PHONOLOGICAL CONTENT IMPACT ON WRONGFUL CONVICTIONS IN FORENSIC VOICE COMPARISON CONTEXT

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

[1]University of Avignon, LIA-CERI, Avignon, France, [2] Univ. Grenoble-Alpes, LIG , F-38000 Grenoble, France
[3]Laboratoire 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 evidence reports [1, 2]. In this framework, the *likelihood ratio* (LR) is used to present the results of the forensic expertise. The LR not only supports one of the hypothesis but also quantifies the strength of its support. The LR is calculated using Equation 1.

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

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 [3]. Even though ASR systems have achieved significant progresses in the past two decades and have reached impressive low error rates ($\approx 1\%$ [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 impunity" 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 18^{th} 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 [7]. 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 20^{th} 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 [12].

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, 24, 26] 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 /ʁ/ also performing fairly well. In [31], /s/, /t/ and /b/ 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. Phoneme classification

To conduct our work, we propose to use phoneme classes in place of individual phonemes. Working on phoneme classes presents two main advantages in the context of our study. First, to study the effect of 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 phoneme classification is describe below: oral vowels (OV) {/i/, /u/, /y/, /e/, /ɛ/, /ø/, /œ/, /o/, /ɔ/, /ɑ/.}, nasal vowels (NV) {/ɑ̃/, /ɔ̃/, /ũ/, /ɛ̃/}, nasal consonants (NC) {/m/, /n/}, plosive (P) { /p/, /t/, /k/, /b/, /d/, /g/}, fricatives (F) {/f/, /s/, /ʃ/, /v/, /z/, /ʒ/} and liquids (L) {/l/, /ɹ/}. This phoneme 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 $(100 \times 100 \times 29 = 290k)$ non-targets pairs. More details could be found in [33].

3.2. Evaluation metric

We use the C_{llr} largely used in FVC to evaluate the LR. C_{llr} is not based on hard decisions like, for example, *equal error rate* (EER) [17, 34, 35]. C_{llr} has the meaning of a cost or a loss: lower the C_{llr} is, better is the performance. C_{llr} could be calculated as follows:

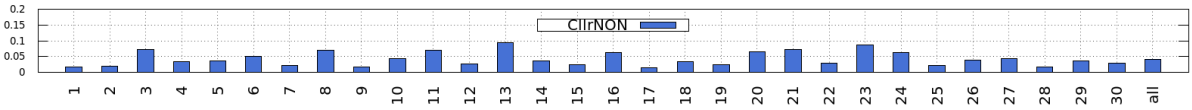

Fig. 1. *bar-plot of C_{llr}^{NON} per speaker and for "all" (all speakers pooled together).*

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

As shown in Equation 2, C_{llr} can be decomposed into the sum of two parts: C_{llr}^{NON}, which is the average information loss related to non-target trials. C_{llr}^{TAR}, which is the average information loss related to target trials. In FVC applications, the first componentwill give an idea about the risk of "wrongful conviction" and the second component will express the risk of "impunity".

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

3.3. Phoneme filtering protocol for data selection

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

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

$$C_{llr}^{R} = \frac{Cllr^{Random} - Cllr^{Specific}}{Cllr^{Random}} \times 100\% \quad (3)$$

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

3.4. Baseline LIA Systems

3.4.1. LIA speaker recognition system

In all experiments, we use as baseline the LIA_SpkDet system presented in [36].This system is developed using the ALIZE/SpkDet open-source toolkit [37, 38]. 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 [39] is applied.

3.4.2. LIA transcription system

FABIOLE database has been automatically transcribed thanks to Speeral, 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}^{R} 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 results, at the end of the speaker discrimination power scale. The results are quite consistent between the 30 target speakers, with limited variations. This outcome corroborates results of [25, 24, 26, 29], where oral vowels and nasals are found to be particularly speaker specific information.

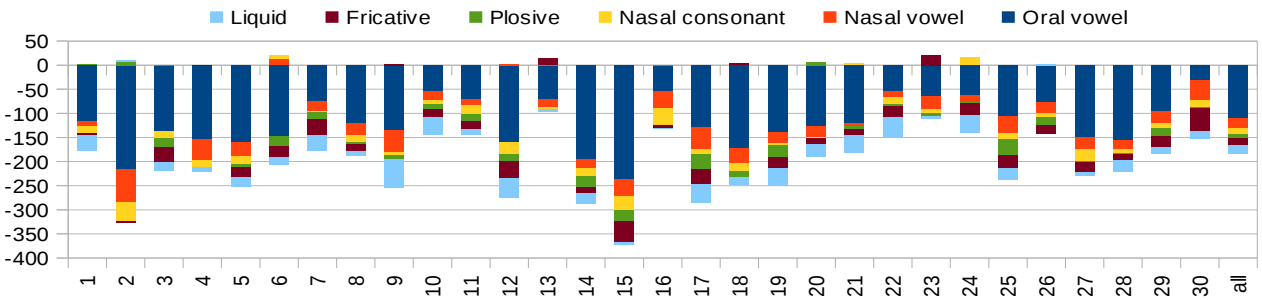

Fig. 2. Stacked bar chart of C_{llr}^{R} computed on C_{llr}^{NON} (non-target trials) per speaker and for "all".

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_{llr}^{NON} for all pairs (C_{30}^{2}=435 pairs) shows that even though the vast majority of pairs (more than 90%) present a very low C_{llr}^{NON} (<0.01), there exist few pairs who present a quite high C_{llr}^{NON} 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_{llr}^{NON} value.

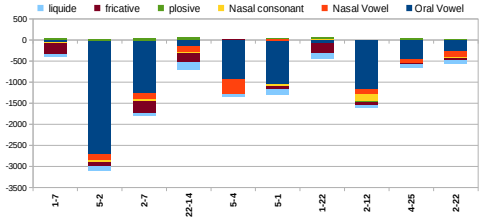

Fig. 3. Stacked bar chart of relative C_{llr}^{NON} for the 10 "best" speaker pairs, spk^{i}-spk^{j}, in term of discrimination power (mean C_{llr}^{NON}=5.10^{-5}).

The mean C_{llr}^{NON} for the 10 best speaker pairs is equal to 5×10^{-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_{llr}^{NON} 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_{llr}^{NON} 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 [43, 44] 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 [45]. 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 [46] 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.

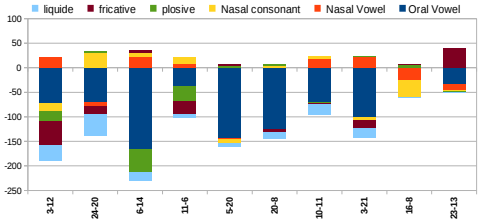

Fig. 4. Stacked bar chart of relative C_{llr}^{NON} for the 10 "worst" speaker pairs, spk^{i}-spk^{j}, in term of discrimination power (mean C_{llr}^{NON}=0.52).

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_{llr}^{NON} 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_{llr}^{NON}. 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_{llr}^{NON} (<0.01) while there exist few pairs presenting a quite high C_{llr}^{NON} 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.

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] 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.

[4] 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.

[5] Jean-François Bonastre, Frédéric Bimbot, Louis-Jean Boë, Joseph P Campbell, Douglas A Reynolds, and Ivan Magrin-Chagnolleau, "Person authentication by voice: a need for caution.," in *INTERSPEECH*, 2003.

[6] GALE Group et al., "Wests encyclopedia of american law," 2008.

[7] Joshua A Jones, "Wrongful conviction in the american judicial process: History, scope, and analysis," *Student Pulse*, vol. 4, no. 08, 2012.

[8] C Ronald Huff and Martin Killias, *Wrongful convictions and miscarriages of justice: causes and remedies in North American and European criminal justice systems*, Routledge, 2013.

[9] Gary Edmond, "Impartiality, efficiency or reliability? a critical response to expert evidence law and procedure in australia," *Australian Journal of Forensic Sciences*, vol. 42, no. 2, pp. 83–99, 2010.

[10] Adam I Kaplan, "Case for comparative fault in compensating the wrongfully convicted, the," *UCLA L. Rev.*, vol. 56, pp. 227, 2008.

[11] Craig M Cooley and Gabriel S Oberfield, "Daubert, innocence, and the future of forensic science: Increasing forensic evidence's reliability and minimizing wrongful convictions: Applying daubert isn't the only problem," *Tulsa L. Rev.*, vol. 43, pp. 285–933, 2007.

[12] C Ronald Huff, Arye Rattner, and Edward Sagarin, *Convicted but innocent: Wrongful conviction and public policy*, Sage publications, 1996.

[13] Brandon L Garrett and Peter J Neufeld, "Invalid forensic science testimony and wrongful convictions," *Virginia Law Review*, pp. 1–97, 2009.

[14] Adele Bernhard, "Justice still fails: A review of recent efforts to compensate individuals who have been unjustly convicted and later exonerated," *Drake Law Review*, vol. 52, 2004.

[15] C Huff, "Wrongful convictions: The american experience," *Canadian Journal of Criminology and Criminal Justice*, vol. 46, no. 2, 2004.

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

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

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

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

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

[21] 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.

[22] 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.

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

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

[25] 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, 2006.

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

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

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

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

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

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

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

[33] Moez Ajili, Jean-François Bonastre, Juliette Kahn, Solange Rossato, and Guillaume Bernard, "Fabiole, a speech database for forensic speaker comparison," *LREC*, 2016.

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

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

[36] 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.

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

[38] 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.

[39] 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.

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

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

[42] Moez Ajili, Jean-François Bonastre, Solange Rossato, and Juliette Kahn, "Inter-speaker variability in forensic voice comparison:a preliminary evaluation," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016.

[43] James T Wright, "The behavior of nasalized vowels in the perceptual vowel space," *Experimental phonology*, pp. 45–67, 1986.

[44] Solange Rossato, Gang Feng, and Rafaël Laboissière, "Recovering gestures from speech signals: a preliminary study for nasal vowels.," in *ICSLP*, 1998.

[45] Jianwu Dang and Kiyoshi Honda, "Acoustic characteristics of the human paranasal sinuses derived from transmission characteristic measurement and morphological observation," *The Journal of the Acoustical Society of America*, vol. 100, no. 5, pp. 3374–3383, 1996.

[46] Mark Kac, "Can one hear the shape of a drum?," *The american mathematical monthly*, vol. 73, no. 4, pp. 1–23, 1966.

[47] Carolyn Gordon, David L Webb, and Scott Wolpert, "One cannot hear the shape of a drum," *Bulletin of the American Mathematical Society*, vol. 27, no. 1, pp. 134–138, 1992.