Homogeneity Measure for Forensic Voice Comparison: A Step Forward Reliability

Moez Ajili[1,2(✉)], Jean-François Bonastre[1], Solange Rossato[2], Juliette Kahn[3], and Itshak Lapidot[4]

[1] Laboratoire Informatique d'Avignon (LIA), University of Avignon, Avignon, France
[2] Laboratoire Informatique de Grenoble (LIG), University of Grenoble, Grenoble, France
[3] Laboratoire National de Métrologie et d'Essais (LNE), Paris, France
[4] Afeka Center for Language Processing (ACLP), Tel Aviv, Israel

Abstract. In forensic voice comparison, it is strongly recommended to follow the Bayesian paradigm to present a forensic evidence to the court. In this paradigm, the strength of the forensic evidence is summarized by a *likelihood ratio* (LR). But in the real world, to base only on the LR without looking to its degree of reliability does not allow experts to have a good judgement. This work is mainly motivated by the need to quantify this reliability. In this concept, we think that the presence of speaker specific information and its homogeneity between the two signals to compare should be evaluated. This paper is dedicated to the latter, the homogeneity. We propose an information theory based homogeneity measure which determines whether a voice comparison is feasible or not.

Keywords: Forensic voice comparison · Reliability · Homogeneity · Speaker recognition

1 Introduction

In forensic comparison, it is strongly recommended to present the forensic evidence to the court following the Bayesian paradigm [1]: *Speaker recognition* (SR) systems should calculate for a given trial a *likelihood ratio* (LR) which represents the degree of support for the prosecutor hypothesis (the two speech extracts are pronounced by the same speaker) rather than the defender hypothesis (the two speech extracts are pronounced by different speakers). Theoretically, a good LR is assumed to contain by itself all the needed information including reliability: in good conditions, the LR should be far from 1 to support comfortably one of the two hypothesis (big LR values, about 10^{10} support H_0 and low values, about 10^{-10} support H_1) while in bad conditions the LR is close to one and consequently, it does not allow a good discrimination between the two hypothesis. But in the real world, forensic processes are working only with an empirical estimation of LRs that could be far from theoretical ones. In this case, LRs are unable to embed reliability information furthermore as there is no concrete evaluation of the disagreement between theoretical and empirical LR. It is particularly true for SR systems which are working as black boxes: They calculate

A. Pardo and J. Kittler (Eds.): CIARP 2015, LNCS 9423, pp. 135–142, 2015.
DOI: 10.1007/978-3-319-25751-8_17

a sort of score in all situations without verifying if there is enough reliable information present in the two records. Then, those scores will be calibrated (i.e. normalized) to be viewed as a LR [2][3]. So, it could be misleading to the court if experts report only the LR and not its degree of reliability. Presently, these issues of validity and reliability are of great concern in forensic science [4] [5] [6] [7] [8] [9]. To cope with this problem, it is interesting to define a *confidence measure* (CM) that indicates the reliability of a system output. Several solutions were proposed in [6] [7] [8] [9][10] [11] where the CM is estimated for each trial from both system decision score and the two speech extracts of a given voice comparison, S_A-S_B. An alternative consists in studying the losses in LR quality which are related to: (i) A lack of discriminative information (as shown in [12]) in S_A and/or S_B. (ii) Sufficient discriminative information are available but the system is unable to output a meaningful (LR) due for example to the mismatch between elements used to build the system (UBM, total variability matrix, PLDA,...) and the pair of voice records S_A-S_B [13][14]. In brief, the loss could be divided into two origins. Our interest concerns the case (i) detailed before.

The final objective of our work is to define a "*Feasibility measure*" (FM) able to measure the presence of speaker discriminant cues and the homogeneity of this information between the pair of voice records S_A-S_B. So, this measure is estimated only from the two in-interest voice records. If it is obvious that the presence of speaker specific information inside S_A and S_B is mandatory, it is not sufficient: examples tied with the same class of cues should be included in both speech recordings in order to be useful.

In this paper, we address more specifically the problem of the evaluation of the homogeneity of two speech signals in terms of information classes, at the acoustic level. We propose an information theory-based homogeneity criterion able to quantify this homogeneity.

This paper is structured as follows. Section 2 presents our new homogeneity measure and details the algorithm to compute it. Section 3 describes the LIA baseline system and presents experiments and results. Then, section 4 presents the conclusion and proposes some extends of the current work.

2 Information Theory Based Homogeneity Measure

In this section, we define an information theory (IT) based homogeneity measure denoted HM(). Its objective is to calculate the amount of acoustic information that appertains to the same class between the two voice records.The set of acoustic frames gathered from the two files S_A and S_B is decomposed into acoustic classes thanks to a Gaussian Mixture Model (GMM) clustering. Then the homogeneity is first estimated in terms of bits as the amount of information embedded by the respective "number of acoustic frames" of S_A and S_B linked to a given acoustic class. Each acoustic class is represented by the corresponding Gaussian component of the GMM model. The occupation vector could be seen as the number of acoustic frames of a given recording belonging to each class m. It is noted: $[\gamma_{g_m}(s)]_{m=1}^{M}$.

Given a Gaussian g_m and two posterior probability vectors of the two voice records S_A and S_B, $[\gamma_{g_m}(A)]_{m=1}^{M}$ and $[\gamma_{g_m}(B)]_{m=1}^{M}$, we define also:

- $\chi_A \cup \chi_B = \{x_{1A}, ..., x_{NA}\} \cup \{x_{1B}, ..., x_{NB}\}$ the full data set of S_A and S_B with cardinality N=N_A +N_B
- $\gamma(m)$ and $\omega(m)$ are respectively the occupation and the prior of Gaussian m where $\omega(m) = \frac{\gamma(m)}{\sum_{k=1}^{M}\gamma(k)} = \frac{\gamma(m)}{N}$
- $\gamma_A(m)$ (respectively $\gamma_B(m)$) is the partial occupations of the m^{th} component due to the voice records S_A (respectively S_B).
- p_m is the probability of the Bernoulli distribution of the m^{th} bit (due to the m^{th} component), B(p_m). p_m=$\frac{\gamma_A(m)}{\gamma(m)}$, $\overline{p}_m = 1 - p_m = \frac{\gamma_B(m)}{\gamma(m)}$.
- $H(p_m)$ the entropy of the m^{th} Gaussian (the unit is bits) given by: $H(p_m) = -p_m log_2(p_m) - \overline{p}_m log_2(\overline{p}_m)$.

The class entropy, H(p_m), has some interesting properties in the context of an homogeneity measure:

* $H(p_m)$ belongs to $[0, 1]$.
* $H(p_m) = 0$ if $p_m = 0$ or $p_m = 1$. It means that when the repartition of the example of a given class m is completely unbalanced between S_A and S_B, $H(p_m)$ is zero (i.e. $H(p_m)$ goes to zero when p_m is close to 0 or 1).
* $H(p_m) = 1$ when $p_m = 0.5$. $H(p_m)$ is maximal when the examples belonging to a given class are perfectly balanced between between S_A and S_B (i.e. $H(p_m)$ goes to the maximum value 1 when the repartition goes to the balanced one).

With these theoretical properties, $H(p_m)$ is definitively a good candidate in order to build a homogeneity measure. Two measures based on $H(p_m)$ are proposed hereafter. The first measure is a normalized version. It ignores the size of the frame sets (i.e. the duration of the recordings) when the second ones 'non-normalized' takes this aspect into account.

The normalized HM denoted "HM_{BEE}" is calculated as shown in Equation 1. It measures the Bit Entropy Expectation (BEE) with respect to the multinomial distribution defined by GMM's priors $\{\omega(m)\}_{i=1}^{M}$.

$$HM_{BEE} = \sum_{m=1}^{M} \frac{\gamma(m)}{N} H(p_m) = \sum_{m=1}^{M} \omega_m H(p_m) \tag{1}$$

By definition HM_{BEE} contains the percentage of the data-homogeneity between S_A and S_B. It does not take into account the quantity of the homogeneous information between the two speech extracts. To integrate this information, a *Non-normalized Homogeneity Measure* (NHM) is proposed. NHM calculates the quantity of homogeneous information between the two voice records as shown in Equation 2. The amount of information is defined in term of number of acoustic frames. NHM measures the BEE with respect of the quantity of information present in each acoustic class $\{\gamma(m)\}_{i=1}^{M}$.

$$NHM_{BEE} = \sum_{m=1}^{M} (\gamma_A(m) + \gamma_B(m))H(p_m) = \sum_{m=1}^{M} \gamma(m)H(p_m) \quad (2)$$

As mentioned before, a GMM presenting the different acoustic classes is mandatory to estimate both homogeneity measure. So, it will be reasonable to estimate HM using different representation of the acoustic space. Several avenues are explored in this paper. First, we use a GMM trained only on the two speech signals. The major advantage of this representation is its independence toward the system. Nevertheless, the amount of data involved in the two signals is not always quite sufficient to build stable acoustic classes. An alternative consists in to use a stable representation of the acoustic space, UBM. As it is learnt on a very large data set and its high ability to model the whole acoustic space, the estimation quality of the UBM could be higher than the GMM A-B (learnt only on the two speech recordings).

3 Experiments and Results

In order to evaluate the homogeneity measures presented in section 2, we propose several experiments based on NIST SRE framework.

3.1 Baseline LIA System

In all experiments, we use as baseline the LIA_SpkDet system presented in [15]. This system is developed using the ALIZE/SpkDet open-source toolkit [16]. It uses I-vector approach [17].

Acoustic features are composed of 19 MFCC parameters, its derivatives, and 11 second order derivatives (the frequency window is restricted to 300-3400 Hz). A normalization file-based process is applied, so that the distribution of each cepstral coefficient is 0-mean and 1-variance for a given utterance.

The *Universal Background Model* (UBM) is trained on Fisher database on about 10 millions of speech frames. It has 512 components whose variance parameters are floored to 50% of the global variance (0.5). The total variability matrix T is trained using 15660 sessions from 1147 speakers (using NIST_SRE 2004, 2005, 2006 and Switchboard data). Speaker models are derived by Bayesian adaptation of the Gaussian component means, with a relevance factor of 14. The same database is used to estimate the inter-session matrix W in the I-vector space. The dimension of the I-Vectors in the total factor space is 400.

For scoring, PLDA scoring model [18] is applied. The speaker verification score given two I-vectors w_A and w_B is the likelihood ratio described by:

$$score = log \frac{P(w_A, w_B|H_p)}{P(w_A, w_B|H_d)} \quad (3)$$

where the hypothesis H_p states that inputs w_A and w_B are from the same speaker and the hypothesis H_d states they are from different speakers.

3.2 Experimental Protocol

All the experiments presented in this work are performed based upon the NIST-SRE 2008 campaign, all trials (det 1), "short2-short3", restricted to male speakers only (referred to as 2008 protocol). This protocol is composed by 39433 tests (8290 target tests, the rest are impostor trials). The utterances contain 2.5 minutes of speech in average.

As seen in section 2, for each trial the set of acoustic frames is clustered thanks to a GMM. This GMM has 512 components and is trained by EM/ML (with a variance flooring ≈ 0).

3.3 Evaluation Process of the Homogeneity Measure

In order to evaluate the proposed homogeneity measures, we apply it on all the trials of our evaluation set and sort the set accordingly. We are expecting that lowest values of homogeneity are correlated with the lowest performance of the speaker recognition system, as well as the opposite behaviour for high values. To compute the speaker recognition performance, we select the *log-likelihood-ratio cost* (C_{llr}), largely used in forensic voice comparison because it is based on likelihood ratios and not on hard decisions like, for example, *equal error rate* (EER) [6,19]. C_{llr} has the meaning of a cost or a loss: lower the C_{llr} is, better is the performance. In order to withdraw the impact of calibration mistakes, we use the minimum value of the C_{llr}, noted C_{llr}^{min}. If a C_{llr} could be computed for a given trial, it makes sense to average the values on a reasonably large set of trials. So, we apply a 1500 trials sliding window, with a step of 1000, on the trials sorted by homogeneity values. On each window, we compute the averaged C_{llr}^{min} to be compared with the HM value (computed here as the median value on the window). To work on such number of trials allows also to compute the percentage of *false rejection* (FR) and *false acceptance* (FA). FR and FA are computed using a threshold estimated onto the whole test set and tuned to correspond at the EER. The C_{llr}^{min} baseline system computed on all trials is equal to 0.2241.

3.4 Evaluation of Homogeneity Measures

- GMM A-B. In this subsection, we use a GMM learnt on the pair of speech signals (GMM A-B). From Figure 1, it can be seen that HM_{BEE} value does not have a remarkable impact on C_{llr}^{min}. It is confirmed by a not significant low correlation with C_{llr}^{min} , evaluated to a R^2 equal to -0.39 (p=0.16). It seems that to focus only on BEE, ignoring the involved quantity of examples does not allow to build an homogeneity measure with the desired characteristics.

Experimental results obtained using NHM_{BEE} are reported in Figure 2. The shape of the curve is interesting with C_{llr}^{min} varying from 0.309 to 0.122, indicating a high correlation between NHM_{BEE} and C_{llr}^{min} (R^2 = -0.942, p <0.01). Moreover, it seems that NHM_{BEE} brings new information compared to the system outputs. The result is confirmed with a lower R^2 of 0.55 (to be compared with a R^2 equal to 0.73 in the case of HM_{BEE}). Further experiments have been done using NHM_{BEE} only.

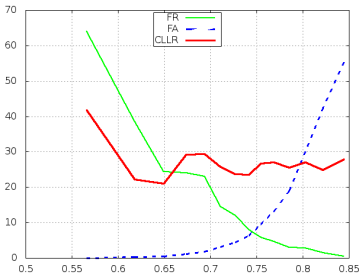

Fig. 1. HM_{BEE} behaviour, estimated with GMM-AB.

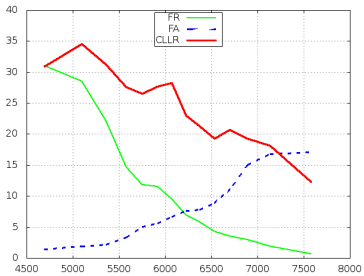

Fig. 2. NHM_{BEE} behaviour, estimated with GMM-AB.

- UBM Model. In Figure 3, we present the results obtained using NHM_{BEE} like the previous one but here, we use directly the UBM in order to cluster the acoustic frames of the pair of speech recordings. With a C_{llr}^{min} varying between 0.3 and 0.089 and its high correlation with NHM_{BEE}, evaluated to R^2 equal to -0.950 ($p < 0.01$), this variant seems to outperform the previous one.

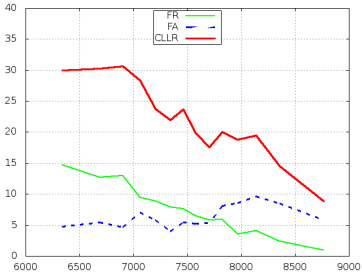

Fig. 3. NHM_{BEE} behaviour, estimated with UBM.

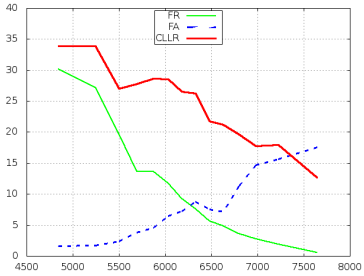

Fig. 4. NHM behaviour using GMM A-B initialized with UBM.

Two more experiences are realized. In Figure 4, we report results when the GMM A-B is now initialized with the UBM (case A), and in Figure 5, we use the UBM mean-adapted (using MAP) by the two speech recordings S_A and S_B (case B). In both cases, NHM is highly correlated with the SR system performance, C_{llr}^{min} (A: R^2 = -0.963, $p < 0.01$; B: R^2 = -0.973, $p < 0.01$). Moreover, it seems to be more dependent to the system output compared to the previous one in which we use only the UBM (case A: $R^2 = 0.57$, p<0.01; case B: $R^2 = 0.37$, p<0.01; UBM $R^2 = 0.29$, p<0.01). We notice that using the mean-adapted UBM model to estimate NHM_{BEE} adds more stability to C_{llr}^{min} variation. It can be explained by the fact that adapted UBM model preserves the good modeling of the whole acoustic space and at the same time, takes into account the characteristics of a given trial. Whereas in case A, it is clear that using a GMM A-B with UBM

initialization is very close to the case in which we use GMM A-B (without initialization). This result is clear when we see the big similarity between the two NHM behavioural curve (Figure 4 and 2).

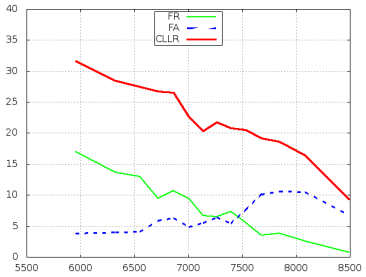

Fig. 5. NHM behaviour estimated using the UBM adapted by S_A and S_B.

4 Discussion and Conclusions

In this paper, we have proposed an IT-based data Homogeneity Measure denoted NHM_{BEE} where the quantity of homogeneous examples presented in both speech extracts is taken into account. NHM_{BEE} belongs to the Bit Entropy Expectation (BEE) computed on a Gaussian Mixture Model view of the couple of speech recordings which composes a given voice comparison trial. A first variant from this measure uses GMM as a model trained by the pair of recordings. It showed interesting properties with a nice relation between the homogeneity values and the C_{llr}^{min}, varying from (HM=4689,C_{llr}^{min}=0.309) to (HM=7579, C_{llr}^{min}=0.1227). A second variant of NHM_{BEE} uses directly the UBM model in order to cluster the pair of speech recordings (without training or adaptation of the UBM). This version has a similar behaviour than the previous one but outperformed it with a behavioural curve moving from (HM=6341, C_{llr}^{min}=0.3) to (HM=8762, C_{llr}^{min}=0.089). In the same direction, the use of a UBM mean-adapted by the pair of recordings adds more stability to the C_{llr}^{min}, varying quite consistently from (HM=5953,C_{llr}^{min}=0.31) to (HM=8490, C_{llr}^{min}=0.09). This result shows that the way to cluster the pair of speech recordings is important. Moreover, the different variant of NHM_{BEE} showed a low correlation with the scores issued by the speaker recognition system. The behavioural curves of NHM_{BEE} and this low correlation encourage us strongly to conclude that NHM_{BEE} is a good candidate in order to measure the data homogeneity between a pair of speech recordings, in the view of voice comparison reliability.

This work will firstly extended by working on other representation of acoustic classes in order to estimate NHM_{BEE}. In addition to this point, the behaviour of our measures depending on the session variability factors should be explored more deeply. Finally, as expressed in the introduction, data homogeneity is a mandatory first step for a voice comparison feasibility measure and we expect to explore this new avenue.

References

1. Champod, C., Meuwly, D.: The inference of identity in forensic speaker recognition. Speech Communication, 193–203 (2000)
2. Brummer, N., van Leeuwen, D.A.: On calibration of language recognition scores. In: Speaker and Language Recognition Workshop, pp. 1–8. IEEE Odyssey (2006)
3. Brummer, N., Doddington, G.: Likelihood-ratio calibration using prior-weighted proper scoring rules (2013). arXiv preprint arXiv:1307.7981
4. Bonastre, J.F., Bimbot, F., Boë, L.J., Campbell, J.P., Reynolds, D.A., Magrin-Chagnolleau, I.: Person authentication by voice: a need for caution. In: INTERSPEECH (2003)
5. Campbell, J.P., Shen, W., Campbell, W.M., Schwartz, R., Bonastre, J.-F., Matrouf, D.: Forensic speaker recognition. Institute of Electrical and Electronics Engineers (2009)
6. Morrison, G.S.: Forensic voice comparison and the paradigm shift. Science & Justice, 298–308 (2009)
7. Rose, P.: Technical forensic speaker recognition: Evaluation, types and testing of evidence. Computer Speech & Language, 159–191 (2006)
8. Morrison, G.S., Zhang, C., Rose, P.: An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system. Forensic science international, 59–65 (2011)
9. Morrison, G.S.: Measuring the validity and reliability of forensic likelihood-ratio systems. Science & Justice, 91–98 (2011)
10. Campbell, W.M., Reynolds, D.A., Campbell, J.P., Brady, K.: Estimating and evaluating confidence for forensic speaker recognition. In: ICASSP, pp. 717–720 (2005)
11. Mengusoglu, E., Leich, H.: Confidence Measures for Speech/Speaker Recognition and Applications on Turkish LVCSR. PhD Faculte Polytechnique de Mons (2004)
12. Rao, W., Mak, M.W.: Boosting the performance of i-vector based speaker verification via utterance partitioning. IEEE Transactions on Audio, Speech and Language Processing, 1012–1022 (2013)
13. Greenberg, C.S., Stanford, V.M., Martin, A.F., Yadagiri, M., Doddington, G.R., Godfrey, J.J., Hernandez-Cordero, J.: The 2012 NIST speaker recognition evaluation. In: INTERSPEECH, pp. 1971–1975 (2013)
14. Kahn, J., Audibert, N., Rossato, S., Bonastre, J.F.: Intra-speaker variability effects on speaker verification performance. In: Odyssey, p. 21 (2010)
15. Matrouf, D., Scheffer, N., Fauve, B.G., Bonastre, J.F.: A straightforward and efficient implementation of the factor analysis model for speaker verification. In: INTERSPEECH, pp. 1242–1245 (2007)
16. Larcher, A., Bonastre, J.F., Fauve, B.G., Lee, K.A., Lévy, C., Li, H., Parfait, J.Y.: ALIZE 3.0-open source toolkit for state-of-the-art speaker recognition. In: INTERSPEECH, pp. 2768–2772 (2013)
17. Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P.: Front-end factor analysis for speaker verification. IEEE Transactions on Audio, Speech, and Language Processing, 788–798 (2011)
18. Prince, S.J.D., Elder, J.H.: Probabilistic linear discriminant analysis for inferences about identity. IEEE 11th International Conference on Computer Vision, ICCV (2007)
19. Brummer, N., du Preez, J.: Application-independent evaluation of speaker detection. Computer Speech and Language, 230–275 (2006)