SPEECH
COMMUNICATION
An international journal of the European Association for Signal Processing (EURASIP)
and of the International Speech Communication Association (ISCA)

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

Geoffrey Stewart Morrison , Ewald Enzinger

PII: S0167-6393(19)30219-5
DOI: https://doi.org/10.1016/j.specom.2019.06.007
Reference: SPECOM 2654

To appear in: *Speech Communication*

Received date: 11 June 2019
Revised date: 26 June 2019
Accepted date: 26 June 2019

Highlights

- Validation of forensic voice comparison systems under casework conditions.
- Summary of published papers.
- Observations on results.
- Reflections on aims and process.
- Acknowledgments.

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

Geoffrey Stewart Morrison [1,2,*], Ewald Enzinger [3]

[1] Forensic Speech Science Laboratory, Institute for Forensic Linguistics, and Centre for Forensic Data Science, Department of Computer Science, Aston University, Birmingham, England, United Kingdom

[2] Forensic Evaluation Ltd, Birmingham, England, United Kingdom

[3] Eduworks Corporation, Corvallis, Oregon, United States of America

★ 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*), [https://www.sciencedirect.com/journal/speech-communication/special-issue/10KTJHC7HNM], Guest Edited by G.S. Morrison and E. Enzinger.

* Corresponding author. E-mail address: geoff-morrison@forensic-evaluation.net

Abstract

This conclusion to the virtual special issue (VSI) “Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (*forensic_eval_01*)” provides a brief summary of the papers included in the VSI, observations based on the results, and reflections on the aims and process. It also includes errata and acknowledgments.

Keywords: Forensic voice comparison; Evaluation; Validity; Reliability; Casework conditions

Declarations of interest: none

Abbreviations

CI	credible interval
C_{llr}	log-likelihood-ratio cost
DNN	deep neural network
EER	equal error rate
GMM	Gaussian mixture model
MSR	Microsoft Research
NIST SRE	National Institute of Standards and Technology Speaker Recognition Evaluation
PLDA	probabilistic linear discriminant analysis
SID	speaker identification
UBM	universal background model
VOCALISE	Voice Comparison and Analysis of the Likelihood of Speech Evidence
VSI	virtual special issue

1. Summary

The present paper serves as a conclusion to the virtual special issue (VSI) “Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (*forensic_eval_01*)”. A set of training data and a set of test data reflecting the conditions of a real forensic case were made available, rules for participation were published in the introduction (Morrison & Enzinger, 2016), and participants were asked to use these data to empirically validate the performance of their forensic voice comparison systems. All papers submitted to the VSI described validations of systems based on automatic speaker recognition technology.

Besides the introduction and the conclusion, the VSI consists of the following papers:

1. van der Vloed (2016, 2017)
   – An evaluation of Batvox 4.1, a commercial GMM i-vector PLDA system. Different ways of optimizing the system using case-specific data were tested.

2. Silva & Medina (2017)
   – An evaluation of the MSR Identity Toolbox, an open source toolbox released by Microsoft Research. Both GMM-UBM and GMM i-vector PLDA systems were evaluated. The systems were trained exclusively on case-specific data. Use of different feature-domain mismatch compensation techniques were tested.

3. Zhang & Tang (2018)
   – An evaluation of Batvox 3.1, a commercial GMM-UBM system. System optimization using different amounts of case-specific data was tested.

4. Jessen, Meir, Solewicz (2019)
   – An evaluation of commercial systems: Nuance Forensics 9.2, a GMM i-vector PLDA system; and Nuance Forensics 11.1, a GMM i-vector + DNN senone posterior i-vector PLDA system. Different ways of optimizing the systems using case-specific data were tested.

5. Jessen, Bortlík, et al. (2019)
   – An evaluation of commercial systems: Phonexia SID-XL3, a GMM i-vector + DNN bottleneck PLDA system; and Phonexia SID-BETA4, a DNN embedding (x-vector) PLDA system. Uncalibrated outputs of these systems were compared with outputs that had been normalized-calibrated using models trained on case-specific data.

6. Kelly et al. (2019)

   – An evaluation of commercial systems: VOCALISE 2017B, a GMM i-vector PLDA system; and VOCALISE 2019A-Beta-RC1, a DNN embedding (x-vector) PLDA system. System variants were (a) trained on non-case-specific data, (b) trained on non-case-specific data and optimized using case-specific data, and (c) trained on case-specific data only (all included calibration trained on case-specific data).

The performance metrics from the best-performing variant of each system (best performing in terms of $C_{\mathrm{llr}}^{\mathrm{pooled}}$)[1] are presented in Table 1. For the full sets of performance metrics and graphics, see the individual papers.

Table 1. Performance metrics for the best-performing variant of each system.

System	Type	$C_{\mathrm{llr}}^{\mathrm{pooled}}$	$C_{\mathrm{llr}}^{\mathrm{mean}}$	95% CI	$C_{\mathrm{llr}}^{\mathrm{min}}$	$C_{\mathrm{llr}}^{\mathrm{cal}}$	EER
Batvox 3.1	GMM-UBM	0.593	0.473	1.130	0.396	0.198	0.126
MSR GMM-UBM	GMM-UBM	0.576	0.549	0.368	0.444	0.132	0.139
MSR GMM i-vector	GMM i-vector	0.449	0.437	0.479	0.301	0.148	0.085
Batvox 4.1	GMM i-vector	0.365	0.304	1.156	0.317	0.048	0.096
Phonexia XL3	DNN bottleneck	0.294	0.225	1.160	0.231	0.063	0.066
Nuance 9.2	GMM i-vector	0.285	0.258	0.336	0.161	0.124	0.047
VOCALISE 2017B	GMM i-vector	0.267	0.230	1.178	0.239	0.029	0.070
Nuance 11.1	DNN senone	0.255	0.234	0.309	0.124	0.130	0.031
VOCALISE 2019A	x-vector	0.246	0.213	1.040	0.189	0.057	0.053
Phonexia BETA4	x-vector	0.208	0.163	0.779	0.098	0.110	0.022

[1] The Phonexia x-vector variant shown in Table 1 had the second best $C_{\mathrm{llr}}^{\mathrm{pooled}}$, but was only slightly worse than the best variant in terms of $C_{\mathrm{llr}}^{\mathrm{pooled}}$ (0.208 versus 0.207), and was substantially better on other metrics.

The overall pattern of results in the VSI was not surprising to those familiar with the development of automatic speaker recognition technology over the last two decades and particularly over the last few years. The observations made below regarding the performance of different types of system mirror results obtained in non-forensic evaluations, e.g., in NIST SRE. The overall pattern of results in the VSI was as follows:

1. Systems based on newer automatic speaker recognition technology outperformed systems based on older technology:
    a. GMM i-vector PLDA outperformed GMM-UBM
    b. GMM i-vector + DNN senone posterior i-vector PLDA outperformed GMM i-vector PLDA
    c. x-vector PLDA outperformed GMM i-vector PLDA
    d. x-vector PLDA outperformed GMM i-vector + DNN bottleneck PLDA
    e. x-vector PLDA outperformed GMM i-vector + DNN senone posterior i-vector PLDA (inferred from cross-paper comparison, but may have been due to other aspects of system design)
2. With respect to use of case-specific data:
    a. Systems optimized using case-specific data outperformed systems that were trained exclusively on non-case-specific data.
    b. Other than for GMM-UBM, systems initially trained on non-case-specific data and then optimized using case-specific data outperformed systems trained exclusively on case-specific data (the amount of non-case-specific data used for initial training was always much larger than the amount of case-specific data available).
    c. The greater the amount of case-specific data used for optimization, the better the performance.

The observations made above are general, and one may find exceptions.

Above, we used the term “optimization” to cover a range of disparate techniques for adapting models and for normalizing and calibrating scores using case-specific data; the papers published in the VSI do not allow for a systematic comparison of these techniques (and they may have confounded the observations above regarding different types of systems). The designs of the normalization-calibration techniques used in Batvox 3.1 and in the Phonexia systems actually caused their likelihood ratio outputs to be miscalibrated (see the discussions in the conclusions of Zhang & Tang, 2017, and Jessen, Bortlík, et al., 2019).

All the results are based on a single set of validation data reflecting the conditions of a single case. The results are not necessarily generalizable to other conditions in other cases – the relative performance of the different systems may depend on the particular conditions and amount of case-specific training/optimization data available. Even the newest systems tested as part of the VSI may soon be updated or replaced, hence the relative performance of systems produced by different developers may change.

2. Reflections

Our primary aims in proposing and guest editing the VSI were to encourage practitioners to empirically validate their forensic voice comparison systems under casework conditions, and to increase courts' awareness of the need for empirical validation under casework conditions. Ultimately, we hope that validation under conditions reflecting those of the case under investigation will become standard practice for all practitioners and will be demanded by the courts.

The VSI was designed so that practitioners could test their existing systems following procedures that reflected how they would use them in forensic casework. It was not designed for development of new systems. We are happy that the VSI received contributions from a number of practitioners working in operational forensic laboratories. We had hoped that we would have received more contributions from research laboratories (e.g., laboratories that participate in NIST SRE). Perhaps the focus on forensic application and validation of existing systems was not of interest to them. In order to facilitate participation by practitioners and researchers who have many other commitments, the VSI had a long submission window, ~2 years. Half of the submissions did not come in until the end of that submission window. It is not clear whether a shorter window would have resulted in the same number of submissions sooner or in fewer submissions. It was apparent, however, that many of those who did contribute had difficulty finding the time to work on the VSI.

The editorial process we adopted involved a division of labor whereby the first-named guest editor solicited submissions and assisted authors with experimental design and developing pre-submission drafts of their manuscript, and the second-named guest editor then handled post-submission manuscripts including recruiting reviewers and making decisions with respect to acceptance. The second-named guest editor also processed the results and provided the performance metrics and graphics to the authors.

Agreeing to submit a paper to the VSI was a condition of receiving access to the training and test data. Since the VSI is now closed, we will make the data available without this condition. The VSI dataset may therefore be used to test other systems. Practitioners and researchers who wish to request access to the data should complete the request form provided at http://databases.forensic-voice-comparison.net/#forensic_eval_01. The Matlab scripts that calculate the performance metrics and draw the performance graphics are also now provided via the login at that URL.

3. Errata

In Table 1 of Silva & Medina (2017) and Table 1 of Zhang & Tang (2018) the contents of the $C_{\text{llr}}^{\text{min}}$ and $C_{\text{llr}}^{\text{cal}}$ columns were inadvertently transposed. The larger values should have been in the $C_{\text{llr}}^{\text{min}}$ column, not the $C_{\text{llr}}^{\text{cal}}$ column. This was an editorial error for which we apologize. In the relevant rows of Table 1 of the present paper, the transposition has been corrected.

4. Acknowledgments

Many thanks to all of the authors and reviewers of the VSI papers. The work and time contributed by the authors is especially appreciated. We would like to give special thanks to David van der Vloed for acting as a guinea pig in authoring the first paper, and to Michael Jessen and Yosef A Solewicz for coauthoring two papers.

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

5. References

Jessen M., Bortlík J., Schwarz P., Solewicz Y.A. (2019). Evaluation of Phonexia automatic speaker recognition software under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01). *Speech Communication*, 111: 22–28. https://doi.org/10.1016/j.specom.2019.05.002

Jessen M., Meir G., Solewicz Y.A. (2019). Evaluation of Nuance Forensics 9.2 and 11.1 under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01). *Speech Communication*, 110: 101–107. https://doi.org/10.1016/j.specom.2019.04.006

Kelly F., Fröhlich A., Dellwo V., Forth O., Kent S., Alexander A. (2019). Evaluation of VOCALISE under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01). *Speech Communication*.

Morrison G.S., Enzinger E. (2016). Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic_eval_01) - Introduction. *Speech Communication*, 85: 119–126. http://dx.doi.org/10.1016/j.specom.2016.07.006

Silva G.D. da, Medina C.A. (2017). Evaluation of MSR Identity Toolbox under conditions reflecting those of a real forensic case (forensic_eval_01). *Speech Communication*, 94: 42–49. http://dx.doi.org/10.1016/j.specom.2017.09.001

van der Vloed D. (2016). Evaluation of Batvox 4.1 under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01). *Speech Communication*, 85: 127–130. http://dx.doi.org/10.1016/j.specom.2016.10.001

van der Vloed D. (2017). Erratum to “Evaluation of Batvox 4.1 under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01)” [Speech Communication 85 (2016) 127–130]. *Speech Communication*, 92: 23. http://dx.doi.org/10.1016/j.specom.2017.04.005

Zhang C., Tang C. (2018). Evaluation of Batvox 3.1 under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01). *Speech Communication*, 100: 13–17. https://doi.org/10.1016/j.specom.2018.04.008