Impact of rhythm on forensic voice comparison reliability

Moez Ajili, Solange Rossato, Dan Zhang, Jean-François Bonastre

Impact of rhythm on forensic voice comparison reliability

Moez Ajili[1], *Solange Rossato*[2], *Dan Zhang*[2], *Jean-François Bonastre*[1]

[1]University of Avignon, LIA-CERI, Avignon, France
[2]Univ.Grenoble-Alpes, LIG , Grenoble, France

Abstract

It is common to see voice recordings being presented as a forensic trace in court. Generally, a forensic expert is asked to analyze both suspect and criminals voice samples in order to indicate whether the evidence supports the prosecution (same-speaker) or defence (different-speakers) hypotheses. This process is known as Forensic Voice Comparison (FVC). Since the emergence of the DNA typing model, the likelihood-ratio (LR) framework has become the new golden standard in forensic sciences. The LR not only supports one of the hypotheses but also quantifies the strength of its support. However, the LR accepts some practical limitations due to its estimation process itself. It is particularly true when Automatic Speaker Recognition (ASpR) systems are considered as they are outputting a score in all situations regardless of the case specific conditions. Indeed, several factors are not taken into account by the estimation process like the quality and quantity of information in both voice recordings, their phonological content or also the speakers intrinsic characteristics, etc. All these factors put into question the validity and reliability of FVC. In our recent study, we showed that intra-speaker variability explains 2/3 of the system losses. In this article, we investigate the relations between intra-speaker variability and rhythmic parameters.

Index terms— Forensic voice comparison, rhythm, reliability, speaker factor, speaker recognition.

1. Introduction

Forensic voice comparison (FVC) is based on the comparison of a recording of an unknown 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 sound 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 the following Equation,

$$\mathrm{LR} = \frac{\mathrm{p}(\mathrm{E} \mid \mathrm{H_p})}{\mathrm{p}(\mathrm{E} \mid \mathrm{H_d})} \tag{1}$$

where E is the trace, $\mathrm{H_p}$ is the prosecutor hypothesis (same origin), and $\mathrm{H_d}$ 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 (EER $\approx 1\%$ [5, 6, 7]), the forensic scenario is still a very challenging one for ASpR for several reasons detailed in [8, 9, 10]. Indeed, ASpR are working as black boxes: they are outputting a score in all situations regardless of the case specific conditions, ignoring a large set of observable factors.

First, trial conditions like the quality and quantity of information in both voice recordings. The speech samples are containing noises, 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 [11, 12, 13], the authors showed that homogeneity of the speaker-specific information between the two recordings of a voice comparison trial is playing an important role and should not be ignored by the LR estimation process.

Second, the phonological content is not used explicitly, as well as the presence or absence of different speaker-specific cues. In state-of-the-art ASpR systems, for example IVector (IV) based ones, a recording is encoded by an unique low dimensional vector. However several research works like [14, 15, 16, 17] agree that speaker specific information is not equally distributed on the speech signal and strongly depends on the phoneme distribution.

And finally, the speaker himself is an important factor [18, 8, 19, 20]. A speaker could 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 native or second language should be taken into account by the forensic experts). The speakers intrinsic characteristics may have a huge impact on the intra-speaker variability and therefore put into question the validity and reliability of FVC [8, 19, 20]. Indeed, in [19], we showed that speakers do not behave the same way in response of similar condition changes: some speakers will be quite robust with limited LR variation when some other are showing a huge variation. In a recent study [8], we showed that intra-speaker variability has a great impact on the system accuracy and it is responsible of about $2/3$ of the system loss (this proportion is higher for some speakers with an intra-speaker variability that can explain more than 95% of the system losses). And it is important to never forget that the speakers are not necessarily cooperative and may disguise their voices, with consequences on performance [21].

In this article, we take our previous analysis [8] a step farther and investigate deeper the relations between intra-speaker variability and rhythmic parameters (changes in speaker speech rhythm may be a factor of intra-speaker variability). First, we propose to analyze whether some rhythmic parameters are de-

pendent on the speaker. Second, we investigate if variation in rhythm may explain the high intra-speaker variability observed for some speakers and therefore explain the difference in performance observed between speakers. Our study is performed based on Fabiole [22], a database where within-speaker variability is strong.

This paper is structured as follows. Section 2 presents a review of research on speech rhythm. 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. The scope of research on speech rhythm

There have been a large number of studies on speech rhythm variability, focusing on different aspects of speech: between-language rhythmic similarities and differences [23, 24, 25, 26], rhythmic characteristics of dialects or vernaculars of a language [27, 28, 29], metrically regular speech [30, 31], pathological speech [32, 33], and more particularly speaker idiosyncratic rhythmic characteristics [34, 35, 36, 37, 38] which is on the scope of this study.

Speech rhythm in terms of durational variability of different levels of phonetic intervals can vary between speakers. A possible rationale motivating rhythmic variability between speakers was derived from the observation that the kinematic properties of the articulators over time are, on the one hand, driven by their individual anatomic characteristics, their spatial dimensions, mass and accelerations [39], and, on the other hand, by the individual ways speakers acquired to operate their articulators [36, 40]. The individual steering of the articulators should then result in individual temporal characteristics of speech. [36, 37, 41] showed that durational measures of speech rhythm could vary strongly and significantly between speakers. [36] further revealed that the most likely sources of this variability are articulatory factors varying between speakers.

3. Experimental protocol

This section presents firstly the database used, FABIOLE and the evaluation metrics applied in this study. The rest of the section is dedicated to the methodology retained to evaluate the impact of rhythmic parameters 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 [42], ESTER 1, ESTER 2 [43] and ETAPE [44]. This characteristic allows to use these databases as a source of training data. More details could be found in [22].

3.2. Evaluation metric

We use the $\mathrm{C_{llr}}$ and the minimum value of the $\mathrm{C_{llr}}$, denoted $\mathrm{C^{min}_{llr}}$, 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) [45, 46, 47, 48]. $\mathrm{C_{llr}}$ has the meaning of a cost or a loss: lower the $\mathrm{C_{llr}}$ is, better is the performance. $\mathrm{C_{llr}}$ could be calculated as follows:

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

As shown in Equation 2, $\mathrm{C_{llr}}$ can be decomposed into the sum of two parts:

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

In this paper, we use an affine calibration transformation [49] estimated using all the trial subsets (*pooled condition*) using FoCal Toolkit [50].

3.3. 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 open-source toolkit [52, 53, 54]. It uses I-vector approach [5]. Acoustic features are composed of 19 LFCC parameters, their 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

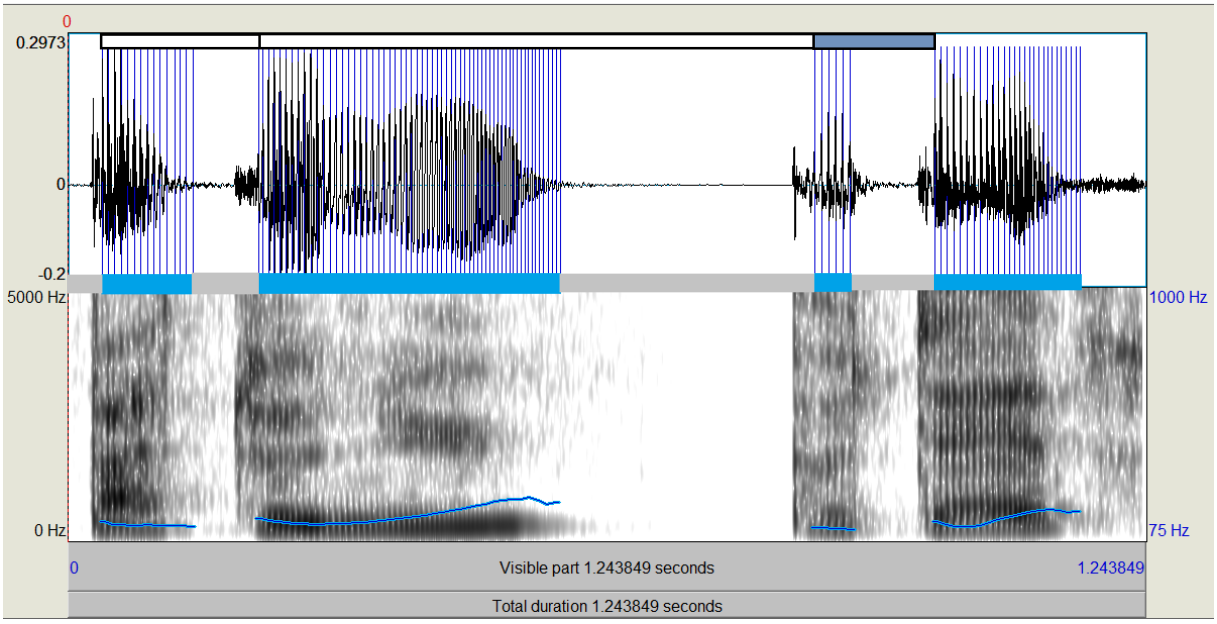

Figure 1: *Visualization of voiced (in blue) and unvoiced intervals (in gray) as well as, the pairs (black framed). Pairs for which the voiced portion is shorter than the unvoiced one are represented by black framed with blue.*

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. Temporal measures applied

In this paper, We use a wide variety of temporal measures that are commonly used in the field of speech rhythm research [56, 35, 36, 37]: we measured durational variability of voiced and unvoiced intervals (including pauses). Seven measures are used in this study:

- The percentage over which speech is voiced $\%VO$ [57];
- The mean voiced interval duration $\overline{VO}$;
- The rate-normalized standard deviation of unvoiced interval durations (VarcoUV [58]):
  VarcoUV= $\frac{\Delta UV}{\overline{UV}}\%$
  where ΔUV is the standard deviation of unvoiced interval durations and $\overline{UV}$ is the mean of unvoiced interval durations.
- The rate-normalized standard deviation of voiced interval durations (VarcoVO [57])
  VarcoVO= $\frac{\Delta VO}{\overline{VO}}\%$
  where ΔVO is the standard deviation of voiced interval durations and $\overline{VO}$ is the mean of voiced interval durations.

For example, a speech recording that contains significant pauses would have a slightly lower %VO and a larger VarcoUV variance coefficient than another speech recording with fewer long pauses. To these classical measures, we also add the mean and the standard deviation of the time interval between the beginning of two successive voiced intervals or a pair [1]. For example, the i^{th} pair is the interval of duration of the i^{th} voiced interval ($\mathrm{dVO_i}$) and the $(\mathrm{i}+1)^{\mathrm{th}}$ unvoiced one ($\mathrm{dUV_{i+1}}$), $\mathrm{dVO_i}$ + $\mathrm{dUV_{i+1}}$. Among these pairs, we estimate the percentage of pairs for which the voiced interval VO_i is shorter than the unvoiced interval UV_{i+1}. Therefore, to the above list three measure are added:

- The percentage of pairs for which the duration of an unvoiced interval is greater than the voiced interval, $\%(\mathrm{UV_{i+1}} > \mathrm{VO_i})$;
- The average duration of pairs, Average(pair);
- The standard deviation of the duration of the pairs, VarcoPair.

These temporal measures were calculated for each file using "ProsodyPro" a script developed by [59] available under [2]. Figure 1 illustrates how rhythmic parameters mentioned above are extracted.

3.5. Statistical significance evaluation

In this subsection, we present the statistical methods used to study the significance of our results. We selected *"analysis of variance"* (ANOVA) one of the most widely used statistical hypothesis tests. A difference in term of $\mathrm{C_{llr}}$, is considered significant if the obtained p-value is below an arbitrary threshold, classically set to 0.05. In order to study the size of an effect, several standardized measures have been proposed. An effect size is a quantitative measure designed to quantify the degree of association between an effect (e.g., a main effect, an interaction, a linear contrast) and the dependent variable [60, 61]. The value of the measure of association is squared and it can be interpreted as the proportion of variance in the dependent variable that is attributable to each effect. Eta squared η^2 [62], one among these measures, is the proportion of the total variance that is attributed to an effect. It is calculated as the ratio

[1] The last voiced interval is not taken into account.

[2] `http://www.homepages.ucl.ac.uk/~uclyyix/ProsodyPro/`

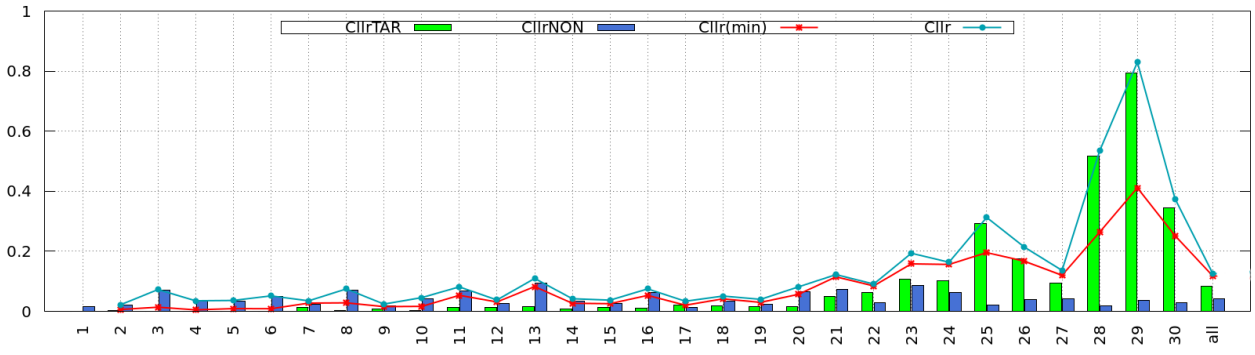

Figure 2: C_{llr}, $C_{\text{llr}}^{\text{min}}$, $C_{\text{llr}}^{\text{TAR}}$, $C_{\text{llr}}^{\text{NON}}$ *per speaker and for "all" (data from all the speakers are pooled together) [8].*

of the effect variance ($\text{SS}_{\text{effect}}$) to the total variance (SS_{total}). As shown in Equation 3, η^2 can be interpreted as the ratio of variance explained by the factor of interest.

$$\eta^2 = \frac{\text{SS}_{\text{effect}}}{\text{SS}_{\text{total}}} \tag{3}$$

A larger value of Eta-squared η^2, always indicates a stronger effect. A commonly used interpretation, mentioned in [63, 64] (pp. 283−287), is to refer to effect sizes as:

- Small when $\eta^2 \approx 1\%$.
- Medium when $\eta^2 \approx 6\%$.
- Large when $\eta^2 \approx 14\%$.

4. Results and discussion

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. Performance variability due to speaker factor

Figure 2 presents C_{llr} estimated individually for each T speaker (the results are presented following the same ranking as [19], which was based on general C_{llr} performance). In this figure, C_{llr} is divided into two components, $\text{C}_{\text{llr}}^{\text{TAR}}$ and $\text{C}_{\text{llr}}^{\text{NON}}$, 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 $\text{C}_{\text{llr}}^{\text{NON}}$) presents a quite small variation regarding speakers while there is a huge variation of the information loss related to target trials (measured by $\text{C}_{\text{llr}}^{\text{TAR}}$). The information loss coming from target trials (computed by $\text{C}_{\text{llr}}^{\text{TAR}}$) is mainly responsible of the reported high costs obtained for some speakers (such as speaker 28, 29 and 30).

4.2. Speaker factor effect on rhythm

In order to quantify the effect of speaker on rhythm, we first extract our seven rhythmic parameters on each file (100 speech recordings per speaker). One-way ANOVA is then performed with speaker as fixed factor and the rhythmic parameter as the dependent variable. This process is done separately for each of the seven rhythmic parameters. Results are reported in Table 1 and illustrated in Figure 4 for better visualization. Figure 4 is a radar chart which shows the size of variability between speakers explained for each rhythmic parameters. Each radius represents one temporal measure. The length of the radius is proportional to the magnitude of the speaker factor.

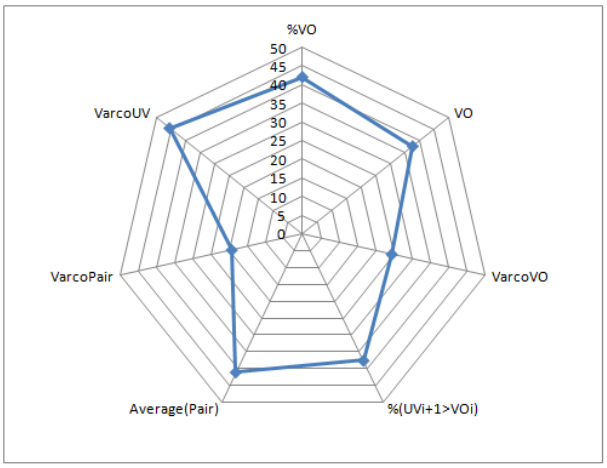

Figure 4: *Radar-chart presenting the speaker effect size explained in terms of η^2 for each rhythmic parameters.*

Table 1: Speaker factor effect on the 7 rhythmic parameters explained in terms of Eta-square η^2. (*) represents the significance level. "bold" indicates a high effect.

Variable	$\eta 2$	p-value
%VO	**42.05**	***
$\overline{\text{VO}}$	**37.72**	***
VarcoVO	**25.49**	***
VarcoUV	**45.48**	***
%(UV_{i+1}>VO_i)	**37.68**	***
Average(pair)	**41.14**	***
VarcoPair	**19.39**	***

In this experiment, all observed differences between speakers on rhythmic parameters are significant with p-value <0.001.

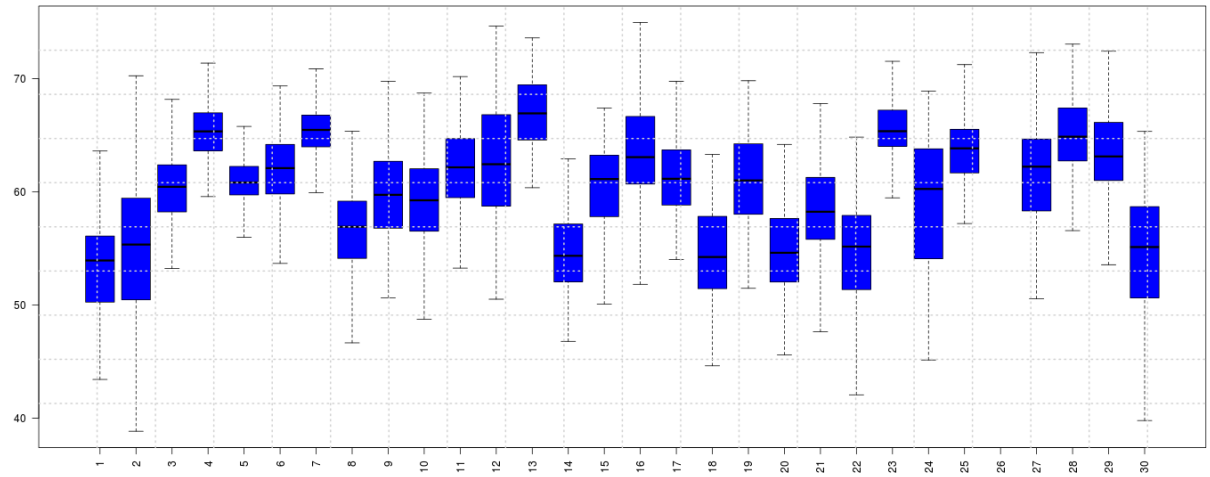

Figure 3: *%VO distribution for each speaker.*

The speaker factor has a significant effect on all the measured rhythmic parameters. If speaker factor always shows a large effect on the rhythmic parameters, this effect is varying from 19.39% for VarcoPair to 45.48% for VarcoUV. This result suggests that rhythmic parameters are highly influenced by the speaker factor.

To illustrate this finding, we present a focus on one of the parameters, %VO. Figure 3 is a box-plot showing the distributions of %VO for each speaker of Fabiole set T. When %VO computed on all the speakers is 60.2% (All), this proportion varies from 53.7% (speaker 1) to 65.5% (speaker 13).

4.3. Can rhythmic parameters explain the difference of performance between speakers?

In order to investigate the effect of rhythmic parameters on the performance, we select the three "Best" and the three "Worst" speakers of set T based on C_{llr} issued from the experiment presented in Figure 2. We wish to investigate the differences between the two speaker subsets according to each of the seven rhythmic parameters. The "worst" contains speakers 28, 29 and 30 (Average C_{llr} = 0.573bits) while the "Best" groups speakers 1, 2 and 3 (Average C_{llr} =0.036 bits).

To quantify the effect of speaker class ("Worst" or "Best") on each rhythmic parameter, one-way ANOVA is performed with speaker class as fixed factor and the rhythmic parameter as the dependent variable. Results are reported in Table 2.

Table 2: Speaker class effect size on the 7 rhythmic parameters explained in terms of $\eta 2$. (*) represents the significance level. "bold", "italic" and "normal" indicate respectively a high, medium and small effect.

Variable	$\eta 2$	Significance
%VO	*13.10*	***
$\overline{VO}$	**36.20**	***
VarcoVO	**23.11**	***
VarcoUV	1.40	**
%(UV_{i+1}>VO_i)	**35.94**	***
Average(pair)	**34.55**	***
VarcoPair	*8.50*	***

In this experiment, all observed differences between the two classes ("Best" and "Worst") on rhythmic parameters are significant. The p-value is <0.001 for all parameters, except for the VarcoUV case. The speaker class ("Best" and "Worst") factor has a large effect on $\overline{VO}$, %(UV_{i+1}>VO_i), Average(pair) and VarcoVO. %VO and VarcoPair are less variable across the speakers with a medium effect. A small effect is obtainable for VarcoUV measure.

For a deeper investigation, we present in Figure 5 the $\overline{VO}$ and Average(Pair) distributions and in Figure 6, the VarcoVO and %(UV_{i+1}>VO_i) distributions for both "Best" and "Worst" speaker classes. "All" conditions is also presented for comparative purposes.

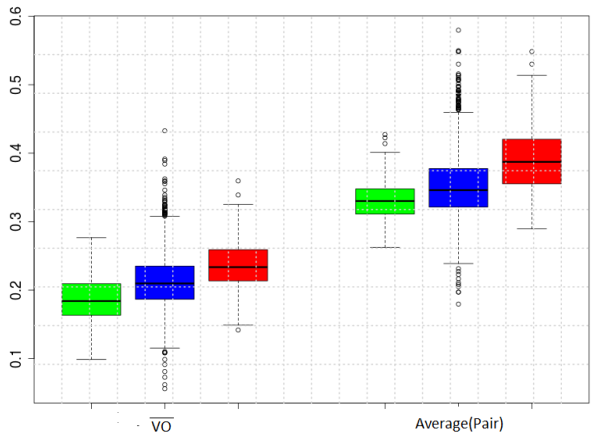

Figure 5: $\overline{VO}$ and Average(Pair) distributions for "Best" (green), "All" (blue) and "Worst" (red) speaker classes.

The VarcoVO and %(UV_{i+1}>VO_i) mean values for "Best" speakers (99.01, 47.40) are significantly higher than mean values of the "Worst" speakers (86.27, 36.39). $\overline{VO}$ and Average(Pair) mean values of "Best" class (0.18, 0.33) are significantly lower than those for "Worst" class (0.23, 0.39). It means that the "Worst" speakers have longer voiced segments than the "Best" ones. This result suggests that longer voiced intervals al-

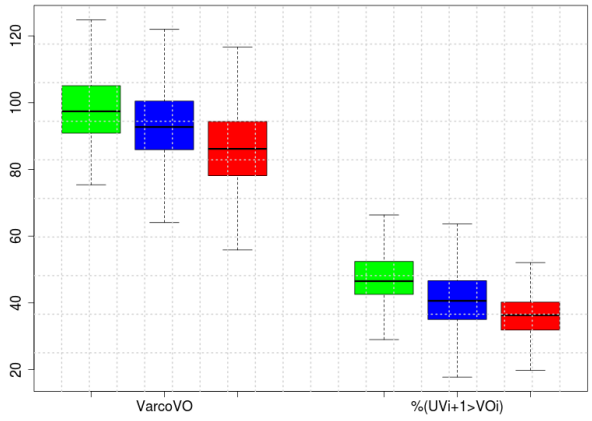

Figure 6: VarcoVO and $\%(UV_{i+1}>VO_i)$ distributions for "Best" (green), "All" (blue) and "Worst" (red) speaker classes.

low a larger articulatory flexibility in their production. This may explain a larger intra-speaker variability for "Worst" speakers and higher information losses.

5. Conclusion

This article is a complementary study to our previous research works published in [19, 8]. In the first work, we showed that speakers do not behave similarly even if the experimental conditions are well controlled (thanks to Fabiole database). In the second one, we showed a large influence of the phonetic content on ASpR performance. Furthermore, we observed a large variability depending on the speakers.

In this article, we explored the influence of rhythmic parameters on ASpR performance using Fabiole database and an ANOVA framework. We studied seven rhythmic parameters $\overline{VO}$, Average(pair), VarcoVO, %VO, VarcoPair, $\%(UV_{i+1}>VO_i)$ and VarcoUV.

In a first step, we examined the influence of the speaker on rhythmic parameters. We found that our seven rhythmic parameters (All based on temporal characteristics of speech intervals) revealed highly significant differences between speakers: The part of the speaker variance explained by a rhythmic parameters varies from 19.391% for VarcoPair to 45.485% for VarcoUV.

In a second step, we focused on the relations between these rhythmic parameters and the difference of performance between speakers. We first selected two subsets of speakers that maximized the differences in term of C_{llr}, denoted "Best" and "Worst". Then, we investigated the differences between the two speaker subsets according to each of our seven rhythmic parameters. We found that the speaker subset -i.e. the difference in performance between the speakers- has a significant effect on each rhythmic parameter. This effect is large for four parameters ($\overline{VO}$, $\%(UV_{i+1}>VO_i)$, Average(pair) and VarcoVO), medium for two (%VO and VarcoPair) and small for one (VarcoUV). We found that the worse speakers (in terms of C_{llr}) have longer voiced segments than the best performers ones. This result suggests that longer voiced intervals allow a larger articulatory flexibility, which may explain a larger intra-speaker variability for low performer speakers (and, therefore, higher information losses).

Concerning the latter results, some caution should be expressed. Firstly, the presented experiments were done on Fabiole, which present small variations in terms of speaking style as well as on sociocultural characteristics. Moreover, Fabiole contains only 30 speakers, a relatively small number which could highlight a speaker specificity (known or unknown). Nevertheless, the obtained results could be a consequence of speakers anatomical configurations, which in turn are governed by neurological motor patterns in the brain of the speaker.

In order to answer to the latter questions, we wish to enlarge the dataset, including more speakers and more speaking styles. It will also allow to increase significantly the size of the best/worse subsets (only 3 speakers each in this work), which will decrease the risk to take into account too much the speakers themselves.

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] 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.

[9] Moez Ajili, "Reliability of voice comparison for forensic applications," Avignon, 2017.

[10] Jean-François Bonastre, Juliette Kahn, Solange Rossato, and Moez Ajili, "Forensic speaker recognition: Mirages and reality," *S. Fuchs/D*, p. 255, 2015.

[11] 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.

[12] 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.

[13] Moez Ajili, Jean-François Bonastre, Waad Ben Kheder, Solange Rossato, and Juliette Kahn, "Homogeneity measure impact on target and non-target trials in forensic voice comparison," *Proc. Interspeech 2017*, pp. 2844–2848, 2017.

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

[20] 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 *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017.

[21] 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.

[22] 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.

[23] David Abercombie, *Elements of general phonetics*, Aldine Pub. Company, 1967.

[24] Esther Grabe and Ee Ling Low, "Durational variability in speech and the rhythm class hypothesis," *Papers in laboratory phonology*, vol. 7, no. 515-546, 2002.

[25] Anastassia Loukina, Greg Kochanski, Burton Rosner, Elinor Keane, and Chilin Shih, "Rhythm measures and dimensions of durational variation in speech," *The Journal of the Acoustical Society of America*, vol. 129, no. 5, pp. 3258–3270, 2011.

[26] Sam Tilsen and Amalia Arvaniti, "Speech rhythm analysis with decomposition of the amplitude envelope: characterizing rhythmic patterns within and across languages," *The Journal of the Acoustical Society of America*, vol. 134, no. 1, pp. 628–639, 2013.

[27] Sónia Frota and Marina Vigário, "On the correlates of rhythmic distinctions: The european/brazilian portuguese case," *Probus*, vol. 13, no. 2, pp. 247–275, 2001.

[28] Low Ee Ling, Esther Grabe, and Francis Nolan, "Q uantitative characterizations of speech rhythm: Syllable-timing in singapore english," *Language and speech*, vol. 43, no. 4, pp. 377–401, 2000.

[29] Tamara V Rathcke and Rachel H Smith, "Speech timing and linguistic rhythm: On the acoustic bases of rhythm typologies," *The Journal of the Acoustical Society of America*, vol. 137, no. 5, pp. 2834–2845, 2015.

[30] Victoria Leong and Usha Goswami, "Impaired extraction of speech rhythm from temporal modulation patterns in speech in developmental dyslexia," *Frontiers in human neuroscience*, vol. 8, pp. 96, 2014.

[31] Michael ODell and Tommi Nieminen, "Coupled oscillator model of speech rhythm," in *Proceedings of the XIVth international congress of phonetic sciences*. University of California Berkeley, 1999, vol. 2, pp. 1075–1078.

[32] Victoria Leong, Michael A Stone, Richard E Turner, and Usha Goswami, "A role for amplitude modulation phase relationships in speech rhythm perception," *The Journal of the Acoustical Society of America*, vol. 136, no. 1, pp. 366–381, 2014.

[33] Julie M Liss, Laurence White, Sven L Mattys, Kaitlin Lansford, Andrew J Lotto, Stephanie M Spitzer, and John N Caviness, "Quantifying speech rhythm abnormalities in the dysarthrias," *Journal of speech, language, and hearing research*, vol. 52, no. 5, pp. 1334–1352, 2009.

[34] Volker Dellwo and Jacques Koreman, "How speaker idiosyncratic is measurable speech rhythm," in *Abstract presented at the annual IAFPA meeting*, 2008.

[35] Volker Dellwo, Adrian Leemann, and Marie-José Kolly, "Speaker idiosyncratic rhythmic features in the speech signal," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[36] Volker Dellwo, Adrian Leemann, and Marie-José Kolly, "Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors," *The Journal of the Acoustical Society of America*, vol. 137, no. 3, pp. 1513–1528, 2015.

[37] Adrian Leemann, Marie-José Kolly, and Volker Dellwo, "Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison," *Forensic science international*, vol. 238, pp. 59–67, 2014.

[38] Athanasios Lykartsis, Stefan Weinzierl, and Volker Dellwo, "Speaker identification for swiss german with spectral and rhythm features," in *Audio Engineering Society Conference: 2017 AES International Conference on Semantic Audio*. Audio Engineering Society, 2017.

[39] Pascal Perrier, "Gesture planning integrating knowledge of the motor plant's dynamics: A literature review from motor control and speech motor control," 2012.

[40] Pär Wretling and Anders Eriksson, "Is articulatory timing speaker specific?–evidence from imitated voices," in *Proc. FONETIK*. Citeseer, 1998, vol. 98, pp. 48–52.

[41] Lukas Wiget, Laurence White, Barbara Schuppler, Izabelle Grenon, Olesya Rauch, and Sven L Mattys, "How stable are acoustic metrics of contrastive speech rhythm?," *The Journal of the Acoustical Society of America*, vol. 127, no. 3, pp. 1559–1569, 2010.

[42] 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.

[43] 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.

[44] 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.

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

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

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

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

[49] 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.

[50] Niko Brummer, "Focal toolkit," *Available in http://www. dsp. sun. ac. za/nbrummer/focal*, 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] Volker Dellwo and Adrian Fourcin, "Rhythmic characteristics of voice between and within languages," *Revue Tranel (Travaux neuchâtelois de linguistique)*, vol. 59, pp. 87–107, 2013.

[57] Volker Dellwo, Adrian Fourcin, and Evelyn Abberton, "Rhythmical classification based on voice parameters," in *International Conference of Phonetic Sciences (ICPhS)*, 2007, pp. 1129–1132.

[58] Volker Dellwo, "Rhythm and speech rate: A variation coefficient for c," *Language and language-processing*, pp. 231–241, 2006.

[59] Yi Xu, "Prosodyproa tool for large-scale systematic prosody analysis," Laboratoire Parole et Langage, France, 2013.

[60] Daniël Lakens, "Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and anovas," *Frontiers in psychology*, vol. 4, pp. 863, 2013.

[61] Catherine O Fritz, Peter E Morris, and Jennifer J Richler, "Effect size estimates: current use, calculations, and interpretation.," *Journal of Experimental Psychology: General*, vol. 141, no. 1, pp. 2, 2012.

[62] Timothy R Levine and Craig R Hullett, "Eta squared, partial eta squared, and misreporting of effect size in communication research," *Human Communication Research*, vol. 28, no. 4, pp. 612–625, 2002.

[63] Jacob Cohen, "Statistical power analysis for the behavioral sciences (revised ed.)," 1977.

[64] Jacob Cohen, "Statistical power analysis for the behavior science," *Lawrance Eribaum Association*, 1988.