Speech Communication 85 (2016) 119–126

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Editorial

Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (*forensic_eval_01*) – Introduction☆

Geoffrey Stewart Morrison[a,b,*], Ewald Enzinger[a]

[a] *Morrison & Enzinger, Independent Forensic Consultants, Vancouver, British Columbia, Canada & Corvallis, Oregon, United States of America*
[b] *Department of Linguistics, University of Alberta, Edmonton, Alberta, Canada*

ARTICLE INFO

Article history:
Available online 4 August 2016

Keywords:
Forensic voice comparison
Evaluation
Validity
Reliability
Casework conditions

ABSTRACT

There is increasing pressure on forensic laboratories to validate the performance of forensic analysis systems before they are used to assess strength of evidence for presentation in court. Different forensic voice comparison systems may use different approaches, and even among systems using the same general approach there can be substantial differences in operational details. From case to case, the relevant population, speaking styles, and recording conditions can be highly variable, but it is common to have relatively poor recording conditions and mismatches in speaking style and recording conditions between the known- and questioned-speaker recordings. In order to validate a system intended for use in casework, a forensic laboratory needs to evaluate the degree of validity and reliability of the system under forensically realistic conditions. The present paper is an introduction to a Virtual Special Issue consisting of papers reporting on the results of testing forensic voice comparison systems under conditions reflecting those of an actual forensic voice comparison case. A set of training and test data representative of the relevant population and reflecting the conditions of this particular case has been released, and operational and research laboratories are invited to use these data to train and test their systems. The present paper includes the rules for the evaluation and a description of the evaluation metrics and graphics to be used. The name of the evaluation is: *forensic_eval_01*.

1. Introduction

There is increasing pressure on forensic laboratories to validate the performance of forensic analysis systems before they are used to assess strength of evidence for presentation in court (*Daubert v Merrell Dow Pharmaceuticals* [1993, 509 US 579]; National Research Council, 2009; Forensic Science Regulator, 2014; National Commission on Forensic Science, 2016). In forensic voice comparison, calls for validity and reliability to be empirically tested under casework conditions date back to the 1960s (see Morrison, 2014, for a review), but still go widely unheeded.

Different forensic voice comparison systems may use different approaches, and even among systems using the same general approach there can be substantial differences in operational details. From case to case, the relevant population, speaking styles, and recording conditions can be highly variable, but it is common to have relatively poor recording conditions and mismatches in speaking style and recording conditions between the known- and questioned-speaker recordings. In order to validate a system intended for use in casework, a forensic laboratory needs to evaluate the degree of validity and reliability of the system under forensically realistic conditions. To contribute to this, we have released a set of training and test data representative of the relevant population and reflecting the conditions of an actual forensic voice comparison case, and are organizing an evaluation based on these data. Practitioners and researchers from operational and research laboratories are invited to participate in the evaluation. The data are available to researchers and practitioners who agree to abide by the rules of the evaluation as described below. The data will allow operational and research laboratories to run tests of their system or systems, the results of which will be comparable with the test results from other systems. The evaluation is named *forensic_eval_01*.

The present paper serves as an introduction to a virtual special issue (VSI) in *Speech Communication*. It describes the data provided for training and testing, the rules to be followed by participants, and the metrics and graphics which will be used to describe the

☆ This paper is part of the Virtual Special Issue entitled: Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic_eval_01), [http://www.sciencedirect.com/science/journal/01676393/vsi], Guest Edited by G. S. Morrison and E. Enzinger.

* Corresponding author.
E-mail address: geoff-morrison@forensic-evaluation.net (G.S. Morrison).

http://dx.doi.org/10.1016/j.specom.2016.07.006
0167-6393/

performance of the systems. The submission period for the VSI is expected to be two years. Within that time period, submissions are solicited of papers which:

1. describe the forensic voice comparison systems employed and the manner in which they were employed in sufficient detail that another suitably qualified and equipped forensic practitioner or researcher could potentially replicate what was done; and
2. present the results of empirically testing the systems on the test data provided.

Submitted papers are expected to be relatively succinct. Introductions should be brief. Rather than re-describing the data and test protocol, reference should be made to the descriptions in the present paper. Discussion and conclusions should highlight noteworthy observations based on the results, and may include recommendations for practice. Later submissions may compare their results with selected results from papers that have already been published. At the end of the submission period the plan is to write a conclusion paper including a comparative summary of results from across the papers published in the VSI.

In line with the position expressed by Stoel et al. (2016) in a recent *Science & Justice* editorial, participants in *forensic_eval_01* will be required to commit to publishing the results of the tests of their forensic voice comparison systems irrespective what those systems' levels of performance turn out to be.

The likelihood ratio framework for the evaluation of evidence has been adopted for *forensic_eval_01*, and participants must submit test results as numeric log likelihood ratios.

The test protocol treats each system as a black box. The evaluation is therefore neutral with respect to the approach to forensic voice comparison used and details of implementation. As a practical matter, however, the number of test pairs for which a response is required makes it unlikely that a system will be tested if it requires substantial human input for each test pair.

We wish to emphasise that *forensic_eval_01* is based on the relevant population and conditions (speaking styles and technical recording conditions) of one forensic voice comparison case, and neither the relative nor absolute degree of performance observed for any system will necessarily be generalizable to the relevant population and conditions of any other case. Given the great diversity in relevant population and conditions from case to case, we would advise courts not to base decisions regarding admissibility or weight directly on the results published in this VSI. Instead, when a forensic voice comparison analysis is proffered, we recommend that the court consider and make enquiries as to whether the degree of validity and reliability of the forensic voice comparison system employed in that particular case has been empirically tested using data that are sufficiently representative of the relevant population and sufficiently reflective of the conditions in that particular case (results published in the VSI should only be deemed relevant for a particular case if they meet these criteria).[1] We also caution that although some systems may require less human input than others, all systems require some degree of human input. The human expert is part of the system, and hence the demonstrated degree of performance of a method employed by one forensic practitioner may not be the same as when the same method is employed by another forensic practitioner.

If *forensic_eval_01* is judged a success, then in the future we plan to run subsequent evaluations based on the conditions of other real forensic cases.

2. Training and test data

The training and test data consist of audio recordings which, based on careful selection of data and a simulation of case conditions, are intended to represent the relevant population and reflect the speaking styles and recording conditions found in an actual forensic voice comparison case. The case involved a questioned-speaker recording (recording of a speaker of questioned identity, the offender) in which the speech signal had been transmitted via a landline telephone to a call centre, there was background office noise (babble and typing noises) at the call centre, and the recording was saved in a compressed format. The call included verbal exchange of information, i.e., names, addresses, numbers, and letters, and the duration of speech from the speaker of interest was 46 s. The known-speaker recording (recording of a speaker of known identity, the suspect) was of a police interview recorded in a room with substantial reverberation, there was ventilation-system noise, and the recording was saved in a different compressed format. The speaker on each recording was clearly an adult male who spoke English with an Australian accent.

The training and test data are similar to those previously described in Enzinger et al. (2016).[2] Enzinger et al. (2016) described in detail the procedures used to select training and test data representative of the relevant population and speaking styles, and the signal processing procedures used to simulate the technical recording conditions of the known-speaker and questioned-speaker recordings. The training and test recordings came from a database of audio recordings of 500+ Australian English speakers (Morrison et al., 2015). The database included multiple non-contemporaneous recordings of most speakers (about half had three recording sessions, a quarter two recording sessions, and a quarter one recording session), and each recording session included a simulated police interview task and a task which involved exchange of information over the telephone.[3] The original recordings consisted of high quality audio recorded in sound treated booths with the speakers wearing head mounted microphones. In the information exchange task there was one speaker per channel. Each speaker was in a different sound booth and they communicated via a telephone system. In the interview task the interviewer was face to face with the interviewee, but relatively far from the microphone and they avoided speaking at the same time as the interviewee. See Morrison et al. (2012) for additional details of the data collection protocol.

Inter-utterance pauses and any interlocutor speech and extraneous noises were removed using automatic procedures followed

[1] Likewise, if an analysis based on a system which has not been tested in *forensic_eval_01* is proffered to a court, we recommend that the court consider whether the degree of validity and reliability of that system has been empirically tested using data that are sufficiently representative of the relevant population and sufficiently reflective of the conditions in that particular case, and that no inference be drawn simply from the fact that it was not tested in *forensic_eval_01*.

[2] The differences are as follows: In Enzinger et al. (2016) the data consisted of specified numbers of feature vectors (MFCCs + deltas): 4,137 feature vectors were extracted from each questioned-speaker-condition recording and 10,452 from each known-speaker-condition recording. In *forensic_eval_01*, the data are provided as concatenated utterances of total duration 46 s for each questioned-speaker-condition recording and 125.694 s for each known-speaker-condition recording. The length of each utterance and the number of utterances concatenated vary from recording to recording, and there are pauses between each utterance which are nominally 200 ms long. Even if the same feature extraction process as was used in Enzinger et al. (2016) were applied to the *forensic_eval_01* recordings, the exact number of feature vectors extracted from the recordings would therefore differ from recording to recording and from the numbers of feature vectors used in Enzinger et al. (2016). Also, in Enzinger et al. (2016) a questioned-speaker-condition recording was inadvertently omitted for one of the 61 speakers in the test set. In the *forensic_eval_01* test data, each and every speaker has a questioned-speaker-condition recording.

[3] A whole recording session usually lasted less than an hour and included, among other things, both the information exchange task and the interview task.

by manual confirmation and correction. The portions of a recording corresponding to the utterances (plus 100 ms before and after) were then concatenated into a single recording. Signal processing techniques were applied to replicate the effects of transmission through telephone systems, of being saved in compressed file formats, and also to add noise and reverberation (a detailed description of these procedures is provided in Enzinger et al., 2016). This resulted in one set of recordings which reflected the speaking style and recording conditions of the questioned-speaker recording from the case, and another set which reflected the speaking style and recording conditions of the known-speaker recording from the case. The questioned-speaker-condition recordings were truncated to 46 s long and the known-speaker-condition recordings to 125.694 s long (368,000 and 5,543,105 samples at sampling frequencies of 8000 and 44,100 Hz respectively, the latter being the length of the shortest interview-task recording available in the database).

The training and test data for use in the evaluation came from a total of 166 speakers, 88 recorded in three non-contemporaneous recordings sessions (at intervals of approximately a week), 35 recorded in two non-contemporaneous recordings sessions, and 44 recorded in only one session.[4] Training data consist of a total of 423 recordings from 105 speakers (191 recordings in questioned-speaker condition and 232 in known-speaker condition), and the test data consist of a total of 223 recordings from 61 speakers (61 recordings in questioned-speaker condition and 162 in known-speaker condition). There is only one questioned-speaker-condition recording per test-set speaker, which is always from the first recording session. Each questioned-speaker-condition recording in the test set (from session 1) is to be compared with each available session 2 and session 3 known-speaker-condition recording from the same speaker (all same-speaker comparisons are non-contemporaneous), and with each available session 1, session 2, and session 3 known-speaker-condition recording from each and every other speaker in the test set (every speaker in the test set has recordings from at least two recording sessions). This results in 111 same-speaker pairs of recordings (from 61 unique speakers), and 9720 different-speaker pairs of recordings (from 3660 unique pairs of speakers, counting speaker A in questioned-speaker condition versus speaker B in known-speaker condition separately from speaker B in questioned-speaker condition versus speaker A in known-speaker condition).

3. Evaluation rules, instructions, and materials

1. Participants must commit to make public a description of their systems and how they used those systems, and to make public the test results irrespective of how those results may turn out.
2. The evaluation is conducted on an honour basis. Participants must not engage in any activities which could be considered cheating.
3. Those who wish to participate should go to http://databases.forensic-voice-comparison.net/#forensic_eval_01 and click on the link to indicate their agreement to abide by the rules and to request a user name and password to allow them to access the evaluation materials.
   a. The evaluation materials will only be made available to those who agree to abide by the rules of the evaluation.
   b. The evaluation materials must not be redistributed and must not be used for commercial purposes (if there is any doubt as to whether a proposed use would constitute commercial use, please contact the organizers of the evaluation for clarification).
4. Before submitting the test results for their system or systems (Rule 12 below) participants must write and submit a manuscript including the introduction and methodology sections.
   a. Submit the manuscript to the VSI via the *Speech Communication* editorial system: http://www.evise.com/evise/faces/pages/navigation/NavController.jspx?JRNL_ACR=SPECOM
   b. Select "SI: forensic_eval_01" as the article type.
   c. The manuscript will be sent out for review, a decision of either "reject" or "revise and resubmit" will be made by the editors, and reviewers' comments will be returned to the participant.
   d. Assuming the decision was "revise and resubmit", the participant should complete the evaluation of their system or systems, add the results (Rule 13), and discussion and conclusion, and submit the revised version of their manuscript via the editorial system (Rule 14).
5. The evaluation materials consist of the following:
   a. "forensic_eval_01_train.zip" (2.2GB) contains ".wav" recordings in questioned-speaker and known-speaker conditions. These may be used for training and optimization.
   b. "forensic_eval_01_test.zip" (1.6GB) contains ".wav" recordings in questioned-speaker and known-speaker conditions. These must only be used to test the system (see cross-validation exception in Rule 7 below).
   c. Questioned-speaker-condition recordings are named using the following convention:
      i. nnnn(m)_fax.wav
      ii. "nnnn" is a four digit speaker ID, and "m" is a single digit recording session ID.
   d. Known-speaker-condition recordings are named using the following convention:
      i. nnnn(m)_int.wav
      ii. "nnnn" is a four digit speaker ID, and "m" is a single digit recording session ID.
   e. "forensic_eval_01_train_boundarylabels.zip" and "forensic_eval_01_test_boundarylabels.zip" contain markers indicating the boundaries between the utterances which were concatenated to make the ".wav" files.
      i. Participants may make use of these markers. This is optional, not required.
      ii. The ".zip" archives contain Matlab files, extension ".mat", one corresponding to each ".wav" file and using the same naming convention as for the ".wav" files. These files contain markers indicating the start of each utterance (variable name: "section_boundaries" which contains numeric values indicating number of samples from the beginning of the ".wav" file), plus markers to indicate any sections which have been zeroed out (variable names: "zero_start_boundaries", "zero_end_boundaries").
      iii. Channel effects and background noise (and reverberation for known-speaker-condition recordings) have been added across utterance boundaries and over the zeroed out sections. Speech nominally begins 100 ms after an utterance boundary marker, and ends 100 ms before the next utterance boundary marker. There are no such gaps before or after zeroed out sections. Speech begins 100 ms after the beginning of the recording. Since the training and test data were simply truncated when they reached 368,000 and 5,543,105 samples for questioned-speaker-condition and known-speaker-condition recordings respectively, there is no extra 100 ms at the end of the truncated recordings.

[4] Since non-contemporaneous same-speaker test comparisons could not be constructed for the 44 speakers who only participated in one recording session, recordings from all these speakers were included in the training set.

   f. "forensic_eval_01_test_pairs.csv" is a spreadsheet file containing a list of pairs of test recordings to be compared.
      i. The file name in the first column is that of the questioned-speaker-condition recording.
      ii. The file name in the second column is that of the known-speaker-condition recording.
6. The training data provided, and any other data participants may have, can be used to train and optimize their systems.
   a. Participants may make any use they wish of the training data, but are advised to avoid over-training or over-optimizing their systems. In particular, participants should not test a large factorial combination of different system settings and submit a large number of results files to the organizers. There may be a limited number of system variants that a participant has a priori reasons to want to test, and when writing their paper they should justify their choice.
7. Test data must not be used to train the system, with the following exception: Cross-validated use of the test data to train a final score to likelihood ratio conversion, aka calibration, is allowed (it is optional, it is not required).
   a. If used, cross validation must leave out all recordings from the speaker or speakers being tested (the one speaker for a same-speaker comparison or the two speakers for a different-speaker comparison), not just the two recordings actually being compared.
8. Under no circumstances should the test data be used for optimizing the system. Cross validation using test data is not allowed for system optimization.
9. Participants will have access to speaker IDs for the test set. Participants must not make any use of this information when training or testing their system (apart from to ensure that the same data are not used for training and testing if using cross-validation).
   a. If a participant tests a system in which the strength of evidence conclusion is based on human judgement, they must devise a procedure to mask the test-set speaker IDs from the human practitioner.
10. Participants must produce a results file which is a copy of the two columns of "forensic_eval_01_test_pairs.csv" plus an additional third column which contains the test results.
    a. The result from comparing each test pair must appear in the corresponding row.
    b. A test result must be entered for every row.
    c. Each result must be entered as a log-base-10 likelihood ratio.
    d. There is no restriction on the number of significant figures to enter.
11. The results file must be named "forensic_eval_01_results_ParticipantID_SystemID_yyyy-mm-ddx.csv", where "ParticipantID" should be the (short) name of the participating laboratory, and "SystemID" should be the (short) name of the system tested. "yyyy-mm-ddx" is a version control string which consists of a four digit year, a two digit month, a two digit day, and an alphabetic within-day code ("a" for the first version of that day, "b" for the second, etc.), e.g., "2017-01-01a".
12. The results files must be e-mailed directly to the organizers at <forensic_eval_01@forensic-evaluation.net>
    a. This is required so that the organizers can ensure that performance metrics and graphics are calculated and drawn in exactly the same way for all systems submitted by all participants.
    b. Participants must not submit results files until after they have received a "revise and resubmit" decision on their manuscript.
    c. A separate file must be sent for each system tested.
    d. A participant must send results files for all systems or system variants tested all at once. They must not submit the results from a system, get results, modify the system, and then submit a new set or results.
    e. Participants may submit the results of and write a paper based on one system at one point in time and then submit the results of and write a paper based on a substantially different or substantially revised system at another point in time. Permissible examples could include an i-vector system versus a GMM-UBM system, or a major new version of a piece of commercial software. If, however, a participant's practice is to use multiple systems concurrently and fuse the results, they should submit the results from the component systems and/or fused system all together rather than as a series of papers. The organizers and editors reserve the right to reject submissions which they consider to be trivial variants of systems previously submitted by the same participant.
13. After the organizers have received the results file from a participant, the organizers will generate and provide that participant with the performance metrics and graphics based on their results.
    a. The organizers hope to be able to provide each participant with their results within a few days of submission, but this may take longer depending on other commitments the organizers may have at the time.
14. After the participant has received the performance metrics and graphics, they must produce a revised version of their manuscript including a results section and discussion and conclusions, and submit it, along with a list of changes and rebuttals, via the *Speech Communication* editorial system.

4. Evaluation metrics and graphics

In the literature on evaluation of forensic evidence in general and forensic voice comparison in particular there are a number of different metrics and graphics used for reporting the results of empirical tests of system performance. Some of these may be intended only for technical system development purposes, but others may also be intended for presentation in court (along with adequate explanation). For *forensic_eval_01* we have chosen to use a selection of popular metrics and graphics, without necessarily endorsing or recommending the use of particular metrics and graphics. We have included equal error rate (EER), although we do not think it is consistent with the forensic practitioner's role in the likelihood ratio framework. We have included metrics and graphics related to the precision of the likelihood ratio output of the forensic systems. It is a matter of some controversy whether this is appropriate and there is currently an ongoing debate on the topic (Morrison, 2016). We have also included metrics and graphics which would be acceptable to those who take the position that assessing precision is not appropriate. In order to allow comparison across all papers in the VSI, all the metrics and graphics listed below must be included in the reported results from all systems (authors of individual papers do not have the option of not including particular metrics and graphics, but they may add comments regarding their opinion of the suitability of particular metrics and graphics).

Examples are provided of each type of graphical plot listed below, and a table shows example values for each metric. The data used to create these plots and table were artificial and generated for illustrative purposes only. Results from two artificial systems are shown, one with better and one with poorer performance.

Table 1
Exact values of performance metrics.

System	C_{llr}^{pooled}	C_{llr}^{mean}	95% CI	C_{llr}^{min}	C_{llr}^{cal}	EER
1	0.548	0.529	0.498	0.475	0.073	0.150
2	0.101	0.071	0.988	0.080	0.022	0.026

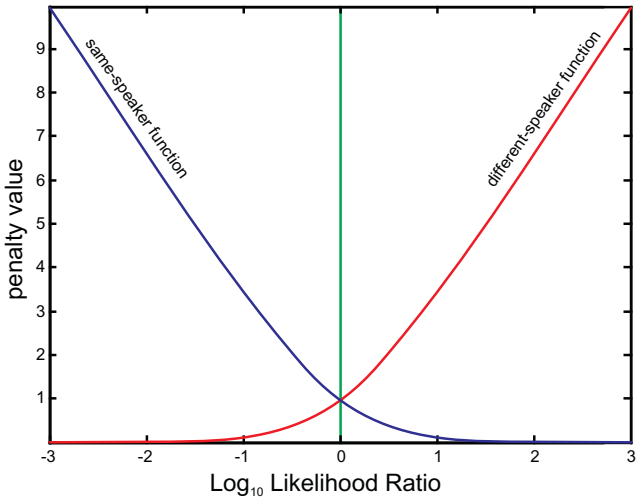

Fig. 1. Plot of log likelihood ratio versus penalty value for calculating C_{llr}.

Metrics:

Example values for all the metrics are given in Table 1.

- Log likelihood ratio cost (C_{llr}^{pooled})[5]

  This is a single value summary of system performance. If the test pair is a same-speaker pair the higher the likelihood ratio the better the performance. A penalty value is assigned to the result from each same-speaker comparison. This penalty value is small for large positive log likelihood ratios (likelihood ratios much greater than 1), high for large negative log likelihood ratios (likelihood ratios much less than 1), and intermediate for log likelihood ratios close to 0 (likelihood ratios close to 1). Mutatis mutandis for a different-speaker pair, for which the smaller the likelihood ratio the better the performance. Fig. 1 shows a plot of log likelihood ratios versus penalty functions. A penalty value is assigned to the result from each test comparison, and C_{llr}^{pooled} is an average of all the penalty values. The formula for calculating C_{llr}^{pooled} is given in Eq. 1. A system which provides no useful information and always responds with a likelihood ratio of 1 will result in a C_{llr}^{pooled} value of 1. As performance improves the value of C_{llr}^{pooled} decreases towards 0. Smaller C_{llr}^{pooled} values indicate better performance. Descriptions of this metric can be found in Brümmer and du Preez (2006), van Leeuwen and Brümmer (2007), González-Rodríguez, et al. (2007), Morrison (2011), Drygajlo et al. (2015), Meuwly et al. (2016). Calculations will be performed using the implementation in the FoCal toolkit (Brümmer, 2005).

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

LR_{ss} and LR_{ds} are likelihood ratios calculated for same-speaker and different-speaker test pairs, respectively. N_{ss} and N_{ds} are the numbers of same-speaker and different-speaker comparisons.

- 95% credible interval (95% CI)

  This is a measure of the precision (reliability) of the output of the system. The test data include groups of comparisons consisting of a single questioned-speaker-condition recording versus multiple known-speaker-condition recordings where the latter all belong to the same speaker (who may be the same as the questioned speaker or a different speaker from the questioned speaker). The 95% CI measures the variability of the resulting multiple likelihood ratios corresponding to each group, including multiple known-speaker recordings but always the same pair of speakers. The 95% CI is averaged over all groups (including same-speaker and different-speaker pairs). 95% CI values will be reported on a scale of ± orders of magnitude (= $\log_{10}$ scale). This metric will be calculated using the parametric procedure described in Morrison (2011).

- Log likelihood ratio cost, accuracy only (C_{llr}^{mean})

  This is a measure of the accuracy (validity) of the output of the system. This is the same as the C_{llr}^{pooled} metric, but whereas all the test results were pooled to calculate C_{llr}^{pooled}, for C_{llr}^{mean} the calculations were performed on the means of the groups defined in the description of the 95% CI metric. Whereas C_{llr}^{pooled} conflates accuracy and precision, C_{llr}^{mean} and 95% CI attempt to separate out accuracy and precision (if there were no groups, C_{llr}^{pooled} and C_{llr}^{mean} would be identical metrics of accuracy, and precision would not be measurable). This metric is described in Morrison (2011).

- Discrimination loss (C_{llr}^{min}) and calibration loss (C_{llr}^{cal})

  The minimum log likelihood ratio cost (C_{llr}^{min}) is the same as C_{llr}^{pooled}, but calculated after the likelihood ratio values from the test results have been optimised using the non-parametric pool-adjacent-violators (PAV) algorithm. Note that this optimisation is the result of training and testing on the same data. The level of performance indicated by C_{llr}^{min} is not, therefore, the level of performance expected when the system is applied to new test data. C_{llr}^{min} is called "discrimination loss". The difference between C_{llr}^{min} and C_{llr}^{pooled} is called "calibration loss" (C_{llr}^{cal}). Descriptions of these metrics can be found in Brümmer and du Preez (2006), van Leeuwen and Brümmer (2007), González-Rodríguez, et al. (2007), Drygajlo et al. (2015), Meuwly et al. (2016).

- Equal error rate (EER)

  If the likelihood ratio test results are combined with prior odds to produce posterior odds and a same-speaker versus different-speaker decision is made by imposing a threshold on the posterior odds, some proportion of the results from same-speaker comparisons will be classified as different-speaker (misses or false rejections) and some proportion of the results from different-speaker comparisons will be classified as same-speaker (false alarms or false acceptances). In general, the priors and/or decision threshold can be adjusted so that the proportion of false alarms decreases but the proportion of misses concomitantly increases, or vice versa misses decrease but false alarms increase. If the priors and/or decision threshold are adjusted such that the proportion of false alarms and the proportion of misses are equal, the resulting proportion is the equal error rate. EER will be calculated using the Receiver Operator

[5] Many publications simply use the abbreviation C_{llr}, we add the superscript suffix to unambiguously indicate C_{llr}^{pooled} as opposed to C_{llr}^{mean}, see below.

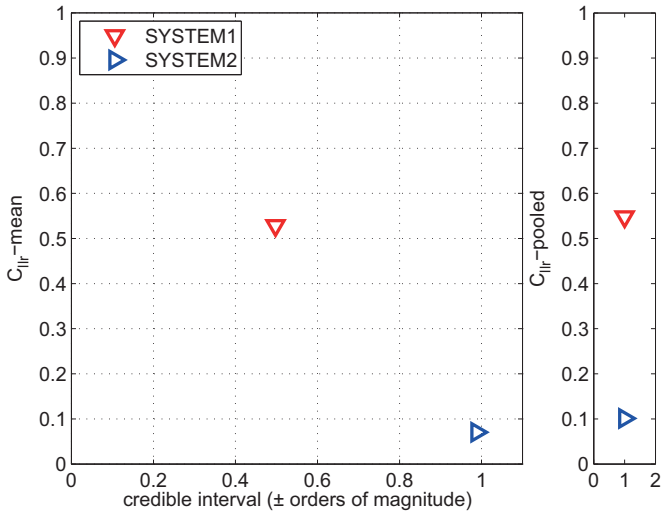

Fig. 2. Example plot showing C_{llr}^{mean} versus 95% CI (left panel) and C_{llr}^{pooled} (right panel).

Characteristic Convex Hull method (see Brümmer and de Villiers, 2013).

Graphics:

- Accuracy and precision metric plot

  An example of this type of plot is given in Fig. 2. The left panel consists of a two dimensional scatter plot with C_{llr}^{mean} on the *y* axis and 95% CI on the *x* axis. The performance of each system is represented as a point in this two-dimensional space. The right panel is one dimensional and plots a point corresponding to the C_{llr}^{pooled} value of each system.

- Tippett plot (no precision)

  Examples of this type of plot are given in Fig. 3. For results from same-speaker comparisons the *y* axis indicates the cumulative proportion of log likelihood ratios with values equal to or less than the value on the *x* axis. For results from different-speaker comparisons the *y* axis indicates the cumulative proportion of log likelihood ratios with values equal to or greater than the value on the *x* axis. Rather than plotting individual points as such, it is customary to draw a line between each point, resulting in a curve which rises to the right for same-speaker results and to the left for different speaker results. As an empirical plot of results from all test trials, a Tippett plot is information rich. In general, however, the further to the right the same-speaker curve and the further to the left the different-speaker curve, the better to the performance of the system. EER can be read off as the *y* axis value corresponding to the point where the two curves cross. This graphic is described in Meuwly (2001), Morrison (2010), Drygajlo et al. (2015), Meuwly et al. (2016).

- Tippett plot (with precision)

  Examples of this type of plot are given in Fig. 4. This is a variant of the Tippett plot in which instead of pooling all the same-speaker results and pooling all the different-speaker results to draw the two curves, the solid curves are based on the mean values from each of the groups defined in the description of the 95% CI metric. The dashed lines to the left and right of the solid lines indicate the 95% CIs. This graphic is described in Morrison et al. (2010).

- Detection error tradeoff (DET) plot

  An example of this type of plot is given in Fig. 5. The following description is deliberately written in a way which relates the DET plot to the Tippett plot. The Tippett plot (no precision) has log likelihood ratio values indicated on the *x* axis. If one were to use equal priors (prior odds of 1, log prior odds of 0), the value of the log posterior odds would be the same as the log likelihood ratio, and the *x* axis could be re-labelled log posterior odds. One could then sweep a decision threshold along the *x* axis and at each point read off the false alarm rate and miss rate as the values on the *y* axis corresponding to the same-speaker curve and different-speaker curve respectively. For this purpose, prior odds and decision thresholds are interchangeable, increasing the prior odds and decreasing the threshold value have the same effect on false alarm and miss rates. The decision threshold could be fixed at log posterior odds of zero and the *x* axis could be re-labelled as minus log prior odds (increasing the prior odds would move the log posterior probability curves to the right, but instead the curves can be left where they are and the *x* axis re-labelled as log prior odds in-

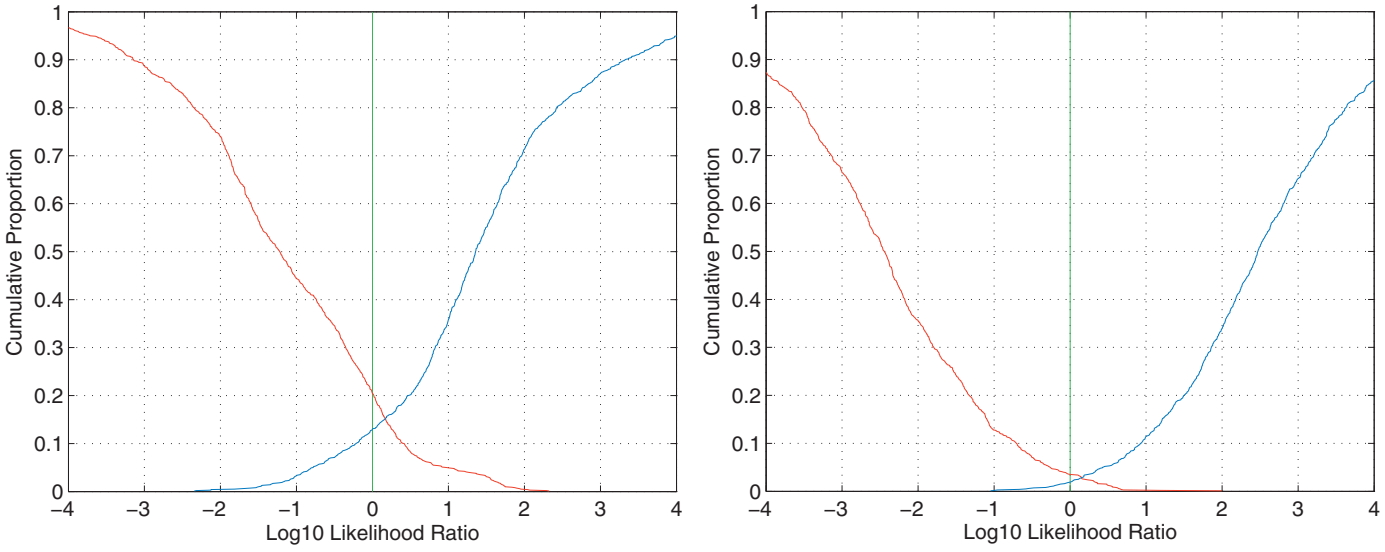

Fig. 3. Examples of Tippett plots (no precision). Left panel: System 1. Right panel: System 2.

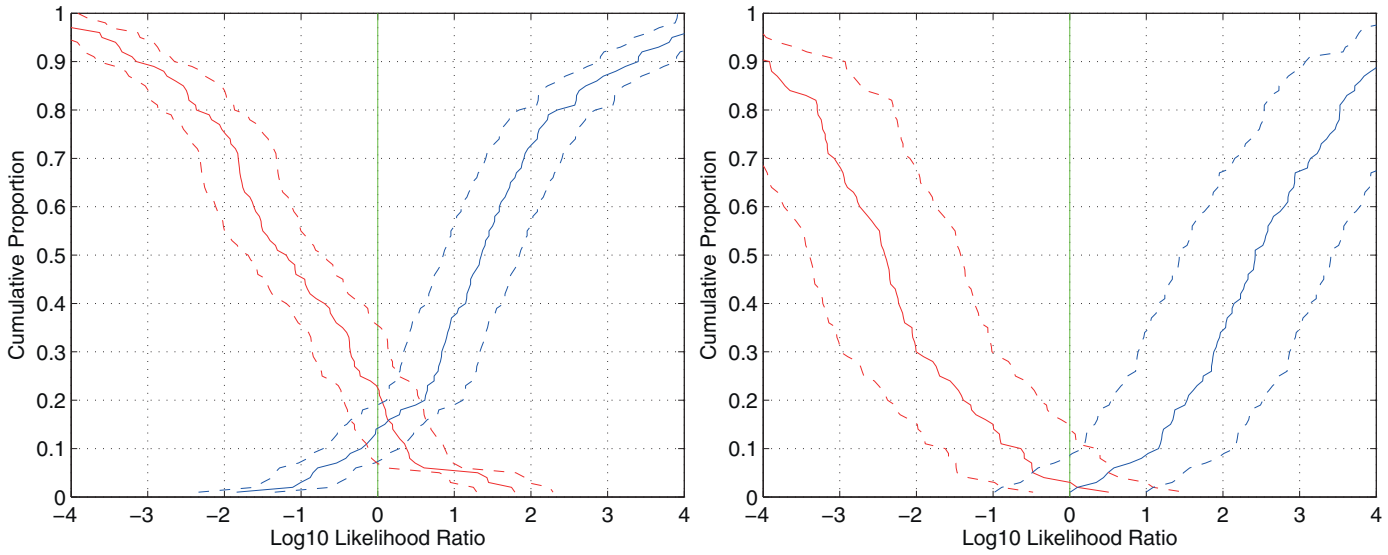

Fig. 4. Examples of Tippett plots (with precision). Left panel: System 1. Right panel: System 2.

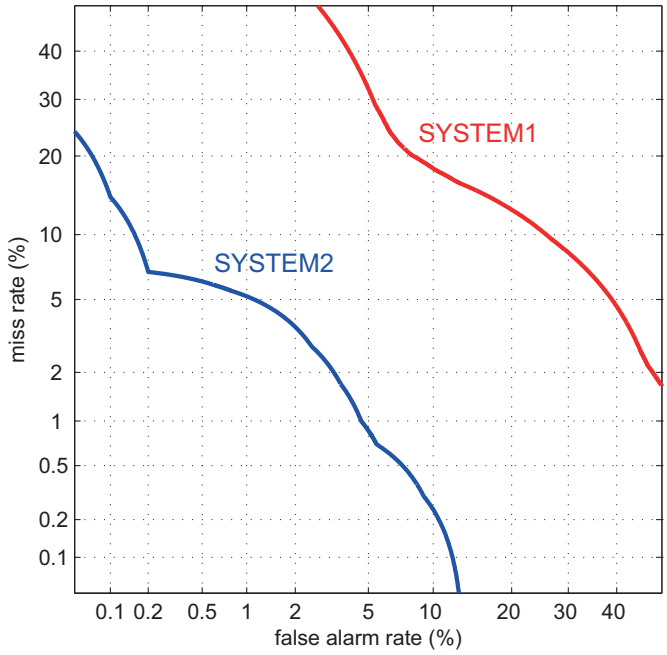

Fig. 5. Example of a DET plot.

creasing to the left). In general, the sweep can cover a range of combinations of prior odds and thresholds that a trier of fact could potentially use. A DET plot does not show the prior odds or decision threshold values, but shows the resulting combinations of false alarm and miss rates. The rates are plotted on logarithmic scales resulting in a roughly straight diagonal line (top left to bottom right) for each system. In general, the closer the line to the origin the better the performance of the system. ERR can be read off from either axis as the value corresponding to the point where the results line crosses a diagonal bottom left to top right line which passes through the origin. This graphic is described in Martin et al. (1997), Drygajlo et al. (2015), Meuwly et al. (2016). The plots are drawn using the Receiver Operator Characteristic Convex Hull method (see Brümmer and de Villiers, 2013).

- Empirical cross entropy (ECE) plot

Examples of this type of plot are given in Fig. 6. $C_{\text{llr}}^{\text{pooled}}$ is calculated using likelihood ratios, this is equivalent to using posterior odds if the prior odds are even (log prior odds are 0). ECE is an extension of $C_{\text{llr}}^{\text{pooled}}$ calculated on posterior odds using a specified prior odds value and the likelihood ratio output from the test trials. The formula for calculating ECE is given in Eq. 2. The ECE plot is produced by sweeping over a range of prior odds and plotting the resulting ECE values, log prior odds are indicated on the x axis and ECE on the y axis. The solid curve on an ECE plot represents ECE values calculated using the likelihood ratios from the test results (its height at log prior odds = 0 is $C_{\text{llr}}^{\text{pooled}}$). The dotted curve represents the ECE values calculated using a system which always outputs a likelihood ratio of 1. The dashed curve represents the ECE values after the likelihood ratios from the test results have been optimised using the non-parametric pool-adjacent-violators (PAV) algorithm (its height at log prior odds = 0 is $C_{\text{llr}}^{\text{min}}$). Note that this optimisation is the result of training and testing on the same data. The level of performance indicated by the dashed line is not therefore the level of performance of the actual system tested, the performance of the actual system tested is indicated by the solid line. ECE plots can reveal calibration problems. A miscalibrated system may work well within a particular range of prior-odds values, but poorly within a different range of prior-odds values – performance within the latter range could be even worse than a system that gave no information and always responded with a likelihood ratio of 1. This graphic is described in Ramos Castro (2007), Ramos and González-Rodríguez (2013), Ramos et al. (2013), Drygajlo et al. (2015), Meuwly et al. (2016).

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

P_{ss} and P_{ds} are the prior probabilities of the same-speaker and different-speaker hypotheses. LR_{ss}, LR_{ds}, N_{ss}, and N_{ds} are as previously defined.

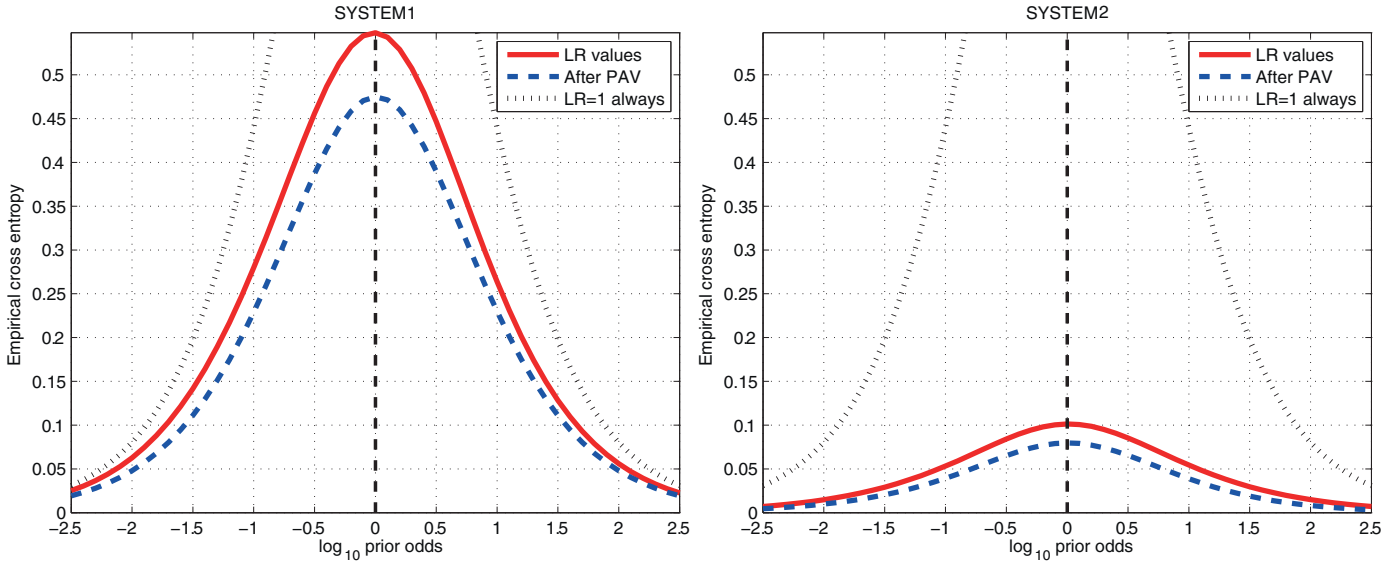

Fig. 6. Examples of ECE plots. Left panel: System 1. Right panel: System 2.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Brümmer, N. 2005. FoCal toolkit. http://www.dsp.sun.ac.za/nbrummer/focal.

Brümmer, N., de Villiers, E., 2013. The BOSARIS toolkit: theory, algorithms and code for surviving the new DCF. arXiv:1304.2865.

Brümmer, N., du Preez, J., 2006. Application independent evaluation of speaker detection. Comput. Speech Lang. 20, 230–275. http://dx.doi.org/10.1016/j.csl.2005.08.001.

Drygajlo, A., Jessen, M., Gfroerer, S., Wagner, I., Vermeulen, J., Niemi, T., 2015. Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition, Including Guidance on the Conduct of Proficiency Testing and Collaborative Exercises. European Network of Forensic Science Institutes, Wiesbaden, Germany http://www.enfsi.eu/documents/methodological-guidelines-best-practice-forensic-semiautomatic-and-automatic-speaker.

Enzinger, E., Morrison, G.S., Ochoa, F., 2016. A demonstration of the application of the new paradigm for the evaluation of forensic evidence under conditions reflecting those of a real forensic-voice-comparison case. Sci. Justice 56, 42–57. http://dx.doi.org/10.1016/j.scijus.2015.06.005.

Forensic Science Regulator, 2014. Codes of practice and conduct for forensic science providers and practitioners in the criminal justice system (Version 2.0). Forensic Science Regulator, Birmingham, England. https://www.gov.uk/government/publications/forensic-science-providers-codes-of-practice-and-conduct-2014.

González-Rodríguez, J., Rose, P., Ramos, D., Toledano, D.T., Ortega-García, J., 2007. Emulating DNA: rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. IEEE Trans. Audio Speech Lang. Process. 15, 2104–2115. http://dx.doi.org/10.1109/TASL.2007.902747.

van Leeuwen, D.A., Brümmer, N., 2007. An introduction to application-independent evaluation of speaker recognition systems. In: Müller, C. (Ed.), Speaker Classification I. Fundamentals, Features, and Methods, 2007. Springer-Verlag, Heidelberg, Germany, pp. 330–353. http://dx.doi.org/10.1007/978-3-540-74200-5_19

Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M., 1997. The DET curve in assessment of detection task performance. In: Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech). Rhodes, Greece, pp. 1895–1898.

Meuwly, D., 2001. Reconnaissance de locuteurs en sciences forensiques: l'apport d'une approche automatique. University of Lausanne PhD dissertation.

Meuwly, D., Ramos, D., Haraksim, R., 2016. A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. Forensic Sci. Int. http://dx.doi.org/10.1016/j.forsciint.2016.03.048.

Morrison, G.S., 2010. Forensic voice comparison. In: Freckelton, I., Selby, H. (Eds.), Expert Evidence. Thomson Reuters, Sydney, Australia ch. 99 http://expert-evidence.forensic-voice-comparison.net.

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

Morrison, G.S., 2014. Distinguishing between forensic science and forensic pseudoscience: testing of validity and reliability, and approaches to forensic voice comparison. Sci. Justice 54, 245–256. http://dx.doi.org/10.1016/j.scijus.2013.07.004.

Morrison, G. S., 2016. Special issue on measuring and reporting the precision of forensic likelihood ratios: Introduction to the debate. Sci. Justice. http://dx.doi.org/10.1016/j.scijus.2016.05.002.

Morrison, G.S., Rose, P., Zhang, C., 2012. Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice. Aus. J. Forensic Sci. 44, 155–167. http://dx.doi.org/10.1080/00450618.2011.630412.

Morrison, G.S., Thiruvaran, T., Epps, J., 2010. Estimating the precision of the likelihood-ratio output of a forensic-voice-comparison system. In: Cernocký, H., Burget, L. (Eds.), Proceedings of Odyssey 2010: The Language and Speaker Recognition Workshop. Brno, Czech Republic. International Speech Communication Association, pp. 63–70.

Morrison, G.S., Zhang, C., Enzinger, E., Ochoa, F., Bleach, D., Johnson, M., Folkes, B.K., De Souza, S., Cummins, N., Chow, D., 2015. Forensic database of voice recordings of 500+ Australian English speakers http://databases.forensic-voice-comparison.net/#australian_english_500.

National Commission on Forensic Science, 2016. Universal accreditation. Policy Recommendation. https://www.justice.gov/ncfs/file/477851/download.

National Research Council, 2009. Strengthening Forensic Science in the United States: A Path Forward. National Academies Press, Washington, DC http://www.nap.edu/catalog.php?record_id=12589.

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

Ramos, D., González-Rodríguez, J., 2013. Reliable support: measuring calibration of likelihood ratios. Forensic Sci. Int. 230, 156–169. http://dx.doi.org/10.1016/j.forsciint.2013.04.014.

Ramos, D., González-Rodríguez, J., Zadora, G., Aitken, C.G.G., 2013. Information-theoretical assessment of the performance of likelihood ratio computation methods. J. Forensic Sci. 58, 1503–1518. http://dx.doi.org/10.1111/1556-4029.12233.

Stoel, R.D., Mattijssen, E.J.A.T., Berger, C.E.H., 2016. Building the research culture in the forensic sciences: announcement of a double blind testing program. Sci. Justice 56, 155–156. http://dx.doi.org/10.1016/j.scijus.2016.04.003.