Measuring Performance in Forensic Automatic Speaker Recognition: VQ, GMM-UBM, i-vectors

Rudolf Haraksim[1] and Andrzej Drygajlo[1]

Abstract: The purpose of this paper is to evaluate the performance of different methods used for forensic automatic speaker recognition (FASR). For this purpose, three different methods were chosen (VQ, GMM and i-vectors). The most recent efforts in automatic speaker recognition (ASR) have resulted in the development of i-vectors, a method based on factor analysis and eigenvoice decomposition. In addition to the i-vectors, two traditionally used text-independent speaker recognition methods, namely Vector Quantization (VQ) and Gaussian Mixture Model – Universal Background Model (GMM-UBM) based are used. Although it has been proven that the VQ and GMM methods perform well in matched recording conditions, their performance is degrading in mismatched conditions and they require the use of additional compensation techniques. The performance of the three methods in both cases (matched and mismatched recording conditions) is evaluated in accordance with the methodological guidelines for best practice in forensic semiautomatic and automatic speaker recognition.

Keywords: Likelihood Ratios, Performance, Measurement, Forensic, FASR, ASR

1 Introduction

The purpose of forensic automatic speaker recognition (FASR) is to provide information of evidential weight if speech in the questioned speaker biometric voice sample (BVS) originates from suspected speaker or not. Best-practice FASR process is associated with the Bayesian interpretation framework and calculation of the strength of evidence (likelihood ratio (LR)). In this process, the feature vectors extracted from the questioned speaker BVS are compared against the feature vectors extracted from the suspected speaker BVS [Dr07, Ro02] using deterministic or statistical methods [Ro06].

For the performance evaluation experiments we have chosen three different representatives of them. Two of these methods, deterministic (VQ) and statistical (GMM) have been shown not to be significantly affected by the length of speech utterances. These two methods usually perform well in the matched recording conditions but for the mismatched training and testing conditions require additional techniques [ABD04, AD12]. The i-vectors method, on the other hand, has the channel compensation functionality embedded. An overview of different approaches evaluating the strength of evidence in forensic speaker recognition (FSR) in the Bayesian interpretation framework

[1] Speech Processing and Biometrics Group, Swiss Federal Institute of Technology Lausanne, Switzerland, haraksim@gmail.com, andrzej.drygajlo@epfl.ch

using likelihood ratios (LR) is presented for example in [Ki09]. While the VQ [Ki09] has the advantage of not requiring many speech utterances, the outcome of this method is a distance based on the k-nearest neighbors and additional procedure is required to compute the LRs from the distance measures. The GMM [CM00] output similarity scores and the i-vectors [De11] produce log likelihood ratios.

Performance evaluation is done using a dataset with a known ground truth by analysing the accuracy and the discriminating power of all methods, as described in [Dr16].

2 Methods used

All three methods chosen (VQ, GMM-UBM and the i-vectors) have one point in common. From the signal processing point of view, the high-dimensional feature vectors (13 Mel Frequency Cepstral Coefficients (MFCCs) and their first and second order derivatives) extracted from the speech signal are transformed into low-dimensional representation , the codebook in case of the VQ, the method parameters in case of the GMM-UBM and the i-vectors, which are used to compare the feature vectors extracted from the questioned speech utterances with speaker models in FASR.

The UBM was created by pooling all the training data using the expectation maximization (EM) algorithm. Given the relatively small size of the training dataset and in order to avoid over/under fitting we have created one 64-mixture GMM from the training dataset. The same amount - 64 codewords - was used in the VQ to create a codebook. The reader may consider using higher mixture dimensionality, our ambition was a fair comparison across the ASR models [Re00]. Further 13 suspected speaker models consisting of 64-mixture GMMs were created and adapted from the UBM using maximum a-posteriori (MAP) estimation.

The i-vectors method was proposed by Dehak et al. [De11]. The i-vector implementation used in our experiments is based on the Microsoft (MSR) Identity Toolbox [SSH13], which is based on the eigenvoice decomposition and factor analysis and uses the probabilistic linear discriminant analysis (PLDA). The process of extraction of i-vectors from the feature vectors includes cepstral mean variance normalization, creation of the GMM-UBM from the training dataset, MAP adaptation of the evaluation dataset from the UBM, learning the total variability subspace [De11], i-vectors extraction, learning PLDA [PE07] and computation of similarity scores using the PLDA model. 64 Gaussian mixtures are used to produce GMMs and UBM prior to i-vectors extraction.using fixed set of parameters. Following parameters are used in the process of constructing i-vectors in our experiments: number of GMM mixtures (64), UBM iterations (10), total variability subspace iterations (10), PLDA iterations (5) and the dimensionality of the eigenvoice subspace (1).

3 Dataset used

A set of French speaking persons, recorded by the Institute de Police Scientifique (IPS) at the University of Lausanne (UNIL) and the Speech Processing and Biometrics Group at the Swiss Federal Institute of Technology Lausanne (EPFL) is used. The dataset contains BVSs of 61 male speakers aged between 18 and 50, where the majority of speakers are aged between 18 and 30. The speech is recorded in two different conditions (Public Switched Telephone Network - PSTN, Global System for Mobile communication - GSM). 17 BVS per individual were recorded covering text dependent as well as spontaneous speech, ranging from 3 to 240 seconds. The dataset was split into the training (48 speakers) and performance evaluation (13 speakers) subsets.

The training dataset, used to compute the UBM parameters in the case of GMM, the VQ codebook and the i-vector parameters (GMM model and T-matrix) contains 699 variable-length (15-45 sec.) speech utterances of 48 speakers. The "reference" subset is used to create suspected speaker models and contains 4 variable-length BVSs (15-45 sec.) per each of the 13 speakers, which are used to compute the codebook (VQ) or the parameters (GMM-UBM / i-vectors) for the suspected speakers. The questioned BVSs "trace" subset contains 11 BVSs per speaker ranging from 15 to 45 seconds. Feature vectors extracted from the questioned BVSs are subsequently used to compute the LRs.

All BVS are subject to voice activity detection (VAD) [MY14], which was suppolied as a part of the VOICEBOX package [Br]. 39-dimensional feature vectors are extracted from each of the BVS [KL10].

3.1 Bayesian interpretation framework

The strength of evidence *E* of the observed BVS is evaluated using two competing hypotheses H_0 and H_1 and background information:

H_0: Suspected speaker is the source of the questioned BVS
H_1: Suspected speaker is not the source of the questioned BVS

It is possible to use alternative propositions, keeping in mind that each change in hypotheses might induce a change in the reference population, recording conditions, etc. The strength of speech evidence is evaluated using the *LR* formula:

$$LR = \frac{P(E \mid H_0)}{P(E \mid H_1)} \qquad (1)$$

where evidence *E* represents features extracted from the questioned BVS. All available implementations of the three methods, including the VQ, which traditionally outputs similarity scores based on distances between the codebook and the feature vectors, are treated to output LRs using a leave-one-out [CT07] linear logistic regression calibration (LLR) [Ra07, chapter 6.5.4]. The left-out score plays the role of the evidence and the rest of the LRs depending on the source of origin of the questioned BVS and the suspected speaker model are used for calculating the parameters of the LLR.

4 Performance evaluation

As proposed in [Dr16, chapter 4.2] the performance is measured in terms of accuracy, discriminating power and calibration. The performance is presented using Tippett plot [ME01] and the associated accuracy metric of Probability of Misleading Evidence (PME). The Proportions Trade-off curves with the accociated metric of Equal Proportion Probability (EPP) measure the discriminating power. The Empirical Cross-Entropy (ECE) [RG13] plots with the associated metric of log likelihood ratio cost [BP06] measure the accuracy (Cllr), discriminating power ($\mathrm{Cllr}^{\mathrm{min}}$) and calibration ($\mathrm{Cllr}^{\mathrm{cal}}$).

For the matched conditions the performance of the three methods is evaluated using two different recording conditions (GSM and PSTN), as presented in Section 3. Under the matched conditions there is no difference in the the training dataset and evaluation dataset recording conditions. In the mismatched conditions the recording condition of the questioned speaker BVSs is different from the recording condition of the suspected speaker BVS (e.g., questioned BVSs are recorded in the PSTN condition and the suspected speaker BVSs are in the GSM condition and vice versa). The training dataset contains speaker BVSs in both conditions (combination of GSM and PSTN BVSs).

5 Experimental results

Similar parameterization conditions are used across the three methods under evaluation to ensure a fair comparison of the methods (64 Gaussian mixtures, 64-word codebook).

5.1 Matched conditions

In the matched conditions the training and the evaluation dataset are recorded with the PSTN or GSM. Figures 1 and 2 show the Proportions Trade-off curves, Tippett plots and the ECE plots for two matched conditions and all three methods. Proportions trade-off curves (Figure 1 – left) show, that the best discriminating power measured by Equal Proportion Probability (EPP), which is found at the intersection with the proportions trade-off curves and the main diagonal, was observed for the GMM-UBM method for the PSTN (fixed line) dataset (EPP = 0.0145), while the worst EPP was observed for the VQ for the GSM dataset (EPP = 0.0985%). On the Tippett plots (Figure 1 – right) we measure the performance of the methods using the PMEs, which are found at the inverse cumulative distribution functions for logLR = 0. Here lowest PMEs are observed for the GMM-UBM method and the PSTN (fixed line) dataset ($\mathrm{PME}_{\mathrm{H0}}$=0.014, $\mathrm{PME}_{\mathrm{H1}}$=0.03), while the highest PMEs are observed for the VQ for the GSM dataset ($\mathrm{PME}_{\mathrm{H0}}$= 0.099,

$\mathrm{PME}_{\mathrm{H1}}$ = 0.11).

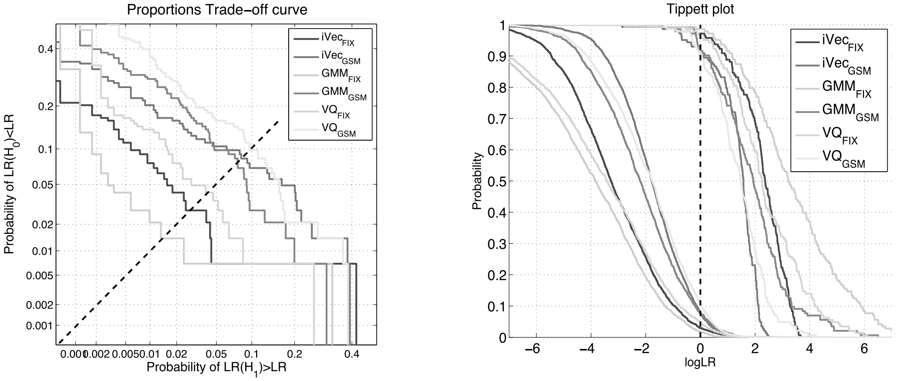

Fig. 1: Proportion trade-off curves on the left, Tippett plots on the right

Figure 2 shows the graphical representations of ECE in which we can explore the Cllr values, Cllr at the intersection of the red-solid curve and the dashed line corresponding to Prior($\log_{10}$odds) = 0 and Cllr^{min} at the intersection of the blue dashed curve and the dashed line corresponding to Prior($\log_{10}$odds) = 0.

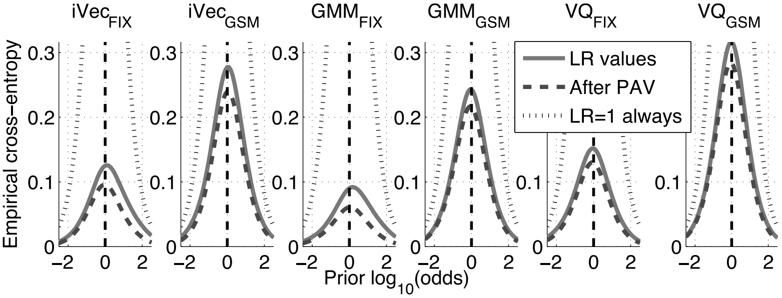

Fig. 2: Empirical Cross-Entropy plots

Best discriminating power measured by the Cllr^{min} and the best accuracy measured by the Cllr is achieved by the GMM-UBM method on the PSTN (fixed line) dataset (Cllr^{min} = 0.06, Cllr = 0.084), while the worst discriminating power and the worst accuracy is achieved by the VQ method on the GSM dataset (Cllr^{min} = 0.279, Cllr = 0.315). Detailed results are presented in Table 1.

The LR values produced by the three methods were subjected to the leave-one-out linear logistic regression calibration. The calibration measured by the Cllr^{cal} (Cllr^{cal} = Cllr - Cllr^{min}) indicates the best calibration for the VQ for the PSTN dataset and the worst calibration for the i-vectors on the GSM dataset.

	EPP	PME_{H0}	PME_{H1}	*Cllr*	$Cllr^{min}$	$Cllr^{cal}$
GMM_PSTN	**0.0145**	**0.014**	**0.018**	**0.084**	**0.06**	0.024
GMM_GSM	0.077	0.07	0.077	0.242	0.22	0.022
iVec_PSTN	0.0285	0.028	0.03	0.124	0.095	0.029
iVec_GSM	0.0845	0.098	0.077	0.275	0.238	**0.037**
VQ_PSTN	0.0425	0.035	0.048	0.151	0.13	**0.021**
VQ_GSM	**0.0985**	**0.098**	**0.107**	**0.315**	**0.279**	0.036

Tab. 1: Experimental results obtained by the three methods on the PSTN and GSM datasets

5.2 Mismatched conditions

In the mismatched conditions setting we assume that the recording condition of the questioned speaker BVSs is different from the recording condition of the suspected speaker. The questioned speaker recorded in the PSTN condition and the suspected speaker recorded in the GSM condition is labelled as *C1*, while the questioned speaker recorded in the GSM condition and the suspected speaker recorded in the PSTN is labelled as *C2*. We further assume that the training dataset contains both recording conditions (e.g., GSM and PSTN). Figures 3 and 4 show the Proportions Trade-off curves, Tippett plots and the ECE plots for the mismatched conditions.

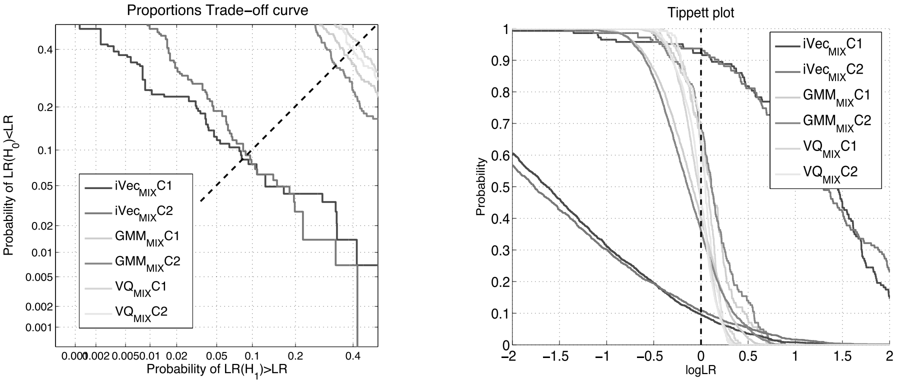

Fig. 3: Proportion trade-off curves on the left, Tippett plots on the right

The proportions trade-off curves (Figure 3 – left) indicate, that the best discriminating power measured by the EPP was achieved by the i-vectors (EPP_C1 = 0.0845, EPP_C2 = 0.0915). The EPP measured for the VQ and GMM-UBM in both mismatched conditions would likely improve by using channel compensation techniques [ABD04, AD12]. The discriminating power for the VQ and GMM-UBM methods resulted in EPP > 0.3 for both mismatched conditions.

For easier interpretation, the logLR scale of the Tippett plots (Figure 3 right) was modified in Figure 4. Due to the absent channel compensation, the GMM-UBM and the VQ do not extract as much evidential information from the speech fragments as the i-vectors. For the i-vectors we observe the lowest PMEs (PME_{H0}> 0.063, PME_{H1} < 0.109) while the highest PMEs are produced by the VQ (PME_{H0}> 0.36, PME_{H1} < 0.46).

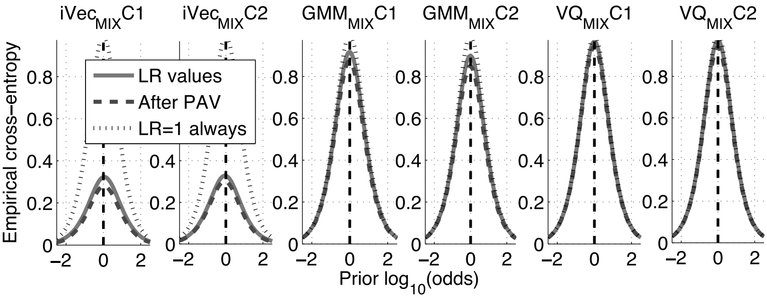

Fig. 4: Empirical Cross-Entropy plots (different scales than in Figure 2)

The best discriminating power and accuracy, as presented in the ECE plots (Figure 4) are obtained by the i-vectors method for both mismatched conditions ($Cllr^{min}$ < 0.3, Cllr < 0.322). The VQ and the GMM-UBM result in accuracy (Cllr > 0.9) and discriminating power ($Cllr^{min}$>0.85). The summary of the results is presented below in Table 2.

	EPP	*PME_{H0}*	*PME_{H1}*	*Cllr*	*$Cllr^{min}$*	*$Cllr^{cal}$*
GMM_C1	0.378	0.343	0.4	0.915	0.875	0.04
GMM_C2	0.336	0.32	0.36	0.9	0.85	0.05
iVec_C1	**0.0845**	**0.077**	**0.096**	**0.32**	**0.285**	**0.035**
iVec_C2	**0.0915**	**0.063**	**0.109**	**0.322**	**0.3**	**0.022**
VQ_C1	0.42	0.36	0.46	0.965	0.94	0.025
VQ_C2	0.399	0.32	0.46	0.97	0.937	0.033

Tab. 2: Evaluation results obtained by three methods in the C1 and C2 conditions

6 Conclusions

In the matched recording conditions all of the three evaluated methods obtained comparable results and merit their place in the FASR, keeping in mind that different settings (e.g., larger code-books, different training parameters) would have produced different, potentially better results. The question of: *"Which one should be used in a particular case?"* we leave to the discretion of the forensic expert.

From the results obtained in the mismatched conditions in terms of discriminating power measured by the EPP and $Cllr^{min}$, accuracy measured by the Cllr and calibration measured by the $Cllr^{cal}$ we can conclude, that the i-vectors method performs within boundaries comparable to its performance in the matching condition on the GSM dataset. The results presented in [ABD04, AD12] indicate, that the use of channel compensation techniques would improve the results of the GMM-UBM.

Acknowledgements

The work presented was supported by the Grant C12.0117 from the Swiss State Secretariat for Education, Research and Innovation entitled "Biometric Evidence in Forensic Automatic Speaker Recognition" in the framework of COST IC 1106 Action.

References

[ABD04] Alexander A.; Botti F.; Drygajlo A.: Handling mismatch in corpus-based forensic speaker recognition, Proceedings of ODYSSEY 2004 (Toledo), pp. 69-74, 2004

[AD12] Alonso-Moreno V.; Drygajlo A.: A joint factor analysis model for handling mismatched recording conditions in forensic automatic speaker recognition, Proceedings of the International Conference on Biometrics (ICB 2012) , New Delhi, pp. 484-489, 2012

[BP06] Brümmer N.; du Preez J.: Application-independent evaluation of speaker detection, Computer Speech & Language, Vol. 20, Issue 2-3, pp. 230-275, April-July 2006

[Br] Brookes M., VOICEBOX: Speech processing toolbox for Matlab, internet source: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[CM00] Champod C.; Meuwly D.: The inference of identity in forensic speaker identification, Speech Commun 31(2–3):193–203, 2000

[CT07] Cawley G.C.; Talbot N.L.C., Preventing over-fitting in model selection via Bayesian regularization of the hyperparameters, Journal of Machine Learning Research, 8: p. 841-861, 2007

[De11] Dehak N. et.al.: Front-end factor analysis for speaker verification, IEEE TASLP, vol. 19, pp. 788-798, May 2011

[Dr16] Drygajlo A. et.al.: Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition, Verlag für Polizeiwissenschaft, 2015

[Dr07] Drygajlo A.: Forensic automatic speaker recognition, IEEE Signal Process Mag 24(2):132–135, 2007

[Ki09] Kinnunen T. et.al: Comparative evaluation of maximum a posteriori Vector Quantization and Gaussian Mixture Models in speaker verification, Pattern Recognition Letters, Vol. 30, n. 4, pp. 341-347, 2009

[KL10] Kinnunen T.; Li H.: An overview of text-independent speaker recognition: From features to supervectors, Speech Communication 52: 12-40, 2010

[Me01] Meuwly D., Reconnaissance de Locuteurs en Sciences Forensiques: L'apport d'une approche automatique, PH1 thesis, 2001

[MY14] Mak M.W.; Yu H.B: A study of voice activity detection techniques for NIST speaker recognition evaluations, Computer Speech and Language 28: 295-313, 2014

[PE07] Prince S.J.D; Elder J.H: Probabilistic linear discriminant analysis for inferences about identity, in Proc. IEEE ICCV, Rio de Janeiro, Brazil, Oct. 2007

[Ra07] Ramos D., Forensic Evaluation of the Evidence using Automatic Speaker Recognition Systems, PhD thesis, Madrid, 2007

[Re00] Reynolds D.A. et.al.: Speaker Verification Using Adapted Gaussian Mixture Models, Digital Signal Processing 10, pp. 19–41, 2000

[RG13] Ramos D.; Gonzalez-Rodriguez J.: Reliable Support: Measuring Calibration of Likelihood Ratios, Forensic Science International, Vol. 230, pp. 156-169, May 2013

[Ro02] Rose P.: Forensic speaker identification, Taylor and Francis, London, 2002

[Ro06] Rose P.: Technical forensic speaker recognition: evaluation, types and testing of evidence, Computer Speech Lang 20(2–3):159–191, 2006

[SSH13] Sadjadi S.O.; Slaney M.; Heck L.: MSR Identity Toolbox v1.0: A MATLAB Toolbox for Speaker-Recognition Research, in Speech and Language Processing Technical Committee Newsletter, IEEE, 2013