exoneration of the innocent, a great benefit to the legal system and to society at large. Demonstrably reliable FVC performed at the behest of law-enforcement agencies (including our partner organisations) during investigations may also lead to the refocussing of the investigation on other suspects, saving the cost of further investigation of a suspect who may turn out to be innocent, or in other cases may help convince a guilty suspect to confess, saving the costs associated with a trial.

APPROACH AND TRAINING

We will collect a large database of recordings of Australian English voices, and use it to conduct research on improving reliability and reducing the time taken to perform FVC. We will do this by

combining the strengths of the acoustic-phonetic and the automatic approaches. The database will also be used in testing the reliability of the FVC system we develop, and will become an essential component of the system which will ultimately be used to perform casework.

Combining the strengths of acoustic-phonetic and automatic approaches

We will conduct research on combining the strengths of acoustic-phonetic and automatic approaches so as to obtain an optimal maximisation of system reliability and minimisation of human labour. We expect that this will be achieved via a human-supervised semi-automatic system rather than a fully automatic system.

Front-end tasks consist of the selection of phonetic units and the extraction of acoustic information from those units. In the acoustic-phonetic approach the identification and location of phonetic units is currently performed manually. We will explore using automatic phone recognisers to help automate and speed up this task. This will include an assessment of existing techniques and adapting them to our phonetic-unit task. The identity and the boundary locations of the phonetic units found by the automatic phone recognisers will be checked and, when necessary, corrected by a human expert. This will allow us to more rapidly process a number of phonetic units in the database both for research purposes and in anticipation of using these phonetic units in casework. In casework the automatic system could immediately give the phonetician a rough count of available phonetic units so that they can immediately focus on and manually check the phonetic units which appear to have sufficient tokens.

In order to provide useful information for FVC, phonetic units must have acoustic properties which have a relatively small within-speaker variation and a relatively large between-speaker variation. These acoustic properties should, ideally, also be relatively robust to transmission-channel effects. One set of candidates which we will test in our system are the *formant trajectories* of vowels. Simple models including only the initial and final formant values have been found to be highly effective for human listeners' vowel-phoneme identification (Gottfried *et al*, 1993; Nearey & Assmann, 1986). The speaker is therefore free to choose a trajectory between these points which suits their physiology and motor-learning idiosyncrasies. The details of the formant trajectories therefore have the potential to contain substantial information relevant to speaker identity, empirically this information has been found to be highly effective for FVC (Morrison, 2009a), and the second formant is relatively robust to channel effects. We will expand on earlier work in this area by exploring a larger number of phonetic-units (including common words such as "right") and testing on our much larger database of voices recorded under more forensically realistic conditions. We will test several automatic formant tracking algorithms (e.g., Rudoy *et al*, 2007; Vallabha & Tuller, 2004) to determine which is most reliable for our application (preliminary work on a small database shows promising results, de Castro, Ramos, & González-Rodríguez, 2009). We will also test other procedures for the extraction of acoustic information (e.g., *frame-averaged frequency modulation*, FM; Nosratighods *et al*, 2009; Thiruvaran, Ambikairajah, & Epps, 2008), to determine whether they perform better than formant trajectories or provide complementary information which can be combined with formant trajectory information to increase system performance.

Another phonetic-unit and acoustic-property combination which may be effective for FVC is the *spectra of nasal consonants*. Nasal consonants are resonants with a complete closure of the oral cavity, e.g., a bilabial /m/, alveolar /n/, or velar /ŋ/ closure, and where the velum is lowered so that air flows through the nasal passages. This means that nasals are produced using main and branch resonators (the latter consisting of the oral cavity and nasal sinuses) and therefore have complex spectra. The complexity of nasal cavities allows for large variation between speakers, but, apart from when the speaker has a cold, etc., they are static structures which do not vary from occasion to occasion. The acoustic spectra of nasal consonants may therefore have relatively high between-speaker variation and low within-speaker variation, making them efficient features for FVC. We will explore the efficacy of different quantifications of nasal spectra (O'Shaughnessy, 2000, pp. 209–210, 452), and the degree to which they are robust to transmission channel effects. To out knowledge this would be the first investigation of the effectiveness of nasals for quantitative FVC.

Once the system's front end has extracted vectors of numbers which characterise the acoustic properties of the voice recordings, the purpose of the back end is to use these to calculate forensic likelihood ratios. We will explore the effectiveness of different procedures when applied to acoustic-

phonetic data, e.g., the multivariate kernel-density formula (MVKD, Aitken & Lucy, 2004), and the Gaussian mixture model - universal background model (GMM-UBM, Reynolds *et al*, 2000).

Reliability testing

We will manually process a portion of the database (data from 100 speakers) using existing acoustic-phonetic techniques and use this as a benchmark against which to test the new system's components.

We will test the robustness of each feature and system component, and the system as a whole, by comparing results from the original high-quality recordings with the results after passing the recordings through filters which mimic the landline and mobile telephone transmission channels which are typical of audio recordings in casework (see Guillemin & Watson, 2008, for this technique).

Likelihood ratios greater than one favour the same-speaker hypothesis and likelihood ratios less than one favour the different-speaker hypothesis; however, FVC is not a binary decision task, rather it is the task of determining the strength of evidence with respect to the same-speaker versus different-speaker hypotheses, i.e., the extent to which likelihood ratios are greater than or less than one, equivalently the extent to which log likelihood ratios are greater than or less than zero. Different systems can be compared according to which produces larger positive log likelihood ratios from test data known to be same-speaker comparisons and larger negative log likelihood ratios for test data known to be different-speaker comparisons. A metric which captures the gradient goodness of a set of likelihood ratios derived from test data, and which can be used to directly compare different systems on the same test data is the log-likelihood-ratio cost, C_{llr} (Brümmer & du Preez, 2006). This metric has been adopted in the NIST Speaker Recognition Evaluations (SRE). C_{llr} provides a measure of the reliability of the system, and we will use it to compare different permutations of our system as it is developed, and to compare the final system with other existing systems.

A question raised by those who resist the adoption of the new paradigm for FVC is how large a database must one collect if a suitable database is not already available? This is a question about the accuracy and precision of the likelihood ratios as the size of the population sample used to calculate the probability density for the denominator of the likelihood ratio is varied (Aitken, 1991). Ishihara & Kinoshita (2008) and Becker *et al* (2009) have begun to look at the accuracy issue using randomisation tests, but the generalisability of these tests is limited because of the small sizes non-overlapping randomised subgroups which can be extracted from the relatively small databases examined. To date, there has been no published work on evaluating the precision of FVC likelihood-ratio results, but we consider the ability to provide a credible interval for a likelihood ratio presented in court as an essential aspect of demonstrating reliability (NRC, 2009, identifies "the reporting of a measurement with an interval that has a high probability of containing the true value" as part of a scientific approach, p. 4-8). We will perform empirical evaluations, using our large database to determine the parameters for a hierarchical probability density model from which we will generate synthetic samples in Monte Carlo simulations. Synthetic population samples of different sizes will be generated, and changes in the likelihood ratios which the system calculates for a set of test data will be observed (a infinite number of synthetic samples can be generated, Monte Carlo simulation is not subject to the same constraints as randomisation tests). Since the distribution of the model used to generate the synthetic data is known, we will be able to directly assess the accuracy and precision of the results. If asymptotic behaviour occurs once a particular size of population sample has been reached, then this can be taken as an indicator that for practical purposes one may not need to use larger databases. In the same way we will also examine the effect of changing the number tokens extracted from each recording.

Database

Since typicality must be considered when calculating a forensic likelihood ratio, it is essential to have a database of voice samples from members of the relevant population. A large database of voice samples is also necessary to perform research in order to find ways of extracting information from voice samples which will lead to more reliable results. A database of voice samples is also needed to test the reliability of FVC systems.

The database used in casework must be reflective of the potential population of offenders, which must be determined on a case-by-case basis; however, a useful starting point would be a database which contains voice recordings of speakers of the same gender, language, and dialect. In Australia a reasonable starting point is therefore a database of recordings of speakers of Australian English. If the

database is sufficiently large and representative of the range of variation in Australian English voices, then for many cases it should be possible to extract a suitable subset of voices from the larger database. We will therefore collect a database containing multiple audio recordings of the voices of approximately 1000 speakers of Australian English. Voice recordings will cover three speaking styles which are common in FVC casework: 1. casual telephone conversation, 2. information exchange via telephone, and 3. simulated police interview. Each speaker will be recorded on two separate occasions to allow for same-speaker comparisons. Note that existing and proposed databases (such as ANDOSL and Burnham *et al*.'s LIEF) do not provide the appropriate combination of speaking styles, number of speakers, non-contemporaneous recordings of each speaker, and of representing Australian English as spoken today.

Data collection for the proposed project will in part replicate the procedures developed by the Guardia Civil for the collection of their initial database of 750 Spanish voices. These procedures have been adopted and expanded by Rose & Morrison, and Zhang & Morrison for the collection of databases of the voices of approximately 100 Australian English and 60 Chinese speakers respectively. The latter projects will be completed before the beginning of the proposed project and the procedures will therefore have been throughly piloted. In order to elicit natural speech, the participants will be given three tasks to perform. The procedures for the first two tasks are as follows: Two speakers who know each other sit in different rooms and are recorded while holding a telephone conversation. Each speaker wears a high-quality flat-frequency-response lapel microphone connected to a soundcard and computer in a third room, and the signal from each microphone is recorded on a separate recording channel (high quality recordings can subsequently be degraded to test the impact of different channel effects). In the first task speakers have a conversation about whatever they want. In the second task each speaker receives a poorly reproduced fax, some information is illegible for one speaker and other information is illegible for the other. They converse over the phone in order to exchange the missing information. In the third task each speaker is independently interviewed in person (not via the telephone) by a researcher. The researcher asks questions which will simulate a police interview and which the speaker can answer naturally without rôle playing. Each task lasts 5–10 minutes in order to obtain ample data from each speaker. So as to allow for same-speaker comparisons, each speaker is recorded on two occasions separated by approximately two weeks.

We will record approximately 1000 speakers, 250 from each recording site: Canberra (AFP), Melbourne (VicPol), Perth (WAPol), Sydney. Data collection at the first three sites will be facilitated by the partners, data collection in Sydney will be conducted at UNSW. The only requirements for speaker participation will be that they are English-speaking adults who reside in Australia. Via this pseudo-random procedure we aim to obtain a representative sample of Australian English voices, including those of males and females, and speakers from different regions and social backgrounds.

We will not be targeting immigrant or minority groups, and the resulting database may have limited utility for casework involving speakers from such groups, but it should be of high utility if suspects and offenders come from majority groups. We consider this to be an initial database, and in the future, in continued collaboration with our partners, we hope to expand the size and coverage of the database.

In order to begin work on generalising our system to other languages and dialects, the Guardia Civil will provide us with a copy of their Spanish database, and we will also work on applying our system to Zhang's smaller databases of Chinese voices. The US National Institute of Standards and Technology (NIST) is actively developing a Human-Assisted Speaker Recognition Evaluation (HASR). We will participate and work on the General American database which they supply.

Training

The proposed project will offer the world's first PhD research programme in combined acoustic-phonetic and automatic approaches to FVC. We will recruit one student with a strong background in phonetic science and another with a strong background in engineering speech processing. They will be trained to become highly competent and knowledgeable forensic scientists. Both will receive comprehensive training in acoustic-phonetic and automatic FVC, and in the evaluation of forensic evidence. Training in signal processing and automatic speech processing will be provided by Ambikairajah and Epps (they will also provide training in this area to Morrison). Training in acoustic phonetics, forensic science, and applied statistics will be provided by Morrison. Training

in issues related to the admissibility of forensic evidence will be provided by Edmond. In order to give the students a thorough understanding of the nature of voice data, both students will work on the collection of the database at the Sydney site, and will manually process a portion of the database.

Each student will produce a dissertation on a different aspect of the proposed project's research programme. One is expected to focus on the extraction of information from nasal spectra, the automation of this process, and its integration into the FVC system; and the other on automatically locating phonetic units of interest, and on comparing the effectiveness of different procedures for extracting information from vowels including formant trajectories and FM features.

To assist training within the Australian police partner organisation, at the beginning and end of the project Morrison will visit and give seminars on FVC in general and on the purpose and results of the project in particular (at the end of the project this will be done jointly with the PhD students).

In order to provide training and communicate the research results to the broader law-enforcement and judicial communities, NIFS will organise seminars and workshops. Speakers will include academic researchers, forensic scientists from police labs, and legal experts.

NATIONAL BENEFIT

The Australian Government has established *Safeguarding Australia* as a National Research Priority. Our project is clearly part of "protecting Australia from terrorism and crime" with the potential to assist the criminal justice system during late investigative stages and during the trial stage.

If successful this project will make demonstrably reliable FVC a practical everyday reality in Australia. Forensic scientists will be able to efficiently produce accurate and precise strength-of-evidence statements for presentation in court when the suspect and offender are speakers of Australian English, and be able to do this at reasonable cost within a reasonable time frame. This will assist the courts in reaching their verdict and thus be of benefit to the Australian justice system (law-enforcement agencies, prosecutors, defence counsel, innocent defendants and victims in criminal cases, and claimants and respondents in civil cases), and by extension will be of benefit to Australian society in general. In addition, the FVC system and database we develop will be of benefit during pre-trial investigation: Application of forensic voice comparison during criminal investigations may in some instances lead to the refocussing of resources, saving the substantial costs associated with continuing to investigate and prosecute a suspect who turns out to be innocent, and in other instances may help leverage guilty pleas, saving the substantial costs associated with criminal trials.

Researchers in Australia are already world leaders in acoustic-phonetic likelihood-ratio approaches to FVC, but work in Australia on automatic approaches to FVC only began in 2007. The proposed project would allow researchers in Australia to play a more prominent role in automatic FVC and make them world-leaders in hybrid acoustic-phonetic/automatic approaches. Our participation in the NIST HASR would place Australia at the forefront of international evaluation of FVC systems. If successful our bids to run a tutorial and special session on FVC at the *Pan-American/Iberian Meeting on Acoustics*, and out bids to host the *NIST Speaker Recognition Workshop*, the *ISCA Odyssey Speaker and Language Recognition Workshop*, and the *International Association of Forensic Phonetics and Acoustics Conference* would also give Australia a high profile in the scientific research community.

In terms of investment in research and infrastructure, Australia lags more than a decade behind Spain, but collaboration in the proposed project between Australian and Spanish law-enforcement agencies and university researchers would allow us to rapidly close the gap and make Australia one of the top two countries in the world in the implementation of FVC. We believe that our FVC system will exceed the reliability of existing systems while still allowing us to perform casework at a reasonable cost in a reasonable time frame, making us the Australian and potentially the international leaders in providing FVC service.

PARTNER ORGANISATION COMMITMENT AND COLLABORATION

The main group of partner organisations consists of the forensic laboratories of the *Australian Federal Police* (AFP), *Victoria Police* (VicPol), and *Western Australia Police* (WAPol). These partners are providing substantial financial and in-kind support and will contribute primarily by facilitating the compilation of the database, including assisting with the recruitment of speakers and the recording of voice samples at their sites. The availability of a practical reliable FVC system for use with Australian English voices will enable forensic scientists to conduct faster and cheaper FVC and better serve the

police services, who are the largest potential customers for this service (it will also allow us to better serve other clients such as defence counsel). We hope that the proposed project will be only the beginning of collaboration between the Australian police partners and the researchers, and that, as with the Guardia Civil in Spain, they will continue to invest in and derive benefit from future research on FVC. The Australian police partners are currently concerned with assessing reliability of all branches of forensic science in which they work. We are developing tools and expertise for reliability testing in FVC which will also be applicable to other branches of forensic science.

The *National Institute of Forensic Science* (NIFS) is providing financial and in-kind support, and will contribute primarily by organising workshops and seminars to inform and educate the law-enforcement and judicial communities about FVC in general and the proposed project in particular. They will also work with the *Australasian Speech Science and Technology Association* (ASSTA) Forensic Speech Science Committee (FSSC) on developing national standards for FVC.

The *Guardia Civil* will support our research by providing us with their database of Spanish voices, which we will use to test the generalisability of the systems we develop for Australian English. They will also provide advice to the researchers and the Australian police partners regarding the implementation of FVC in casework. The Guardia Civil are interested in supplementing their existing automatic system with acoustic-phonetic analyses, and we will be able to provide them with advice and assistance on this.

In addition to working on standards, ASSTA will contribute financial support and co-sponsor our application to host the international workshops/conferences. The database we compile will be of great value to forensic speech scientists and other acoustic-phonetics and speech-processing researchers who are members of ASSTA. All the Australia-based applicants are members of ASSTA.

COMMUNICATION OF RESULTS

Throughout the project we will maintain a website publicising its goals and progress.

In order to communicate the methodologies and results to the research community, papers will be submitted to top-level international journals in acoustic phonetics, speech processing, and forensic science such as *Journal of the Acoustical Society of America*; *Speech Communication*; *IEEE Transactions on Audio, Speech and Language Processing*; and *Forensic Science International*. Papers will also be presented at top international conferences and Australasian conferences (see list of conferences in section E1 below). In addition, Morrison is organising a special session and a tutorial on FVC at the *Pan-American/Iberian Meeting on Acoustics*, Nov 2010. We will also apply to host the *NIST Speaker Recognition Workshop* (NIST SRE), the *ISCA Odyssey Speaker and Language Recognition Workshop*, and the *International Association of Forensic Phonetics and Acoustics* (IAFPA) *Conference* in 2012, running these three back-to-back so as to bring together automatic speaker recognition researchers and auditory, acoustic-phonetic, and automatic FVC researchers.

Communication of results to the law-enforcement and legal communities will be achieved via seminars at the police partner sites, and seminars and workshops run by NIFS (see *Training* above).

In order to communicate the research results to members of the general public, at appropriate junctures we will publish press releases and offer to give radio, television, and newspaper interviews.

ROLE OF PERSONNEL

•Morrison will coordinate the project as a whole making regular contact with each of the researchers and partner organisations. He will oversee the collection of the database, setting-up the recording equipment and training the personnel at each site. He will perform part of the manual processing of the database. He will work with other team members on automating the location of phonetic units, extraction of information from the dynamic spectral properties, on the reliability-testing research, and on the expansion of the system to other languages and dialects. •Ambikairajah and Epps will develop and maintain the system back-end into which the front-end acoustic information extraction component will be integrated. They will also work on extraction of information using FM features. •González-Rodríguez and Ramos will work on automatic formant tracking and its integration into the system, and will work with Morrison on the expansion of the system to Spanish. •Zhang will work with Morrison on the expansion of the system to Chinese. •Edmond will advise on legal issues related to the admissibility of forensic evidence in court, and ensure that the system developed conforms to Australian and international admissibility standards. •The PhD students will perform part of the manual

processing of the database, and work on portions of the project as described above under *Training*. •Temporary research assistants and the PhD students will recruit, schedule, and record speakers, and add the recordings into the database including all necessary record keeping. •The following eminent forensic scientists have agreed to act as occasional consultants on the project, advising on their respective areas of expertise, acoustic-phonetic FVC, and evaluation of forensic evidence and forensic-science education: ·Dr Philip Rose, internationally recognised expert on acoustic-phonetic FVC; ·Prof Claude Roux, Director, Centre for Forensic Science, University of Technology Sydney, and Chair, Australian and New Zealand Association of Forensic Science Educators.

REFERENCES

•Aitken CGG (1991) Populations and samples. In Aitken CGG, Stoney DA (eds) *The Use of Statistics in Forensic Science*. Chichester, UK: Ellis Horwood. 51–82. •Aitken CGG, Lucy D (2004) Evaluation of Trace Evidence in the Form of Multivariate Data. *Applied Statistics*, 54, 109–122. •Aitken CGG, Taroni F (2004) *Statistics and the Evaluation of Forensic Evidence for Forensic Scientists* (2nd ed). Chichester, UK: Wiley. •Assoc Forensic Science Providers (2009) Standards for the formulation of evaluative forensic science expert opinion. *Science & Justice*, 49, 161–164. •Balding DJ (2005) *Weight-of-evidence for Forensic DNA Profiles*. Chichester, UK: Wiley. •Becker T, Jessen M, Grigoras C (2009) Speaker verification based on formants using Gaussian mixture models. *Proceedings of NAG/DAGA International Conference on Acoustics, Rotterdam*. •Brümmer N, du Preez J (2006) Application Independent Evaluation of Speaker Detection. *Computer Speech and Language*, 20, 230–275. •Buckleton J (2005) A framework for interpreting evidence. In Buckleton J, Triggs CM, Walsh SJ (eds) *Forensic DNA Evidence Interpretation*, 27–63. Boca Raton, FL: CRC. •Champod C, Meuwly D (2000) The inference of identity in forensic speaker recognition. *Speech Communication*, 31, 193–203. •de Castro A, Ramos D, González-Rodríguez J (2009) Forensic speaker recognition using traditional features comparing automatic and human-in-the-loop formant tracking. *Proceedings of Interspeech 2009, Brighton*, 2343–2346. •Edmond G (2008) Specialised knowledge, the exclusionary discretions and reliability. *UNSW Law Journal*, 31, 1–55. •Evett IW (1998) Towards a uniform framework for reporting opinions in forensic science case-work. *Science & Justice*, 38, 198–202. •Evett IW (2009) Evaluation and professionalism. *Science & Justice*, 49, 159–160. •Foreman LA, Champod C, Evett IW, Lambert JA, Pope S (2003) Interpreting DNA evidence: A review. *International Statistics Journal*, 71, 473–473. •French JP, Harrison P (2007) Position statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases. *International Journal of Speech Language and the Law*, 14: 137–144. •González-Rodríguez J, Drygajlo A, Ramos-Castro D, García-Gomar M, Ortega-García J (2006) Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. *Computer Speech and Language*, 20, 331–335. •González-Rodríguez J, Rose P, Ramos D, Torre D, Ortega-García J (2007) Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15, 2104–2115. •Gottfried M, Miller JD, Meyer DJ (1993) Three approaches to the classification of American English diphthongs. *Journal of Phonetics*, 21, 205–229. •Guillemin BJ, Watson C (2008) Impact of the GSM mobile phone network on the speech signal: Some preliminary findings. *International Journal of Speech, Language and the Law*, 15, 193–218. •Ishihara S, Kinoshita Y (2008) How many do we need? exploration of the population size effect on the performance of forensic speaker classification. *Proceedings of Interspeech 2008 Brisbane*, 1941–1944. •Jessen M (2008) Forensic phonetics. *Language and Linguistics Compass*, 2, 671–711. •Kinnunen T, Li H (2010) An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*, 52, 12–40. Available Oct 2009 at www.sciencedirect.com •Law Commission of England and Wales (2009) *The admissibility of expert evidence in criminal proceedings in England and Wales: A new approach to the determination of evidentiary reliability*. Consultation Paper No 190. •Lucy D (2005) *Introduction to Statistics for Forensic Scientists*. Chichester, UK: John Wiley. •Morrison GS (2009a) Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs. *Journal of the Acoustical Society of America*, 125, 2387– 2397. •Morrison GS (2009b) Forensic voice comparison and the paradigm shift. *Science & Justice*, 49(4). Available Oct 2009 at www.sciencedirect.com •National Research Council (2009) *Strengthening forensic science in the United States: A path forward*. Washington, DC: National Academies Press. •Nearey TM, Assmann PF (1986) Modeling the role of vowel inherent spectral change in vowel identification. *Journal of the Acoustical Society of America*, 80, 1297–1308. •Nosratighods M, Thiruvaran T, Epps J, Ambikairajah E, Ma B, Li H (2009) Evaluation of a fused FM and cepstral-based speaker recognition system on the NIST 2008 SRE. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei*, 4233–4236. •O'Shaughnessy D (2000) *Speech Communications: Human and Machine*. New York: IEEE Press. •Reynolds DA, Quatieri TF, Dunn RB (2000) Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10, 19–41. •Rose P (2002) *Forensic Speaker Identification*. London & New York: Taylor and Francis. •Rose P (2006) Technical forensic speaker recognition: Evaluation, types and testing of evidence. *Computer Speech and Language*, 20, 159–191. •Rose P, Morrison GS (2009). A response to the UK position statement on forensic speaker comparison. *International Journal of Speech, Language and the Law, 16*, 139–163. •Rudoy D, Spendley DN, Wolfe PJ (2007) Conditionally linear Gaussian models for estimating vocal tract resonances. *Proceedings of Interspeech 2007, Antwerp.* 526–529. •Saks MJ, Koehler JJ (2005) The coming paradigm shift in forensic identification science. *Science*, 309, 892–895. •Shriberg E, Stolke A (2008) The case for automatic higher-level features in forensic speaker recognition. *Proceedings of Interspeech 2008, Brisbane*. 1509–1512. •Thiruvaran T, Ambikairajah E, Epps J (2008) FM features for automatic forensic speaker recognition. *Proceedings of Interspeech 2008, Brisbane*. 1497–1500. •Vallabha GK, Tuller B (2004) Choice of filter order in LPC analysis of vowels. *Proceedings of the Sound to Sense: 50+ Years of Discoveries in Speech Communication*, C-203–C-208.