Questions, propositions and assessing different levels of evidence: forensic voice comparison in practice

Vincent Hughes[1] (Corresponding author)
Richard Rhodes[2,1]

[1]University of York

Department of Language and Linguistic Science
University of York
York
YO105DD
UK

[2]J P French Associates

J P French Associates
86 The Mount
York
YO24 1AR
UK

vincent.hughes@york.ac.uk
richard.rhodes@jpfrench.com

Abstract

This paper contributes to the ongoing discussion about the distinction between observations and propositions in forensic inference, with a specific focus on forensic voice comparison casework conducted in the UK. We outline both linguistic and legal issues which make the evaluation of voice evidence and the refinement of propositions problematic in practice, and illustrate these using case examples. We will argue that group-level observations from the offender sample will always be evidential and that the value of this evidence must be determined by the expert. As such, a proposal is made that experts should, at least conceptually, think of voice evidence as having two levels, both with evidential value: group-level and individual-level. The two rely on different underlying assumptions, and the group-level observations can be used to inform the individual-level propositions. However, for the sake of interpretability, it is probably preferable to present only one combined conclusion to the end user. We also wish to reiterate points made in previous work: in providing conclusions, the forensic expert must acknowledge that the value of the evidence is dependent on a number of assumptions (propositions and background information) and these assumptions must be made clear and explicit to the user.

1 Introduction

In their 2015 paper in *Science and Justice*, Hicks et al [1] discussed the formulation of propositions and the evaluation of forensic evidence. Specifically, they argued that it is essential that observations which form forensic evidence are not used to define propositions, upon which the evidence is conditional. Subsequent responses to this paper [2,3,4] have examined these issues more specifically in the context of forensic voice evidence. We wish to thank the authors for their stimulating discussion. Our paper is not intended to be a formal response to Hicks et al or Morrison et al, but rather a contribution to the wider scientific debate. Here, we present our views on the issues of evidence, propositions, and background information from the perspective of forensic voice analysts working in the context of the justice systems in England and Wales, and Northern Ireland. Specifically, we outline issues arising from the nature of the voice as a form of forensic evidence, exemplifying these with case examples, and present a framework for thinking about observations and propositions when evaluating voice evidence.

1.1 The likelihood ratio

A forensic likelihood ratio (LR) is an expression of the weight or strength of the evidence under the competing propositions of the prosecution and defence (for further discussion see [5,6,7]). It is expressed as:

$$\frac{p(E|H_p,I)}{p(E|H_d,I)}$$

where p is probability, H_p is the prosecution proposition, H_d is the defence proposition and I is background information in the case. The probability of the evidence (E) is conditional on the propositions and the background information, and in this way, the LR is the answer to a specific question. Appropriately defining the propositions, and in particular the defence (or alternative) proposition, is a crucial issue in forensic inference. This is because the defence proposition defines the relevant population which forms a baseline against which the expert assesses the typicality of the evidence. This is the same whether using statistical methods which require empirical data from a sample of the relevant population or more subjective methods based on published studies and experience.

1.2 General background

Hicks et al [1] argue that forensic observations (i.e. evidential analytical findings) should not be included in propositions. Evidence is evaluated by the forensic expert and is defined by two properties; firstly, whether the observations have some probative value, and secondly, whether expert knowledge is required to determine the value of the observations. Propositions (i.e. two mutually exclusive scenarios representing the prosecution and defence views on the evidence), however, are evaluated by the Court, and thus should not be "findings led" (p. 521). If the observations have no value or if the value of the observations can be determined without expert knowledge, then Hicks et al argue that they can be incorporated into the propositions.

In one example, they point to shoemark comparisons, where a 'common sense' approach taken by many experts is to assess the strength of evidence based on the alternative proposition that the shoemark must have been left by a trainer of a similar brand - e.g. on the basis that the print came from another Nike Air Max. However, they argue that this ignores the evidential value of narrowing down the shoemark as having come from this particular brand and model of trainer, as opposed to any other trainer or type of shoe. The example which prompted the response in [2] related to voice comparison evidence, where Hicks et al argue that, unless the *accent* of the offender is agreed by all parties, the alternative proposition should not include the expert's observations about group-level characteristics, such as regional background, age and gender (again, unless the court can be expected to assess and evaluate these aspects without expert knowledge).

Morrison et al [2] disagree with this position, demonstrating empirically that without a well-defined alternative proposition, experts will not be able to accurately and reliably carry out voice comparison, and might grossly misrepresent the strength of evidence (while also reducing the validity of the system). If the reference sample does not match the questioned samples well (e.g. for age, sex and language spoken), the magnitude of the LR will be inflated. Further, if the relevant population is too widely defined, and subsequently too large, it will not be possible to adequately sample the population for a case. They contend that, if the assumptions made are clear, the expert can select propositions based on group-level characteristics through a pre-analytical screening exercise. Further, they argue that the court will *usually* be able to reliably determine the regional background (defined by country) and sex of the offender and make an inference about the evidential value of these group-level characteristics. Following this approach, Morrison et al do not generally include the evidential value of group-level characteristics into their conclusion.

In Hicks et al's [3] rejoinder, they come to some agreement that the expert is rightly expected to form well-defined propositions, but that this process has evidential value and might be evaluated formally as an LR (if it requires expert knowledge). In an online reply, Morrison et al [4] claim that these characteristics will usually be obvious to all parties, and thus their assessment is usually outside the expert's domain as it does not require expert knowledge. Therefore, this can be used to form the alternative proposition, rather than being assessed as part of the evidence.

We agree with many of the points raised in this series of papers, that:

- The LR is the answer to a specific question
- The expert must carefully consider propositions in each case
- Group-level characteristics narrow down the pool of possible offenders
- Analysis methods are better, i.e., more valid and more reliable, when the relevant population matches the offender sample well
- Forensic evidence should be compatible with reasonable expectations of users
- Assumptions should, therefore, be made clear to those users

However, we would like to expand on these and further points, particularly in the context of forensic voice comparison evidence in England and Wales, and Northern Ireland. Morrison et al [4] express their satisfaction that the Hicks et al rejoinder "mostly resolves apparent disagreements between us". However, there remain fairly key questions to be addressed: should the expert incorporate the evidential value of group-level characteristics (age-group/gender/accent type etc.) into their conclusion, or can this be safely left to the court to assess? Does this require expert knowledge? Further, how and when can an expert include this information, and in what circumstances are they in a better position to do so than the Court? And fundamentally, is the expert's conclusion answering the question that the justice system is asking? We address these questions below and in section 4 provide example cases which illustrate these issues.

2 Practical issues in forensic voice comparison

Forensic voice comparison accounts for the vast majority (c. 70%; [8]) of work carried out by forensic speech scientists in legal and civil cases in the UK. Such cases usually involve the comparison of a voice in a recording of an unknown offender (e.g. a threatening telephone call) and a recording of a known suspect (e.g. a police interview). For a detailed overview of forensic voice comparison methods see [9,10,11]. In such cases, the prosecution proposition will be, straightforwardly, that the criminal recording and known recording contain the voice of the same speaker. At the most general level, the defence proposition is that the recordings contain the voices of different speakers. In the following sections we outline issues with the refinement of the defence proposition for forensic voice comparison evidence (for further discussion see [12,13]) relating to the nature of voice evidence and its evaluation in practice.

2.1 The nature of voice evidence

2.1.1 The voice as a carrier of group and individual information

Unlike other forms of forensic evidence (e.g. fingerprints), information about the groups of which the offender is a member is available via an evidential recording of his/her voice. The speech signal encodes information about both the individual speaker and the group(s) to which that speaker belongs. This theoretical dichotomy between individual- and group-level information is convenient, but notoriously problematic in linguistics (see [14]). Indeed, the complexity of the relationship between individual- and group-level information is one factor which makes speech a difficult form of forensic evidence (as discussed in [15]), especially when discussing the distinction between evidence and propositions. There are a number of reasons for this. The phonetic features which indicate a speaker's group memberships are often referred to as the speaker's *accent* (although within the field of forensic voice comparison, e.g. in Morrison et al [2], and outside of linguistics, the term is generally used restrictively to refer to a speaker's regional background). However, *accent* is multidimensional in terms of the regional and social groups which define it. In forensic voice comparison, there is generally a focus on 'regional background' (often defined broadly on a country level, e.g. Australian English; see [2]) and 'sex' (binary male or female). However, accent is much more than geography and sex. There may be many other relevant factors including socially-defined gender, socioeconomic class, ethnicity and geographical mobility (to name but a few). In many ways it is more appropriate to define a speaker's *accent* in terms of the point of overlap between numerous groups. Defined narrowly enough, this intersection between multiple groups may itself be individualising (i.e. it may reduce the population of potential offenders down to an extremely small number, or even a single person). *Accents* are also multidimensional in terms of the linguistic and phonetic features which characterise them. Speakers are often variable in speech production, even for features which are stereotypical of a certain region or social group (e.g. style shifting). What it means to be a member of any single group (with the exception of biologically fixed factors such as sex) is also fluid, dependent on a speaker's attitudes and stance, the topic of conversation and the interlocutor. Thus, there is no sense is which we can talk about a *uniform accent* which is the same across all members of a community. Finally, certain linguistic and phonetic features can encode both group- and individual-level information, and to varying degrees. For more discussion on the complex nature of between-speaker variation see [16,17,18].

2.1.2 Group-level information is evidential

We are of the view that, in the vast majority of forensic voice comparison cases, group-level information observed in the offender sample will be evidential, and that such observations should not necessarily be restricted to broadly-defined regional background and sex. We would argue that the value of this group-level evidence must be determined by the expert, not lay people, and be incorporated in some way into the expert's conclusion. Indeed, in some cases, group-level observations may provide the greatest probative value for the court in answering the fundamental question of whether the recordings contain the voice of the same speaker or not. This is why, just as with any other forensic process, analysts should be properly qualified to, and validated in their ability to, make accurate group-level observations and assess their value.

Using the criteria outlined in Hicks et al [1], there are a number of reasons why we consider group-level characteristics evidential in forensic voice comparison:

Making group-level observations requires expert knowledge

Morrison et al [2] argue that "it will usually be obvious to a forensic speech scientist whether the questioned speaker is male or female, what language they are speaking, and broadly what accent they are speaking. These properties will usually also be perceptually salient to all parties" (p. 493). However, there is no empirical evidence to support this contention. In fact, published studies suggest that lay people perform extremely poorly when attempting to determine even general information such as a speaker's regional background, even at the level of country of origin [19]. Performance is considerably worse when trying to identify more fine-grained regional or social groupings [20,21]. This is even more concerning if one takes into account the factors which can affect the quality of recordings in a forensic case; in particular, Clopper and Bradlow [22] show that even moderate occlusion by noise reduces listeners' ability to correctly determine regional groups. As highlighted above, the voice also encodes other group-level information than regional background and sex. Therefore, experts are in a position to provide considerably more evidential, group-level observations than lay people are.

The observations have value

All group-level observations will have some value to the court. This may be the case even for something as broad as language - particularly if the language spoken by the offender would be unfamiliar to lay people. Observations about the regional and social groups to which the unknown speaker belongs necessarily reduces the population of potential offenders. The more detailed the picture that can be formed of these groups, the smaller that population becomes. The value of these observations could therefore be considerable (see the case examples below).

To infer the value of the observations also requires expert knowledge

Even if the many decision makers in a court process were able to determine the characteristics of a voice, they will not have the knowledge and training to be able to make a forensic inference about the value of these observations and the effect on the strength of evidence. In order to empirically estimate the typicality (or rarity) of a speaker's *accent*, it may in some cases be useful to use census data (as described in Morrison et al's [2] Australia/New Zealand example). However, in most cases this is not appropriate. This is due primarily to the complexity of group- and individual-level patterns of speech production, as outlined in 2.1.1. Specifically, there is no direct mapping between geographical or social boundaries (or categories) and linguistic production (as highlighted in [18]). Thus, census data may not in any sense capture linguistically meaningful distinctions between groups. Any linguistic information captured by census data will, therefore, necessarily be statistically imprecise. Further, census information is generally focused on large-scale group factors (e.g. regional background). This may be fine if, like Morrison et al [2], you only consider the regional background and sex of the speaker, but will be incomplete for many other group-level observations that a sociolinguistically informed expert may make. Therefore, we would argue that the expertise of someone who understands sociolinguistic variability and its complexity is required in order to assess typicality, and infer the value of the observations in the context of the case.

2.2 Analysis of voice evidence in practice

2.2.1 Evidence in practice

The academic debate surrounding the probabilistic evaluation of evidence often oversimplifies how forensic evidence is used by the justice system. In most of the literature, the ‘Court’ or the ‘trier-of-fact’ are represented as the only stated user of a forensic report, and they will be able to read the report and have it explained to them by the authoring scientist. A well-formed, testable alternative proposition will be handed to the expert by the defence team. It is also expected that all parties may ‘agree’ vital pieces of information, sufficiently in advance of the trial so that the expert can use this information to develop the appropriate alternative proposition. In the context of the discussions between Hicks et al and Morrison et al, this might mean the recordings in a voice comparison case will be available to all parties and they will be able to agree properties of the voice as part of a defence position. However, in reality things are not so straightforward.

In our casework experience, a well-formed defence proposition is provided in less than 5% of criminal cases, and in some of these cases it is not possible to test those provided[1]. If a defence proposition is provided, it is generally produced in a defence statement very close to the time of the trial (sometimes the week before, in rare cases this comes earlier). The forensic report is generally produced much earlier in the process, however (normally anywhere between 6 weeks to 18 months before the trial). This time difference also means it

[1] For example, a defence hypothesis that states ‘it was not the defendant who made the call, it was person X’, where there is no recording of person X’s voice, or person X does not exist.

is very unlikely that the expert will be able to rely on ‘agreed’ information about the voice or the circumstances of the case at the stage when the forensic work is being carried out. The status of information in a criminal case is often fairly transient; it can change before the trial, and will most likely be questioned - and is therefore open to change - throughout the trial process, particularly if it relies on other witness testimony. One method suggested by the AFSP [23] is to provide different conclusions based on differing sets of propositions, but this might be seen as the expert improperly coaching the defence (i.e. the defence could pick the conclusion which best supports their case, ignoring the effect of changing priors). In most cases, therefore, the expert will have to formulate the alternative proposition him- or herself, and might have to revisit the analysis and conclusion if provided with a defence statement closer to the time of trial.

Further, forensic reports have many audiences and inform multiple decision stages before a trial starts. We should consider all stages rather than viewing a singular audience of ‘the Court’. For a police investigation, the report may come before or after arrest or charge, and the forensic evidence may affect decisions made by police officers and forensic managers, along with the Crown Prosecution Service (CPS), about whether the suspect is charged. It may also affect a decision by the CPS on which offences to prosecute and how to go about those prosecutions. It can have an impact on what strategy is employed by the suspect, with or without a solicitor, in interview, or later down the line in deciding whether to offer a guilty plea. It will affect what approaches are taken by prosecution and defence barristers, before and during the trial. This is particularly important if the expert evidence is ‘agreed’ between prosecution and defence (for example, if the conclusion offered by the prosecution witness is accepted by the defence, who either agree the conclusion offered although it does not support their version of events, or change their version of events to fit the forensic conclusion), which is often the case with forensic evidence. Furthermore, for most of these decisions, the expert is not present to explain the assumptions that are made; many of the decision-makers are receiving the information second hand without direct access to the report, as presenting evidence in court is actually a relatively rare activity for the forensic expert. More importantly, in the light of the present discussion, many of the decision-makers will not have access to the recordings in question. With so many people involved in the pre-trial process, and with many not having listened to the recordings, we question the idea that all parties in a case can be expected to share reasonable beliefs and expectations about group-level characteristics (as stated in Morrison et al [2,4]), *even if* they had the skill to make those judgments.

After all of these decisions are made, the report may be presented to a judge and/or jury. However, in the UK systems jurors will not have direct access to a forensic report (i.e. they are not given a copy to read), but the information is presented through the expert's oral testimony or by a barrister. In most cases, the expert is not present in court to explain the conclusion or its underlying assumptions. Where the expert's conclusion has been agreed, the report or its conclusions may simply be read to the jury by a barrister. It is vital, then, that the report incorporates a useful answer to the question being asked, and can stand alone without explanation of the underlying assumptions made by the expert. This is particularly difficult when many of the decision-makers have an expectation that forensic evidence will provide a binary 'match/non-match' result.

In summary, the idea of a defence proposition being agreed by all parties to a criminal trial according to information derived from a voice sample does not reconcile with our practical experience; we have simply never come across such a case. The usual position is that nothing is provided or agreed. The expert will therefore be responsible for forming the alternative proposition. They may have access to some conditioning information which will assist in forming propositions, but this is rarely certain and may change in the days leading up to and during the trial. Further, the expert may not be present to explain the assumptions underlying their conclusion, meaning the report must make these clear.

2.2.2 The current situation in forensic voice comparison

Based on discussions with colleagues around the world, it appears relatively common for voice experts to follow a procedure akin to the one described by Morrison et al [2], in which the expert refines the relevant population based on observations from the offender sample. This is typically restricted to decisions about the regional background and sex of the offender, but may include judgments about other group-level factors. The examination for the purposes of the report then consists of analysing speaker-specific properties of the voice (rather than group-level characteristics) and assessing their value relative to the already refined relevant population. However, as highlighted above, this approach fails to recognise that the observations made in refining the population will necessarily have evidential value, and in some cases considerable value (see the examples below).

A recurring issue for experts in the field is that the specific assumptions upon which the evidence is conditional are very rarely, if ever, made explicit to the user (Morrison et al [2] appear to be an exception to this rule). We think that the reason for this lies in the issues we raised in 2.2.1. In most cases, and this appears to be true outside of the UK as well, a coherent and agreed set of propositions (and specifically a defence proposition) is rarely given. Therefore, the expert is responsible for refining the relevant population in order to evaluate the evidence (leaving to one side, for now, whether these observations themselves have evidential value, we completely agree with Morrison et al [2] that some refinement of the population is required). However, since these assumptions are not formally agreed by the defence and that the defence's version of events can change at any time, there is a fear amongst experts that the assumptions could be used as a form of defence, in and of themselves. For instance, if the expert has refined the population to speakers of Newcastle (North East England) English based on the offender displaying linguistic features consistent with Newcastle, the defence may be able to argue that their client was born somewhere else, even if they grew up in Newcastle or had family from Newcastle. This is because sharing linguistic features of a region does not necessarily mean a person was born or lives there. Of course, such reasoning is flawed in that the alternative proposition is defined according to properties of the offender, not the suspect; however, it could be persuasive to the Court and may undermine the expert's testimony.

3 Assessing different levels of voice evidence

Given the issues raised above, we believe that it is useful to view voice evidence as having two levels: (1) group-level evidence, and (2) individual-level evidence. We hope this idea of 'levels of evidence' can become the default framework for conceptualising voice evidence. The implementation of this framework will necessarily be different in every case, and we therefore propose this as a way of thinking about voice evidence, and not a rigid structure in which conclusions should be presented; this does not preclude generating or presenting conclusions in the different ways currently found in forensic speech laboratories.

The value of these different sets of observations must be assessed using different assumptions. Group-level evidence requires very general assumptions of the kind described in Hicks et al's [1] criteria for propositions. These group-level observations can then be used to refine the population for evaluating individual-level evidence.

We believe that experts should consider both levels of evidence in all cases, acknowledging that they are each associated with different assumptions, and where possible assess the evidential value of both levels. This approach is also suggested by Morrison et al [2]: "If need be, two likelihood ratios can be presented, one based on demographic information and related to the refinement of the relevant population, and the other based on an acoustic and statistical analysis using the already refined relevant population". However, we think that it is best to provide one conclusion, and to do this using aspects of both the Hicks et al [1,3] and the Morrison et al [2] approaches.

3.1 Level 1: Group-level characteristics

As highlighted in 2.1, many group-level characteristics can be inferred from a recording of an offender's voice. These are essentially categorical judgments about the regional and social groups of which the offender is a member. The responsibility for making such observations and assessing their value must lie with the expert. Of course, there may be uncertainty associated with such observations which could be incorporated into the overall conclusion using fully Bayesian methods (see [24,25]). In order to formally assess the value of the group-level observations only very general assumptions about the alternative proposition are required (e.g. it wasn't the defendant, it was another person in the UK). The LR would resemble a random match probability (like for DNA), whereby the numerator would be 1 if both the suspect and offender were members of the same linguistic groups and the denominator would be the proportion of other people in the UK in those same linguistic groups. For example, if two percent of speakers in the UK speak with a certain accent, the LR would be:

$$\frac{1}{2/100} = 50$$

Therefore, based on the accent-level evidence, it would be 50 times more likely to find this accent assuming the sample were produced by the suspect than if it were produced by another person in the UK. It is of course possible for suspect and offender samples to display different regional or social patterns even when they are the same person - in cases involving voice disguise, bidialectalism or bilingualism, for example - and here the numerator would not be 1.

3.2 Level 2: Individual-level characteristics

The observations made as part of level 1 can then be used to refine the relevant population for evaluating individual-level characteristics of the speakers. Morrison et al [2] refer to these individual-level characteristics as “the measured acoustic properties of the voice”. However, this is a relatively narrow definition. Such characteristics could, in principle, consist of any auditorily-assessed or acoustic features within the speech signal (see French et al [26] and French [8] for more details on the features commonly analysed in forensic voice comparison). Although features at this level may indicate membership of certain groups, they will, separately and in combination, have value at level 2 if they are not uniformly represented within that group. For example, although many speakers of London area varieties of English might replace ‘th′ sounds in words like ‘thing’ with an ‘f’ sound, this still has some value if there is variation within the group. Once the group level is defined, feature typicality is assessed within that group.

3.3 Combining level 1 and level 2

As suggested in [2], the appropriate way in principle for the expert to express their conclusion based on two levels of evidence would be to provide two LRs. However, we have concerns about how this would work in practice. There is a growing body of research showing the difficulties that the courts, and especially juries made up of lay people (and therefore other lay users of expert evidence), as well as forensic scientists [7], have in interpreting expert evidence expressed in the form of a LR [27,28]. The presentation of two LRs is likely to be even more confusing, and potentially lead to much greater misunderstanding and misrepresentation of expert conclusions. This is because it is less clear how the separate LRs relate to each other and how both help to answer the question of whether two recordings contain the voices of the same or different speakers; this equates to giving two half-answers to what the legal system sees as a simple question.

We think it is therefore preferable, from the perspective of understandability and clarity, to present a single conclusion only (be that a verbal statement or a numerical value). Whether that LR comprises group- or individual-level evidence, or both, depends on the case itself. This is perhaps, theoretically, not an ideal solution. However, the single LR approach represents a compromise between the current situation whereby group-level evidence is generally not evaluated and not commented on as a means of interpreting the expert's LR, and the two LR approach, which is almost certainly not understandable by the average users of forensic evidence.

There are cases involving such atypical group-level factors (see 4.1 below), that the LR presented to the court/user would incorporate or relate only to these observations. This relies on an expert's subjective judgment about whether the group-level evidence is strong enough to constitute an answer to the user's question. This is, of course, problematic (and is why we think there should be a greater focus on testing relating to level 2 factors). However, more commonly, we envisage that the LR would relate to individual, level 2 evidence (see 4.2). In such cases, it should be made clear what conditioning assumptions have been made when evaluating the individual-level evidence (just as with any form of expert evidence) and a qualitative statement should be made about the group-level observations. In this way, the group-level evidence provides context for interpreting the value of the conclusion. In other cases, (see 4.3) it may be appropriate to offer a conclusion which incorporates evidential inferences at both level 1 and level 2.

In some cases (see 4.4), a completely different approach might be needed and the question of how best to combine evidence from each level remains open.

4 Example cases

In this section we describe some example cases which illustrate these issues, particularly where the definition of the defence proposition is problematic. These examples are intended to illustrate some of the points we make in sections 2 and 3, and are adapted from or motivated by a number of real cases. Although they might seem out of the ordinary, such cases are not unusual, due in part to the nature of voice evidence and linguistic variation. For the sake of convenience, let us assume that it is known and agreed, that the cases concern the UK only.

4.1 Hybrid accent

This example is used to illustrate the differences between the methods put forward by Hicks *et al* [1,3] and Morrison *et al* [2,4].

Mr Smith is suspected of leaving a threatening voicemail message. Both he and the offender have hybrid American and Geordie (i.e. from Newcastle in the North East of England) accents of English. By hybrid, we mean that the speech patterns share features commonly found in American English (e.g. production of /r/ in words like *car*) and Geordie English (e.g. production of the vowel in *face* as [ɪə], such that *face* sounds to the ears of outsides like *fierce*). The different conceptualisations of evidence included in propositions are:

Morrison et al

I the case is based in the UK

E acoustic & phonetic features of the voice

H_p Mr Smith was the offender leaving the voicemail

H_d The offender was not Mr Smith, it was another adult man with a hybrid Geordie-American accent of English

Conclusion: the evidence offers limited value, since for the population of Geordie-American accented male speakers of English, Mr Smith is, probably, very typical. Further, it is very difficult to assess the strength of this evidence because the population is extremely small; it would not be possible to collect a sufficient sample of the population for statistical testing. The user is expected to be able to determine the accent type, assess its rarity and the impact this might have on the evidence.

Hicks et al

I the case is based in the UK

E features of the voice[2]

H_p Mr Smith was the offender leaving the voicemail

H_d [Level 1] The offender was not Mr Smith, it was another person in the UK

H_d [Level 2] The offender was not Mr Smith, it was another adult man with a Geordie-American accent of English

Conclusion: the evidence incorporates the rarity of the hybrid accent as part of the level 1 analysis, and is therefore very strong. As above, establishing typicality for level 2 is very difficult due to problems with sampling such a minority group. The user doesn't need to make any further inference.

In this case, the division of voice features into those defined by group and those defined according to the individual would be far from straightforward, as different speakers of a hybrid accent might have different influences from each source accent. Further, it might be impossible to assess the typicality of voice features at an individual level because it is impossible to sample a reference population for hybrid Geordie-American speakers; what if no others exist? Taken overall, however, our preference would be for a model similar to that put forward by Hicks et al, where the user is not required to make a further (potentially unguided) inference from the expert's evidence.

4.2 An 'unremarkable' case

In this scenario, offender recordings (telephone calls) in a bank fraud are submitted for comparison with a suspect, Mr Jones. The offence is broadly linked to the Greater Manchester (GM) area as the recordings relate to bank accounts held within GM; however, the phone calls were made from an unknown location. The expert can establish from the fraudulent call recordings that the offender is a man in early adulthood (c. 20 - 45 years old) with an accent from the GM area; Mr Jones's voice also matches this profile. It is hard to assess the value of group-level (level 1) observations because the potential pool of offenders is not clearly delimited by the case information. It could be the UK, the North West of England, Greater Manchester, or a different area altogether. In this case, therefore, the expert

[2] This is not defined specifically in [1]

could state the group-level characteristics derived from the offender recordings and how they have shaped the following hypotheses, but make the analysis on a level 2 basis:

H_p		Mr Jones was the offender making the telephone calls
H_d	[Level 2]	The offender was not Mr Smith, it was another young adult man with a GM accent of English

The level 2 analysis is made by assessing the similarity of the samples, and assessing the typicality of the features of the questioned voice against a model of other young adult, male speakers of GM-accented English. The conclusion is given on the level 2 analysis, with the conditioning information from level 1 described alongside. It is then up to the user to consider the impact of group level information as it is not taken into account in the expert's conclusion - it might be of relatively low value if other information in the trial leads them to consider that GM is the relevant area.

4.3 Rare accent

Imagine another case in Greater Manchester, where a series of hoax 999 recordings across a number of months come from telephone boxes across the GM area. The caller is apparently a young adult man with an Australian accent. A suspect with a similar profile - Mr Douglas - is arrested and interviewed, and the recordings are submitted for comparison. In this case, the group-level evidence could be much more important and relevant than the individual-level evidence. At level 2, the propositions would be:

H_p		Mr Douglas was the offender making the telephone calls
H_d	[Level 2]	The offender was not Mr Douglas, it was another young adult man with an Australian accent of English

It seems inappropriate that the expert in this instance answers the question 'how strong is the evidence for or against Mr Douglas being the speaker in the calls?' with 'slightly more likely than it being any other young, male Aussie speaker' - this pool of potential offenders includes, mainly, men in Australia who have little opportunity to abuse phone boxes in Greater Manchester. Rather, the evidence should address the likelihood of the caller being Mr

Douglas against the likelihood of the caller being ‘another young male Aussie speaker *in the GM area*’, incorporating both level 1 and level 2 evidence:

I The offending is linked to the GM area

H_p Mr Douglas was the offender making the telephone calls

H_d [Level 1] The offender was not Mr Douglas, it was another person in GM

H_d [Level 2] The offender was not Mr Douglas, it was another young adult man with an Australian accent of English

*E*1 The rarity of the Australian accent in GM

*E*2 Any individual-level features according to H_d Level 2

See Hicks et al [3] for E1 / E2 descriptions

How this analysis and interpretation is carried out might depend on the expert’s interpretation and sampling method, and the availability of relevant demographic/ migration statistics. However, this might be a case where census data could be employed to give a broad estimate of the value of (at least the geographical origin element of) level 1 evidence (*E*1). The value of each LR could be calculated and reported separately (a method we do not support). More practically, the LRs from each level could be combined into one conclusion (as suggested in [3]), especially for interpretative methods which do not require empirical data sampling (which, as above, would be difficult for the small population represented in level 2 in this example). In this way, the analysis uses a well-matched reference population and takes into account the evidential value of group-level observations in the context of the case.

In truth, the division in the last two examples is an over-simplification - there are no clear distinctions between ‘typical’ and ‘rare’ accents, these fall on a spectrum. For example, in reference to 4.1 above, dissecting the population along a few simple demographic lines (age, gender, accent type) reveals that this ‘unremarkable’ young adult male GM accent group is still a minority (around 5-10% of the population). Further, even those accents which one might view as typical are, in reality, usually more multifaceted than these descriptions account for. However, the examples serve the point of demonstrating different approaches that might be taken, and how they depend on different case circumstances.

4.4 Positive level 1, Negative level 2

It is relatively straightforward to imagine different types of cases where the evidence provides support for the defence hypothesis: i.e., the view that the speakers are different people. This could be due to differences at level 1: if the speakers have different genders, speak different accents/dialects of English, and/or are clearly of different ages (in contemporaneous recordings). In contrast, differences in voice features between speakers within a similar group could be found at level 2. However, interpreting differences might become problematic where level 1 and level 2 analyses offer contrasting results.

Envisage a case where a sexual assault is video-recorded using a mobile phone. The assault takes place in a village in the Scottish Highlands. The male offender has a London-area accent. An interview sample is retrieved from the suspect, Mr Brown, who has a similar accent type. Given the rarity of this accent type and the background information (i.e. that the crime took place in the Highlands), the level 1 evidence provides fairly strong evidence for the same-speaker hypothesis. However, the analysis at the individual level reveals moderate-strength evidence that the speakers are different people. If the two (level 1 and level 2) LRs are presented in tandem or combined, the evidence may very well support the same-speaker view, despite the two samples showing differences at the individual level. The common sense answer seems clear: that a combined level 1 + level 2 evaluation accounts better for this situation, or that the level 2 negative LR should 'override' the level 1 positive LR (however, this requires post-hoc interpretation, which is not ideal once the testing propositions have been set.) Alternatively, the expert might explain the outcomes from both levels in detail.

This type of case might raise the warning that no one-size-fits-all solution exists. Rather, experts should be aware of, and competent in applying, logical frameworks for assessing evidence according to a whole range of circumstances, and those which may be particular to their field.

5 Conclusions

We intend that this paper will contribute to the ongoing debate about the nature of evidence and propositions in forensic science, and particularly the implications for forensic voice comparison casework. Below we summarise our key points:

The nature of voice evidence

- The voice is a complex biometric encoding a considerable amount of group-level (as well as individual-level) information, which is accessible when listening to forensic recordings
- Making observations about group-level characteristics should always be within the expert's domain because users in the legal system may not have access to recordings, and even if they do, non-experts cannot be expected to make accurate group-level observations or assess their evidential value
- It is useful to conceptualise voice evidence as having two levels: group-level and individual-level evidence
    - Experts should consider, and where possible, assess both levels of evidence. This means including observations in propositions, but in a logically coherent way
    - Experts should be aware of the different assumptions that underlie their evaluation, and clearly explain these to users
    - In most cases, group-level observations frame propositions and are integral to the analysis
    - In some cases, group-level evidence may be central to the expert's conclusion
    - Experts should therefore be tested to validate their ability to make and evaluate group-level observations
- In the interests of clarity, we suggest presenting only one conclusion
    - This will usually be based on level 2 evidence, with a statement about the level 1 assumptions conditioning the evidence
    - This is preferable to the two LR approach suggested by Morrison et al [2], given the difficulties that users have with interpreting even a single piece of forensic evidence
    - In some cases this will be on the basis of combined level 1 and level 2 evaluations; i.e., assessing both group- and individual-level characteristics
    - In rarer cases, this may be based only on level 1 evaluations (where the *accent* itself is extremely unusual; such as in the Geordie-American example above)

Expert evidence in practice

- The end-user of a forensic report is not always a Court, and very rarely are experts called to explain to the Court the assumptions they used in evaluating the evidence
- Given the nature of the (UK's) legal systems and processes, it is extremely unrealistic to assume that the expert will be given a formal defence proposition or information about the voice samples which is 'agreed' by both sides. Therefore, the forensic expert must make pragmatic decisions about the conditioning information used to evaluate the strength of the voice evidence
- These assumptions should be made explicit by the expert in their report
- It is essential that users of forensic evidence understand an expert's conclusion is the answer to a specific question; a forensic conclusion is not interpretable in isolation, rather it is conditional on propositions and information, such that changes to those conditioning assumptions necessarily change the expert's conclusion. With more open discourse between forensic scientists and the legal community, there may be more fundamental shifts in practice such that specific propositions to test are provided by the defence

We wish to reinforce the point made by Hicks et al that "open scientific debate, in an atmosphere of mutual respect, is a key enabler to progress, especially when it comes to the complexity of interpretative issues in forensic science" (p. 402). However, as we have hopefully made clear in this paper, debate solely within the academic community does not necessarily resolve many of the practical or discipline-specific issues facing forensic science. We would like to encourage much greater debate between academics, forensic practitioners from different specialisms and the legal community. As with all aspects of the application of the LR framework to forensic evidence, it is only through interdisciplinary communication that we will be able to apply theoretically logical frameworks for evaluating all kinds of forensic evidence, and for them to be understood and accepted by the legal system.

6 References

[1] Hicks, T., Biedermann, A., de Koeijer, J. A., Taroni, F., Champod, C., & Evett, I. W. (2015). The importance of distinguishing information from evidence/observations when formulating propositions. *Science & Justice 55*(6): 520-525.

[2] Morrison, G. S., Enzinger, E., & Zhang, C. (2016). Refining the relevant population in forensic voice comparison–A response to Hicks et alii (2015). *Science & Justice* 56(6): 492-497.

[3] Hicks, T., Biedermann, A., de Koeijer, J. A., Taroni, F., Champod, C., & Evett, I. W. (2017). Reply to Morrison et al. (2016) Refining the relevant population in forensic voice comparison – A response to Hicks et alii (2015) The importance of distinguishing information from evidence/observations when formulating propositions. *Science & Justice* 57(5): 401-402.

[4] Morrison, G. S., Enzinger, E., & Zhang, C. (2017). Reply to Hicks et alii (2017) Reply to Morrison et alii (2016) Refining the relevant population in forensic voice comparison – A response to Hicks et alii (2015) The importance of distinguishing information from evidence/observations when formulating propositions. Unpublished letter at URL (accessed 09/06/2017): [http://geoff-morrison.net/documents/Morrison,%20et%20al%20(2017)%20reply%20to%20Hicks,%20et%20al%20(2017)%20-%202017-04-25a.pdf]

[5] Robertson, B. and Vignaux, G. A. (1995) *Interpreting Evidence - Evaluating Forensic Science in the Courtroom*. Wiley: Chichester.

[6] Evett, I. W., Jackson, G., Lambert, J. A. and McCrossan, S. (2000) The impact of the principles of evidence interpretation on the structure and content of statements. *Science & Justice* 40: 233-239.

[7] Aitken, C. G. G. and Taroni, F. (2004) *Statistics and the Evaluation of Evidence for Forensic Scientists (2nd edition)*. Wiley: Chichester.

[8] French, J. P. (2017) Developmental history of speaker comparison. *Festschrift for the Eminent Phonetician Jack Windsor Lewis on the Occasion of his 90th Birthday, English Phonetics* 21.

[9] Foulkes, P. and J. P. French (2012). Forensic speaker comparison: a linguistic-acoustic perspective. In P. Tiersma and L. Solan. (Eds.) *Oxford Handbook of Language and the Law*. Oxford: Oxford University Press, pp. 557–572.

[10] Gold, E. and J. P. French (2011). International practices in forensic speaker comparison. *International Journal of Speech, Language and the Law* 18(2): 293–307.

[11] Morrison, G. S., Enzinger, E. and Zhang, C. (2017) Forensic speech science. In I. Freckleton and H. Selby (Eds.) *Expert Evidence*. Sydney: Thomson Reuters, ch. 99.

[12] Hughes, V. (2014) The definition of the relevant population and the collection of data for likelihood ratio-based forensic voice comparison. Unpublished PhD Thesis, University of York, UK.

[13] Morrison, G. S., Ochoa, F. and Thiruvaran, T. (2012) Database selection for forensic voice comparison. In *Proceedings of Odyssey 2012: The Language and Speaker Recognition Workshop*. Singapore, pp. 74-77.

[14] Garvin, P. L. and Ladefoged, P. (1963) Speaker identification and message identification in speech recognition. *Phonetica* 9: 193-199.

[15] Gold, E. & Hughes, V. (2013) Issues and opportunities: the application of the numerical likelihood ratio framework to forensic speaker comparison. *Science & Justice* 54(4): 292–299.

[16] Foulkes, P. and Docherty, G. J. (1999) The social life of phonetics and phonology. *Journal of Phonetics* 34: 409-438.

[17] Patrick, P. (2008) The speech community. In J. K. Chambers, P. Trudgill and N. Schilling-Estes (Eds.) *Handbook of Language Variation and Change*. Oxford: Wiley-Blackwell, pp.573-597.

[18] Britain, D. (2013) Space, diffusion and mobility. In J. K. Chambers and N. Schilling (Eds.) *Handbook of Language Variation and Change (2nd Edition)*. Oxford: Wiley-Blackwell, pp. 471-500.

[19] Van Bezooijen, R. and Gooskens, C. (1999). Identification of language varieties: The contribution of different linguistic levels. *Journal of Language and Social Psychology 18*(1): 31-48.

[20] Williams, A., Garrett, P. and Coupland, N. (1999) Dialect recognition. In D. Preston (Ed.) *Handbook of Perceptual Dialectology*. Philadelphia: John Benjamins, pp. 345-358.

[21] Clopper, C. G. and Pisoni, D. B. (2004). Some acoustic cues for the perceptual categorization of American English regional dialects. *Journal of Phonetics 32*(1): 111-140.

[22] Clopper, C. G. and Bradlow, A. R. (2008). Perception of dialect variation in noise: Intelligibility and classification. *Language and speech 51*(3), 175-198.

[23] [UK] Association of Forensic Science Providers. (2009). Standards for the formulation of evaluative forensic science expert opinion. *Science & Justice* 49: 161-164.

[24] Brümmer, N. (2011) Fully Bayesian LR: extending the paradigm shift. *Invited talk at the Netherlands Forensic Institute (NFI)*. 19-20 October 2011. https://sites.google.com/site/nikobrummer/NFI_BayesianLR.pdf?attredirects=0 (accessed: 26th February 2018)

[25] Brümmer, N. and Swart, A. (2014) Bayesian calibration for forensic evidence reporting. In *Proceedings of the 15th Annual Conference of the International Speech Communication Association (Interspeech),* Singapore, pp. 388-392.

[26] French, J. P., Nolan, F., Foulkes, P., Harrison, P. and McDougall, K. (2010) The UK position statement on forensic speaker comparison: a rejoinder to Rose and Morrison. *International Journal of Speech, Language and the Law* 17(1): 138-183.

[27] Mullen, C. *et al* (2013) Perception problems of the verbal scale. *Science & Justice* 54(2): 154-158.

[28] Martire, K. A. *et al* (2014) On the interpretation of likelihood ratios in forensic science evidence: presentation formats and the weak evidence effect. *Forensic Science International* 240: 61-68.