ORGANIZATION OF SCIENTIFIC AREA COMMITTEES (OSAC)
FOR FORENSIC SCIENCE

SPEAKER RECOGNITION SUBCOMMITTEE

Essential scientific literature for human-supervised automatic approaches to forensic speaker recognition

Prepared by Scientific Literature Working Group,
Speaker Recognition Subcommittee

1st edition

April 2021

Preface

This document provides a structured bibliography of essential scientific literature for human-supervised automatic approaches to forensic speaker recognition. The bibliography complies with criteria for foundational scientific literature published by the US National Commission on Forensic Science. The scope of the document is restricted to forensic speaker recognition conducted for the purpose of presenting testimony in court, as opposed to conducted for purely investigative purposes.

Keywords

essential scientific literature; method validation; forensic speaker recognition; automatic speaker recognition; forensic speaker identification; forensic speaker comparison; forensic voice comparison; evaluation of evidence

Abbreviations

ASR	Automatic Speaker Recognition
DMSAC	Digital Multimedia Scientific Area Committee
DNN	deep neural network
GMM-UBM	Gaussian mixture model - universal background model
GSV-SVM	Gaussian supervectors - support vector machines
JFA	joint factor analysis
NCFS	National Commission on Forensic Science
NIST SRE	National Institute of Standards and Technology Speaker Recognition Evaluation
OSAC	Organization of Scientific Area Committees for Forensic Science
OSAC SR	Speaker Recognition Subcommittee of OSAC
p.	page
pp.	pages
PCAST	President's Council of Advisors on Science and Technology
PLDA	probabilistic linear discriminant analysis
UK	United Kingdom of Great Britain and Northern Ireland
US	United States of America

Table of Contents

1. Introduction

1.1. Scope

This document provides a bibliography of some of the essential scientific literature related to human-supervised automatic approaches to forensic speaker recognition. The scope is restricted to forensic speaker recognition conducted for the purpose of presenting testimony in court. Investigative applications are out of scope.

1.2. Automatic speaker recognition technology

Automatic speaker recognition technology is rapidly developing. Since around 2000, approaches have evolved from "Gaussian mixture model - universal background model" (GMM-UBM), to "Gaussian supervectors - support vector machines" (GSV-SVM), to "joint factor analysis" (JFA), to "i-vectors" used in conjunction with "probabilistic linear discriminant analysis" (PLDA). Current state of the art is based on "x-vectors" (also called "embeddings") derived from "deep neural networks" (DNNs).

The National Institute of Standards and Technology's Speaker Recognition Evaluations (NIST SRE), held every few years since 1996, have provided regular opportunities for blind testing of technological advances (Greenberg et al., 2020). Newer approaches have empirically demonstrated better performance.

Automatic speaker recognition has multiple applications, only one of which is forensic speaker recognition.

1.3. Forensic speaker recognition

Forensic speaker recognition is the process of comparing the properties of a recording of a speaker of questioned identity with the properties of one or more recordings of a speaker of known identity in preparation for testifying to a court of law deciding whether the recordings are of the same speaker or not. Sometimes there is no known-speaker recording and the task is to compare multiple questioned-speaker recordings.

Forensic speaker recognition may also be conducted to assist with law enforcement agency investigations. Functional requirements for investigative applications can differ substantially from those for evidential applications and are not addressed here. This document solely applies to evidential applications.

1.3.1. Human-supervised automatic approaches

There are multiple approaches to forensic speaker recognition, including "auditory", "spectrographic", "acoustic phonetic", and "human-supervised automatic", as well as combinations of these. This document focuses on human-supervised automatic approaches, which make quantitative measurements of acoustic properties of speech recordings and use those measurements as input to statistical models that provide probabilistic output related to same-

speaker versus different-speaker propositions. The measurements are made automatically, and the statistical models run automatically, but human expertise is required to make key decisions. These decisions include selecting appropriate recordings for training and testing forensic analysis systems.

1.3.2. Transparency and reproducibility

Human-supervised automatic approaches to forensic speaker recognition have high degrees of transparency and reproducibility. The analytical methods and procedures can be described in detail, and in principle the software and data used to conduct an analysis can be shared. Although some details of commercial systems may be trade secrets, descriptions of the general methods they employ are accessible in the peer-reviewed literature.

1.3.3. Evaluation of evidence

The likelihood-ratio framework has been used in conjunction with the human-supervised automatic approaches to forensic speaker recognition since around 2000 as a basis for evaluating the strength of evidence. While other methods of forensic inference exist, this document focuses on likelihood-ratio,[1] as a means of evaluating forensic evidence (e.g.: Aitken et al., 2011; Morrison et al., 2017). The use of this evaluative framework is recommended by many organizations, including:

- the American Statistical Association (Kafadar et al., 2019);[2]
- the Association of Forensic Science Providers of the UK and the Republic of Ireland (Association of Forensic Science Providers, 2009);
- and the European Network of Forensic Science Institutes
    - for forensic science in general (Willis et al., 2015),
    - and for forensic speaker recognition in particular (Drygajlo et al., 2015).

1.3.4. Reduction of cognitive bias

Cognitive bias is a recognized concern[3] in forensic science. The 2016 report by the President's Council of Advisors on Science and Technology (PCAST) repeatedly recommended the replacement of subjective methods with more objective methods, such as procedures that can be performed by automatic systems.[4] It stated that:

---

[1] For simplicity, this document uses the term "likelihood ratio" throughout, but this is not intended to exclude Bayesian concepts and methods.

[2] "To evaluate the weight of any set of observations made on questioned and control samples, it is necessary to relate the probability of making these observations if the samples came from the same source to the probability of making these observations if the questioned sample came from another source in a relevant population of potential sources." (Kafadar et al., 2019, p. 2)

[3] See Dror (2020) for an overview of cognitive bias in decision making

[4] For PCAST's definitions of "subjective" and "objective", see President's Council of Advisors (2016, p. 5).

> Objective methods are, in general, preferable to subjective methods. Analyses that depend on human judgment (rather than a quantitative measure of similarity) are obviously more susceptible to human error, bias, and performance variability across examiners. In contrast, objective, quantified methods tend to yield greater accuracy, repeatability and reliability, including reducing variation in results among examiners. Subjective methods can evolve into or be replaced by objective methods. (President's Council of Advisors, 2016, pp. 46–47)

Although not perfectly objective, the automatic nature of human-supervised automatic approaches provides a degree of resistance to cognitive biases that may arise (intentionally or unintentionally) either from the practitioner's knowledge of task-irrelevant information or from the practitioner's expectations or preferences regarding the outcome of the analysis. For human-supervised automatic approaches, a practitioner will exercise subjective judgment in selecting representative data for training and testing in a particular case (i.e. the "human-supervised" portion of the process). After the initial choices are made, however, the practitioner will be unable to predict or further control how the selection will affect specific results of the forensic analysis system, as the system objectively extracts quantitative measurements of the acoustic properties of voice recordings and inputs them directly to statistical models. The practitioner does not control this portion of the process, and hence cannot influence the results.

For a particular case, practitioners must make subjective judgments on issues such as what the relevant population is; whether the data are sufficiently reflective of that relevant population; and whether the data used for training and testing the forensic analysis system are sufficiently reflective of the speaking and recording conditions. So long as the practitioner makes the critical subjective judgments without knowing how they will affect the output of the forensic analysis system, those judgments are less likely to be affected by any expectations or preferences the practitioner might hold regarding the outcome of the analysis. It is to be expected that:

1. when the system is applied to the questioned- and known-speaker recordings in the case, the system's output will be a reasonable answer to the question posed by the propositions adopted for the case; and
2. the validation results will be reasonably informative as to the expected performance of the system when it is applied in the case.

1.3.5. Validation

The automatic nature of human-supervised automatic approaches facilitates validation in that it makes it easy to compare large numbers of test recordings. Nevertheless, the need for empirical validation under casework conditions has been emphasized by a number of organizations:

- National Research Council of the US National Academy of Sciences (National Research Council, 2009);
- US President's Council of Advisors on Science and Technology (President's Council of Advisors, 2016);

- European Network of Forensic Science Institutes (Drygajlo et al., 2015);

For a review of this topic, see Morrison (2014).

1.4. Criteria for foundational scientific literature

The U.S. National Commission on Forensic Science (2015b, pp. 1-2) stated that "each forensic discipline must have an underlying foundation that is the result of a rigorous vetting process and that is ultimately captured in the peer-reviewed scientific literature", and that "To strengthen confidence in results obtained in forensic examinations, each forensic discipline must identify resources that are scientifically credible, valid and with a clear scientific foundation". "The term 'foundation' was used ... to emphasize that each forensic discipline must have a scientifically robust and validated basis to its methods, its technologies, and its process of interpreting data".

The National Commission on Forensic Science (NCFS) proposed a number of criteria for "foundational scientific literature supportive of forensic practice". Those criteria included:

- That the literature be:
    - peer-reviewed original research;
    - peer-reviewed substantive reviews of the original research;
    - or reports of consensus-development conferences.
- That the literature be published in books authored by recognized experts, or published in journals which:
    - employ rigorous peer review by independent reviewers to assess consistency with the norms of scientific practice;
    - encourage ethical conduct in research and publication practices;
    - have recognized experts on their editorial boards;
    - are indexed in databases of scientific literature that are available through academic libraries.

In the field of automatic speaker recognition, peer-reviewed conference-proceedings papers can be highly influential. Such papers are therefore appropriately included in this bibliography of foundational scientific literature.
In addition to the NCFS criteria, decisions as to what to include in this document took into account the extent to which the literature was consistent with guiding principles that had already been adopted by the Organization of Scientific Area Committees - Speaker Recognition Subcommittee (OSAC SR). A statement of those guiding principles is provided in the Appendix (Sec. 3).

2. Bibliography of foundational scientific literature

The bibliography of foundational scientific literature is divided into two sections: "Automatic speaker recognition technology" (Sec. 2.1), and "Forensic speaker recognition" (Sec. 2.2). Each of these sections is divided into thematic subsections. Within each subsection, references are listed according to the year of publication. References published in the same year are listed in alphabetical order.

2.1. Automatic-speaker-recognition technology (ASR)

Listed below are references to publications that describe automatic-speaker-recognition technology that has been or currently is used in forensic applications. For the most part, however, these publications are not specific to forensic application.

2.1.1. Reviews

The following resources provide an overview of the field and place the remaining literature in context.

Campbell, J. P., Shen, W., Campbell, W. M., Schwartz, R., Bonastre, J.-F., & Matrouf, D. (2009). Forensic speaker recognition. *IEEE Signal Processing Magazine*, *26*(2), 95–103. https://doi.org/10.1109/msp.2008.931100

Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*, *52*(1), 12–40. https://doi.org/10.1016/j.specom.2009.08.009

Hansen, J. H. L., & Hasan, T. (2015). Speaker Recognition by Machines and Humans: A tutorial review. *IEEE Signal Processing Magazine*, *32*(6), 74–99. https://doi.org/10.1109/msp.2015.2462851

Matějka, P., Plchot, O. ř., Glembek, O. ř., Burget, L. š., Rohdin, J., Zeinali, H., Mošner, L., Silnova, A., Novotný, O. ř., Diez, M., & "Honza" Černocký, J. (2020). 13 years of speaker recognition research at BUT, with longitudinal analysis of NIST SRE. *Computer Speech & Language*, *63*, 101035. https://doi.org/10.1016/j.csl.2019.101035

Villalba, J., Chen, N., Snyder, D., Garcia-Romero, D., McCree, A., Sell, G., Borgstrom, J., García-Perera, L. P., Richardson, F., Dehak, R., Torres-Carrasquillo, P. A., & Dehak, N. (2020). State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and Speakers in the Wild evaluations. *Computer Speech & Language*, *60*, 101026. https://doi.org/10.1016/j.csl.2019.101026

2.1.2. ASR Factors

There are a number of factors that must be taken into account when considering specific ASR systems, including data selection, acoustic features, calibration, and validation.

2.1.2.1. Data selection

Hansen, J. H. L., & Bořil, H. (2018). On the issues of intra-speaker variability and realism in speech, speaker, and language recognition tasks. *Speech Communication*, *101*, 94–108. https://doi.org/10.1016/j.specom.2018.05.004

2.1.2.2. Acoustic features

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*(4), 357–366. https://doi.org/10.1109/tassp.1980.1163420

Reynolds, D. A. (1994). Experimental evaluation of features for robust speaker identification. *IEEE Transactions on Speech and Audio Processing*, *2*(4), 639–643. https://doi.org/10.1109/89.326623

Mammone, R. J., Xiaoyu Zhang, & Ramachandran, R. P. (1996). Robust speaker recognition: a feature-based approach. *IEEE Signal Processing Magazine*, *13*(5), 58. https://doi.org/10.1109/79.536825

Pelecanos, J., & Sridharan, S. (2001, June). Feature warping for robust speaker verification. *ODYSSEY-2001*, 213–218. https://www.isca-speech.org/archive_open/odyssey/odys_213.html

Jin, Q., & Zheng, T. F. (2011). Overview of front-end features for robust speaker recognition. *Proceedings of Asia-Pacific Signal and Information Processing Association (APSIPA) Annual Summit and Conference*. http://www.apsipa.org/proceedings_2011/pdf/APSIPA335.pdf

Sadjadi, S. O., & Hansen, J. H. L. (2015). Mean Hilbert envelope coefficients (MHEC) for robust speaker and language identification. *Speech Communication*, *72*, 138–148. https://doi.org/10.1016/j.specom.2015.04.005

2.1.2.3. Normalization, calibration, and fusion

Auckenthaler, R., Carey, M., & Lloyd-Thomas, H. (2000). Score Normalization for Text-Independent Speaker Verification Systems. *Digital Signal Processing*, *10*(1–3), 42–54. https://doi.org/10.1006/dspr.1999.0360

Pigeon, S., Druyts, P., & Verlinde, P. (2000). Applying Logistic Regression to the Fusion of the NIST'99 1-Speaker Submissions. *Digital Signal Processing*, *10*(1–3), 237–248. https://doi.org/10.1006/dspr.1999.0358

Brümmer, N., Burget, L., Černocký, J., Glembek, O., Grézl, F., Karafiát, M., van Leeuwen, D. A., Matějka, P., Schwarz, P., & Strasheim, A. (2007). Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(7), 2072–2084. https://doi.org/10.1109/tasl.2007.902870

Cumani, A., Batzu, P. D., Colibro, D., Vair, C., Laface, P., & Vasilakakis, V. (2011). Comparison of speaker recognition approaches for real applications. *Proceedings of Interspeech*, 2365–2368. https://www.isca-speech.org/archive/interspeech_2011/i11_2365.html

Ferrer, L., Nandwana, M. K., McLaren, M., Castan, D., & Lawson, A. (2019). Toward Fail-Safe Speaker Recognition: Trial-Based Calibration With a Reject Option. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(1), 140–153. https://doi.org/10.1109/taslp.2018.2875794

2.1.2.4. Validation procedures, metrics, and graphics

Brümmer, N., & du Preez, J. (2006). Application-independent evaluation of speaker detection. *Computer Speech & Language*, *20*(2–3), 230–275. https://doi.org/10.1016/j.csl.2005.08.001

2.1.3. ASR algorithms

ASR algorithms have evolved over the years. The papers listed below provide a better understanding of each algorithm. Practitioners should be aware of the technologies encompassed in their ASR systems, and the relative strengths and weaknesses therein.

2.1.3.1. Gaussian mixture model - universal background model (GMM-UBM)

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, *10*(1–3), 19–41. https://doi.org/10.1006/dspr.1999.0361

2.1.3.2. Gaussian supervectors - support vector machine (GSV-SVM)

Campbell, W. M., Sturim, D. E., & Reynolds, D. A. (2006). Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Processing Letters*, *13*(5), 308–311. https://doi.org/10.1109/lsp.2006.870086

2.1.3.3. Joint factor analysis (JFA)

Kenny, P., Boulianne, G., Ouellet, P., & Dumouchel, P. (2007). Joint Factor Analysis Versus Eigenchannels in Speaker Recognition. *IEEE Transactions*

on Audio, Speech and Language Processing, *15*(4), 1435–1447. https://doi.org/10.1109/tasl.2006.881693

2.1.3.4. i-vectors

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-End Factor Analysis for Speaker Verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(4), 788–798. https://doi.org/10.1109/tasl.2010.2064307

Lei, Y., Scheffer, N., Ferrer, L., & McLaren, M. (2014). A novel scheme for speaker recognition using a phonetically-aware deep neural network. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1695–1699. https://doi.org/10.1109/icassp.2014.6853887

McLaren, M., Lei, Y., & Ferrer, L. (2015). Advances in deep neural network approaches to speaker recognition. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4814–4818. https://doi.org/10.1109/icassp.2015.7178885

2.1.3.5. Probabilistic linear discriminant analysis (PLDA)

Prince, S. J., & Elder, J. H. (2007). Probabilistic Linear Discriminant Analysis for Inferences About Identity. *Proceedings of the IEEE 11th International Conference on Computer Vision*. https://doi.org/10.1109/iccv.2007.4409052

Kenny, P. (2010). Bayesian speaker verification with heavy tailed priors. *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*. https://www.isca-speech.org/archive_open/odyssey_2010/od10_014.html

García-Romero, D., & Espy-Wilson, C. Y. (2011). Analysis of i-vector length normalization in speaker recognition systems. *Proceedings of Interspeech*, 249–252. https://www.isca-speech.org/archive/interspeech_2011/i11_0249.html

Sizov, A., Lee, K. A., & Kinnunen, T. (2014). Unifying Probabilistic Linear Discriminant Analysis Variants in Biometric Authentication. *Lecture Notes in Computer Science*, 464–475. https://doi.org/10.1007/978-3-662-44415-3_47

2.1.3.6. x-vectors

Snyder, D., García-Romero, D., Povey, D., & Khudanpur, S. (2017). Deep Neural Network Embeddings for Text-Independent Speaker Verification. *Proceedings of Interspeech*, 999–1003. https://doi.org/10.21437/interspeech.2017-620

Snyder, D., García-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-Vectors: Robust DNN Embeddings for Speaker Recognition. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5329–5333. https://doi.org/10.1109/icassp.2018.8461375

2.2. Forensic speaker recognition

Listed below are references to publications dealing specifically with forensic speaker recognition. Sec. 2.2.1 lists reviews that provide an overview of the field and place the remaining literature in context. Sec. 2.2.2 lists seminal publications that introduced the use of the likelihood-ratio framework to the field. Secs. 2.2.3, 2.2.4, and 2.2.5 list publications dealing with the key topics of data selection, calibration, and validation in the context of the likelihood-ratio framework. Sec. 2.2.6 lists publications that include reports of empirical validation conducted under conditions reflecting forensic casework conditions. They include several papers from a journal special issue in which different forensic-speaker-recognition systems were all tested on the same data. Conditions vary from case to case, hence additional validation will often be necessary prior to using a particular forensic-speaker-recognition system in a particular case.

2.2.1. Reviews

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. https://enfsi.eu/wp-content/uploads/2016/09/guidelines_fasr_and_fsasr_0.pdf

Morrison G.S., Enzinger E., Zhang C. (2018). Forensic speech science. In Freckelton I., Selby H., eds., *Expert Evidence*, chapter 99. Thomson Reuters, Sydney, Australia. http://expert-evidence.forensic-voice-comparison.net/

Morrison G.S., Enzinger E., Ramos D., González-Rodríguez J., Lozano-Díez A. (2020). Statistical models in forensic voice comparison. In Banks D.L., Kafadar K., Kaye D.H., Tackett M. (Eds.) *Handbook of Forensic Statistics*, chapter 20. CRC, Boca Raton, FL. http://handbook-of-forensic-statistics.forensic-voice-comparison.net/

2.2.2. Likelihood-ratio framework

A likelihood ratio expresses the probability of the observations if one proposition were true versus the probability of the observations if an alternative proposition were true. The propositions must be mutually exclusive. In human-supervised automatic approaches, the observations are quantitative measurements of the acoustic properties of recordings of speakers' voices, and the propositions are some version of the following:

The voices of interest on each of two or more audio recordings were produced by the same speaker.

versus

The voices of interest on each of two or more audio recordings were produced by different speakers, each from the same population.

The following papers provide additional information on the application of the likelihood ratio framework to forensic speaker recognition:

Champod, C., & Meuwly, D. (2000). The inference of identity in forensic speaker recognition. *Speech Communication*, *31*(2–3), 193-203. https://doi.org/10.1016/s0167-6393(99)00078-3

Rose, P. (2002). Forensic Speaker Identification. London: CRC Press, https://doi.org/10.1201/9780203166369

Gold, E., & Hughes, V. (2014). Issues and opportunities: the application of the numerical likelihood ratio framework to forensic speaker comparison. *Science & Justice : Journal of the Forensic Science Society*, *54*(4), 292-299. https://doi.org/10.1016/j.scijus.2014.04.003

2.2.3. Data selection

Morrison, G. S., Enzinger, E., & Zhang, C. (2016). Refining the relevant population in forensic voice comparison – A response to Hicks et alii (2015) The importance of distinguishing information from evidence/observations when formulating propositions. *Science & Justice*, *56*(6), 492–497. https://doi.org/10.1016/j.scijus.2016.07.002

Hughes, V., & Rhodes, R. (2018). Questions, propositions and assessing different levels of evidence: Forensic voice comparison in practice. *Science & Justice*, *58*(4), 250–257. https://doi.org/10.1016/j.scijus.2018.03.007

Hughes, V., & Rhodes, R. (2018a). Corrigendum to 'Questions, propositions and assessing different levels of evidence: Forensic voice comparison in practice.' *Science & Justice*, *58*(5), 384. https://doi.org/10.1016/j.scijus.2018.06.006

2.2.4. Calibration and fusion

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 Transactions on Audio, Speech and Language Processing*, *15*(7), 2104–2115. https://doi.org/10.1109/tasl.2007.902747

Morrison, G. 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. https://doi.org/10.1080/00450618.2012.733025

Morrison, G. S., & Poh, N. (2018). Avoiding overstating the strength of forensic evidence: Shrunk likelihood ratios/Bayes factors. *Science & Justice*, *58*(3), 200–218. https://doi.org/10.1016/j.scijus.2017.12.005

2.2.5. Validation procedures, metrics, and graphics

Morrison, G. S. (2011). Measuring the validity and reliability of forensic likelihood-ratio systems. *Science & Justice*, *51*(3), 91–98. https://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. *Science & Justice*, *54*(3), 245–256. https://doi.org/10.1016/j.scijus.2013.07.004

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. https://doi.org/10.1016/j.specom.2016.07.006

Meuwly, D., Ramos, D., & Haraksim, R. (2017). A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. *Forensic Science International*, *276*, 142–153. https://doi.org/10.1016/j.forsciint.2016.03.048

Wang, B., Hughes, V., & Foulkes, P. (2019). Effect of score sampling on system stability in likelihood ratio based forensic voice comparison. *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS)*, 3065–3069. https://vincehughes.files.wordpress.com/2019/04/effect-of-score-sampling-on-system-stability-in-likelihood-ratio-based-forensic-voice-comparison_full-paper.pdf

2.2.6. Validation studies

Solewicz Y.A., Becker T., Jardine G., Gfroerer S. (2012). Comparison of speaker recognition systems on a real forensic benchmark. In *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, pp. 85–91. https://isca-speech.org/archive/odyssey_2012/od12_086.html

van der Vloed D., Bouten J., van Leeuwen D. (2014). NFI-FRITS: A forensic speaker recognition database and some first experiments. In *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, pp. 6–13. http://cs.uef.fi/odyssey2014/program/pdfs/21.pdf

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. *Science & Justice*, *56*(1), 42–57. https://doi.org/10.1016/j.scijus.2015.06.005

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. https://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. https://doi.org/10.1016/j.specom.2017.04.005

Enzinger, E., & Morrison, G. S. (2017). Empirical test of the performance of an acoustic-phonetic approach to forensic voice comparison under conditions similar to those of a real case. *Forensic Science International*, *277*, 30–40. https://doi.org/10.1016/j.forsciint.2017.05.007

da Silva, D. G., & 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. https://doi.org/10.1016/j.specom.2017.09.001

Morrison, G. S. (2018). The impact in forensic voice comparison of lack of calibration and of mismatched conditions between the known-speaker recording and the relevant-population sample recordings. *Forensic Science International*, *283*, e1–e7. https://doi.org/10.1016/j.forsciint.2017.12.024

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

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*, *112*, 30–36. https://doi.org/10.1016/j.specom.2019.06.005

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

van der Vloed D., Kelly F., Alexander A. (2020). Exploring the effects of device variability on forensic speaker comparison using VOCALISE and NFI-FRIDA, a forensically realistic database. In *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, pp. 402–407.

3. Appendix: Statement of guiding principles

OSAC SR has adopted the following as guiding principles in the development of standards and other documents related to forensic speaker recognition conducted for the purpose of presenting testimony in court.

3.1. Transparency and reproducibility

The forensic practitioner must clearly describe the materials analyzed, and the observations made on those materials. The forensic practitioner must clearly describe each of the propositions that they set out to evaluate.
The procedures, methods, and data used in the forensic analysis must be described in sufficient detail that another suitably qualified forensic practitioner could reproduce the process.

3.2. Framework for evaluation of evidence

The forensic practitioner should assess the relative probabilities of the observations given two competing propositions. These propositions must be mutually exclusive.

3.3. Reduction of cognitive bias

Appropriate procedures should be adopted to reduce the potential for cognitive bias, including procedures to prevent the forensic practitioner from being unnecessarily exposed to task-irrelevant information.

3.4. Validation

The system used to conduct the forensic analysis must be empirically tested using sufficient ground-truth data reasonably representative of the conditions of the case. To the extent possible, validation data must replicate real life case conditions. The system as a whole must be tested, not just its component parts. The system includes all methods and procedures implemented by the forensic practitioner.

Details of validation protocols, validation data, and validation results must be made available to all parties in the case.

4. Publications cited in the Introduction (Sec. 1)

Aitken C.G.G., Berger C.E.H., Buckleton J.S., Champod C., Curran J.M., Dawid A.P., Evett I.W., Gill P., González-Rodríguez J., Jackson G., Kloosterman A., Lovelock T., Lucy D., Margot P., McKenna L., Meuwly D., Neumann C., Nic Daéid N., Nordgaard A., Puch-Solis R., Rasmusson B., Redmayne M., Roberts P., Robertson B., Roux C., Sjerps M.J., Taroni F., Tjin-A-Tsoi T., Vignaux G.A., Willis S.M., Zadora G. (2011). Expressing evaluative opinions: A position statement. *Science & Justice*, 51: 1–2. http://dx.doi.org/10.1016/j.scijus.2011.01.002

Association of Forensic Science Providers (2009). Standards for the formulation of evaluative forensic science expert opinion. *Science & Justice*, 49: 161–164. http://dx.doi.org/10.1016/j.scijus.2009.07.004

Dror, Itiel E. (2020). Cognitive and human factors in expert decision making: six fallacies and the eight sources of bias. *Analytical Chemistry 2020 92* (12), 7998-8004 https://doi.org/10.1021/acs.analchem.0c00704

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. http://enfsi.eu/wp-content/uploads/2016/09/guidelines_fasr_and_fsasr_0.pdf

Greenberg C.S., Mason, L.P., Sadjadi, S.O., Reynolds D.A. (2020). Two decades of speaker recognition evaluation at the national institute of standards and technology. *Computer Speech & Language*, 60: article 101032. https://doi.org/10.1016/j.csl.2019.101032

Kafadar K., Stern H., Cuellar M., Curran J., Lancaster M., Neumann C., Saunders C., Weir B., Zabell S. (2019). *American Statistical Association position on statistical statements for forensic evidence*. American Statistical Association. https://www.amstat.org/asa/files/pdfs/POL-ForensicScience.pdf

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

Morrison G.S., Thompson W.C. (2017). Assessing the admissibility of a new generation of forensic voice comparison testimony. *Columbia Science and Technology Law Review*, 18: 326–434.

National Commission on Forensic Science (2015b). *Scientific literature in support of forensic science and practice*. https://www.justice.gov/archives/ncfs/file/786591/download

National Research Council (2009). *Strengthening forensic science in the United States: A path forward*. National Academies Press, Washington, DC. https://doi.org/10.17226/12589

President's Council of Advisors on Science and Technology. (2016). *Forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods*. https://obamawhitehouse.archives.gov/administration/eop/ostp/pcast/docsreports/

Willis S.M., McKenna L., McDermott S., O'Donell G., Barrett A., Rasmusson A., Nordgaard A., Berger C.E.H., Sjerps M.J., Lucena-Molina J.J., Zadora G., Aitken C.G.G., Lunt L., Champod C., Biedermann A., Hicks T.N., Taroni F. (2015). *ENFSI guideline for evaluative reporting in forensic science*. European Network of Forensic Science Institutes. http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf