Comparison of Speaker Recognition Systems on a Real Forensic Benchmark

Yosef A. Solewicz[1], Timo Becker[2], Gaëlle Jardine[3] and Stefan Gfroerer[2]

[1]National Police, Israel
[2]Federal Criminal Police Office, Germany
[3]Police Technique et Scientifique, France

solewicz@police.gov.il, timo.becker@bka.bund.de, gaelle.jardine@interieur.gouv.fr, stefan.gfroerer@bka.bund.de

Abstract

This paper analyses the performance of several automatic speaker recognition systems using a real forensic database. The systems evaluated have been tested or are currently in use by forensic institutes. A comprehensive error analysis is performed in order to assess the each system's behaviour to real casework. We further investigate compensation techniques aimed at minimising the performance gap between laboratory development and application on real forensic data. While unrestricted application of automatic systems in the forensic domain is still not a reality, our experiments suggest that automatic systems can be a valuable support in decision-making for the forensic examiner.

1. Introduction

Automatic Speaker Recognition is benchmarked by the National Institute of Standards and Technology (NIST) [1]. NIST campaigns address a broad range of speaker recognition applications, which include, but go far beyond, forensics.

In the forensic scenario, the examiner, possibly assisted by automatic systems, should state how similar two speakers are, expressed, for instance, as likelihood ratios. Any quantitative inference involves statistics that are based on reference speakers and express the strength of the forensic evidence. Realistic forensic evaluations therefore require that the databases used be representative of real forensic scenarios, as for example in [2]. Another peculiarity of the forensic world is the way one measures system performance. While most systems are commonly evaluated in terms of general performance figures, the forensic examiner is concerned with individual comparisons and is interested, ultimately, in knowing under which circumstances some system could be helpful and under which circumstances results of some system may be misleading.

In this paper, we analyse the performances of several speaker recognition systems from a forensic perspective. All evaluated systems bar one are currently used in forensic laboratories: at the German Federal Criminal Police Office (Bundeskriminalamt – BKA), the French Police Technique et Scientifique (PTS) and the Israeli National Police (INP). The corpus that is used for the evaluation is a collection of telephone taps taken from real cases.

What we are interested in are the systems' responses to outlier trials. We establish a protocol to identify individual systems' strengths and weaknesses in light of particular types of interferences that are typically present in forensic evidence. We further conduct post evaluation experiments to identify these interferences and reduce their effects.

The paper is organised as follows. Sections 2 and 3 respectively describe the corpus and systems used in these experiments. Section 4 describes the experiments performed, followed by analyses of the results in Section 5. A post evaluation of the results is described in Section 6. Section 7 concludes the paper and discusses future plans.

2. Corpus

The corpus used in these experiments is the "GFS1.0-Corpus (German Forensic Speech Corpus Version 1.0)". The corpus is property of the BKA and comprises a selection of forensic telephone taps edited by the German Forensic Science Institute, Section KT54-1. GFS was produced as part of the EU project *Correlation between phonetic–acoustic–auditory and automatic approaches in forensic speaker identification* in order to make authentic forensic data available to forensic institutes[1]. The corpus contains spontaneous German speech recordings of male individuals. The current version consists of two protocol variants, one of which is more oriented towards analyses by automatic systems. This is the one used in the present evaluations. It is organised as follows.

- 39 offender recordings with a minimum duration of 30 seconds, originating from 24 speakers.
- 21 suspect recordings with a minimum duration of 60 seconds, originating from 21 speakers.
- 49 reference population recordings with a minimum duration of 60 seconds, originating from 49 speakers.

3. System descriptions

In this section, we briefly describe the seven systems used in the analyses. The systems consist of two groups. The ISR systems are the result of in-house development at the INP. The EUR group consists of two systems developed by the BKA and two commercial products in use or tested by the PTS. EUR systems are currently used in a collaborative

[1] Since it contains case recordings, albeit anonymised, the GFS1.0-Corpus is available only to official forensic science institutes.

exercise of the European Network for Forensic Science Institutes (ENFSI) and are therefore anonymised.

3.1. ISR systems

ISR1 is a GMM-SVM-NAP system, described in more detail in [3]. The development data used was obtained from NIST evaluations 2004 and 2006. SVM models are trained using the 49 background speakers from GFS.

ISR2 uses i-vector features extracted through maximum a posteriori adaptation followed by linear discriminant analysis (LDA) as a second processing layer [4]. The scoring stage is performed within the Mahalanobis metrics [5]. Data from NIST 2004, 2006 and 2008 evaluations was used for system development.

ISR3 is similar to ISR2, except that the i-vectors are extracted through principal component analysis (PCA). The data used to build the PCA matrix is previously processed through two-wire NAP [3]. The PCA and LDA mappings are then combined into a single matrix. The within-class-covariance weighting used in the scoring step is omitted since the i-vectors are derived from a previously channel-normalised eigenspace.

3.2. EUR systems

The EUR systems consist of forensic systems used or tested in forensic laboratories in Germany and France and are either commercial products or the result of autonomous development. We prefer to look at the performances of the four EUR systems as a group and do not reveal the systems' names here[1]. The order of the following systems' descriptions does not match the order of the EUR numbers.

One EUR system, known as SPES, is a RASTA-PLPCC UBM-GMM system [6]. For the experiments described here, we used Version 7.2.3, where RABM as described in [6] is not applied. The system is the result of cooperation between the BKA, Koblenz University of Applied Sciences and the Department of Phonetics at the University of Trier. The UBM is based on 1,792 recordings from 780 speakers. Over half are in German, the rest is mixed language. The background population consists of 470 recordings from 185 speakers. None of these recordings include the 49 reference speakers of GFS.

Another EUR system, known as VoCS [7], is a MFCC-RASTA UBM-GMM system. It is also in use at the BKA. The UBM consists of 23 recordings from 23 speakers (German, English, Arabic). VoCS uses the 49 reference speakers of the GFS corpus for T-NORM. The calibration parameters are estimates based on an evaluation of the AHUMADA corpus [8].

The two remaining EUR systems are commercially available products, one of which is used at the PTS. These systems are Batvox [9] version 3.1.2 Basic and LVIS [10] version 6.5 Pre-Forensic.

4. Experiments

The systems introduced in Section 3 are tested on the GFS corpus as described in Section 2. Each of the 39 offender recordings is compared with each of the 21 suspect recordings, resulting in 36 target trials and 783 impostor trials. Due to this limited number of trials, some caution must be used in the interpretation of the results.

System performances are shown in Table 1. Performance level is presented in terms of two metrics, namely the Equal Error Rate (EER) and NIST's Detection Cost Function (DCF) [1]. EER is an application-independent metric defined by the operating point at which the probability of miss detections equals that of false alarms. Since these two probabilities may differ somewhat, they are further averaged in our EER estimation. The DCF is an application-dependent metric commonly used in automatic speaker recognition evaluations. Due to lack of additional development data, both metrics are evaluated a posteriori, meaning that optimum thresholds are chosen for either operating point based on the test data. In addition, DET plots [1] for the respective systems are shown in Figure 1. DET plots assess the quality of the systems at different operating points, which allows for an easier comparison between systems. Note that the reliability of these performance metrics should not be overemphasised because of the low number of comparisons involved.

The issue of score calibration [11] is not addressed here since not all the systems produce calibrated scores, though we will use calibration in Section 5 in the context of error analysis.

Table 1: System performances

System	EER (%)	DCF (x10^2)
ISR1	11.0	3.8
ISR2	8.8	2.8
ISR3	**8.3**	**2.5**
EUR1	13.9	3.8
EUR2	**8.3**	3.3
EUR3	11.2	4.8
EUR4	10.8	3.8

[1] The aim of this paper is not to run a competition among the systems evaluated, but to learn about their commonalities and differences. As we will see below, the differences in regards to potential applicability for forensic expertise are generally small, so we would not benefit from a general ranking of the systems.

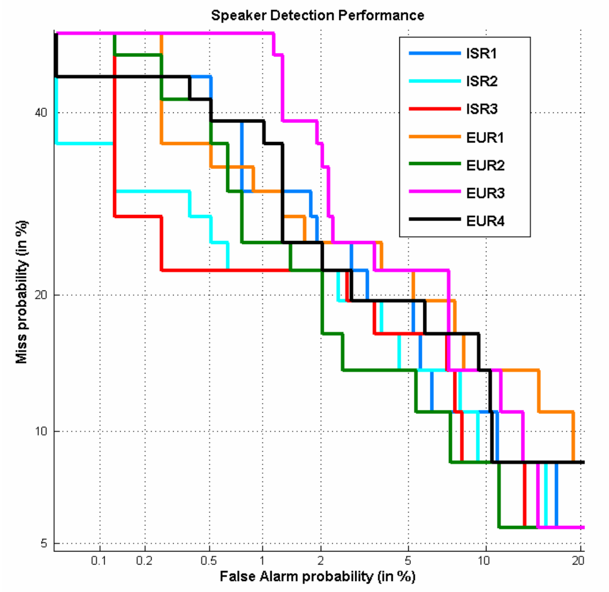

Figure 1: DET plots.

5. Discussion

5.1. Error criterion

In this section we conduct a detailed error analysis for the tested systems. Ideally, we should like to find, for each error, a link between the audio quality of the recording in question and the system's methodologies and development data.

In order to compare all systems, we need to create a common ground among the systems' outcomes. We also need to define what is an error from a forensic perspective. To standardise the results, we calibrated all systems and estimated log-likelihood-ratio costs ($\boldsymbol{C_{llr}}$) [11]. The calibration assists the examiner in making the appropriate conclusion: likelihood ratios reflect the probability of the evidence matching either of two competitive hypotheses, namely, *the samples originating from the same speaker* versus *the samples originating from different speakers*. The $\boldsymbol{C_{llr}}$ is defined as [11]:

$$C_{llr} = \frac{1}{2}\left(\frac{1}{N_t}\sum_{i=1}^{N_t}\log_2\left(1+\frac{1}{S_{t_i}}\right) + \frac{1}{N_n}\sum_{j=1}^{N_n}\log_2\left(1+S_{n_j}\right)\right) \quad (1)$$

where the first summation is over all target trials and the second over all non-target trials. $\boldsymbol{N_t}$ and $\boldsymbol{N_n}$ are the total numbers of target and non-target trials, respectively, and $\boldsymbol{S}$ represents a trial's likelihood ratio.

The minimum of this function, which is reached through an optimally calibrated system, can be used to compare overall performance among different systems. In our case, instead of searching for the global minimum in the overall cost, we look at the costs of individual trials, which is more appropriate in the context of a forensic analysis. In particular, we can infer the quality of individual target and non-target trials by isolating the corresponding members under the summations in Eq. 1. Therefore, the cost of individual target or non-target trials will be given by:

$$C_{llr}^{trial}(S) = \begin{cases} \log_2\left(1+\dfrac{1}{S}\right), & \text{target} \\ \log_2(1+S), & \text{non-target} \end{cases} \quad (2)$$

The proposed error criterion equally weights and symmetrically scales false positive and false negative flaws. High costs for either error type indicate a bad performance for the specific trial.

By means of this criterion we expect to spot and compare errors within and across different systems. Note that the scores in (1) must be expressed in terms of likelihood ratios. Since not all the systems provide likelihood ratios, we calibrate all of them using each system's test results. Though this is an optimal and possibly not realistic calibration, it seems justified in the context of a comparative analysis. In the following sections we analyse the performance attained by the systems both in isolation and comparatively with respect to the $\boldsymbol{C_{llr}}$.

5.2. Individual error analysis

Before we go into detailed individual error analysis, we start by presenting the overall performances of the systems. The minimum $\boldsymbol{C_{llr}}$ (1) attained by each calibrated system is presented in Figure 2. Recall that due to the limited number of trials in our evaluation the differences of the systems under investigation should be treated with care. We observe that $\boldsymbol{C_{llr}}$ performance metric correlates quite well with the estimated EER and DCF metrics listed in Table 1.

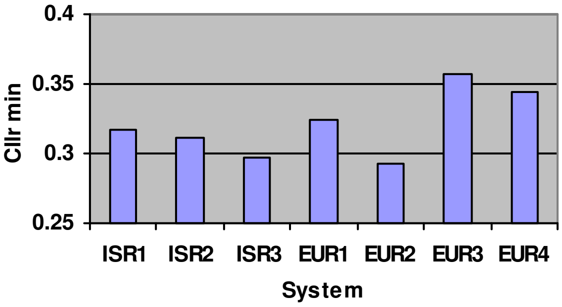

Figure 2: Minimum Cllr for the systems.

Regarding individual errors, we arbitrarily label as errors those trials that are associated with a cost greater than 1 (2). We prefer to offer independent analyses for target and non-target trials, since the attribution of cost decisions in a forensic scenario is considered to be outside the examiner's scope. Figures 3 and 4 respectively show the costs (on a logarithmic scale to include outliers) of the greatest false-negative and false-positive errors produced by each individual system. Note that after log scaling, the hard limit for misleading scores is zero.

Generally speaking, cautious systems, while also outputting a low overall $\boldsymbol{C_{llr}}$, will avoid individual high-cost outputs, which obviously diminishes their usefulness as a decision-making support tool. In this respect, EUR1 stands apart from the other systems: while it displays stable but relatively high costs in false negative trials, apart from two gross errors, it shows low costs in false-positive decisions.

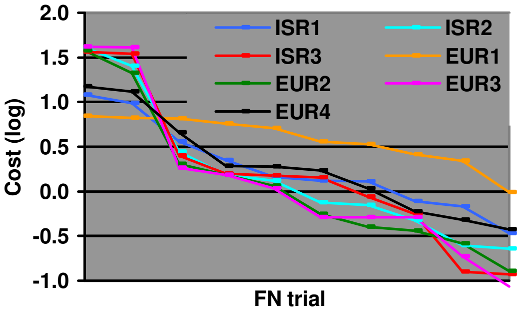

Figure 3: Most costly false negative trials per system (10 greatest errors displayed).

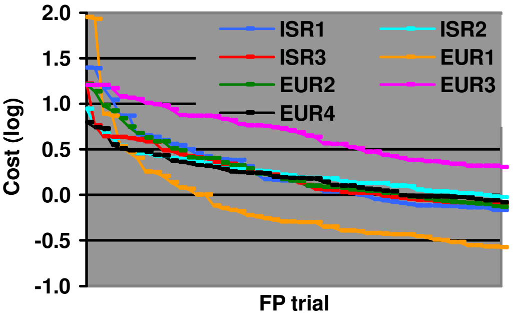

Figure 4: Most costly false positive trials per system (50 greatest errors displayed).

5.3. Relative error analysis

In Section 5.2, we examined the systems' individual cost performances and focused on their errors. We shall now look at their relative performances across all trials.

For each trial, we pick the system that has the lowest $\boldsymbol{C_{llr}}$ and calculate the difference between that cost and each of the other systems' costs. These differences are added for each system and normalised for the number of trials. A system that has a low average difference is expected to produce decisions which are often close to the best possible outcome. We further differentiate between target and impostor trials and arbitrarily split the trials into *"easy"* and *"difficult"* categories. An easy trial is one where all systems produce a C_{llr} below 1; all other trials are considered to be difficult. Roughly 80% of the impostor trials and 70% of the target trials are considered to be easy.

The average differences from optimum C_{llr} for each system are shown in Figure 5 for target trials, and in Figure 6 for impostor trials. Notwithstanding the small number of trials, especially of target trials, these figures seem to indicate different types of behaviour for different kinds of trials. Most remarkably, EUR1 seems to perform well on difficult impostor trials but poorly on difficult target trials; EUR3 seems to be misleading when dealing with difficult impostor trials. These trends were also observed when the systems were analysed separately in Section 5.2. Another interesting observation is that systems that were not ranked among the best in terms of classical overall performance measures (cf. Table 1 and Figure 2), like ISR1 and EUR4, seem to perform well in terms of difference from the lowest decision cost.

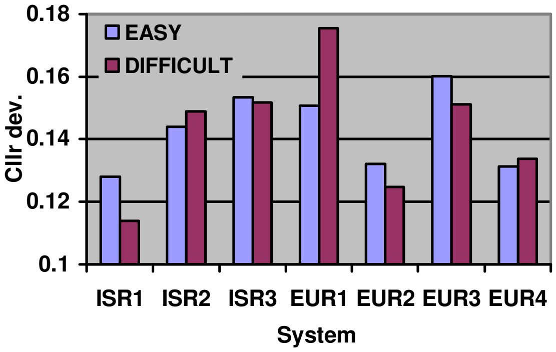

Figure 5: Average Cllr difference from best system for "easy" and "difficult" target trials.

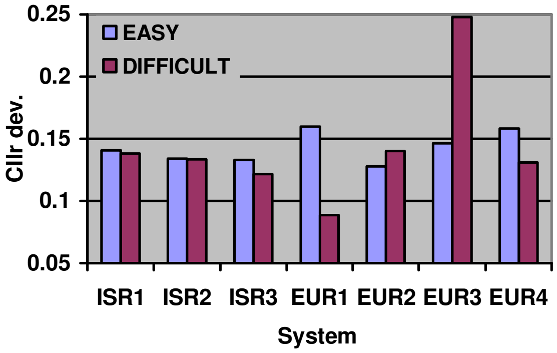

Figure 6: Average C_{llr} difference from best system for "easy" and "difficult" impostor trials.

5.4. Auditory correlation

Another important topic in forensic speaker comparison is the degree of correlation between human and automatic examinations [13]. In this context we performed a preliminary auditory analysis of a few dozen of the "difficult" trials as defined in Section 5.3. Special attention was paid to the following aspects: 1) the channel mismatch, assessed by means of distortions observed on the long-term spectrum envelope of the recordings, and 2) the examiner's auditory impression of the speaker's speaking style, accent, dialect and vocal effort.

In most of these trials, auditory and automatic results do not correlate. While there were a number of "difficult" trials that phonetic experts, too, had trouble assessing correctly, there were many "difficult" trials in which the speakers' (dis)similarity was easily recognised auditorily, e. g. in terms of voice quality and accent. The most blatant of these automatic errors may sometimes, but not always, have been caused by severe channel mismatch. Phonetic experts therefore seem to have an advantage over automatic systems in trials with prominent vocal or linguistic features, or accompanied by severe channel mismatch. The advantage of automatic systems over phonetic experts could be studied by looking at trials with consistently correct results and high scores across all systems and comparing them to human performance. This could be part of future research.

6. Post Evaluation

Automatic speaker comparison technology should not be regarded as a foolproof stand-alone solution by forensic examiners. To begin with, the system development should be oriented towards forensic scenarios. This is not trivial, given the limitations on "forensic" data available for development. Moreover, our current experiments revealed that even though systems generally seem to perform similarly on a chosen data set, there are differences concerning individual trials. The typical attributes of forensic material sometimes lead to unpredictable results that are not necessarily consistent among the systems investigated. The goal should be to know the effects of certain acoustic parameters on certain system settings and be able to work with a controlled fusion of systems.

As an initial step in this direction, a series of post-evaluations were carried out with ISR systems in order to investigate to what extent characteristics that are specific to the GFS corpus impact on different systems. In particular, we investigated the relevance of background model, channel and utterance length. A better understanding of these phenomena will help us in developing systems more suitable to the forensic reality.

6.1. Background modeling

Mainly due to the NIST campaigns, speech data used for system development is predominantly made up of informal conversations in English. The GFS corpus, however, is made up of German-language recordings and includes emotional speech. We performed some partial experiments in order to investigate the effects of the mismatch between development and application data on the performance of the systems.

Mismatch in SVM-GMM training was evaluated by using two different SVM background models for ISR1: GFS and NIST recordings. In principle, it was observed that performance with GFS was considerably better (DCF of 0.038) than with NIST (DCF of 0.052). We shall see in Section 6.3 that this effect disappears when length compensation is applied.

The relevance of background mismatch was also evaluated within the i-vector framework by employing a score normalisation scheme explicitly dependent on statistics of a set of background conversations [12]. First- and second-order statistics were obtained for both the native (i.e., GFS) and the NIST background models and used in ISR2 and ISR3 systems. The choice of background proved to be irrelevant for the i-vector systems.

6.2. Intersession compensation

In these experiments, ISR1 was used with several NAP configurations. Although completely trained on NIST English data, NAP significantly increased recognition performance. In particular, the two-wire NAP [3] version led to further improvement: using GFS background and no length compensation (see Section 6.3), NAP decreased its DCF from 0.057 to 0.043. Two-wire NAP further improved the DCF to 0.038.

Interestingly, even though two of the EUR systems do not use intersession compensation techniques, they are still comparable to ISR1 performance which uses NAP. This could be due to the fact that EUR systems are trained with non-English recordings also, whereas ISR systems are trained with English recordings only. We therefore expect the incorporation into GMM systems of both intersession compensation techniques and appropriate development data to lead to further improvement.

6.3. Utterance length

The final mismatch factor we investigated concerns utterance length. Utterance length has been extensively evaluated in NIST campaigns and is known to affect speaker recognition performance in several ways. First, varied lengths may bias results concerning speaker identity, and forensic systems should be able to react appropriately. The GFS recordings happen to be homogenous in terms of length, which makes length compensation between trials unnecessary. Second, length inconsistency between development and application may cause systems to work sub-optimally. In this regard, all ISR system components are optimised to NIST evaluation protocols, which are mainly based on about five-minute dialogues for training and testing. In contrast, the GFS corpus contains 60-second training conversations and 30-second testing conversations. Therefore, the primary concern posed by GFS is the shortness of the length of the evaluation segments rather than mismatch in utterance length between development and evaluation segments.

Several score normalisation procedures have been proposed to reduce mismatch bias [14], most of them data-driven. For obvious reasons, this is not a ready solution for forensic applications. We would prefer model-based compensation techniques, which are not dependent on additional data, such as D-norm [14]. In the following experiments, we will use another simple model-domain procedure to compensate for the lack of data in the relatively short test segments found in GFS. This technique was successfully applied in another set-up in which test recordings had to be segmented into very short chunks [15]. This technique pursues a looser GMM model adaptation to cope with scarce data by reducing the relevance factor used in testing MAP adaptation from 16 (used for the longer training conversations) to 0.5. Note that the ISR3 system, based on i-vectors derived from PCA following MAP adaptation is also affected by the proposed length compensation.

Our experiments suggested that length compensation has a great impact on GMM-SVM methodology. Firstly, ISR1 matched or even outperformed the best systems evaluated. Secondly, the gain obtained using native background for training reported in Section 6.1 disappeared. (DCF of 0.026 using GFS background and 0.024 using NIST.) As discussed earlier, this is very important for forensic applications where background speakers in comparable settings are not abundantly available from casework. In contrast, length compensation had only a minor impact on ISR3.

Figure 7 shows the effects of length compensation on ISR1 and ISR3 compared to the best EUR system. The length-compensated systems use NIST background, either for SVM training in ISR1 or for score normalisation in ISR3.

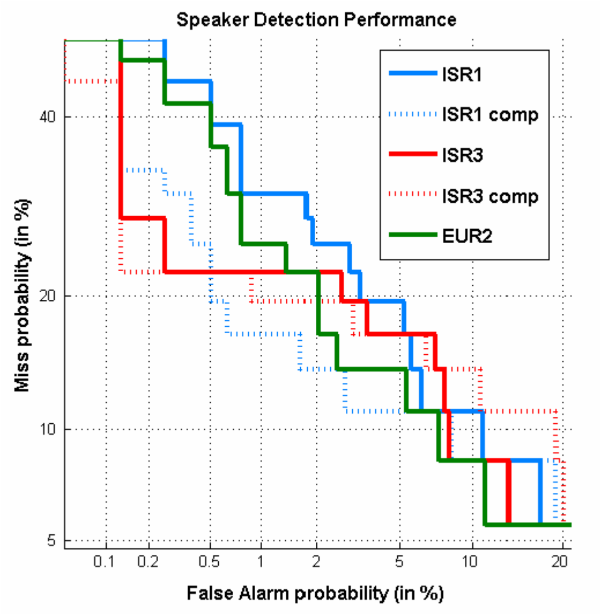

Figure 7: DET plots for length-compensated and non-compensated systems.

7. Conclusions and future work

This paper reports the evaluation of several automatic voice recognition systems on real forensic data. Most of the systems are currently in use by forensic institutions and their performance is analysed under a forensic perspective. Particular attention was paid to systems' errors concerning individual comparisons. In this sense, the systems are often complementary, reacting differently and to some extent unpredictably to different types of interferences while the systems' general performances are comparable. We then applied compensation procedures attempting to minimise data mismatch. We found that length compensation was especially successful for one of the systems, while the others were generally less affected by the compensation techniques.

The results obtained are encouraging, even though they should be interpreted with care due to the limited number of tests. We intend to extend the scope of these evaluations and look for correlations between different automatic recognition methods and forensic auditory analysis. By understanding the limitations of technology in different scenarios, we can increase the role that these systems can play in assisting the forensic examiner.

Acknowledgements

The authors are grateful to Ondrej Glembek for advice concerning i-vector implementation.

8. References

[1] A. Martin and M. Przybocki. The NIST speaker recognition evaluation series, National Institute of Standards and Technology's Web site [Online]. Available: <http://www.nist.gov/speech/tests/sre>.

[2] Van Leeuwen, D.A. and Bouten, J.S., "The NFI/TNO forensic speaker recognition evaluation plan", Available: <http://speech.tm.tno.nl/aso/evalplan-2003.pdf>.

[3] Solewicz Y. and Aronowitz, H., "Two-wire Nuisance Attribute Projection", Proc. of Interspeech, 928-931, 2009.

[4] Dehak, N., Kenny, P., Dehak, R., Dumouchel, P. and Ouellet P., "Front-end factor analysis for speaker verification," IEEE Transactions on Audio, Speech, and Language Processing, 19(4), 788-798, 2011.

[5] Bousquet, P.-M., Matrouf, D. and Bonastre J. F., "Intersession compensation and scoring methods in the i-vectors space for speaker recognition", Proc. of Interspeech, 485-488, 2011.

[6] Becker, T., Jessen, M., Alsbach, S., Broß F. and Meier, T., "SPES: The BKA Forensic Automatic Voice Comparison System", Proc. of Odyssey 2010.

[7] Becker, T., "Automatischer forensischer Stimmenvergleich", University Trier (unpublished), 2011.

[8] Ramos, D., Gonzalez-Rodriguez, J., Gonzalez-Dominguez, J. and Lucena-Molina, J. J., "Addressing database mismatch in forensic speaker recognition with AHUMADA III: a public real-casework database in Spanish", Proc. of Interspeech 2008.

[9] www.agnitio.es

[10] www.loquendo.com

[11] Brümmer, N. and du Preez, J., "Application independent evaluation of speaker detection", Comput. Speech Lang., Vol. 20, 2006, 230–275.

[12] N. Dehak, R. Dehak, J. Glass, D. Reynolds, and P. Kenny, "Cosine similarity scoring without score normalization techniques", Proc. of Odyssey, 2010.

[13] Greenberg, C., Martin, A, Brandschain, L., Campbell, J., Cieri, C., Doddington, G. and Godfrey J., "Human assisted speaker recognition in NIST SRE10", Proc. of Odyssey, 2010.

[14] Bimbot, F., Bonastre, J.-F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-Garcia, J., Petrovska-Delacretaz, D., and Reynolds, D., "A tutorial on text-independent speaker verification", EURASIP Journal on Applied Signal Processing 2004, 4, 430-451.

[15] Solewicz, Y. A. and Aronowitz, H. "Implicit segmentation in two-wire speaker recognition", Proc. of Interspeech, 377-380, 2011.