Phonological content impact on wrongful convictions in Forensic Voice Comparison context
Moez Ajili, Jean-François Bonastre, Waad Ben Kheder, Solange Rossato, Juliette Kahn

To cite this version:

HAL Id: hal-01962627
https://hal.archives-ouvertes.fr/hal-01962627
Submitted on 20 Dec 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Phonological Content Impact on Wrongful Convictions in Forensic Voice Comparison Context

Conference Paper · March 2017
DOI: 10.1109/ICASSP.2017.7952536

5 authors, including:

Jean-François Bonastre
Université d’Avignon et des Pays du Vaucluse
245 PUBLICATIONS 4,642 CITATIONS

Solange Rossato
Université Grenoble Alpes
85 PUBLICATIONS 234 CITATIONS

Waad Ben Kheder
Vocapia
30 PUBLICATIONS 83 CITATIONS

Some of the authors of this publication are also working on these related projects:

Project
Forensic Voice Comparison View project

Project
Iban ASR View project

All content following this page was uploaded by Moez Ajili on 29 May 2018.
The user has requested enhancement of the downloaded file.
PHONOLOGICAL CONTENT IMPACT ON WRONGFUL CONVICTIONS IN FORENSIC VOICE COMPARISON CONTEXT

Moez Ajili¹, Jean-François Bonastre¹, Waad Ben Kheder¹, Solange Rossato², Juliette Kahn³

¹University of Avignon, LIA-CERI, Avignon, France, ²Univ. Grenoble-Alpes, LIG, F-38000 Grenoble, France ³Laboratoire National de metrologie et d’Essais, LNE, Paris, France

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. Nevertheless, 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. It is particularly true when FVC is considered using Automatic Speaker Recognition (ASR) systems. Indeed, in the LR estimation performed by ASR systems, different factors are not considered such as speaker intrinsic characteristics, denoted “speaker factor”, the amount of information involved in the comparison as well as the phonological content and so on. This article focuses on the impact of phonological content on FVC involving two different speakers and more precisely the potential implication of a specific phonemic category on wrongful conviction cases (innocents are send behind bars). We show that even though the vast majority of speaker pairs (more than 90%) are well discriminated, few pairs are difficult to distinguish. For the “best” discriminated pairs, all the phonemic content play a positive role in speaker discrimination while for the “worst” pairs, it appears that nasals have a negative effect and lead to a confusion between speakers.

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

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 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:

\[ LR = \frac{p(E|H_{ph})}{p(E|H_{dh})} \]

where \( E \) is the evidence or trace, \( H_{ph} \) is the prosecutor hypothesis (same origin), and \( H_{dh} \) is the defender hypothesis (different origins).

Automatic Speaker Recognition (ASR) is considered as one of the most appropriate solution when LR framework is involved. Even though ASR systems have achieved significant progresses in the past two decades and have reached impressive low error rates (≈ 1% [4]), using these systems to assess the strength-of-evidence in court remains inconclusive and raises some doubts about their reliability [5]. In other words, if the LR value is used by a court in order to help to take a decision, inevitably, “errors of impurity” or even wrongful conviction [6]—considered as the most outrageous in the miscarriage of justice—will occur. Wrongful convictions have been seen since a long time in the judicial process [7]. This pushed several jurists to highlight this serious phenomenon such as William Blackstone, an 18th century English jurist known by this famous quote: “It is better that ten guilty persons escape than that one innocent suffer”. Up to now, wrongful conviction cases are staggering and can not be prevented by the criminal judicial system [8, 9, 10, 11]. Serious study of this phenomenon began less than a decade after an unquantifiable number of wrongfully convicted persons that have served prison sentences or even been executed for crimes committed by others [9]. An example cemented in memory and highlighted by the mass media, is the tragic wrongful conviction of “George Stinney” who was the youngest person in the U.S. in the 20th century to be sentenced to death and to be executed (14 years old) when he was innocent or “Ricky Jackson” who was released recently after spending 39 years behind bars. Many others wrongful conviction cases are mentioned in [22].

At the beginning of 1992 and thanks to the “Innocence Project” which has led to the exoneration of a significant number of innocent previously convicted (196 cases until now), forensic practices have received much attention [8, 13, 14, 15]. FVC, like some other forensic disciplines, is not infallible and therefore this scenario is still a very challenging one for ASR for several reasons: First, the computation of the LR by an ASR system is only approximated by an extraction process and therefore, despite its nice theoretical aspects, will accept some limitations coming from imperfections of this estimation. It is particularly true if we take for example the “calibration” process [16, 17, 18] where ASR system is outputting a score and using different normalization steps to see this score as a LR.

Second, the appealing high accuracy reached by ASR system should be taken with caution. Indeed, the evaluation protocols (for example in NIST-SRE [19]) focus on global performance using a brute-force strategy and take into account the averaged behavior of ASR systems. Consequently, these protocols may ignore many sensitive cases which represent several distinct specific situations where the ASR systems may show a specific behavior due, for example, to the speech samples that could be recorded in different situations, the noises, the content of the recordings or the speakers themselves. Several research works emphasized the limits of the underlined evaluation protocols [20, 21].

Third, the phonological content play an important role in speaker comparison. Several research works like [22, 23, 24] agree that...
speaker specific information is not equally distributed on the speech signal and particularly depends on the phoneme distribution. For example, in [25] authors show that nasal and oral vowel convey the most important speaker specific information than the other phonemic categories. However, for ASR system based on I-vector, a recording is encoded by one low dimensional vector and thereby the phonological content of a recording is not used explicitly, as well as the presence or absence of different speaker-specific cues, which could be a key in some forensic cases.

Despite the apparent richness of the above literature review, the analyze of the phonological content impact on speaker comparison are still dedicated to ASR and do not take into account the specific context of FVC: there is no evaluation of the impact of the phonological content regarding different speakers and/or an assessment of a specific phonological category impact on the speaker discrimination where only two different speakers are involved. This is mainly due to a lack in terms of number of recording per speaker in the used databases.

This paper is dedicated to focus on the third point of the highlighted lacks. We investigate the impact of phonological content on the comparison process in order to determine if there are some phonological classes that are bringing more speaker specific cues than others, or also the potential implication of a specific phonemic category on the confusion between two speakers which leads to wrongful conviction in forensic context. This work could be done thanks to Fabiole database which provide a high number of speech recordings per speaker and therefore the impact of phonological content per speaker as well as per speaker pairs could finally be investigated.

2. PHONOLOGICAL 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 ASR 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 [27, 28, 29]. The authors agreed that vowels and nasals provide the best discrimination between speakers. [30] presents a ranking of 24 isolated German phonemes, which indicates nasals as providing the best SR performance, with the voiced alveolar fricative /z/ and the voiced uvular fricative /h/ also performing fairly well. In [31], /l, l/, /l/ and /h/ are found to perform worse than vowels and nasals. [27, 28] strongly promote the nasals and vowels as best performers. The influence of the phonological content of both voice recordings was also evaluated in [22] 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 [25, 26, 24, 29], nasals and vowels were found to be particularly speaker specific information and nasal vowels are more discriminant than oral vowels. Finally, [23] and, more recently, [32], 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 phonological 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 ASR 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 ASR system.

2.2. Phonoeme 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 phonological content, a phoneme transcription/alignment process is mandatory. If the classification is well chosen, the use of phoneme classes will 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 phonoem classification is describe below: oral vowels (OV) {/i/, /u/, /l/, /e/, /æ/, /ɛ/, /æ/, /l/, /ɔ/, /ɔ/, /u/, /u/, /u/, /u/, /u/, /u/, /u/, /u/, /u/}, nasal vowels (NV) {/n/, /ɛ/, /n/, /l/, /l/}, nasal consonants (NC) {/n/, /l/}, plosive (P) {/p/, /t/, /k/, /h/, /l/, /l/, /l/, /g/}, fricatives (F) {/f/, /θ/, /θ/, /l/, /l/, /l/, /l/, /l/} and liquids (L) {/l/, /l/}. This phonoem classification will be adopted in all experiments in this paper.

3. EXPERIMENTAL PROTOCOL

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 phonological content on FVC.

3.1. Corpus

This work is conducted using FABIOLE database dedicated to investigate the reliability of ASR-based FVC. FABIOLE is primarily designed for studies on speaker effect while 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 native speakers divided into two sets:

- Set T: 30 target speakers each associated with 100 recordings.
- Set I: 100 impostor speakers each associated with 1 recording.

As we focused on wrongful conviction cases, we are interested on non-target trials. Only set T is used in order to provide more than 4.5M non-matched pairs. The trials are divided into 30 subsets, one for each T speaker. These subsets 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 × 100 × 29 = 290k) non-targets pairs. More details could be found in [33].

3.2. Evaluation metric

We use the Cllr, largely used in FVC to evaluate the LLR. Cllr, is not based on hard decisions like, for example, equal error rate (EER) [17, 34, 35]. Cllr has the meaning of a cost or a loss: lower the Cllr is, better is the performance. Cllr could be calculated as follows:
In all experiments, we use as baseline the LIA speaker recognition system presented in [38]. This system is developed using the ALIZE/SpkDet open-source toolkit [37, 38]. It uses I-vector approach [2]. 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) 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 [39] is applied.

### 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 [38]. This system is developed using the ALIZE/SpkDet open-source toolkit [37, 38]. It uses I-vector approach [2]. 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 UBMs 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 [39] is applied.

### 3.4.2. LIA transcription system

FABIOLE database has been automatically transcribed thanks to SperaL, LIA automatic transcription system [40]. 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% [41].

### 4. RESULTS

The global $C_{llr}^{NON}$ (computed using all the non-target trial subsets put together) which is expected to be primarily linked to speaker discrimination power, is equal to 0.04 bits. The reported performance level is close to the level showed during the large evaluation campaigns (like the NIST’s ones) and therefore many details of the comparison are still hidden.

#### 4.1. Phonological content impact on voice comparison per speaker

Figure 1 presents $C_{llr}^{NON}$ estimated individually for each $T$ speaker (the results are presented following the same ranking as [42], which was based on general $C_{llr}$ performance.). Results show that $C_{llr}^{NON}$ per speaker presents a significant variation among speakers: The lowest $C_{llr}^{NON}$ value, 0.013, is seen for spk.17 while the highest value, 0.093, is seen for spk. 13 (almost multiplied by 7).

Figure 2 is a stacked bar chart which displays the impact of the phonemic classes per speaker as well as for “all” condition (averaged on all the speakers), in terms of relative $C_{llr}^{NON}$ ($C_{llr}^{Random}$, computed on $C_{llr}^{NON}$). All phonemic classes appear to embed speaker discrimination power since their absence leads to a $C_{llr}^{NON}$ degradation compared to the “Random” case (indicated by a negative value of the relative $C_{llr}^{NON}$). Moreover, a consistent variation of performance is observed between the 6 phonemic classes with different extent: Withdrawing 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 comparison are still hidden.
4.2. Phonological content impact on speaker comparison per speaker pairs

In order to better suit with FVC context, we study the impact of phonological content for speaker pairs. The computation of $C^{NON}_{llr}$ for all pairs ($C_{20}=435$ pairs) shows that even though the vast majority of pairs (more than 90%) present a very low $C^{NON}_{llr}$ ($<0.01$), there exist few pairs who present a quite high $C^{NON}_{llr}$ which can reach 0.9. In Figure 3 and 4, two subsets of speaker pairs are selected, according to speaker discrimination power, in order to better visualize our results: Figure 3 (respectively Figure 4) uses a form similar to Figure 2 to present the 10 “best” (respectively “worst”) speaker pairs in term of $C^{NON}_{llr}$ value.

The mean $C^{NON}_{llr}$ for the 10 best speaker pairs is equal to $5 \times 10^{-5}$ while it is equal to 0.52 for the 10 worst pairs. Figure 3 shows that almost all the phonemic categories embed a speaker discrimination power with different extent (there is no misleading phonemic class) and more precisely oral vowels appear to convey the most important part of speaker-specific cues: withdrawing these phonemes increases the $C^{NON}_{llr}$ as shown for example for the pair “5-2”. On the other side, Figure 4 shows different outcomes: even if almost all phonemic categories still play a positive role in speaker discrimination, withdrawing nasals, vowels or consonant, from the recordings leads, surprisingly, to an improvement of $C^{NON}_{llr}$ for most of speaker pairs. For example, the pairs “24-20” and “3-21”, show a relative win of 40% and 25% respectively when nasals are withdrawn. This finding could be explained by the hypothesis of a nasal signature which corresponds to the transfer function of nasal cavities and sinuses. This nasal signature reflects mainly anatomical differences, as the speaker can only connect or not theses cavities to the vocal tract, without any controlled changes on them. Despite such inter-speaker anatomical differences, it may be possible that, for a pair of speakers, both acoustic spectra be similar. On a mathematical point of view, the question was asked for a 2-D resonator in and answered in 47 where authors found two different shapes with the same acoustic spectra. For the pair “23-13”, fricatives appears to convey a significant part of the LR performance loss. This could be explained by the use of a narrow band which exclude fricative’s speaker-specific cues in high frequencies. Another time, the global tendencies are shadowing potential speaker-specific effects.

5. CONCLUSION

This paper is dedicated to study the phonological content impact on voice comparison process in order to prevent wrongful convictions. It uses an ASR system as measurement instrument and, more particularly, the $C^{NON}_{llr}$ variations. We analyzed the influence of 6 phonemic classes: oral vowel, nasal vowel, nasal consonant, fricative, plosive and liquid using FABIOLE database, a corpora with a large number of speech recordings per speaker. In a first step, we investigated the impact of each phonemic class on speaker discrimination performance measured by $C^{NON}_{llr}$. Results showed that the 6 phonemic classes appear to embed speaker discrimination power. Moreover, a consistent variation of performance is observed between the classes with different extent: Oral vowels first, followed by nasals, are the most important classes conveying speaker specific cues. Then, liquids, fricative and plosive. This outcome is quite consistent between the 30 target speakers, with limited variations. In a second step, we explored deeply the phonological impact by focusing on speaker pairs. We proved that the vast majority of pairs (more than 90%) present a very low $C^{NON}_{llr}$ ($<0.01$) while there exist few pairs presenting a quite high $C^{NON}_{llr}$ which could reach 0.9. We showed that: (i) For the “best” discriminated pairs, all the phonemic content still play a positive role in speaker discrimination. (ii) For the “worst” speaker pairs, it appears that nasals have a negative effect and convey a significant part of LR performance loss. This could be explained by the hypothesis of similarity of acoustic spectra of nasal cavities, of the two speakers in question.

6. ACKNOWLEDGEMENTS

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


