INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Homogeneity Measure Impact on Target and Non-target Trials in Forensic Voice Comparison

Moez Ajili[1], *Jean-François Bonastre*[1], *Waad Ben Kheder*[1], *Solange Rossato*[2], *Juliette Kahn*[3]

[1]University of Avignon, LIA-CERI, Avignon, France
[2]University of Grenoble, LIG , Grenoble, France
[3]Laboratoire Nationale de métrologie et d'Essai, LNE, Paris

Abstract

It is common to see mobile recordings being presented as a forensic trace in a court. In such cases, a forensic expert is asked to analyze both suspect and criminal's voice samples in order to determine the strength-of-evidence. This process is known as *Forensic Voice Comparison* (FVC). The *Likelihood ratio* (LR) framework is commonly used by the experts and quite often required by the expert's associations "best practice guides". Nevertheless, the LR accepts some practical limitations due both to intrinsic aspects of its estimation process and the information used during the FVC process. These aspects are embedded in a more general one, the lack of knowledge on FVC reliability. The question of reliability remains a major challenge, particularly for FVC systems where numerous variation factors like duration, noise, linguistic content or... within-speaker variability are not taken into account. Recently, we proposed an information theory-based criterion able to estimate one of these factors, the homogeneity of information between the two sides of a FVC trial. Thanks to this new criterion, we wish to explore new aspects of homogeneity in this article. We wish to question the impact of homogeneity on reliability separately on target and non-target trials. The study is performed using FABIOLE, a publicly available database dedicated to this kind of studies with a large number of recordings per target speaker. Our experiments report large differences of homogeneity impact between FVC genuine and impostor trials. These results show clearly the importance of intra-speaker variability effects in FVC reliability estimation. This study confirms also the interest of homogeneity measure for FVC reliability.

Index Terms:Forensic voice comparison, homogeneity, intra-speaker variability, reliability, speaker recognition.

1. Introduction

Forensic voice comparison (FVC) is based on the comparison of a recording of an unknown criminal's voice (the question piece or trace) and a recording of a known suspect's voice (the comparison piece). It aims to indicate whether the evidence supports the prosecution (H_p, the two speech excerpts are pronounced by the same speaker) or the defence (H_d, the two speech excerpts are pronounced by two different speakers) hypotheses. In FVC, as well as in several other forensic disciplines, the Bayesian paradigm is recognized as the logical and theoretically sounded framework to model forensic problems [1, 2, 3]. In this framework, the *likelihood ratio* (LR) is used to present the results of the forensic expertise. The LR not only supports one of the hypotheses but also quantifies the strength of its support. The LR is calculated using Equation 1.

$$LR = \frac{p(E/H_p)}{p(E/H_d)} \quad (1)$$

where E is the trace, H_p is the prosecution hypothesis (same origin), and H_d is the defence hypothesis (different origins). The LR's numerator corresponds to a numerical statement about the degree of similarity of the evidence with respect to the suspect and the denominator to a numerical statement about the degree of typicality with respect to the relevant population [4].

Theoretically, the LR provides a founded value of the relative strength of its support to the prosecutor or the defender hypothesis and thus, is self-sufficient and does not need any confidence measure to take into account characteristics of a specific voice comparison trial. In practice the situation is different: the LR is approximated using a specific process and this process accepts some limitations. It is particularly true when automatic FVC is considered, as the ASR systems are outputting a score in all situations regardless the case specific conditions. Moreover, the ASR FVC systems use different normalization steps to see a score as a LR, including the so-called "calibration". Several variability factors are not taken into account explicitly such as trial conditions (quantity and quality of information involved in both voice recordings), the phonetic content [5] or speaker intrinsic characteristic [6, 7]. Despite the huge impact of intra-speaker variability in ASR system, this factor is still not well addressed in the different evaluation campaigns such as NIST framework due to the low number of available utterances per speaker.

Recently, we introduced FABIOLE [8], a database dedicated to FVC reliability estimation. It provides a large number of recordings per speaker and hence each speaker has a large number of target trials. In [9], we proposed *NHM*, a measure of the homogeneity of the acoustic information between the two voice records of a voice comparison trial. We highlighted a large performance variability depending on the level of homogeneity estimated using NHM.

In this paper, we wish to investigate more deeply the impact of homogeneity on FVC reliability using FABIOLE database. Particularly, we are taking advantage of FABIOLE to highlight NHM behavioral differences when target versus non-target LRs are on spotlights, knowing that misleading non-target LRs do not have the same potential consequences than for target LRs (innocent convicted vs criminal unpunished).

This paper is structured as follows. Section 2 reminds the homogeneity measure. Section 3 describes the experimental protocol, Fabiole database and the baseline ASR system used in this article. Section 4 presents the experimental results. Then, section 5 presents the conclusion and proposes some extends of the current work.

http://dx.doi.org/10.21437/Interspeech.2017-152

2. Information theory based homogeneity measure

In this section, we present briefly *NHM*, the information theory (IT) based acoustic homogeneity measure presented in [9, 10]. Its objective is to calculate the amount of acoustic information that appertains to the same (acoustic) 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 estimated based on the number of acoustic frames of S_A and S_B linked to the same 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: $H(p_m) = -p_m log_2(p_m) - \overline{p}_m log_2(\overline{p}_m)$.

NHM measures the (non normalized) *Bic Entropy Expectation* BEE with respect of the quantity of information present in each acoustic class $\{\gamma(m)\}_{i=1}^{M}$:

$$\begin{aligned} NHM_{BEE} &= \sum_{m=1}^{M} (\gamma_A(m) + \gamma_B(m)) H(p_m) \\ &= \sum_{m=1}^{M} \gamma(m) H(p_m) \end{aligned} \tag{2}$$

3. Experimental protocol

3.1. Corpus

FABIOLE is a speech database created within the ANR-12-BS03-0011 FABIOLE project. The main goal of this database is to investigate the reliability of ASR-based FVC. FABIOLE is primarily designed to allow studies on intra-speaker variability and the other factors are controlled as much as possible: channel variability is reduced as all the excerpts come from French radio or television shows; the recordings are clean in order to decrease noise effects; the duration is controlled with a minimum duration of 30 seconds of speech; gender is "neutralized" by using only recordings from male speakers; and, finally, the number of target and non-target trials per speaker is fixed. FABIOLE database contains 130 male French native speakers divided into two sets:

- Set T: 30 target speakers each associated with 100 recordings.
- Set I: 100 impostor speakers. Each impostor pronounced one recording. These files are used mainly for non-target trials.

FABIOLE allows to propose more than $150,000$ matched pairs (target trials) and more than $4.5millions$ non-matched pairs (non-target trials). In this paper, we use only the T set. The trials are divided into 30 subsets, one for each T speaker. For one subset, the voice comparison pairs are composed of at least one recording pronounced by the corresponding T speaker. It gives for a given subset 294950 pairs of recordings distributed as follows: 4950 same-speaker pairs and $290k$ different-speakers pairs. The target pairs are obtained using all the combinations of the 100 recordings available for the corresponding T speaker (C_{100}^2 targets pairs). Non-targets pairs are obtained by pairing each of the target speaker's recording (100 are available) with each of the recordings of the 29 remaining speakers, forming $100 \times 100 \times 29 = 290k$ non-target pairs.

FABIOLE contains recordings gathered from different kind of speakers, including journalists, announcers, politicians, chroniclers, interviewers, etc. FABIOLE material is close to the one of REPERE [11], ESTER 1, ESTER 2 [12] and ETAPE [13]. This characteristic allows to use these databases as a source of training data. More details can be found in [8].

3.2. Baseline LIA speaker recognition System

In all experiments, we use as baseline the LIA_SpkDet system presented in [14].This system is developed using the ALIZE/SpkDet open-source toolkit [15, 16, 17]. It uses the I-vector approach [18]. Acoustic features are composed of 19 LFCC parameters, its derivatives, and 11 second order derivatives. The bandwidth is restricted to 300-3400 Hz in order to suit better with FVC applications.

The *Universal Background Model* (UBM) has 512 components. 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 the 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 "$3,410$" sessions from "617" speakers. The dimension of the I-Vectors in the total variability space is 400. For scoring, PLDA scoring [19] is applied.

3.3. Evaluation metric

We use the C_{llr} widely used in forensic voice comparison as it is designed to evaluate the LR and is not based on hard decisions like, for example, *equal error rate* (EER) [20, 21, 22, 23]. C_{llr} has the meaning of a cost or a loss: the lower the C_{llr} is, better is the performance. C$_{llr}$ can be calculated as follows:

$$C_{llr} = \underbrace{\frac{1}{2N_{tar}} \sum_{LR \in \chi_{tar}} \log_2\left(1 + \frac{1}{LR}\right)}_{C_{llr}^{TAR}} + \underbrace{\frac{1}{2N_{non}} \sum_{LR \in \chi_{non}} \log_2(1 + LR)}_{C_{llr}^{NON}} \tag{3}$$

As shown in Equation 3, C$_{llr}$ can be decomposed into the sum of two parts:

- C_{llr}^{TAR}, which is the average information loss related to target trials.
- C$_{llr}^{NON}$, which is the average information loss related to non-target trials.

This decomposition allows to quantify the information loss relative to both kind of comparison. In this paper, we use an affine calibration transformation [24] estimated using all the trial subsets (*pooled condition*) using FoCal Toolkit [25].

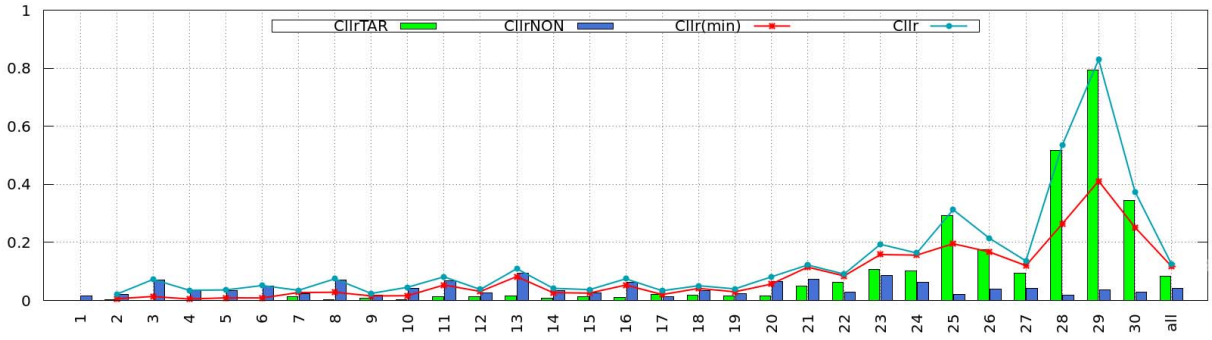

Figure 1: C_{llr}, C_{llr}^{min}, C_{llr}^{TAR}, C_{llr}^{NON} *per speaker and for "all" (data from all the speakers are pooled together).*

C_{llr} can also be seen as a sum of two errors: error relative to the calibration process, denoted C_{llr}^{cal}, and error relative to the discrimination power of the system, denoted C_{llr}^{min}. The quality of the calibration (i.e., the mapping from score to log-likelihood-ratio) could be evaluated by calculating C_{llr}^{cal} given by the following Equation:

$$C_{llr}^{cal} = C_{llr} - C_{llr}^{min}. \quad (4)$$

A system is deemed well-calibrated when it has a low mis-calibration cost and is, therefore, able to provide more reliable likelihood ratio values.

4. Results

The global C_{llr} (computed using all the trial subsets put together) is equal to $0.12631 bits$ and the corresponding global EER is 2.88%. The performance level is similar to the level showed during the large evaluation campaigns (like the NIST's ones). This global representation hides, among other things, the information loss related to target and non-target trials.

4.1. Information loss for target and non-target trials

In order to highlight variability of information loss between target and non-target trials, we present, in Figure 1, C_{llr} estimated individually for each T speaker subset (the result are presented following the same ranking as done in [6]). In this figure, C_{llr} is divided into C_{llr}^{TAR} and C_{llr}^{NON} as shown in Equation 3. A deeper look at the relative weight of target and non-target trials in the global C_{llr} shows that target trials bring in general about two third of C_{llr} loss (0.67 vs 0.33). This proportion is significantly higher (up to 0.94 vs 0.06) for the speakers who present the largest C_{llr} loss contribution. Results also show that information loss related to non-target trials (measured by C_{llr}^{NON}) presents a quite small variation regarding speakers while there is a huge inter-speaker variation of the information loss related to target trials (measured by C_{llr}^{TAR}).

4.2. HM impact on target and non-targets trials

Figure 2 presents for all the trials, the homogeneity measure, NHM, in function of C_{llr} (as well as C_{llr}^{min} and C_{llr}^{cal}). In order to compute the C_{llr} corresponding to a given NHM value, we apply on the trials sorted by homogeneity values a sliding window containing $1/10$ of all trials, moved using a step of N values (N is equal to the size of the window). On each window, we compute the averaged C_{llr} to be compared with the NHM value, computed here as the median value on the window.

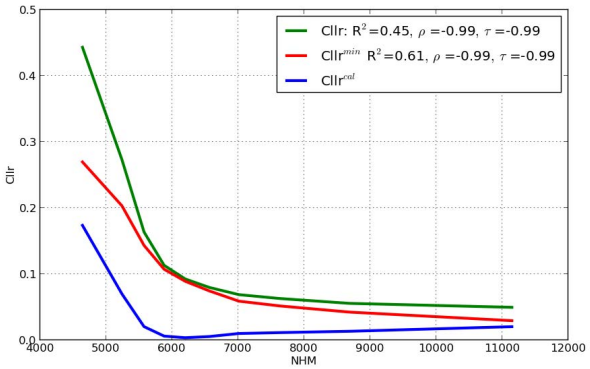

Figure 2: *NHM behavioral curve for the pooled condition (all the comparison tests taken together)*

The shape of the curve brings interesting comments. The C_{llr} is decreasing in function of *NHM* with a quite consistent evolution from (*NHM*=4650, C_{llr}=0.44) to (*NHM*=11140, C_{llr}=0.04bits). A large significant correlation between the homogeneity values *NHM* and the C_{llr} is observed, confirmed by a large correlation coefficients, Spearman ρ=-0.99, Kendall's τ=-0.99 and a p-value < 0.001 (R^2 is also provided for comparison purpose).

Figure 2, shows also that the calibration error, measured by C_{llr}^{cal}, decreases in function of *NHM* and reaches its minimum value when *NHM* is about 6000. For larger NHM, C_{llr}^{cal} shows a tiny degradation remaining quite low (C_{llr}^{cal} does not exceed 0.02bits). It is important to note that the largest calibration error, C_{llr}^{cal}=0.16, is observed at the lowest homogeneity value.

NHM appears to be able to predict correctly the performance class in terms of C_{llr} using only the two speech recordings of a voice comparison trial.

In Figure 2, the impact of homogeneity on genuine and impostor LRs is not visible. To investigate this point, we present in Figure 3 a scatter point corresponding to (NHM,C_{llr}) as well as box-plot for better visualization. The vizualization is proposed separately for target (a and c) and non-target (b and d) trials. Several important comments could be extracted from Figure 3 :

- For target trials, it is interesting to notice that both C_{llr} average values and standard deviation are decreasing when the homogeneity is increasing. The general shape of the curve is

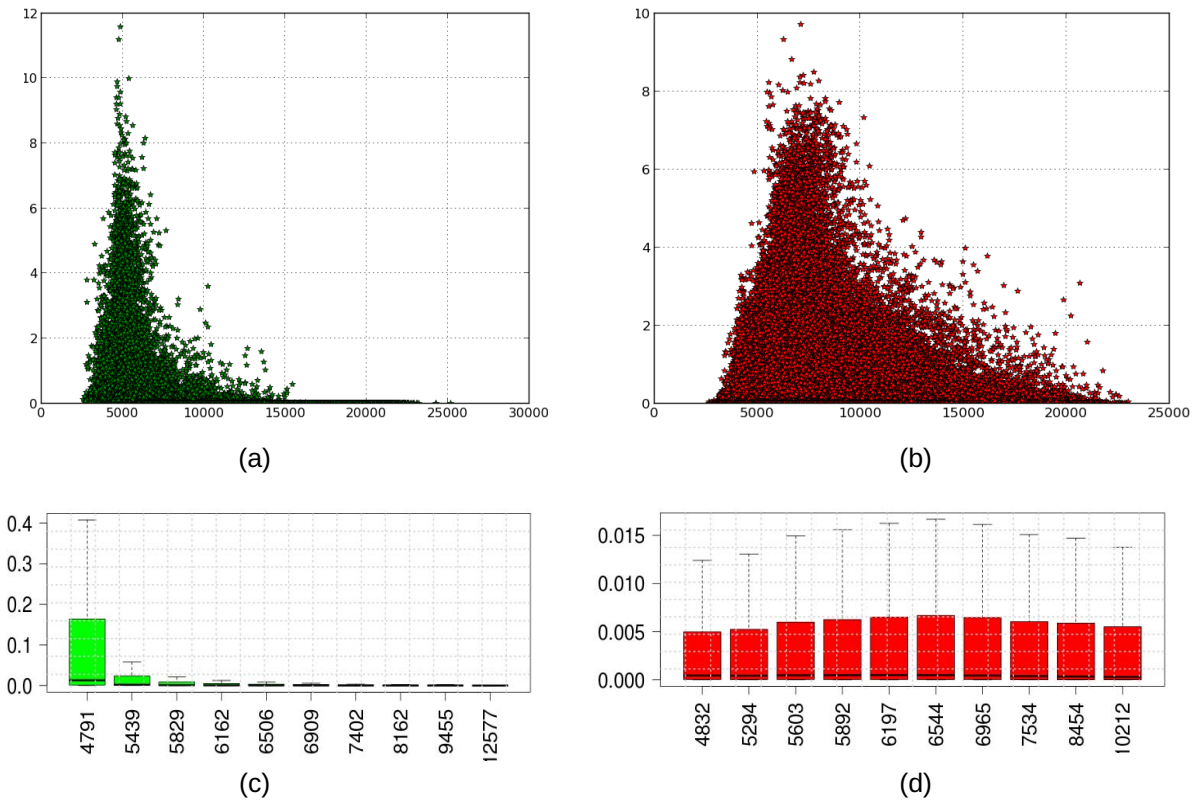

Figure 3: *C_{llr} against homogeneity criterion. (a)/(c) are respectively a scatter plot and box-plot (10 bins without outliers) of Cllr against NHM for target trials. (b)/(d) present the same information for non-target trials.*

clearly exponential. Indeed, C_{llr} average value and standard deviation ($\overline{C_{llr}}$, SD) vary quite consistent between (0.406, 1.06) and (0.004, 0.04). Starting on "bin 4", C_{llr} is becoming infinitesimal as *NHM* increases. This observation indicates that the LR accuracy and reliability are directly linked to the acoustic homogeneity between the two files of a voice comparison trial.

• For non-target trials, the situation is less easy to interpret. First, the right part of the plot shows a behavior comparable to the previous case: after a specific value, the C_{llr} is decreasing when *NHM* is increasing. But the left part of the curve doesn't show at all this direct relation between NHM and C_{llr}. Interestingly, these homogeneity values correspond to the largest calibration losses shown in Figure 2. For non-target trials, if homogeneity is still a required parameter, it seems that it is not a sufficient criterion to predict the reliability of a voice comparison trial. We could hypothesize that the presence of the same acoustic criterion in both audio files is needed in order to separate two speakers, but is not sufficient: cues able to discriminate these two speakers are also needed. This hypothesis is corroborated by [26] where we showed that the phonological information used to discriminate a pair of speakers depends on the speakers themselves.

5. Conclusion

This work is focused on forensic context and investigates differences in reliability between target and non-target LRs. This work took advantage of FABIOLE, a database designed for this purpose.

We showed firstly that the target trials bring about two third of C_{llr} loss compared to non-target trials (0.67 vs. 0.33). This proportion is significantly higher for some speakers (up to 0.94 vs 0.06). These results suggest strongly the presence of a high intra-speaker variability effect in FVC. This factor should be taken into account in reliability evaluation.

In a second step, we investigate the relations between the acoustic homogeneity measured using NHM and C_{llr}. NHM confirmed its ability to predict the C_{llr} class based only on the two speech recordings of a given voice comparison trial. The impact of homogeneity was also examined separately on genuine and impostor trials. For target trials, we found that a good NHM is directly linked to a low level of C_{llr} loss. It says that an acceptable amount of homogeneous acoustic information is enough to authorize the system to evaluate if the two files come from the same source. For the 70% highest NHM, the corresponding C_{llr} are $\approx$0. For non-target trials, the same behavior than for target trials is observed for the 70% highest NHM. For the 30% lowest NHM, there is no clear link between NHM and C_{llr}. We could hypothesized that for non-target trials having common acoustic material is not enough and the presence of adequate cues for the in-interest pair of speakers is mandatory. Further work is needed, on a larger number of speakers, in order to precise the scope of this behavior.

6. Acknowledgments

The research reported here was supported by FABIOLE (ANR-12- BS03-0011) and ALFFA (ANR-13-BS02-0009) projects.

7. References

[1] AOFS Providers, "Standards for the formulation of evaluative forensic science expert opinion," *Sci. Justice*, vol. 49, pp. 161–164, 2009.

[2] Christophe Champod and Didier Meuwly, "The inference of identity in forensic speaker recognition," *Speech Communication*, vol. 31, no. 2, pp. 193–203, 2000.

[3] Colin GG Aitken and Franco Taroni, *Statistics and the evaluation of evidence for forensic scientists*, vol. 10, Wiley Online Library, 2004.

[4] Andrzej Drygajlo, Michael Jessen, Stefan Gfroerer, Isolde Wagner, Jos Vermeulen, and Tuija Niemi, "Methodological guidelines for best practice in forensic semiautomatic and automatic speaker recognition," *European Network of Forensic Science Institutes*, 2015.

[5] Moez Ajili, Jean-François Bonastre, Waad Ben Kheder, Solange Rossato, and Juliette Kahn, "Phonetic content impact on forensic voice comparison," in *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, 2016, pp. 210–217.

[6] Moez Ajili, Jean-Fran¸cois Bonastre, Solange Rossatto, and Juliette Kahn, "Inter-speaker variability in forensic voice comparison: a preliminary evaluation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 2114–2118.

[7] George Doddington, Walter Liggett, Alvin Martin, Mark Przybocki, and Douglas Reynolds, "Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation," Tech. Rep., DTIC Document, 1998.

[8] Moez Ajili, Jean-François Bonastre, Solange Rossato, Juliette Kahn, and Guillaume Bernard, "Fabiole, a speech database for forensic speaker comparison," *International Conference on Language Resources, Evaluation and Corpora*, 2016.

[9] Moez Ajili, Jean-François Bonastre, Solange Rossato, Juliette Kahn, and Itshak Lapidot, "An information theory based data-homogeneity measure for voice comparison," in *Interspeech 2015*, 2015.

[10] Moez Ajili, Jean-François Bonastre, Solange Rossato, Juliette Kahn, and Itshak Lapidot, "Homogeneity measure for forensic voice comparison: A step forward reliability," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pp. 135–142. Springer, 2015.

[11] Aude Giraudel, Matthieu Carré, Valérie Mapelli, Juliette Kahn, Olivier Galibert, and Ludovic Quintard, "The repere corpus: a multimodal corpus for person recognition.," in *LREC*, 2012, pp. 1102–1107.

[12] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J-F. Bonastre, and G. Gravier, "The ester phase ii evaluation campaign for the rich transcription of french broadcast news," in *European Conference on Speech Communication and Technology*, 2005, pp. 1149–1152.

[13] Guillaume Gravier, Gilles Adda, Niklas Paulson, Matthieu Carré, Aude Giraudel, Olivier Galibert, et al., "The etape corpus for the evaluation of speech-based tv content processing in the french language," *International Conference on Language Resources, Evaluation and Corpora*, 2012.

[14] Driss Matrouf, Nicolas Scheffer, Benoit GB Fauve, and Jean-François Bonastre, "A straightforward and efficient implementation of the factor analysis model for speaker verification.," in *INTERSPEECH*, 2007, pp. 1242–1245.

[15] Jean-François Bonastre, Frédéric Wils, and Sylvain Meignier, "Alize, a free toolkit for speaker recognition.," in *ICASSP (1)*, 2005, pp. 737–740.

[16] Jean-François Bonastre, Nicolas Scheffer, Driss Matrouf, Corinne Fredouille, Anthony Larcher, Alexandre Preti, Gilles Pouchoulin, Nicholas WD Evans, Benoit GB Fauve, and John SD Mason, "Alize/spkdet: a state-of-the-art open source software for speaker recognition.," in *Odyssey*, 2008, p. 20.

[17] Anthony Larcher, Jean-François Bonastre, Benoit GB Fauve, Kong-Aik Lee, Christophe Lévy, Haizhou Li, John SD Mason, and Jean-Yves Parfait, "Alize 3.0-open source toolkit for state-of-the-art speaker recognition.," in *INTERSPEECH*, 2013, pp. 2768–2772.

[18] Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

[19] Simon JD Prince and James H Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.

[20] Geoffrey Stewart Morrison, "Forensic voice comparison and the paradigm shift," *Science & Justice*, vol. 49, no. 4, pp. 298–308, 2009.

[21] Niko Brümmer and Johan du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 2, pp. 230–275, 2006.

[22] Daniel Ramos Castro, *Forensic evaluation of the evidence using automatic speaker recognition systems*, Ph.D. thesis, Universidad autónoma de Madrid, 2007.

[23] Joaquin Gonzalez-Rodriguez and Daniel Ramos, "Forensic automatic speaker classification in the coming paradigm shift," in *Speaker Classification I*, pp. 205–217. Springer, 2007.

[24] Niko Brümmer, Lukáš Burget, Jan Honza Černockỳ, Ondřej Glembek, František Grezl, Martin Karafiat, David A Van Leeuwen, Pavel Matě, Petr Schwarz, and Albert Strasheim, "Fusion of heterogeneous speaker recognition systems in the stbu submission for the nist speaker recognition evaluation 2006," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 7, pp. 2072–2084, 2007.

[25] Niko Brummer, "Focal toolkit," *Available in http://www. dsp. sun. ac. za/nbrummer/focal*, 2007.

[26] Moez Ajili, Jean-François Bonastre, Waad Ben Kheder, Solange Rossato, and Juliette Kahn, "Phonological content impact on wrongful convictions in forensic voice comparison context," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017.