In this presentation I argue 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. I focus 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. I consider the aural-spectrographic approach to forensic voice comparison (also known as "voiceprint" or "voicegram" examination) in light of this principle, and also the currently widely practiced auditory-acoustic-phonetic approach (these two approaches do not appear to be mutually exclusive). Finally, I challenge the audience members to consider what each of them thinks constitutes the relevant principles regarding acceptable practice, and then consider their own approach to forensic-acoustic analysis in light of those principles.
INTRODUCTION

This proceedings paper is an abridged version of a full-length paper which has been submitted for publication elsewhere. The present version does not include discussion of the auditory-acoustic-phonetic approach mentioned in the abstract, although it is discussed in the conference presentation itself.

The title of this special session, and of this opening presentation, was deliberately chosen to be provocative, but is probably somewhat (if not highly) inaccurate: I don’t expect any of the presentations to actually provide a definition which could be used to include everything one wants to count as science and to exclude everything one doesn’t want to count as science, a problem known in philosophy of science as the demarcation problem. I do, however, hope that the presenters will provide a discussion of what they consider to be relevant principles governing acceptable practice in forensic science in general and forensic acoustics 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.

The title is also inaccurate with respect to this particular presentation in that it will be both more general and more specific: I will be talking about issues affecting forensic science in general and considering examples from forensic voice comparison in particular.

There are serious problems with current practice in forensic science, as documented in the National Research Council report on Strengthening forensic science in the United States: A path forward (NRC, 2009). Although this report is from the United States, I would be very surprised if similar problems did not exist in Canada and in other parts of the world.

The message of the NRC report could be summarized as “forensic science should be more scientific”, and it explicitly calls for the adoption of a “scientific culture” (NRC, 2009, p. 125). From a philosophy and sociology of science perspective, Cole (2010) 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” (Cole, 2010, 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 (NRC, 2009, pp. 22–23), will be the focus of my presentation, and can be summarized as: The validity and reliability of forensic analyses should be tested.

This recommendation is also an element of a paradigm for the evaluation of forensic evidence which I have been advocating for several years. It consists 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. (2012) and Morrison (2012). Details of my thoughts on selecting an appropriate database for forensic-voice-comparison cases appear in Morrison, Ochoa, & Thiruvaran (2012), and details of my thoughts on appropriate metrics and methodology for testing validity and reliability for forensic comparison in general appear in Morrison (2011).

TESTING OF VALIDITY AND RELIABILITY UNDER CONDITIONS REFLECTING THOSE OF THE CASE UNDER INVESTIGATION USING DATA DRAWN FROM THE RELEVANT POPULATION

Introduction

Why is it essential to measure the validity and reliability of a forensic analysis under conditions reflecting those of the case under investigation using samples drawn from the relevant population? Quite simply such testing is the
only way to demonstrate the degree to which a forensic system does what it is claimed to do, and to demonstrate the degree of consistency with which it does that.

**Lack of Testing of Experience-based Systems**

It is my impression that practitioners of experience-based approaches are often unable or unwilling to undergo validity and reliability testing. I have even heard one practitioner of such an approach claim in court that the validity and reliability of forensic voice comparison cannot be tested, and another claim that their approach to forensic voice comparison was scientifically valid because it was reproducible and testable but without presenting any evidence that their analysis had in fact been reproduced or that their ability to do what they claimed to be able to do had in fact been tested.

Some of this is likely due to a practical issue: systems based on data, quantitative measurements, and statistical models are often wholly or substantially automated and once the system has been built, tailored, and optimized to the relevant population and the conditions of the case under investigation it may be relatively easy to run a large number of test trials; in contrast, an experience-based system may have to start from scratch on each trial. There may be a large investment in setting up the former, but then little additional cost for each test trial, whereas for the latter there may be moderate investment on the first test trial and the same moderate investment on every other test trial resulting in a rapidly increasing total investment as the number of test trials increases. The outcome is that experience-based practitioners can perform one trial, the actual comparison of the suspect and offender sample, relatively cheaply and relatively quickly compared to the cost of building, tailoring, and optimizing a system based on data, quantitative measurements, and statistical models.

**Lack of Testing / Lack of Appropriate Testing of Systems Based on Data, Quantitative Measurements, and Statistical Models**

Leaving aside the issue of whether I am critical of the design of any particular system based on data, quantitative measurements, and statistical models, I have seen such systems used inappropriately in both research and casework. The principal problems are inappropriate selection of the relevant population and a sample thereof, and no testing or inappropriate testing of validity and reliability. A likelihood ratio as a forensic strength-of-evidence statement is the answer to a particular question, a question defined in part by the defense hypothesis which in turn defines the relevant population from which to draw a sample. This sample is then used to calculate the denominator of the likelihood ratio. Samples for testing the system should also be drawn from the same population. Pairs of samples used for testing the system should also reflect the same recording conditions and mismatched in recording conditions (speaking style, background noise, telephone transmission, etc.) as the actual suspect and offender recordings. Testing using data from a different population and from different recording conditions, may not be informative as to the performance of the forensic-voice-comparison system under the conditions of the case under investigation. These are issues which we discussed at length in Morrison, Ochoa, & Thiruvaran (2012), and so I do not discuss them further here.

**THE SPECTROGRAPHIC/AURAL-SPECTROGRAPHIC APPROACH**

**Introduction**

I appeared as an expert witness in two cases in Australian courts last year (2012) where I was asked to critique reports submitted by experts who used the aural-spectrographic approach. The aural-spectrographic approach (aka “voiceprinting” and “voicegram identification”) consists of listening to suspect and offender recordings (and depending on the protocol also recordings of foil speakers), and looking at spectrograms made from each of those same recordings. Using both auditory and visual examination, the practitioner forms an experience-based subjective opinion as to whether the suspect and offender recordings were produced by the same speaker. The spectrographic approach is visual-mode only, but was supplanted by the aural-spectrographic approach at the beginning of the 1970s. In both approaches the practitioner’s opinion is based directly and wholly on their subjective experience-based judgment.

There has been much debate as to whether these approaches are appropriate. Unsupported claims of near infallibility have been made, and at times the debate has been acrimonious (see, for example, Hollein, 2002, pp. 24–25 and ch. 6; and Koenig, 2002). The most comprehensive balanced review of the use of the approaches and the

Reports Including Consideration of Principles for Determining Acceptable Practice

In 1968 Peter B. Denes, at that time the Chair of the Speech Communication Technical Committee (SCTC) of the Acoustical Society of America (ASA), appointed six SCTC members (including himself) to investigate the use of the spectrographic approach (see Pickett, 1969). In a sense this study group was the forerunner of our current ASA Forensic Acoustics Subcommittee (FAS), although there was approximately a 40 year gap between the study group being active and the formation of the FAS. At the time, a visual-mode only approach was prevalent and this was the focus of the study group’s investigation. The study group presented a draft report at the SCTC meeting on 9 April 1969, at which the SCTC endorsed the report. The final version (Bolt et al., 1970) was published in the Journal of the Acoustical Society of America (JASA) in 1970 with a footnote that the views expressed were those of the authors as individuals. The following quotes from the report are of particular interest in relation to my topic of principles governing acceptable practice in forensic science:

What kinds of evidence would convince scientists of the reliability of speaker identification based on voice patterns?

The usual basis for the scientific acceptance of any new procedure is an explicit description of experimental methods and of results of relevant tests. The description must be sufficient to allow the replication of experiments and results by other scientists. . . .

Lacking explicit knowledge and procedures, can individuals nevertheless acquire such expertise in identification from voice patterns that their opinions could be accepted as reliable? . . . Validation of this approach to voice identification becomes a matter of replicable experiments on the expert himself, considered as a voice identifying machine.

Thus, voice identification might be accomplished either on the basis of explicit knowledge and procedure available to anyone, or on the basis of the unexplained expertise of individuals. In either case, validation requires experimental assessment of performance on relevant tasks.

. . .

It may be objected that this minimal set of tests is unreasonably arduous. We do not believe that it is. As scientists we could accept no less in checking the reliability of a “black box” supposed to perform speaker identification. (Bolt et al., 1970, pp. 601–602)

Court determinations may also depend on the apparent validity of exhibits brought in evidence. (Bolt et al., 1970, p. 602)

We conclude that the available results are inadequate to establish the reliability of voice identification by spectrograms. . . . Procedures exist, as we have suggested, by which the reliability of voice identification methods can be evaluated. We believe that such validation is urgently required. (Bolt et al., 1970, p. 603)

The principles expressed in Bolt et al. (1970) parallel the second and third elements of the paradigm I promote: highly preferred use of approaches based on quantitative measurements, databases representative of the relevant population, and statistical models; and obligatory testing of validity and reliability under conditions reflecting those of the case under investigation using test data drawn from the relevant population. The first element, obligatory use of the likelihood-ratio framework, was not introduced to forensic voice comparison until the late 1990’s (see Morrison, 2009, for a history).

The principles and conclusions expressed in Bolt et al. (1970) were also similar to principles and conclusions expressed on pages 60–62 of a 1979 NRC report on the aural-spectrographic approach prepared at the request of the Federal Bureau of Investigations (FBI, the committee who authored the report included both proponents and opponents of the approach), on pages 12–13 and 22 of the previously-mentioned 2009 NRC report on forensic science in general, and on pages 21–22 and Recommendation 6.3 (p. 209) of the US National Institute of Standards and Technology (NIST) and National Institute of Justice (NIJ) 2012 report on forensic fingerprint analysis (Expert Working Group on Human Factors in Latent Print Analysis, 2012). A clear pattern emerges across all these sober reports issued over the last four-and-a-half decades: The key scientific concern regarding any approach to forensic analysis (including the spectrographic or aural-spectrographic approach for forensic voice comparison) is whether it has been demonstrated to be sufficiently valid and reliable under casework conditions. It is up to forensic scientists to demonstrate the degree of validity and reliability of the approach under casework conditions, and it is up to the
legal authorities, who may not have a good understanding of the approach itself, to decide whether the demonstrated
degree of validity and reliability is sufficient.1

Tests of Validity

Over the years there have been a number of tests of the validity of the spectrographic and aural-spectrographic
approached. Some of these are summarized in NRC (1979), Gruber & Poza (1995), and elsewhere, but one research
report draws our attention because it described the largest study conducted and it had the greatest impact on practice
and (for a number of years) admissibility. Tosi et al. (1972) reported on a study conducted between 1968 and 1970
using recordings of 250 US-English speakers, and just under 35 thousand experiment trials performed by 29
examiners. These experiments were conducted in visual-only mode. Prior to performing the test, as part of the
research protocol the examiners received one month of training in the spectrographic approach (the participants were
not previously-trained professional practitioners of the approach). In general, the examiners were presented with
spectrograms from a recording of one speaker, and a set of spectrograms from recordings of multiple other speakers
one of whom might be the same speaker. Some trials were closed set where the examiner knew that the target
speaker was included, but others were open set where the examiner had to either choose a recording as being
produced by the same speaker as on the first recording or say that none of the other recordings were produced by the
same speaker. The examiners also had to indicate how confident they were in their decision.

According to Tosi et al. (1972), in the most forensically realistic condition tested (open set and non-
contemporaneous recordings) the “false elimination” rate was 13% and the “wrong matches” rate was 6%. If only
the 74% of decisions when the examiners were “fairly certain” or “almost certain” were included, these rates
dropped to 5% and 2% respectively. These values were averaged across fixed and random word context, but results
for the more forensically realistic random context were said to be poorer than those for the fixed context. Tosi et al.
(1972) ended their paper by speculating as to how their results might be relevant for casework conditions.

Tosi et al. (1972) was immediately criticized by the ASA SCTC study group (Bolt et al., 1973; see also Gruber &
Poza, 1995).2 The primary criticisms were that the laboratory tests were methodologically flawed and did not reflect
casework conditions, and that Tosi et al.’s attempt to extrapolate to casework conditions was based on dubious
assumptions.3

Tosi et al. (1972) also made reference to a “field study” conducted at the Michigan Department of State Police
Crime Laboratory. This was a review of 673 aural-spectrographic cases conducted between 1967 and 1970. The
results, as reported in Tosi et al. (1972), where that no decision was made in 59% of cases because of poor audio
quality or quantity (as compared with 26% “almost uncertain” and “fairly uncertain” responses in the laboratory
study), and of the remainder, 38% were declared to be the same speaker and 62% different speakers. It was further
reported that of the same-speaker conclusions 29% were confirmed because the suspects “admitted culpability or
were convicted by evidence other than that produced by their voice” (p. 2042). The latter clearly has a danger of
circularity: a suspect may plead guilty even if they are innocent, and if the voice evidence is presented one cannot
determine the extent to which this contributed to a jury’s decision (it may be that the voice evidence was not
presented in these legal trials).

The argument in Tosi et al. (1972) appears to have been that the use of the aural-spectrographic approach (visual
and auditory) by professionals taking as much time as they need, as opposed to the use of the spectrographic
approach (visual only) by amateurs with only one month of training performing the task in a limited amount of time,
combined with the use of a “no decision” option will lead to higher correct-decision rates. As an argument in favor
of the use of the aural-spectrographic approach (not necessarily in contrast with the spectrographic approach), I find
this unconvincing. I need to see the results of tests under conditions reflecting those of casework, and for which
there can be no dispute as to the same-speaker or different-speaker status of every test pair. I do, however, believe it

---

1The brief pre-Daubert discussion of legal acceptability in Bolt et al. (1970) focused on the court’s determination of whether the approach had

gained general acceptability within the scientific community, which in turn they argued should be based on demonstrated degree of validity.

2In contrast to the opinions expressed by the ASA SCTC study group, Greene (1975, p. 189–190) and Tosi (1979, p. 144) quote extracts from a

letter written on 23 or 28 March 1973 by the then President of the ASA, Karl D. Kryter, which appears to give conditional support to the use of

the aural-spectrographic approach:

contrary to the resolution [Bolt et al., 1973?], it can be stated, in my opinion, that by scientific tests it has been proven within normal

standards of scientific reliability and validity, that voiceprints for some speakers, under certain conditions and with certain analysis

procedures, can provide positive identification . . . (Karl D. Kryter quoted in Tosi, 1979, p. 144)

There is no record of this letter in the minutes of ASA Executive Council meetings (p.c. Elaine Moran, ASA Office Manager, 13 November

2012) so it appears that Kryter wrote this letter in his personal capacity rather than in his capacity as President of the ASA.

3 Rather than discuss the “Tosi extrapolation” here, I recommend the review of this issue in Gruber & Poza (1995, part B). Extrapolation from

laboratory studies to casework was also discussed at length in the 1979 NRC report.
is self evident that if one avoids making a decision in cases one judges to be difficult and removes these from the statistics, then one will be left with cases which are generally easier and one will therefore achieve a better correct-decision rate. I also believe that in fact all practitioners, irrespective of their approach, decline to perform analyses when a priori they believe that the quantity or quality of the recorded material is such that their system is unlikely to produce a high strength of evidence in either direction. I believe that it would actually be unethical to proceed with a full analysis without at least making the client aware of the likely limitations. An attack on this practice per se I would not consider appropriate, but neither would I consider it appropriate to make unsubstantiated claims that this practice will eliminate or virtually eliminate errors. If this practice is part of casework conditions, then it should be included when assessing the degree of validity and reliability of a forensic system under casework conditions.

**Is the Fact That a Spectrogram Is Used a Key Aspect of the Criticism of the Approach?**

With respect to the spectrographic and aural-spectrographic approaches, is the fact that a spectrogram is used a key aspect of the criticism of the approach? I would argue that it is not. As I understand it, sober criticisms have always centered on the issue of whether the degree of validity and reliability of the approach has been demonstrated under conditions reflecting those of casework. If a forensic scientist did not use spectrograms, but, for example, instead measured formant values from tokens of a number of vowel phonemes, plotted these on a first-formant by second-formant (F1 by F2) plot, and then on the basis of looking at these plots made an experience-based subjective opinion, this approach would be subject to the same criticisms. Whether such an approach were deemed acceptable should depend on the principle of whether it had been tested under casework conditions and found to be sufficiently valid and reliable. The same criteria should apply even if a graphic representation is not used at all. The same criteria should apply to a purely auditory approach, to looking at a table of numbers, and, as argued earlier, to an approach based on quantitative measurements, databases representative of the relevant population, and statistical models.

**CONCLUSION**

I have argued that if one wants to determine whether a particular forensic analysis approach is acceptable, one should first specify what one considers to be the principles governing what would be acceptable. Once this has been done, the same principles can be applied to all approaches which one may want to consider.

One of the key principles in my opinion is that the validity and reliability of the approach be empirically tested under conditions reflecting those of the case under investigation using test samples drawn from the relevant population. This (or a very similar principle) was also proposed in each of the Acoustical Society of America Speech Communication Technical Committee study group’s report on *Speaker identification by speech spectrograms: a scientists’ view of its reliability for legal purposes* (Bolt et al., 1970), the National Research Council report *On the theory and practice of voice identification* (NRC, 1979), the National Research Council report on *Strengthening forensic science in the United States* (NRC, 2009), and the National Institute of Standards and Technology / National Institute of Justice report on *Latent print examination and human factors* (Expert Working Group on Human Factors in Latent Print Analysis, 2012).

I have considered the aural-spectrographic approach to forensic voice comparison from the perspective of this principle. In the end I will refrain from making an explicit statement as to whether I think this approach is acceptable. What I want to emphasize instead is that this decision should be based on principles, particularly whether the approach has been tested under conditions reflecting those of the case under investigation using test samples drawn from the relevant population, and been found to be sufficiently valid and reliable. I also want to challenge you, the members of the audience, to consider what principles you think are relevant, and then consider whether the approach you employ for forensic-acoustic analysis is acceptable given those principles.

**ACKNOWLEDGMENTS**

This research was supported by The Australian Research Council, Australian Federal Police, New South Wales Police, Queensland Police, National Institute of Forensic Science, Australasian Speech Science and Technology Association, and the Guardia Civil through Linkage Project LP100200142. Unless otherwise explicitly attributed, the opinions expressed are those of the author and do not necessarily represent the policies or opinions of any of the above mentioned organizations.
REFERENCES


