Voice Comparison and Rhythm: Behavioral Differences between Target and Non-target Comparisons

Moez Ajili, Jean-François Bonastre, Solange Rossato

Interspeech 2018
2-6 September 2018, Hyderabad

Voice Comparison and Rhythm: Behavioral Differences between Target and Non-target Comparisons

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

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

name.surname@univ-avignon.fr, name.surname@univ-grenoble-alpes.fr

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 criminal's 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 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. In our recent study, we showed the importance of the phonemic content and we highlighted interesting differences between inter-speakers effects and intra-speaker's ones. In this article, we wish to take our previous analysis a step farther and investigate the impact of rhythm variation separately on target and non-target trials.

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 defence (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]. In this framework, the *likelihood ratio* (LR) is used to present the results of the forensic expertise. The LR supports one of the hypothesis but also quantifies the strength of its support.

Automatic Speaker Recognition (ASpR) is considered as one of the most appropriate solution [3]. Even if impressive low error rates ($\approx 1\%$ [4, 5]) were reported in the last years, the forensic scenario is still very challenging for ASpR for several reasons [6].

The first factor is the trial conditions like the quality and quantity of information in both voice recordings. The speech samples contain noises, may be very short. Their content cannot be controlled (at least for the trace) and may not contain enough relevant information. In [7, 8, 9], we showed that homogeneity of the speaker-specific information between the two recordings of a voice comparison trial is playing an important role.

Second, the speaker himself is an important factor [10, 6, 11]. A speaker could be ill, or under the influence of stress, alcohol or other drugs. The social and linguistic environment of the unknown speaker is unknown by construction.Speaker's intrinsic characteristics may have a huge impact on the intra-speaker variability [6, 11]. Indeed, [11] showed that speakers do not behave the same way in response of similar conditions: some speakers will be quite robust with limited LR variation when some other are showing a huge variation. In [6], we showed that intra-speaker variability has a great impact on the system accuracy and is responsible of about $2/3$ of the system loss (This proportion is higher for some speakers with an intra-speaker variability able to 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 [12].

Finally, the phonological content is not exploited explicitly, as well as the presence or absence of different speaker-specific cues. However several research works like [13, 14, 15, 16] agree that speaker specific information is not equally distributed on the speech signal and strongly depends on the phoneme distribution. [6] showed that the phonological content has a different impact on target than on non-target comparisons.

In this article, we take our previous analysis [6] a step farther and investigate deeper the impact of rhythm variation separately on target and non-target comparisons. First, we propose to analyze whether some rhythmic parameters are dependent to 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 [17], a database where within-speaker variability is strong.

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 [18, 19], rhythmic characteristics of dialects or vernaculars of a language [20, 21], metrically regular speech [22, 23], pathological speech [24, 25], and more particularly speaker idiosyncratic rhythmic characteristics [26, 27, 28, 29] which is on the scope of this study.

Speech rhythm in terms of durational variability of different levels of phonological intervals can vary between speakers. A possible rationale motivating rhythmic variability between speakers was derived from the observation that the kinematic

10.21437/Interspeech.2018-61

properties of the articulators over time are, on the one hand, driven by their individual anatomic characteristics, their spatial dimensions, mass and accelerations [30], and, on the other hand, by the individual ways speakers acquired to operate their articulators [27, 31]. The individual steering of the articulators should then result in individual temporal characteristics of speech. [27, 28, 32] showed that durational measures of speech rhythm could vary strongly and significantly between speakers. [27] 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 contains 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 $(100 \times 100 \times 29 = 290\text{k})$ non-targets pairs.

FABIOLE contains recordings gathered from different kinds of speakers, including journalists, announcers, politicians, chroniclers, interviewers, etc. More details could be found in [17].

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) [33]. $\mathrm{C_{llr}}$ has the meaning of a cost or a loss: the lower the $\mathrm{C_{llr}}$ is, the better the performance is. $\mathrm{C_{llr}}$ could be calculated as follows:

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

As shown in Equation 1, $\mathrm{C_{llr}}$ 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 using FoCal Toolkit [34].

3.3. LIA speaker recognition system

In all experiments, we use as baseline the LIA_SpkDet system [35].This system is developed using the ALIZE/SpkDet open-source toolkit [36, 37]. It uses I-vector approach [4]. 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 [38] is applied.

3.4. Temporal measures applied

In this paper, we use a variety of temporal measures that are commonly used in the field of speech rhythm research [39, 26, 27, 28]: 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 [40];
- The mean voiced interval duration $\overline{\mathrm{VO}}$;
- The rate-normalized standard deviation of unvoiced interval durations (VarcoUV [41]); VarcoUV= $100 \times \frac{\Delta \mathrm{UV}}{\overline{\mathrm{UV}}}$ where ΔUV and $\overline{\mathrm{UV}}$ are the standard deviation and the mean of unvoiced interval durations.
- The rate-normalized standard deviation of voiced interval durations (VarcoVO [40]); VarcoVO= $100 \times \frac{\Delta \mathrm{VO}}{\overline{\mathrm{VO}}}$ where $\overline{\mathrm{VO}}$ and ΔVO are respectively the mean and standard deviation of voiced interval durations.

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. For example, the $\mathrm{i^{th}}$ pair is the interval of duration of the $\mathrm{i^{th}}$ 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 $\mathrm{VO_i}$ is shorter than the unvoiced interval $\mathrm{UV_{i+1}}$. 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);

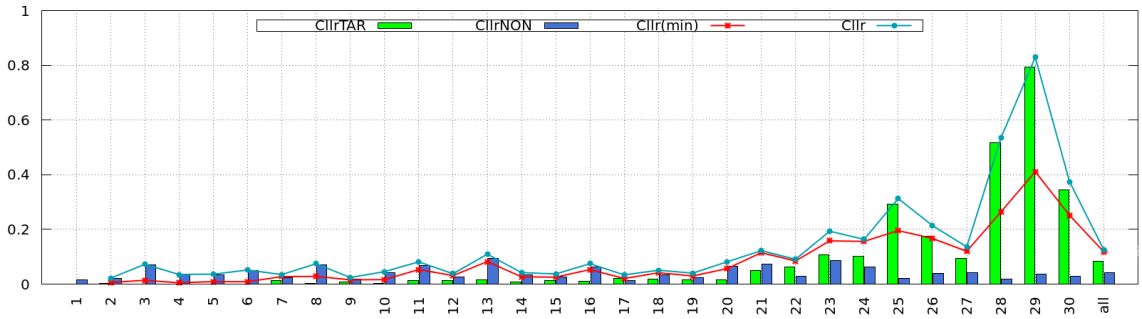

Figure 1: C_{llr}, C_{llr}^{min}, C_{llr}^{TAR}, C_{llr}^{NON} *per speaker and for "all" (data from all the speakers are pooled together) [6].*

- The standard deviation of the duration of the pairs, VarcoPair.

These temporal measures were calculated for each file using "ProsodyPro" a script developed by [42] available under `http://www.homepages.ucl.ac.uk/~uclyyix/ProsodyPro/`.

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 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 [43, 44]. 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 [45], one among these measures, is the proportion of the total variance that is attributed to an effect. It is calculated as the ratio of the effect variance (SS_{effect}) to the total variance (SS_{total}). As shown in Equation 2, η^2 can be interpreted as the ratio of variance explained by the factor of interest.

$$\eta^2 = \frac{SS_{effect}}{SS_{total}} \quad (2)$$

A larger value of Eta-squared η^2, always indicates a stronger effect. A commonly used interpretation, mentioned in [46] (pp. 283−287), is to refer to effect sizes as Small when $\eta^2 \approx 1\%$, Medium when $\eta^2 \approx 6\%$ and 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 0.12631bits 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 1 presents C_{llr} estimated individually for each T speaker (the results are presented following the same ranking as [11], 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 (such as speaker 28, 29 and 30).

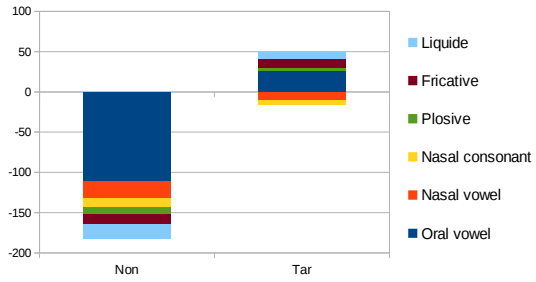

Figure 2: Stacked bar chart of C_{llr}^{R} *computed on* C_{llr}^{TAR} *(target trials) and* C_{llr}^{NON} *(non-target trials). Positive* C_{llr}^{TAR} *indicates that phoneme category in question have a negative effect in FVC and vice versa [6].*

In Figure 2, we remind the different behaviors of the phonological content between target versus non-target comparisons reported in our previous study [6]: when all the phonological classes play a positive role in speaker discrimination for non-target comparisons. Only nasals, vowels and consonants, appeared to be conveying speaker-specific information for target comparisons.

4.2. Speaker factor effect on rhythm

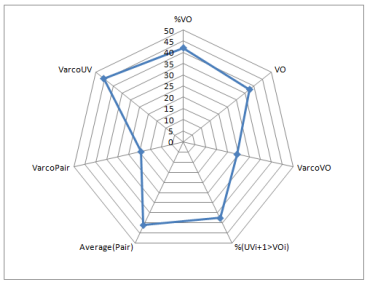

Figure 3: Radar-chart presenting the speaker effect size explained in terms of η^2 *for each rhythmic parameters.*

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 3 for better visualization. Figure 3 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.

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.

Measure	$\eta 2$	p-value
%VO	**42.056**	***
$\overline{VO}$	**37.727**	***
VarcoVO	**25.492**	***
VarcoUV	**45.485**	***
%(UV$_{i+1}$>VO$_i$)	**37.686**	***
Average(pair)	**41.142**	***
VarcoPair	**19.391**	***

In this experiment, all observed differences between speakers on rhythmic parameters are significant with p-value <0.001. 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 19.391% for VarcoPair to 45.485% for VarcoUV. This result suggests that rhythmic parameters are highly influenced by the speaker.

4.3. Can rhythmic parameters explain the difference of performance between target and non-target comparisons?

In this section, C_{llr}^{TAR} and C_{llr}^{NON} are calculated for each file. Indeed, for a given file, C_{llr}^{TAR} and C_{llr}^{NON} are estimated using target and non-target trials involving the file in question.

In order to investigate the effect of rhythmic parameters on the information loss for target comparisons, we select the $x\%$ "Best" and the $x\%$ "Worst" files based on C_{llr}^{TAR}. We wish to investigate the differences between the two file subsets according to each of the seven rhythmic parameters. The same strategy is applied for non-target comparisons using C_{llr}^{NON}. In this experiment, each subset corresponds to 500 files.

Table 2: file 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.

	TAR		NON	
	$\eta 2$	p-value	$\eta 2$	p-value
%VO	3.88	***	2.5	***
$\overline{VO}$	3.77	***	0.42	*
VarcoVO	**14.79**	***	0.27	n.s
VarcoUV	0.21	n.s	0.04	n.s
%(UV$_{i+1}$>VO$_i$)	**13.26**	***	1.51	***
Average(pair)	2.20	***	0.13	n.s
VarcoPair	2.30	***	0.10	n.s

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

For target comparisons, all observed differences between the two subsets ("Best" and "Worst") on rhythmic parameters are significant except for VarcoUV. The file class ("Best" and "Worst") factor has a large effect on $VarcoVO$ and %(UV$_{i+1}$>VO$_i$) while it is small for the remaining rhythmic parameters.

For non-target comparisons, only the differences on $\%VO$ and %(UV$_{i+1}$>VO$_i$) between the two subsets ("Best" and "Worst") are significant. The effect of the file class on the remaining rhythmic parameters is non significant. A small effect size of the file class is observed for all the rhythmic parameters.

Taken together, the results suggest that rhythm variation has a significant impact on target comparison's accuracy while for non-target comparison, this variation seems to not have an impact on the FVC accuracy.

5. Conclusion

This article is a complementary study to our previous research works published in [11, 6]. 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 phonological 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 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 target versus non-target comparisons. We first selected two subsets of files that maximized the differences in term of C_{llr}^{TAR}, denoted "Best" and "Worst". Then, we investigated the differences between the two subsets according to each of our seven rhythmic parameters. This strategy is applied also on non-target comparisons using C_{llr}^{NON}. We found that the file subset -i.e. the difference in performance between excerpts- has a significant effect on each rhythmic parameter for target comparisons. This effect is large for VarcoVO and $\%(UV_{i+1} > VO_i)$ and small for the remaining parameters. On the other side, the file subsets ("Best" and "Worst" in terms of C_{llr}^{NON}) seem to not have a significant difference according to the studied rhythmic parameters. This result suggests that rhythm variation has essentially a significant impact on target comparisons while for non-target ones it does not seem to impact accuracy.

6. References

[1] C. Champod and D. Meuwly, "The inference of identity in forensic speaker recognition," *Speech Communication*, 2000.

[2] C. G. Aitken and F. Taroni, *Statistics and the evaluation of evidence for forensic scientists*. Wiley Online Library, 2004, vol. 10.

[3] E. Gold and P. French, "An international investigation of forensic speaker comparison practices," in *Proceedings of the 17th International Congress of Phonetic Sciences*, 2011.

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

[5] J. H. Hansen and T. Hasan, "Speaker recognition by machines and humans: a tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[6] M. Ajili, J.-F. Bonastre, W. Ben Kheder, S. Rossato, and J. Kahn, "Phonetic content impact on forensic voice comparison," in *Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 210–217.

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

[8] ——, "Homogeneity measure for forensic voice comparison: A step forward reliability," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Springer, 2015, pp. 135–142.

[9] M. Ajili, J.-F. Bonastre, W. B. Kheder, S. Rossato, and J. Kahn, "Homogeneity measure impact on target and non-target trials in forensic voice comparison," *Proc. Interspeech 2017*, pp. 2844–2848, 2017.

[10] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds, "Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation," DTIC Document, Tech. Rep., 1998.

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

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

[13] K. Amino, T. Sugawara, and T. Arai, "Idiosyncrasy of nasal sounds in human speaker identification and their acoustic properties," *Acoustical science and technology*, 2006.

[14] M. Antal and G. Toderean, "Speaker recognition and broad phonetic groups." in *SPPRA*, 2006, pp. 155–159.

[15] M. Ajili, J.-F. Bonastre, W. B. Kheder, S. Rossato, and J. Kahn, "Phonological content impact on wrongful convictions in forensic voice comparison context," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017.

[16] M. Ajili, "Reliability of voice comparison for forensic applications." Avignon, 2017.

[17] M. Ajili, J. Bonastre, J. Kahn, S. Rossato, and G. Bernard, "Fabiole, a speech database for forensic speaker comparison," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC, May 23-28.*, 2016.

[18] A. Loukina, G. Kochanski, B. Rosner, E. Keane, and C. 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.

[19] S. Tilsen and A. 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.

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

[21] T. V. Rathcke and R. 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.

[22] V. Leong and U. Goswami, "Impaired extraction of speech rhythm from temporal modulation patterns in speech in developmental dyslexia," *Frontiers in human neuroscience*, vol. 8, p. 96, 2014.

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

[24] V. Leong, M. A. Stone, R. E. Turner, and U. 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.

[25] J. M. Liss, L. White, S. L. Mattys, K. Lansford, A. J. Lotto, S. M. Spitzer, and J. N. Caviness, "Quantifying speech rhythm abnormalities in the dysarthrias," *Journal of speech, language, and hearing research*, vol. 52, no. 5, pp. 1334–1352, 2009.

[26] V. Dellwo, A. Leemann, and M.-J. Kolly, "Speaker idiosyncratic rhythmic features in the speech signal," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[27] ——, "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.

[28] A. Leemann, M.-J. Kolly, and V. Dellwo, "Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison," *Forensic science international*, vol. 238, pp. 59–67, 2014.

[29] A. Lykartsis, S. Weinzierl, and V. 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.

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

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

[32] L. Wiget, L. White, B. Schuppler, I. Grenon, O. Rauch, and S. 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.

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

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

[35] D. Matrouf, N. Scheffer, B. G. Fauve, and J.-F. Bonastre, "A straightforward and efficient implementation of the factor analysis model for speaker verification." in *INTERSPEECH*, 2007.

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

[37] A. Larcher, J.-F. Bonastre, B. G. Fauve, K.-A. Lee, C. Lévy, H. Li, J. S. Mason, and J.-Y. Parfait, "Alize 3.0-open source toolkit for state-of-the-art speaker recognition." in *INTERSPEECH*, 2013.

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

[39] V. Dellwo and A. Fourcin, "Rhythmic characteristics of voice between and within languages," *Revue Tranel (Travaux neuchâtelois de linguistique)*, vol. 59, pp. 87–107, 2013.

[40] V. Dellwo, A. Fourcin, and E. Abberton, "Rhythmical classification based on voice parameters," in *International Conference of Phonetic Sciences (ICPhS)*, 2007, pp. 1129–1132.

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

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

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

[44] C. O. Fritz, P. E. Morris, and J. J. Richler, "Effect size estimates: current use, calculations, and interpretation." *Journal of Experimental Psychology: General*, vol. 141, no. 1, p. 2, 2012.

[45] T. R. Levine and C. 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.

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