

## The UK position statement on forensic speaker comparison: a rejoinder to Rose and Morrison

*Peter French, Francis Nolan, Paul Foulkes,  
Philip Harrison and Kirsty McDougall*

### Introduction

An article 'Position Statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases' (hereon PS 2007) was published in vol 14.1 (2007) on behalf of UK forensic phoneticians (French & Harrison 2007). The editors invited responses to PS 2007; the first and only response to date, by Rose and Morrison (hereon R&M 2009), was published in vol 16.1 (2009). The present article is a rejoinder to that response from five of those responsible for the formulation of PS 2007. In addressing the response we confine ourselves to the central issues that trouble us in the position it represents. First, however, we confirm our agreement with some matters of general principle.

### Points of agreement

Like Rose and Morrison, we take the view that the claims embodied in the positive judgements of impressionistic likelihood scales (e.g., '(highly) likely to be the same person') are logically unsustainable. The crucial insight of expressing a conclusion via a likelihood ratio (LR) is that it matters not only how likely the

---

#### Affiliation

Peter French, Paul Foulkes, Philip Harrison: University of York & JP French Associates

Francis Nolan, Kirsty McDougall: University of Cambridge

email: jpf501@york.ac.uk

---

evidence is on the hypothesis that the suspect was responsible for leaving it, but also how likely it is given the alternative hypothesis that it was left by someone else. This was a central motivation for the discussions that led to PS 2007. The framework set out in that document serves to remind forensic phoneticians of the need to judge the distinctiveness of the features found in the criminal and suspect samples, and this implies comparison with a broader population, albeit informally via the analyst's experience and general linguistic knowledge rather than formally and quantitatively.

The LR conceptual framework also rightly takes the onus of identification away from the expert, whose role it is to assess the strength of the evidence, leaving identification to the triers of fact. This is again reflected in PS 2007 (but see below on the issue of elimination). In this respect PS 2007 also advocated the term 'comparison' rather than 'identification'. It was in relation to this specific point that we spoke of bringing forensic speech analysis into line with modern thinking in forensic science and DNA analysis (PS 2007: 138).

Finally, as we stated in PS 2007, 'we accept in principle the desirability of considering the task of speaker comparison in a likelihood ratio (including Bayesian) conceptual framework. However, we consider the lack of demographic data along with the problems of defining relevant reference populations as grounds for precluding the quantitative application of this type of approach in the present context'. Whilst we take no issue with introducing overall qualitative LR expressions into the formulation of conclusions, it is the feasibility of applying a fully quantitative LR approach derived from demographic data – even in the long term – that lies at the centre of our disagreement with R&M 2009.

## Points of divergence

### Samples which are 'not consistent' with each other

PS 2007 allows the forensic phonetic expert (in effect) to eliminate speakers by deeming their speech sample to be 'not consistent' with the unknown sample. This is clearly at odds with a strict LR approach. We also acknowledge that PS 2007 incorrectly claimed that there was 'no logical flaw' in allowing for categorical exclusion in this way. However, whilst to some extent we share R&M's unease at the way in which this imposes a discrete decision, in contrast to the continuous range of LRs available in a purely quantitative approach, we believe that in practical terms it is justified. Moreover, the same flaw is, in fact, found in R&M's response. They acknowledge that instances do occur where exclusion appears to be the justified outcome:

Of course, we can imagine some conditions under which a voice comparison could result in a *definitive exclusion*, e.g., the vocal tract of a young child

could not produce the lower formants of a typical adult male, but in such cases the voices are likely to sound so different that it would be highly unlikely that a forensic expert would be consulted. (p. 150, emphasis added)

We take this to mean that, in this admittedly unlikely scenario, R&M would themselves reach an absolute ('definitive') conclusion (thus akin to our term 'not consistent') rather than arriving at a ratio to indicate that the lower formant values found in the samples overwhelmingly support the hypothesis that the speaker was a man rather than a child.

In our extensive combined experience of speaker comparisons (several thousands) it is in fact quite common to have putatively matching samples that survive a long way through the law enforcement process even though they are quite distinct in internally-consistent ways. The distinctness might well, for instance, be at the level of micro-dialect; but we are familiar with cases where even quite salient dialect, sex/gender or age differences have been overlooked by the law enforcement authorities. We have, for example, been asked to compare samples where the suspect was a middle-aged female heroin addict with pathological voice features (including very low pitch, and harsh and creaky phonation) but the disputed sample transpired to be spoken by her five year-old daughter. In several other cases, where there has been no question of disguise, we have encountered gross mismatches in dialect when comparing reference and disputed samples (Welsh/Liverpool, Manchester/Northern Irish, Manchester/Southern Irish, Panjabi/Jamaican, Nigerian/Caribbean).

In such cases of dialect mismatch it can always be objected that there exist consummately bi-dialectal speakers. However, establishing within the LR framework (or indeed any framework) that two recordings with internally-consistent different dialects in fact come from a bi-dialectal speaker will be well-nigh impossible, because the differences that could be measured will be ambiguous in terms of their source: are they the result of a different speaker's vocal tract, or just the dialect?

In practical terms it would be a waste of everyone's time to proceed when demonstrable acoustic or linguistic mismatches exist between the samples. We would concede that there is a substantive issue over how distinct samples have to be to allow a 'not consistent' (or 'exclude') opinion, and would expect R&M to use their exclusion option more reluctantly than is our practice; but in the present state of knowledge we regard this option as an essential one. There is no ontological difference between R&M making a 'definitive exclusion' in an extreme case such as the little girl/adult male example they cite, and those where, in the less extreme circumstances exemplified above, we would deem the samples 'not consistent'.

Furthermore, we note that during the course of speaker comparison analysis, discrete decisions are routinely and unavoidably made without the support of LRs. For example, in editing a multi-speaker recording for phonetic analysis the analyst excludes from their composite sample those voices that they consider to be from speakers other than the target. Conversely, positive identifications are also made in order to establish which sections of the recording can be concatenated in order to provide a continuous working sample of the target's voice.

### The absence of population statistics

R&M recognise the criticism that the proportion of speech features for which appropriate population statistics exist is vanishingly small, particularly once the issue of what constitutes a relevant population in a particular case is considered. In their work two main solutions are proposed for dealing with this. One is to be found in Rose (2007), the other in R&M (2009). These are discussed in turn below.

### Gathering the population data

Rose (2007) has suggested a strategy that involves the analyst putting together a reference population to meet the needs of each particular case. This may entail making appropriate recordings from other speakers with a similar demographic and linguistic background to the suspect, followed by analysis of the corpus.

Rose argues: '[w]e have, as Traditional LR-based FSR experts, to be prepared to go and get a suitable reference sample for each case' (2007: 52). He notes that this task is 'considerably easier' for experts working in automatic speaker recognition. However, he provides an illustration of how it is possible to gather reference data based on analysis of formant trajectories of /je-/ from the word 'yes', where samples were collected from 30 speakers.

Whilst this example illustrates the general principle neatly, it is essential to keep in mind the potentially vast range of acoustic, phonetic, broader linguistic, and indeed non-linguistic features that are very often relevant in a given forensic comparison. (The fact that some of these features do not pertain to voice but to language and non-linguistic behaviours provides part of the basis for our referring to the forensic task as 'speaker comparison' rather than 'voice comparison'.) Amongst the features commonly considered in speaker comparison cases are the following:

1. Vocal setting and voice quality. Full analysis (using a version of the Laver VPA scheme; Laver 1980, 1994) distinguishes phonation features, overall muscular tension features and vocal tract features, with up to 38 individual elements to be considered.

2. Intonation, potentially including analysis of tone unit nuclei, heads and tails.
3. Pitch, measured as average and variation in fundamental frequency.
4. Articulation rate.
5. Rhythmical features.
6. Connected speech processes such as patterns of assimilation and elision.
7. A large set of consonantal features, including energy loci of fricatives and plosive bursts, durations of nasals, liquids, and fricatives in specific phonological environments, voice onset time of plosives, presence/absence of (pre-)voicing in lenis plosives, and discrete sociolinguistic variables.
8. A large set of vowel features, including acoustic patterns such as formant configurations, centre frequencies, densities, and bandwidths, and auditory qualities of sociolinguistic variables.
9. Higher-level linguistic information including use and patterning of discourse markers, lexical choices, morphological and syntactic variants, pragmatic behaviour such as turn-taking and telephone call opening habits, aspects of multilingual behaviour such as code-switching.
10. Evidence of speech impediment, voice and language pathology.
11. Non-linguistic features characteristic of the speaker, for example patterns of audible breathing, throat-clearing, tongue clicking, and both filled and silent hesitation phenomena.

It is clear that this list extends very substantially beyond a subset of formant frequencies or trajectories. To attempt to collect and analyse adequate reference data that would include this range of features would be prohibitively difficult. This becomes clear when one considers that a detailed and comprehensive analysis of a known recording of a single suspect, similar analysis of a questioned, criminal sample and the interpretation of findings never takes less than many hours and may take days. Given, then, the likely additional time and costs involved in collecting and analysing reference data, it might well be that those instructing forensic speech experts would refuse to underwrite the exercise. In a recent presentation by the National Manager of Forensic and Data Centres of the Australian Federal Police, warnings were issued against experts turning each forensic case into a research project (Robertson 2007).

A potential solution to this difficulty would be to limit the analysis to the features for which reference data already exist or could be obtained relatively easily. However, it is a forensic imperative for the analyst to consider all aspects of the materials in question. Even though research has established that, for example, some vowel formant patterns have a certain discriminative power, this does not absolve the analyst from considering the many other factors which could, in principle, radically alter the outcome of the comparison, either to reinforce or to overturn a conclusion drawn on the basis of a limited set of features. Indeed, as Rose points out in relation to a formant-based approach ‘there are plenty of different-speaker comparisons that yield LR<sub>s</sub> greater than unity – that is, the difference between the samples has to be interpreted as more likely had they come from the same speaker – and there are also plenty of same-speaker comparisons with LR < 1. This is the way speakers and their formants behave.’ (2006: 5). This observation, with which we unreservedly agree, argues strongly against the restriction of analytic focus to the small subset of features for which reference data are available.

Further difficulties arise when we consider how to delimit the relevant population for comparison. It is commonly assumed that the population should be controlled for aspects of the speaker’s regional and social background, since these factors may significantly affect the patterning of linguistic and phonetic features relevant to a case. We further note that population data must also be controlled for a wide range of other factors which are known to have significant effects on aspects of speech, voice and/or language. These include environmental effects of the recording situation (e.g. use of telephone, Lombard speech, transmission medium, recording hardware, distance from and orientation to the microphone) and short-term effects on the speaker resulting from e.g. smoking, intoxicants, or health problems. The range of potential factors to be controlled for is in fact very large indeed.

It should also be borne in mind that reference data are likely to have a limited ‘shelf-life’ in terms of their suitability. All languages and dialects are in a constant state of flux, and rates of change are unpredictable. Comparing forensic samples against outdated reference material thus runs the risk that key diagnostic features for the given case might not be adequately represented in the reference corpus. A recent study by Loakes (2006), for instance, attempted a likelihood ratio analysis on data from eight male twins, Australian English speakers from Melbourne recorded in 2002–3. As the reference population she used the Bernard (1967, 1970) dataset, consisting of 170 adult males from New South Wales. Loakes found the results to be ‘severely restricted’ (2006: 237). For most of the variables examined, the 8 twins’ data fell outside the range covered by all of the reference speakers due to sound change that had operated in Australia in the intervening years.

Self-evidently, dialect variation, in addition to historical change, limits the applicability of any one set of linguistic-phonetic reference data. We do not, incidentally, regard dialect diversity as a peculiarly British phenomenon. Although it is true that New World Englishes tend to be more homogeneous than those of the British Isles, there is nonetheless considerable variation in other varieties of English deriving not only from geographical separation but also, increasingly, from ethnic background (e.g., for Australian English see Clyne, Eisikovits and Tollfree 2001, Cox 1999, Cox and Palethorpe 2001, 2006).

### Qualitative approach as short-term solution

As an alternative to assembling and analysing reference corpora to meet the needs of each case, R&M (2009) propose the use of qualitative estimates of the strength of evidence, e.g., 'From my experience I think you would be much more likely to get the differences I have listed between the offender and suspect speech samples assuming that they had come from the same speaker, rather than different speakers' (pp 158–159). However, these are suggested as a purely short-term solution for the UK context and are described as a 'poor substitute' for quantitative LRs (p 59). Whilst we have no issue with the framing of impressionistic conclusions in such terms (see below), we do not regard qualitative expressions purely as a stopgap measure to be adopted pending the introduction of fully quantitative LRs. For the reasons outlined in the previous section, we cannot envisage the compilation of appropriate reference population statistics for anything approaching the full range of features relevant to any speaker comparison case as an achievable end, even in the very long term. This applies not only to cases arising in the UK, but universally.

### Conclusions

We conclude with a summary reiteration of the points developed in this rejoinder.

First, given the continuing absence of population statistics, the proponents of a quantitative LR approach have proposed attempting to assemble background population statistics on a case-by-case basis. But in order for this to be practicable, only a very limited range of features can be taken into account and quantified. This, in our view, could amount to an irresponsible neglect of the richness and complexity of spoken communication and the analytic possibilities they allow. By ignoring the vast majority of features available for analysis the expert runs the very real risk of producing an opinion that could lead to a miscarriage of justice.

Second, we are of the view that it is unrealistic to see it as merely a matter of time and research before a rigorously and exclusively quantitative LR approach

can be regarded as feasible, let alone reliable, within the auditory-acoustic forensic speaker comparison methods predominantly in use at present. The multi-faceted nature of speech communication makes it highly implausible that relevant population statistics will ever be available for the majority of features that present themselves in a particular case. This is not to say that we would exclude the possibility of incorporating quantitative LRs into a subset of the analysis in forensic casework. We regard research in this field as essential for the development of forensic speech science. We recognise the ongoing work of R&M as making a very valuable contribution, and three of the five present authors are also engaged in research of this nature.<sup>1</sup>

Finally, whilst we take issue with R&M on the above points, we do not see the UK 2007 framework as set in stone. We are in agreement with R&M that it constitutes a step towards the adoption of Bayesian concepts, and we maintain that it is broadly conceptually compatible with a Bayesian LR approach. We are, furthermore, fully open to the possibility that as our understanding of speaker characterisation progresses and our techniques advance, the framework will need to be modified. Indeed, we can envisage the adoption of a yet more explicitly Bayesian framework for the statement of conclusions, involving an overall qualitative LR along the lines recommended by R&M. We appreciate the contribution made by R&M to this ongoing discussion, and acknowledge the contribution that their critique makes in furthering the debate. For the present, however, we feel that the 2007 framework provides a sound practical framework for the expression of conclusions arrived at through auditory-acoustic phonetic comparison.

### **About the authors**

Peter French is Chairman of JP French Associates Forensic Speech and Acoustics Laboratory and Honorary Professor in the Department of Language Science at the University of York where he undertakes and supervises research and teaches postgraduate courses in Forensic Speech science. He is President of IAFPA and a founding editor of the International Journal of Speech, Language and the Law. Over the past twenty-five years he has been involved in around five thousand forensic cases involving sound, speech and language.

Francis Nolan is Professor of Phonetics in the Department of Linguistics at the University of Cambridge. His research interests range over phonetic theory, intonation, connected speech processes and speaker characteristics. His long-standing involvement in forensic phonetics covers both fundamental research and casework, including speaker comparison and the use of voice parades. He believes that forensic phonetic practice needs to be underpinned by advances in phonetic theory. He is a founding member of IAFPA.

Paul Foulkes has held academic posts at the Universities of Leeds and Newcastle and is currently Reader in the Department of Language and Linguistic Science at the University of York where he directs the MSc degree in Forensic Speech Science. A former editor of the *International Journal of Speech, Language and the Law*, he undertakes research in forensic phonetics in addition to his continuing work on language variation and change. He is a consultant with JP French Associates, in which role he carries out forensic casework.

Philip Harrison is a forensic consultant and Director of JP French Associates and has worked on over 1000 cases in the areas of speaker comparison, disputed utterance analysis, transcription, authentication and enhancement as well as many miscellaneous cases. He has an undergraduate degree in Acoustical Engineering from the Institute of Sound and Vibration Research (ISVR) at the University of Southampton and a postgraduate degree in Phonetics and Phonology at the University of York. He is currently conducting doctoral research on the measurement and analysis of formants, and frequently lectures on forensic speech science to universities and other organisations in the UK and abroad.

Kirsty McDougall is a British Academy Postdoctoral Fellow in the Department of Linguistics, University of Cambridge, having previously worked as a Research Associate on the forensic phonetic projects DyViS and VoiceSim. She has a B.A. in linguistics and a B.Sc. in mathematics and statistics from the University of Melbourne, and an M.Phil. and Ph.D. in linguistics from the University of Cambridge. Her research interests include speaker characteristics, theories of speech production, and phonetic realisation of varieties of English. She is a member of IAFPA.

## Notes

- 1 Bayesian Biometrics for Forensics, funded by the Marie Curie FP7 Initial Training Network, 2010–2014.

## References

- Bernard, J. R. (1967) Some measurements of some sounds of Australian English. Ph.D. Dissertation, University of Sydney.
- Bernard, J. R. L. (1970) Toward the acoustic specification of Australian English. *Zeitschrift für Phonetik* 23: 113–128.
- Clyne, M. Eisikovits, E. and Tollfree, L.F. (2001) Ethnic varieties of Australian English. In D. Blair and P. Collins (eds) *English in Australia* (pp. 223–238). Amsterdam: John Benjamins.
- Cox, F. M. (1999) Vowel change in Australian English. *Phonetica* 56: 1–27.
- Cox, F. and Palethorpe, S. (2001) The changing face of Australian English vowels. In D. Blair and P. Collins (eds) *English in Australia* (pp. 17–44). Amsterdam: John Benjamins.

- Cox, F. M. and S. Palethorpe (2006) Some phonetic characteristics of Lebanese Australian English. Paper presented at the Australian Linguistic Society Conference. Brisbane, 7–9 July.
- French, J. P. and Harrison, P. (2007) Position statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases. *International Journal of Speech Language and the Law* 14: 137–144.
- Laver, J. (1980) *The Phonetic Description of Voice Quality*. Cambridge: Cambridge University Press.
- Laver, J. (1994) *Principles of Phonetics*. Cambridge: Cambridge University Press.
- Loakes, D. (2006) A Forensic Phonetic Investigation into the Speech Patterns of Identical and Non-Identical Twins. Ph.D. Dissertation, University of Melbourne.
- Robertson, J. (2007) Forensic Speech Science from a Police Perspective. Paper presented at the Australian Research Council Network in Human Communication Science Workshop: FSI not CSI: Perspectives in State-of-the-Art Forensic Speaker Recognition, Sydney.
- Rose, P. (2006) Accounting for Correlation in Linguistic-Acoustic Likelihood Ratio-based Forensic Speaker Discrimination, *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey* pp.1–8, 28–30 June 2006.
- Rose, P. (2007) Going and getting it – Forensic speaker recognition from the perspective of a traditional practitioner-researcher. Paper presented at the Australian Research Council Network in Human Communication Science Workshop: FSI not CSI – Perspectives in State-of-the-Art Forensic Speaker Recognition, Sydney. <http://forensic-voice-comparison.net/documents>.
- Rose, P., & Morrison, G. (2009). A response to the UK Position Statement on forensic speaker comparison. *International Journal of Speech Language and the Law* 16(1): 139–163.