PHONETIC CONTENT IMPACT ON 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 metreologie et d'essai, LNE, Paris

ABSTRACT

Forensic Voice Comparison (FVC) is increasingly using the *likelihood ratio* (LR) in order to indicate whether the evidence supports the prosecution (same-speaker) or defender (different-speakers) hypotheses. In addition to support one hypothesis, the LR provides a theoretically founded estimate of the relative strength of its support. Despite this nice theoretical aspect, the LR accepts some practical limitations due both to its estimation process itself and to a lack of knowledge about the reliability of this (practical) estimation process. In a large set of situations, a lack in reliability at the estimation process level potentially destroys the reliability of the resulting LR. It is particularly true when automatic FVC is considered, as *Automatic Speaker Recognition* (ASpR) systems are outputting a score in all situations regardless of the case specific conditions. Furthermore, ASpR systems use different normalization steps to see their scores as LR and these normalization steps are potential sources of bias. In the LR estimation done by ASpR systems, different factors are not taken into account such as the amount of information involved in the comparison, the phonemic content and finally the speaker intrinsic characteristics, denoted here "speaker factor". Consequently, a more complete view of reliability seems to be a mandatory point for FVC, even if a LR-like approach is used. This article focuses on the impact of phonemic content on FVC performance and variability. The experimental part is using FABIOLE database. This database is dedicated to this kind of studies and allows to examine both inter-speaker variability and intra-speaker variability. The results demonstrate the importance of the phonemic content and highlight interesting differences between inter-speakers effects and intra-speaker's ones.

Index Terms**—** Forensic voice comparison, phonemic category, reliability, speaker factor, speaker recognition.

1. INTRODUCTION

Forensic voice comparison (FVC) is based on the comparison of a recording of an unknown criminal's voice (the evidence or trace) and a recording of a known suspect's voice (the comparison piece). It aims to indicate whether the evidence supports the prosecution (the two speech excerpts are pronounced by the same speaker) or defender (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 denoted as the logical and theoretically sounded framework to model and represent forensic evidence reports [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 hypothesis but also quantifies the strength of its support. The LR is calculated using Equation 1.

$$LR = \frac{p(E/H_{ph})}{p(E/H_{dh})} \tag{1}$$

where E is the trace, H_{ph} is the prosecutor hypothesis (same origin), and H_{dh} is the defender 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. Automatic Speaker Recognition (ASpR) is considered as one of the most appropriate solution when LR framework is involved [4]. Even though ASpR systems have achieved significant progresses in the past two decades and have reached impressive low error rates ($\approx 1\%$ [5, 6, 7]), the forensic scenario is still a very challenging one for ASpR for several reasons:

- Trial conditions; The speech recordings may be recorded in different situations and at least one situation is partially or completely unknown (the trace recording situation). The speakers are not necessarily cooperative and may disguise their voices, with consequences on performance [8]. A speaker could also be ill, or under the influence of stress, alcohol or other factors. The social and linguistic environment of the unknown speaker is unknown by construction (so, for example, an unknown mother or second language should be taken into account by the forensic experts). The speech samples will most likely contain noise, may be very short, their content can't be controlled (at least for the trace) and may not contain enough relevant information for comparative purposes. In their "need for caution 2009 paper" [9] (which inspired a large part of this paragraph) the authors said in the conclusion: "Each of these variables, in addition to the known variability of speech in general, makes reliable discrimination of speakers a complicated and daunting task". This sentence remains in 2016 a nice synthesis of FVC's challenging aspects (for FVC in general and not only for ASpR-based FVC).

- Real-life LR approximation/estimation processes; As said before, the LR provides a theoretically founded value of the relative strength of its support to the prosecutor or the defender hypothesis. So, it appears to be self-sufficient and does not need any confidence measure or confidence interval to take into account the characteristics of a specific voice comparison trial. But, in real world, the LR is *approximated* by an automatic process and, despite its nice theoretical aspects, will accept some limitations coming from imperfections of its estimation process. It is particularly true when automatic FVC is considered, as the ASpR systems are outputting a *score* and use different normalization steps to *see* this score as a LR. The main normalization process is the so-called "calibration" process [10, 11, 12, 13, 14, 15, 16]. Several other normalization steps are used, like at the acoustic parameterization level [17, 18] , at the iVector level [19, 6, 20] or at the score level [21, 22]. A "reference population" is also often used to evaluate the "typicality" [23]. A large majority of the involved normalization approaches are based on training data and represent a potential source of biases as a mismatch between the training speech material and a given forensic trial could be large. Moreover, the amount of mismatch is often unknown as the trial conditions could be partially or largely unknown and as the training conditions are not always well defined.

- Limits of the performance evaluation; If the desire to use ASpR in FVC is not novel, due to intrinsic interests of automatic processes in this field, this desire has increased significantly during the past years as the performance level reached by speaker recognition systems has become very attractive. The performance is measured thanks to international evaluation campaigns like NIST-SRE's ones [24, 25]. If the pros of such evaluation campaigns are well established, several research works emphasized the limits of the underlined evaluation protocols [26, 27] or [28, 29]. Moreover, the classical evaluation criterion and protocols used in ASpR are not designed for FVC. EER and DCF are mainly used, which are based on hard score decisions and not on LR's reliability (even if, more recently, a more adapted criteria, CLLR [30] is also used). The protocols focus on global performance using a brute-force strategy and take into account the averaged behavior of ASpR systems. In the same time, they ignore many sensitive cases which represent several distinct specific situations where the ASpR systems show a specific behavior due, for example, to the recording conditions, the noises, the content of the recordings or the speakers themselves (and the evaluation databases are still missing a lot of variation factors).

- A lack in recording content analysis; In state-of-the-art ASpR systems, for example IVector(IV)-based ones, a recording is encoded by one low dimensional vector. The phonemic content of a recording is not used explicitly, as well as the presence or absence of different speaker-specific cues. However several research works like [31, 32, 33, 34] agree that speaker specific information is not equally distributed on the speech signal and particularly depends on the phoneme distribution. In [29, 35] the authors showed that homogeneity of the speaker-specific information between the two recordings of a voice comparison trial is also playing an important role and should not be ignored by the LR estimation process.

Despite its apparent richness, the above literature review reveals different lacks. First, the majority of the quoted research works are dedicated to ASpR and do not take into account the specific context of FVC. Second, they do not take into account correctly intra-speaker variability, mainly due to a lack of the used databases in terms of number of recording per speaker.

This paper is dedicated to answer to a part of the highlighted lacks. It investigates the impact of phonetic content on voice comparison process and more precisely at the phonemic level. We propose to analyze whether certain classes of phonemes are bringing more speaker discrimination information than others and if these differences are stable among the speakers. We wish also to analyze deeply how phoneme categories affect both intra- and inter-speaker variability. This work is using FABIOLE database [36], a database dedicated to study intra-speaker variability.

This paper is structured as follows. Section 2 presents a review on phonemic content influence on speaker discrimination and propose a phoneme classification in order to study the impact of phonemic content on voice comparison. Section 3 is dedicated to the experimental protocol. Then, section 4 shows experiments and results. Section 5 concludes the paper and discusses future plans.

2. PHONEMIC CONTENT AND SPEAKER DISCRIMINATION

If everybody agrees on the fact that voice signal is conveying information on the speaker, including speaker's identity, it is less easy to list the different cues which embed this aspect (this is true for both human perception and automatic systems). In this research work, we do not wish to answer to this question but we propose to use an ASpR system in order to investigate the links between phonological content and speaker discrimination abilities.

2.1. A review of literature

Several earlier studies have analyzed the speaker-discriminant properties of individual phonemes or of phoneme classes [37, 38, 39]. The authors agreed that vowels and nasals provide the best discrimination between speakers. [40] presents a ranking of 24 isolated German phonemes, which indicates nasals as providing the best performance, with the voiced alveolar fricative /z/ and the voiced uvular fricative /ʁ/ also performing fairly well. In [41], /s/, /t/ and /b/ are found to perform worse than vowels and nasals. [37, 38] strongly promote the nasals and vowels as best performers. The influence of the phonemic content of both voice recordings was also evaluated in [31] in which authors suggest that glides and liquids together, vowels -and more particularly nasal vowels- and nasal consonants contain more speaker-specific information than phonemically balanced speech utterances. According to [33, 42, 34, 39], nasals and vowels were found to be particularly speaker specific information and nasal vowels are more discriminant than oral vowels. Finally, [32] and, more recently, [43], show that some frequency sub-bands seem to be more relevant to characterize speakers than some others.

It appears clearly from this literature survey that the phonemic content has an impact on speaker recognition performance and that it seems possible to rank the phoneme depending on their abilities in terms of speaker discrimination. It is important to remind that we discuss here results obtained using an ASpR system as a measurement instrument. We are not able to discriminate between the intrinsic characteristics of a cue and the way that this cue is taken into account by an ASpR system.

2.2. Phoneme classification

To conduct our work, we propose to use phoneme classes in place of individual phonemes. Working on phoneme classes presents two main advantages in the context of our study. First, to study the effect of phonemic content, a phoneme transcription/alignment process is mandatory. If the classification is well chosen, the use of phoneme classes will allow to reduce the effect of potential errors done at the transcription level. Second, the speech extracts involved in FVC trials are usually of a relatively short duration. To work at phoneme level presents a risk of piecemeal or inconsistent results, due to insufficient amount of speech material for some phonemes. Working with a short set of phoneme classes will allow to overcome this risk. In this work, We propose to classify the speech content into 6 phoneme categories based on phonological features. The phoneme classification is describe below:

- Oral vowels (OV) which includes /i/, /u/, /y/, /e/, /ɛ/, /ø/, /œ/, /o/, /ɔ/, /ɑ/.
- Nasal vowels (NV) which includes /ɑ̃/, /ɔ̃/, /ũ/, /ɛ̃/.
- Nasal consonants (NC) which includes /m/, /n/.
- Plosive (P) which includes /p/, /t/, /k/, /b/, /d/, /g/.

- Fricatives (F) which includes /f/, /s/, /ʃ/, /v/, /z/, /ʒ/.
- Liquids (L) which includes /l/, /ɹ/[1].

This phoneme classification will be adopted in all experiments in this paper.

3. EXPERIMENTAL PROTOCOL

In order to investigate the phonemic content impact on voice comparison, we conduct several experiments. This section presents firstly the database used, FABIOLE. The rest of the section is dedicated to the methodology retained to evaluate the impact of the phonemic content on FVC.

3.1. Corpus

FABIOLE is a speech database created inside the ANR-12-BS03-0011 FABIOLE project. The main goal of this database is to investigate the reliability of ASpR-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 "controlled" by using only recordings from male speakers; and, 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 each associated with at least 100 recordings.
- Set I: 100 impostor speakers. Each impostor pronounced one recording. These files are used mainly for non-targets trials.

FABIOLE allows to organize more than $150,000$ matched pairs (target trials) and more than $4.5M$ 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 with 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^2_{100} targets pairs). Whereas, 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 consequently $(100 \times 100 \times 29 = 290k)$ non-targets pairs.

FABIOLE contains recordings gathered from different kinds of speakers, including journalists, announcers, politicians, chroniclers, interviewers, etc. FABIOLE material is close to the one of REPERE [44], ESTER 1, ESTER 2 [45] and ETAPE [46]. This characteristic allows to use these databases as a source of training data. More details could be found in [36].

3.2. Evaluation metric

We use the C_{llr} and the minimum value of the C_{llr}, denoted C_{llr}^{min}, largely used in forensic voice comparison as they wish to evaluate the LR and are not based on hard decisions like, for example, *equal error rate* (EER) [47, 30, 48, 13]. C_{llr} has the meaning of a cost or a loss: lower the C_{llr} is, better is the performance. C_{llr} could be calculated as follows:

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

As shown in Equation 2, 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.

Within our experimental conditions, we make the hypothesis that C_{llr}^{TAR} reflects mainly *intra-speaker variability*. We base this statement as it appears in [49] that the between-speakers differences in terms of target-score distributions are mainly observed on the standard deviation values while the corresponding means are homogeneous. Respectively, we assume that C_{llr}^{NON} is mainly linked to *inter-speakers variability*. In commercial ASpR applications, the first component will give an idea of the risk to see an impostor spoofing the authentication system and the second component will express the risk to reject a client.

In this paper, we use an affine calibration transformation [50] estimated using all the trial subsets (*pooled condition*) using FoCal Toolkit [11].

3.3. Phoneme filtering protocol for data selection

In order to study the influence of a specific phonemic class (detailed in Subsection 2.2), we use a knock-out strategy: the in-interest information is withdrawn from the trials and the amount of performance loss indicates the influence of the corresponding speech material. So, we perform several experiments where the speech material corresponding to a given class is removed from the two speech recordings of each trials. This condition is denoted here "**Specific**". Since the amount of speech material is largely unbalanced (for example, in our experiments, nasal consonants represent 6% of the speech material and oral vowels 36%), in order to avoid a potential bias, we create a control condition denoted "**Random**", where the corresponding amount of speech material is randomly withdrawn. More precisely, for each speech signal, when a certain percentage of speech frames is withdrawn for the "**Specific**" condition, the same percentage of frames is randomly withdrawn for the "**Random**" condition. This process is repeated 20 times, creating 20 times more trials in "**Random**" condition than in "**Specific**" one.

The impact of a specific phonemic class is quantified by estimating the relative C_{llr}^{R} given by Equation 3.

$$C_{llr}^{R} = \frac{Cllr^{random} - Cllr^{specific}}{Cllr^{random}} \times 100\% \tag{3}$$

A positive value of C_{llr}^{R} indicates that the speech material related to the corresponding phonemic class brings a larger part of the speaker-discriminant loss than averaged speech material. A negative value says the opposite: the corresponding phonemic class reduces the discriminant loss compared to averaged phonemic content.

3.4. Baseline LIA Systems

3.4.1. LIA speaker recognition system

In all experiments, we use as baseline the LIA_SpkDet system presented in [51].This system is developed using the ALIZE/SpkDet

[1] / ɹ/ is very variable in French, it may be according to the context [ɹ], [ʁ] or [χ]

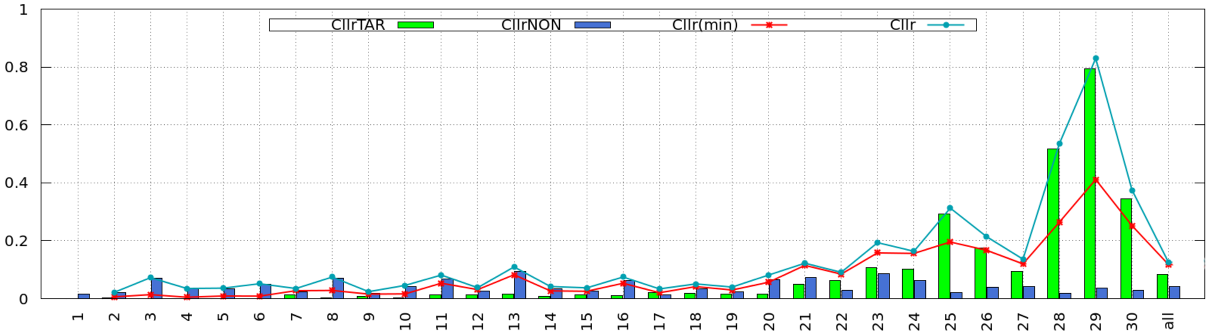

Fig. 1. *C_{llr}, C_{llr}^{min}, C_{llr}^{TAR}, C_{llr}^{NON} per speaker and for "all" (data from all the speakers are pooled together).*

open-source toolkit [52, 53, 54]. It uses I-vector approach [5]. 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 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 factor space is 400. For scoring, PLDA scoring model [55] is applied.

3.4.2. LIA transcription system

FABIOLE database has been automatically transcribed thanks to Speeral, LIA automatic transcription system [56]. This system was used to transcribe REPERE development set (which contains speech recordings close to FABIOLE excerpts) with an overall Word Error Rate of 29$\%$ [57].

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 close to the level showed during the large evaluation campaigns (like the NIST's ones).

4.1. Phonemic content impact on voice comparison

Table 1 shows the impact of the 6 phonemic categories on C$_{llr}$ for "Specific" and "Random" conditions (C$_{llr}^{min}$ results are also provided for comparison purposes). It gives also the amount of speech frames per phoneme class (mean and deviation over the trials). The results are given using the "pooled" condition (averaged on all the speakers). A large variation is observed between the phonemic classes: the withdrawal of nasal vowels, nasal consonants or oral vowels leads to a loss of information compared to the "**Random**" case while the absence of plosive, liquid and fricative does not seem to have an influence on the system accuracy or leads to a small improvement of the performance.

This outcome corroborates results of [33, 34, 42, 39], where nasals and vowels are found to be particularly speaker specific information and more precisely nasal vowels to be more informative than oral vowels. The result we obtained for the fricatives -the class exhibits low speaker discriminating properties- is clearly in conflict with [43]'s finding. An explanation could be that [43] uses a wideband while in this paper a narrow band (300-3400Hz) is applied.

Table 1. C$_{llr}$ and C$_{llr}^{min}$ for "Specific" and "Random" conditions (baseline results are: C$_{llr}$=0.126 and C$_{llr}^{min}$=0.117. Mean and SD of the duration per class are provided.

	C$_{llr}$		C$_{llr}^{min}$		**Duration (s)**	
Category	**Withdrawn**		**Withdrawn**		**Mean**	**SD**
	Specific	**Random**	**Specific**	**Random**		
NV	0.14689	0.12941	0.13498	0.11975	3.14	1.56
NC	0.13713	0.12815	0.12728	0.11897	2.05	1.03
OV	0.15396	0.14689	0.14601	0.12819	13.00	5.50
L	0.12966	0.13032	0.12173	0.12029	4.03	1.96
P	0.13278	0.13431	0.12244	0.12228	7.72	3.40
F	0.12703	0.13238	0.12007	0.12135	5.84	2.68

4.2. Speaker factor

Figure 1 presents C_{llr} estimated individually for each T speaker (the results are presented following the same ranking as [49], which was based on general C_{llr} performance). In this figure, C_{llr} is divided into two components, C_{llr}^{TAR} and C_{llr}^{NON}, in order to quantify separately the information loss relative to target and non-target trials. The results 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 variation of the information loss related to target trials (measured by C_{llr}^{TAR}). The information loss coming from target trials (computed by C_{llr}^{TAR}) is mainly responsible of the reported high costs obtained for some speakers.

Figure 2 is a stacked bar chart which shows the contribution of each phonemic class to the C_{llr}^{R}, depending on the speaker. The same general tendency than in table 1 appears clearly: C_{llr}^{R} results for nasal vowels and nasal consonants are negative and indicates that their absence brings generally a degradation of FVC performance. But a large variability depending on the speaker is also present for all the phonemic classes. For example, speaker 2 shows a relative loss of 175$\%$ when oral vowels are withdrawn while speaker 28 shows a relative win of about 40$\%$ in the same situation. Another time, the global tendencies are shadowing potential speaker-specific effects.

These results reinforce the "speaker factor" hypothesis presented in [49], where it is assumed that -in the view of an ASpR-based voice comparison system- all the speakers do not behave the same way in response of similar condition changes: some speakers will be quite

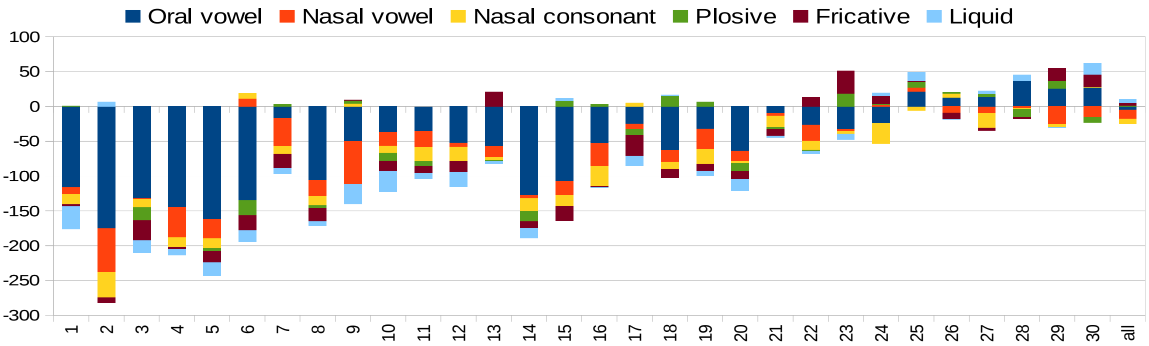

Fig. 2. Stacked bar chart of C^{R}_{llr} (computed on both target- and non-target trials) per speaker and for ”all”.

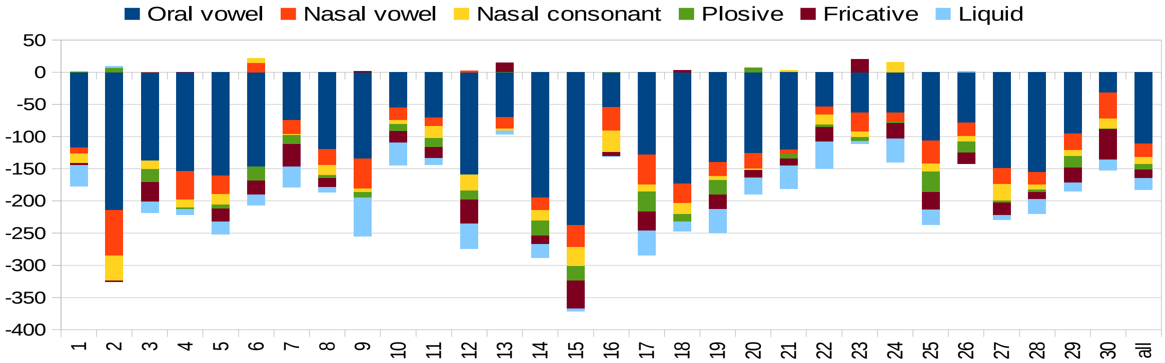

Fig. 3. Stacked bar chart of C^{R}_{llr} computed on C^{NON}_{llr} (non-target trials) per speaker and for ”all”.

robust with limited LR variation (leads to low C_{llr} values) when some other are showing a huge variation (leads to high C_{llr} values).

4.3. Phonemic content impact on intra- and inter-speaker variability

Figure 3 is a stacked bar chart which displays the impact of the phonemic classes per speaker, in terms of relative C^{NON}_{llr} (C^{R}_{llr} computed on C^{NON}_{llr}). C^{NON}_{llr} is computed only on non target trials and is expected to be primarily linked to speaker discrimination power (i.e. inter-speakers variability). The 6 phonemic classes appear to embed speaker discrimination power since their absence leads, in almost all the cases, to a C_{llr} degradation. To withdraw the oral vowels causes the largest accuracy loss, ranking in top this phonemic class in terms of speaker discrimination power with a large margin with the next class. Nasals, vowels first and consonants second, appear to convey the most discrimination power after the oral vowels. Liquid, fricative and plosive obtain similar results, at the end of the speaker discrimination power scale. The results are quite consistent between the 30 target speakers, with limited variations.

Figure 4 uses a form similar to Figure 3. It presents the impact of the phonemic classes per speaker, in terms of relative C^{TAR}_{llr} (C^{R}_{llr} computed on C^{TAR}_{llr}). C^{TAR}_{llr} is computed only using target trials and is expected to be primarily linked to intra-speaker variability. The first outcome differs significantly from the previous case: to withdraw the oral vowels from the recordings leads to an improvement of C_{llr} for about 70% of the speakers: for the target trials, oral vowels are tied with higher $Cllr$. Fricative, liquid and plosive classes have a same behavior than oral vowels. Instead, the nasals (and particularly the nasal vowels) still play a positive role: to withdraw these phonemes increases the C_{llr}. Taken together, the results using relative C^{TAR}_{llr} and relative C^{NON}_{llr} bring to us some remarks :

- The nasal phonemes effectiveness for speaker comparison could be explained by the important contribution of nasal and paranasal cavities. This morphological aspect of these phonemes is something that the speakers cannot unintentionally or voluntarily control by themselves and allows low within-speaker and high between-speaker variability [58, 59].

- A same phonemic class, the oral vowels, brings the largest part in terms of speaker discrimination but presents in the same time a large intra-speaker variability which conveys a significant part of the LR performance loss.

- A deeper look at the relative weight of target and non-target trials in the global C_{llr}, as plotted in Figure 1, shows that intra-speaker variability brings in general about two third of C_{llr} loss (0.66 vs 0.33). This proportion is significantly higher (until 0.94 vs 0.06) for the speakers who present the largest contribution to the C_{llr} loss. It is interesting to link this finding with two facts: almost all studied phoneme classes are helping for speaker discrimination for all the speakers (Figure 3); all the speakers accept some phoneme classes which are degrading the target part of C_{llr} when some other classes are performing well (Figure 4). For the latter remark, it is interesting to remark that the same phoneme class could have a very different

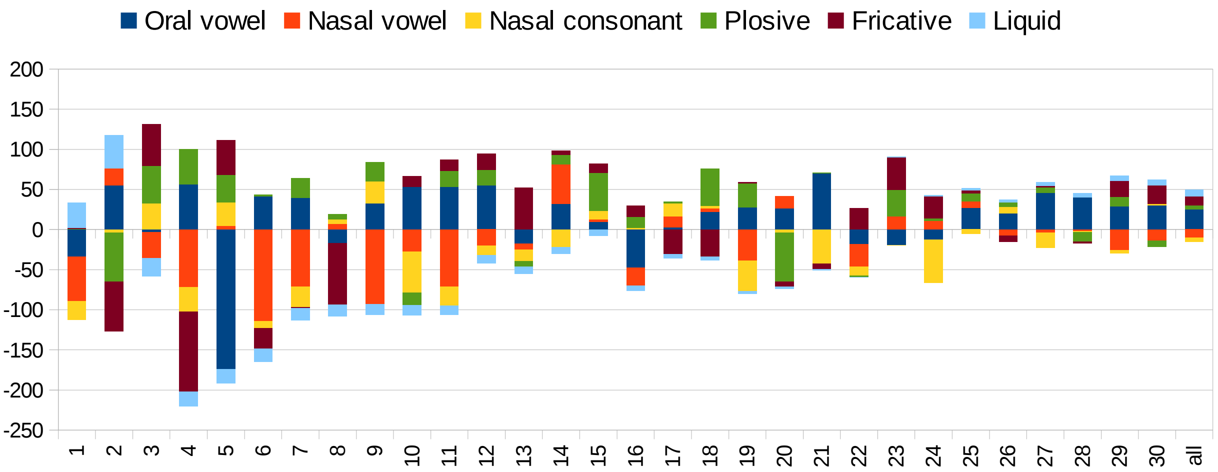

Fig. 4. Stacked bar chart of C_{llr}^{R} computed on C_{llr}^{TAR} (target trials) per speaker and for "all".

behavior depending on the speaker.

5. CONCLUSION

This paper is dedicated to investigate the impact of phonemic content on voice comparison process. It uses an ASpR system as measurement instrument and, more particularly, the C_{llr} variations. We analyzed the influence of 6 phonemic classes: nasal vowel, oral vowel, nasal consonant, fricative, plosive and liquid. The work was done using FABIOLE database, a corpora dedicated to voice comparison reliability experiments with a large number of speech extracts per speaker.

In a first step, we investigated the impact of each phonemic class on voice comparison performance measured with C_{llr}. The results showed that oral vowels, nasal vowels and nasal consonants are better than averaged phonemic content in terms of voice comparison performance, as already seen in the literature. The fricatives do not seem to perform better than an averaged content, which is surprising compared to the literature, but this could be explained by some differences in the experimental choices (like the bandwidth) to better suit the FVC context.

In a second step, we explored intra- and inter-speaker variability aspects by focusing on target and non target parts of C_{llr}. For inter-speaker variability, we showed that all the phonemic content play an important role in terms of speaker discrimination power. The oral vowels are the largest contributors, followed by nasals and liquids and this finding is consistent among most speakers. When we focused on intra-speaker variability, oral vowels appeared to be tied with a high intra-speaker level of C_{llr}. We saw previously that this phonemic class was bringing a large part of the speaker discrimination power but it appears also very sensitive to intra-speaker variability. In contrast, nasals showed a high capacity for speaker discrimination and at the same time appeared to be robust for intra-speaker variability.

In this article, we highlighted at several steps the importance of speaker factor which is a denomination that reflects mainly intra-speaker variability and differences between the speakers according to this variability. It also includes differences linked to the speaker of ASpR systems responses to a same stimulus. We observed large variations of C_{llr} and C_{llr}^{TAR} between our 30 speakers. We also observed large variations per speaker of the system's responses to different phonemic classes, in terms of relative C_{llr}^{TAR}.

As a consequence of these findings, the main takeaway of the presented work is the fact that ASpR usual evaluation protocols are mainly selecting the best features in terms of speaker discrimination (C_{llr}^{NON}) and are largely missing intra-speaker variability when the latter is a key factor for numerous application scenarios. This is particularly true for FVC scenario and it appears mandatory to work more on intra-speaker variability as well as on speaker factor in order to estimate the reliability of a solution in this domain.

The results presented in this article remain preliminary and incomplete. First, if FABIOLE offers an opportunity to open the door for research on intra-speaker variability, this database is still very poor, with only 100 ×30s of speech per speaker, with few contextual variability and with only 30 male speakers. In future work we would like to conduct similar analysis on a 10 times larger database (1000 recordings per speakers and several hundreds of speakers) and to consider individual phonemes instead of phoneme classes.

Finally, this work lets us dreaming about an ASpR FVC system that thoroughly analyzes the phonetic content of the speech extracts and details consequently its outputs in a language understandable by an expert.

6. ACKNOWLEDGMENTS

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

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] Erica Gold and Peter French, "An international investigation of forensic speaker comparison practices," in *Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong, China*, 2011, pp. 1254–1257.

[5] 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.

[6] Pierre-Michel Bousquet, Jean-François Bonastre, and Driss Matrouf, "Exploring some limits of gaussian plda modeling for i-vector distributions," in *Odyssey: The Speaker and Language Recognition Workshop*, 2014.

[7] John HL Hansen and Taufiq Hasan, "Speaker recognition by machines and humans: a tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[8] Sachin S Kajarekar, Harry Bratt, Elizabeth Shriberg, and Rafael De Leon, "A study of intentional voice modifications for evading automatic speaker recognition," in *2006 IEEE Odyssey-The Speaker and Language Recognition Workshop*. IEEE, 2006, pp. 1–6.

[9] Joseph P Campbell, Wade Shen, William M Campbell, Reva Schwartz, J-F Bonastre, and Driss Matrouf, "Forensic speaker recognition," Institute of Electrical and Electronics Engineers, 2009.

[10] Niko Brummer and David A van Leeuwen, "On calibration of language recognition scores," in *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The*. IEEE, 2006, pp. 1–8.

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

[12] Joaquin Gonzalez-Rodriguez, Andrzej Drygajlo, Daniel Ramos-Castro, Marta Garcia-Gomar, and Javier Ortega-Garcia, "Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition," *Computer Speech & Language*, vol. 20, no. 2, pp. 331–355, 2006.

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

[14] Joaquin Gonzalez-Rodriguez, Phil Rose, Daniel Ramos, Doroteo T Toledano, and Javier Ortega-Garcia, "Emulating dna: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 7, pp. 2104–2115, 2007.

[15] Yuko Kinoshita and Shunichi Ishihara, "Background population: how does it affect lr-based forensic voice comparison?," *The International Journal of Speech, Language and the Law*, vol. 21, no. 2, pp. 191–224, 2014.

[16] Andreas Nautsch, Rahim Saeidi, Christian Rathgeb, and Christoph Busch, "Robustness of quality-based score calibration of speaker recognition systems with respect to low-snr and short-duration conditions," Odyssey, 2016.

[17] Jason Pelecanos and Sridha Sridharan, "Feature warping for robust speaker verification," 2001.

[18] Sriram Ganapathy, Jason Pelecanos, and Mohamed Kamal Omar, "Feature normalization for speaker verification in room reverberation," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 4836–4839.

[19] Pierre-Michel Bousquet, Anthony Larcher, Driss Matrouf, Jean-François Bonastre, and Oldrich Plchot, "Variance-spectra based normalization for i-vector standard and probabilistic linear discriminant analysis.," in *Odyssey*, 2012, pp. 157–164.

[20] Daniel Garcia-Romero and Carol Y Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems.," in *Interspeech*, 2011, pp. 249–252.

[21] Roland Auckenthaler, Michael Carey, and Harvey Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1, pp. 42–54, 2000.

[22] Ondrej Glembek, Lukas Burget, Najim Dehak, Niko Brummer, and Patrick Kenny, "Comparison of scoring methods used in speaker recognition with joint factor analysis," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 4057–4060.

[23] 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.

[24] Craig S Greenberg, Alvin F Martin, Bradford N Barr, and George R Doddington, "Report on performance results in the nist 2010 speaker recognition evaluation," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[25] Craig S Greenberg, Vincent M Stanford, Alvin F Martin, Meghana Yadagiri, George R Doddington, John J Godfrey, and Jaime Hernandez-Cordero, "The 2012 nist speaker recognition evaluation.," in *INTERSPEECH*, 2013, pp. 1971–1975.

[26] 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.

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

[28] Juliette Kahn, Nicolas Audibert, Solange Rossato, and Jean-François Bonastre, "Intra-speaker variability effects on speaker verification performance.," in *Odyssey*, 2010, p. 21.

[29] 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.

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

[31] Ivan Magrin-Chagnolleau, Jean-Francois Bonastre, and Frédéric Bimbot, "Effect of utterance duration and phonetic content on speaker identification usind second order statistical methods," in *Proceedings of EUROSPEECH*, 1995.

[32] Laurent Besacier, Jean-François Bonastre, and Corinne Fredouille, "Localization and selection of speaker-specific information with statistical modeling," *Speech Communication*, vol. 31, no. 2, pp. 89–106, 2000.

[33] Kanae Amino, Tsutomu Sugawara, and Takayuki Arai, "Idiosyncrasy of nasal sounds in human speaker identification and their acoustic properties," *Acoustical science and technology*, vol. 27, no. 4, pp. 233–235, 2006.

[34] Margit Antal and Gavril Toderean, "Speaker recognition and broad phonetic groups.," in *SPPRA*, 2006, pp. 155–159.

[35] 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.

[36] Moez Ajili, Jean-François Bonastre, Juliette Kahn, Solange Rossato, and Guillaume Bernard, "Fabiole, a speech database for forensic speaker comparison," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.*, 2016.

[37] Jared J Wolf, "Efficient acoustic parameters for speaker recognition," *The Journal of the Acoustical Society of America*, vol. 51, no. 6B, pp. 2044–2056, 1972.

[38] M Sambur, "Selection of acoustic features for speaker identification," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, no. 2, pp. 176–182, 1975.

[39] Julian P Eatock and John S Mason, "A quantitative assessment of the relative speaker discriminating properties of phonemes," in *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*. IEEE, 1994, vol. 1, pp. I–133.

[40] U Hofker, "Auros-automatic recognition of speakers by computers: phoneme ordering for speaker recognition," in *Proc. 9th International Congress on'Acoustics, Madrid*, 1977, pp. 506–507.

[41] R Kashyap, "Speaker recognition from an unknown utterance and speaker-speech interaction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 6, pp. 481–488, 1976.

[42] Kanae Amino, Takashi Osanai, Toshiaki Kamada, Hisanori Makinae, and Takayuki Arai, "Effects of the phonological contents and transmission channels on forensic speaker recognition," in *Forensic Speaker Recognition*, pp. 275–308. Springer, 2012.

[43] Laura Fernández Gallardo, Michael Wagner, and Sebastian Möller, "I-vector speaker verification based on phonetic information under transmission channel effects.," in *INTERSPEECH*, 2014, pp. 696–700.

[44] 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.

[45] 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.

[46] 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.

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

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

[49] M. Ajili, J. f. Bonastre, S. Rossatto, and J. Kahn, "Inter-speaker variability in forensic voice comparison: A preliminary evaluation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 2114–2118.

[50] 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.

[51] 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.

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

[53] 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.

[54] 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.

[55] 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.

[56] Georges Linares, Pascal Nocéra, Dominique Massonie, and Driss Matrouf, "The lia speech recognition system: from 10xrt to 1xrt," in *International Conference on Text, Speech and Dialogue*. Springer, 2007, pp. 302–308.

[57] Benjamin Bigot, Grégory Senay, Georges Linares, Corinne Fredouille, and Richard Dufour, "Combining acoustic name spotting and continuous context models to improve spoken person name recognition in speech.," in *INTERSPEECH*, 2013, pp. 2539–2543.

[58] KN Stevens, "Acoustic phonetics. 1998," 1999.

[59] C Schindler and C Draxler, "The influence of bandwidth limitation on the speaker discriminating potential of nasals and fricatives," *International Association for Forensic Phonetics and Acoustics (IAFPA)*, 2013.