2015-12-04a

President's Council of Advisors on Science and Technology

https://www.whitehouse.gov/administration/eop/ostp/pcast

https://www.whitehouse.gov/webform/pcast-forensic-science-solicitation-questions-0

PCAST Forensic Science Questions

PCAST consists of 20 of the nation's leading scientists and engineers, appointed by the President to provide direct advice to him and the White House on important matters of science and technology. PCAST has recently begun to explore how best to ensure the quality of forensic science, based on reliable scientific principles and methods, within the criminal justice system. PCAST members are interested in hearing from the broad stakeholder community on each of the questions listed below in an effort to better understand the landscape of this topic.

Please note that any comments are subject to public release under the Freedom of Information Act, and may be archived consistent with the Federal Records Act and Presidential Records Act, as applicable.

This form will accept submissions until Wednesday, December 23, 2015.

Name:

Dr Geoffrey Stewart Morrison

Title and Affiliation:

Independent Forensic Consultant;

Adjunct Associate Professor, Department of Linguistics, University of Alberta

E-mail:

geoff-morrison@forensic-evaluation.net

Existing Forensic Techniques and Practices

Question 1:

What studies have been published in the past 5 years that support the foundational aspects of each of the pattern-based forensic science methods, including (but not limited to) latent print analysis; firearms/toolmarks; shoe/tire prints; bitemark analysis; questioned documents? What studies are needed to demonstrate the reliability and validity of these methods?

With limited resources to commit to the current response to the questions asked, I restrict my comments primarily to material published by my colleagues and myself, primarily in the area of forensic speech science; however, much of what I have to say is I believe also applicable across other branches of forensic science.

Forensic Science is undergoing a paradigm shift (Saks & Koehler, 2005; Morrison, 2009). Over several years my colleagues and I have developed a formulation of a paradigm for the evaluation of forensic evidence which includes the following key elements:

- Use of the likelihood ratio framework for the evaluation and presentation of the strength of forensic evidence.

    This is the logically correct framework for the evaluation of forensic evidence (e.g., Robertson & Vignaux, 1995; Balding, 2005; Buckleton, 2005; Association of Forensic Science Providers, 2009; Morrison, 2010; Evett et al, 2011; Berger et al, 2011; Redmayne et al, 2011; Robertson et al, 2011; Morrison, 2012), is was adopted as standard for DNA in the mid 1990's (Foreman et al, 2003), is gradually being adopted in other branches of forensic science (including forensic voice comparison: Morrison, 2009; Morrison & Enzinger, 2013), and is recommended in the recently-published European Network of Forensic Science Institutes' Guideline for Evaluative Reporting in Forensic Science (Willis et al, 2015).

    An aspect of the likelihood ratio framework that is often not well understood is that a likelihood ratio is the answer to a specific question specified by both the prosecution and the defense hypotheses. In a forensic voice comparison case, the prosecution hypothesis is usually that the speaker of questioned identity is the defendant. The defense hypothesis is usually that the speaker of questioned identity is not the defendant. The defense hypothesis is not, however, that the speaker of questioned identity is any other speaker on the planet, but rather that it is a person selected at random from a relevant population – this population will be restricted by information which can be gleaned from the recording of the speaker of questioned identity and will include properties such as the gender of speaker and the language spoken (see Rose, 2002; Morrison, Ochoa, Thiruvaran, 2012). Case circumstances in other branches of forensic science may also restrict the relevant population. The forensic scientist must make transparent what specific question they have set out

to answer so that the judge at an admissibility hearing or the trier of fact at trial can decide whether the forensic scientist has set out to answer an appropriate question, and also so that they can understand the answer that the forensic scientist provides to that question.

- Use of relevant data, quantitative measurements, and statistical models to calculate likelihood ratios.

This approach is transparent and replicable – the forensic scientist can describe what they did in sufficient detail that another suitably qualified forensic scientist can repeat what they did. In contrast, if the strength of evidence statement made by the forensic scientist is based primarily or directly on a subjective judgment, then this is not transparent or replicable.

An approach based on relevant data, quantitative measurements, and statistical models does involve subjective elements. For example: What is the relevant population for this case? Is the sample of the population sufficiently representative of the relevant population and does it adequately reflect the conditions of the case under investigation? (see discussion under "testing" below) These are pre-empirical questions which should be debated before the judge at an admissibility hearing or the trier of fact at trial. If the judge or trier of fact is satisfied, then the remainder of the analysis is objective. In the first instance it is up to the forensic scientist to satisfy themself that the data they use for training and testing their statistical models are sufficiently representative of the relevant population and casework conditions, but ultimately it is the judge and trier of fact who must be satisfied.

This approach is intrinsically much more robust to the potential effects of cognitive bias than an approach in which the strength of evidence statement made by the forensic scientist is based primarily or directly on subjective judgment. The subjective elements in an approach based on relevant data, quantitative measurements, and statistical models are far removed from the final decision as to the strength of the evidence. (See Found, 2015, for a recent introduction to cognitive bias in forensic science.)

Approaches based on relevant data, quantitative measurements, and statistical models are also practically easier to test than approaches based on subjective judgment. The former can quickly and cheaply provide responses to hundreds or thousands of test trials, whereas each subjective judgment is time consuming for a human expert.

Further discussion of all of the above is provided in Morrison & Stoel (2014).

- Empirical testing of the validity and reliability of the forensic analysis system under conditions reflecting those of the case under investigation.

Such testing is the only way to demonstrate that the forensic analysis system is actually fit for purpose (National Research Council, 2009; Forensic Science Regulator, 2014a, 2014b). Such testing treats the system as a black box and is not prejudiced against any approach whether it be based on relevant data, quantitative measurements, and statistical models, or based directly on subjective judgment. If the judge or trier of fact is satisfied that the test data are sufficiently representative of the relevant population and casework conditions, and that the demonstrated level of performance of the system using these test data is adequate, then there need be no debate regarding the internal workings of the system.

In Morrison (2011) I described suitable metrics for testing validity and reliability within the likelihood ratio framework. In Morrison (2014) I reviewed calls going back to the 1960s for the validity and reliability of forensic voice comparison to be empirically tested under conditions reflecting those of the case under investigation. The conditions of audio recordings in forensic voice comparison are highly variable from case to case:

-The relevant population.

-The speaking style on the questioned voice recording and on the known voice recording, e.g., casual conversation, formal speech, responses to police interview questions, whispering, shouting.

-Presence, volume, and type of background noise, e.g., music, traffic noise, ventilation system noise, babble.

-Presence and details of reverberation, e.g., if the recording is made in a small room with hard walls.

-Transmission of the speech signal through different telecommunications channels which affect the properties of the signals in different (generally deleterious) ways, e.g., landline telephone, mobile telephone, Voice over Internet Protocol.

-Saving audio recordings in formats which distort and lose information, e.g. MP3.

The range of possibilities is essentially infinite such that a separate test of validity and reliability is required for each case. The performance of a forensic voice comparison system under one set of conditions will not necessarily be informative as to the performance of that system under another set of conditions. For example, a system which works well under studio-recording conditions, may perform quite poorly if the audio has been transmitted through a telephone system, especially poorly if it has been transmitted through a mobile telephone system (Zhang et al, 2013). Taking the results of testing a system under one set of conditions and presenting them as informative as to the performance of the system under another set of conditions may be quite misleading. It is not appropriate to run a single validation test and then assume that this is applicable across a range of

different casework conditions. Empirical tests of validity and reliability must be run in a case by case basis. To a greater or lesser extent this may also be true in other branches of forensic science.

To date, my colleagues and I have published two research papers describing the implementation of this paradigm under conditions based on those of real cases (Enzinger et al, 2015; Enzinger & Morrison, 2015). These papers include empirical tests of the validity and reliability of different forensic voice comparison systems under conditions reflecting those of two different cases. Additional papers testing the performance of other systems and other sets of casework conditions are submitted and in preparation.

References:

Association of Forensic Science Providers (2009). Standards for the formulation of evaluative forensic science expert opinion. *Science & Justice*, 49, 161–164. http://dx.doi.org/10.1016/j.scijus.2009.07.004

Balding, D.J. (2005). *Weight-of-Evidence for Forensic DNA Profiles*. Chichester UK: Wiley.

Berger, C.E.H., Buckleton, J., Champod, C., Evett, I.W., Jackson, G. (2011). Evidence evaluation: A response to the Court of Appeal judgment in R v T. *Science & Justice*, 51, 43–49. http://dx.doi.org/10.1016/j.scijus.2011.03.005

Buckleton, J. (2005). A framework for interpreting evidence. In: J. Buckleton, C.M. Triggs, S.J. Walsh (Eds.), *Forensic DNA Evidence Interpretation* (pp. 27–63). Boca Raton, FL: CRC.

Enzinger, E., Morrison, G.S. (2015). Mismatched distances from speakers to telephone in a forensic-voice-comparison case. *Speech Communication*, 70, 28–41. http://dx.doi.org/10.1016/j.specom.2015.03.001

Enzinger, E., Morrison, G.S., Ochoa, F. (2015). A demonstration of the application of the new paradigm for the evaluation of forensic evidence under conditions reflecting those of a real forensic-voice-comparison case. *Science & Justice*. http://dx.doi.org/10.1016/j.scijus.2015.06.005

Evett, I.W., Aitken, C.G.G., Berger, C.E.H., Buckleton, J.S., Champod, C., Curran, J.M., Dawid, A.P., Gill, P., González-Rodríguez, J., Jackson, G., Kloosterman, A., Lovelock, T., Lucy, D., Margot, P., McKenna, L., Meuwly, D., Neumann, C., Nic Daeid, N., Nordgaard, A., Puch-Solis, R., Rasmusson, B., Radmayne, M., Roberts, P., Robertson, B., Roux, C., Sjerps, M.J., Taroni, F., Tjin-A-Tsoi, T., Vignaux, G.A., Willis, S.M., Zadora, G. (2011). Expressing evaluative opinions: A position statement. *Science & Justice*, 51, 1–2. http://dx.doi.org/10.1016/j.scijus.2011.01.002

Foreman, L.A., Champod, C., Evett, I.W., Lambert, J.A., Pope, S., 2003. Interpreting DNA evidence: A review. *International Statistical Review*, 71, 473–495. http://dx.doi.org/10.1111/j.1751-5823.2003.tb00207.x

Forensic Science Regulator (2014a). Codes of practice and conduct for forensic science 590 providers and practitioners in the criminal justice system (Version 2.0). Birmingham, UK: Forensic Science Regulator. https://www.gov.uk/government/publications/forensic-science-providers-codes-of-practice-and-conduct-2014

Forensic Science Regulator (2014b). Draft guidance: Digital forensics method validation. Birmingham, UK: Forensic Science Regulator. https://www.gov.uk/government/consultations/digital-forensics-method-validation-draft-guidance

Found, B. (2015). Deciphering the human condition: The rise of cognitive forensics. *Australian Journal of Forensic Sciences*, 47, 386–401. http://dx.doi.org/10.1080/00450618.2014.965204

Morrison, G.S. (2009). Forensic voice comparison and the paradigm shift. *Science & Justice*, 49, 298–308. http://dx.doi.org/10.1016/j.scijus.2009.09.002

Morrison, G.S. (2010). Forensic voice comparison. In I. Freckelton, & H. Selby (Eds.), *Expert Evidence* (Ch. 99). Sydney, Australia: Thomson Reuters. Stable URL: http://expert-evidence.forensic-voice-comparison.net/

Morrison, G.S. (2011). Measuring the validity and reliability of forensic likelihood-ratio systems. *Science & Justice*, 51, 91–98. http://dx.doi.org/10.1016/j.scijus.2011.03.002

Morrison, G.S. (2012). The likelihood-ratio framework and forensic evidence in court: A response to R v T. *International Journal of Evidence and Proof*, 16, 1–29. http://dx.doi.org/10.1350/ijep.2012.16.1.390

Morrison, G.S. (2014). Distinguishing between forensic science and forensic pseudoscience: Testing of validity and reliability, and approaches to forensic voice comparison. *Science & Justice*, 54, 245–256. http://dx.doi.org/10.1016/j.scijus.2013.07.004

Morrison, G.S., Enzinger, E. (2013). Forensic speech science – Review: 2010–2013. In: NicDaéid, N. (Ed.), *Proceedings of the 17th International Forensic Science Managers' Symposium* (pp. 616–623, 629–635). Lyon, France: INTERPOL.

Morrison, G.S., Ochoa, F., Thiruvaran, T. (2012). Database selection for forensic voice comparison. In *Proceedings of Odyssey 2012: The Language and Speaker Recognition Workshop, Singapore* (pp. 62–77). International Speech Communication Association.

Morrison, G.S., Stoel, R.D. (2014). Forensic strength of evidence statements should preferably be likelihood ratios calculated using relevant data, quantitative measurements, and statistical models – a response to Lennard (2013) Fingerprint identification: How far have we come? *Australian Journal of Forensic Sciences*, 46, 282–292. http://dx.doi.org/10.1080/00450618.2013.833648

National Research Council (NRC), 2009. *Strengthening Forensic Science in the United States: A Path Forward*. Washington, DC: National Academies Press. http://www.nap.edu/catalog.php?record_id=12589

Redmayne, M., Roberts, P., Aitken, C.G.G., Jackson, G. (2011). Forensic science evidence in question. *Criminal Law Review*, 5, 347–356.

Robertson, B., Vignaux, G.A. (1995). *Interpreting Evidence*. Chichester UK: Wiley.

Robertson, B., Vignaux, G.A., Berger, C.E.H. (2011). Extending the confusion about Bayes. *Modern Law Review*, 74, 444–455.

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

Saks, M.J., Koehler, J.J. (2005). The coming paradigm shift in forensic identification science, *Science*, 309, 892–895.

Willis, S.M., McKenna, L., McDermott, S., O'Donell, G., Barrett, A., Rasmusson, B., Nordgaard, A., Berger, C.E.H., Sjerps, M.J., Lucena-Molina, J., Zadora, G., Aitken, C.C.G., Lunt, L., Champod, C., Biedermann, A., Hicks, T.N., Taroni, F. (2015). *ENFSI guideline for evaluative reporting in forensic science*. European Network of Forensic Science Institutes.

Zhang, C., Morrison, G.S., Enzinger, E., Ochoa, F. (2013). Effects of telephone transmission on the performance of formant-trajectory-based forensic voice comparison – female voices. *Speech Communication*, 55, 796–813. http://dx.doi.org/10.1016/j.specom.2013.01.011

Question 2:

Have studies been conducted to establish baseline frequencies of characteristics or features used in these pattern-based matching techniques? If not, how might such studies be conducted? What publicly accessible databases exist that could support such studies? What closed databases exist? Where such databases exist, how are they controlled and curated? If studies have not been conducted, what conclusions can and cannot be stated about the relationship between the crime scene evidence and a known suspect or tool (e.g., firearm)?

My interpretation of this question is that it is asking about the availability of data for training statistical models which calculate the denominator of the likelihood ratio – see in the answer to Question 1 the

discussion of the likelihood ratio framework, and the use of relevant data, quantitative measurements and statistical models. See also the papers referenced in my response to Question 1, which include studies which make such calculations.

At present, lack of suitable data is the biggest practical impediment to performing forensic voice comparison casework. Technical recording conditions (noise, communication system transmission, etc.) can potentially be simulated if one has a database of high-quality audio recordings, but the relevant population and the speaking styles cannot be simulated. Morrison, Rose, Zhang (2012) described a protocol for data collection which elicits natural speech in three different speaking styles which are common in forensic casework – telephone conversation, information exchange over the telephone, and simulated police interview. This protocol was used to collect a database of multi-session (non-contemporaneous) audio recordings of 500+ Australian English speakers (Morrison et al, 2015). The database is available to researchers and forensic practitioners upon request. It allows casework to be conducted when the relevant population is either male or female Australian English speakers and when the speaking styles in the recordings of the known and questioned speakers are similar to those included in the database, but not if the population or speaking styles differ. Few, or no, other existing databases fulfill the criteria necessary to conduct general casework of being reasonably large, and of having multiple high-quality non-contemporaneous recordings of each speaker in speaking styles common in forensic casework. To the best of my knowledge no database exists that would be generally suitable for performing casework involving US English. Although it is theoretically practical to collect databases of some populations and speaking styles in anticipation of performing casework and such databases may allow a relatively large proportion of cases to be conducted, there will always be populations and conditions which cannot be anticipated or for which it is not an efficient use of resources to collect databases on the off chance that such cases will occur. If the case is important enough and the voice evidence important enough in the case, then resources may be available to collect relevant data on an as-needed basis. In practice there are few cases in which such case-specific data collection is performed. If relevant data are not available and not collected, then a proper forensic voice comparison cannot be conducted. Even if one had a method (e.g., a subjective judgment approach) which did not require training data, data for empirical testing would still be required and those data would have to be sufficiently representative of the relevant population and sufficiently reflective of the conditions of the known and questioned speaker recordings in the case.

References:

Morrison, G.S., Rose, P., Zhang, C. (2012). Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice. *Australian Journal of Forensic Sciences*, 44, 155–167. http://dx.doi.org/10.1080/00450618.2011.630412

Morrison, G.S., Zhang, C., Enzinger, E., Ochoa, F., Bleach, D., Johnson, M., Folkes, B.K., De Souza, S., Cummins, N., Chow, D. (2015). *Forensic database of voice recordings of 500+ Australian English speakers*. Available: http://databases.forensic-voice-comparison.net/

Question 3:

How is performance testing (testing designed to determine the frequency with which individual examiners obtain correct answers) currently used in forensic laboratories? Are performance tests conducted in a blind manner? How could well-designed performance testing be used more systematically for the above pattern-based techniques to establish baseline error rates for individual examiners? What are the opportunities and challenges for developing and employing blind performance testing? What studies have been published in this area?

A human operator who performs part of the analysis is part of the system, and the whole system must be tested including that particular operator. As explained in my response to Question 1, testing should be performed on a case by case basis. At present there is a great dearth of such testing on forensic voice comparison practice. See the papers on testing referenced in my response to Question 1, particularly: Morrison (2011), Morrison (2014), Enzinger et al (2015), Enzinger & Morrison (2015).

New Technology

Question 4:

What are the most promising new scientific techniques that are currently under development or could be developed in the next decade that would be most useful for forensic applications? Examples could include hair analysis by mass spectrometry, advances in digital forensics, and phenotypic DNA profiling.

The problems in forensic voice comparison are not at this stage primarily technological, they are primarily the lack of relevant databases, the lack of understanding of the logically correct framework for the evaluation of evidence, and the lack of empirical testing of validity and reliability under casework conditions. Lack of understanding of the logically correct framework and lack of understanding of what constitutes appropriate testing affects all branches of forensic science, and is widespread among forensic scientists, lawyers, and judges. Current priorities need to be database collection and training, not new technology.

Question 5:

What standards of validity and reliability should new forensic methods be required to meet before they are introduced in court?

All forensic methods, not just new ones, must be empirically tested under conditions reflecting those of the case under investigation using data sampled from the relevant population. In branches of forensic science such as forensic voice comparison where the relevant population and conditions are highly variable from case to case, testing of validity and reliability must be conducted on a case by case basis – there should be a *Daubert* / Rule 702 hearing for every case. See my response to Question 1 and the papers on testing referenced there, particularly: Morrison (2011), Morrison (2014), Enzinger et al (2015), Enzinger & Morrison (2015).

Additional Expertise

Question 6:

Are there scientific and technology disciplines other than the traditional forensic science disciplines that could usefully contribute to and/or enhance the scientific, technical and/or societal aspects of forensic science? What mechanisms could be employed to encourage further collaboration between these disciplines and the forensic science community?

At present, the major problem is not a need for contributions from scientific and technology disciplines outside of forensic science, rather it is a need for forensic scientists to adopt a new paradigm based on advancement which have already been made in the field of forensic inference and statistics. Within many branches of forensic science (including forensic voice comparison) there is substantial resistence to change (Curran, 2013) – this is in the nature of a paradigm shift (Kuhn, 1962). Courts and bodies with powers to regulate forensic science must insist on the adoption of a new paradigm which includes logically correct reasoning and empirical demonstration of validity and reliability under casework conditions.

References:

Curran, J.M. (2013). Is forensic science the last bastion of resistance against statistics? *Science & Justice*, 52, 251–252. http://dx.doi.org/10.1016/j.scijus.2013.07.001

Kuhn, T.S. (1962). *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press.

Other

Question 7:

Please share any additional comments.

My former titles and affiliations include:

Scientific Counsel, Office of Legal Affairs, INTERPOL General Secretariat;

Director, Forensic Voice Comparison Laboratory, School of Electrical Engineering & Telecommunications, University of New South Wales;

Chair, Forensic Acoustics Subcommittee, Acoustical Society of America.

Any opinions I express are my own and do not necessarily represent the opinions or policies of any of the organizations with which I am or have been affiliated.

A formatted version of my text is provided as File 1: "PCAST forensic science questions - GSM - 2015-12-04a.pdf".

Files

Please upload any related materials (limit of 3 documents).

File 1: PCAST forensic science questions - GSM - 2015-12-04a.pdf

Max file size: 2 MB

File 2:

Max upload size: 2 MB

File 3:

Max upload size: 2 MB