Issues and opportunities for the application of the numerical likelihood ratio framework to forensic speaker comparison

Erica Gold, Vincent Hughes

Department of Language and Linguistic Science, University of York, Heslington, York, YO10 5DD, United Kingdom

erica.gold@york.ac.uk, vh503@york.ac.uk

Abstract

Across forensic speech science, the likelihood ratio is increasingly becoming accepted as the "logically and legally correct" [1] framework for the expression of expert conclusions. However, there remain a number of theoretical and practical shortcomings in the procedures applied for computing LRs based on speech evidence. In this paper we review how the LR is currently applied to speaker comparison evidence and outline three specific areas which deserve further investigation; namely statistical modelling, issues relating to the *relevant population* [2] and the combination of LRs from correlated parameters. We then consider future directions for confronting these issues and discuss the implications for forensic comparison evidence more generally.

Keywords

Likelihood ratio; Phonetics; Forensic speech science; Modelling; Relevant population; Correlations

1. Introduction

This paper provides an overview of the fundamental theoretical principles and practical procedures involved in the application of the numerical likelihood ratio (LR) to forensic speaker comparison (FSC) evidence. More importantly, this paper considers three specific areas for concern with current LR practises as applied to FSC and explores potential directions for future work. These issues relate to statistical modelling, the definition of the *relevant population* and collection of reference data, and the combination of correlated parameters into an overall LR (OLR). We believe that these concerns also have broader implications for the application of the numerical LR across all forensic disciplines involving comparison evidence.

1.1 Auditory-acoustic forensic speaker comparison

Forensic speaker comparison (FSC) typically involves a disputed recording of an unknown offender (e.g. threatening phone call, recording of an assault) and a known sample of the suspect (commonly a police interview). The expert is instructed to conduct an assessment and comparison of the speech patterns in both samples. On the basis of the expert testimony the trier of fact is responsible for assessing whether the disputed and known samples contain the voice of the same or different individuals. This decision informs the *ultimate issue* [3] of the innocence or guilt of the defendant. In the UK, FSC accounts for the majority of casework undertaken by forensic speech scientists (c. 70%, French p.c.).

FSC evidence is analysed using a range of different procedures (for an overview see Gold and French [4]). Of the FSC evidence currently admitted in the Courts, a large proportion (in Europe [4]) involves the analysis of traditional linguistic-phonetic parameters by forensic phoneticians (FPs) using auditory-phonetic plus acoustic-phonetic techniques (AuPA + AcPA [4]). AuPA + AcPA is a combined approach based on detailed analytical listening and quantification using acoustic analysis. Commonly, FPs would consider a range of segmental (incl. production of individual consonants and vowels), suprasegmental (incl. intonation, pitch, rhythm, speech rate), pathological (incl. stuttering etc.) higher-order linguistic (incl. lexical choice, syntax) and non-linguistic (incl. clicks, hesitation phenomena) parameters in a FSC case.

1.2 Expression of conclusions

Currently, FPs implement a variety of conclusion frameworks as laid down by laws, regulations and employers in different jurisdictions. The majority of experts (38% of 34 practitioners in [4]) use some form of hierarchical classical probability scale (CPS). CPSs involve an expression of the probability of the recordings containing the same or different speaker(s) given the combined weight of the linguistic parameters analysed (below).

Table 1

Typical classical probability scale using for expressing conclusions in FSC cases (from [5, 6])

POSITIVE IDENTIFICATION	NEGATIVE IDENTIFICATION
sure beyond reasonable doubt	*probable*
there can be very little doubt	*quite probable*
highly likely	*likely*
likely	*highly likely*
very probable	
probable	
quite possible	
possible	
… that they are the same person	… that they are different people

Consistent with developments in other forensic disciplines, arguments have been made for the implementation of a more scientifically grounded conclusion framework in FSC [5, 7] which avoids the expression of "inappropriate" [8] conclusions made on the basis of information beyond that of the evidential recordings. The move away from expressions of posterior probability has been termed the "paradigm shift" [9] and relates primarily to the adoption of the Bayesian likelihood ratio (LR).

2. Background

2.1 The Likelihood Ratio (LR)

General descriptions of the LR and the use of Bayes' theorem for assessing uncertainty in the Courtroom can be found in Aitken and Stoney [10], Robertson and Vignaux [11] and Aitken and Taroni [2]. The validity of the LR in terms of its application to forensic comparison evidence can be found in responses to the England and Wales Appeal Court ruling in *R v T* [12] (Evett et al. [13], Morrison [14], Berger et al. [15]). The conceptual validity of the LR for the expression of expert FSC conclusions has also been outlined by Champod and Meuwley [16], Rose [17] and Morrison [8].

The LR provides a gradient estimation of the strength of the evidence based on the ratio of probability (*p*) of the evidence (E) given (|) the prosecution hypothesis (H_p) and the probability of evidence given the defence hypothesis (H_d). The odds form of the LR is defined in Eq.(1):

$$\frac{p(\mathbf{E}|\mathbf{H_p})}{p(\mathbf{E}|\mathbf{H_d})} \tag{1}$$

Applied to FSC evidence, the likelihood ratio (LR) consists of an assessment of the similarity between the KS and DS with regard to a given parameter and the typicality of those values [17, 18] within the wider, *relevant population* [2]. Quantification of the typicality of between- and within-speaker variation is conducted based on models generated from a sample of the *relevant population*. Such a sample is termed the ‘reference data’.

The outcome is a value centred on one, such that LRs of greater than one offer support for H_p, and LRs of less than one offer support for H_d [19]. The magnitude of the LR determines how much more likely the evidence would be given H_x than H_y [20]. A LR of five, for instance, should be interpreted as the evidence being five times more likely assuming the hypothesis that the recordings contain the voice of the same speaker than assuming the hypothesis that different-speakers were involved.

2.2 Application of the LR to AuPA + AcPA FSC

There has been a relatively strong representation of the LR in FSC, with roughly 20% of experts using either a numerical or verbal LR [4]. Development of the LR framework in the field of FP is largely thanks to a small community of researchers. The work in this area has predominantly focused on two main areas: assessing speaker discrimination using a numerical LR and the overall improvement in methodologies for the computation of LRs.

The application of the LR framework to AuPA + AcPA FSC has focused almost exclusively on vowels. In terms of their speaker-discriminatory potential, mid-point approaches [21, 22, 23, 24, 25], formant trajectories [26, 27, 28, 29] and long-term formant distributions [30, 31] have all been analysed. Comparatively, the same attention has not been given to consonants and additional parameters, although a limited set of studies has considered non-vocalic parameters [32, 33, 34, 35, 36, 37, 38]. Such studies highlight that considerably more research is needed into the speaker discriminatory value of consonants and non-segmental parameters using a numerical LR framework.

Methodological advances for the computation of LRs have also been made. In particular studies have considered the appropriateness and performance of different statistical modelling techniques [25, 35, 39, 40]. Research has also focused on issues of correlation [41, 42, 43, 44, 45], delimitation of the *relevant population* [46, 47, 48], the amount of reference data needed [49, 46, 50], system calibration [51, 44, 52] as well as measures of validity and reliability [44, 52].

2.3 Complexity of speech evidence

However, a substantial issue for the application of the LR to FSC, and one which has been consistently overlooked in the literature, is the complexity of naturally occurring spontaneous speech as evidence. Speech is inherently variable, such that no two utterances produced by the same individual are identical. Such *intra*-speaker variability separates speech from other forms of forensic evidence (such as DNA) in that $p(E|H_{ss})$ can never be 1. Variability in the speech of the same individual can be caused by numerous extraneous factors (incl. time of day, emotional state, interlocutor and topic) and the maximal extent of potential intra-speaker variability is not well understood by current models of phonetics, phonology or sociolinguistics.

Variation between speakers is also highly conditional on a number of factors. *Inter*-speaker variation is determined by biological and anatomical factors (vocal tract length accounts in part for fundamental frequency), as well as systemic-phonological (in the South of England the words *bath* and *trap* are produced with different vowels, whereas in the North both are [a]) and social factors (incl. regional background, age, sex, socio-economic background, ethnicity) [54, 55, 56]. Such grouping variables interact with each other and affect different linguistic-phonetic parameters in different ways.

Given potentially high within-speaker variation and the numerous grouping factors which determine the distribution of between-speaker variation, it is unsurprising that many traditional linguistic-phonetic parameters offer relatively low speaker-discriminatory potential. As such FPs generally prefer a componential approach to AuPA + AcPA FSC. However, of the potentially numerous parameters a FP may analyse in a given case many form highly correlated sub-systems due to physiological, phonological and social factors. The correlation structure of the data must be appropriately accounted for in order for the resulting OLR to be a meaningful representation of the strength of evidence.

Further, unlike other forms of forensic evidence the range of parameters analysed in a componential AuPA + AcPA approach means that FPs deal with both discrete (frequency counts primarily for consonantal and higher-order parameters) and continuous (incl. formant frequencies and f0 (pitch)) data. Although much of the continuous data are assumed to be normally distributed, non-normally distributed data are also common in FSC. In addition, values within-speakers may display a different distribution to those between-speakers even for the same parameter.

Finally, forensic speech samples are an inherently complex form of linguistic data. Under forensic conditions the recording/transmission technology of speech and the conditions under which speech is obtained are often compromised. Incriminating samples are increasingly recorded via mobile telephone. Both telephone bandwidth restrictions and mobile phone codecs have been shown to artificially attenuate the speech signal (particularly affecting formant frequencies: [57, 58, 59, 60]), which is not present in direct, high quality recordings made during police interviews. Low signal-to-noise ratio in criminal recordings, caused by high levels of background noise or overlapping speech, is also problematic for analytical purposes. Further, there is typically a mismatch in the conditions in which the criminal and suspect speech samples are elicited. Criminal samples are often made in situations of elevated emotional states (e.g. arguments), considerable physical activity (e.g. during an assault), or intoxication. These samples are then compared to suspect recordings that have been made in very different conditions (i.e. a different setting, with a different interlocutor, sober).

3. Issues

Despite the considerable progress made over the last 12 years, the complexities of speech as a form of comparison evidence have largely been overlooked or oversimplified. As such there remain issues with methodological procedures. In particular we highlight three primary areas for development: alternative modelling techniques, defining the *relevant population* and collecting reference data, and accounting for correlations between parameters.

Whilst these issues have previously been considered to varying extents, the complexity of the problems has often been overlooked and the solutions offered are not without fault. In order to ensure a valid and reliable estimation of the strength of evidence (in line with Daubert [61]), it is important that further attention is directed towards more appropriate and ultimately successful solutions. In the following sections we assess each issue in turn considering the evolution of current procedures, their shortcomings and potential directions for future research.

3.1 Statistical modelling

When calculating a LR the conceptual framework remains consistent, but there are a number of mathematical procedures that can be used to arrive at a numerical value. In FP the different procedures used are selected in relation to the characteristics of the data distributions. Aitken and Taroni [2] state that "for any particular type of evidence the distribution of the characteristic [parameter] is important. This is so that it may be possible to determine the rarity or otherwise of any particular observation." Therefore, it is important to use the model that best fits the distribution of the data in the calculation of LRs in order to represent the strength of evidence as accurately as possible.

As outlined in §2.3, AuPA + AcPA FSC is a particularly complicated form of expert evidence which deals with both continuous and discrete data in various different forms. Clearly for FP, not all parameters behave in the same manner. That is to say that not all parameters are distributed in the same way when it comes to capturing patterns in the population. For example, one parameter may have a higher degree of occasion-to-occasion variation while another parameter is much more consistent in its variation. For this reason, selecting the best fit model for data is a vital part in any LR calculation.

At this point in time, there have been a limited number of models used in the calculation of LRs in FP and these have primarily been restricted to continuous parameters. The most commonly employed models are Lindley [62], multivariate kernel density (MVKD; Aitken and Lucy [63]), and the Gaussian mixture model – universal background model (GMM-UBM; Reynolds et al. [64]) procedures. The Lindley model is used for calculating LRs with univariate data and assumes that values both within- and between-speakers are normally distributed. Much of the early LR literature used Lindley to compute strength of evidence in order to assess the discriminatory potential [21, 27, 65]. The appropriateness of Lindley for modelling traditional acoustic-phonetic data (primarily vowel formants) has also been assessed directly in [66, 67].

The MVKD model proposed by Aitken and Lucy [63] is used in the calculation of LRs for multivariate data. It is able to account for variance within a group as well as between groups and assumes normality for variation within a group; however, the estimation of variance for between groups uses a kernel density model [39]. The MVKD approach is the formula which has been applied most often in the FP LR literature, primarily due to the fact that linguistic data is commonly multivariate (i.e. parameters consist of multiple elements). MVKD has almost exclusively been used to compute LRs using vowel formant data [26, 28], but has also been applied to F0 parameters [36] and articulation rate [38].

The mathematical procedure for MVKD bases its estimates of both the within- and between-group (group = speaker) variance on distributions in the population database (reference database), which includes sets of measurements from the *relevant population*. Group means are used in the MVKD model and “the between group distribution is modelled via a summation of a set of equally-weighted kernels … Each kernel is a Gaussian whose covariance matrix is a scaled version of the pooled within-group covariance matrix” (Morrison [39]).

The third model (GMM-UBM) is commonly applied in ASR (i.e. [67, 69]) but has also been used to calculate LRs in traditional FSC [30, 24, 31]. GMM-UBM, like MVKD, can accommodate multivariate data; however, GMM-UBM models the data differently to MVKD. GMM-UBM utilises GMMs to characterise distributions instead of kernel densities (i.e. MVKD). The most significant difference between GMM-UBM and MVKD is that the background model for GMM-UBM is person independent and compared against a model of person-specific parameter characteristics when comparing same (SS) and different (DS) speaker pairs [64]. Morrison [39] further explains that:

> The UBM is trained using the expectation–maximisation (EM) algorithm. Rather than building the suspect model from scratch using the often limited amount of known-voice data, it is created by copying the UBM then using a maximum a posteriori (MAP) procedure to adapt its weights, mean vectors, and covariance matrices towards a better fit to the known-voice data (in the most widely used variant of the procedure only the means are adapted).

Previous research [39, 70] has shown GMM-UBM to perform both better and worse than the MVKD model, and success is dependent on trialling the model on data sets and comparing results.

Selecting an appropriate model requires careful examination of the data distributions for a given parameter. Although the MVKD is currently selected most frequently for use in LR calculations, it has been found that the MVKD model is not sufficient for capturing distributions accurately for all parameters, namely some of those which are discrete. In response to this problem, new models have been proposed for clicks (a discrete parameter in FP) by Aitken and Gold [40]. The continued development of such models allows for a more thorough implementation of all types of speech data, because at present not all parameters can be put into the LR framework.

3.2 Defining the *relevant population* and collecting reference data

Theoretically the *relevant population* in a given case is defined by the defence hypothesis (H_d). However, a significant problem is that in many jurisdictions the defence may offer a non-specific alternative hypothesis, or no alternative at all. In such cases, the expert requires a default defence hypothesis which is more specific than 'it was someone else in the population'. Similarly, in LR-based research it is necessary to have an underlying (if not explicit) default definition of the *relevant population* in order to assess typicality.

The concept of *logical relevance* (Kaye [71, 72]) is an approach which has previously been adopted in FSC research. It refers to broad grouping variables which may affect the frequency of the given parameter(s) under analysis in the wider population. Following *logical* *relevance*, Rose claims that the *relevant population* should be controlled for language (broadly defined as regional background) and sex [19] (i.e. 'it wasn't the suspect, it was another male speaker of Australian English), and that within-speaker variation should be modelled using non-contemporaneous samples from speakers in the reference data [73, 74, 75]. This default H_d has been reflected in the majority of LR-based studies using traditional acoustic-phonetic parameters [32, 21, 42, 66]

However, the Rose [19] default fails to capture the considerable multidimensionality and complexity of within- and between-speaker variation (§2.3). Despite Loakes' [76] claim that "tighter constraints on social variables might also need to be applied to population selection", such additional potentially *logically relevant* factors have consistently been overlooked by the community of FVC researchers working within the numerical LR framework. Further, given the numerous factors known to affect intra-speaker variability any model of within-speaker variation based solely on non-contemporaneity will necessarily be underoptimistic relative to the facts of the case at trial.

A conceptual problem with *logical relevance* is the paradox that without knowing who the offender is, it is not possible to be certain what the population is of which he is a member (in terms of regional background, sex, age, etc.). As an alternative, Morrison et al. [47] propose that the reference data should consist of similar sounding speakers to the offender as judged by lay listeners, since the evidential samples are submitted for expert analysis by a lay listener (usually a police officer) on the basis of perceived similarity across the recordings. As such, the *relevant population* is defined by a single subjective grouping variable ('similarity').

The primary issue with this approach is the assumption that speaker similarity should be judged by a panel of lay listeners. Perceptual dialectology [77] and ear witness reliability research (Bull and Clifford [78, 79], Watt [80]) suggests listeners' own linguistic backgrounds (amongst other things) affect judgments of speaker similarity. Therefore Morrison et al.'s [47] claim that the panel of lay listeners should consist of "police officer(s) (or other appropriate listener(s))" fails to account for the numerous factors which are known to affect listeners' similarity judgments. Studies [80, 81] also reveal that even listeners from the same speech community are sensitive to different elements of the acoustic signal and as such are often linguistically erratic when assessing similarity. As such the *relevant* *population* is likely to be inconsistent with regard to demographic features such as sex and age; an assumption which is unsustainable if the *relevant population* is to be consistent across all forms of forensic evidence in a given case.

A consequence of the linguistically erratic nature of lay listeners' similarity judgements is also that the resulting set of reference data is not replicable and the decisions upon which an assessment of sufficient similiarity are not transparent for the trier-of-fact. Finally, in determining which samples to present to the panel of lay listeners the expert necessarily makes subjective, categorical decisions over the demographic make-up of the *relevant population* even if such controls are only as broad as language or general regional background [47]. The motivation for controlling certain factors and ignoring others before presenting the samples to the panel is unclear.

Finally, there are a number of practical issues related to the collection of reference data from the *relevant population*. The first relates to the lack of available reference data from which to quantify typicality [82]. Broadly there are two alternatives to this problem: using tailored reference data collected on a case-by-case basis [22] or use existing 'off-the-shelf' databases (large (non-)forensic corpora). However, there will inevitably be some mismatch between the reference data and the facts of the case at trial with regard to social and stylistic factors. The extent to which such mismatch affects the resulting LR has received relatively little attention in the literature (other than Hughes and Foulkes [46], Hughes [83]).

The second issue is the amount of reference data needed. A small number of FSC studies have addressed the effect of small samples, with both Ishihara and Kinoshita [49] and Hughes and Foulkes [46] finding that LRs are generally unstable and misrepresentative with small amounts of reference data. Aside from Hughes [83] and Rose [84], relatively little attention has been directed towards addressing how large the set of reference data needs to be in order to achieve a precise estimate of strength of evidence.

3.3 Correlations

According to naïve Bayes [85], numerical LRs from separate biometric parameters may be combined using simple multiplication provided there is an assumption of mutual independence. This is because unless parameters are truly independent there is a risk of overestimating the strength of evidence by measuring the same parameter multiple times. However, as outlined above a significant problem for AuPA + AcPA FSC is the multitude of potential correlations both within- and between-parameters.

The issue of how best to account for correlations has received considerable attention in the FSC LR literature. In early LR-based studies using traditional linguistic-phonetic parameters, the problem of applying naïve Bayes assumptions of independence to multivariate strength of evidence was often acknowledged, but ultimately ignored. Kinoshita [32] used naïve Bayes to combine LRs based on the best performing set of formant predictors from a set of short vowels, /m/, and /ʃ/ into a single expression of posterior probability. Similarly, Alderman [21] generated an overall LR (OLR) from different vowel formant predictors using naïve Bayes in order to compare the speaker-discriminatory performance of different combinations of parameters (and individual features of parameters).

Rose, Osanai and Kinoshita [65] display a more overt awareness of the issues surrounding correlation within- and between-parameters. In their study of the discriminatory performance of formants compared with segmental cepstra from /ɔː ɕ ɴ/, linear regression was applied to assess the degree of correlation between individual formants and individual cepstral coefficients extracted from a single phoneme. Whilst LRs were combined into an OLR using an assumption of independence, between-parameter correlation was not explicitly tested because “it was assumed that, given the very different phonetic nature of the three segments used, there was unlikely to be much correlation between all but their highest formants” [65].

The development of LR modelling techniques has brought with it the capability to deal with more appropriately with the complexities of correlation. Aitken and Lucy’s [63] MVKD formula treats the set of data upon which LRs are computed as multivariate data and as such is able to account for within-segment correlation. Rose et al. [41] investigated the comparative performance of the multivariate LR approach and the naïve Bayes assumption of independence. The naïve Bayes approach was shown to overestimate the strength of SS and DS LRs compared with the more conservative MVKD model. Performance, in terms of the proportion of errors, was also better assuming independence, leading Rose et al. to conclude that “the ‘correct’ formula is still not exploiting all the discriminability in the speech data (and as such) the Idiot's approach [naïve Bayes] is still preferable" [41], irrespective of the fact that it produces misrepresentative estimates of the strength of evidence.

Most recently logistic regression fusion has been adopted as the primary means of accounting for potential correlation between linguistic-phonetic parameters in LR-based FSC. Fusion is a form of "back-end processing" [24] which attaches weights to parameters based on correlations between LRs from individual parameters, rather than considering correlations in the data. Fusion was developed within the field of automatic speaker recognition (ASR) (Brümmer et al. [86], Gonzalez-Rodriguez et al. [87], Ramos Castro [88]) and has since been applied in a number of studies using traditional phonetic parameters [23, 28, 44, 75] leading Rose and Winter [24] to claim that fusion is one of the "main advances" to have emerged from automatic methods.

Fusion is currently the only alternative to a naïve Bayes approach for LR-based FP analysis. However, there are also a number of potential problems with fusion. Firstly, back-end processing, as the name suggests, deals with correlations after the generation of numerical LRs. Therefore, as suggested by Rose [23] "it is … possible … that two segments which are not correlated by virtue of their internal structure and which therefore should be naively combined, nevertheless have LRs which do correlate". Equally the reverse is also possible, whereby correlated parameters generate non-correlated LRs. More broadly, there is also an issue of efficiency. Since fusion is implemented after the generation of LRs, the original analysis may unnecessarily include a number of highly correlated parameters which offer limited combined strength of evidence.

4. Discussion

The aim of this paper has not been to critique the work that has been previously carried out, but to raise awareness to those areas in LR methods that would benefit from further research. In this section we consider directions for potential alternative to current procedures and directions for future work.

4.1 Statistical modelling

In respect of modelling, it is of our opinion that research in this area would benefit from closer work with forensic statisticians. This would allow for informed decisions to be made at different stages of the LR modelling process. Such collaborative work should primarily focus on developing models for parameters which are not currently considered [as in 40] within the LR framework. Given that the combination of multiple parameters in FSC is preferred [4], it is important to include as many possible parameters to achieve a combined estimation of the strength of evidence.

This is considered preferable to restricting FSC to those parameters which can conveniently be modelled using existing procedures. Through a collaborative effort it is hoped that models can be created to fit the distributions of specific phonetic and linguistic parameters rather than shoe-horning FP parameters into existing models. Finally, it is also important to consider the limitations and strengths of the models currently being used in FSC through further empirical testing and analysis.

4.2 Defining the *relevant population*

We are broadly in agreement with Morrison et al. [47] that the *relevant population* should consist of similar-sounding speakers to the offender (Brümmer and de Villers [89]). However, we are of the opinion that lay-listeners should not judge similarity, but rather that similarity is judged by a linguistically informed expert. By allowing the expert to make informed judgements of similarity the outcome will be a set of reference data which is coherent in terms of sociolinguistically, *logically relevant* grouping variables.

The overarching message is that whilst social and stylistic variability makes speech an unusually complex forensic medium, it should not be overlooked or ignored. Rather structured variation offers an invaluable resource for assessing typicality and, unlike other forms of forensic evidence, is available to the expert directly from the speech signal itself. *Speaker similarity* judged by an expert is also consistent with the underlying assumption of a single *relevant population* across all expert evidence presented to the Court and is in line with current procedures in ASR and other forensic disciplines (e.g. DNA).

Future research should concentrate on how a nuanced view of the social and stylistic structure of within- and between-speaker variation may provide a more meaningful estimation of the strength of evidence. Developing from Hughes and Foulkes [46] it is essential that there is more testing of the extent to which social and stylistic factors are *logically relevant* for given parameters in narrowly defined speech communities. Future work should also develop on procedures outlined in Morrison, Rose and Zhang [48] towards compiling forensically realistic, openly available databases for extracting reference data and quantifying typicality. This is particularly important in the British English context, since population statistics are currently only available for a small number of parameters (f0 [90], articulation rate [38], click rate [91]) in a limited number of speech communities (primarily young male Standard Southern British English (SSBE) speakers).

4.3 Correlations

Given the potential limitations of fusion as highlighted by Rose and Winter [24], it is first and foremost important to empirically test whether back-end processing captures the same correlations as those in the original linguistic/phonetic data. The results of such empirical tests should be used to determine whether fusion it is an appropriate solution to the combination of correlated parameters in FSC. If it is the case that correlations found in the data are comparable to those that exist between LRs, we nonetheless believe that it remains preferable to develop a form of ‘front-end processing’ to account for correlations. As such, there would be a conscious awareness of correlations between the parameters under analysis, prior to the extraction of any numerical data. An alternative avenue to fusion could lie with the development of graphical models or a complete Bayesian network for speech evidence (Aitken and Taroni [2], Taroni et al. [92]). A Bayesian network would aim to create a ‘front-end’ mathematical model of interdependencies between speech parameters in order to appropriately combine parameters.

A model of front-end processing would need to consider three factors. Firstly, predicted correlations made on the basis of established linguistic and phonetic theory. Secondly, there should be quantification of overall between-parameter, between-speaker correlations within homogeneous communities of speakers. Thirdly, a more fine-grained assessment of relationships between parameters must also be considered which assesses the individual rather than the group. This is because in real casework the magnitude of correlations between parameters in the criminal sample and the suspect sample must be determined on the basis of prior testing. If the correlations within individuals are unknown, this could lead to misguided combinations of parameters, and a misrepresentative estimation of the strength of evidence. Crucially the issue of correlation applies both to experts working in a LR framework, who must account for naïve Bayes, as well as those working in other frameworks where the expert personally selects parameters to consider and combine in casework.

5. Conclusion

It is our view that given the current state of the field, work on the three issues highlighted in this paper is central to the development of the numerical LR framework as applied to FSC. However, that is not to say that future development should be isolated to only these issues. Rather, there remain both theoretical and practical difficulties involved with the application of the LR to speech evidence. With the growth of LR work it is our responsibility that existing procedures are continually reviewed and improved. Even if acceptable performance is achieved there will always be room for improvement.

Whilst we have outlined specific problems and areas for future growth, it is essential that all future work continues to acknowledge and consider the importance of the linguistic and phonetic principles underlying the data. It is our view that when carrying out research under an LR framework it is vital that methodological decisions are informed by basic linguistic knowledge. By building on the work already carried out, acknowledging limitations of current procedures, and seeking linguistically and phonetically informed solutions we are able to aim for a more transparent and flexible LR framework. This will improve the extent to which we are able to capture the true complexity of linguistic data and ultimately provide more meaningful and accurate estimations of the strength of evidence.

Finally, such issues are not exclusively restricted to FSC and are readily discussed in other forensic disciplines [93, 94, 95]. Whilst the complexity of speech as evidence causes considerable difficulties for the application of the LR framework to FSC, it also means that speech is the ideal testing ground for the development of procedures which will be of value to other forensic disciplines. In this way the complexity of speech offers a unique opportunity for speech to be at the fore front of the “paradigm shift” towards the application of the LR to all forms of forensic evidence.

Acknowledgements

This research has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 238803 and the Economic Social Research Council (ESRC). Considerable thanks to Peter French, Paul Foulkes and Michael Jessen for comments on previous drafts of this paper. We are also grateful to Colin Aitken and Tereza Neocleous for their input.

References

[1] P. Rose, G. S. Morrison, A response to the UK position statement on forensic speaker comparison, International Journal of Speech, Language and the Law 16 (2009), 139-163.

[2] C. G. G. Aitken, F. Taroni, Statistics and the evaluation of evidence for forensic scientists, 2nd ed, Wiley, Chichester, UK, 2004.

[3] M. Lynch, R. McNally, ‘Science’, ‘common sense’ and DNA evidence: a legal controversy about the public understanding of science, Public Understanding of Science 12 (2003), 83-103.

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

[5] A. P. A. Broeders, Some observations on the use of probability scales in forensic identification, Forensic Linguistics 6 (1999), 228-241.

[6] J. Baldwin, J. P. French, Forensic phonetics, Pinter, London, UK, 1990.

[7] C. Champod, I. W. Evett, Commentary on A.P.A. Broeders (1999) 'Some observations on the use of probability scales in forensic identification', Forensic Linguistics 7 (2000), 238-243.

[8] G. S. Morrison, The place of forensic voice comparison in the ongoing paradigm shift. Written version of an invited presentation at the 2nd International Conference on Evidence Law and Forensic Science, 25-26 July (2009). Beijing, China, 1-16.

[9] M. J. Saks, J. J. Koehler, The coming paradigm shift in forensic identification science, Science 309 (2005), 892-895.

[10] C. G. G. Aitken, D. A. Stoney, The use of statistics in forensic science, Ellis Horwood, London, UK, 1991.

[11] B. Robertson, G. A. Vignaux, Interpreting evidence: evaluating forensic science in the courtroom, Wiley, Chichester, UK, 1995.

[12] R v T [2010] EWCA Crim 2439.

[13] I. W. Evett et al. Expressing evaluative opinions: A position statement, Science & Justice 51 (2011), 1–2.

[14] G. S. Morrison, The likelihood-ratio framework and forensic evidence in cout: a response to R v T, International Journal of Evidence and Proof 16 (2012), 1-29.

[15] C. E. H. Berger, J. Buckleton, C. Champod, I. W. Evett, G. Jackson, Evidence evaluation: a response to the court of appeal judement in R v T, Science and Justice 51 (2011), 43-49.

[16] C. Champod, D. Meuwly, The inference of identity in forensic speaker recognition, Speech Communication 31 (2000), 193-203.

[17] P. Rose, Forensic Speaker Identification, Taylor and Francis London, UK, 2002.

[18] G. S. Morrison, Forensic voice comparison, in: I. Freckelton, H. Selby (Eds.), Expert Evidence, Thomson Reuters, Sydney, Australia, Ch 99.

[19] P. Rose, Technical Forensic Speaker Identification from a Bayesian Linguist's Perspective, Keynote paper, Forensic Speaker Recognition Workshop, Speaker Odyssey ’04. 31 May - 3 June 2004, Toledo, Spain (2004), 3-10.

[20] I. W. Evett, L. Foreman, G. Jackson, J. Lambert, DNA profiling: a discussion of issues relating to the reporting of very small match probabilities, Criminal Law Review (2000), 341–355.

[21] T. Alderman, The Bernard data set as a reference distribution for Bayesian likelihood ratio-based forensic speaker identification using formants, Proceedings of the 10th Australian International Conference on Speech Science & Technology, Macquarie University, Australia. 8-10 December (2004), 510-515.

[22] P. Rose, Forensic speaker discrimination with Australian English vowel acoustics, Proceedings of the 16th International Congress of Phonetic Sciences. Saarbrücken, Germany. 6-10 August (2007), 1817-1820.

[23] P. Rose, Bernard’s 18 – vowel inventory size and strength of forensic voice comparison evidence, Proceedings of the 13th Australian International Conference on Speech and Technology. Melbourne, Australia. 14-16 December (2010), 30-33.

[24] P. Rose, E. Winter, Traditional forensic voice comparison with female formants: Gaussian mixture model and multivariate likelihood ratio approaches, Proceedings of the 13th Australasian International Conference on Speech Science and Technology, Melbourne, Australia. 14-16 December (2010), 42-45.

[25] C. Zhang, G. S. Morrison, P. Rose, Forensic speaker recognition in Chinese: a multivariate likelihood ratio discrimination on /i/ and /y/, Proceedings of Interspeech 2008 Incorporating SST 2008, International Speech Communication Association (2008), 1937–

1940.

[26] P. Rose, Y. Kinoshita, T. Alderman, Realistic extrinsic forensic speaker discrimination with the diphthong /aɪ/, Proceedings of the 11th Australasian International Conference on Speech Science and Technology, University of Auckland, New Zealand, 6-8 December (2006), 329-334.

[27] Y. Kinoshita, T. Osanai, Within-speaker variation in diphthongal dynamics: what can we compare, Proceedings of the 11th Australasian International Conference on Speech Science and Technology, University of Auckland, New Zealand. 6-8 December (2006), 112-117.

[28] G. S. Morrison, Likelihood-ratio voice comparison using parametric representations of the formant trajectories of diphthongs, Journal of the Acoustical Society of America 125 (2009), 2387-2397.

[29] E. Enzinger, Characterizing formant tracks in Viennese diphthongs for forensic speaker comparison. Proceedings of the AES 39th International Conference – Audio Forensics, Hillerød, Denmark, (2010), 47–52.

[30] T. Becker, M. Jessen, C. Grigoras, Forensic speaker verification using formant features and Gaussian Mixture Models, Proceedings for Interspeech Special Session: Forensic Speaker Recognition – Traditional and Automatic Approaches, Brisbane, Australia, (2008).

[31] J. P. French, P. Foulkes, P. Harrison, L. Stevens, Vocal tract output measures: relative efficacy, interrelationships and limitations, International Association of Forensic Phonetics and Acoustics Conference, Santander, Spain, (2012).

[32] Y. Kinoshita, Use of likelihood ratio and Bayesian approach in forensic speaker identification, Proceedings of the 9th Australian International conference on Speech Science and Technology. 2-5 December (2002), Melbourne, Australia, 297-302.

[33] C. Kavanagh, Speaker discrimination using English nasal durations and formant dynamics, International Association of Forensic Phonetics and Acoustics Conference, Trier, Germany, (2010).

[34] C. Kavanagh, Intra- and inter-speaker variability in duration and spectral properties of English /s/, Acoustical Society of America Conference, San Diego, USA, (2011).

[35] Y. Kinoshita, Does Lindley's LR estimation formula work for speech data? Investigation using long-term F0, International Journal of Speech, Language and the Law 12 (2005), 235-254.

[36] Y. Kinoshita, S. Ishihara, P. Rose, Exploring the discriminatory potential of F0 distribution parameters in traditional forensic speaker recognition, International Journal of Speech, Language and the Law 16 (2009): 91-111.

[37] E. Gold, Collecting population statistics: the discriminant power of clicks. Acoustical Society of America Conference, San Diego, USA, (2011).

[38] E. Gold, Articulation rate as a discriminant in forensic speaker comparisons. UNSW Forensic Speech Science Conference. Sydney, Australia, 3 December (2012).

[39] G. S. Morrison, A comparison of procedures for the calculation of forensic likelihood ratios from acoustic–phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model–universal background model (GMM–UBM), Speech Communication 53 (2011), 242-256.

[40] C.G.G. Aitken, E. Gold, Evidence evaluation for discrete data. Forensic Science International 230 (2013), 147-155.

[41] P. Rose, D. Lucy, T. Osanai, Linguistic-acoustic forensic speaker identification with likelihood ratios from a multivariate hierarchical random effects model: a "non-idiot's Bayes" approach, Proceedings of the 10th Australasian Conference on Speech Science and Technology. Macquarie University, Australia, 8-10 December (2004), 492-497.

[42] P. Rose, The intrinsic forensic discriminatory power of diphthongs, Proceedings of the 11th Australasian International Conference on Speech Science and Technology. 6-8 December (2006), University of Aukland, New Zealand, 64-69.

[43] P. Rose, The effect of correlation on strength of evidence estimates in forensic voice comparison: uni- and multivariate likelihood ratio-based discrimination with Australian English vowel acoustics, International Journal of Biometrics 2 (2010), 316-329.

[44] G. S. Morrison, T. Thiruvaran, J. Epps, An issue in the calculation of logistic-regression calibration and fusion weights for forensic voice comparison, Proceedings of the 13th Australasian International Conference on Speech Science and Technology, Melbourne, Australia. 14-16 December (2010), 74-77.

[45] E. Gold, V. Hughes, Defining interdependencies between speech parameters, BBfor2 Short Summer School in Forensic Evidence Evaluation and Validation, Madrid, Spain 18 June (2012).

[46] V. Hughes, P. Foulkes, Effects of variation on the computation of numerical likelihood ratios for forensic voice comparison, Paper presented at International Association of Forensic Phonetics and Acoustics conference, Universidad Internacional Menedez Pelayo, Santander, 5th - 8th August (2012).

[47] G. S. Morrison, F. Ochoa, T. Thiruvaran, Database selection for forensic voice comparison, in: Proceedings of Odyssey 2012: The Language and Speaker Recognition Workshop, Singapore, International Speech Communication Association (2012).

[48] G. S. Morrison, P. Rose, C. Zhang, Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice, Australian Journal of Forensic Sciences (2012), 155-167.

[49] S. Ishihara, Y. Kinoshita, How many do we need? Exploration of the Population Size Effect on the performance of forensic speaker classification, Paper presented at the 9th Annual Conference of the International Speech Communication Association (Interspeech), Brisbane, Australia, (2008),1941-1944.

[50] Y. Kinoshita, S. Ishihara, The effect of sample size on the performance of likelihood ratio-based forensic voice comparison, Proceedings of the 14th Australasian International

Conference on Speech Science and Technology, 3-6 December (2012), Macquarie University, Australia.

[51] G. S. Morrison, Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio, Australian Journal of Forensic Sciences 45 (2013), 173-197.

[52] G. S. Morrison, Y. Kinoshita, Automatic-type calibration of traditionally derived likelihood ratios: Forensic analysis of Australian English /o/ formant trajectories. Proceedings of Interspeech 2008 International Speech Communication Association, 1501–1504.

[53] G. S. Morrison, Measuring the validity and reliability of forensic likelihood-ratio systems, Science & Justice 51 (2011), 91–98.

[54] J. K. Chambers, Sociolinguistic theory, 2nd ed, Blackwell, Oxford, UK, 2005.

[55] P. Eckert, Linguistic variation as social practise, Blackwell, Oxford, UK, 2000.

[56] R. Wardaugh, An introduction to sociolinguistics, 5th ed, Blackwell, Oxford, UK, 2006.

[57] C. Byrne, P. Foulkes, The mobile phone effect on vowel formants, International Journal of Speech, Language and the Law 11 (2004), 83-102.

[58] E. Enzinger, Measuring the Effects of the Adaptive Multi-Rate (AMR) codecs on formant tracker performance, Second Pan-American/Iberian Meeting on Acoustics, November 15–19 (2010) Cancún, México.

[59] E. Gold, The effects of video and voice recorders in cellular phones on vowel formants and fundamental frequency, Unpublished Master's Thesis, University of York, 2009.

[60] H. J. Künzel, Beware of the 'telephone effect': the influence of telephone transmission on the measurement of formant frequencies, International Journal of Speech, Language and the Law 8 (2001), 80-99.

[61] Daubert v. Merrell Dow Pharmaceuticals 113 S Ct 2786 1993.

[62] D. V. Lindley, A problem in forensic science, Biometrika 64 (1977), 207-213.

[63] C. G. G. Aitken, D. Lucy, Evaluation of trace evidence in the form of multivariate data, Applied Statistics 54 (2004), 109-122.

[64] D. A. Reynolds, T. F. Quatieri, R. B. Dunn, Speaker verification using adapted Gaussian mixture models, Digital Signal Processing 10 (2001), 19–41.

[65] P. Rose, T. Osanai, Y. Kinoshita, Strength of forensic speaker identification evidence: multispeaker formant- and cepstrum-based segmental discrimination with a Bayesian likelihood ratio as threshold, Forensic Linguistics 10 (2003), 179-202.

[66] Y. Kinoshita, Testing realistic forensic speaker identification in Japanese: a likelihood ratio-based approach using formants, PhD Dissertation, Australian National University, 2001.

[67] T. Alderman, Forensic speaker identification: a likelihood ratio-based approach using vowel formants, LINCOM Studies in Phonetics, Lincom, Munich, Germany, 2005.

[68] D. Meuwly, A. Drygajlo Forensic speaker recognition based on a Bayesian framework and Gaussian mixture modelling (GMM), Proceedings of Odyssey (2001), 145-150.

[69] A. Alexander, A. Drygajlo, Scoring and direct methods for the interpretation of evidence in forensic speaker recognition. Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP), Jeju, Korea, (2004).

[70] C. Zhang, G. S. Morrison, T. Thiruvaran, Forensic voice comparison using Chinese /iau/, Proceedings of the 17th International Congress of Phonetic Sciences, 17-21 August (2011), Hong Kong, 2280-2283.

[71] D. H. Kaye, Logical relevance: problems with the reference population and DNA mixtures in People v. Pizarro, Law, Probability and Risk 3 (2004), 211-220.

[72] D. H. Kaye, DNA probabilities in People v. Prince: When are racial and ethnic statistics relevant? in: T. Speed, D. Nolan, D. (Eds.) Probability and Statistics: Essays in Honour of David A Freedman, Institute of Mathematical Statistics, Beachwood, OH, 289-301.

[73] P. Rose, A forensic phonetic investigation in non-contemporaneous variation in the F-pattern of similar-sounding speakers, in R. Mannell, J. Robert-Ribes (Eds.) Proceedings of 5th International Conference on Spoken Language Processing, ASSTA, Canberra (1998), 49-52.

[74] P. Rose, Long- and short-term within-speaker differences in the formants of Australian hello, Journal of the International Phonetic Association 29 (1999), 1-31.

[75] P. Rose, Forensic voice comparison with Japanese vowel acoustics – a likelihood ratio-based approach using segmental cepstra, in W. S. Lee, E. Zee (Eds.) Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong, 17-21 August (2011), 1718-1721.

[76] D. Loakes, A forensic phonetic investigation into the speech patterns of identical and non-identical twins, PhD Dissertation, University of Melbourne, Australia, 2006.

[77] C. Montgomery, Northern English dialects: a perceptual approach, PhD Thesis, University of Sheffield, 2007.

[78] R. Bull, B. R. Clifford, Earwitness voice recognition accuracy, in G. L. E. Wells, F. Loftus (Eds.) Eyewitness Testimony: Psychological Perspectives, Cambridge University Press, Cambridge, UK, 1984, 92-123

[79] R. Bull, B. R. Clifford, Earwitness testimony, in A. Heaton-Armstrong, E. Shepherd, D. Wolchover (Eds.) Analysing Witness Testimony: A Guide for Legal Practitioners and Other Professionals, Blackstone Press, London, UK, 1999, 194-206.

[80] D. Watt, The identification of the individual through speech, in C. Llamas, D. Watt (Eds.) Language and Identities. Edinburgh University Press, Edinburgh, 2010, 76-85.

[81] F. Nolan, K. McDougall, T. Hudson, Some acoustic correlates of perceived (dis)similarity between same-accent voices, in W. S. Lee, E. Zee (Eds.) Proceedings of the

17th International Congress of Phonetic Sciences, Hong Kong. 17-21 August (2011), 1506-1509.

[82] J. P. French, P. Harrison, Position Statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases, International Journal of Speech, Language and the Law 14 (2007), 137-144.

[83] V. Hughes, The effects of variability on the outcome of numerical likelihood ratios for forensic voice comparison [working title], PhD in progress. University of York, UK.

[84] P. Rose, The likelihood ratio goes to Monte Carlo: the effect of reference sample size on likelihood-ratio estimates, Proceedings of the 14th Australasian International Conference on Speech Science and Technology. 3-6 December (2012), Macquarie University, Australia.

[85] I. Kononenko, Comparison of inductive and naïve Bayesian capitalised learning approaches to automatic knowledge acquisition, in B. Wielinga et al. (Eds.) Current trends in knowledge acquisition, IOS Press, Amsterdam, Netherlands, 1990.

[86] N. Brümmer, L. Burget, J. H. Cernocký, O. Glembek, F. Grézl, M. Karafiát, D. A. van Leeuwen, P. Matejka, P. Schwarz, A. Strasheim, Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST SRE 2006, IEEE Transactions on Audio Speech and Language Processing 15 (2007) 2072-2084.

[87] J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. T. Toledano, J. Ortega-Garcia, Emulating DNA: rigorous quantification of evidential weight in transparent and testable forensic speaker recognition, IEEE Transactions of Audio, Speech and Language Processing 15 (2007), 2104-2115.

[88] D. Ramos Castro, Forensic evaluation of the evidence using automatic speaker recognition systems, PhD dissertation, Universidad Autónoma de Madrid, Madrid, Spain, 2007.

[89] N. Brümmer, E. de Villers, What is the 'relevant population' in Bayesian forensic inference? https://sites.google.com/site/nikobrummer/ (accessed 03/11/11), 2011.

[90] T. Hudson, G. de Jong, K. McDougall, P. Harrison, F. Nolan, F0 statistics for 100 young male speakers of Standard Southern British English, in J. Trouvain, W. Barry (Eds.) Proceedings of the 16th International Congress of Phonetic Sciences, 6-10 August (2007), Saarbrücken, 1809-1812.

[91] E. Gold, P. French, P. Harrison, Clicking behaviour as a speaker discriminant in English, Journal of the International Phonetic Association (in press).

[92] F. Taroni, C. G. G. Aitken, P. Garbolino, A. Biedermann, Bayesian networks and probabilistic inference in forensic science, Wiley, Chichester, UK, 2006.

[93] D. J. Balding, R. A. Nichols, DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands, Forensic Science International 64 (1994), 125-140.

[94] D. J. Balding, R. A. Nichols, Significant genetic correlations among Caucasians at forensic DNA loci, Heredity 78 (1997), 583-589.

[95] National Research Council (Committee on DNA Forensic Science) The evaluation of forensic DNA evidence, National Academy Press, Washington, USA, 1996.