Science and Justice 54 (2014) 245–256

Contents lists available at ScienceDirect

Science and Justice

journal homepage: www.elsevier.com/locate/scijus

Review

Distinguishing between forensic science and forensic pseudoscience: Testing of validity and reliability, and approaches to forensic voice comparison ☆

Geoffrey Stewart Morrison *

Forensic Voice Comparison Laboratory, School of Electrical Engineering & Telecommunications, University of New South Wales, UNSW Sydney, NSW 2052, Australia

ARTICLE INFO

Article history:
Received 12 January 2013
Received in revised form 28 May 2013
Accepted 17 July 2013

Keywords:
Validity
Reliability
Forensic voice comparison
Aural
Spectrographic
Acoustic–phonetic

ABSTRACT

In this paper it is argued that one should not attempt to directly assess whether a forensic analysis technique is scientifically acceptable. Rather one should first specify what one considers to be appropriate principles governing acceptable practice, then consider any particular approach in light of those principles. This paper focuses on one principle: the validity and reliability of an approach should be empirically tested under conditions reflecting those of the case under investigation using test data drawn from the relevant population. Versions of this principle have been key elements in several reports on forensic science, including forensic voice comparison, published over the last four-and-a-half decades. The aural–spectrographic approach to forensic voice comparison (also known as "voiceprint" or "voicegram" examination) and the currently widely practiced auditory–acoustic–phonetic approach are considered in light of this principle (these two approaches do not appear to be mutually exclusive). Approaches based on data, quantitative measurements, and statistical models are also considered in light of this principle.

Contents

1. Introduction . . . 246
   1.1. The 2009 National Research Council report's versus Cole's concept of forensic "science" . . . 246
   1.2. Paradigm . . . 246
   1.3. The likelihood-ratio framework . . . 246
   1.4. Approaches based on quantitative measurements, databases representative of the relevant population, and statistical models . . . 246
2. Testing of validity and reliability under conditions reflecting those of the case under investigation using data drawn from the relevant population . . . 247
   2.1. Introduction . . . 247
   2.2. Procedures for measuring validity and reliability . . . 247
   2.3. Lack of testing of experience-based systems . . . 248
   2.4. Lack of testing/lack of appropriate testing of systems based on data, quantitative measurements, and statistical models . . . 248
   2.5. Pre-testing or case-by-case testing? . . . 248
3. The spectrographic/aural–spectrographic approach . . . 249
   3.1. Introduction . . . 249
   3.2. Legal admissibility . . . 249
   3.3. Reports including consideration of principles for determining acceptable practice . . . 250
   3.4. Tests of validity . . . 251
   3.5. Conversions and reasons for conversion . . . 252
   3.6. Is the fact that a spectrogram is used a key aspect of the criticism of the approach? . . . 252
   3.7. The IAFPA resolution . . . 253
4. The auditory–acoustic–phonetic approach . . . 254
5. Conclusion . . . 255
Acknowledgments . . . 255
References . . . 256

☆ This is a version of the opening presentation of the Special Session on *Distinguishing Between Science and Pseudoscience in Forensic Acoustics* <http://montreal2013.forensic-acoustics.net/> at ICA 2013: 21st International Congress on Acoustics/165th Meeting of the Acoustical Society of America/52nd Meeting of the Canadian Acoustical Association, Montréal, 2–7 June 2013 <http://www.ica2013montreal.org/>. An abridged written version appears in the conference proceedings under the title "Distinguishing between science and pseudoscience in forensic acoustics". The present written version maintains some of the oral character of the original presentation.

* Now Forensic Consultant, Vancouver, British Columbia, Canada. Tel.: +1 604 637 0896, +44 191 645 0896, +61 2 800 74930.
E-mail address: geoff-morrison@forensic-evaluation.net.

1355-0306/$ – see front matter
http://dx.doi.org/10.1016/j.scijus.2013.07.004

1. Introduction

The title of this paper was deliberately chosen to be provocative, but is probably somewhat (if not highly) inaccurate: I don't plan to actually provide a definition which could be used to include everything one wants to count as science and to exclude everything one does not want to count as science, a problem known in philosophy of science as the *demarcation problem*. See Edmond & Mercer [1] on the problems with the "junk science" versus "good science" debate. I will, however, provide a discussion of what I consider to be relevant principles governing acceptable practice in forensic science in general and forensic voice comparison in particular. I believe that it is more productive to focus on and potentially debate principles and then consider different approaches in light of these principles, rather than immediately attempt to critique the approaches. I believe that a focus on principles will help us to understand what really matters.

There are serious problems with current practice in forensic science, as documented in the 2009 National Research Council (NRC) report on *Strengthening forensic science in the United States: A path forward* [2], in the 2012 Frontline documentary *The real CSI: How reliable is the science behind forensic science?* [3], and elsewhere. Although both the aforementioned report and documentary are from the United States, I would be very surprised if similar problems did not exist in Canada and in other parts of the world.

1.1. The 2009 National Research Council report's versus Cole's concept of forensic "science"

The message of the 2009 NRC report [2] could be summarized as "forensic science should be more scientific", and it explicitly calls for the adoption of a "scientific culture" [2, p. 125]. From a philosophy and sociology of science perspective, Cole [4] is critical of the NRC report's portrayal of science and scientific culture, arguing among other things that it focused on "discovery science" whereas the majority of forensic science practice is what he calls "mundane science". Discovery science can be exemplified by the recently completed process of hypothesizing the existence of the Higgs boson then designing and running an experiment to test this hypothesis, whereas mundane science can be exemplified by "laboratory technicians performing routine assays, industrial scientists seeking to refine a product or process, and even physicians trying to diagnose patients or engineers trying to design a safer bridge" [4, p. 447]. Cole points out, however, that the NRC report never claimed that forensic science was "not science", "unscientific", or "pseudoscience", and that it instead made a number of specific claims and recommendations. One of these recommendations, Recommendation 3 [2, pp. 22–23], will be the focus of my presentation, and can be summarized as:

The validity and reliability of forensic analysis approaches and procedures should be tested.

1.2. Paradigm

For several years I have been advocating a paradigm for the evaluation of forensic evidence consisting of the following three elements:

1. obligatory use of the likelihood-ratio framework
2. highly preferred use of approaches based on quantitative measurements, databases representative of the relevant population, and statistical models
3. obligatory testing of validity and reliability under conditions reflecting those of the case under investigation using data drawn from the relevant population.

Recent summaries of the paradigm appear in Morrison, Evett, et al. [5] and Morrison [6]. Details of my thoughts on selecting an appropriate database for forensic-voice-comparison cases appear in Morrison, Ochoa, & Thiruvaran [7], and details of my thoughts on appropriate metrics and methodology for testing validity and reliability for forensic comparison in general appear in Morrison [8].

Below I briefly discuss the first two elements, then discuss the third element in greater detail.

1.3. The likelihood-ratio framework

I (and many others) consider the likelihood-ratio framework to be the logically correct framework for the evaluation and interpretation of forensic evidence irrespective of the approach adopted (several approaches to forensic voice comparison are discussed below). There is increasing support for this position: In 2011, 31 experts in the field signed a position statement that included an affirmation that they consider the likelihood-ratio framework to be the most appropriate framework for the evaluation of evidence (Evett et al. [9]), and this position statement was endorsed by the Board of the European Network of Forensic Science Institutes (ENFSI), representing 58 laboratories in 33 countries.

In the context of forensic voice comparison, the forensic scientist must assess the likelihood of getting the acoustic properties of the recording of a speaker of questioned identity had it been produced by a speaker of known identity (similarity) versus had it been produced by some other speaker from the relevant population (typicality).[1] The likelihood-ratio framework requires the forensic scientist to consider both similarity and typicality, and to consider what constitutes the relevant population. Much has been written and said about the likelihood-ratio framework, and I will not focus on this element of the paradigm in the current paper.[2]

1.4. Approaches based on quantitative measurements, databases representative of the relevant population, and statistical models

Approaches based on quantitative measurements, databases representative of the relevant population, and statistical models are highly preferred over more human–expert–experience-based approaches because they are more transparent, more easily replicated, and as a practical matter more easily subjected to validity and reliability testing.[3] They are more transparent and more easily replicated because it is possible to describe the data used, measurements made, and statistical models applied in sufficient detail that another suitably qualified and equipped forensic scientist can copy what was done — the first forensic scientist can even provide the second with the data and software which they used. If there are major discrepancies in results, these can potentially be traced back to mistakes in the application of the procedures (e.g., measuring the wrong sample or misrecording a measurement) or genuine disagreements with respect to issues such as what constituted the relevant population. Complete objectivity is unachievable, and it may be reasonable to expect that differences in subjective decisions will typically be the cause

[1] The speaker of questioned identity is usually the offender and the speaker of known identity is usually a suspect. This is not always the case, for example the speaker of questioned identity could be a victim, and the recording of the speaker of known identity a recording of a missing person who it is believed could be that victim. For simplicity, I will use the terms "offender" and "suspect" hereafter, rather than the more widely applicable but periphrastic "speaker of questioned identity" and "speaker of known identity".

[2] Introductions to the likelihood-ratio framework include Robertson & Vignaux [10], Balding [11], and Morrison [12].

[3] Systems with the output based directly on human expert judgments can be fused with systems based on data, quantitative measurements, and statistical models (see Morrison [13]); however, because such a fused system would include a system with the output based directly on human expert judgments it would still be less transparent, harder to replicate, and harder to test than a system based on data, quantitative measurements, and statistical models. Note that the use of systems based on data, quantitative measurements, and statistical models is preferred rather than obligatory within the paradigm, the paradigm does not absolutely preclude the use of systems whose output is based directly on human expert judgments. As discussed in Section 2.1 below, whatever the approach used, the validity and reliability of the system should be tested under conditions reflecting the condition of the case under investigation and the best performing system should be used irrespective of whether it is based directly on human expert judgments, based on data, quantitative measurements, and statistical models, or a fusion of the two.