# INTER-SPEAKER VARIABILITY IN FORENSIC VOICE COMPARISON: A PRELIMINARY EVALUATION

*Moez Ajili*[1], *Jean-françois Bonastre*[1], *Solange Rossato*[2] *and Juliette Kahn*[3]

[1] University of Avignon, Laboratoire Informatique d'Avignon (LIA), France
[2] University of Grenoble, Laboratoire Informatique de Grenoble (LIG), France
[3] Laboratoire National de métrologie et d'Essais (LNE), France

## ABSTRACT

In forensic voice comparison, it is strongly recommended to follow Bayesian paradigm. In this paradigm, the strength of the forensic evidence is summarized by a *likelihood ratio* ($LR$). The $LR$ magnitude quantifies the strength of the evidence: far from unity for a meaningful $LR$ (a $LR$ which supports strongly one of the hypothesis); close to unity when the evidence is next to useless. Despite this nice theoretical aspect, the $LR$ does not embed the reliability of its estimation process itself. And, in various cases, a lack in reliability inside the estimation process is able to destroy the reliability of the resulting $LR$. It is particularly true when voice comparison is considered, as Speaker Recognition ($SR$) systems are outputting a score in all situations regardless of the case specific conditions. Furthermore, $SR$ systems use different normalization steps to see their scores as $LR$ and these normalization steps are clearly a potential source of bias. Consequently, a complete view of reliability should be taken into account for forensic voice comparison. This article focuses on one part of this question, the "speaker factor", the characteristics and the behaviors of the two speakers involved in a voice comparison trial.

***Index Terms*—** forensic voice comparison, inter-speaker variability, speaker profile, speaker recognition.

## 1. INTRODUCTION

In the past two decades, *speaker recognition* ($SR$) systems have achieved significant progresses. Particularly, $SR$ systems based on I-vector [1] have reached impressive low error rates in realistic conditions ($\approx 1\%$). The robustness of the evaluation for a given test condition comes by the fact of using a large number of voice comparison samples. The number of samples per speaker as well as the characteristics of the speakers themselves are not taken into account. This global nature of evaluation ignores many other important factors that could be on a huge impact on the recognition strength and thereby made $SR$ system not only weak but also meaningless in some forensic cases where every trial represents a specific situation that should be analyzed carefully and independently of the other trials. In forensic voice comparison, it is strongly recommended to present the forensic technical report to the court following the Bayesian paradigm [2, 3, 4]. 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 excerpts are pronounced by the same speaker) rather than the defender hypothesis (the two speech excerpts are pronounced by two different speakers).

By definition, a $LR$ is assumed to synthesize the conclusion of a voice comparison forensic technical report, for the discrimination between the two hypothesis as well as for reliability's aspects. The $LR$ magnitude quantifies the strength of the evidence: far from unity for a meaningful $LR$ (a $LR$ which supports strongly one of the two hypothesis); close to unity when the evidence is next to useless. But in the real world, a $LR$ is only approximated using an extraction process. The reliability of this process should be taken into account by the experts and sometime questioned. It is particularly true when there is a: (i) insufficient quantity of information inside both voice records; (ii) bad quality of speaker specific information in the trial [5, 6]; (iii) insufficient homogeneity of the information between both records [7]. Presently, these issues of validity and reliability are of great concern in forensic science [5, 6, 8, 9, 10, 11, 12, 13]. Several solutions were proposed in [12, 13, 6, 5, 14, 15, 16] where reliability is estimated for each trial from both system decision output and the two speech extracts of a given voice comparison, S$_A$-S$_B$. In our previous work, we showed that homogeneity of the acoustic information between the two voice records is important in order to have meaningful $LR$s [7]. In this work, we focus on the "speaker factor". We know from [17] that varying the speech extract used to represent a given speaker has a huge impact on $SR$ systems' performance. Based on this results, we assume that -in the view of a voice comparison automatic system- all the speakers do not behave the same way in response of similar condition changes: some speakers will be quite robust with limited $LR$ variation when some other are showing a huge variation.

This paper is dedicated to this speaker factor, which put the reliability issues mentioned in (i), (ii) and (iii) into perspective. We wish to propose a ”speaker profile” concept which classifies speakers depending on this ”speaker” aspect. It is important to remind that this ”speaker profile” notion as well as ”speaker factor” itself should be taken with caution as the effects we are working on are always seen using a $SR$ system as glasses.

This paper is structured as follows. Section 2 presents the speaker profile concept in the context of forensic voice comparison. Section 3 describes the LIA baseline system and shows experiments and results. Then, section 4 presents the conclusion and proposes some extends of the current work.

## 2. INTER-SPEAKER VARIABILITY: SPEAKER PROFILE CONCEPT

Forensic voice comparison is a delicate task that should be treated with much vigilance and caution [9, 18]; every detail should be taken into account to allow a reliable forensic technical report. If the presence of enough speaker specific information inside the two voice excerpts is mandatory, looking on speaker himself and his characteristics should not be forgotten as well. In the famous [19] article, the authors characterized the different speakers in terms of their error tendencies. Speakers for which the system has a normal behavior are denoted as “sheep”. Speakers who cause a proportionately high number of false rejection errors are called “goats”. Speakers who tend to cause false acceptance errors because they are accepting too much impostors are “lambs”, and those who tend to cause false acceptance errors as the impostor speaker are called “wolves”. [17] investigated more deeply this ”speaker factor” notion. In this article, the authors showed that speaker recognition systems performance depends significantly on which speech extract is used in order to represent a given speaker. The reported error rates are five time larger when ”worst” speech extracts are used for all the target speakers, compared to the rates obtained using ”best” speech extracts. In these two articles, the different classifications are based on percentage of both $FA$ and $FR$ which depend on a hard decision fixed by a threshold. the threshold is tuned in function of prior probabilities to have a genuine or an impostor trial and the costs (commercially speaking) of each kind of errors.

In forensic cases, the situation is different. First, following the Bayesian paradigm, there is no hard decision and an automatic system is expecting to present the results as a fully meaningful $LR$. Second, forensic experts should not take into account the hypothesis’ priors. So it is less important to know if a speaker is a “sheep”, a “goat”, a “lamb” or a “wolf” than to know how he behaves, viewed by a voice comparison framework. Therefore, we propose to classify the speakers into two main classes:

- “well-behaved” profile. This class groups the speakers which show a normal behavior. It corresponds mainly to the “sheep” speakers.
- “hybrid” profile. This class groups all the other speakers.

To characterize a ”speaker profile” accordingly to speaker’s characteristics is not as simple as it seems to be. The main point is to study the inter-speaker differences in terms of performance variations when some factors are changing, and to classify correspondingly the speakers into the profiles. This process involves many variation factors as speaker accent or dialect, sex, speaking style, prosody, emotion and even speaker age [20][21]. The factors mentioned before should be studied deeply in order to define properly the speaker profile. It is surprising that much less attention were paid to study the effect of intrinsic speaker variability compared to extrinsic factors as noise, and channel or microphone effects.

$SR$ systems are working as black boxes: scores are calculated in all situations regardless of the relevant information present in the two records and then calibrated (i.e. normalized) to be viewed as a $LR$ [22, 23]. The latter could be meaningless in some cases mentioned before. In Figure 1, we present a schematic view of a $LR$ estimation which takes as input homogeneity and speaker profile.

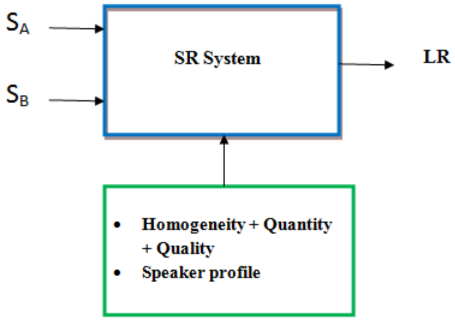

**Fig. 1**. Schematic view of meaningful LR estimation.

For a given speaker comparison trial composed of two speech recordings, S$_A$-S$_B$, the SR system estimates the corresponding LR taking into account two main factors:

- The homogeneity of the speaker specific information in S$_A$-S$_B$ [7].
- The speaker profile.

By adding these notions, the outputted LR is meaningful in itself.

## 3. EXPERIMENTS AND RESULTS

In order to show the impact of inter-speaker variability in terms of ”speaker” behavior, we propose several experiments based on FABIOLE database framework. These experiments will aim to illustrate the speaker profile concept presented in section 2,

### 3.1. Corpus

FABIOLE is a new speech database created inside the ANR-12-BS03-0011 FABIOLE project. The main goal of this database is to investigate the speaker factor, including intra-speaker variability, so we tried to control as much as possible the other factors. First, channel variability is reduced as all the excerpts come from French radio or television shows. Second, for most pairs, the quality of recordings are high in order to decrease noise effects. Third, all the speech files have a minimum duration of 30 seconds of speech. Then, we selected only male speakers. Finally, the number of targets and non targets trials per speaker is fixed.

FABIOLE database contains 130 male French native speakers divided into two sets:

- Set $T$: 30 targets speakers who everyone has at least 100 test files. Hence, each speaker can be associated with a large number of targets trials, which is a clear advantage compared to various other databases.

- Set $I$: 100 impostors who everyone has one file (one session). These test files are used essentially to create non-targets trials. It allows to associate a given impostor recording with all the $T$ speakers, removing one of the frequent bias in NIST-based experiments.

FABIOLE contains different speakers, including journalists as “Olivier Truchot”, announcers as “Thomas Soulie”, politicians as “Manuel Valls”, chroniclers as “Serge Hefez”, interviewers as “Fernand Tavares”, etc. Some speakers appear only in one emission as “Arnaud Ardoin” and “Michel Ciment” whereas others appear in several emissions as “Manuels Valls”. With the characteristics mentioned before, FABIOLE database seems to be well suited to study the impact of speaker factor.

FABIOLE material is close to the one of REPERE [24], ESTER 1, ESTER 2 [25] and ETAPE [26]. This characteristic allows to use these databases as a source of training data.

### 3.2. BASELINE LIA SYSTEM

In all experiments, we use as baseline the LIA_SpkDet system presented in [27]. This system is developed using the ALIZE/SpkDet open-source toolkit [28] [29] [30]. It uses I-vector approach [1].

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

The *Universal Background Model* ($UBM$) has 512 components and is trained by EM/ML. The $UBM$ and the total variability matrix, $T$, are trained on Ester 1&2, REPERE and ETAPE databases on male speakers that do not appear in FABIOLE database. They are estimated using “$7,690$” sessions from “$2,906$” speakers whereas the inter-session matrix $W$ is estimated on a subset (selected by keeping only the speakers who have pronounced at least two sessions) using “$3,798$” sessions from “$672$” speakers. The dimension of the I-Vectors in the total factor space is $400$.

For scoring, PLDA scoring model [31] is applied. The speaker verification score given two I-vectors $w_A$ and $w_B$ is the likelihood ratio described by:

$$score = log\frac{P(w_A, w_B|H_p)}{P(w_A, w_B|H_d)} \quad (1)$$

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.

### 3.3. Experimental protocol

All the experiments presented in this work are performed based upon FABIOLE database. FABIOLE proposes more than $150,000$ targets trials and $300,000$ non-targets trials divided into 30 subsets, one for each $T$ speaker (the speakers of the set $T$). So, for one subset, all the voice comparison pairs are composed with at least one recording pronounced by the corresponding $T$ speaker. It gives for a given subset $14950$ pairs of recordings distributed as follows: $4950$ same-speaker pairs and $10,000$ different-speakers pairs. The target pairs were obtained from all the combinations of the $100$ recordings available for the corresponding $T$ speaker ($C^2_{100}$ targets pairs). Whereas, non-targets pairs are obtained by pairing each of the $T$ speaker's recording (100 are available) with each of the $100$ speakers of the $I$ set, forming consequently $(100 \times 100 = 10000)$ non-targets pairs.

In order to study the inter-speaker variations and to emphasize the ”speaker factor”, we compute the *log-likelihood-ratio cost* ($C_{llr}$) independently on each of the 30 trials subsets. We selected the $C_{llr}$, largely used in forensic voice comparison, because it is based on likelihood ratios and not on hard decisions like *equal error rate* (EER) [12, 32, 33, 34]. $C_{llr}$ has the meaning of a cost or a loss: lower the $C_{llr}$ is, better is the performance. As our main interest is the ”speaker factor”, we use the minimum value of the $C_{llr}$ (denoted $C^{min}_{llr}$) in order to withdraw the impact of calibration mistakes.

For comparison, *False Reject* rate ($FR$) and *False Alarm* rate ($FA$) are also computed using a threshold estimated onto the whole test set and tuned to correspond at the global $EER$.

## 4. PRELIMINARY ANALYSIS OF RESULTS

The global $C^{min}_{llr}$ (computed using all the trial subsets put together) is equal to $0.1765$ $bits$ and the corresponding global $EER$ is $4.5\%$. Figure 3 presents the corresponding target and non-target score distributions.

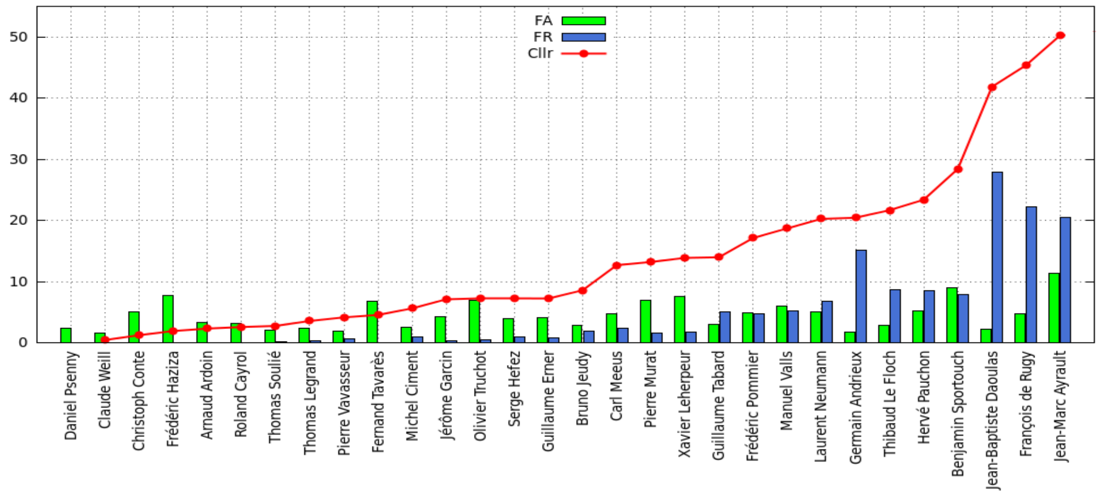

**Fig. 2**. $C_{llr}^{min} \times 100$, $FA\%$, $FR\%$ for all speakers.

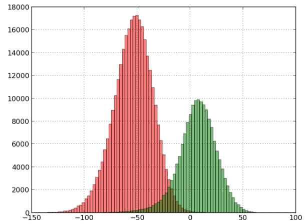

**Fig. 3**. Target and non-target score distributions for the pooled condition (all the comparison tests taken together).

This global representation hides the impact of the inter-speaker differences due to the speaker factor. In Figure 2, we present $C_{llr}^{min}$ estimated individually for each $T$ speaker subset. The subsets are ranked from the lowest to the highest values of $C_{llr}^{min}$. $FR\%$ and $FA\%$ are also provided.

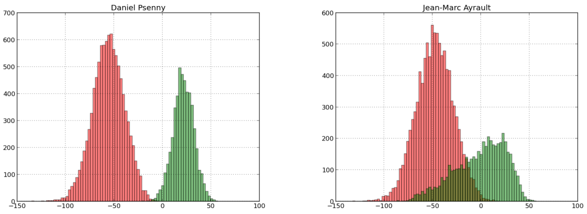

**Fig. 4**. Examples of target and non-target score distributions for a well-behaved speaker (left) and an hybrid speaker (right).

This experiment confirms our hypothesis: even if the trial subsets are mainly similar (number of recordings, duration, signal quality, channel variability, etc.) and if the impostor examples (in terms of speaker as well as in recordings) are strictly identical for all the subsets, a large variability is present which means that speakers do not behave the same way. We notice that 3 speakers show a $C_{llr}^{min}$ higher than $0.4$ $bits$, when 16 speakers present a $C_{llr}^{min}$ lower than $0.09$ $bits$ while the remaining speakers present a medium cost close to the global one. This observation suggests strongly the existence of speaker profiles.

To illustrate the concept of ”speaker profile” we are promoting, we present in Fig.4 examples of speakers presenting a well-behaved and an hybrid profiles, in a form corresponding to Fig.3. Here, extreme speakers from Fig.2 are selected.

## 5. CONCLUSION

This work is focused on forensic context and investigates inter-speaker variability in terms of ”speaker behavior”, as viewed by an automatic SR system. It took advantage of a new database, FABIOLE, designed specifically for this kind of work. When the global $C_{llr}^{min}$ (computed using all the trial subsets put together) is equal to $0.1765$ $bits$, we observed that about half of the speakers obtain significantly better $C_{llr}^{min}$ (lower than $0.09$ $bits$) and about $10\%$ of the speakers present very high $C_{llr}^{min}$ (higher than $0.4$ $bits$) compared to the average cost. This result supports strongly our hypothesis that speakers could be classified into ”speaker profiles”. When previous works like [19][17] proposed several speaker profiles, we defined only two profiles well suited for a forensic context: well-behaved speakers and hybrid speakers, as for us only the first ones could provide meaningful $LR$s.

In future works, it will be of a particular interest to try to predict the speaker profile based on the speaker characteristics. We want also to see if, combined with other information like the homogeneity developed in [7], the speaker profile could be used in order to help an automatic voice comparison system to propose meaningful $LR$s in various situations.

## 6. ACKNOWLEDGMENTS

The research reported here was supported by ANR-12-BS03-0011 FABIOLE project.

## 7. REFERENCES

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

[2] AFSP; Standards for the formulation of evaluative forensic science expert opinion. Science&Justice, pp. 161-164 (2009).

[3] Champod, Christophe, and Didier Meuwly. ”The inference of identity in forensic speaker recognition.” Speech communication 31.2 (2000): 193-203.

[4] Aitken, Colin GG, and Franco Taroni. Statistics and the evaluation of evidence for forensic scientists. Vol. 16. Chichester: Wiley, 2004.

[5] G. S. Morrison, Measuring the validity and reliability of forensic likelihood-ratio systems, Science & Justice, vol. 51, no. 3, pp. 9198, 2011.

[6] G. S. Morrison, C. Zhang, and P. Rose, An empirical estimate of the precision of likelihood ratios from a forensic-voice comparison system, Forensic science international, vol. 208, no. 1, pp. 5965, 2011.

[7] M. Ajili, JF. Bonastre, S. Rossato, J. Kahn and I. Lapidot. An information theory based data-homogeneity measure for voice comparison (Interspeech 2015).

[8] J.-F. Bonastre, F. Bimbot, L.-J. Boe, J. P. Campbell, D. A.Reynolds, and I. Magrin-Chagnolleau, Person authentication byvoice: a need for caution. in INTERSPEECH, 2003.

[9] J. P. Campbell, W. Shen, W. M. Campbell, R. Schwartz, J.-F.Bonastre, and D. Matrouf, Forensic speaker recognition. Institute of Electrical and Electronics Engineers, 2009.

[10] Daubert v. merrell dow pharmaceuticals, inc. p. 579, 1993.

[11] M. J. Saks and J. J. Koehler, The coming paradigm shift in forensic identification science, Science, vol. 309, no. 5736, pp. 892895, 2005.

[12] G. S. Morrison, Forensic voice comparison and the paradigm shift, Science & Justice, vol. 49, no. 4, pp. 298308, 2009.

[13] P. Rose, Technical forensic speaker recognition: Evaluation, types and testing of evidence, Computer Speech & Language, vol. 20, no. 2, pp. 159191, 2006.

[14] W. M. Campbell, D. A. Reynolds, J. P. Campbell, and K. Brady, Estimating and evaluating confidence for forensic speaker recognition. in ICASSP, 2005, pp. 717720.

[15] E. Mengusoglu and H. Leich, Confidence measures for speech/speaker recognition and applications on turkish lvcsr, 2004.

[16] N. R. Council, Strengthening forensic science in the united states: A path forward, 2009.

[17] Kahn, J., Audibert, N., Rossato, S. and Bonastre, J. F. Intra-speaker variability effects on Speaker Verification performance. In Odyssey p. 21 (2010).

[18] Bonastre, J. F., Kahn, J., Rossato, S., Ajili, M. . Forensic Speaker Recognition: Mirages and Reality. In Fuchs, S., Pape, D., Petrone, C., Perrier, P. (2015). Individual Differences in Speech Production and Perception.

[19] Doddington, G., Liggett, W., Martin, A., Przybocki, M., & Reynolds, D. (1998). Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation.

[20] F Kelly, JHL Hansen. The effect of short-term vocal aging on automatic speaker recognition performance (INTERSPEECH 2015).

[21] Matveev, Yuri. ”The problem of voice template aging in speaker recognition systems.” Speech and Computer. Springer International Publishing, 2013. 345-353.

[22] Ramos-Castro, D., J. Gonzalez-Rodriguez, and J. Ortega-Garcia. ”Likelihood ratio calibration in a transparent and testable forensic speaker recognition framework.” Speaker and Language Recognition Workshop, IEEE Odyssey 2006.

[23] Doddington, George. ”The role of score calibration in speaker recognition.” Thirteenth Annual Conference of the International Speech Communication Association. 2012.

[24] Giraudel, A., Carr, M., Mapelli, V., Kahn, J., Galibert, O., & Quintard, L. (2012, May). The REPERE Corpus: a multimodal corpus for person recognition. In LREC (pp. 1102-1107).

[25] Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J. F., & Gravier, G. (2005, September). The ESTER phase II evaluation campaign for the rich transcription of French broadcast news. In Interspeech (pp. 1149-1152).

[26] Gravier, G., Adda, G., Paulson, N., Carr, M., Giraudel, A., & Galibert, O. (2012). The ETAPE corpus for the evaluation of speech-based TV content processing in the French language. In LREC-Eighth international conference on Language Resources and Evaluation (p. na).

[27] Matrouf, D., Scheffer, N., Fauve, B. G., Bonastre, J. F. A straightforward and efficient implementation of the factor analysis model for speaker verification. INTERSPEECH (2007).

[28] Bonastre, J. F., Wils, F., Meignier, S. ALIZE, a free toolkit for speaker recognition. In ICASSP (2005) (pp. 737-740).

[29] Bonastre, J. F., Scheffer, N., Matrouf, D., Fredouille, C., Larcher, A., Preti, A.,Mason, J. S. ALIZE/spkdet: a state-of-the-art open source software for speaker recognition. In Odyssey (2008).

[30] Larcher, A., Bonastre, J. F., Fauve, B. G., Lee, K. A., Lvy, C., Li, H., Parfait, J. Y. . ALIZE 3.0-open source toolkit for state-of-the-art speaker recognition. In INTERSPEECH (2013)

[31] Prince, Simon JD, and James H. Elder. ”Probabilistic linear discriminant analysis for inferences about identity.” Computer Vision. ICCV, IEEE 11th International Conference (2007).

[32] Brummer, N., and du Preez, J. Application-independent evaluation of speaker detection. Computer Speech and Language, pp.230-275 (2006).

[33] Castro, D. R. Forensic evaluation of the evidence using automatic speaker recognition systems (Doctoral dissertation, Universidad autnoma de Madrid), (2007).

[34] Gonzalez-Rodriguez, J., Ramos, D. Forensic automatic speaker classification in the Coming Paradigm Shift. In Speaker Classification I (pp. 205-217). Springer Berlin Heidelberg, (2007).