FABIOLE, a Speech Database For Forensic Speaker Comparison

Moez Ajili, Jean-François Bonastre, Juliette Kahn, Solange Rossato, Guillaume Bernard
University of Avignon, National Metrology and Testing Laboratory , University of Grenoble
Avignon, Paris, Grenoble
{moez.ajili, jean-francois.bonastre}@univ-avignon.fr, {juliette.kahn,guillaume.bernard}@lne.fr, solange.rossato@imag.fr

Abstract

A speech database has been collected for use to highlight the importance of "speaker factor" in forensic voice comparison. FABIOLE has been created during the *FABIOLE project* funded by the French Research Agency (ANR) from 2013 to 2016. This corpus consists in more than 3 thousands excerpts spoken by 130 French native male speakers. The speakers are divided into two categories: 30 target speakers who everyone has 100 excerpts and 100 "impostors" who everyone has only one excerpt. The data were collected from 10 different French radio and television shows where each utterance turns with a minimum duration of $30s$ and has a good speech quality. The data set is mainly used for investigating *speaker factor* in forensic voice comparison and interpreting some unsolved issue such as the relationship between speaker characteristics and system behavior. In this paper, we present FABIOLE database. Then, preliminary experiments are performed to evaluate the effect of the "speaker factor" and the show on a voice comparison system behavior.

Keywords: forensic voice comparison, intra-speaker variability, speaker recognition, FABIOLE database.

1. Introduction

Speaker recognition (SR) system comparison is made possible by the organization of evaluation campaigns such as those organized by the *National Institute of Standards and Technology* (NIST), NIST-SRE. Since the first organization in 1996, SR systems have achieved significant progresses and have reached low error rate ($\approx 1\%$). The robustness of this kind of evaluation is ensured by using a very large number of voice comparison samples. The number of samples per speaker as well as the characteristics of the speakers themselves, except for their native language and sex, are not taken into account in the evaluation plan. A major challenge for present SR systems is their tolerance to speaker differences and variations in context. Differences in speaker characteristics are a major source of inter- and intra-speaker variation which should be taken into consideration. Some research have attempted to deal with some aspects of this variability (Kahn et al., 2010; Doddington et al., 1998) in which authors have shown that speaker recognition performance depends on "speaker factor". If inter-speaker variability is an important factor, intra-speaker variability is not less important (Kahn et al., 2010). It involves many factors as speaker accent or dialect, speaking style, prosody, emotion and even speaker age. Regarding the latter, (Matveev, 2013) show a clear trend of degradation of the performance of automatic speaker recognition systems in a time interval of up to 4 years. Consequently, both inter- and intra-speaker variability should be studied in order to have a reliable assessment. These information could prove useful in some sensitive applications such as forensic voice comparison.

The availability of good speech databases is crucial for "speaker factor" assessment. SR evaluation is emphasized by numerous evaluation campaigns over the last decades, among which the annual NIST SRE evaluations since 1996. Although, it is the most commonly used framework for the evaluation, it does not allow to have a deep study on the speaker factor impact. We take for example NIST'SRE 2008 where there are a large number of speaker (≈ 300) but every speaker disposes only 3 speech files in average, inter-speaker variability is high but intra-speaker variability can not be studied because of the low number of utterances per speaker. Moreover, even forensic databases (Ramos et al., 2008; van der Vloed et al., 2014; Morrison et al., 2015), -that should pay attention to all important factors- could not bring out the importance of this factor for the same reason (low utterance number per speaker). We take for example a recent forensic database (Morrison et al., 2015) where a large number of speakers (301 male and 231 females) are provided but every one disposes a limited number of speech recordings (only 5 male and 19 female have been registered more than 3 times). Intra-speaker variability is very important and should not be neglected anymore especially for forensic voice comparison. In this context, "FABIOLE" [1], a large French speech database has been collected for use in order to allow a robust study of the inter- and intra-speaker variability as it is a major issue in scientific research generally and in forensic voice comparison particularly. The corpus consists of more than 3100 speech utterances of 130 French speakers distributed as follow: 30 "targets speakers who everyone has 100 speech recordings and 100 "impostors" who everyone has only one utterance. Speakers could be politicians, interviewers, chronicles, etc. FABIOLE data were collected from French radio and television shows.

The corpus differs from previous corpora in three ways:

a) The number of speech file per speaker is high. So, the number of target and non-target trials per speaker would be quite large to have a reliable investigation on intra-speaker variability.

b) Utterance conditions: excerpts are quite long and have a good quality. Each utterance turns with a minimum duration of $30s$.

c) Presence of speaker with different role and variation in context. This allow to investigate the impact of "speaker's role" on the strength system behavior.

The aim of this study is not dedicated to decrease the per-

[1] New corpus that will be easily accessible to the scientific community.

formance measure or to improve the recognition strength as much as it is dedicated to focus on the relatively unexplored issue, dependence of system performance on intrinsic speaker characteristics, that could be conclusive in some forensic cases.
This paper is structured as follows. Section 2 presents FABIOLE database. Section 3 describes a first example of FABIOLE use. Then, section 4 presents first conclusion and some perspectives.

2. Database Description

FABIOLE is a new speech database created during the ANR-12-BS03-0011 FABIOLE project. The main goal of this database is to investigate the speaker factor, including intra-speaker variability. That is why 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 as our main interest is the speaker factor. Third, all the speech files have a minimum duration of 30 seconds of speech (short utterance is no longer a matter). Then, we selected only male speakers, female are not selected because we does not find 30 women who have enough excerpts with the desired characteristics (see subsection 2.1. for the details). Finally, the number of targets and non targets trials per speaker is fixed. A common design technique, used in many databases, such as Switchboard-1, is to record a smaller set of speakers in many sessions and a separate, larger set of speakers in a single session. With this technique, one can achieve both good sampling of inter-speaker variability in potential impostor speakers, and of intra-speaker variability in client speakers. Fabiole database is based on the same technique as shown in the following subsection 2.1..
Furthermore, the database should not be too far from other French data that can be used as training data to build state-of -art models and enable us to perform automatic transcriptions. In the literature the following databases are available:

- ESTER 1 (Galliano et al., 2006): About 100 hours of transcribed data make up the corpus, recorded between 1998 and 2004 from six French speaking radio stations: France Inter, France Info, RFI, RTM, France Culture and Radio Classique. Shows last from 10 minutes up to 60 minutes. They consist mostly of prepared speech such as news reports, and a little conversational speech (such as interviews).
- ESTER 2 (Galliano et al., 2009) comes to supplement ESTER 1 corpus with about 100 hours of transcribed broadcast news recorded from 1998 to 2004.
- REPERE (Kahn et al., 2012; Galibert and Kahn, 2013) It currently contains 60 hours of video with multimodal annotations. The systems have to answer the following questions: Who is speaking? Who is present in the video? What names are cited? What names are displayed? The challenge is to combine the various information coming from the speech and the images.
- ETAPE corpus (Gravier et al., 2012) consists of 30 hours of TV and radio broadcasts, selected to cover a wide variety of topics and speaking styles, emphasizing spontaneous speech and multiple speaker areas.

FABIOLE contains the same type of records as those contained in these databases. The list of speakers of REPERE, ETAPE and ESTER that must be excluded to avoid biasing the system is provided in the package FABIOLE.

2.1. Speaker information

Regarding inter-speaker variability, it is of course desirable to include many speakers to achieve a good sampling of a speaker population. At the same time, it is often desirable to get a good sampling of individual speakers over multiple sessions to study intra-speaker variability. Note that a single-session database is not useful for estimating the absolute level of performance for a speaker recognition system in a practical application because it does not include intra-speaker variability. Consequently, every speaker should be represented in the database by a large number of speech file recorded in different shows. With only limited resources for database creation, a trade-off between the number of speakers, number of sessions and total cost is often necessary.
The challenge consists in having different speakers with sufficient number of test files each one in order to study inter- and intra-speaker variability. FABIOLE database contains 130 male French native speakers divided into two sets:

- Set T: 30 targets speakers who everyone has at least 100 speech files. Hence, each speaker can be associated with a large number of targets trials, which is a clear advantage compared to various other databases in which the number of target trials per speaker is very low.
- Set I: 100 impostors who everyone has one speech 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.

Table 1 presents the amount of data per broadcaster for every speaker. The 30 speakers are a homogeneous group of native male speakers, speak the same dialect and range in age from 30 to more than 70 years old. Speakers have different professions and their collected speech is totally used for public speaking. We find interviewers as *Olivier Truchot*, politicians as *Jean-Marc Ayrault*, chroniclers as *Daniel Psenny*, debater as *Bruno Jeudy*, etc. Hence, some speakers exhibit a relatively small variation in profession.
We can see also that some speakers appear only in one show as *Arnaud Ardoin* and *Michel Ciment* whereas others appear in more than one show as *Manuels Valls* as shown in Table 1: 13 speakers were recorded from 1 kind of show, 10 speakers from 2 kinds of shows and 7 speakers from 3-5 kinds of shows. The effective duration of speech in each call is approximately 30 seconds.

2.2. Shows and sampling

All the speech files are sampled to 16 kHz and they are collected from 10 different sources. FABIOLE includes

Table 1: Amount of data per broadcaster for every speaker.

Spk vs show	cvgdinfo	cvgddebat	entreligne	parlinfo	topquestions	bfmstory	temPauch	MsqPlum	ComParle	Spublic
Arnaud Ardoin	0	2:13:41	0	0	0	0	0	0	0	0
Daniel Psenny	0	0	0	0	0	0	0	0	1:04:15	0
Guillaume Tabard	0	0	0	0	0	0:0:31.18	0	0	1:02:17.13	0
Michel Ciment	0	0	0	0	0	0	0	1:11:31.54	0	0
Thomas Legrand	0	0	0	0:0:47.90	0	0	0	0	1:17:23.49	0
Fernand Tavarès	0:0:38.80	0	0	1:02:15.73	0	0	0	0	0	0
Hervé Pauchon	0	0	0	0	0	0	0:53:50.50	0	0	0
Olivier Truchot	0	0	0	0	0	1:13:24.20	0	0	0	0
Thomas Soulié	0:9:01.94	0	0	1:21:18.63	0	0	0	0	0	0
Benjamin Sportouch	0	0	0	0:15:26.83	0	0	0		0:43:13.07	0
François de Rugy	0:12:05.82	0	0	0:36:18.00	0:3:45.29	0:5:39.76	0	0	0	0
Jean Baptiste Daoulas	0:17:19.52	0:0:34.13	0	0:39:32.99	0	0	0	0	0	0
Pierre Murat	0	0	0	0	0	0	0	0:54:40.01	0:11:36.74	0
Bruno Jeudy	0	0	0	0	0	0:5:18.27	0	0	1:01:39.57	0
Frédéric Haziza	0	0	1:05:48.02	0	0	0:1:23.94	0	0	0	0
Jean-Marc Ayrault	0:1:28.30	0	0	0:1:49.59	0:52:47.12	0	0	0	0	0
Pierre Vavasseur	0	0	0	0	0	0	0	0	1:08:09.38	0
Carl Meeus	0	0	0:59:40.36	0:3:47.07	0	0	0		0	0
Frédéric Pommier	0	0	0	0	0	0	0	0	1:07:10.21	0
Jérôme Garcin	0	0	0	0	0	0	0	1:58:45.01	0	0
Roland Cayrol	0	0	0	1:18:15.50	0	0	0	0	0	0
Christophe Conte	0	0	0	0	0	0	0	0	1:26:08.53	0
Germain Andrieux	0:9:28.61	0:1:52.86	0	0:48:11.11	0	0	0	0	0	0
Laurent Neumann	0	0	0:33:39.01	0:4:09.54	0	0:29:18.56	0	0	0	0
Serge Hefez	0	0	0	0	0	0	0	0	0	1:28:54.79
Claude Weill	0	0	0:58:00.13	0:13:47.99	0	0	0	0	0	0
Guillaume Erner	0	0	0	0	0	0	0	0	0	2:58:14
Manuel Valls	0:2:35.97	0	0:1:21.95	0:2:18.82	1:49:02.31	0:4:12.24	0	0	0	0
Thibaud Le Floch	0:20:16.74	0:2:20.55	0	0:34:34.68	0	0	0	0	0	0
Xavier Leherpeur	0	0	0	0	0	0	0	1:09:27.19	0	0
100locuteurs	0	0	0:1:06.40	0:4:21.74	1:29:44.85	0:7:15.74	0:1:36.96	0:1:33.17	0:4:10.96	0:19:43

speech recordings from television and radio sequences of 10 programs (same shows used in databases mentioned above) recently recorded (from 2013 to 2014). Excerpts come from:

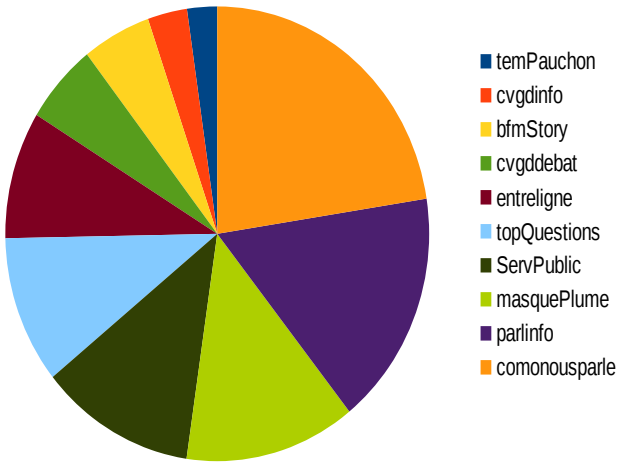

Figure 1: *Proportion of shows contribution in FABIOLE database ranged from the lowest (Blue) to the highest (Orange) contribution.*

- Debates such as “Ca vous regarde - Le débat” (*cvgddebat*), “Entre les Lignes” (*entreligne*), “Le masque et la plume” (*MsqPlum*).
- Chronic as “Service public” (*Spublic*), “Comme on nous parle” (*ComParle*).
- Interviews as “Un Temps de Pauchon” (*temPauch*).
- Parliamentary jousting as “Top Questions” (*topquestions*).
- News such as “BFM Story” (*bfmstory*), ”LCP Info” (*parlinfo*), ”Ca vous regarde-l’Info (*cvgdinfo*)”.

This variety can involve different speaking style and thereby it allow to study its effect on the system behavior. The amount of data per broadcaster is showed in Fig 1. We can see that excerpts come from 10 different shows with high difference of contribution. For example, *Comme on nous parle* is 22.5% of Fabiole while *temps de Pauchon* is only 2.2% from all the database. We are aware of the limits of this distribution in terms of analysis but it is very difficult to find an equal distribution of speakers and shows since speakers tend to always address the same show.

2.3. Others parameters: Age, Speaking style and profession

As our main interest are speakers of set T, we present in Table 2 detailed information related to their age and profession while for speakers of set I information is given globally. Table 2 shows that the majority of target speakers are 40 to 60 years old and the most of them are interviewers (10 Spks), chroniclers(7 Spks) and debaters(7 Spks).
Note that for some speakers as *Bruno Jeudy*, age information is missing.

2.4. Data transcription

FABIOLE database has been entirely orthographically transcribed. The transcription is done as follow: 10% of FABIOLE are transcribed manually whereas the remaining part

Table 2: Speakers ages and professions.

Spk vs age	20-30	30-40	40-50	50-60	60-70	>70	Profession
Arnaud Ardoin	0	0	1	0	0	0	interviewer
Daniel Psenny	0	0	0	1	0	0	chronicler
Guillaume Tabard	0	0	0	1	0	0	debater
Michel Ciment	0	0	0	0	0	1	chronicler
Thomas Legrand	0	0	0	1	0	0	chronicler
Fernand Tavarès	0	0	0	0	0	1	interviewer
Hervé Pauchon	0	0	0	1	0	0	envoyé spécial
Olivier Truchot	0	0	1	0	0	0	interviewer
Thomas Soulié	0	0	1	0	0	0	interviewer
Benjamin Sportouch	0	1	0	0	0	0	debater
François de Rugy	0	0	1	0	0	0	politician
Jean Baptiste Daoulas	0	1	0	0	0	0	envoyé spécial
Pierre Murat	0	0	0	0	1	0	debater
Bruno Jeudy	-	-	-	-	-	-	debater
Frédéric Haziza	0	0	0	1	0	0	interviewer
Jean-Marc Ayrault	0	0	0	0	1	0	politician
Pierre Vavasseur	0	0	0	0	1	0	debater
Carl Meeus	0	0	1	0	0	0	debater
Frédéric Pommier	0	0	1	0	0	0	chronicler
Jérôme Garcin	0	0	0	1	0	0	interviewer
Roland Cayrol	0	0	0	0	0	1	scientist
Christophe Conte	0	0	1	0	0	0	chronicler
Germain Andrieux	0	1	0	0	0	0	interviewer
Laurent Neumann	0	0	0	1	0	0	chronicler
Serge Hefez	0	0	0	0	1	0	interviewer
Claude Weill	0	0	0	1	0	0	chronicler
Guillaume Erner	0	0	1	0	0	0	interviewer
Manuel Valls	0	0	0	1	0	0	politician
Thibaud Le Floch	0	0	1	0	0	0	interviewer
Xavier Leherpeur	0	0	0	1	0	0	debater
100locuteurs	1	9	30	27	30	2	all profession

are transcribed thanks to Speeral, LIA automatic transcription system (Linares et al., 2007).

This system used for the transcription of REPERE development set (contains speech recordings close acoustically to FABIOLE excerpts) reaches an overall *Word Error Rate* (WER) of 29% (Bigot et al., 2013). These high WER are mainly due to speech disfluencies and to adverse acoustic environments (for example, calls from noisy streets with mobile phones).

Transcriptions are provided in LIA transcription encoding and follow common guidelines of orthographic transcription. From the transcription, task specific information can be exported into appropriate formats (stm, mdtm, etc). We show in Fig 2, an example of a transcription of 1 speech recording.

The transcription does not have a speaker turn because every file correspond entirely to a single speaker. The transcription file includes information on the name and gender of the speaker, if he is a native speaker or not, the speech type, the channel if it is a telephone vs. studio (In our case, all excerpts are from studio). The transcription file is further divided into speech segments.

The transcription includes punctuation and is case-sensitive. In particular, the transcription indicates disfluencies such as the word “euh” in French.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE Trans SYSTEM "trans-14.dtd">
<Trans scribe="oh" audio_filename="Arnaud_Ardoin---cavousregardeledebat_2013-10-24---3369---00.wav"
version="1" version_date="051107">
<Speakers>
<Speaker id="S0M" name="S0M" check="yes" type="male" dialect="native" accent="" scope="local"/>
</Speakers>
<Episode>
<Section type="report" startTime="0" endTime="48.23">
<Turn speaker="S0M" startTime="0" endTime="16.43" mode="" fidelity="" channel="studio">
<Sync time="0"/>
j'_ai pas sur_la dans nos dans_le fichier informatique pour_la de1couvrir celle_de
Se1gole2ne_Royal qui tient..
</Turn>
<Turn speaker="S0M" startTime="16.43" endTime="27.97" mode="" fidelity="" channel="studio">
<Sync time="16.43"/>
tre2s inte1ressant datait photos un_peu originale de Se1gole2ne_Royal sait ce sondage BVA le je...
</Turn>
<Turn speaker="S0M" startTime="27.97" endTime="37.51" mode="" fidelity="" channel="studio">
<Sync time="27.97"/>
interdiction licencieux entreprise faisant des be1ne1fices cinquante pour_cent quarante sept...
</Turn>
</Section>
</Episode>
</Trans>
```

Figure 2: *An example of a transcription of one utterance.*

3. Examples of FABIOLE use

This section present first analysis conducted using the FABIOLE database. These experiments goal to analyze the influence of the intra-speaker variability and the influence of the show on a baseline voice comparison system.

3.1. BASELINE Voice comparison System

In all experiments, we use as baseline the LIA_SpkDet system presented in (Matrouf et al., 2007). This system is developed using the ALIZE/SpkDet open-source toolkit (Bonastre et al., 2005) (Bonastre et al., 2008) (Larcher et al., 2013). It uses I-vector approach (Dehak et al., 2011).
Acoustic features are composed of 19 LFCC parameters (cepstral parameters using a linear scale) issued from a frequency window restricted to 300-3400 Hz [2], 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 (Prince and Elder, 2007) 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.

3.2. Experimental protocol

All the experiments presented in this paper are performed based upon FABIOLE database. FABIOLE proposes $150,000$ targets trials and about 5 $millions$ non-targets trials. In this paper, we use only a subset of comparisons trials. We adopt the following protocol. We use all the target trials and only $300,000$ non-targets trials. The selection will be detailed below. The trials are 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.
As we mentioned before, this database will be used to study the inter- and intra-speaker variability. A first step forward is to focus on an eventual link between speaker and system performance. To do so, 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 a loss in terms of likelihood ratio discrimination power which does not require threshold and hard decisions like *equal error rate* (EER) (Brümmer and Du Preez, 2006; Castro, 2007; Gonzalez-Rodriguez and Ramos, 2007; Morrison, 2009). C_{llr} has the meaning of a cost or a loss: lower is the C_{llr}, better is the performance. We use the calibrated C_{llr} and the minimum value of the C_{llr} (denoted respectively C^{cal}_{llr} and C^{min}_{llr}). C^{cal}_{llr} involves calibration loss while C^{min}_{llr} contains only discrimination loss. We can judge the quality of the calibration Q_{cal} (i.e., the mapping from score to log-likelihood-ratio which is actually present in the detector) by:

$$Q_{cal} = C^{cal}_{llr} - C^{min}_{llr}. \quad (2)$$

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.
Note that parameters of calibration are estimated based on the pooled conditions (all the subset are put together) using FoCal Toolkit (Brummer, 2007).

3.3. Speaker factor

The global C^{min}_{llr} (respectively C^{cal}_{llr}) (computed using all the trial subsets put together) is equal to 0.1765 $bits$ (respectively 0.1867 $bits$) and the corresponding global EER is 4.52%. Fig 3 presents the corresponding target and non-target score distributions.

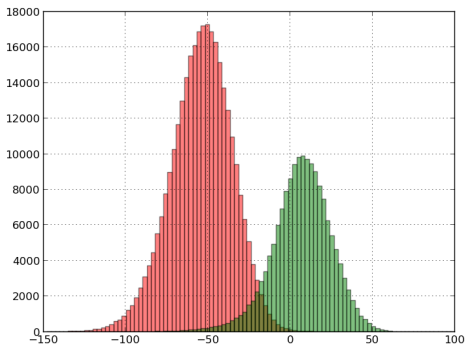

Figure 3: Target and non-target score distributions for the pooled condition (all the comparison tests taken together.

This global representation is close to performance estimation approach of the main evaluation campaigns (like the NIST's ones). It hides the impact of the inter-speaker differences due to the speaker factor.
In order to highlight this aspect, in Fig 4, we present C^{min}_{llr} estimated individually for each T speaker subset. The subsets are ranked from the lowest to the highest values of C^{min}_{llr}. $FR\%$ and $FA\%$ are also provided.
Fig.5 is the score distributions for the two extreme speakers from Fig.4 (in a form corresponding to Fig.3). These examples illustrate two speaker with different behavior. In forensic case, it should be treated differently. For more details, you could see (Ajili et al., 2016) where we detail a speaker profile concept.

[2] we select only a restricted frequency band because there are some utterance taken by telephone.

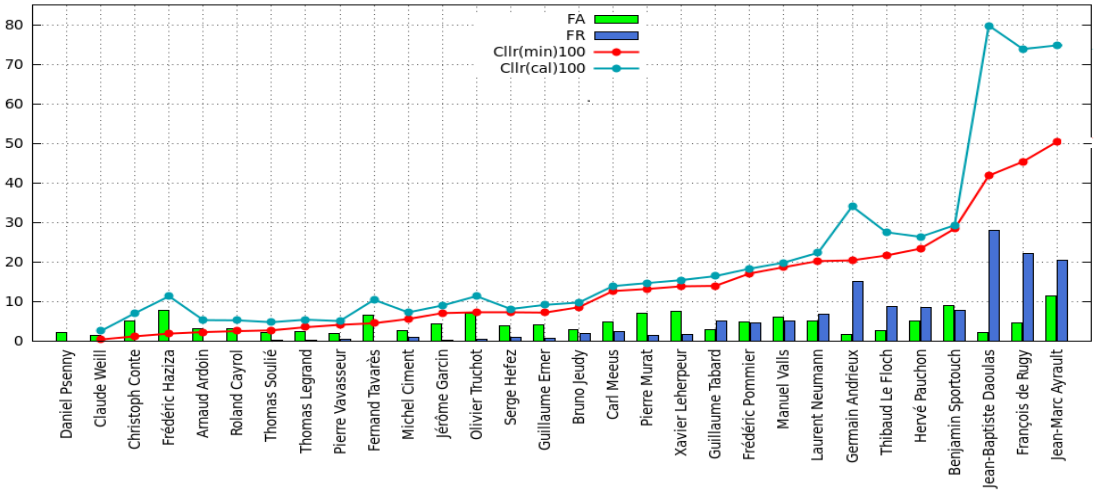

Figure 4: $C_{llr}^{min} \times 100$, $C_{llr}^{cal} \times 100$, $FA\%$, $FR\%$ for all speakers of set T.

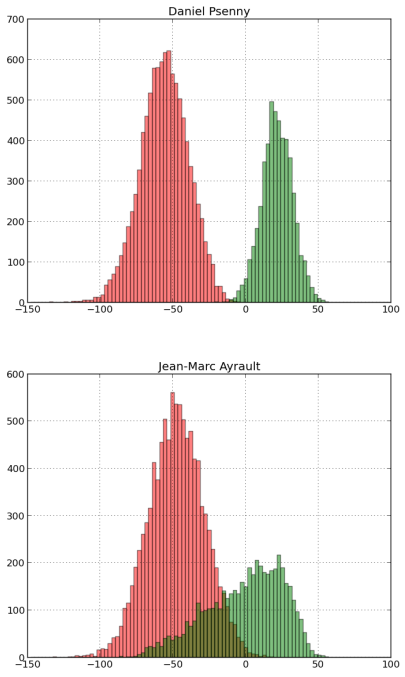

Figure 5: Examples of target and non-target score distributions for two speaker who imply different SR behavior.

The large difference between the two set of distributions highlight the importance of speaker factor: 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 speakers as well as in recordings) are strictly identical for all the subsets, a large performance variability is still present between the test sets. As the main difference between the sets is the T target speaker, it seems to indicate that speakers do not behave the same way.

To investigate more deeply the speaker effect, we come back to Fig 4: 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.

3.4. Show impact

To study the effect of the show on the performance, trials where the two recordings come from the same show are put together. We obtain at the end 10 subsets corresponding to our different shows. Then, C_{llr}^{cal}, C_{llr}^{min} and EER are calculated for each subset. Results are presented in Table 3.

Table 3: Performance per show. N_T and N_N are respectively numbers of target and non target trials.

Emiss. vs perf.	C_{llr}^{cal}	C_{llr}^{min}	EER	N_T	N_N
ServPublic	0.1136	0.0924	2.51	18811	5607
temPauchon	0.2026	0.0857	2.5	4950	200
ComonNouParle	0.0986	0.0904	2.42	36616	4704
masqPlume	0.3	0.207	6.32	19549	790
bfmStory	0.1531	0.0983	2.54	5933	1848
cvgddebat	0.0056	-	-	4959	0
cvgdinfo	0.67	-	-	1413	0
entreligne	0.1322	0.089	2.9	13721	654
parlinfo	0.2269	0.209	5.82	24762	3822
topquestions	0.466	0.36	11.2	8214	9400

Ça vous regarde le débat and *Ça vous regarde l'info* are excluded from the analysis because of the low number of non-target comparisons. Table 3 shows that SR behavior is influenced by the show. The "show's performance" presents a large variability: C_{llr}^{min} is varying from 0.085 $bits$ for *Temps de Pauchon* to 0.36 $bits$ for *Top Questions*.

4. Conclusions and perspectives

In this paper, we present FABIOLE corpus created during the FABIOLE project funded by the French Research Agency (ANR). This corpus would help to increase significantly the work dedicated to highlight the importance of speaker factor in forensic process.
FABIOLE database includes 30 "target" speakers who everyone disposes 100 speech utterances and 100 "impostors" who everyone has only 1 utterance. The excerpts come

from 10 different shows including different speaking style such as debates, political speech, news, etc.
Speakers have a large variability in profession. Indeed, FABIOLE includes interviewers, politicians, chroniclers, debaters, etc. FABIOLE also presents data from speakers with large age differences (from 30 to more than 70 years old).
To highlight the importance of "speaker factor", we presented preliminary results that shows the variation in C_{llr} according to the 30 target speaker. We showed that SR system presents different behavior: high discrimination power for the majority of speakers (low C_{llr}), while for 3 speakers its discrimination power degrades (high C_{llr}). Moreover, radio and TV shows influence the SR system behavior: Trials where both voice recordings came from *Comme on nous parle* have a low C_{llr} (0.09 bits) while those from *Top Questions* present a high C_{llr} (0.46 bits).
It is our hope and belief that this database will be a useful speech resource and contribute to the advances of state-of-the-art in forensic speaker comparison long after the end of the project.

5. Acknowledgements

The research reported here was supported by ANR-12-BS03-0011 FABIOLE project. Authors thank all thartiicipants in this Working Group in particular Olivier Galibert.

6. Bibliography

Ajili, M., Bonastre, J.-F., Rossato, S., and Kahn, J. (2016). Inter-speaker variability in forensic voice comparison:a preliminary evaluation. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE.

Bigot, B., Senay, G., Linares, G., Fredouille, C., and Dufour, R. (2013). Combining acoustic name spotting and continuous context models to improve spoken person name recognition in speech. In *INTERSPEECH*, pages 2539–2543.

Bonastre, J.-F., Wils, F., and Meignier, S. (2005). Alize, a free toolkit for speaker recognition. In *ICASSP (1)*, pages 737–740.

Bonastre, J.-F., Scheffer, N., Matrouf, D., Fredouille, C., Larcher, A., Preti, A., Pouchoulin, G., Evans, N. W., Fauve, B. G., and Mason, J. S. (2008). Alize/spkdet: a state-of-the-art open source software for speaker recognition. In *Odyssey*, page 20.

Brümmer, N. and Du Preez, J. (2006). Application-independent evaluation of speaker detection. *Computer Speech & Language*, 20(2):230–275.

Brummer, N. (2007). Focal toolkit. *Available in http://www. dsp. sun. ac. za/nbrummer/focal*.

Castro, D. R. (2007). *Forensic evaluation of the evidence using automatic speaker recognition systems*. Ph.D. thesis, Universidad autónoma de Madrid.

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

Doddington, G., Liggett, W., Martin, A., Przybocki, M., and Reynolds, D. (1998). Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation. Technical report, DTIC Document.

Galibert, O. and Kahn, J. (2013). The first official repere evaluation. In *SLAM@ INTERSPEECH*, pages 43–48.

Galliano, S., Geoffrois, E., Gravier, G., Bonastre, J.-F., Mostefa, D., and Choukri, K. (2006). Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news. In *Proceedings of LREC*, volume 6, pages 315–320.

Galliano, S., Gravier, G., and Chaubard, L. (2009). The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. In *Interspeech*, volume 9, pages 2583–2586.

Gonzalez-Rodriguez, J. and Ramos, D. (2007). Forensic automatic speaker classification in the "coming paradigm shift". In *Speaker Classification I*, pages 205–217. Springer.

Gravier, G., Adda, G., Paulson, N., Carré, M., Giraudel, A., and 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*, page na.

Kahn, J., Audibert, N., Rossato, S., and Bonastre, J.-F. (2010). Intra-speaker variability effects on speaker verification performance. In *Odyssey*, page 21.

Kahn, J., Galibert, O., Quintard, L., Carré, M., Giraudel, A., and Joly, P. (2012). A presentation of the repere challenge. In *Content-Based Multimedia Indexing (CBMI), 2012 10th International Workshop on*, pages 1–6. IEEE.

Larcher, A., Bonastre, J.-F., Fauve, B. G., Lee, K.-A., Lévy, C., Li, H., Mason, J. S., and Parfait, J.-Y. (2013). Alize 3.0-open source toolkit for state-of-the-art speaker recognition. In *INTERSPEECH*, pages 2768–2772.

Linares, G., Nocéra, P., Massonie, D., and Matrouf, D. (2007). The lia speech recognition system: from 10xrt to 1xrt. In *Text, Speech and Dialogue*, pages 302–308. Springer.

Matrouf, D., Scheffer, N., Fauve, B. G., and Bonastre, J.-F. (2007). A straightforward and efficient implementation of the factor analysis model for speaker verification. In *INTERSPEECH*, pages 1242–1245.

Matveev, Y. (2013). The problem of voice template aging in speaker recognition systems. In *Speech and Computer*, pages 345–353. Springer.

Morrison, G., Zhang, C., Enzinger, E., Ochoa, F., Bleach, D., Johnson, M., Folkes, B., De Souza, S., Cummins, N., and Chow, D. (2015). Forensic database of voice recordings of 500+ australian english speakers. *URL:) http://databases. forensic-voice-comparison. net*.

Morrison, G. S. (2009). Forensic voice comparison and the paradigm shift. *Science & Justice*, 49(4):298–308.

Prince, S. J. and Elder, J. H. (2007). Probabilistic linear discriminant analysis for inferences about identity. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE.

Ramos, D., Gonzalez-Rodriguez, J., Gonzalez-Dominguez, J., and Lucena-Molina, J. J. (2008). Addressing database mismatch in forensic speaker recognition with

ahumada iii: a public real-casework database in spanish. In *Interspeech*. International Speech Communication Association.

van der Vloed, D., Bouten, J., and van Leeuwen, D. A. (2014). Nfi-frits: A forensic speaker recognition database and some first experiments.