INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Null-Hypothesis LLR: A proposal for Forensic Automatic Speaker Recognition

Yosef A. Solewicz[1], *Michael Jessen*[2], *David van der Vloed*[3]

[1]National Police, Israel
[2]Bundeskriminalamt, Germany
[3]Netherlands Forensic Institute, Netherlands

solewicz@police.gov.il, michael.jessen@bka.bund.de, d.van.der.vloed@nfi.minvenj.nl

Abstract

A new method named Null-Hypothesis LLR (H_0LLR) is proposed for forensic automatic speaker recognition. The method takes into account the fact that forensically realistic data are difficult to collect and that inter-individual variation is generally better represented than intra-individual variation. According to the proposal, intra-individual variation is modeled as a projection from case-customized inter-individual variation. Calibrated log Likelihood Ratios (LLR) that are calculated on the basis of the H_0LLR method were tested on two corpora of forensically-founded telephone interception test sets, German-based GFS 2.0 and Dutch-based NFI-FRITS. Five automatic speaker recognition systems were tested based on the scores or the LLRs provided by these systems which form the input to H_0LLR. Speaker-discrimination and calibration performance of H_0LLR is comparable to the performance indices of the system-internal LLR calculation methods. This shows that external data and strategies that work with data outside the forensic domain and without case customization are not necessary. It is also shown that H_0LLR leads to a reduction in the diversity of LLR output patterns of different automatic systems. This is important for the credibility of the Likelihood Ratio framework in forensics, and its application in forensic automatic speaker recognition in particular.

Index Terms: automatic speaker recognition, forensic speech science, likelihood ratios¸ calibration

1. Introduction

When automatic speaker recognition is applied in the forensic domain, the success of the method should be validated on a corpus of speech that is directly representative of casework in forensic voice comparison [1]. Forensically realistic data are also relevant in the final stages of the automatic analysis (scoring and in particular the calculation of calibrated Likelihood Ratios) in which domain-specific input is important. Despite these demands, it is a common problem in Forensic Automatic Speaker Recognition (FASR) that case-relevant data are difficult to collect and that the number of speakers and recordings is very low compared to the very large amounts of data that are used to train state-of-the-art speaker recognition systems [2, 3; see 4 for overview] using domain-independent data from international corpora. Within that overall difficulty, it is easier to sample isolated recordings from different speakers than to collect data in which two or more recordings from the same speaker are available. As a result, there are more resources in FASR available for the modeling and testing of inter-speaker variation than intra-speaker variation. Different automatic systems deal with the intra-speaker variation problem in various ways, including the use of domain-independent resources or certain score adjustment strategies. In the present paper, a proposal is made that relies entirely on domain-specific (forensic) data. The proposal elaborates upon [5] and shows that the intra-speaker distribution of scores can be estimated by projecting from the inter-speaker distribution. Since the latter reflect the null hypothesis in the LLR calculation process (i.e. questioned speaker and suspect are non-identical), the LLR output obtained with the proposed method is called “Null-Hypothesis LLR” and will be abbreviated as H_0LLR. It is investigated in this paper whether the H_0LLR method, with its advantage of full domain-specificity, performs on a level with (or even better than) other, system-internal, LLR methods or whether the advantages of H_0LLR have to be traded in for a reduction in performance.

A second intention of this study is to improve upon the following situation. When different systems are tested using the same corpus (results to be shown in this paper), the range of LLRs reported by the different systems can differ substantially. For example, the same comparison might give a value of LLR=5 in system-1 but LLR=3 in system-2. This is an undesired situation because it will be difficult for courts to understand why LLRs are so variable despite existing best practice proposals that emphasize the solidness of LLRs and the LR-framework [6]. Although some of the differences between systems result from differences deep in the architecture and are hard to bridge, it will be investigated here whether the difference between the LLR output of different systems can at least be reduced, if at the LLR-calculation stage the system-internal methods are replaced with H_0LLR-based LLR-calculation.

2. Method

2.1. Prerequisites to H_0LLR

One frequently used procedure for arriving at calibrated LLRs in FASR is the scoring method ([1] for overview), which works roughly as follows. The comparison between the questioned-speaker recording and the suspect recording yields a score. In order to obtain the numerator of the LLR, the score is compared with the intra-speaker distribution and for the denominator the score is compared with the inter-speaker distribution. The inter-speaker distribution is supplied by comparing the questioned speaker with the speakers of a Reference Population (RP), i.e. a database of recordings of about 30 speakers or more that is representative for the conditions of the case ([7] for an example of the methods in a

Copyright © 2017 ISCA

http://dx.doi.org/10.21437/Interspeech.2017-1023

commercial system). This way of supplying the inter-speaker distribution can be considered a *customized* method (i.e. it is adapted to the conditions of the case by including data from the case, here the questioned speaker). There is less agreement on how to supply the intra-speaker distribution. One method is to arrive at a customized intra-speaker distribution by comparing several recordings or recording-segments of the suspect with each other and applying some adjustment (because the scores are often biased when stemming from the same recording) (cf. [8]). Another, non-customized, method is to use intra-speaker data from comparisons that are unrelated to the case and perhaps domain-independent for forensics. It is also possible to combine both of these methods by starting from a non-customized distribution as a prior and adapt it with data from the suspect if enough suspect data is available [9]. What these methods have in common is that either non-customized data are necessary or strategies of score adjustment (or both). In contrast, the method proposed here will rely entirely on customized data. As made explicit in section 2.2, the method arrives at an inter-speaker distribution by comparing the questioned speaker and the suspect with the RP and it projects the intra-speaker distribution from this inter-speaker distribution of scores.

H_0LLR builds upon the concept of the "normalized p-value" (see also [5]). The idea is to obtain the inter-speaker distribution H_0 (in ways shown above, using a reference population *r*) and determine the p-value, which is the probability of obtaining scores higher than the comparison score, given H_0. The comparison score is the result of the comparison between questioned-speaker (*x*) and suspect (*y*) (for an illustration, this score is labeled "S" in Figure 1). Subsequently, the user (forensic expert) inputs her/his expected EER (Equal Error Rate) estimate for the system, given the environment conditions (channel, noise, duration, etc.) of *r*. This estimate is informed by previous forensic validations and publicly available NIST-like SRE reports approximating the conditions present in a case.[1] The following formula is proposed for the normalized p-value:

$$\overline{\text{p-value}}(x, y, r) = 1 - \frac{\text{p-value}(x, y, r)}{\text{EER}(r)}$$

Normalized p-values' range can be conveniently confined to [0,1], if negative values are truncated to zero. The rationale behind it is to normalize the probability of rejecting the null hypothesis by the uncertainty involved in the comparison. Furthermore, this complement of the ratio between p-value and EER gives a sense of recognition score (the closer to one, the stronger the evidence for speaker identity).

According to the normalized p-value idea, a p-value close to (or greater than) the expected EER leads to a low (or zero by definition) normalized p-value, suggesting that the null hypothesis of an impostor match should not be rejected. If, on the other hand, the p-value is very low with respect to the EER, the normalized p-value will be high, suggesting the alternative hypothesis of a target match can probably be accepted.

Since the uncertainty involved in the decision is expressed in terms of EER, the alternative hypothesis is assessed indirectly and therefore there is a move from the frequentist notion of the p-value into the direction of the Bayesian LR approach. The H_0LLR method, made explicit in the 2.2, completes this move by bringing the frequentist notion of the normalized p-value fully in line with the Bayesian LR-approach [10, 11]. This is achieved by projecting a virtual alternative hypothesis distribution across the score corresponding to the expected EER to obtain the LR numerator and therefore an approximation of the LLR.

2.2. H_0LLR method: Calculations

The proposed H_0LLR method proceeds along the following sequence of steps. The major results of applying these steps are illustrated in Figure 1.

1. Calculate the mean and standard deviation of impostor (i.e. inter-speaker) scores, specifically the scores obtained by comparing the *questioned speaker* with the speakers of the reference population, arriving at μ_q, σ_q.

2. Calculate the mean and standard deviation of impostor scores, specifically the scores obtained by comparing the *suspect* with the speakers of the reference population, arriving at μ_s, σ_s.

3. Estimate $h_0 = N(\mu, \sigma^2)$, the null hypothesis (inter-speaker) PDF, where $\mu = \frac{1}{2}(\mu_q + \mu_s)$ and $\sigma = \frac{1}{2}(\sigma_q + \sigma_s)$. This step acts similar to Symmetric Normalization (S-norm) [12]. The null hypothesis PDF is shown in Fig. 1 as the lefthand Gaussian labeled H_0.

4. Assess p_{EER}, the expected EER for the comparison and find E, the corresponding score w.r.t h_0, by computing H_0^{-1}, the inverse CDF of h_0 at the complement of p_{EER}:

$$E = H_0^{-1}((1 - p_{EER})|\mu, \sigma^2) \tag{1}$$

As mentioned in 2.1, the concept behind p_{EER} is that the user of a FASR system makes an estimate of about how well the automatic system will perform in terms of EER under the conditions of the case.

5. Calculate S, the score obtained when the questioned speaker is compared with the suspect.

6. Project S across E to find the projected score, S':

$$S' = 2E - S \tag{2}$$

7. Estimate H_0LLR as the logarithm of the ratio between the H_0 PDF at S' and S, respectively:

$$H_0\text{LLR} = \log_{10}\left(\frac{h_0(S'|\mu,\sigma^2)}{h_0(S|\mu,\sigma^2)}\right) \tag{3}$$

This expression further results in an affine score-to-LLR mapping.

We assume the same σ^2 for H_1 (LLR numerator). This decision is motivated from previous experimentation in which the equal variance assumption promised robustness against mismatch situations. It also avoids modeling artifacts as discussed in [13]. Equal variance in the intra- and inter-speaker distribution is also proposed in [14].

As shown with the righthand Gaussian in Fig. 1, an intra-speaker distribution has been virtually projected from the inter-speaker distribution through the procedure shown here.

2.3. Testing H_0LLR on forensic corpora

The H_0LLR proposal expressed in 2.2 was tested on two forensic corpora, GFS 2.0 and NFI-FRITS. GFS 2.0 (German

[1] It can be shown that H_0LLR is relatively robust against deviations around the optimum (the "true") EER settings for the case conditions (but the technicalities are beyond of the scope of this paper).

Forensic Speech Corpus) is similar to, and partially overlapping, with the corpus GFS 1.0 that had been released by the Bundeskriminalamt (BKA) in 2011 for a collaborative exercise within ENFSI (European Network of Forensic Science Institutes) [15, 16]. Compared to GFS 1.0, GFS 2.0 is more rigidly focused on natural telephone conversations from telephone interceptions. 23 German-speaking test speakers (with two recordings per speaker from different sessions) were used with net duration ranging between about 20 and 60 seconds.

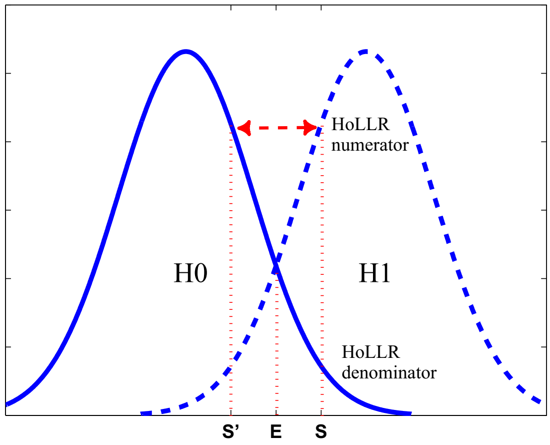

Figure 1: *Illustration of the Null-Hypothesis LLR proposal. X-axis: score level. Y-axis probability density. Both scales are represented abstractly.*

NFI-FRITS is a corpus based on telephone-interception recordings containing data from several languages [17]. Out of the full NFI-FRITS corpus, a subset of 23 Dutch-speaking test speakers with two different-session recordings per speaker was used. Net duration was fixed at 30 seconds for all recordings.

A total of five forensic i-vector-based speaker recognition systems was tested, four of them commercial. The systems are Batvox 4.1 (Agnitio), iVocalise (Oxford Wave Research), Nuance Forensics 9.2, Voice Inspector (Phonexia) and VBS [18]. Of these, three provide calibrated LLRs. Since calibrated LLRs are needed for a baseline test, to be explained shortly, the main focus of the study was on these three. They will be referred to as system-1 to system-3. System-1 and -2 require a reference-population (RP) for LLR-calculation, whereas system-3 has no option of specifying a RP when operating in batch mode (which is necessary when, as here, making hundreds of comparisons). In addition to providing LLRs, system-1 and system-2 offer the option of extracting the scores before they are converted into LLRs.

Three different types of analysis were carried out for each of the systems (as far as possible given the design differences just mentioned).

In the first type, called baseline, the system is used with the regular LLR output of the system. The RPs required for system-1 and -2 consist of single recordings of 30 speakers (distinct from the tested speakers), drawn from the resources of GFS 2.0 when testing with GFS 2.0 and NFI-FRITS when testing with NFI-FRITS.

In the second type of analysis, called H_0LLR, the system is used with the score output of the system. These scores are used as input to the H_0LLR method shown in 2.2. The RP needed for that purpose (see step 1 and 2 in 2.2) is the one mentioned for the baseline condition. This second analysis type was not applicable for system-3 because it has no separate score output that is different from the LLR output.

In the third type of analysis, called H_0LLR-recal, the system is used for regular LLR output, as in the baseline condition. This output is used as input to the H_0LLR method. Since H_0LLR operates on calibrated LLRs, H_0LLR can be considered as a re-calibration process in this type of analysis. (Note that in principle re-calibration would have no effect on already well-calibrated LLRs due to the LR idempotence property [14].) Since a RP is needed for system-1 and -2 when calculating LLRs, and a RP is also needed for H_0LLR, there are two different RPs that need to be kept distinct. The RP used for H_0LLR is the dedicated one that was mentioned for the other two types of analysis. The RP needed for the LLR calculation in system-1 and system-2 is one consisting of 30 landline-phone speakers of German (when working with GFS 2.0) and Dutch (when working with NFI-FRITS). These RP data were kindly provided by the manufacturer of one of the systems. For the second and third type of analysis, the estimated EER value needed for H_0LLR was 5% (see step 4 in 2.2), which is a representative performance level for the type of recordings and automatic systems tested here.

3. Results and Discussion

The results of the three types of analysis applied to the three FASR systems (with the mentioned exception of the second analysis type on system-3) are shown in Table 1. The table shows how each of these tests (rows) performs in terms of different well-known performance descriptors, including EER, Cllr_cal, Cllr_min [19, 20] and NIST's minimum detection cost function (DCF) with SRE'08 parameters [21].

Table 1: *Results of tests with three FASR systems based on the corpus GFS 2.0 (above) and NFI-FRITS (below).dcf08 is multiplied with 100, cllr_cal and cllr_min with 10.*

GFS 2.0	EER	dcf	Cllr	Cllr_
	(%)	08	_cal	min
Syst-1 baseline	6.9	1.9	2.9	1.4
Syst-1 H_0LLR	6.3	1.5	2.4	1.3
Syst-1 H_0LLR-recal	7.9	1.5	2.3	1.3
Syst-2 baseline	8.7	3.8	3.6	2.0
Syst-2 H_0LLR	8.7	2.6	3.5	2.2
Syst-2 H_0LLR-recal	8.7	3.6	3.9	2.5
Syst-3 baseline	4.0	2.2	2.2	0.8
Syst-3 H_0LLR-recal	1.6	1.5	2.2	0.4
NFI-FRITS	EER	dcf	Cllr	Cllr_
	(%)	08	_cal	min
Syst-1 baseline	4.3	2.5	3.1	1.0
Syst-1 H_0LLR	4.3	2.3	2.6	0.9
Syst-1 H_0LLR-recal	4.3	2.2	2.5	1.0
Syst-2 baseline	11.7	4.5	4.3	2.9
Syst-2 H_0LLR	7.5	2.9	3.8	2.1
Syst-2 H_0LLR-recal	8.3	3.5	4.0	2.1
Syst-3 baseline	4.2	1.6	2.7	0.7
Syst-3 H_0LLR-recal	4.3	1.5	2.5	0.7

From the results in Table 1 it is possible to compare the discrimination performance (focusing the discussion on EER and Cllr_min) and the calibration performance (Cllr_cal) between the baseline application of the three systems on the one hand and the application of H_0LLR (both score-based and in re-calibration mode as far as possible) on the other hand.

As for discrimination (looking at EER and Cllr_min here), H_0LLR leads to better performance than baseline where system-3 is tested with GFS 2.0 and where system-2 is tested with NFI-FRITS. For the other tests the discrimination differences when H_0LLR is compared with baseline are less systematic or very small.

As for calibration (looking at Cllr_cal), there is an advantage of H_0LLR compared to baseline where system-1 is tested with GFS 2.0 and NFI-FRITS and system-2 (and slightly system-3) with NFI-FRITS.

Although there are several details emerging from these test results, the very least that can be said about them is that the application of the H_0LLR method does not lead to any systematically worse discrimination and calibration performance than the calibrated-LLR calculation methods that are intrinsic to the tested systems (baseline).

Figure 2 shows Tippett plots of the different types of analysis applied to the corpora GFS 2.0 and NFI-FRITS.[1]

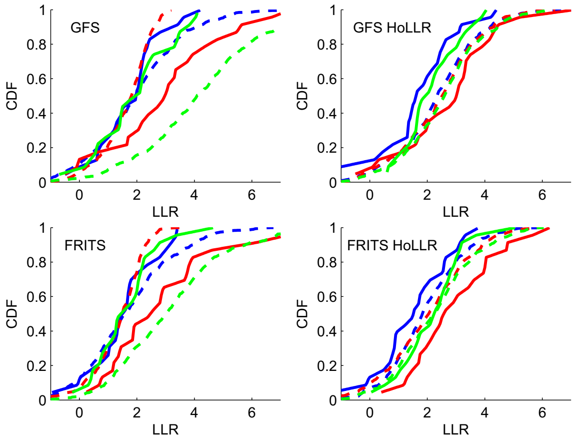

Figure 2: *Tippett plots based on tests with GFS 2.0 (upper two plots) and NFI-FRITS (lower two) in the baseline analysis (left two) and analysis with H_0LLR (right two). Red lines: system-1, blue: system-2, green: system-3. Solid lines: target trials, dashed: impostor trials (inverted). LLR expressed in log10.*

In order to facilitate the discussion, the way the Tippett plots are shown in Fig. 2 differs from the more common ways. Specifically, starting from "type II" Tippett plots mentioned in [1], the impostor trials (results from different-speaker comparisons), which (mostly) show negative LLRs, are inverted into the (mostly) positive range of LLRs that is also occupied by the target trials (same-speaker comparisons).

The Tippett plots in Fig. 2 show that in baseline operation (lefthand plots) the different systems differ quite visibly in the range of LLRs. Within systems-1 and -3 there is also a marked difference between the distributions of the target trials and the impostor trials. This diversity of LLRs exemplifies the problem mentioned in the Introduction that the same comparison might yield quite different LLRs when analyzed with different systems. It also demonstrates that for some systems target and impostor trials are not distributed symmetrically.

The righthand plots in Figure 2 show that the application of the H_0LLR method proposed in this paper has the effect of bringing the LLR ranges occupied by different systems into closer correspondence with one another. H_0LLR application also has the effect that the difference in the shapes between target and impostor patterns (asymmetry) that occurred in the baseline application of two of the systems is now considerably reduced. In work in progress by the authors, this visual convergence of patterns from different systems that is achieved by H_0LLR will be quantified with reference to a range of LLRs commonly used in forensics.

Figure 2 also shows that the patterns found with the two tested corpora are very similar, both in baseline application and in terms of the effect of H_0LLR. This similarity might result from the similarity of the conditions (telephone-intercepted speech) and the same number of tested speakers. The language difference between GFS 2.0 (German) and NFI-FRITS (Dutch) did not seem to matter for the LLR results nor did the different duration schemes mentioned in 2.3.

The two systems, among the total of five tested, that do not provide LLRs, resulted in patterns similar to the other three when H_0LLR was applied to the scores provided by the two systems.

4. Conclusion

A new method for the calculation of calibrated LLRs in automatic speaker recognition named H_0LLR was introduced. With this method the intra-individual variation that is necessary for the numerator of the LLR is derived as a projection from inter-individual variation that can be modeled much more robustly given the sparsity of forensically realistic data. For LLR calculation the H_0LLR method does not require any data from outside the domain of forensics and it relies on customized (case-adapted) data only. Based on tests with forensically realistic corpora and various i-vector-based automatic systems it was shown that the advantages of domain-specificness and case customization are not accompanied by any loss of performance compared to the inbuilt LLR calculation methods of the tested systems that rely on different strategies (Table 1).

When the results of the tests are represented as Tippett plots (Figure 2) it turns out that the application of H_0LLR reduces the differences in the LLR values and ranges that are given by different automatic systems and it increases the symmetry between target and impostor distributions. Such a result improves the credibility of the LR-framework in forensics generally [6] and specifically the way (L)LRs are calculated in automatic speaker recognition. The intention behind this research is to increase the level of transparency, reproducibility and forensic relevance when working with different automatic systems.

5. Acknowledgements

A large proportion of the research presented here was made possible by travel funding within the German-Israeli Cooperation on Counter-Terrorism Technology.

[1] Among the H_0LLR-based analyses (righthand plots in Fig. 2), the results for Syst-1 H_0LLR-recal and for Syst-2 H_0LLR-recal are omitted for readability and only the results from the score-based analysis types Syst-1 H_0LLR and Syst-2 H_0LLR are shown. For system-3 the result shown is the only one available, i.e. Syst-3 H_0LLR-recal.

6. References

[1] A. Drygajlo, M. Jessen, S. Gfroerer, I. Wagner, J. Vermeulen, and T. Niemi, *Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition*. Frankfurt: Verlag für Polizeiwissenschaft, 2015. [also accessible at http:// enfsi.eu/wp-content/uploads/2016/09/guidelines _fasr_and_fsasr_0.pdf].

[2] N. Dehak, *Discriminative and Generative Approaches for Long- and Short-Term Speaker Characteristics Modeling: Application to Speaker Verification*, Ph.D. thesis, Ecole de Technologie Superieure de Montreal, 2009.

[3] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 788–798, 2011.

[4] J. H. L. Hansen and T. Hasan, "Speaker recognition by machines and humans," *IEEE Signal Processing Magazine*, vol. 32, pp. 74–99, 2015.

[5] Y. A. Solewicz, G. Jardine, T. Becker, and S. Gfrörer, "Estimated intra-speaker variability boundaries in forensic speaker recognition casework," *Proceedings of Biometric Technologies in Forensic Science (BTFS)* (Nijmegen), pp. 31–33, 2013.

[6] ENFSI Guideline for evaluative reporting in forensic science. http://enfsi.eu/news/enfsi-guideline-evaluative-reporting-forensic-science/

[7] D. van der Vloed, "Evaluation of BatVox 4.1 under conditions reflecting those of a real forensic voice comparison case (*forensic_eval_01),*" *Speech Communication,* vol. 85, pp. 127–130, 2016.

[8] J. Gonzalez-Rodriguez, A. Drygajlo, D. Ramos-Castro, M. Garcia-Gomar, and J. Ortega-Garcia, "Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition," *Computer Speech and Langua*ge, vol. 20, pp. 331–355, 2006.

[9] D. Ramos-Castro, J. Gonzalez-Rodriguez, A. Montero-Asenjo, and J. Ortega-Garcia, "Suspect-adapted MAP estimation of within-source distributions in generative likelihood ratio estimation," *Proceedings of ODYSSEY* (San Juan), pp. 1–5, 2006.

[10] A. Drygajlo, D. Meuwly, and A. Alexander, "Statistical methods and Bayesian interpretation of evidence in forensic automatic speaker recognition*," Proceedings of EUROSPEECH* (Geneva), pp. 689–692, 2003.

[11] J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. Toledano and J. Ortega-Garcia, "Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 2104–2115, 2007.

[12] S. Shum, N. Dehak, R. Dehak, and J. Glass, "Unsupervised speaker adaptation based on the cosine similarity for text-independent speaker verification," *Proceedings of Odyssey* (Brno), pp. 76–82, 2010.

[13] G. S. Morrison, "Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio," *Australian Journal of Forensic Science,* vol. 45, pp. 173–197, 2013.

[14] D. A. van Leeuwen and N. Brümmer, "The distribution of calibrated likelihood-ratios in speaker recognition," *Proceedings of INTERSPEECH* (Lyon), pp. 1619–1621, 2013.

[15] Y. A. Solewicz, T. Becker, G. Jardine, and S. Gfrörer, "Comparison of speaker recognition systems on a real forensic benchmark," *Proceedings of ODYSSEY* (Singapore), pp. 86–91, 2012.

[16] T. Becker, Y. A. Solewicz, G. Jardine, and S. Gfrörer, "Comparing automatic forensic voice comparison systems under forensic conditions." *Proceedings of the Audio Engineering Society* (Denver), pp. 197–202, 2012.

[17] D. van der Vloed, J. Bouten and D. A. van Leeuwen, "*NFI-FRITS:* A forensic speaker recognition database and some first experiments," *Proceedings of ODYSSEY* (Joensuu), pp. 6–13, 2014.

[18] http://voicebiometry.org/

[19] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech and Language*, vol. 20, pp. 230–275, 2006.

[20] D. A. van Leeuwen and N. Brümmer, "An introduction to application-independent evaluation of speaker recognition systems," in: C. Müller (ed.) *Speaker Classification I: Fundamentals, Features, and Methods*. Berlin: Springer, pp. 330–353, 2007.

[21] http://www.itl.nist.gov/iad/mig/tests/sre/