The Speaker and Language Recognition Workshop (Odyssey 2024)
18-21 June 2024, Quebec City, Canada

Exploring speaker similarity based selection of relevant populations for forensic automatic speaker recognition

Linda Gerlach, Finnian Kelly, Kirsty McDougall, Anil Alexander

Oxford Wave Research
Oxford, UK
Phonetics Laboratory
University of Cambridge, UK
lg589@cam.ac.uk

Abstract

This study investigates various strategies for selecting relevant populations and their impact on forensic automatic speaker recognition in a likelihood ratio framework. Besides random and demographic metadata-based selection, it explores perceptual voice similarity as a potential criterion. Using a database controlled for gender, language, and recording conditions, an automatic approach based on phonetic features was used to select the most and least perceptually similar speakers to questioned speaker recordings in mock cases. It compares the strength of evidence obtained using these strategies for male and female mock cases and across different relevant population sizes. Random and metadata-based selections converge in log-likelihood ratio cost (Cllr) as the selected population size nears 50. While using perceptually similar speakers improves the overall Cllr, in individual cases effects may vary based on the degree of perceptual similarity or dissimilarity between recordings.

1. Introduction

Forensic Automatic Speaker Recognition (FASR) is the process of comparing speech samples using automatic speaker recognition technology in order to assist legal decision-making about whether the speakers on the samples are the same or different. Typically, FASR involves the comparison of a speech sample from a known speaker with that of an unknown (or questioned) speaker. As the underlying speaker recognition technology has matured, FASR is used by an increasing number of forensic practitioners, e.g. [1, 2, 3, 4].

A likelihood ratio (LR) is a means of assigning evidential weight to the output of an FASR system. Given a comparison between two speech samples, the LR expresses the system output under two competing hypotheses, for example, 1) the speech in the two samples originates from the same speaker (the same-speaker or H0 hypothesis), versus 2) the speech in the two samples originates from different speakers (the different-speaker or H1 hypothesis). There is widespread endorsement of the LR framework in FASR, and in forensic science more generally [5, 6, 7].

The LR framework takes a Bayesian approach, whereby the prior odds of the H0 or H1 hypotheses being true are multiplied by the LR given the evidence (E) to yield the posterior odds. The prior odds are assigned by the court, taking into account background knowledge of the case. The LR itself is calculated by dividing the probability of E given hypothesis H0 by the probability of E given a competing hypothesis H1. An LR larger than 1 provides stronger evidence in favour of H0 and increases the posterior odds, while an LR smaller than 1 offers stronger support for H1 and decreases the posterior odds [6, 8, 7].

While there are different ways in which the LR can be practically calculated in FASR [6], they all depend on the notion of a relevant population (also referred to as a reference population/suspect population/background data), which is a group of speakers representative of the speech samples under comparison. A typical approach is to represent the H0 hypothesis by the distribution of same-speaker comparison scores within the relevant population, and the H1 hypothesis by the distribution of different-speaker comparison scores within the relevant population [6]. The relevant population therefore allows the typicality of the compared speech features to be assessed. The definition of the relevant population (and the selection of data representative of that population) is central to the resultant LR, and consequently has been a topic of much discussion and debate.

1.1. Relevant population selection

It is widely accepted that the relevant population should generally be defined based on H1 while taking into account background knowledge of the case, and should remain the same for all other evidence to be assessed [6]. A challenge arises when the defence proposition is not sufficiently specific or not provided at all; a situation that is common [9, 10, 8, 11]. In court, the value of evidence for which a large, unspecific relevant population is selected may be very limited due to the resulting small prior odds as well as potentially small LRs [12, 13]. As a result, a relevant population must be defined in a pragmatic manner by the forensic analyst.

On the one hand, without a clear defence proposition, the assumed H1 often is that the questioned speaker's voice is not that of the known speaker but of another individual of the same gender speaking that language [11]. Relying on logical relevance means that the relevant population is restricted in terms of speaker characteristics to those individuals that could theoretically have produced the speech sample [14, 15]. Such characteristics are either known based on other evidence, or must be estimated, and may include gender and language spoken, but also age group, social class, or accent [16, 17]. In addition, the conditions within the known and questioned samples (i.e. speaker-related factors such as speaking style, and technical factors such as the recording device) must be accounted for when selecting data for the relevant population.

On the other hand, there is an argument that if two audio samples are sent for forensic analysis, it is likely because they

10.21437/odyssey.2024-4

sound similar enough to a naïve listener, e.g. a police officer. Morrison et al. [13] propose that the default H1 should assume that the questioned speaker's voice is not that of the known speaker but someone sounding sufficiently similar to prompt such comparison. The authors suggested that a panel of lay listeners could be consulted to judge the similarity between the questioned speaker and the nominated relevant population speakers, taking into account speaking styles and channel conditions of the recordings for the selection of a relevant population. Morrison et al. further tested an MFCC-based GMM-UBM automatic speaker recognition system in a small-scale study replacing the suggested lay listener panel to select perceptually similar speakers for 25 male mock questioned speakers, controlling also for language and recording condition. Results showed that the similarity-based relevant population yielded a lower Cllr than the vast majority of the random selections.

1.2. Motivation

Recent research [18, 19] suggests that automatically extracted phonetic features perform better than MFCCs in approximating lay listener ratings of perceived voice similarity. The present study therefore aims to build on Morrison et al.'s [13] study by using an ASR system based on phonetic features to select a similar-sounding relevant population.

This study considers three options that a forensic analyst could use to select a relevant population: a random selection of speakers from a database that roughly corresponds to the H1 hypothesis, a selection based on perceptually similar sounding speakers (either using the most similar or least similar speakers), or simply using any known metadata (e.g., region of upbringing, age, etc).

Although intuitively using a random selection of speakers may seem attractive, the number and diversity of speakers in the source database from which the selection is drawn, as well as the size of the selection, can create significant variations in results. If choosing a different random selection can significantly impact the LR, then this is a cause for concern. This study will consider the impact of selecting different random population subsets on the LRs obtained. This will include same-speaker and different-speaker cases.

Another option is to consider using the most (or alternatively least) similar-sounding speakers to the questioned speaker to form a relevant population. Choosing the most similar-sounding speakers (using an automatic approach) would create a relevant population of speakers who are also arguably similar sounding to each other. This could make different speakers in this set score more highly in comparisons against each other, and potentially reduce the LRs for same-speaker comparisons.

The third option of choosing the relevant population on the basis of metadata descriptors accompanying the database such as gender, language, age, and other such demographic factors can be considered a pragmatic approach. As long as these demographic factors are taken into consideration in defining the H1 hypothesis (e.g. female speakers of German from Bavaria between the age of 40 and 60), this is acceptable in the forensic likelihood ratio framework. However, this approach would require a huge database of speakers to make selections from so that once all these demographic selections are considered, there is still a sufficient number of speakers in the relevant population (e.g. above 50).

This study investigates whether these different selection methods can improve the discrimination and calibration performance of the system and how these choices affect the LR. Specifically, it compares selecting the relevant population based on perceived speaker similarity with random or metadata-based selections. Furthermore, the study examines how factors such as relevant population size and speaker gender interact with these different selection methods.

2. Experiment

An experiment was conducted to explore the impact of different relevant population selection strategies on the strength of evidence in a set of 'mock cases', constructed from a forensically-relevant dataset.

2.1. Automatic speaker recognition system

The ASR system used in this study is VOCALISE forensic automatic speaker recognition software [20]. Following Gerlach et al. [19], perceptual similarity was assessed using x-vectors based on automatically-extracted phonetic (auto-phonetic) features, i.e. long-term formant distributions F1 to F4. Forensic speaker comparisons were conducted using an x-vector model based on MFCCs with 22 dimensions (including energy) extracted via 23 Mel filters between 20 and 3,700 Hz, followed by LDA and PLDA, each of 150 dimensions. Prior to all speaker modelling, voice activity detection (VAD) was applied.

2.2. Speaker data and mock case creation

The GBR-ENG speaker database [21] was used as a source of forensically-relevant data in this study. GBR-ENG contains 6000 spontaneous landline and mobile telephone conversations in English from 600 male and female speakers recorded across England. Recording durations varied between 3 to 6 minutes with a mean duration of 1 min 53 s after VAD. Demographic metadata included speaker ID, speaker gender, age, and region of upbringing, among others. Speakers were aged 18 to 60 years. Only landline recordings from three regions of upbringing (North, Midlands, and South of England; GB-NOR, GB-MID, GB-SOU) were selected and separated according to speaker gender. The recordings were then split into a test set, i.e. mock case files, and a relevant population superset containing recordings from all three regions, from which potential relevant populations were selected.

To create a set of mock cases, 5 male speakers and 5 female speakers from each of the three regions of upbringing were randomly selected resulting in 15 mock offenders per gender. The set of mock case speakers was not assessed for their level of similarity. Two recordings per speaker were used as the mock case files (i.e. to serve as known and questioned samples), leading to the comparison of 15 x 15 recordings, thereof 15 same-speaker (SS) comparisons and 210 different-speaker (DS) comparisons per gender. All remaining speakers of the same gender with at least two files formed part of the relevant population superset ($N_{male} = 155$, $N_{female} = 177$ speakers).

2.3. Relevant population selection approaches

Relevant populations were selected either randomly, based on ranked automatically estimated perceived voice similarity, or on metadata provided with the database.

2.3.1. Random selection

Ten random relevant populations each were sampled in increments of ten (10, 20, 30, ..., 100), resulting in 100 random rel-

evant populations. The speakers within each random relevant population were of the same gender, but varied across region of upbringing. See Section 3.2 for further details.

2.3.2. Selection based on ranked automatically obtained estimates of perceived voice similarity

The questioned speaker samples from each mock case were compared to the same-gender relevant population superset using the VOCALISE x-vector auto-phonetic model. The resulting scores were subjected to cross-validation calibration in BioMetrics [22] (which applies linear logistic regression to convert scores to logLRs) to normalise their numerical range and allow comparability across datasets. The scores were then ranked for each questioned speaker. All mock known speaker samples were ignored for the selection of relevant populations based on ranked similarity. Male mock cases overall yielded slightly higher similarity logLRs ($M = -5.521$, $SD = 2.561$) compared to females ($M = -6.477$, $SD = 3.712$) and were less variable. For each mock questioned speaker, top (most similar) and bottom (least similar) relevant populations were selected at increments of ten from a relevant population size of ten to 100. Note that due to the sizes of the relevant population supersets for male and female speakers, increasing overlap between top and bottom selections if $N > 77$ (male) and $N > 88$ (female) is expected.

2.3.3. Metadata-based selection

For each gender, relevant populations solely based on metadata were selected to form a baseline for the random and ranked methods of relevant population selection. The metadata-based selection considered speaker gender and region of upbringing ("meta-region") in addition to recording condition, i.e. landline. Within the male and female relevant population supersets, there are proportionally more GB-SOU speakers (N_{male} = 98, N_{female} = 92 speakers) than GB-NOR or GB-MID speakers (27 and 52 respectively). Therefore, the metadata-based relevant population size was limited to approximately $N = 30$, at the low end of recommended relevant population sizes [23, 24, 25, 1]. For each speaker region containing more than 30 speakers in the superset, ten different relevant populations were sampled.

2.4. Data processing and analysis

Within all selections that were made for the different relevant populations, all speakers were compared to each other in multiple pairwise comparisons using the VOCALISE x-vector MFCC model to obtain scores. Similarly, each mock questioned speaker was compared to all same-gender mock known speakers using the x-vector MFCC model. The resulting mock case scores were then calibrated using each of the relevant population selections to obtain logLRs via logistic regression, and Cllrs were calculated [26, 27].

3. Results

3.1. Individual variability

The distribution of SS (H0) and DS (H1) comparison scores for each of the different relevant population selections (with a population size of 30 speakers) for an example female mock case are shown in Figure 1. A similar trend is observed for the combined set of female mock cases.

In this example case, the H0 distributions are similar across all relevant population selections, while there are some noticeable shifts in the H1 distributions. For the top (most similar) selection, there is a H1 shift to the right, relative to the random and meta-region distributions. For the bottom (least similar) selection, there is a H1 shift to the left, relative to the random and meta-region distributions. This behaviour can be expected, given that higher different-speaker scores are more likely to be observed if those different speakers sound similar to one another, and, similarly, lower different-speaker scores are more likely to be observed if those different speakers sound dissimilar to one another.

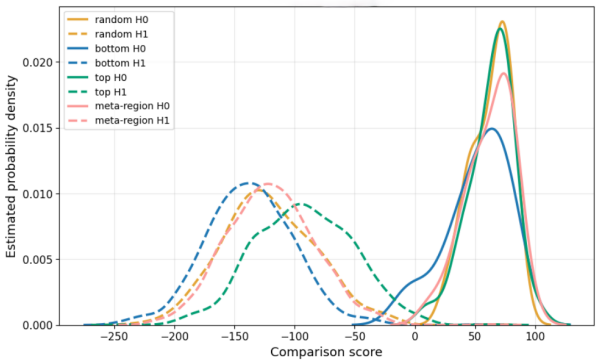

Figure 1: Kernel density estimation for H0 and H1 scores in each relevant population selection at a relevant population size of 30 for a female example speaker (747).

Therefore, if a mock case comparison involving two similar-sounding different speakers is considered, a lower logLR would generally be expected given the top selection than with the bottom selection. Similarly, if a mock case comparison involving two dissimilar-sounding different speakers is considered, a lower logLR would generally be expected given the bottom selection than with the top selection. For the different-speaker comparisons considered here then, selecting a relevant population based on the relative similarity of the samples in the case may increase the strength of evidence. For same-speaker comparisons however, the situation is not as clear, as there is relatively little difference between the H0 distributions for the different selections.

Referring to the average logLRs across all female mock cases in Table 1, it is evident that if two samples from different speakers are compared, this results, on average, in a lower DS logLR when using the top relevant population than when using the bottom relevant population. However, when two samples from the same speaker are compared, this results, on average, in a higher SS logLR when using the bottom relevant population than when using the top relevant population. Overall, these results suggest increased separability between SS and DS logLRs when a similar-sounding relevant population is chosen. It is worth noting that the analysis does not address the relative similarity between the samples in each case, which could provide further insights into the outcomes.

To consider the performance of the relevant populations across different sizes, Cllrs are examined next.

3.2. Global trends

The Cllrs calculated from the logLRs for each of the mock cases are shown in Figures 2 (female) and 3 (male). Across all relevant population sizes, very low Cllrs can be observed, indicating

Table 1: Average same-speaker (SS) and different-speaker (DS) LogLRs and difference between SS and DS logLRs across all 15 female mock cases based on the different relevant populations with 30 speakers each.

	random	**meta-region**	**top**	**bottom**
SS LogLR	8.698	9.950	9.057	12.640
DS LogLR	-10.225	-13.189	-19.275	-12.858
Difference	18.922	23.140	28.331	25.498

good discrimination and calibration performance for all relevant populations.

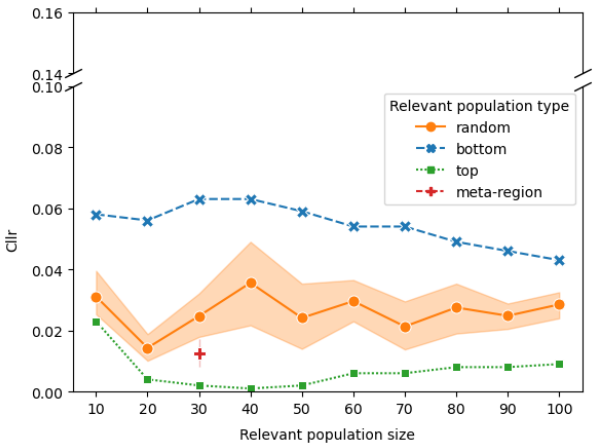

Figure 2: Cllrs for random, top, bottom, and metadata-based relevant population selections at size increments for female mock questioned speakers. Average Cllr and 95% CI for random and metadata-based relevant populations. Note the y-axis is not continuous.

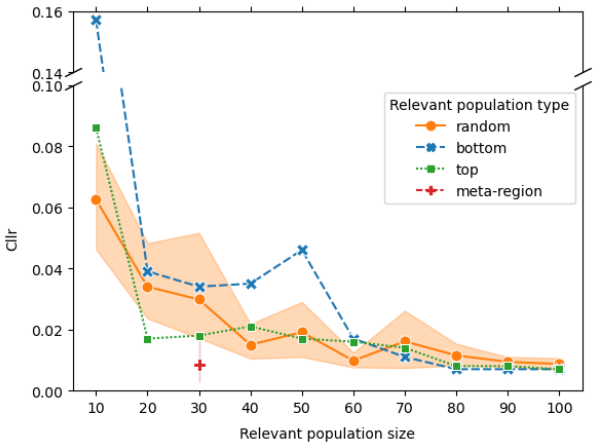

Figure 3: Cllrs for random, top, bottom, and metadata-based relevant population selections at size increments for male mock questioned speakers. Average Cllr and 95% CI for random and metadata-based relevant populations. Note the y-axis is not continuous.

Across male and female mock cases, the bottom relevant population selection (Figure 2 and 3 , blue cross, dashed line) has a higher Cllr than the top selection (green square, dotted line) at nearly all relevant population sizes, indicating worse discrimination and calibration. The random relevant population (orange circle, solid line) has Cllrs largely in the range between or overlapping with those of bottom and top selections. The metadata-based selection at $N = 30$ (red plus, no line) has a low Cllr competing with that of the top relevant population selection.

Overall, the patterns of Cllrs with increasing relevant population size vary across male and female mock cases. While doubling the relevant population size from 10 to 20 leads to a striking reduction of the Cllr in male cases, especially for the bottom selection, in female cases this is followed by a comparatively small decrease of the Cllr. In general, increasing the population size for female cases seems to have little impact on the discrimination and calibration of the system. Further, the difference in Cllrs of the different relevant populations is clearer in female cases ($M_{top} = 0.007$, $M_{random} = 0.026$, $M_{bottom} = 0.055$) and the top selection outperforms both random and bottom at all increments. While the meta-region relevant population outperforms other selections for the male mock cases, note that its logLRs are predominantly based on the mock cases from one region of upbringing (GB-SOU) and not evenly across all three regions.

The interpretation of results for relevant populations above a size of 50 becomes more complex. Due to the increasing relevant population size, there is an increasing overlap of speakers in the different relevant population types. While a larger relevant population size overall leads to a lower and more stable Cllr, the differences between the top, bottom, and random selections wane.

The results indicate that the top relevant population generally performs well, and converges quickly to low Cllr. The bottom relevant population is consistently worse than a top selection for female speakers, but comparatively better for males. The random selection is subject to variation, but is generally better than the bottom selection. The meta-region relevant population is effective, however, note that this requires metadata and a sufficient number of speakers.

To provide further insight into performance at the mock case level, Figure 4 shows the average absolute differences between SS and DS logLRs for each relevant population type containing 30 speakers across all female questioned speakers. It is evident that there exists a high degree of variability across the individual cases. With respect to the performance of the top versus bottom relevant population, for over half of the mock questioned speakers the difference between SS and DS logLRs is higher when using the top selection. Regarding the metadata-based selections, it can be seen that these did not necessarily yield high differences between SS and DS logLRs, which may be attributed to the rough labelling of region of upbringing. While for some speakers there were clear differences of the extent of the discrimination of the different relevant populations (e.g. 254, 747), for other speakers the differences between the different selections were small (e.g. 989).

4. Discussion

The analysis of the logLRs across the individual mock cases showed overall good discrimination throughout all relevant population types at a relevant population size of 30. It was found that the bottom relevant population with the least similar-sounding speakers compared to the mock questioned speaker on

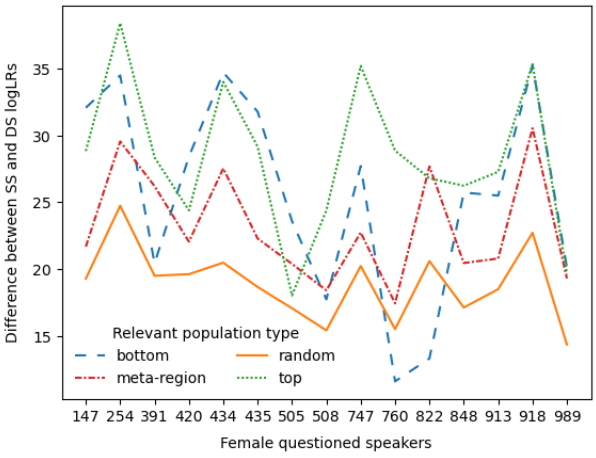

Figure 4: Absolute difference between SS and DS logLRs for female questioned speakers and relevant population selection at a relevant population size of 30.

average yielded higher logLRs for DS comparisons and higher logLRs for SS comparisons than the top relevant population with the perceptually most similar speakers to the mock questioned speaker. Further, the magnitude of the difference between SS and DS logLRs was found to be the highest for the top relevant population when compared to all relevant population selections. LogLRs calculated based on the random selection performed well and the metadata-based selection yielded logLRs with a similar magnitude to those of the random selections. Across the 15 mock cases per speaker gender, some variability regarding the effectiveness of the different relevant populations could be observed, likely also due to the composition of the database which contained more speakers from the South of England than from the other two regions.

Overall, Cllrs in this study were at a very low level, indicating that the system used was able to discriminate well between SS and DS comparisons and well-calibrated even at small relevant population sizes. The analysis of the Cllrs more clearly showed that the bottom relevant population had worse outcomes compared to the other selections. This was particularly true for the relevant population sizes of up to 50 in male mock cases and across all sizes for female mock cases. This may be explained by the larger relevant population superset available for female speakers, resulting in more similar-sounding top and more dissimilar-sounding bottom relevant populations.

This study suggests that even within different relevant populations that are all logically relevant (based on gender, language, and recording conditions) the strength of evidence can vary. This may depend on the perceptual similarity of the relevant population to the questioned recording or whether accent-labels are taken into account. Future research should consider the selection of relevant population speakers based on a combination of metadata and perceptual similarity with a suitable database and a larger number of mock cases.

In non-forensic applications of ASR, for example, commercial or investigative use-cases, there is also a requirement for calibration in order to reliably interpret the output of a system and make decisions about identity. To this end, there have been research efforts towards automatic selection of calibration data for a given comparison, e.g. trial-based calibration [28]. This is similar in spirit to the approach taken in the present study, with a focus on conditions rather than speakers. We note that what sets forensic applications apart is the requirement to satisfy the hypotheses in the case (e.g. ensuring logical relevance of the relevant population) in addition to any automatically derived selection of data.

5. Conclusion

This study explored different strategies for selecting relevant populations and their impact on the estimation of the strength of the evidence. These strategies included random and speaker-demographic metadata-based selection as well as perceptual voice similarity. From a database controlled for gender, language, and recording conditions, an automatic approach using phonetic features was used to select the most and least perceptually similar sounding speakers to questioned speaker recordings in mock cases. The strength of evidence was compared for male and female mock cases and across different relevant population sizes in terms of logLRs and Cllrs. Random and similarity-based selections show convergence in Cllr as the selected population size approaches 50. It was observed that using perceptually similar speakers improves the overall Cllr for both male and female mock cases. On an individual case level, considerable variability in the effect on the estimation of the strength of evidence was observed and the effects on the strength of evidence may depend on how perceptually similar or dissimilar the recordings within the mock case are. Further work focusing on systematically controlling how similar the samples are in each case and evaluating a larger set of cases could provide additional insight into the impact of using similar-sounding relevant populations.

6. References

[1] David van der Vloed and Tina Cambier-Langeveld, "How we use automatic speaker comparison in forensic practice," *International Journal of Speech, Language and the Law*, vol. 29, no. 2, pp. 201–224, 2023.

[2] Colleen Kavanagh, Peter Milne, and Emily Lawrie-Munro, "Forensic voice comparison in Canada," in *Proceedings of the Annual Conference of the International Association for Forensic Phonetics and Acoustics*, Zürich, Switzerland, July 2023, p. 42.

[3] Katharina Klug, Michael Jessen, Yosef A. Solewicz, and Isolde Wagner, "Collection and analysis of multi-condition audio recordings for forensic automatic speaker recognition," *Proceedings of XVII AISV (Associazione Italiana Scienze della Voce) conference: 'Speaker Individuality in Phonetics and Speech Sciences: Speech Technology and Forensic Applications'*, vol. 8, pp. 57–76, 2021.

[4] Erica Gold and Peter French, "International practices in forensic speaker comparisons: Second survey," *International Journal of Speech, Language and the Law*, vol. 26, no. 1, pp. 1–20, 2019.

[5] Geoffrey Stewart Morrison, "Advancing a paradigm shift in evaluation of forensic evidence: The rise of forensic data science," *Forensic Science International: Synergy*, vol. 5, pp. 100270, 2022.

[6] Andrzej Drygajlo, Michael Jessen, Stefan Gfroerer, Isolde Wagner, Jos Vermeulen, and Tuija Niemi, *Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition*, 2015.

[7] Philip Rose, *Forensic speaker identification*, Taylor & Francis e-Library, 2002.

[8] Bernard Robertson, G. A. Vignaux, and Charles E. H. Berger, *Interpreting evidence: evaluating forensic science in the courtroom*, Wiley, 2 edition, 2016.

[9] Erica Gold and Vincent Hughes, "Issues and opportunities: The application of the numerical likelihood ratio framework to forensic speaker comparison," *Science and Justice*, vol. 54, no. 4, pp. 292–299, 2014.

[10] Vincent Hughes and Richard Rhodes, "Questions, propositions and assessing different levels of evidence: Forensic voice comparison in practice," *Science and Justice*, vol. 58, no. 4, pp. 250–257, 2018.

[11] Phil Rose, "Technical forensic speaker identification from a bayesian linguist's perspective," in *Proceedings of The Speaker and Language Recognition Workshop (Odyssey 2004)*, Toledo, Spain, May 2004, pp. 3–10.

[12] Vincent Hughes, *The definition of the relevant population and the collection of data for likelihood ratio-based forensic voice comparison*, Ph.D. thesis, University of York, 2014.

[13] Geoffrey Stewart Morrison, Felipe Ochoa, and Tharmarajah Thiruvaran, "Database selection for forensic voice comparison," in *Proceedings of The Speaker and Language Recognition Workshop (Odyssey 2012)*, Singapore, June 2012, pp. 62–77.

[14] David H. Kaye, "Logical relevance: Problems with the reference population and DNA mixtures in people v. pizarro," *Law, Probability & Risk*, vol. 3, pp. 211–220, 2004.

[15] David H. Kaye, "DNA probabilities in people v. prince: When are racial and ethnic statistics relevant?," *Institute of Mathematical Statistics*, vol. 2, pp. 289–301, 2008.

[16] Vincent Hughes and Paul Foulkes, "The relevant population in forensic voice comparison: Effects of varying delimitations of social class and age," *Speech Communication*, vol. 66, pp. 218–230, 2015.

[17] Vincent Hughes and Paul Foulkes, "What is the relevant population? Considerations for the computation of likelihood ratios in forensic voice comparison," in *Proceedings of Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 3772–3776.

[18] Linda Gerlach, Kirsty McDougall, Finnian Kelly, Anil Alexander, and Francis Nolan, "Exploring the relationship between voice similarity estimates by listeners and by an automatic speaker recognition system incorporating phonetic features," *Speech Communication*, vol. 124, pp. 85–95, 2020.

[19] Linda Gerlach, Kirsty McDougall, Finnian Kelly, and Anil Alexander, "Automatic assessment of voice similarity within and across speaker groups with different accents," in *Proceedings of the 20th International Congress of Phonetic Sciences*, Prague, Czech Republic, August 2023, pp. 3785–3789.

[20] Finnian Kelly, Oscar Forth, Samuel Kent, Linda Gerlach, and Anil Alexander, "Deep neural network based forensic automatic speaker recognition in VOCALISE using x-vectors," in *Proceedings of the AES International Conference*, Porto, Portugal, June 2019, p. Paper 27.

[21] GBR-ENG database, "A telephonic speech database collected for the UK Government for evaluating speech technologies," 2019.

[22] "Bio-Metrics 1.8 Performance Metrics Software," 2019.

[23] Shunichi Ishihara and Yuko Kinoshita, "How many do we need? Exploration of the population size effect on the performance of forensic speaker classification," in *Proceedings of Interspeech 2008*, Brisbane, Australia, September 2008, pp. 1941–1944.

[24] Yuko Kinoshita and Shunichi Ishihara, "Background population: How does it affect LR-based forensic voice comparison?," *International Journal of Speech, Language and the Law*, vol. 21, no. 2, pp. 191–224, 2014.

[25] Vincent Hughes, "Sample size and the multivariate kernel density likelihood ratio: How many speakers are enough?," *Speech Communication*, vol. 94, pp. 15–29, 2017.

[26] Niko Brümmer and Edward de Villiers, "The bosaris toolkit user guide: Theory, algorithms and code for binary classifier score processing," Tech. Rep., 2011.

[27] Niko Brümmer and Edward de Villiers, "The bosaris toolkit: Theory, algorithms and code for surviving the new dcf," Tech. Rep., 2013.

[28] Luciana Ferrer, Mahesh Kumar Nandwana, Mitchell McLaren, Diego Castan, and Aaron Lawson, "Toward Fail-Safe Speaker Recognition: Trial-Based Calibration With a Reject Option," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 140–153, 2019.