



# An Issue in the Calculation of Logistic-Regression Calibration and Fusion Weights for Forensic Voice Comparison

Geoffrey Stewart Morrison<sup>1,2</sup>, Tharmarajah Thiruvaran<sup>1</sup>, Julien Epps<sup>1,3</sup>

<sup>1</sup>Forensic Voice Comparison Laboratory, School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, Australia

<sup>2</sup>School of Language Studies, Australian National University, Canberra, Australia

<sup>3</sup>National ICT Australia, Sydney, Australia

geoff-morrison@forensic-voice-comparison.net, thiru@ee.unsw.edu.au, j.epps@unsw.edu.au

## Abstract

Logistic regression is a popular procedure for calibration and fusion of likelihood ratios in forensic voice comparison and automatic speaker recognition. The availability of multiple recordings of each speaker in the database used for calculation of calibration/fusion weights allows for different procedures for calculating those weights. Two procedures are compared, one using pooled data and the other using mean values from each speaker-comparison pair. The procedures are tested using an acoustic-phonetic and an automatic forensic-voice-comparison system. The mean procedure has a tendency to result in better accuracy, but the pooled procedure always results in better precision of the likelihood-ratio output.

**Index Terms:** logistic regression, calibration, fusion, weights, forensic voice comparison, likelihood ratio

## 1. Introduction

### 1.1. The issue

Logistic regression has become a popular technique for calibration and fusion of scores in automatic speaker recognition and in forensic voice comparison [1–6].

It is a requirement of the procedure that the data used to calculate the weights for calibration/fusion include at least two recordings of each speaker (hereafter *A* and *B*) such that same-speaker (target) scores can be calculated. It is essential in forensic voice comparison that these be non-contemporaneous recordings so as to achieve a reasonable estimate of within-speaker variability – in casework the suspect and offender recordings are non-contemporaneous. The pairs of recordings should also ideally be matched to the speaking styles and channels of the suspect- and offender-recording pair.

Having two recordings per speaker affords the opportunity for multiple different-speaker (non-target) comparisons, as shown in Table 1. If we assume that there are differences in speaking style and/or channel between recording *A* and recording *B*, and we want to maintain the distinction for consistency with the suspect and offender recordings, then we will limit our use of different-speaker pairs to those for which the suspect model is based on a recording *A* and the offender data comes from a recording *B* (those marked with an asterisk in Table 1). Only making use of these two pairs also has the additional advantage that there is no overlap in membership between the pairs and they can therefore be treated as statistically independent, see [7].

Having two non-overlapping pairs of different-speaker comparisons presents a choice as to which scores to use for the calculation of weights for logistic regression calibration/fusion; those from the first pair, those from the

second pair, those from the first and second pair pooled together, or the means of the scores from the two pairs. It may be that common practice is to pool the scores without giving the issue any thought; however, there are theoretical and potentially empirical factors to consider. The key point of this paper is that one should consider the issue and make an informed conscious decision.

Table 1: Possible different-speaker comparison pairs

	Suspect model	Recording	Offender data	Recording
	Spk01	<i>A</i>	Spk02	<i>A</i>
*	Spk01	<i>A</i>	Spk02	<i>B</i>
	Spk01	<i>B</i>	Spk02	<i>A</i>
	Spk01	<i>B</i>	Spk02	<i>B</i>
	Spk02	<i>A</i>	Spk01	<i>A</i>
*	Spk02	<i>A</i>	Spk01	<i>B</i>
	Spk02	<i>B</i>	Spk01	<i>A</i>
	Spk02	<i>B</i>	Spk01	<i>B</i>

## 2. Theoretical arguments

### 2.1. Argument favoring the mean procedure

In many automatic-speaker-recognition applications the ultimate task may be to make a categorical decision, and post-decision the value of the score or likelihood ratio (*LR*) used to make that decision is no longer of interest. In forensic voice comparison, however, accurate and precise estimation of the *LR* is essential. It is an *LR* which the forensic scientist presents to the court as a strength-of-evidence statement which the trier of fact will then consider when making their determination on the ultimate issue of guilt or innocence. The *LR* expresses the relative likelihood of obtaining the acoustic differences between voices on the suspect and offender recordings under the hypothesis that they were both produced by the same speaker versus under the hypothesis that the voice on the offender recording was produced by someone other than the suspect.

If there are two *LR* values, each based on a non-overlapping pair of recordings, then each can be considered an independent estimate of the true *LR* for the comparison of the pair of speakers (or single speaker). According to the central-limit theorem, a more accurate estimate of the true *LR* can be obtained by taking the mean of the two individual estimates. There is therefore an argument to be made that the appropriate set of log-likelihood-ratio scores to use for calculating the weights for logistic-regression calibration/fusion should be the (more accurate) means of each pair.

## 2.2. Potential empirical differences between the mean and pooled procedures

How would using the means of the pairs differ empirically from pooling all the scores without reference to the knowledge that they are paired? We can think about this by considering the function to be minimized in logistic regression, the deviance statistic, which, if equal priors are assumed, is the same as the log-likelihood-ratio cost ( $C_{llr}$ ).  $C_{llr}$  is calculated using Equation 1 (where  $N_{ss}$ ,  $N_{ds}$ ,  $LR_{ss_i}$ ,  $LR_{ds_j}$  are the number of same-speaker and different-speaker comparisons, and the LRs from same-speaker and different-speaker comparisons respectively). Figure 1 provides a plot of the contribution of a different-speaker comparison.

$$C_{llr} = \frac{1}{2} \left( \frac{1}{N_{ss}} \sum_{i=1}^{N_{ss}} \log_2 \left( 1 + \frac{1}{LR_{ss_i}} \right) + \frac{1}{N_{ds}} \sum_{j=1}^{N_{ds}} \log_2 \left( 1 + LR_{ds_j} \right) \right) \quad (1)$$

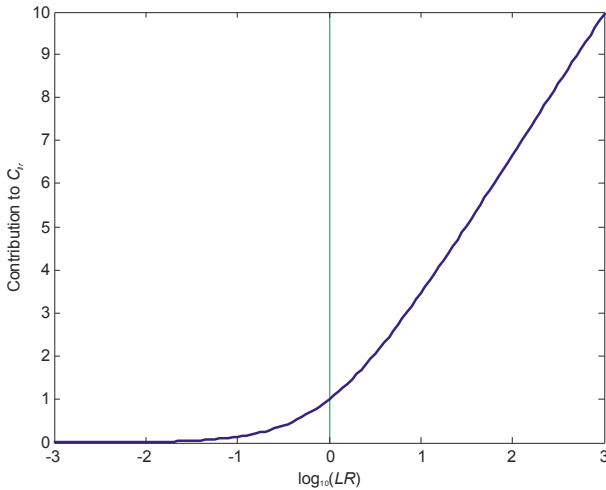


Figure 1: Plot of  $\log_2(1 + LR_{ds_j})$ , the contribution of a different-speaker likelihood ratio to  $C_{llr}$ .

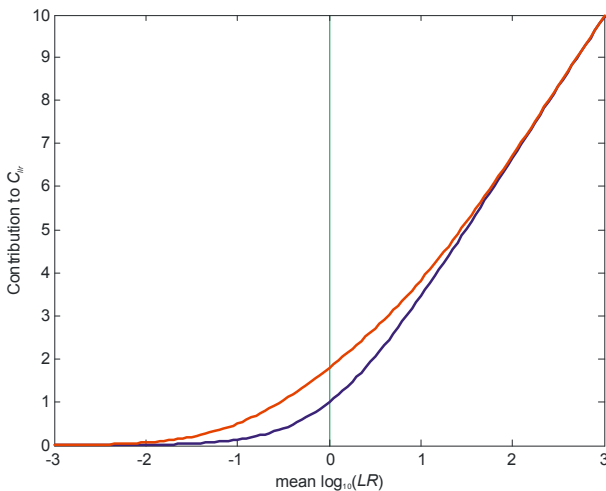


Figure 2: Plot of the contribution to  $C_{llr}$  of the mean procedure (blue line) and pooled procedure (red line) for a pair of different-speaker  $\log_{10}(LR)$  scores separated by 2.

For sake of argument, let us imagine that we have a pair of  $\log_{10}(LR)$  scores which are  $-1$  and  $+1$ . If they were pooled their contribution to  $C_{llr}$  would be the mean of their individual

$C_{llr}$  component values,  $(0.14 + 3.46)/2 = 1.80$ ; however, if their contribution to  $C_{llr}$  were based on their mean value (i.e., 0) that contribution would instead be 1. Figure 2 shows a plot of the  $C_{llr}$  component values from the pooled versus the mean procedures, with the difference between each member of the pair fixed at 2. All else being equal, the difference in the  $C_{llr}$  component values between the two procedures decreases as the pair of LRs move away from a log-likelihood-ratio value of 0 and provide greater support for either the consistent-with-fact or the contrary-to-fact hypothesis, but the pooled procedure always results in a larger contribution to  $C_{llr}$  and hence a larger leverage on the logistic-regression calibration/fusion weights.

## 2.3. Argument favoring the pooled procedure

An explanation for the greater leverage of the pooled technique is that the pooled procedure penalizes any lack of precision in the system whereas the mean procedure does not. If, for sake of argument, the system were perfectly precise, then the two values in each pair would be exactly the same and the two procedures would give the same result. (Given the existence of within-speaker variability at the source, this is of course impossible, and there will always be some degree of imprecision.) All else being equal, as the separation between the members of each pair increases, i.e., as the precision decreases, the leverage of the pooled procedure will increase whereas that of the mean procedure will remain unchanged. There is therefore an argument that, because it takes imprecision into account, the pooled procedure should be preferred over the mean procedure for the calculation of calibration/fusion weights.

## 2.4. Outline of remainder of the paper

In the remainder of the paper, we present empirical tests of which of the two procedures results in the best accuracy and precision of the final likelihood-ratio output of forensic comparison systems (as distinct from the log-likelihood-ratio scores used to calculate the calibration/fusion weights). Tests are conducted using two forensic-voice-comparison systems, one acoustic-phonetic and the other automatic, each applied to a different data set.

## 3. Empirical tests of the mean versus pooled procedures

### 3.1. Systems and data

The *acoustic-phonetic system* was a GMM-UBM system applied to the coefficients of third-order DCTs fitted to the second-formant trajectories of tokens of /aI/, /eI/, /oU/, /aU/, and /oI/ produced by 27 male speakers of Australian English. There were two non-contemporaneous recordings (*A* and *B*) of each speaker, and 16–24 tokens per phoneme per recording. Recording *A* was used to build the suspect model and recording *B* as offender data. Scores were calculated for each phoneme and then the parallel sets of scores were fused. Testing was conducted using a strict cross-validation procedure such that no data from the speaker/speakers being tested were included in the UBM or in the scores used to calculate the fusion weights. The back-end of the system is described in greater detail in [8] and the front-end in [6].

The *automatic system* was a GMM-UBM system applied to MFCCs (16 coefficients plus delta coefficients, with cumulative density mapping for feature normalization) extracted from portions of the signal selected using an energy-

based voice activity detector. No channel/session compensation procedures were applied. The UBM was based on the 750 longest recordings of US English speakers from the NIST SRE 2006 8conv database (101 speakers with 4 to 8 recordings per speaker). The data used to calculate scores for calculating the calibration weights consisted of two non-contemporaneous recordings ( $A$  and  $B$ ) from each of 32 speakers of US English in the NIST SRE 2008 8conv database. These had speech-active duration ranging from 84 s to 131 s with a median of 110 s. The test data consisted of four non-contemporaneous recordings ( $A$ ,  $B$ ,  $C$  and  $D$ ) from each of 100 speakers of US English in the NIST SRE 2008 8conv database. These had speech-active duration ranging from 84 s to 159 s with a median of 109 s. The system is described in greater detail in [7].

### 3.2. Results

In forensic science, it is important to be able to present measures of both the accuracy and precision of a forensic-voice-comparison system. The National Research Council report to Congress on Strengthening Forensic Science in the United States [9] urged that procedures be adopted which include “the reporting of a measurement with an interval that has a high probability of containing the true value; . . . [and] the conducting of validation studies of the performance of a forensic procedure” (p. 121); the latter requiring the use of “quantifiable measures of the reliability and accuracy of forensic analyses” (p. 23).

Whereas accuracy may be of primary concern when the judge decides whether evidence based on the system should be admitted, precision is also important, especially when the strength of evidence is actually presented in court. For example, the trier of fact could (and probably should) interpret a  $\log_{10}(LR)$  of +3 with a 95% credible interval ( $CI$ ) of  $\pm 1$  very differently from  $+3 \pm 3$ . In general, a more precise system will be of greater value to the court.

Measures of accuracy,  $C_{lr}$ , for each of the fused systems are provided in Table 2. The mean procedure, rather than the pooled procedure, was applied when calculating  $C_{lr}$  on the test results. This provides the best estimate of accuracy [7]. For the acoustic-phonetic system, there was only one likelihood-ratio value output for each same-speaker comparison, but two for each different-speaker comparison and the means of the different-speaker  $A$ - $B$  pairs were used for calculating  $C_{lr}$  (in one member of each pair the  $A$  recording was from the lower-numbered speaker and the  $B$  recording from the higher-numbered speaker, and in the other member of the pair was the  $A$  recording was from the higher-numbered speaker and the  $B$  recording from the lower-numbered speaker, see Table 1). Similarly for the automatic system, for which there were  $A$ - $B$  and  $C$ - $D$  pairs,  $C_{lr}$  was calculated using the means of the four values output for each different-speaker comparison and the means of the two values output for each same-speaker comparison.

Measures of precision, the estimated  $\log_{10}(LR)$  95%  $CI$ , for each of the fused systems are provided in Table 3. These were calculated using the parametric procedure described in [7], summarized in Equation 2:

$$CI = \pm t_{1-\frac{\alpha}{2}, df} \hat{\sigma} \quad (2)$$

$$df = \sum_i (n_i - 1)$$

$$\hat{\sigma}^2 = \frac{1}{df} \sum_i \left( \sum_{j=1}^{n_i} (\bar{x}_i - x_{ij})^2 \right)$$

Where  $\alpha$  was set to 0.05,  $n_i$  is the number of  $\log_{10}(LR)$  estimates for speaker comparison  $i$  (i.e.; 2 for each different-

speaker comparison in the acoustic-phonetic system, for which only different-speaker comparisons were used to calculate the  $CI$ ; and 4 for each different-speaker comparison and 2 for each same-speaker comparison in the automatic system), and  $x_{ij}$  is the  $j$ th estimate of the  $\log_{10}(LR)$  for speaker-comparison  $i$ . Uniform priors were assumed, hence the sample variance was substituted for the posterior variance.

Table 2:  $C_{lr}$  values for output of systems using different procedures for calculating fusion/calibration weights.

system	procedure	
	mean	pooled
acoustic-phonetic	0.047	0.040
automatic	0.142	0.151

Table 3: Estimated 95%  $CI$ s, in  $\log_{10}(LR)$ , for output of systems using different procedures for calculating fusion/calibration weights.

system	procedure	
	mean	pooled
acoustic-phonetic	$\pm 3.19$	$\pm 2.81$
automatic	$\pm 1.89$	$\pm 1.55$

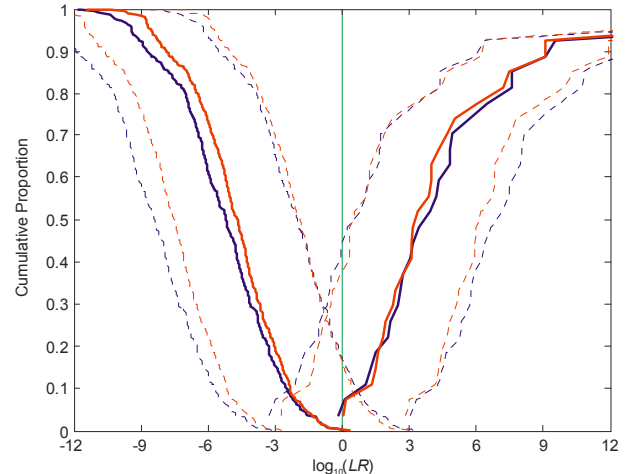


Figure 3: Tippett plots of the output of the acoustic-phonetic system using the mean procedure (blue lines) and pooled procedure (red lines).

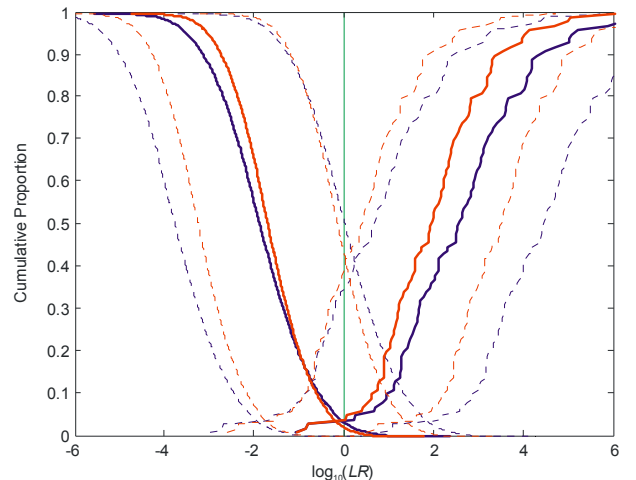


Figure 4: Tippett plots of the output of the automatic system using the mean procedure (blue lines) and pooled procedure (red lines).

Figures 3 and 4 provide Tippet plots of the results using the mean procedure and the pooled procedure for calculating the calibration/fusion weights. The solid lines are based on the application of the mean procedure to the test results (see description of calculation of  $C_{lr}$  above). The dashed lines to the left and right of the solid lines show the estimates of the 95%  $CI$ . Note that across figures the  $x$  axes do not have the same scale, and since the two systems were tested on different test databases their results should not be directly compared.

To explore the robustness of the results from the automatic system, a series of 250 randomization tests were conducted randomly selecting which 32 speakers from the NIST SRE 2008 8conv database to use for weight training and which 100 to use for testing. Histograms based on the difference in  $C_{lr}$  and 95%  $CI$  between the mean and the pooled procedures for each randomization are provided in Figures 5 and 6 respectively. Positive values in Figure 5 indicate that the pooled procedure had a lower  $C_{lr}$  (better accuracy) than the mean procedure, and positive values in Figure 6 indicate that the pooled procedure had a narrower 95%  $CI$  (better precision) than the mean procedure. (In both plots the axes have been limited and extreme outliers are not shown.)

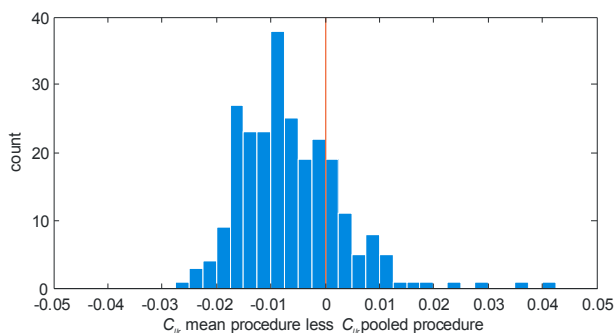


Figure 5: Histogram of the difference in  $C_{lr}$  between the mean and pooled procedures in the randomization tests.

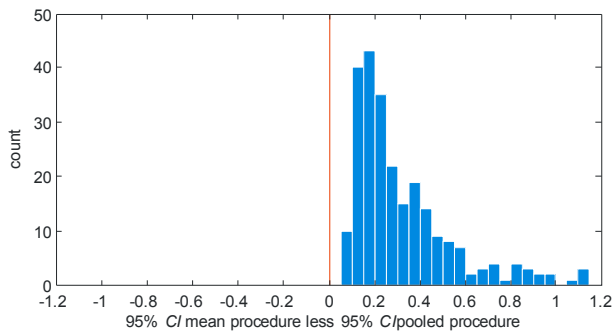


Figure 6: Histogram of the difference in the 95%  $CI$  between the mean and pooled procedures in the randomization tests.

### 3.3. Discussion

In terms of accuracy, the test results did not indicate that one procedure for calculating the calibration/fusion weights was clearly superior to the other, but for the automatic system there was a tendency for the mean procedure to lead to better performance than the pooled procedure. On the test of the acoustic-phonetic system,  $C_{lr}$  for the pooled procedure was 15% less than that for the mean procedure, but on the initial test of the automatic system,  $C_{lr}$  for the mean procedure was 5% less than that for the pooled procedure. The results of the randomization tests on the automatic system (Figure 5) indicated a tendency for the mean procedure to result in better accuracy

In terms of precision, the test results from the pooled procedure for calculating the calibration/fusion weights always

lead to better results than the mean procedure. On the test of the acoustic-phonetic system the 95%  $CI$  was 12% narrower for the pooled procedure than for the mean procedure. For the initial test of the automatic system it was 18% narrower. The results of the randomization tests on the automatic system (Figure 6) indicated that the pooled procedure always resulted in better precision.

## 4. Conclusions

Comparison of two procedures for calculating weights for logistic-regression calibration/fusion indicate that the *mean procedure* had a tendency to produce more accurate results but that the *pooled procedure* will always produce more precise results. Researchers and practitioners should consider any potential trade-off between the accuracy and precision of their automatic-speaker-recognition or forensic-voice-comparison system and consciously choose the procedure for calculating calibration/fusion weights accordingly: On a NIST SRE evaluation where the goal is to minimize an accuracy metric such as  $C_{lr}$  the mean procedure would be preferred, but for forensic voice comparison one may decide that it is more important to maximize precision in which case the pooled procedure would be preferred.

## 5. Acknowledgements

This research was funded in part by Australian Research Council Discovery Grant No. DP0774115.

## 6. References

- [1] Pigeon, S., Druyts, P., and Verlinde, P., "Applying logistic regression to the fusion of the NIST'99 1-speaker submissions", *Digit. Signal Process.*, 10: 237–248, 2000. doi:10.1006/dspr.1999.0358
- [2] Brümmer, N., and du Preez, J., "Application independent evaluation of speaker detection", *Comp. Speech Lang.*, 20:230–275, 2006. doi:10.1016/j.csl.2005.08.001
- [3] Brümmer, N., Burget, L., Cernocký, J.H., Glembek, O., Grézl, F., Karafiát, M., van Leeuwen, D.A., Matejka, P., Schwarz, P., and Strasheim, A., "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST SRE 2006", *IEEE Trans. Audio, Speech, Lang. Process.*, 15:2072–2084, 2007. doi:10.1109/TASL.2007.902870
- [4] van Leeuwen, D.A., and Brümmer, N., "An introduction to application-independent evaluation of speaker recognition systems", in C. Müller [Ed], *Speaker Classification I: Fundamentals, Features, and Methods*, 330–353, Springer-Verlag, 2007. doi:10.1007/978-3-540-74200-5\_19
- [5] González-Rodríguez, J., Rose, P., Ramos, D., Toledano, D. T., Ortega-García, J., "Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition", *IEEE Trans. Audio, Speech, Lang. Process.*, 15:2104–2115, 2007. doi:10.1109/TASL.2007.902747
- [6] Morrison, G. S., "Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs", *J. Acoust. Soc. Americ.* 125:2387–2397, 2009. doi:10.1121/1.3081384
- [7] Morrison, G. S., Thiruvanan, T., and Epps, J., "Estimating the precision of the likelihood-ratio output of a forensic-voice-comparison system", *Proceedings of Odyssey 2010: The Language and Speaker Recognition Workshop*, Brno, Czech Republic, 63–70, 2010.
- [8] Morrison, G. S., "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 Commun.*, In Press. doi:10.1016/j.specom.2010.09.005
- [9] National Research Council, *Strengthening Forensic Science in the United States: A Path Forward*, National Academies Press, 2009.