Morrison, G, Lindh, J, Curran, J.

Likelihood ratio calculation for a disputed-utterance analysis with limited available data.

Speech Communication. 2014;58:81-90.

http://dx.doi.org/10.1016/j.specom.2013.11.004

Likelihood ratio calculation for a disputed-utterance analysis with limited available data

Geoffrey Stewart Morrison[1*], Jonas Lindh[2], James M Curran[3]

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

[2]Division of Speech and Language Pathology, Department of Clinical Neuroscience and Rehabilitation, Institute of Neuroscience and Physiology, Sahlgrenska Academy, University of Gothenburg, Box 452 SE405 30, Gothenburg, Sweden

[3]Department of Statistics, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand

*Corresponding author:

Dr Geoffrey Stewart Morrison
803 - 4438 West 10th Avenue
Vancouver BC
V6R 4R8
CANADA
tel: +1 604 637 0896 / +44 191 645 0896 / +61 2 800 74930
e-mail: geoff-morrison@forensic-evaluation.net

Draft of 13 November 2013

Abstract

We present a disputed-utterance analysis using relevant data, quantitative measurements, and statistical models to calculate likelihood ratios. The acoustic data were taken from an actual forensic case in which the amount of data available to train the statistical models was small and the data point from the disputed word was far out on the tail of one of the modelled distributions. A procedure based on single multivariate Gaussian models for each hypothesis led to an unrealistically high likelihood ratio value with extremely poor reliability, but a procedure based on Hotelling's T^2 statistic and a procedure based on calculating a posterior predictive density produced more acceptable results. The Hotelling's T^2 procedure attempts to take account of the sampling uncertainty of the mean vectors and covariance matrices due to the small number of tokens used to train the models, and the posterior-predictive-density analysis integrates out the values of the mean vectors and covariance matrices as nuisance parameters. Data scarcity is common in forensic speech science and we argue that it is important not to accept extremely large calculated likelihood ratios at face value, but to consider whether such values can be supported given the size of the available data and modelling constraints.

Keywords

forensic; likelihood ratio; disputed utterance; reliability; Hotelling's T^2; posterior predictive density

Non-standard abbreviations[1]

1. Introduction

In forensic speech science the amount of data available from casework samples is often limited. In forensic-speech-science research reports and presentations, and in forensic-speech-science casework reports, we have sometimes seen very large likelihood-ratio values presented, with those values being derived from calculations based on small amounts of data. For reasons we will explain in the present paper, we think that the likelihood-ratio values reported could not be supported by the small amount of data available, but it appears that the authors were not aware of this problem and simply reported the calculated values. In the present paper we illustrate this problem using data from an actual disputed-utterance case on which one of us worked. We chose this example as an extreme example in which the problem should be obvious, but one should also be aware that, because by necessity likelihood ratios in forensic speech science often have to be calculated on the basis of small amounts of data, the problem may also have a substantial impact on calculated likelihood-ratio values in cases where it is not so obvious. The same problem may also occur in other branches of forensic science.

[1] All abbreviations used are standard in the field.

In a disputed-utterance analysis the task of the forensic scientist is to calculate the strength of evidence in the form of a likelihood ratio answering the question: How probable are the observed acoustic properties of the disputed utterance if the speaker had said what the prosecution claims they said versus if the speaker had said what the defence claims they said (Morrison and Hoy, 2012). The present paper describes the calculation of a forensic likelihood ratio on the basis of relevant data, quantitative measurements, and statistical models.[2] The data are taken from an actual case in which the amount of data available to train the models was limited. An initial analysis using single-Gaussian models is described, and the problem with this approach in this case is discussed. The problem is due to the fact that not only is there a small amount of training data but the data point for the disputed utterance is also far out on the tail of one of the modelled Gaussian distributions. Revised analyses are then described. One attempts to take account of the sampling uncertainty of the mean vectors and covariance matrices via the use of a Hotelling's T^2 distribution. The other makes use of a Bayesian posterior-predictive-density analysis which integrates out the values of the mean vectors and covariance matrices as nuisance parameters.

Copies of the data and the MATLAB (Mathworks, 2010) and R (R Development Core Team, 2013) scripts used to make the calculations described in this paper are available from http://geoff-morrison.net/#DispUtLimDat

2. Background to the case

In 2008 one of the authors of the current paper, JL, was asked to perform a disputed-utterance analysis in a Swedish murder case [nº B1293-07 of Hovrätten för nedre Norrland, 2008-02-26]. A word on an audio recording of a police interview with an eye witness (a female Swedish speaker) was disputed as being either the pronoun "dom" [dɔm] (*they*) or the name "Tim" [tʰɪm]. The recording also contained 29 undisputed tokens of "dom" and 16 undisputed tokens of "Tim" spoken by the same

[2] We work in a paradigm which requires the use of the likelihood-ratio framework, has a strong preference for the calculation of likelihood ratios on the basis of relevant data, quantitative measurements, and statistical models, and requires empirical assessment of the degree of validity and reliability of the analytical procedures under conditions reflecting those of the case under investigation (we will refer to this as the "new" paradigm). Although numerous papers have been published describing the new paradigm and describing approaches to forensic voice comparison conducted within the new paradigm, as far as we are aware Morrison and Hoy (2012) is the only previously published paper on disputed utterances conducted within this paradigm. The arguments in favour of the use of the new paradigm for disputed-utterance analysis are made in Morrison and Hoy (2012) and are not repeated here since this is not the focus of the present paper. The present paper assumes the new paradigm and addresses a problem which only arises if one's approach is based on data and statistical models. This problem is the focus of the present paper. We think that this is an important problem which potentially affects not just disputed-utterance analysis, but also forensic voice comparison and other branches of forensic science. The disputed-utterance case we describe in this paper happens to provide a good example of the problem. Our primary objective is to make researchers and practitioners aware of this problem and our secondary objective is to describe some potential solutions. The old paradigm for disputed-utterance analysis is based on listening to the speech on the audio recording and making a subjective judgement – the problem which we address in the present paper does not arise in that paradigm.

speaker. The fact that all the tokens came from the same recording meant that there were no recording-channel or speaking-style mismatch issues.

The original speaker was not cooperative at the point in time when the disputed-utterance analysis was performed, and no other recordings of this speaker were provided, hence no additional data from this speaker could be used in the analysis.

3. Acoustic analysis

In 2008 JL measured the voice onset time (VOT) and first and second formant (F1 and F2) values in the disputed token and each of the tokens of the undisputed words ("dom" and "Tim"). The recording had been made on an analogue tape and was digitised with a sampling rate of 16 kHz and 16 bit quantisation. Details of the recording equipment were not provided at the time and cannot be ascertained now. JL subjectively characterised the recoding quality as poor but not extremely bad. The signal to noise ratio was 28 dB.

Measurements were made using PRAAT (Boersma and Weenik, 2008). The VOT of each token was measured three times (analyses below will be based on the mean of the three VOT measurements for each token). Prevoicing and plosive bursts were clearly visible on a waveform display. No substantial speaking rate differences were observed during the relevant potions of the recording.[3] F1 and F2 were measured at a single point in the middle of the vowel in each token.[4] Formants were measured using the Burg autocorrelation linear-predictive-coding (LPC) algorithm (Anderson, 1978). The resulting formant values were overlayed on a broadband spectrogram and the maximum frequency below which to search for formants adjusted if the initial results were obviously erroneous. Otherwise the settings recommended in the PRAAT documentation for female speakers were used.[5]

On a single graph, JL produced histograms of the distribution of the VOT measurements from the tokens of each word ("dom" and "Tim") and indicated the location of the VOT measurement from the disputed token relative to these histograms, see Fig. 1. Likewise he produced a single F1 by F2 graph

[3] Swedish has devoicing of voiced plosives following voiceless consonants, e.g., in "samt dom" *and them*. Only phonetically voiced tokens of [dɔm] were analysed (and included in the count of 29 tokens). The disputed utterance was not in a context which would result in devoicing.

[4] Note that the units of analysis are the words "dom" [dɔm] and "Tim" [tʰɪm], not the vowels [ɔ] and [ɪ] – any nasalisation effect on the vowel due to the following bilabial nasal is therefore automatically included in the formant measurements. Likewise for any differential reduction effect in the personal pronoun "dom" versus the proper noun "Tim".

[5] Recommended settings in the PRAAT documentation for female speakers for "Sound: To Formant (burg)...": Maximum frequency 5500 Hz. Maximum number of formants 5 (i.e, 10 LPC coefficients). Window length 0.025, "Praat uses a Gaussian-like analysis window with sidelobes below −120 dB. . . . the actual Gaussian window duration is 0.050 seconds. This window has values below 4% outside the central 0.025 seconds, and its frequency resolution (−3 dB point) is 1.298 / (0.025 s) = 51.9 Hz" (PRAAT manual). Pre-emphasis 50 Hz "frequencies below 50 Hz are not enhanced, frequencies around 100 Hz are amplified by 6 dB, frequencies around 200 Hz are amplified by 12 dB, and so forth" (PRAAT manual).

containing scatterplots of the formant measurements from the tokens of each undisputed word and indicated the location of the F1-F2 measurement from the disputed token, see Fig 2.[6]

At the time JL did not calculate a numeric likelihood ratio for presentation during the trial. Instead, he used the graphs as a visual support for his conclusion that the measured acoustic properties of the disputed word were much more likely if the speaker had said “dom” rather than if she had said “Tim”. He reported +4 on the Swedish National Laboratory of Forensic Science’s −4 to + 4 scale, corresponding to a likelihood ratio of equal to or greater than one million (Nordgaard et al., 2012).

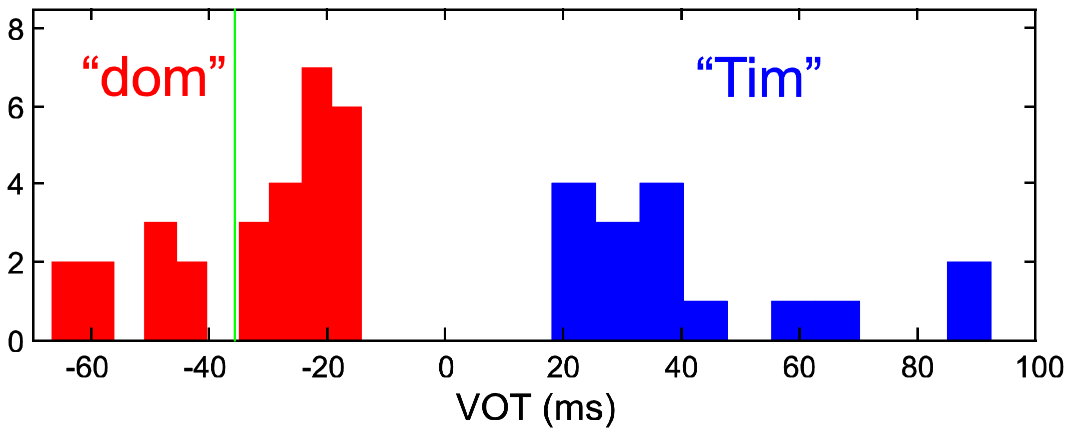

Fig. 1. Histogram of the VOT measurements from the undisputed “dom” and “Tim” tokens. The vertical line indicates the VOT measurement from the disputed word.

[6] Not apparent from the separate Fig.1 and Fig. 2 plots, but apparent from a three-dimensional scatterplot of the data (produced by running the supplied MATLAB script), VOT was negatively correlated with F1 and positively correlated with F2.

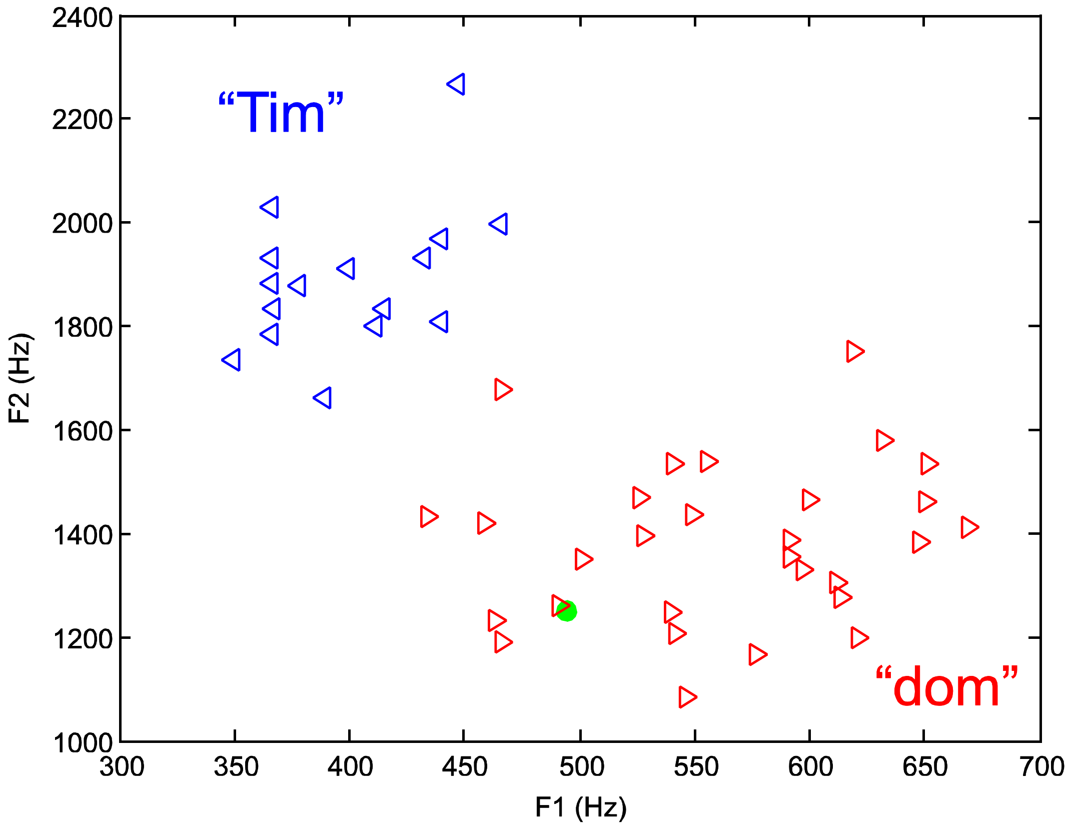

Fig. 2. Scatterplot of the first- and second-formant (F1 and F2) measurements from the undisputed “dom” and “Tim” tokens. The filled circle represents the F1-F2 measurement from the disputed word.

4. First statistical analysis: Single-Gaussian models

In the present paper we reexamine the acoustic data from the 2008 case and calculate a numeric likelihood ratio to assess the strength of the evidence. We used the data from the undisputed “dom” tokens to build a model of the hypothesis that the word spoken was “dom”, and the data from the undisputed “Tim” tokens to build a model of the hypothesis that the word spoken was “Tim”. Each model was a probability-density function (pdf), and given the small amount of training data available it did not make sense to train anything more complex than a single multivariate Gaussian distribution model for each hypothesis. We did, however, first apply a transformation to the VOT data so that it better conformed to the assumption that the data were normally distributed. Details of the transformation and likelihood-ratio calculation are as follows.

4.1. Transformation of VOT data

In Fig. 1 it is obvious that the VOT data, measured in milliseconds, are not normally distributed. Each category has a distribution which is skewed away from zero. In order to reduce the skewness and better conform to the assumption of normality implied by the use of a single-Gaussian model for each hypothesis, we applied an arctangent transformation, Eq. 1.

$$v = \tan^{-1}(\theta v_{\mathrm{ms}}) \tag{1}$$

where v is the transformed VOT value, v_{ms} is the original VOT measurement value in milliseconds, and θ is an angle coefficient (a non-linear scaling coefficient) the value of which was optimised using a Nelder-Mead simplex search which minimised the sum of the absolute values of the per-category skewness statistics from the two categories in the training data (see Lagarias et al., 1998, for details of the search algorithm which was implemented in MATLAB's `fminsearch` function). The minimisation-objective function, C_{g_1}, for the simplex search was that given in Eq. 2.

$$C_{g_1} = |\, g_{1,\mathrm{dom}} \,| + |\, g_{1,\mathrm{Tim}} \,| \tag{2a}$$

$$g_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(v_i - \bar{v}\right)^3}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(v_i - \bar{v}\right)^2\right)^{3/2}} \tag{2b}$$

where $g_{1,\mathrm{dom}}$ and $g_{1,\mathrm{Tim}}$ are the sample skewness statistics of the transformed VOT values from the "dom" and "Tim" categories respectively, v_i is the transformed VOT value measured on the ith token of the category, of which there are n (29 for "dom" and 16 for "Tim"), and $\bar{v}$ is the sample mean of the transformed VOT values across all the tokens of the category.

In this case the optimal θ value was 0.022 radians/ms. Fig. 3 shows histograms of the transformed VOT values. Note the reduction in skewness relative to Fig. 1. The skewness statistics for the VOT data before and after transformation are given in Table 1.

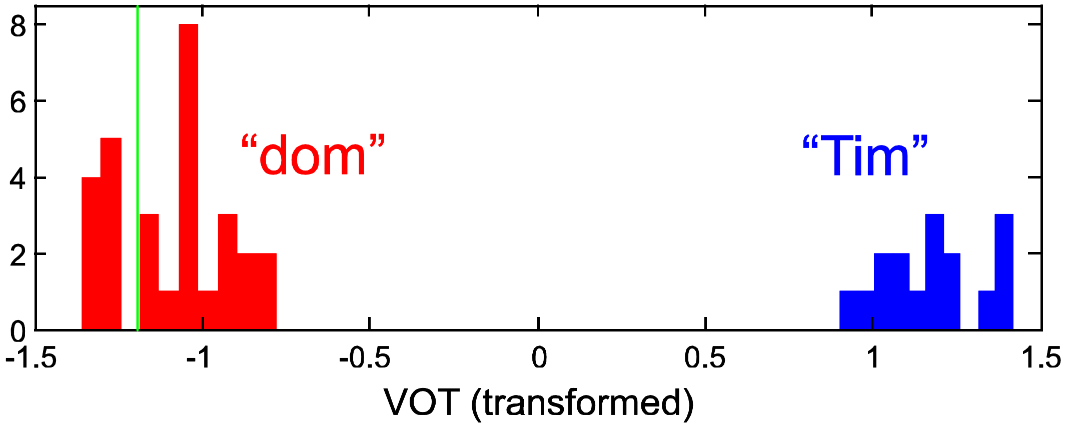

Fig. 3. Histogram of the transformed VOT measurements from the undisputed "dom" and "Tim" tokens. The vertical line indicates the VOT measurement from the disputed word.

Table 1. Skewness statistics, g_1, for the VOT data before and after transformation.

	before	after
“dom”	−0.827	+0.000
“Tim”	+1.223	+0.013

4.2. *Likelihood-ratio calculation*

The pdf value, $f(\cdot)$, of each of the “dom” and “Tim” models was evaluated at the data point corresponding to the measurements taken from the disputed utterance. Each model consisted of a single multivariate Gaussian, Eq 3.

$$f\left(\boldsymbol{y}|\boldsymbol{X}\right)=\frac{1}{\sqrt{|\boldsymbol{S}|(2\pi)^k}}e^{-\frac{1}{2}(\boldsymbol{y}-\bar{\boldsymbol{x}})'\boldsymbol{S}^{-1}(\boldsymbol{y}-\bar{\boldsymbol{x}})} \tag{3a}$$

$$\boldsymbol{S}=\tfrac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})' \tag{3b}$$

where $\boldsymbol{y}$ is the $k \times 1$ vector of F1, F2, and transformed VOT measurements $[v \text{ F1 F2}]'$ from the disputed word; k, the number of variables, is 3; $\boldsymbol{x}_i$ is the $k \times 1$ vector $[v \text{ F1 F2}]'$ of measurements from the ith token of the set of undisputed “dom” or “Tim” words, of which there are n (29 for “dom” and 16 for “Tim”); $\bar{\boldsymbol{x}}$ is the $k \times 1$ sample mean vector of the $k \times n$ matrix $\boldsymbol{X} = [\boldsymbol{x}_i, ..., \boldsymbol{x}_n]$ containing measurements from all tokens of a given undisputed word ($\boldsymbol{X}_{\text{dom}} = [\boldsymbol{x}_{\text{dom},1}, ..., \boldsymbol{x}_{\text{dom},29}]$, $\boldsymbol{X}_{\text{Tim}} = [\boldsymbol{x}_{\text{Tim},1}, ..., \boldsymbol{x}_{\text{Tim},16}]$); and $\boldsymbol{S}$ is the $k \times k$ least-squares sample covariance matrix.

The likelihood ratio, LR, was then calculated as the ratio of the pdf value obtained for the disputed utterance given the “dom” model, and the pdf value obtained for the disputed utterance given the “Tim” model, Eq. 4.

$$LR\left(\boldsymbol{y}|\boldsymbol{X}_{\text{dom}},\boldsymbol{X}_{\text{Tim}}\right)=\frac{f\left(\boldsymbol{y}|\boldsymbol{X}_{\text{dom}}\right)}{f\left(\boldsymbol{y}|\boldsymbol{X}_{\text{Tim}}\right)} \tag{4}$$

For greater numerical stability, the calculations were actually performed by taking the logarithm of each pdf value and then subtracting one from the other, Eq. 5.

$$\log(\ LR(\boldsymbol{y} \mid \boldsymbol{X}_{\text{dom}}, \boldsymbol{X}_{\text{Tim}})\ ) = \log(\ f(\boldsymbol{y} \mid \boldsymbol{X}_{\text{dom}})\ ) - \log(\ f(\boldsymbol{y} \mid \boldsymbol{X}_{\text{Tim}})\ ) \tag{5}$$

4.3. *Likelihood-ratio result and discussion*

The likelihood ratio obtained using the single-Gaussian procedure was 2×10^{77}. This is an extremely large number which cannot be sustained by the small amount of data used to train the model. To give some idea of how large the number is, note that the estimated number of stars in the observable universe is only around 10^{22}.

It could be that the distribution of the population is such that at the acoustic-phonetic level[7] the probability of the evidence given the "Tim" hypothesis is actually zero. To assert that this probability is zero would require this to be assumed since it could never be empirically demonstrated that the value is in fact zero as opposed to some very small non-zero value. Making a categorical decision as to which of the hypotheses is correct is the task of the trier of fact, not of the forensic scientist. Given that the task of the forensic scientist is to evaluate the strength of the evidence, and that there are always limitations on the availability of data directly related to the crime and other relevant data which can be used for statistical modelling of the relevant hypotheses and calculation of likelihood ratios, we agree with the sentiment of Champod and Evett (2000, p. 242) who "do not think that infinite likelihood ratios (or likelihood ratios in the range of thousands of billions) can be scientifically sustained in any forensic discipline." We would, however, not state any particular value, such as 10^{12}, as the limit for acceptability. Consideration should be given as to whether the values can be justified given the amount and complexity of data, and the likely bias and variance of statistical models. We regard this as an issue of the adequacy of the model rather than an issue of the value output by the model, and model adequacy is what we are attempting to address in this paper. See Kaye (2009) for a discussion of this issue in the context of DNA.

How did such a large number result from this analysis? Looking at Figs. 2 and 3 (one can also view a three-dimensional scatterplot of the data by running the supplied MATLAB script) it is apparent that the data point corresponding to the disputed utterance is very far out on the tail of the distribution of the "Tim" data and that that distribution is modelled using a very small number of data points. There is a problem with small *n* and tails: A small amount of training data results in the sample distribution being a poor estimate of the population distribution. Near the centre of the distribution the variability may not be too bad, but any fluctuation in training data (as would result from taking a different sample of the same size from the same population) can have a huge effect on the estimated pdf value far out on the tail.

Fig. 4 provides a simplified one-dimensional example: Imagine that the Gaussian distributions shown were calculated on the basis of data sampled from two relevant populations (each sample is taken as representative of its respective population) and that a likelihood ratio was calculated for a sample of interest which had an *x* value of 45, relatively far out on the tail of the rightmost distribution. In the left panel the rightmost Gaussian model has a mean of 75 and the calculated likelihood ratio is 27. In the right panel imagine that a different sample has been taken from the same population as before (also taken as representative of its respective population) but that this sample resulted in a small change

[7] We do not address here the probability of "slips of the tongue" or of "misspeaking" where a person intended to say one thing but categorically said something else.

in the rightmost model's mean, making it 76 rather than 75 (for simplicity also imagine that the standard deviation for this model has not changed, although in reality we would expect this to change as well). The change in the model is small but the resulting change in the calculated likelihood ratio is relatively large, it now being 37 rather than 27.

The disputed utterance is even further out on the tail of the distribution of the "Tim" data than in the simplified example above. Hence, the estimate of the denominator of the likelihood ratio in this case is expected to have very poor reliability, and by extension the likelihood ratio is expected to be unreliable. We investigate the reliability of this procedure in the next section.

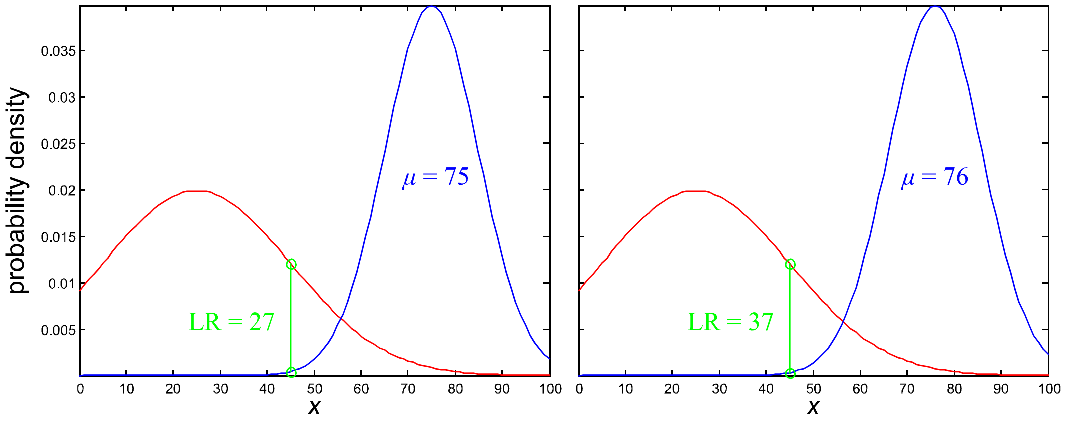

Fig. 4. Example of a relatively large difference in the value of a likelihood ratio calculated on the tail of a distribution due to a small change in that distribution.

4.4. *Evaluation of reliability*

When calculating a forensic likelihood ratio on the basis of relevant data, quantitative measurements, and statistical models, there can be various sources of imprecision (Stoel and Sjerps, 2012), for example: variability in the measurement technique (Zhang et al., 2013); variability in sampling of the relevant population (Curran et al., 2002; Curran, 2005); and, in forensic voice comparison, within-speaker between-recording-session variability, which is generally not represented in the available suspect data (Morrison, 2011). In the current paper we are concerned with imprecision due to a modelling issue: Models are trained on data which are samples from the relevant population. A different set of samples will result in different models and thus different estimates of the likelihood ratio as a quantification of the strength of evidence. This is a general problem which affects all modelling attempts but, as mentioned above, it becomes acute when a pdf value is calculated for a point on the tail of a modelled distribution and the model is trained on a small amount of data.

4.4.1. Reliability evaluation methodology

The reliability of the single-Gaussian modelling procedure was empirically investigated using Monte Carlo simulation to generate multiple samples of 29 “dom” tokens and 16 “Tim” tokens (the same *n* as in the original data). The simulated data were randomly generated from multivariate Gaussian distributions with population mean vectors and covariance matrices equivalent to the sample mean vectors and covariance matrices from the original undisputed “dom” and “Tim” data. Random numbers were generated using MATLAB’s `mvnrnd` function, and the random number stream was first reset so the that same sets of random data could be obtained if the experiment were rerun. A single-Gaussian model of the “dom” hypotheses was trained using 29 tokens sampled from the simulated “dom” population, a single-Gaussian model of the “Tim” hypotheses was trained using 16 tokens sampled from the simulated “Tim” population, and the likelihood ratio for the $[v \text{ F1 F2}]'$ data point from the original disputed utterance was calculated using these models. 10 000 sample sets were generated and 10 000 likelihood ratios calculated.

4.4.2. Reliability results and discussion

Fig. 5 shows a histogram of the results from the Monte Carlo simulations. The reliability of the likelihood-ratio analysis using the amount of available data and the single-Gaussian modelling approach can be seen to be very poor: the spread of the distribution of the Monte Carlo results is very wide. The three vertical lines in Fig. 5 represent, from right to left, the likelihood ratio calculated using the original data (2×10^{77}), the 5th percentile (9×10^{54}), and the 1st percentile (2×10^{44}). We could say that on the basis of our analysis we are 99% certain that the likelihood ratio is at least 2×10^{44}. That is, we are 99% certain that the acoustic properties of the disputed utterance are at least 2×10^{44} times more likely if the speaker had said “dom” rather than “Tim”. We feel, however, that this is still not a number which can be supported by the small amount of training data.

The Monte Carlo simulation approach goes some way towards addressing the small-*n*-and-tails problem, but it suffers from the assumption that the sample statistics from the original data can be used as population parameters for the simulation. Although this may be a reasonable approach for investigating the reliability of the results, it is not a good approach for investigating their validity since the results of the Monte Carlo simulation are necessarily centred around the original sample means. If the sample means and covariances are poor representations of the population means and covariances, this will not be revealed by the Monte Carlo simulation. Thus if the sample means and covariances are such that they result in a much larger (or smaller) likelihood-ratio value than would result if the true population means and covariances were known, the Monte Carlo simulation cannot address this problem.

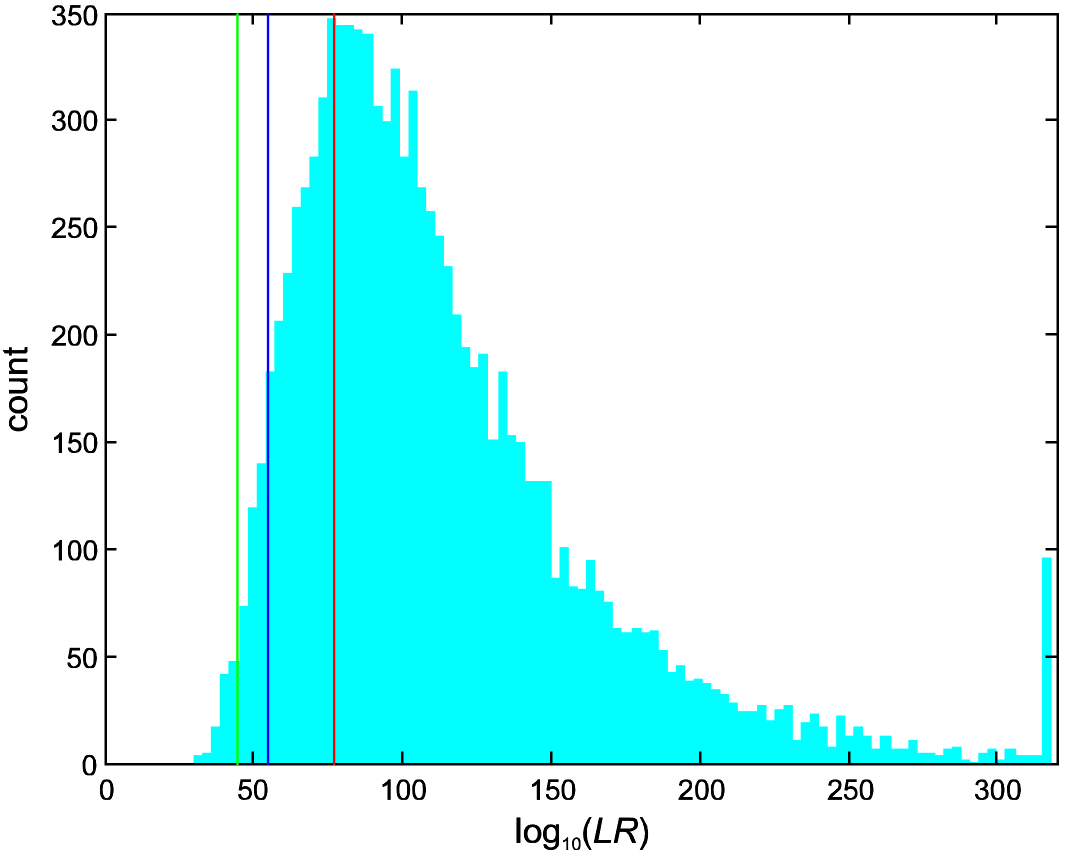

Fig. 5. Histogram of likelihood-ratio results from the Monte Carlo simulations using single-Gaussian models. The vertical lines represent, from right to left, the likelihood ratio calculated using the original data (2×10^{77}), the 5th percentile (9×10^{54}), and the 1st percentile (2×10^{44}). The final bar at the right of the histogram includes the results of simulations for which, given the precision of numbers representable by the computer software, the calculated value of the denominator of the likelihood ratio was zero, hence the calculated value for the likelihood ratio was infinite.

5. Second statistical analysis: Hotelling's T^2 models

The modelling approach based on single Gaussians does not take account of the sampling uncertainty in the mean vectors and covariance matrices. This is especially relevant given the small number of tokens used to train the models. In this section we reanalyse the data using a modelling procedure which does take account of this sampling uncertainty.

A standard approach to account for a small *n* when modelling a multivariate normal distribution is to use Hotelling's T^2 distribution (Hotelling, 1931), the multivariate extension of Student's *t* distribution (Student, 1908). We will therefore repeat our disputed-utterance analysis using a Hotelling's T^2 distribution rather than a Gaussian distribution. Models based on Hotelling's T^2 distribution have previously been used to calculate likelihood ratios for glass evidence (Curran et al., 1997a, 1997b).

5.1. Likelihood-ratio calculation

Hotelling's T^2 statistic is usually used to test the hypothesis that a sample consisting of multiple data points is drawn from a population with a specified mean (one-sample statistic), or that two samples each consisting of multiple data points are drawn from the same population (two-sample statistic). In this case we will use the value of the disputed utterance as one sample (of sample size 1), then calculate the Hotelling's T^2 statistic using the undisputed "dom" data as the second sample, then repeat using the undisputed "Tim" data as the second sample. We will then calculate the pdf value for the T^2 statistic from each of the "dom" and "Tim" models and divide the former by the latter. A way to conceptualise this is that we are calculating the likelihood of obtaining the disputed data point and the undisputed "dom" data given the hypothesis that these come from the same population versus the likelihood of obtaining the disputed data point and the undisputed "Tim" data given the hypothesis that these come from the same population.

For each of each of the undisputed "dom" and "Tim" data, Hotelling's T^2 statistic was calculated as in Eq. 6.

$$T^2 = \frac{n}{n+1}\left(\bar{\boldsymbol{x}} - \boldsymbol{y}\right)' \boldsymbol{S}^{-1}\left(\bar{\boldsymbol{x}} - \boldsymbol{y}\right) \tag{6}$$

Where T^2 is Hotelling's T^2 statistic and the other symbols are as previously defined. Note that we substituted the value 1 for the sample size of the $\boldsymbol{y}$ sample and simplified Eq. 6 (a sample of sample size 1 ends up playing no part on the calculation of the covariance matrix, and the latter is based entirely on the data from the other sample, the $\boldsymbol{X}$ sample of sample size n, as in Eq. 3b).

Multiplying by a constant dependent on the number of dimensions, k, and sample size, n, in the training data for the model, Hotelling's T^2 statistic follows an F distribution with degrees of freedom dependent on, k and n, see Eq. 7.

$$m = n - k \tag{7a}$$

$$z = \tfrac{m}{k(n-1)} T^2 \sim F_{k,m} \tag{7b}$$

The probability-density-function value, the likelihood, $L(\cdot)$, for the F distribution was evaluated at the value z, Eq. 8.

$$L\left(\boldsymbol{y}, \boldsymbol{X}\right) = \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{m}{2}\right)}\left(\tfrac{k}{m}\right)^{\frac{k}{2}} \frac{z^{\frac{k-2}{2}}}{\left(1+\left(\frac{k}{m}\right)z\right)^{\frac{n}{2}}} \tag{8}$$

Where Γ(·) is the gamma function.

Finally, the likelihood ratio was calculated as the ratio of the likelihood obtained for the disputed utterance plus “dom” model, and the likelihood obtained for the disputed utterance plus “Tim” model, Eq. 9.

$$LR\left(\boldsymbol{y}|\boldsymbol{X}_{\mathrm{dom}},\boldsymbol{X}_{\mathrm{Tim}}\right)=\frac{L\left(\boldsymbol{y},\boldsymbol{X}_{\mathrm{dom}}\right)}{L\left(\boldsymbol{y},\boldsymbol{X}_{\mathrm{Tim}}\right)} \tag{9}$$

The reliability of this procedure was evaluated using the same Monte Carlo procedure as had been applied to the single-Gaussian modelling procedure (including using exactly the same random sample sets).

5.2. Results and discussion

The likelihood ratio obtained using the Hotelling’s T^2 procedure was 1×10^{10}, and the 5th and 1st percentiles from the Monte Carlo simulation were 9×10^{8} and 2×10^{8} respectively (see Fig. 6). We think it would be reasonable to report that on the basis of our analysis we are 99% certain that the likelihood ratio is at least 2×10^{8}, i.e., we are 99% certain that the acoustic properties of the disputed utterance are at least approximately two hundred million times more likely if the speaker had said “dom” rather than “Tim”. We think that this is a valid approach which results in little danger of us overstating the strength of the evidence.

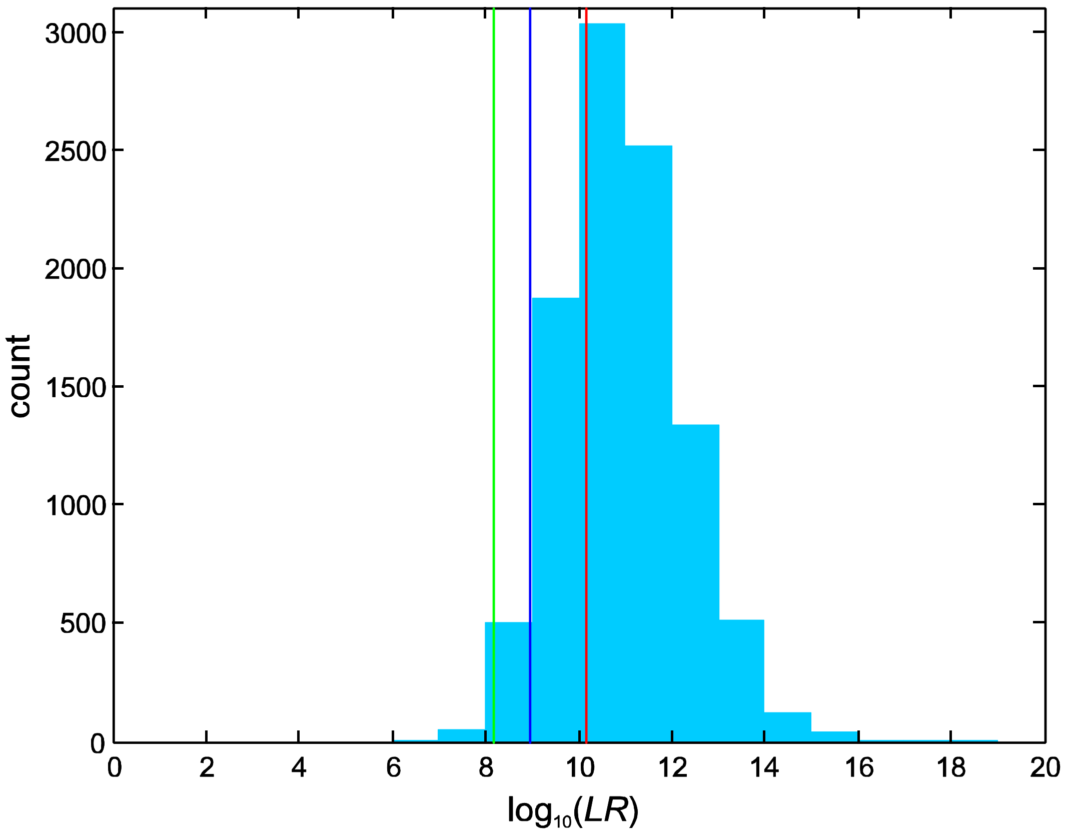

Fig. 6. Histogram of likelihood-ratio results from the Monte Carlo simulation using Hotelling's T^2 models. The vertical lines represent, from right to left, the likelihood ratio calculated using the original data (1×10^{10}), the 5th percentile (9×10^{8}), and the 1st percentile (2×10^{8}).

6. Third statistical analysis: Posterior-predictive-density analysis

The approach to evaluating likelihood ratios in the previous section attempted to deal specifically with the sampling uncertainty in the mean vectors and covariance matrices that arises from the fact that they are estimated from relatively small sets of training data. Some Bayesian statisticians may regard this approach as, at worst, being fundamentally at odds with the whole Bayesian paradigm, and at best, concentrating effort on what are essentially nuisance parameters. Nuisance parameters are parameters required in order to estimate the quantity of interest, but which are not themselves central to the question being addressed. The means and covariances of the Gaussian distributions used in the calculation of the numerator and denominator of the likelihood ratio fall into this category. A traditional approach to dealing with nuisance parameters is to "integrate them out". This is equivalent to averaging the statistic of interest over the distribution of the nuisance parameters. Taking a Bayesian approach to this integration yields the posterior-predictive-density distribution. The posterior-predictive-density distribution is the distribution that a new independent and identically distributed (i.i.d.) data point $\boldsymbol{y}$ would have, given a set of n existing i.i.d. observations $\boldsymbol{X} = [\boldsymbol{x}_i, ..., \boldsymbol{x}_n]$. Such a Bayesian approach has previously been proposed and demonstrated for automatic speaker recognition (Villabla and Brümmer, 2011).

6.1. Likelihood-ratio calculation

The quantity of interest in the present study is the likelihood ratio with the nuisance parameters $\boldsymbol{\mu}_{\mathrm{dom}}$, $\boldsymbol{\mu}_{\mathrm{Tim}}$, $\boldsymbol{\Sigma}_{\mathrm{dom}}$ and $\boldsymbol{\Sigma}_{\mathrm{Tim}}$ integrated out. That is we want the ratio of the posterior predictive density of $\boldsymbol{y}$ given $\boldsymbol{X}_{\mathrm{dom}}$ to the ratio of the posterior predictive density of $\boldsymbol{y}$ given $\boldsymbol{X}_{\mathrm{Tim}}$, Eq. 4. The posterior predictive densities will be calculated by combining prior distributions[8] and sample distributions.

Assuming that $\boldsymbol{X}_{\mathrm{dom}}$ and $\boldsymbol{X}_{\mathrm{Tim}}$ follow normal distributions $\boldsymbol{X} \sim N_k(\boldsymbol{\mu},\boldsymbol{\Sigma})$, and assuming conjugate priors such that (Eq. 10):

$$(\boldsymbol{\mu}|\boldsymbol{\Sigma}) \sim N_k\left(\boldsymbol{\mu}^{prior}, \tfrac{1}{q}\boldsymbol{\Sigma}^{prior}\right) \tag{10a}$$

$$\boldsymbol{\Sigma}^{-1} \sim W^{-1}\left(\boldsymbol{\beta}^{prior}, \alpha^{prior}\right) \tag{10b}$$

where the choice of q corresponds to our belief about the number of times more certain we are about the mean than the variance (q was set to 100, this is a typical value to use, its influence decreases rapidly as the amount of sample data increases, and in this case roving the value of q between 0.001 and 100 resulted in less than an order of magnitude change in the value of the calculated likelihood ratio), W^{-1} is an inverse Wishart distribution, $\boldsymbol{\beta}^{prior}$ is a $k \times k$ symmetric positive definite matrix, and α^{prior} is the degrees of freedom for the Wishart distribution (a constraint is $2\alpha^{prior} > k-1$, α^{prior} was set to the same number as the number of dimensions for the data, i.e., $\alpha^{prior} = k = 3$), then the posterior predictive density of $\boldsymbol{y}$ given $\boldsymbol{X}$ can be calculated as in Eq. 11.

$$f(\boldsymbol{y}|\boldsymbol{X}) = T_k\left(\boldsymbol{y}, \boldsymbol{\mu}^{post}, \boldsymbol{\Sigma}^{post}, \alpha^{post}\right) \tag{11a}$$

$$T_k\left(\boldsymbol{y}, \boldsymbol{\mu}^{post}, \boldsymbol{\Sigma}^{post}, \alpha^{post}\right) = c\left(1 + \frac{1}{\alpha^{post}}\left(\boldsymbol{y}-\boldsymbol{\mu}^{post}\right)' \boldsymbol{\Sigma}^{post}\left(\boldsymbol{y}-\boldsymbol{\mu}^{post}\right)\right)^{-\frac{\alpha^{post}+k}{2}} \tag{11b}$$

[8] These priors, used in calculating the posterior predictive distributions of the acoustic features which are in turn used to calculate the numerator and denominator of the likelihood ratio, should not be confused with the prior odds with respect to the prosecution and defence hypotheses which the trier of fact should theoretically combine with the likelihood ratio in order to arrive at the posterior odds for the hypotheses.

$$c = \frac{\Gamma\left(\frac{\alpha^{post} + k}{2}\right)}{\Gamma\left(\frac{\alpha^{post}}{2}\left(\alpha^{post}\pi\right)^{\frac{k}{2}}\right)} \left|\boldsymbol{\Sigma}^{post^{-1}}\right|^{\frac{1}{2}} \qquad \text{(11c)}$$

$$\boldsymbol{\mu}^{post} = \frac{n\bar{\boldsymbol{x}} + q\boldsymbol{\mu}^{prior}}{n+q} \qquad \text{(11d)}$$

$$\boldsymbol{\Sigma}^{post^{-1}} = \frac{n+q}{n+q+1}\alpha^{post}\boldsymbol{\beta}^{post^{-1}} \qquad \text{(11e)}$$

$$\boldsymbol{\beta}^{post} = \boldsymbol{\beta}^{prior} + \frac{n}{2}\boldsymbol{S} + \frac{1}{2}\frac{nq}{n+q}\left(\bar{\boldsymbol{x}} - \boldsymbol{\mu}^{prior}\right)\left(\bar{\boldsymbol{x}} - \boldsymbol{\mu}^{prior}\right)' \qquad \text{(11f)}$$

$$\boldsymbol{S} = \tfrac{1}{n}\sum\nolimits_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})' \qquad \text{(11g)}$$

$$\alpha^{post} = \alpha^{prior} + \frac{n}{2} + \frac{1}{2}\left(k-1\right) \qquad \text{(11h)}$$

where $T_k(\boldsymbol{y}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha)$ is the multivariate t distribution for k dimensions (Bernardo and Smith, 1994, p. 139], which was calculated using R package `mvtnorm` (Genz and Brents, 2009; Genz et al., 2012). Note that $\boldsymbol{S}$ here is the $k \times k$ maximum-likelihood sample covariance matrix, which differs from Eq. 3 and Eq. 6 where $\boldsymbol{S}$ was the least-squares sample covariance matrix.

The prior distributions for the posterior-predictive-density analysis were based on a published summary of VOT measurements, and on formant measurements from a database of recordings of Swedish speakers. Lundeborg et al. (2012) reported VOT measurements on Swedish plosives. They found no significant difference between men and women and therefore reported the pooled results. Word initial [d] and [t^{h}] preceding a mixture of different vowels were reported to have means (and standard deviations) of −45 (56) and +54 (9) milliseconds respectively. Although the pooled data across multiple speakers and vowel contexts may have a more normal distribution than that observed for the speaker in our case, we remain sceptical that the data are actually normally distributed (there is even the suggestion in Lundeborg et al. (2012) that the distribution for [d] was bimodal); however, these still represented the best information available to inform a prior distribution for VOT in the Swedish words "dom" and "Tim". We converted the mean values from milliseconds to the transformed space (see §4.1), and used standard deviations which had the same standard-deviation-to-mean ratios in the transformed space as they had in the milliseconds space. Prior distributions for F1 and F2 were based on formant measurements on [ɔ] and [ɪ] tokens in a subsection of the Swedia database (Lindh and Eriksson, 2009). The Swedia database was collected using high-quality lapel microphones and recorded on a digital audio tape (DAT) recorder with a sampling rate of 16 kHz and 16 bit quantisation. The subsection of the database consisted of 202 speakers who were all the female speakers from the middle region of Sweden, the same region that the speaker of the disputed utterance came from. These data came from two age groups, 20–30 year olds and 55–75 year olds. The age of speaker of the disputed utterance was not provided to JL in 2008 and we are not able to ascertain it now. The means ($\boldsymbol{\mu}^{prior}$) and standard deviations of the prior distributions are given in Table 2. The diagonals of the covariance matrices ($\boldsymbol{\beta}^{prior}$) for the prior distributions were the squares of the standard deviations given in Table 2, and the off-diagonal elements were set to zero.

Table 2. Means (and standard deviations) of the prior distributions used for the posterior-predictive-density analysis

	VOT [transformed]	F1 [Hz]	F2 [Hz]
"dom"	−1.267 (1.573)	604 (65)	1043 (124)
"Tim"	+1.312 (0.225)	464 (76)	2093 (569)

6.2. Results and discussion

The likelihood ratio calculated via the posterior-predictive-density analysis was 2×10^{11}. This is not far from the value obtained via the Hotelling's T^2 analysis.

Since this is a Bayesian analysis in which the small sample size was taken into account in calculating the posterior predictive distribution, and the mean vector and covariance matrix nuisance parameters were integrated out, a subsequent attempt to assess the effect of the small sample sizes for the training data on reliability would not be theoretically justified.

The use of informative priors was a substantial contributor to the likelihood-ratio value obtained. An additional analysis was performed using "uninformed" priors of zero for every mean in each prior mean vector and the identity matrix for each prior covariance matrix. The resulting likelihood-ratio value was 3×10^{6}.

7. General discussion and conclusion

We have demonstrated a disputed-utterance analysis based on relevant data, quantitative measurements, and statistical models in a case in which, as is typical in forensic speech science (and likely many other branch of forensic science), the amount of data for training models is small. This case may have been extreme in that the data point from the disputed utterance was far out on the tail of a modelled distribution trained using very little data, but it clearly illustrates a problem which always exists to some degree. The importance of this problem may not be immediately obvious in borderline cases. We have attempted to illustrate the importance of not simply taking the likelihood ratio estimate from a standard model at face value, and instead considering whether the value obtained can be supported by the amount of data used to train the model, and considering whether there may be some more appropriate model which can deal with the challenges posed by the available data, or at the very least assessing the reliability of the standard model. Not taking this issue into consideration could lead to hugely overestimating the strength of the evidence in one direction or the other and providing misleading information to the trier of fact.

Decisions as to the appropriateness of any particular model and the choice of one model over another should ideally be made before a likelihood ratio is calculated for the questioned data from the actual forensic case. Before the calculation of the latter, the validity and reliability of any proposed forensic-analysis systems should be empirically tested using test data which reflect the conditions of the case under investigation, and decisions should be made on the basis of these test results. Unfortunately, a problem due to small *n* and tails may not become apparent until one has actually examined or even calculated a likelihood ratio for the questioned data from the case. One would then have to be very careful to justify choosing a different model on the grounds that the initial model employed was faulty and that the other model ameliorates this fault. One would have to be careful to avoid choosing a different model, or giving the impression of choosing a different model, on the basis of the latter giving more favourable results for either the prosecution or for the defence.

For the disputed-utterance case used as an example in the present paper an analysis based on Hotelling's T^2 statistic and a posterior-predictive-density analysis were proposed as more appropriate than an analysis based on multivariate Gaussians. The Hotelling's T^2 approach attempts to take account of the sampling uncertainty of the mean vectors and covariance matrices, and the posterior-predictive-density analysis calculates posterior predictive distributions on the basis of prior distributions and sample data, and integrates out the values of the mean vectors and covariance matrices as nuisance parameters. The Hotelling's T^2 approach was possible and the posterior-predictive-density analysis tractable in this case because the data were assumed to come from a normally distributed population, and also for the latter because relevant information about the variables in this case was available to inform prior distributions – the Hotelling's T^2 approach will be more generally applicable because it can be applied in cases where data to inform priors are not readily available. The scarcity of data essentially forced us to make the normality assumption because fitting more complex models using these amounts of data was a priori not considered appropriate.

The particular solutions employed in the present paper may not be available in situations using more complex models, although a priori one would be expected to have more data before attempting to fit more complex models (such as Gaussian mixture models or kernel density models). Testing of reliability using Monte Carlo simulation could still be a good way of investigating the reliability of more complex models. It could also serve as a diagnostic of whether the complexity of the model can reasonably be supported given the amount of data available. Given a fixed amount of data a potentially high bias but lower variance model may be preferred over a potentially low bias but high variance model. Monte Carlo simulations can provide empirical assessment of reliability related to the interaction between the amount of data and the complexity of the model. At the very least the estimated reliability of the system should be reported to the trier of fact so that they do not mistakenly assume that the single-value likelihood-ratio output of the system is a highly precise number if there is empirical evidence to suggest that it is in fact not.

Acknowledgments

Morrison’s contribution to this research was supported by the Australian Research Council, Australian Federal Police, New South Wales Police, Queensland Police, National Institute of Forensic Science, Australasian Speech Science and Technology Association, and the Guardia Civil through Linkage Project LP100200142.Unless otherwise explicitly attributed, the opinions expressed in this paper are those of the authors and do not necessarily represent the policies or opinions of any of the above mentioned organizations. Curran’s contribution to this research was supported by grant 2011-DN-BX-K541 from the US National Institute of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the US Department of Justice.

References

Anderson, N., 1978. On the calculation of filter coefficients for maximum entropy spectral analysis, in: Childers, D.G. (Ed.), Modern Spectrum Analysis. IEEE Press, New York, 1978, pp. 252–255.

Bernardo, J.M., Smith, A.F.M., 1994. Bayesian Theory. Wiley, Chichester, UK.

Boersma, P., Weenik, D., 2008. Praat: doing phonetics by computer, ver. 5.0.03. Stable URL http://www.praat.org/

Champod, C., Evett, I.W., 2000. Commentary on Broeders (1999) ‘Some observations on the use of probability scales in forensic identification’. Forensic Ling. 7, 238–243.

Curran, J.M., 2005. An introduction to Bayesian credible intervals for sampling error in DNA profiles. Law Prob. Risk. 4, 115–126. doi:10.1093/lpr/mgi009

Curran, J.M., Buckleton, J.S., Triggs, C.M., Weir, B.S., 2002. Assessing uncertainty in DNA evidence caused by sampling effects. Sci. Just. 42, 29–37. doi:10.1016/S1355-0306(02)71794-2

Curran, J.M., Triggs, C.M., Almirall, J.R., Buckleton, J.S., Walsh, K.A.J., 1997a. The interpretation of elemental composition from forensic glass evidence: I. Sci. Just., 37, 241–244. doi:10.1016/S1355-0306(97)72197-X

Curran, J.M., Triggs, C.M., Almirall, J.R., Buckleton, J.S., Walsh, K.A.J., 1997b. The interpetation of elemental composition from forensic glass evidence: II. Sci. Just., 37, 1997, 245–249. doi:10.1016/S1355-0306(97)72198-1

Genz, A., Bretz, F., 2009. Computation of multivariate normal and *t* probabilities: Lecture notes in statistics, vol. 195. Springer-Verlag, Heidelberg, Germany. doi:10.1007/978-3-642-01689-9

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., Hothorn, T., 2012. mvtnorm: Multivariate normal and *t* distributions, R package ver. 0.9-9994. Stable URL: http://CRAN.R-project.org/package=mvtnorm

Hotelling, H., 1931.The generalization of Student's ratio. Annals Math. Stat. 2, 360–378. http://www.jstor.org/stable/2957535

Kaye, D.H., 2009.Trawling DNA databases for partial matches: What is the FBI afraid of? Cornell J. Law Public Policy. 19, 145–171.

Lagarias, J.C., Reeds, J.A., Wright, M.H., Wright, P.E., 1998. Convergence properties of the Nelder-Mead simplex method in low dimensions. SIAM J. Optimization. 9, 112–147. doi:10.1137/S1052623496303470

Lindh, J., Eriksson, A., 2009. The SweDat project and Swedia database for phonetic and acoustic research, in: Proceedings of the 5th IEEE International Conference on e-Science. IEE Computer Society, Los Alamitos, CA, pp. 45–49. doi:10.1109/e-Science.2009.15

Lundeborg, I., Larsson, M., Wiman, S., McAllister, A.M., 2012. Voice onset time in Swedish children and adults. Logopedics Phoniatrics Vocology. 37, 117–122. doi:10.3109/14015439.2012.664654

MathWorks Inc., 2010. Matlab, software release 2010a. MathWorks Inc., Nantick, MA.

Morrison, G.S., 2011. Measuring the validity and reliability of forensic likelihood-ratio systems. Sci. Just. 51, 91–98. doi:10.1016/j.scijus.2011.03.002

Morrison, G.S., Hoy, M., 2012. What did Bain really say? A preliminary forensic analysis of the disputed utterance based on data, acoustic analysis, statistical models, calculation of likelihood ratios, and testing of validity, in: Proceedings of the 46th Audio Engineering Society (AES) Conference on Audio Forensics: Recording, Recovery, Analysis, and Interpretation, Denver, CO, pp. 203–207.

Nordgaard, A., Ansell, R., Drotz, W., Jaeger, L., 2012. Scale of conclusions for the value of evidence, Law Prob. Risk. 11, 1–24. doi:10.1093/lpr/mgr020

R Development Core Team, 2013. R: A language and environment for statistical computing, ver. 2.15.2, R Foundation for Statistical Computing, Vienna, Austria. Stable URL http://www.R-project.org/

Stoel, R.D., Sjerps, M.J., 2012. Interpretation of forensic evidence, in: Roeser, S., Hillerbrand, R., Sandin, P., Peterson M. (Eds.), Handbook of Risk Theory: Epistemology, Decision Theory, Ethics, and Social Implications of Risk. Springer Netherlands, Dordrecht, The Netherlands, pp. 135–158. doi:10.1007/978-94-007-1433-5_6

Student [Gosset, W.S.], 1908. The probable error of a mean. Biometrika. 6, 1–25. doi:10.1093/biomet/6.1.1

Villalba, J., Brümmer, N., 2011. Towards fully Bayesian speaker recognition: Integrating out the between-speaker covariance, in: Proceedings of Interspeech 2011, The 12th Annual Conference of the International Speech Communication Association, Florence, Italy, pp. 505–508.

Zhang, C., Morrison, G.S., Ochoa, F., Enzinger, E., 2013. Reliability of human-supervised formant-trajectory measurement for forensic voice comparison. J. Acoust. Soc. Amer. 133, EL54–EL60. doi:10.1121/1.4773223