REFERENCE SAMPLE SIZE AND THE COMPUTATION OF NUMERICAL LIKELIHOOD RATIOS USING ARTICULATION RATE

VINCENT HUGHES[1], ASHLEY BRERETON[2] & ERICA GOLD[1]

[1]University of York

[2]University of Liverpool

Abstract

This paper explores the effects of variability in the amount of reference data used in quantifying the strength of speech evidence using numerical likelihood ratios (LRs). Monte Carlo simulations (MCS) are performed to generate synthetic data from a sample of existing raw local articulation rate (AR) data. LRs are computed as the number of reference speakers (up to 1000), and the number of tokens per reference speaker (up to 200) is systematically increased. The distributions of same-speaker and different-speaker LRs and system performance (log LR cost (C_{llr}) and equal error rate (EER)) are assessed as a function of the size of the reference data. Results reveal that LRs based on AR are relatively robust to small reference samples, but that system calibration plays an important role in determining the sensitivity of the LRs to sample size.

1. Introduction

Forensic voice comparison (FVC) commonly involves the analysis of the speech patterns in a recording of an unknown criminal's voice and a recording of a known suspect's voice with regard to the competing propositions of the prosecution (same-speaker) and defence (different-speakers). Across forensic disciplines the likelihood ratio (LR) is increasingly accepted as the "logically and legally correct" (Rose & Morrison 2009: 143) framework for the gradient assessment of such comparison evidence (Robertson & Vignaux 1995, Aitken & Taroni 2004). The outcome is gradient in that it not only indicates whether the evidence supports the prosecution or defence, but also provides an estimate of the relative strength of the support. The LR involves an assessment of both the similarity between the evidential samples and the typicality of within- and between-speaker variation in the relevant population (Aitken & Taroni 2004). The assessment of typicality is an essential element of the LR framework, since it allows the expert to estimate the probability of finding the speech evidence assuming the offender is another random member of the relevant population (for issues relating to the definition of the relevant population in speech see Hughes & Foulkes 2012, Morrison et al. 2012). Quantification of typicality is conducted using a set of representative reference data.

However, a practical concern in FVC is the amount of reference data needed to ensure a meaningful estimate of the strength of evidence. Consistent with general sampling principles, the answer is determined by "how accurate the experimenter wishes the estimate to be" (Wackerly *et al.* 2008: 421). That is, the greater the amount of representative data, the more precise the model of the population (Rose 2012). A small number of studies have investigated this issue for FVC. Ishihara and Kinoshita (2008) describe a *population size effect* when computing LRs based on distributional characteristics of fundamental frequency (f0) using small numbers of reference speakers. In their study, same-speaker (SS) and different-speaker (DS) LRs were overestimated by up to 1000 times using 10 reference speakers compared with a baseline using 120 speakers. Hughes and Foulkes (2012) investigated the effect of

different numbers of reference speakers (up to 120) and numbers of tokens per reference speaker (up to 13) on the distribution of LR scores and system performance (i.e. how well the system separates SS pairs from DS pairs) using polynomial representations of formant trajectories from spontaneous GOOSE (/uː/) vowels. Consistent with Ishihara and Kinoshita (2008), scores were found to be more stable and system performance improved with greater than 20 reference speakers. Further, DS LR scores were shown to be very sensitive to the number of tokens per reference speaker, displaying no stability between the 2- and 13-token conditions. Same-speaker pairs were more robust to variation in the amount of data per reference speaker.

The findings of these studies suggest that LRs are generally unstable and misrepresentative when small numbers of reference speakers and small numbers of tokens are used. The results reflect an imprecise estimation of the variation in the population when using small amounts of reference data. That is, the addition of speakers or tokens to a small sample affects the distributions of within- and between-speaker values more than the addition of speakers or tokens to a much larger set of existing data. Consistent with the law of diminishing returns, with a given amount of data the addition of more representative data will have little effect on the overall distribution. Very little work in FVC has considered such an upper limit at which the inclusion of more reference data has an asymptotic effect on LRs and system performance. Yet, the efficiency and cost-effectiveness of the numerical LR approach is, at least to some extent, dependent on knowing how much reference data is enough to produce robust estimates of the strength of evidence.

The relative lack of research in this area is in part due to a lack of sufficiently large amounts of raw data. Monte Carlo simulations (MCS) offer a potential solution to this problem. They involve generating synthetic values from known properties of the distributions of within- and between-speaker variation of a given variable in a given population. Synthetic data can be built from population statistics presented in previous research (e.g. mean and standard deviation (SD) when the distribution is assumed to be normal, although the assumption of normality is not a pre-requisite; as in Rose 2012) or using some existing set of raw data. Whilst Monte Carlo simulations avoid the need for extremely large amounts of raw data, there is a non-trivial *a priori* assumption that the true distribution of the variable in the population is known (or can be well estimated). This is because the distribution of the resulting synthetic data is defined by the properties of the input. In this respect, Monte Carlo simulations are not predictive. Despite having knowledge of the distribution of a variable in a population, MCS are still necessary in investigating how LR performance is affected by sample size since the Multivariate Kernel Density (MVKD) procedure proposed by Aitken & Lucy (2004) includes N speakers as part of the kernel density estimation and N tokens per speaker in determining suspect and offender variance and covariance matrices (Aitken & Lucy 2004: 12-13, Rose 2012, 2013: 94).

Initial exploration of Monte Carlo methods for FVC is offered by Rose (2012), who synthesised F1, F2 and F3 midpoint values for Australian English /aː/ for up to 10,000 speakers based on the distribution of values in Bernard (1967). Using Lindley's (1977) procedure to investigate individual (univariate) formant performance and the MVKD (Aitken & Lucy 2004) formula to investigate combined formant performance, SS LRs were computed based on real case data and assessed as a function of the number of reference speakers. Output was compared against the *true* LR, which is defined as the LR computed using the maximum amount of reference data. As such, the *true* LR is based on the most precise estimation of within- and between-speaker variation in the relevant population. The magnitude of the LR was found to be equivalent to the *true* LR (based on 10,000 speakers) with the inclusion of 30 or more reference speakers. However, Rose's study is limited by the

lack of DS comparisons, since there is no assessment of performance as a function of sample size or calibration based on weights from a development set. The procedures for modelling potential correlations between formants are also not made explicit.

The present study builds on Rose (2012) by using Monte Carlo simulations to investigate the effect of reference sample size on the outcome of numerical LRs based on an analysis of local articulation rate (AR). A set of existing raw data of 59 speakers is firstly analysed with regard to how precisely it estimates patterns in the relevant population, defined in terms of regional background, age, sex and class. MCS are then used to generate normal distributions for 941 synthetic speakers from which up to 200 tokens per speaker are generated. LRs are computed for development (20 speakers) and test (20 speakers) sets extracted from the raw data and the distributions of LRs and system validity are assessed as a function of sample size. Following Rose (2012), results are compared against the *true* LR performance based on the largest set of reference data (up to 1000 speakers, up to 200 tokens per speaker).

2. Methods

2.1. Data

The data consisted of local articulation rate (AR) measurements, quantified as the number of phonological syllables per second within multiple memory stretches (Jessen 2007). AR was chosen primarily because it is a simple, univariate variable, which can be synthesised relatively straightforwardly. However, the extent to which the findings based on AR can be generalised to other variables is limited by the lack of inherent speaker-discriminant value. This is confirmed by a small variance ratio (calculated as the between-speaker SD^2 divided by the within-speaker $mean^2$; Rose *et al.* 2006) of 0.326, which suggests that within-speaker variability in AR is generally higher than between-speaker variation. Whilst a multivariate variable with good speaker discriminatory power would be preferable for investigating sample size, unidimensionality makes AR a good candidate for the preliminary exploration of Monte Carlo methods for speech data. Local AR was chosen over global AR (across a recording) since multiple tokens are needed to estimate within-speaker variation when computing LRs. Further, local AR is a more meaningful forensic resource since it captures individual variability across utterances (Miller *et al.* 1984).

The data were collected as part of Gold (in progress). Local AR was extracted for 100 speakers from Task 2 recordings of the DyViS database (Nolan *et al.* 2009). The DyViS speakers are all young (aged 18-25), male speakers of Standard Southern British English (SSBE) from the University of Cambridge and as such are considered sociolinguistically homogeneous. Task 2 involves a telephone conversation with a mock accomplice, who is demographically matched to the subjects. Since raw AR values are extracted from a single session, the data necessarily provide an under optimistic estimation of the within-speaker variation found in real casework (Rose 2011; Morrison *et al.* 2012). It is expected that LR performance will be poorer in realistic, non-contemporaneous conditions.

Gold (in progress) identified the onset and offset of between 26 and 32 memory stretches per speaker, defined as a period of "fluent speech containing a number of syllables that can easily be retained in short-term memory" (Jessen 2007: 54). Memory stretches were chosen as a unit for measuring AR based on Gold (in progress), who found no significant differences in performance compared with inter-pause stretches. Further, Gold (in progress) suggests that memory stretches are better for FVC as they can be extracted efficiently without requiring

precise measurement of individual pauses. Following Künzel (1997), "fluent speech" was defined as the absence of pauses, hesitation phenomena and repair processes. Each token was then calculated as the total number of phonological syllables divided by the total duration (in seconds) of the memory stretch. Memory stretches generally contained between 7 and 11 syllables.

For each speaker, the first 26 tokens were used in the analysis (the largest number of tokens shared by all speakers). Mean and SD of AR values were calculated by-speaker and converted to z-scores to identify univariate outliers. On the basis of an outlying SD with z > $\pm$3.29 (p < 0.01) (Tabachnick & Fidell 2007: 73), one speaker was removed from the analysis. With this speaker removed z-scores were again calculated, but no values exceeded the $\pm$3.29 threshold. Of the remaining 99 speakers, 20 were selected at random as development data and a further 20 as test data. The remaining 59 speakers were used as reference data from which synthetic reference speakers and tokens were generated.

2.2. Modelling

In this study, Aitken and Lucy's (2004) MVKD formula was used to compute LRs, which models within-speaker variance with a normal distribution and the between-speaker distribution with a kernel density made up of Gaussians from each reference speaker (Morrison 2011a: 243). Using the modelling procedures in the MVKD formula as a starting point, a two-stage process for synthesising data was developed. Firstly, normal distributions for each synthetic speaker were generated by sampling synthetic means and SDs from the raw data. This is because MVKD models within-speaker variation with an assumption of normality. From the synthetic normal distributions $N(\mu, \sigma)$ a second round of simulations were conducted to generate synthetic tokens for each of the synthetic speakers. Prior to conducting MCS, two issues with the raw data were addressed. The first relates to the choice of distribution from which synthetic mean and SD values are sampled. Figures 1 displays the histograms of raw mean and SD values by-speaker fitted with normal distributions.

Skew and kurtosis were calculated to assess how well normality models the data. Following Tabachnick and Fidell (2007: 79), skew was analysed by dividing the skewness (S) by the standard error (S_s), defined as $S_s = \sqrt{6/N}$ where N is the number of observations (in this case 59), to give a z-score (where z > $\pm$1.96 = p < 0.05 and z > $\pm$3.29 = p < 0.01). Skew was non-significant for both means and SDs (p > 0.4). Kurtosis was analysed by dividing the kurtosis statistic by twice the standard error (Tabachnick & Fidell 2007) to generate a z-score. For both sets of data, kurtosis was also found to be non-significant (p > 0.24). Given the statistical assessment of normality and visual inspection of Figures 1 and 2, the normal distribution fits the data sufficiently well.

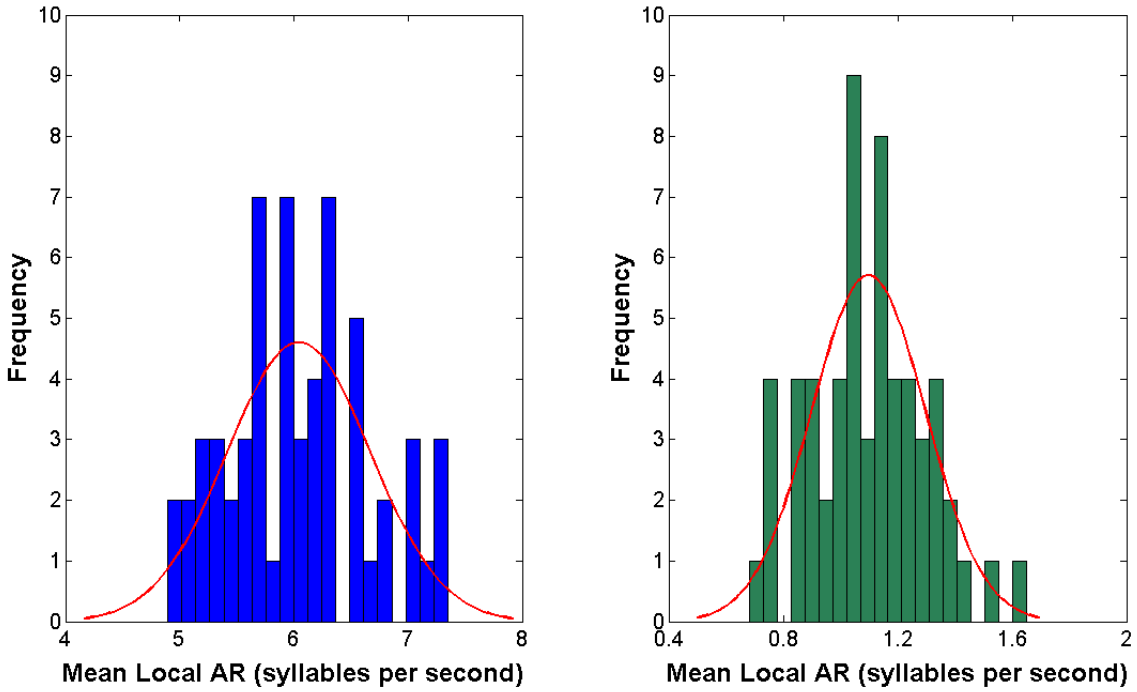

Figure 1: Histograms of mean (left) and SD (right) of AR by speaker fitted with normal distributions

The second issue relating to the raw data is whether the 59 raw speakers provide a sufficiently precise estimate of patterns in the relevant population. To assess how well the sample of raw data approximates the distribution of values in the relevant population, it is necessary to know *a priori* the shape of that distribution. In the absence of this knowledge, the alternative is to assess how the distributions of means and SDs vary as data is added. Independent samples t-tests were calculated for means and SDs using the values for all 59 speakers compared against values for 10 speakers. Values for each speaker were then added consecutively to the smaller data set and the t-test re-run. Welch's Correction (Welch 1947) was applied to account for unequal sample size and an assumption of unequal variance across the sets of data. The results are analysed with regard to the *p*-value where $p = 1$ is equivalent to the two samples having the same normal distribution.

Figure 2 shows that there is no significant difference for AR means in the distribution of values for as few as 10 speakers compared with the distribution using all 59 speakers. Despite an initial dip with small numbers of speakers, *p* increases towards 1 after 25 speakers. For SDs, *p* is relatively low (0.1) with small numbers of speakers, although at no point is the difference between the distributions significant (at either the 1% or 5% levels). Whilst predictably *p* approaches 1 as the number of speakers increases, there is considerable similarity in the distribution of SDs after 40 speakers. Further, the means and SD are consistent with expectations about the range of potential within- and between-speaker variation reported in Goldman-Eisler (1968). As such it is considered that the distributions based on 59 speakers provide a sufficiently precise estimate of the relevant population.

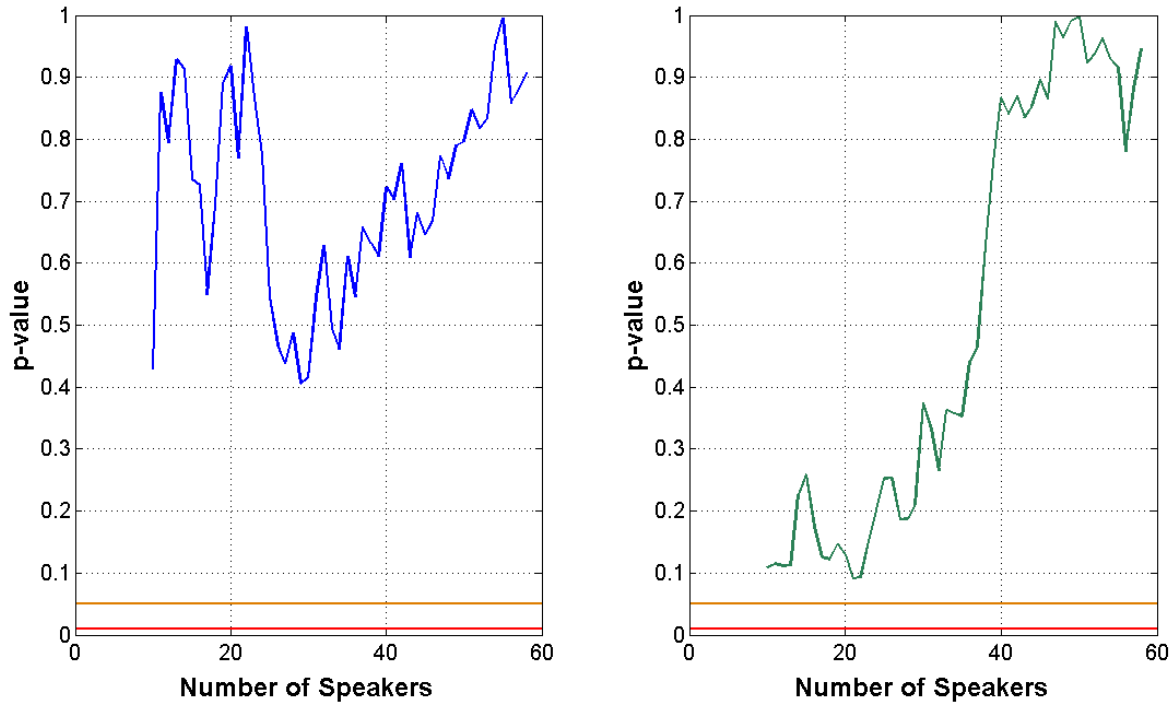

Figure 2: p-values based on t-tests comparing the distributions of means (left) and SDs (right) for the number of speakers on the x axis against that with all 59 raw speakers with 1% (red) and 5% (orange) significance marked

2.3. Monte Carlo simulations (MCS)

The following sections explain the procedures used for creating synthetic tokens of local AR for synthetic speakers based on the set of raw data.

2.3.1. Generating synthetic mean local AR

Mean local AR is denoted by *x*, where x_i is a value for a single speaker ($_i$ is speaker number). Based on the testing of normality in §2.2, the distribution of raw x_i values is converted to a normal probability density function (PDF) with mean of 0 and SD of $\frac{1}{\sqrt{2}}$, $N(0, \frac{1}{2})$, by applying the transformation:

$$z = \frac{(x - \mu_x)}{\sqrt{2}\sigma_x}, \tag{1}$$

where μ_x is the mean of the raw means and σ_x is the SD. This transforms values within the raw *x*-space to normalised values within the *z*-space where the aim of the Monte Carlo simulations is to generate synthetic z_i values from the preferentially scaled PDF. This is done using the inverse of the cumulative distribution function (CDF). The CDF uses integration to calculate the area under the PDF between $-\infty$ and z_i such that:

$$CDF(z) = \int_{-\infty}^{z} N(z, 0, \frac{1}{2})\, dz. \tag{2}$$

Given that the normal distribution is so widely used, a special function called the *error function* (erf) (Wang *et al.* 1989: 333) has been assigned to the integral ($\int$) meaning that it is possible to generate a CDF based on a normal PDF in the following way:

$$\int_{-\infty}^{z} N\left(z, 0, \frac{1}{2}\right) dz = CDF(z) = \frac{1 + \text{erf}(z)}{2} \tag{3}$$

where:

$$\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_{0}^{z} e^{-t^2} dt \tag{4}$$

With the CDF defined as above, normally distributed z_i values can be synthesised using a random variable $Z_i \in [0, 1]$ (i.e. a random number between 0 and 1). Using the inverse CDF ($CDF^{-1}(z)$), a single synthetically generated z_i value is defined as:

$$CDF\left(CDF^{-1}(z)\right) = z = \frac{1 + erf\left(CDF^{-1}(z)\right)}{2} \tag{5}$$

$$\therefore \quad 2z - 1 = erf(CDF^{-1}(z)) \tag{6}$$

$$CDF^{-1}(z) = erf^{-1}(2z - 1). \tag{7}$$

As demonstrated in Figure 3, using a random value for Z_i and with explicit knowledge of the inverse CDF, a synthetic z_i can be generated in the following way:

$$CDF^{-1}(Z_i) = z_i. \tag{8}$$

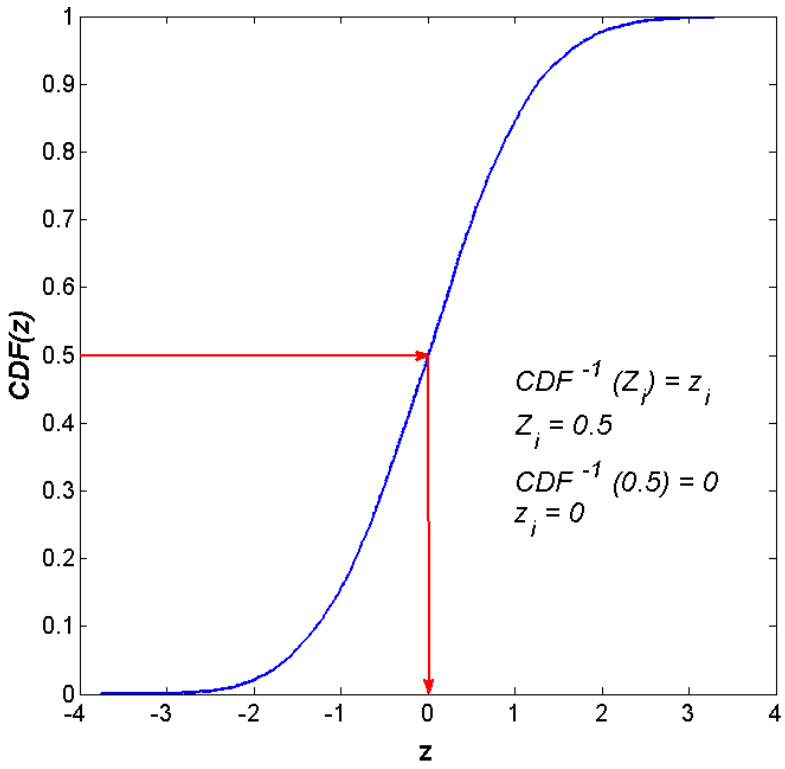

Figure 3: Example of the inverse CDF of mean local AR used to generate a synthetic z_i of 0 based on a random Z_i of 0.5 (z_i = 0 equates to x_i = 6.044 (i.e. the mean of the raw data))

Synthetic z_i values are then transformed back into the linguistically meaningful x-space by:

$$x_i = (\sqrt{2}\sigma_x * z_i) + \mu_x, \tag{9}$$

and used as the mean value for the normal distribution of a single synthetic speaker.

This process is repeated over a number of simulations (n). By the law of large numbers (Wackerly *et al.* 2008: 451), the distribution of $z = (z_1, z_2, \ldots, z_n)$ will converge on $N(0, \frac{1}{2})$ as $n \to \infty$. Therefore, with large n the synthetically generated values will have approximately the same normal distribution as the raw values.

2.3.2. Generating synthetic SD of local AR

The SD of local AR is denoted by y such that y_i is the SD for a single speaker. To generate synthetic y_i values, it is necessary to account for the correlation between the means and SDs in the raw data. Figure 4 reveals a significant (Pearson's r = 0.3964; p = 0.0019), positive correlation such that speakers with higher average AR generally display greater within-speaker variability. Potentially, this is because speakers with higher mean AR are able to exploit a wider range of variability, particularly in higher rates. Since the mean and SD are seemingly not independent a further (simple) projection was incorporated into the simulation of SDs. Rather than sampling from a normal PDF based on the mean and SD of the raw SDs (as with the raw means), $N(a\,x_i + b, \beta)$ was used where the linear trend line determines the mean ($ax_i + b$) and variance around the trend line (residuals) determines the SD (β) (see Figure 4). The mean of the normal distribution from which synthetic SD values are generated, therefore varies as a function of the associated synthetic mean value (x_i).

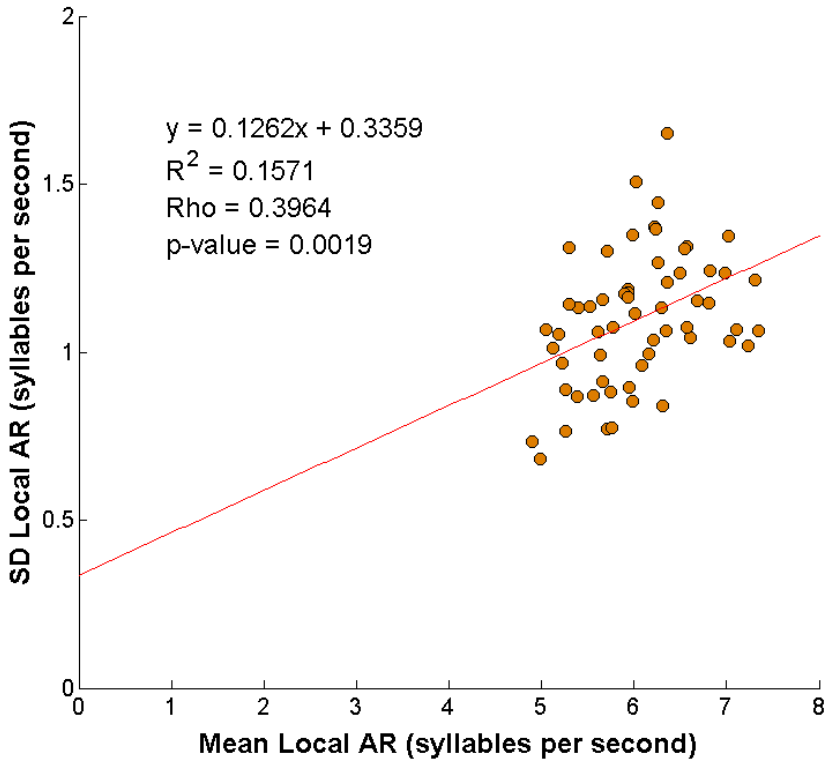

Figure 4: Mean local AR plotted against SD of local AR (syllables/ second) for each of the 59 raw speakers

Following the same procedure as before, synthetic y_i values were generated by converting $N(a\,x_i + b, \beta)$ to a normal PDF for each synthetic x_i. Again, the inverse CDF was used to

transform a random variable $Z^*_i \in [0, 1]$ into normalised z^*_i values, before transforming back to the *y*-space. The synthetic mean and SD values represent the normal distribution $N(x_i, y_i)$ for a new synthetic speaker. From this distribution, individual local AR tokens were synthesised using the same procedures as above. The process of generating synthetic means and SDs was performed 941 times. The synthetic speakers were pooled with the existing 59 raw speakers to create a reference sample of up to 1000 speakers. For the synthetic speakers, up to 200 tokens per speaker were generated. For each of the 59 raw speakers, MCS based on the mean and SD of the 26 raw tokens per speaker were used to generate an additional 174 tokens per speaker.

2.3.3. The synthetic data

The distributions of means and SDs in the raw data, synthetic data and all reference data combined (raw + synthetic) based on 26 tokens per speaker were compared to assess how well MCS approximates patterns in the raw data. Table 1 reveals minimal difference in the mean of the means (μ_x), with the raw data displaying a μ_x 0.023 higher than the synthetic data. The SD of the means in the synthetic data is higher than that in the raw data, although the difference is again negligible (0.015). *p*-values were generated from a comparison of the raw data and the synthetic data, as well as the raw data and all of the reference data, using independent t-tests. The differences between distributions were found to be non-significant, with *p* approaching 1 in both cases (Table 1).

	Mean	**SD**	**t-test (*p*-value)**
Raw data (59 speakers)	6.044	0.627	-
Synthetic data (941 speakers)	6.021	0.642	0.7954
Pooled data (1000 speakers)	6.023	0641	0.8065

Table 1: Mean and SD of mean local AR (syllables/ sec) for the raw data, synthetic data and all reference data

There are marginal differences in the distributions of SD (*y*) values, with μ_y 0.0029 higher for the raw data than for the synthetic data (Table 2). The differences between the sets in terms of σ_y are also marginal with SD in the raw data 0.008 greater than in the synthetic data. Again, paired independent t-tests were performed using the raw data and synthetic data, and the raw data and all reference data combined. In both cases, the differences were non-significant with *p*-values much closer to 1 than for the means.

	Mean	**SD**	**t-test (*p*-value)**
Raw data (59 speakers)	1.098	0.199	-
Synthetic data (941 speakers)	1.095	0.191	0.8989
Pooled data (1000 speakers)	1.095	0.191	0.9049

Table 2: Mean and SD of SD of local AR (syllables/ sec) for the raw data, synthetic data and all reference data

2.4. The present study

A MatLab implementation (Morrison 2007) of Aitken and Lucy's (2004) MVKD formula was used to compute raw LRs. For both the development and test sets independently, 20 SS and 380 DS LRs were computed as the number of speakers in the reference data was systematically increased by 1 starting with 10 and ending with 1000. To test the effect of the number of reference speakers, only the first 26 tokens per speaker were included. Using a random reference sample of 200 speakers, LRs were also computed as a single token per speaker was added to the reference data up to a maximum of 200 tokens. For both experiments, LR scores were transformed using a base-10 logarithm to account for the skew in the distribution of raw LRs. Using log LRs, zero represents the threshold, whereby positive values offer support for the prosecution (same-speaker) and negative values offer support for the defence (different-speaker). The magnitude of the log LR indicates the strength of the support for prosecution of defence.

For the test set, $\log_{10}$ LRs were calibrated based on weights generated from scores for the development set. Calibration was performed using a robust implementation (Morrison 2009) of Brümmer's (2007) logistic-regression-based procedure (Morrison 2013). Following Rose (2012), the effects on log LRs of the number of speakers and tokens per speaker are presented in the form of boxplots, which include the median, interquartile range (1^{st} to 3^{rd} quartile) and overall range (including outliers). The magnitude of LR output is assessed with reference to Table 3. The verbal scale provides a qualitative expression of the strength of numerical LR data which may be better understood by the court. For the purposes of the present study the verbal scale also allows broad differences in LR performance to be assessed.

Range of $\mathbf{Log_{10}}$ LR	**Verbal expression**
$\pm 4 \rightarrow \pm 5$	Very strong evidence
$\pm 3 \rightarrow \pm 4$	Strong evidence
$\pm 2 \rightarrow \pm 3$	Moderately strong evidence
$\pm 1 \rightarrow \pm 2$	Moderate evidence
$0 \rightarrow \pm 1$	Limited evidence

Table 3: Verbal expressions of $\log_{10}$ LRs according to Champod and Evett's verbal scale (2000:240)

Both equal error rate (EER) and log LR cost (C_{llr}) (Brümmer and du Preez 2006) are used as metrics of system validity. Validity refers to how well the system (i.e. the variable and the particular set of data) is able to separate same- (SS) and different-speaker (DS) pairs. EER provides a "hard", "error-based" (Brümmer and du Preez 2006: 230) measure of validity dealing with binary accept-reject decisions EER has an operating point at which the number of false hits (DS pairs offering support for the prosecution) and false misses (SS pairs offering support for the defence) are equal. As such, EER is not a forensically realistic metric of validity since in forensic casework the imperative is to avoid false hits. C_{llr} is a gradient, "soft" detector, which penalises the system for high contrary-to-fact LRs (Rose 2010). In both cases, optimum validity (i.e. complete separation of SS and DS pairs) is 0. C_{llr} becomes sub-optimum as it approaches 1, whilst values of greater than 1 are considered very poor performance (Morrison 2011b).

3. Results

3.1. Number of reference speakers

Figures 5 and 6 display the distributions of calibrated same-speaker (SS) and different-speaker (DS) $\log_{10}$ LRs as a function of the number of reference speakers. Across conditions, SS comparisons predominantly achieve LRs equivalent to 'limited' support for the prosecution, with the majority of $\log_{10}$ LRs approaching 0 (neutral evidence; no support for prosecution or defence). Of the 20 SS comparisons, only one achieves contrary-to-fact support for the defence, although this value never exceeds -0.1. DS comparisons generally perform worse, with over 50% of the 380 comparisons in each condition consistently achieving positive $\log_{10}$ LRs. However, in all conditions the magnitude of the contrary-to-fact DS LRs never exceeds 0.4. The poor overall strength of evidence achieved and the high proportion of false hits, reflects the poor value of AR as a speaker discriminant.

Figure 5 reveals striking consistency in the distribution of $\log_{10}$ LRs as the number of reference speakers increases. The largest difference in the distribution of LRs from the *true* LR (based on 1000 speakers) is found with 50 speakers. However, even in this case the differences are incredibly small (difference in medians = 0.011, difference in ranges = 0.1399). There is marginal underestimation of the median, inter-quartile range and overall range with the smallest amount of reference data (10 speakers). Further, the greatest instability in the distribution of log LRs is found with small amounts of reference data, although the fluctuation is very small.

DS LRs (Figure 6) were also remarkably robust to the effects of differences in reference sample size. Across all conditions the median fluctuates maximally within a range of 0.017. As with SS LRs, there is underestimation of the median (i.e. closer to zero), inter-quartile range and overall range with only 10 speakers compared with the *true* LR. Further, there is minor instability in the distribution of LRs found between the 10- and 50-speaker conditions, with some individual DS LRs increasing by up to 0.17. However, given that the overall range of LRs is always between 'limited' support for the prosecution and 'limited' support for the defence, it is considered that the LRs from the 10-speaker condition adequately capture the *true* distribution of DS LRs for this data set.

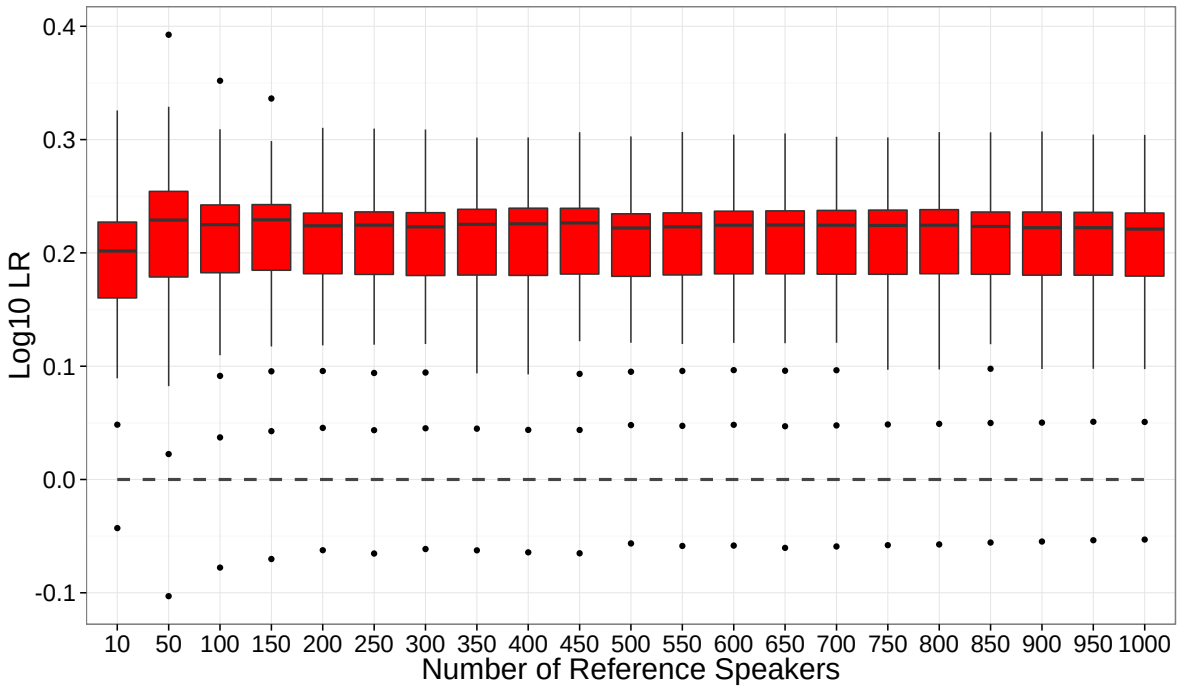

Figure 5: Calibrated SS $\log_{10}$ LRs as a function of the number of reference speakers (where mid line = median, filled box = interquartile range (containing middle 50% of the data), whiskers = scores outside the middle 50%, dots = outliers, dashed line = neutral evidence ('unity')) (Rose 2012)

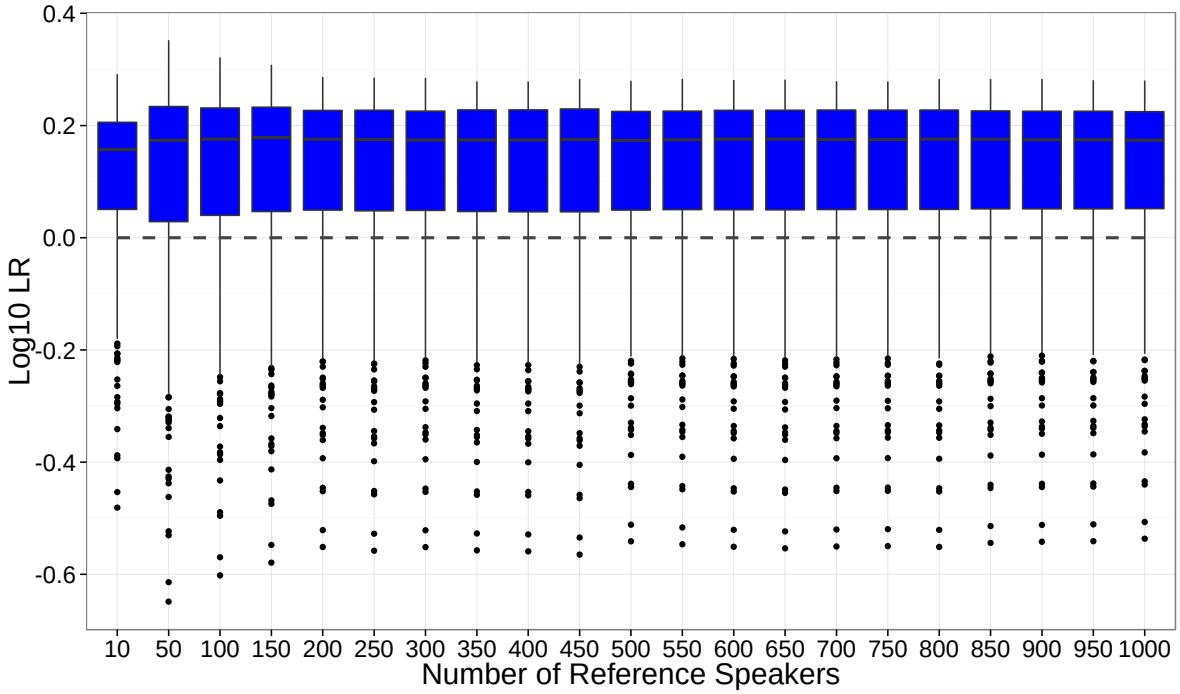

Figure 6: Calibrated DS $\log_{10}$ LRs as a function of the number of reference speakers

Figure 7 shows a line graph of EER as a function of the number of reference speakers, with the *true* EER (LRs based on 1000 reference speakers) plotted as a single line across all N speakers conditions as a means of comparison. The box-like fluctuations in EER are accounted for the fact that EER is a categorical metric with intervals dependent only on the

number of comparisons. Given that the variation in Figure 7 occurs within such a small percentage range, the fluctuations are attributed to a single, or small number of, comparison(s) being correctly/incorrectly categorised at given N speakers conditions. The EER of the *true* LRs is 35.1%. This means that when the proportion of false hits and false misses is equal, the system incorrectly classifies SS as DS and vice versa in 35.1% of pairs. Such poor performance reflects the very high proportion of DS pairs classified as offering support for the prosecution. There is some fluctuation in performance as the number of reference speakers increases. However, the variation appears to be random since the *true* EER is achieved with as few as 17 speakers. Indeed, the maximum extent of the fluctuation in EER performance is just 0.3% across all conditions, suggesting that categorical accept-reject performance of the system is relatively stable across different sample sizes.

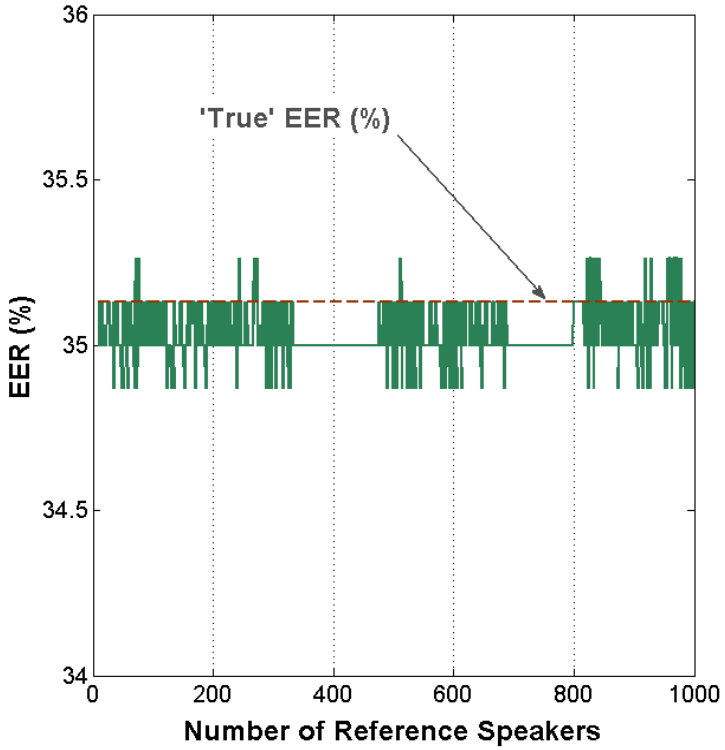

Figure 7: Equal error rate (EER, %) based on calibrated LRs as a function of the number of reference speakers

Figure 8 displays C_{llr} performance as a function of the number of reference speakers. The baseline C_{llr} achieved with the *true* LRs is approaching 1. As with EER, this reflects very poor system validity for AR. Performance based on C_{llr}, as a function of the number of reference speakers, is more systematic than for EER. There is overestimation of performance using C_{llr} with fewer than 200 speakers, such that the lowest value (best validity) is achieved with 57 speakers (0.963). With greater than 200 speakers performance appears asymptotic, although there is still marginal increase in C_{llr}. However, the overall range of C_{llr} remains very small since the C_{llr}s of the systems with very small numbers of speakers (10-20) are almost equivalent to the *true* C_{llr} of 0.971.

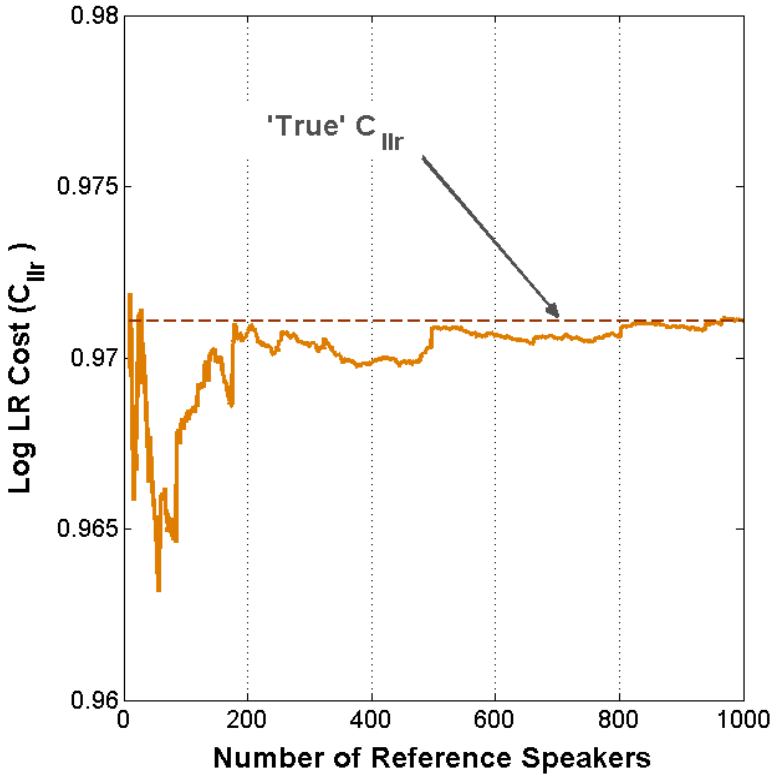

Figure 8: Log LR Cost (C_{llr}) based on calibrated LRs as a function of the number of reference speakers

Uncalibrated SS (Figure 9) and DS (Figure 10) LR scores were also plotted for the same set of test data. The uncalibrated scores display more complex sensitivity to the size of the reference data than calibrated LRs. For SS pairs the interquartile ranges of LR scores are always within one order of magnitude. There is underestimation of the median strength of evidence with small numbers of speakers (the lowest median SS score is achieved with 10 reference speakers). Further, there is marked instability in the interquartile range and overall range with smaller numbers of speakers compared with the *true* scores. The interquartile ranges and overall ranges with fewer than 200 speakers are also consistently underestimated.

Much more significant is the effect of different numbers of reference speaker on individual SS pairs. This is clearly seen in the lowest LR score (offering the most contrary-to-fact support for the defence), which is classed as an outlier. With between 10 and 50 reference speakers this score is around -0.5, equivalent to 'limited' support for the defence. By the inclusion of 150 reference speakers, the score has increased by the equivalent of one order of magnitude (to 'moderate' support for the defence). Assessment of individual scores suggests that certain SS pairs are more affected by the size of the reference sample than others.

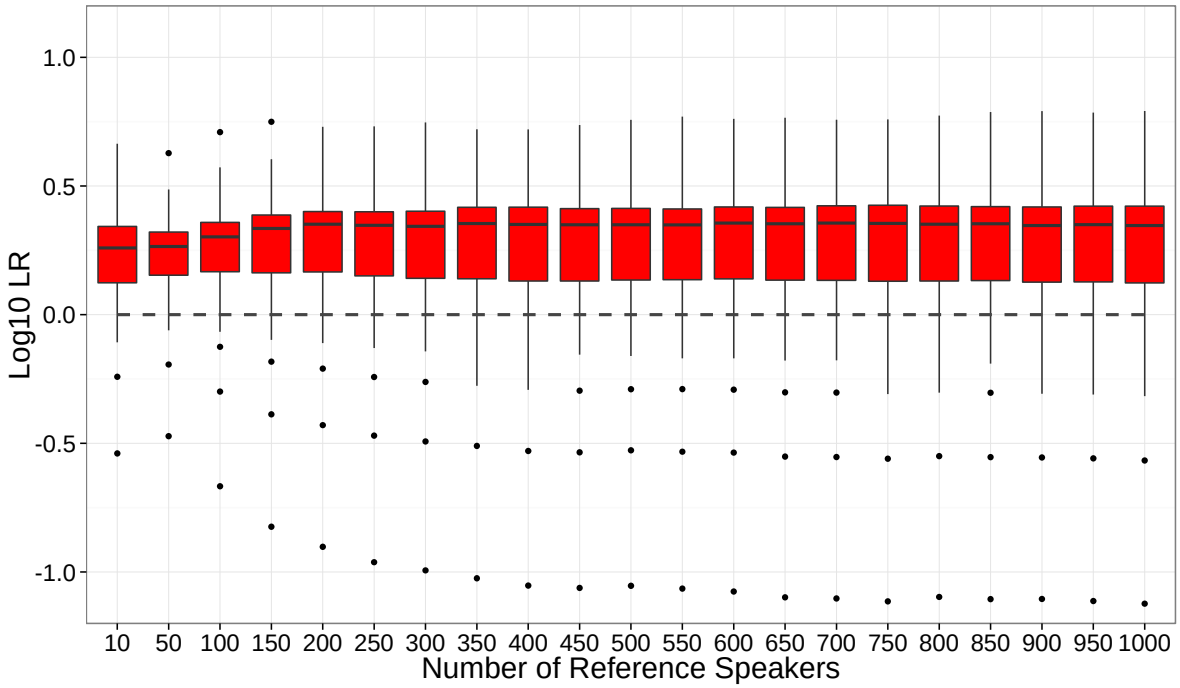

Figure 9: Uncalibrated SS $\log_{10}$ LRs as a function of the number of reference speakers

A similar pattern is found in Figure 10, although the effects are greater for DS pairs. Whilst the median DS score is relatively robust to sample size, the interquartile range and overall range is considerably underestimated with smaller samples. As with SS pairs, this underestimation is greatest in the 50 speaker condition. There is also greater variability in the distributions of scores when using comparatively small amounts of data. Again, the most significant effects are found in the highest, outlying values. For the two most extreme negative DS scores, the strength of evidence increases from less than -2 ('moderate' evidence) to over -3 ('moderately strong' evidence), equivalent to an increase of two orders of magnitude between the 10- and 1000-speaker conditions. Other smaller outliers increase by one order of magnitude as the number of reference speakers increases. This suggests that the magnitude of the score relative to the distribution plays some role in determining sensitivity to sample size.

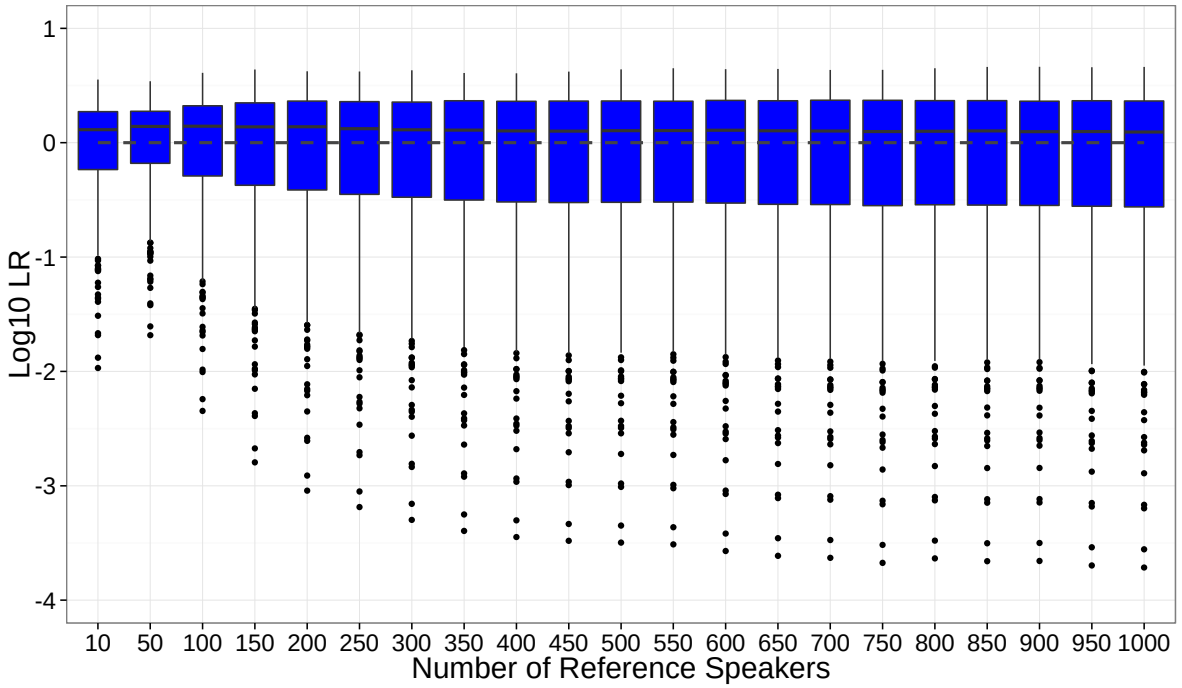

Figure 10: Uncalibrated DS $\log_{10}$ LRs as a function of the number of reference speakers

3.2. Number of tokens per reference speaker

In this section, LRs are computed using a random sample of 200 speakers and between 10 and 200 tokens per speaker. Output is assessed as a function of the number of tokens. Figure 11 shows the distributions of calibrated same-speaker (SS) $\log_{10}$ LRs as the number of tokens per reference speaker increases. There is very minimal overestimation of the median and range when using smaller samples, such that the highest median strength of evidence (0.239) is reached with 5 tokens per speaker and the highest range reached with 6 tokens. As in §3.1, the extent of variation as a function of the number of tokens per speaker is minimal with all bar one of the SS comparisons consistently achieving a LR that supports the prosecution. The magnitude of the calibrated SS LRs is consistently close to the zero turning point (unity).

Figure 12 reveals a similar pattern in terms of calibrated different-speaker (DS) $\log_{10}$ LRs. The median remains essentially the same across all conditions, even when very small numbers of tokens are included in the reference data. The interquartile and overall ranges are marginally overestimated with small numbers of tokens compared with the distribution of *true* LRs. This is reflected in the decrease in the strength of evidence for the two most extreme negative DS LRs, although in absolute terms these LRs increase by less than 0.1 between the 10- and 200-token conditions. In all conditions, DS LRs are maximally spread within a range of two orders of magnitude (between 'limited' support for prosecution and 'limited' support for defence). Further, the middle 50% of DS comparisons consistently achieve LRs offering contrary-to-fact support for the prosecution, although their absolute magnitude is relatively low (never greater than 0.3).

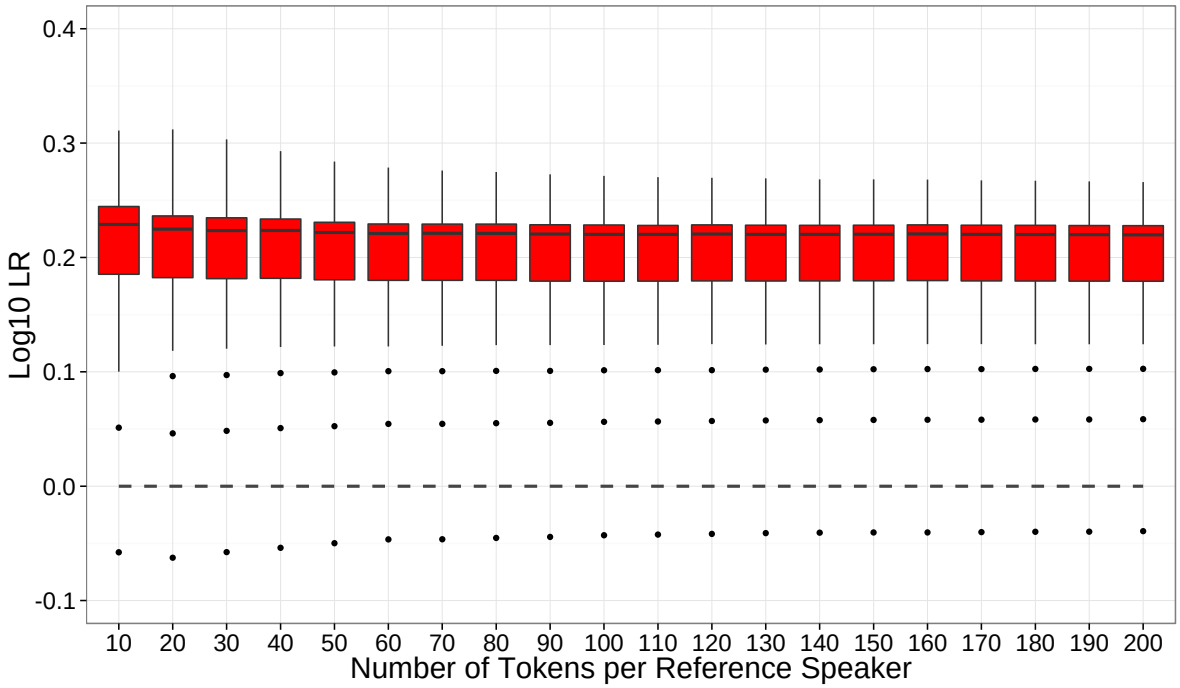

Figure 11: Distributions of calibrated SS $\log_{10}$ LRs as a function of the number of tokens per reference speaker

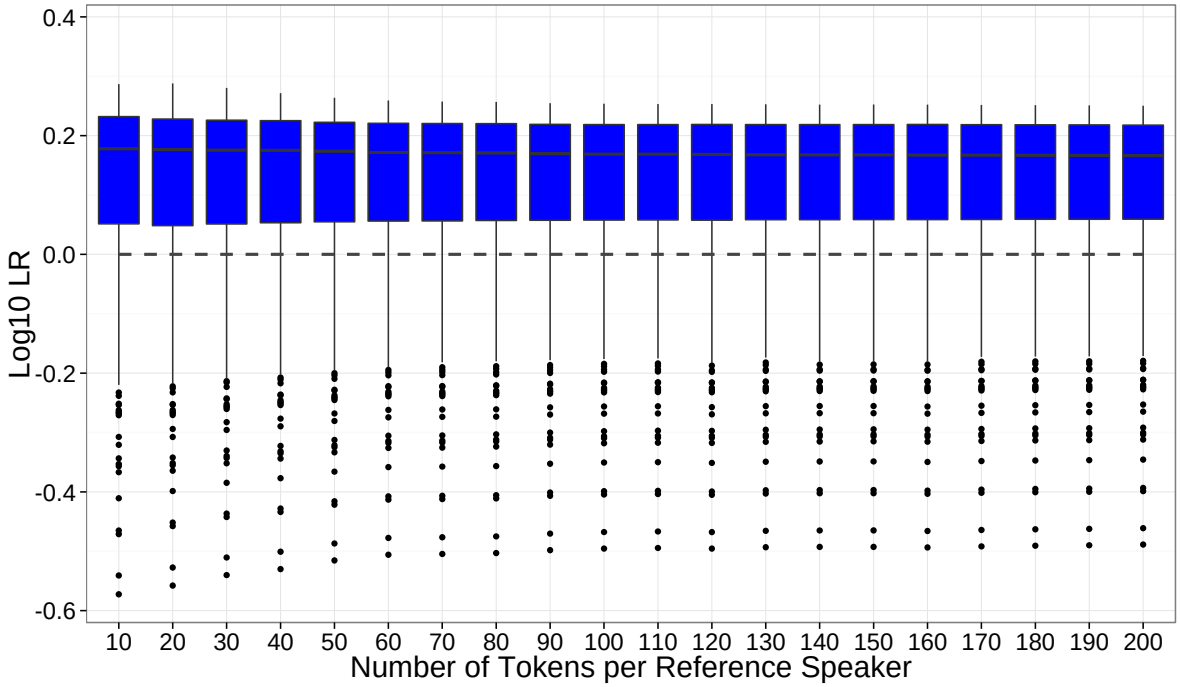

Figure 12: Distributions of calibrated DS $\log_{10}$ LRs as a function of the number of tokens per reference speaker

Consistent with Figures 11 and 12, EER is relatively robust to the number of tokens per reference speaker (Figure 13). EER based on the maximum amount of reference data is 35%. With the inclusion of greater than 96 tokens per speaker EER remains consistent (35%). However, the same EER is achieved with just 2 tokens per speaker suggesting that increasing

the number of tokens does not offer anything in terms of categorical system validity. As with Figure 8, there is variability in EER although this is not systematic. Further, the variability is within a very narrow range (maximally 0.26%) and as such can be assumed to be of little practical interest. Figure 14 displays C_{llr} based on calibrated LRs as a function of the number of tokens per reference speaker. Relative to the *true* LR baseline, there is a small amount of overestimation of how well the system performs when using small amounts of data. The system with the lowest C_{llr} is based on just 6 tokens per speaker. After this point C_{llr} increases until performance appears asymptotic with greater than 100 tokens per speaker, although the range of observed C_{llr} variability is rather small (maximally 0.005).

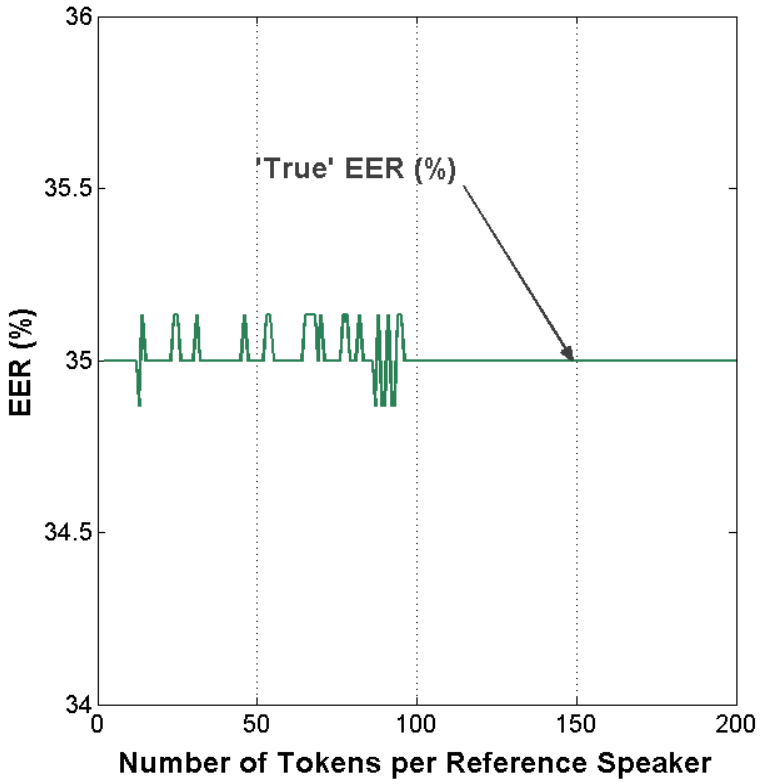

Figure 13: Equal error rate (EER, %) based on calibrated LRs as a function of the number of tokens per reference speaker

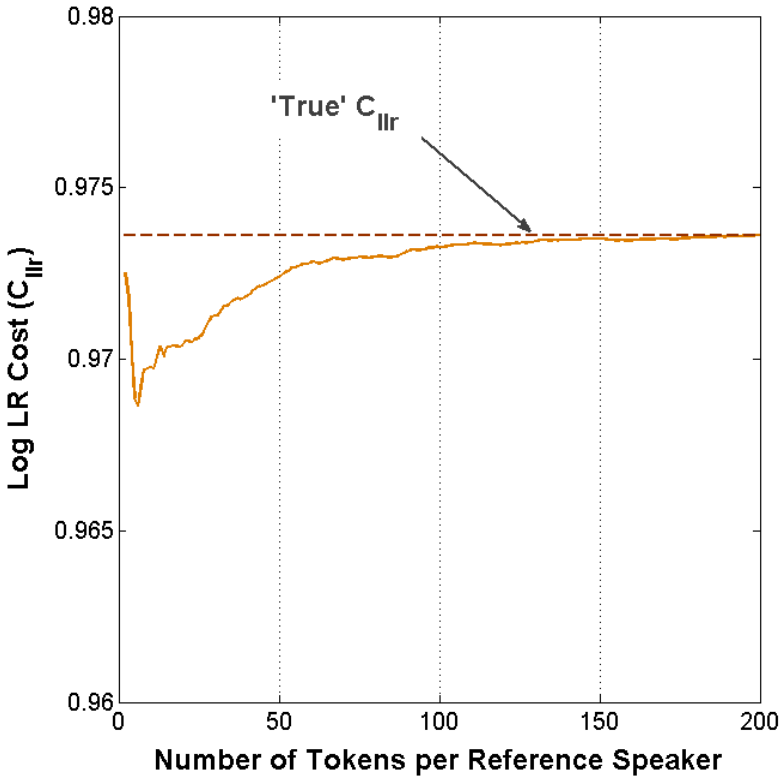

Figure 14: Log LR Cost (C_{llr}) based on calibrated LRs as a function of the number of tokens per reference speaker

As in §3.1.2, calibration plays a role in minimising the effects of small numbers of tokens. Figure 15 reveals considerable differences in the distributions of scores using small reference samples compared with the *true* LR. The median is underestimated (i.e. closer to zero) to a small extent with fewer than 50 tokens. Between the 10- and 50-token conditions there is also greater variability in the median. The interquartile range is narrower with smaller numbers of tokens (and narrowest with 20 tokens). Again the outlying, contrary-to-fact values are affected to the greatest extent. Considering the outlier with the largest negative value, strength of evidence increases by one order of magnitude (increase = 0.97) across all conditions from 'limited' to 'moderate' support for the defence.

The effects of small sample size are more dramatic for DS scores than for SS scores (Figure 16). The median strength of evidence decreases as the number of tokens per speaker increases. As such the median based on 10 tokens is positive whilst the median based on 200 tokens is negative. However, in absolute terms the differences in the medians is relatively small (0.17). As in Figure 15, the interquartile range of DS scores is underestimated when using smaller numbers of tokens per speaker, only stabilising after 100 tokens. Generally, strength of evidence increases (more support for the defence) with larger amounts of data per reference speaker. This is reflected in the magnitude of the most outlying negative scores. Considering one particular outlying DS pair, there is an increase in strength of evidence between the 20 and 200 token conditions equivalent to the difference between 'moderately strong' and 'very strong' support for the defence. In terms of the magnitude of the $\log_{10}$ LR scores this is an increase of two orders of magnitude.

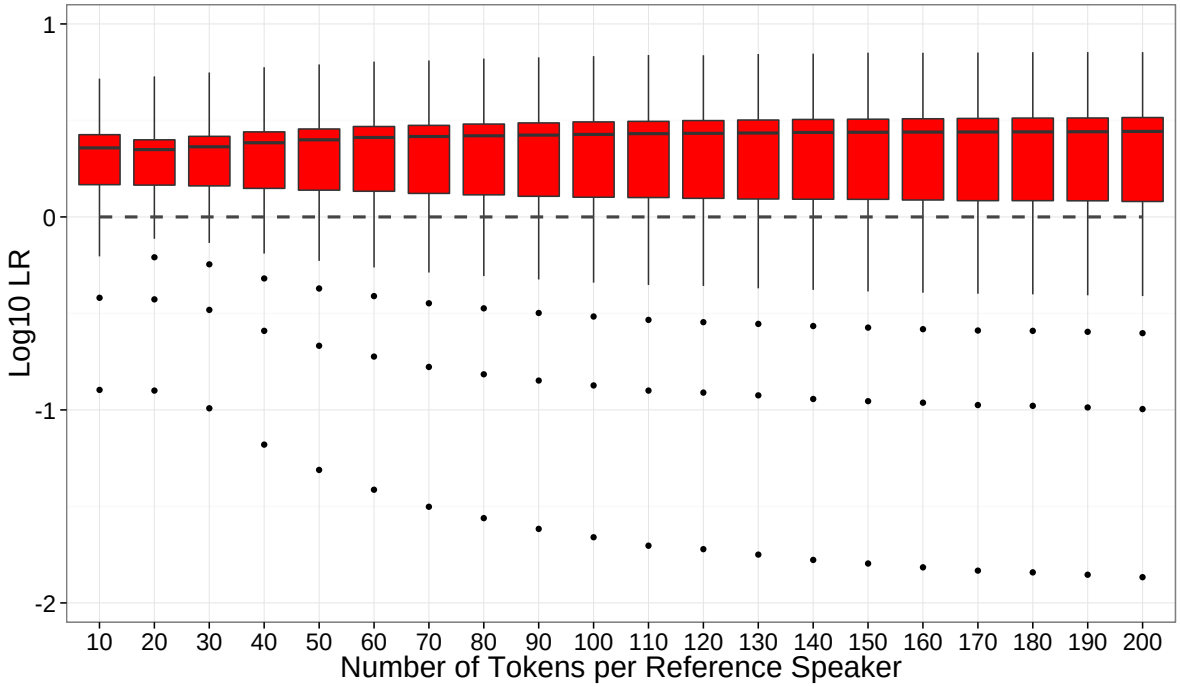

Figure 15: Uncalibrated SS $\log_{10}$ LR scores as a function of the number of tokens per reference speaker

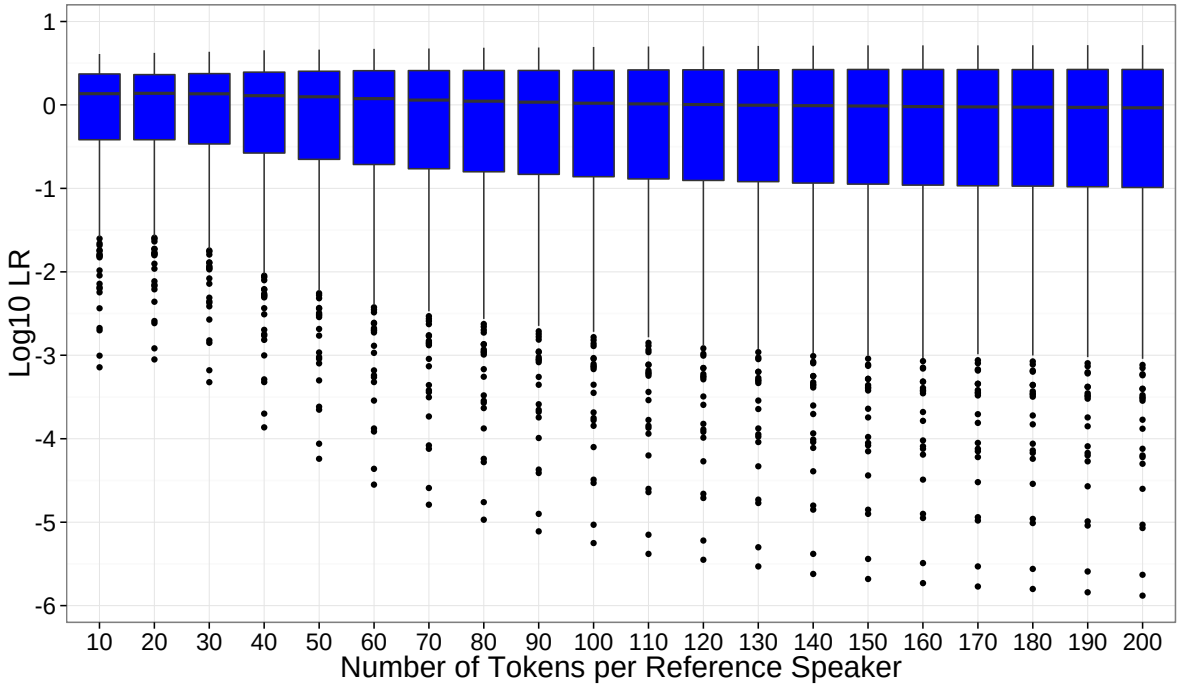

Figure 16: Uncalibrated DS $\log_{10}$ LR scores as a function of the number of tokens per reference speaker

4. Discussion

4.1. Effects of sample size on calibrated LRs

The results in §3.1.1 and §3.2.1reveal that calibrated SS and DS LRs for this set of data are relatively robust to the number of reference speakers and the number of tokens per reference speaker. Limited variability was found in the distributions of LRs with smaller amounts of reference data. However, the medians, interquartile ranges and overall ranges of LRs with the smallest amounts of data were consistently within the same order of magnitude as the distribution of *true* LRs. These findings suggest that a fairly precise estimate of the magnitude of calibrated SS and DS LRs can be achieved using 10 reference speakers and just 2 tokens per speaker. Whilst not considered directly in this study, there are, however, potential interactions between the number of speakers and the number of tokens to consider.

Consistent patterns were also found with regard to the validity of the systems based on calibrated LRs. EER remained stable as the number of speakers and the number of tokens per speaker increased, despite random variation within a very narrow range. Such variability can be explained by the inherently poor performance of AR as a speaker discriminant. Since the calibrated LRs are very close to zero, slight changes in the distributions of within- or between-speaker variation in the reference data can cause marginally positive values to become marginally negative and vice versa. Since EER deals only in absolute false hits and false misses such minor fluctuation has a direct effect on validity. This is an inherent limitation of using EER as a measure of performance, particularly for poor discriminants.

C_{llr} overcomes this limitation of EER by considering the gradient magnitude of the 'errors' made by the system. C_{llr} was found to be at its lowest when using smaller numbers of speakers and tokens per speaker. As such, the C_{llr} validity of calibrated LRs is systematically overestimated with small amounts of data relative to the *true* LRs. This improved performance is attributed to the slightly wider interquartile range for DS comparisons based on small amounts of reference data. Given that the interquartile ranges across all conditions in Figures 6 and 12 are positive (i.e. support for defence), the wider interquartile range means that the first quartile is closer to zero. Since the magnitude of a proportion of the 'errors' is lower when using smaller amounts of data, C_{llr} is also lower.

4.2. The role of calibration

As highlighted in §3.1.2 and §3.2.2, calibration has played an important role in reducing the sensitivity of LRs to small amounts of reference data. As such, the calibrated results in this study are not comparable with Ishihara and Kinoshita (2008), Hughes and Foulkes (2012) or Rose (2012) since these studies did not calibrate scores. The uncalibrated results are very much consistent with previous studies in that scores are misrepresentative and unstable when using small numbers of speakers and tokens. However, whilst previous studies found overestimation of scores within a wider range when using small samples, the scores in this study were underestimated and within a narrower range with small amounts of reference data. This highlights that different variables are affected by sample size in different ways.

The importance of calibration may also be specifically related to AR. The calibration procedure used in this study is configured to improve C_{llr}. For both experiments, the ranges of uncalibrated SS and DS scores increase as a function of the amount of reference data resulting in more contrary-to-fact scores of a higher magnitude when using large amounts of

reference data compared with smaller samples. As such, calibration weights generated for systems based on more reference data are greater than those based on less reference data. Despite calibration improving C_{llr} to different degrees for different conditions, Figures 8 and 14 suggest that AR performance cannot be improved beyond a ceiling close to unity, due to its inherently poor discriminatory value. For better speaker discriminants the role of calibration relative to the size of the reference sample may be different.

The uncalibrated LR results do highlight three important general issues. Firstly, there appears to be an interaction between calibration procedures and the overall sensitivity of LRs to sample size, at least in the present study. Whilst calibration counteracts the effects of small sample sizes, calibrated LRs are spread over a narrower range and are much closer to zero compared with the uncalibrated scores. Secondly, certain pairs of samples are more susceptible to the effects of sample size than others. This may be related to the magnitude of the score relative to the rest of the distribution. Thirdly, the uncalibrated results in Figures 9, 10, 15 and 16 suggest that within- and between-speaker variation are actually very poorly estimated when the background model consists of small amounts of reference data. Therefore, in the absence of calibration, considerable caution should be exercised when interpreting the absolute or relative value of scores generated using a small reference sample.

4.3. MCS procedures for FVC

MCS have provided a valuable resource for investigating the issue of sample size in this study. The procedures implemented have been able to adequately generate a large amount of univariate data which captures the correlation between mean and SD of AR fairly well. In practical terms, MCS are easy to implement and can be used to generate a lot of data quickly and efficiently. Crucially, however, MCS are dependent on the assumption that the underlying distribution of within- and between-speaker variation in the relevant population is known, either through previous research or raw data. As such, caution is advised when implementing MCS procedures using already small sets of raw data. Further, even with larger sets of raw data, procedures for assessing the precision of the representative data should be implemented as a preliminary, exploratory tool.

5. Conclusion

This paper has considered the effects of sample size based on the number of reference speakers and tokens per reference speaker for assessing typicality when computing numerical LRs based on local AR. Calibrated LRs were found to be robust to sample size effects, whilst uncalibrated scores displayed much more sensitivity to the amount of reference data used. Although calibration has been shown to have value in minimising sample size effects for this data set, the generalisability of these results to other FVC variables remains an empirical question. More generally, the results highlight the importance of considering the potential effects of the amount of reference data used when computing LRs both in research and in casework. Future work should focus on testing the sensitivity of much better speaker discriminatory variables to sample size. Attention should also be directed towards developing MCS procedures for synthesising more complex, multivariate FVC variables.

Acknowledgements

This research is funded by a UK Economic and Social Research Council DTC scholarship and the Marie Curie Actions EC Grant Agreement No. PITN-GA-2009-238803 (Bayesian Biometrics for Forensics, BBfor2). We are grateful to Paul Foulkes, Peter French and Dominic Watt for feedback on earlier versions of this paper. Thanks also to Anil Alexander, Esam Alzqhoul, Niko Brümmer, Philip Harrison, Geoffrey Morrison and Balu Nair for scripts which facilitated data analysis. We are thankful to two anonymous reviewers for their valuable comments.

References

AITKEN, COLIN G. G. & LUCY, DAVID. 2004. "Evaluation of trace evidence in the form of multivariate data". *Applied Statistics 54*: 109–122.

AITKEN, COLIN G. G. & TARONI, FRANCO. 2004. *Statistics and the Evaluation of Evidence for Forensic Scientists* (2nd edition). Chichester: John Wiley.

BERNARD, J. R. 1967. "Some measurements of some sounds of Australian English". PhD dissertation, University of Sydney.

BRÜMMER, NIKO. 2007. "FoCal multi-class: toolkit for evaluation, fusion and calibration of multi-class recognition scores".
Online resource: http://sites.google.com/site/nikobrummer/focal

BRÜMMER, NIKO & DU PREEZ, JOHAN. 2006. "Application-independent evaluation of speaker detection". *Computer Speech and Language* 20(2/3): 230–275.

CHAMPOD, CHRISTOPHE & EVETT, IAN W. 2000. "Commentary on A.P.A Broeders (1999) 'Some observations on the use of probability scales in forensic identification". *Forensic Linguistics* 7(2): 238–243.

GOLD, ERICA. In progress. "Calculation of likelihood ratios using phonetic and linguistic features". PhD dissertation, University of York.

GOLDMAN-EISLER, F. 1968. *Psycholinguistics: Experiments in Spontaneous Speech.* London: Academic Press.

HUGHES, VINCENT & FOULKES, PAUL. 2012. "Effects of variation on the computation of numerical likelihood ratios for forensic voice comparison". Paper presented at the *International Association of Forensic Phonetics and Acoustics (IAFPA) Conference*. Universidad Internacional Menedez Pelayo, Santander, Spain. 5–8 August 2012.

ISHIHARA, SHUNICHI & KINOSHITA, YUKO. 2008. "How many do we need? Exploration of the population size effect on the performance of forensic speaker classification". In: *Proceedings of the 9th Annual Conference of the International Speech Communication Association (Interspeech)*. Brisbane, Australia. pp. 1941-1944.

JESSEN, MICHAEL. 2007. "Forensic reference data on articulation rate in German". *Science and Justice* 47: 50–67.

KÜNZEL, HERMAN. J. 1997. "Some general phonetic and forensic aspects of speaking tempo". *International Journal of Speech, Language and the Law* 4(1): 48–83.

LINDLEY, DENNIS. V. 1977. "A problem in forensic science". *Biometrika* 64: 207–213.

MILLER, J., GROSJEAN, F. & LOMANTO, C. 1984. " Articulation rate and its variability in spontaneous speech: a reanalysis and some implications". *Phonetica* 41: 215–225.

MORRISON, GEOFFREY S. 2007. "MatLab implementation of Aitken and Lucy's (2004) forensic likelihood-ratio software using multivariate-kernel-density estimation".
Online resource: http://geoff-morrison.net/Software/multivar_kernel_LR.m

MORRISON, GEOFFREY S. 2009. "Robust version of train_llr_fusion.m from Niko Brümmer's FoCal toolbox".
Online resource: http://geoff-morrison.net/Software/train_llr_fusion_robust.m

MORRISON, GEOFFREY S. 2011a. "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(2): 242–256.

MORRISON, GEOFFREY S. 2011b. "Measuring the validity and reliability of forensic likelihood-ratio systems". *Science and Justice* 51(3): 91-98.

MORRISON, GEOFFREY S. 2013. "Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio". *Australian Journal of Forensic Sciences* 45(2): 173–197.

MORRISON, GEOFFREY S., ROSE, PHILIP & ZHANG, CUILING. 2012. "Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice". *Australian Journal of Forensic Sciences* 44: 155–167.

MORRISON, GEOFFREY S., OCHOA, FELIPE & THIRUVARAN, THARMARAJAH. 2012. "Database selection for forensic voice comparison". *Proceedings of Odyssey 2012: The Language and Speaker Recognition Workshop*. Singapore. 62-77.

NOLAN, FRANCIS, MCDOUGALL, KIRSTY, DE JONG, GEA & HUDSON, TOBY. 2009. "The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research". *International Journal of Speech, Language and the Law* 16(1): 31–57.

ROBERTON, BERNARD & VIGNAUX, G. A. 1995. *Interpreting Evidence: Evaluating Forensic Science in the Courtroom.* Chichester: John Wiley.

ROSE, PHILIP. 2010. "Bernard's 18 – vowel inventory size and strength of forensic voice comparison evidence". In: M. Tabain, J. Fletcher, D. Grayden, J. Hayek & A. Butcher (eds) *Proceedings of the 13th Australasian International Conference on Speech Science and Technology*. Canberra: ASSTA. pp. 30-33.

ROSE, PHILIP. 2011. "Forensic voice comparison with Japanese vowel acoustics – a likelihood ratio-based approach using segmental cepstra. In: *Proceedings of the 17th International Congress of Phonetic Sciences XVII*. Hong Kong, China. pp. 1718-1721.

ROSE, PHILIP. 2012. "The likelihood ratio goes to Monte Carlo: the effect of reference sample size on likelihood ratio estimates". Paper presented at the *UNSW Forensic Speech Science Conference.* Sydney, Australia. 3 December 2012.

ROSE, PHILIP. 2013. "More is better: likelihood ratio-based forensic voice comparison with vocalic segmental cepstra frontends". *International Journal of Speech, Language and the Law* 20(1): 77-116.

ROSE, PHILIP, KINOSHITA, YUKO & ALDERMAN, TONY. 2006. "Realistic extrinsic forensic speaker discrimination with the diphthong /aɪ/". In: P. Warren & C. I. Watson (eds) *Proceedings of the 11th Australasian International Conference on Speech Science and Technology*. Canberra: ASSTA. pp. 1941-1944.

ROSE, PHILIP & MORRISON, GEOFFREY S. 2009. "A response to the UK Position Statement on forensic speaker comparison". *International Journal of Speech, Language and the Law* 16(1): 139–163.

TABACHNICK, BARBARA. G. & FIDDELL, LINDA. S. 2007. *Using Multivariate Statistics* (5th edition). Boston: Pearson.

WACKERLY, DENNIS D., MENDENHALL III, WILLIAM & SCHEAFFER, RICHARD L. 2008. *Mathematical Statistics with Applications* (7th edition). London: Thomson.

WANG, Z. X. & GUO, D. R. 1989. *Special functions.* London: World Scientific.

WELCH, B. L. 1947. "The generalization of student's problem when several different population variances are involved." *Biometrika* 34(1/2): 28–35.

Vincent Hughes
Department of Language and Linguistic Science
University of York
Heslington
York
YO10 5DD
United Kingdom
email: vh503@york.ac.uk

Erica Gold
Department of Language and Linguistic Science
University of York
Heslington
York
YO10 5DD
United Kingdom
email: erica.gold@york.ac.uk

Ashley Brereton
Department of Mathematical Sciences
University of Liverpool
Liverpool
L69 3BX
United Kingdom
email: a.brereton@liverpool.ac.uk