Comparaison des voix dans le cadre judiciaire : influence du contenu phonétique

Moez Ajili[1] Jean-François Bonastre[1] Waad Ben Kheder[1] Solange Rossato[2] Juliette Kahn[3]
(1) Univ. Avignon, LIA, F-84000 Avignon France
(2) Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, F-38000 Grenoble France
(3) LNE, F-78000 Trappes France
(1) prenom.nom@univ-avignon.fr (2) prenom.nom@univ-grenoble.fr, (3) prenom.nom@lne.fr

RÉSUMÉ
En comparaison de voix dans le domaine criminalistique, l'approche Bayésienne est devenue le nouveau "golden standard". Dans cette approche, l'expert exprime ses résultats par un unique nombre, le rapport de vraisemblance (LR). Cet article s'intéresse à l'influence du contenu phonétique sur la fiabilité du LR. Nous nous intéressons particulièrement à la quantité d'information spécifique au locuteur que portent les différents sons de la parole. Cette étude met en évidence des différences importantes entre les phonèmes et, surtout, la forte influence de la variabilité intra-locuteur.

ABSTRACT
phonetic content impact on forensic voice comparison.

Forensic Voice Comparison (FVC) is increasingly using the *likelihood ratio* (LR).This article focuses on the impact of phonemic content on FVC performance and variability. The results demonstrate the importance of the phonemic content and highlight interesting differences between inter-speakers effects and intra-speaker's ones.

MOTS-CLÉS : Reconnaissance du locuteur, comparaison de voix, crimnalistique, fiabilité, contenu phonémique..

KEYWORDS: Forensic voice comparison, phonemic category, reliability..

1 Introduction

Dans les procédures judiciaires la comparaison de voix -ou "Forensic Voice Comparison (FVC)"- est de plus en plus fréquemment employée. L'approche Bayésienne est devenue le nouveau "golden standard" en sciences criminalistiques (Providers, 2009; Champod & Meuwly, 2000; Aitken & Taroni, 2004). Dans cette approche, l'expert exprime le résultat de son analyse sous la forme d'un unique nombre, le rapport de vraisemblance (LR) :

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

où E est la trace, H_p est l'hypothèse de culpabilité (même origine), et H_d est l'hypothèse d'innocence (différentes origines).

10.21437/JEP.2018-4

Ce rapport ne favorise pas seulement une des hypothèses (“culpabilité” ou “innocence”) mais il fournit également le poids de ce support. Cet article poursuit les travaux présentés dans (Ajili *et al.*, 2016b,c). Nous nous visons à hiérarchiser les catégories phonétiques des sons de parole selon la quantité d'information spécifique au locuteur qu'elles contiennent. Le système automatique de reconnaissance du locuteur est utilisé comme outil de mesure dans cette étude.

Cet article présente en section 2 une vue de la littérature consacrée à l'influence du contenu phonétique sur la caractérisation du locuteur. Le protocole expérimental est présenté dans la section 3. La section 4 présente les expériences en découlant et les résultats associés. Un focus est proposé sur l'influence de la bande passante. Enfin, la section présente des conclusions et perspectives.

2 Contenu phonétique et caractérisation du locuteur

L'étude de l'information spécifique du locuteur portées par les phonèmes individuels ou encore des classes de phonèmes a fait l'objet de différents travaux comme (Wolf, 1972; Sambur, 1975; Eatock & Mason, 1994; Hofker, 1977; Kashyap, 1976; Amino *et al.*, 2006, 2012; Antal & Toderean, 2006). Les voyelles orales et les nasales apparaissent en tête en termes de discrimination entre les locuteurs. /s/, /t/ et /b/ sont souvent évalués comme moins porteurs d'information spécifique que les voyelles et les nasales. (Magrin-Chagnolleau *et al.*, 1995) utilise déjà un système automatique de reconnaissance du locuteur pour évaluer le pouvoir discriminant des différents phonèmes. Les auteurs suggèrent que les glissements et les liquides ensemble, les voyelles - et plus particulièrement les voyelles nasales - et les consonnes nasales contiennent plus d'informations spécifiques aux locuteurs qu'un enregistrement vocal phonétiquement équilibré. (Besacier *et al.*, 2000; Gallardo *et al.*, 2014) s'appuient également sur un système automatique. Ils montrent que certaines sous-bandes de fréquence sont plus pertinentes pour caractériser les locuteurs que d'autres, soulignant ainsi l'importance de la bande passante.

3 Protocole expérimental

Cette section est dédiée au protocole expérimental original utilisé pour cette étude, qui s'appuie sur le corpus FABIOLE. Ce corpus est distribué publiquement et a été créé dans le cadre du projet FABIOLE ANR-12-BS03-0011. FABIOLE (Ajili *et al.*, 2016a) présente des caractéristiques dédiées à l'étude de la variabilité intra-locuteur. Les extraits de parole proviennent d'émissions de radio ou de télévision françaises, avec une bonne qualité sonore et présentent une durée minimale de 30 secondes de parole obtenue par concaténation de segments issus de la même émission. FABIOLE est composée d'enregistrements venant de différents types de locuteurs, incluant des journalistes, des présentateurs, des politiciens, etc. Les contenus de FABIOLE sont proches de ceux des bases REPERE (Giraudel *et al.*, 2012), ESTER 1, ESTER 2 (Galliano *et al.*, 2005) et ETAPE (Gravier *et al.*, 2012). Cette caractéristique permet d'utiliser ces bases pour l'entraînement des modèles. FABIOLE contient des enregistrements prononcés par 130 locuteurs masculins, français natifs, répartis en deux ensembles : T : 30 locuteurs cibles associés chacun à 100 extraits de parole ; I : 100 locuteurs imposteurs associés chacun à un seul extrait de parole. Les données de FABIOLE ont été automatiquement transcrites pour réaliser un étiquetage phonétique en utilisant le système Speeral (Linares *et al.*, 2007).

Dans cet article, seul l'ensemble T est utilisé.Pour chaque locuteur de T, 294950 paires d'enregistrements sont constituées. 4950 de ces paires sont des comparaisons cibles et $290k$ des comparaisons imposteurs. Les paires cibles sont obtenues en utilisant toutes les combinaisons des 100 enregis-

trements disponibles pour chaque locuteur alors que les comparaisons non-cibles apparient chaque enregistrement du locuteur cible en question (100 sont disponibles) avec chacun des enregistrements des 29 locuteurs restants, formant par conséquent $(100 \times 100 \times 29 = 290k)$ comparaisons imposteurs.

Le système de reconnaissance du locuteur utilisé est LIA_SpkDet (Matrouf *et al.*, 2007) développé avec ALIZE (Bonastre *et al.*, 2005, 2008; Larcher *et al.*, 2013), qui met en œuvre une approche I-vector (Dehak *et al.*, 2011). Les paramètres acoustiques sont composés de 19 LFCC, de leur dérivées et de 11 dérivées secondes. Une bande passante réduite à la bande téléphonique (300-3400 Hz) est utilisée pour rester proche des conditions classiques en criminalistique. Cependant, à des fins de comparaison, une bande large tirant pleinement partie de la haute qualité des enregistrements d'origine est utilisée pour une des expériences.

Un *Universal Background Model* (*UBM*) de 512 composantes a été entraîné sur Ester 1&2, REPERE et ETAPE, en utilisant des locuteurs hommes qui n'apparaissent pas dans FABIOLE. La "total variability matrix" nécessaire à l'extraction des *i-vectors* a été apprise sur les mêmes données et, enfin, un modèle PLDA est utilisé pour le scoring (Prince & Elder, 2007).

Le C_{llr} et le minimum de C_{llr}, dénoté C_{llr}^{min}, sont utilisés comme mesures de performance (Morrison, 2009; Brümmer & du Preez, 2006; Castro, 2007; Gonzalez-Rodriguez & Ramos, 2007). Le C_{llr} est défini par :

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

C_{llr} a la signification d'un coût ou d'une perte d'information (plus petit est le C_{llr}, meilleure est la performance) et peut être décomposé en deux parties additives :

— $\mathrm{C_{llr}^{TAR}}$, qui correspond à la perte moyenne relative aux paires target (le même locuteur a prononcé les deux enregistrements).

— $\mathrm{C_{llr}^{NON}}$, qui correspond à la perte moyenne relative aux paires non-target (les deux enregistrements proviennent de deux locuteurs différents).

Les scores issus d'un système automatique de reconnaissance de locuteurs doivent être "calibrés" pour devenir des LR. La transformation affine (Brümmer *et al.*, 2007) est employée pour cela, estimée en utilisant toutes les paires disponibles.

Nous avons choisi de travailler au niveau de classes de phonèmes en utilisant la classification suivante : Oral Vowels (OV) (/i/, /y/, /u/, /e/, /ø/, /o/,/ɛ/, /œ/,/ɔ/, /a/) ; Nasal vowels (NV) (/ɑ̃/, /ɔ̃/, /œ̃/, /ɛ̃/.) ; Nasal consonants (NC) (/m/, /n/) ; Plosives (P) (/p/, /t/, /k/, /b/, /d/, /g/) ; Fricatives (F) (/f/, /s/, /ʃ/, /v/, /z/, /ʒ/) et Liquides (L) (qui comprend /l/, /ʁ/).

Pour déterminer l'influence d'une classe phonétique spécifique, nous utilisons une stratégie de "knock-out" : la part de signal correspondant à la catégorie étudiée est retirée des enregistrements et la perte de performance indique alors l'influence de celle-ci. Par conséquence, nous réalisons tout un jeu d'expériences dans lesquelles le matériel de parole correspondant à une classe phonétique spécifique est retiré des deux enregistrements composant une paire de comparaison de voix. La condition expérimentale correspondante est dénommée "**Specific**". La quantité de données correspondant à une catégorie spécifique est fortement variable d'une catégorie à l'autre mais également d'un enregistrement à l'autre (par exemple, dans nos expériences, les consonnes nasales représentent

6% du signal de parole alors que les voyelles orales pèsent pour 36%). Pour pallier ce biais, nous créons une condition de contrôle dénommée "**Random**", où la même quantité de signal est supprimée aléatoirement des enregistrements. Plus précisément, pour chaque enregistrement, quand un certain pourcentage de trames de parole est supprimé pour la condition '**Specific**", le même pourcentage est sélectionné aléatoirement et supprimé pour la condition "**Random**". Par précaution, ce processus est répété 20 fois, créant 20 fois plus de paires dans la condition "**Random**" que pour la condition "**Specific**".

L'impact d'une classe phonétique donnée est estimé relativement par C_{llr}^{R} :

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

Une valeur positive de C_{llr}^{R} indique que la classe phonétique étudiée apporte moins d'information spécifique du locuteur que la condition de controle, une valeur négative montre au contraire que cette classe phonétique apporte plus d'information spécifique du locuteur que le contenu moyen.

4 Expériences et résultats

Nous avons tout d'abord calculé pour référence la performance sur l'ensemble des locuteurs et sur les contenus complets. Le C_{llr} global correspondant est de 0.12631 bits (et l'EER de 2.88%).

TABLE 1 – C_{llr} et $\text{C}_{\text{llr}}^{\text{min}}$ pour les conditions "Specific" et "Random" (C_{llr}=0.126 et $\text{C}_{\text{llr}}^{\text{min}}$=0.117 pour la référence).

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

Nous avons ensuite abordé notre analyse par classe phonétique avec le protocole de "knock-out" décrit précédemment. La table 1 montre l'impact des 6 catégories phonémiques sur C_{llr} pour les conditions "Specific" et "Random"Elle fournit également la quantité de trames de parole pour chaque classe de phonèmes (moyenne et écart par rapport aux extraits de paroles). Une grande variation est observée entre les classes phonémiques : le retrait des voyelles nasales, des consonnes nasales ou des voyelles orales conduit à une perte d'information par rapport à la condition de contrôle ("**Random**") tandis que l'absence de plosives, liquides et fricatives n'a pas d'impact significatif sur la précision du système. Ce résultat est en concordance avec notre revue de la littérature.

Cependant, les fricatives présentent un faible pouvoir discriminatoire, un résultat en contradiction avec la littérature dont, notamment, (Gallardo *et al.*, 2014). La table 2 reprend la même expérience mais en bande large et en se focalisant sur les fricatives et les voyelles orales. Cette fois-ci, les fricatives montrent un pouvoir discriminant supérieur à la moyenne alors que les voyelles orales perdent cette caractéristique. Ce résultat montre clairement -et ce n'est pas réellement une surprise- l'importance de la bande passante. Au vu du contexte "forensique" de ce travail, seuls des résultats obtenus en bande téléphonique sont présentés dans la suite de ce document.

TABLE 2 – Valeurs de C_{llr} et C_{llr}^{min} pour les conditions “Specific” et “Random” pour les fricatives et les voyelles orales en bande large

	C_{llr}		C_{llr}^{min}		Duration (s)	
Category	**Withdrawn**		**Withdrawn**		**Mean**	**SD**
	Specific	**Random**	**Specific**	**Random**		
F	0.11738	0.11334	0.10713	0.10394	5.84	2.68
OV	0.11845	0.12216	0.11127	0.10824	13.00	5.50

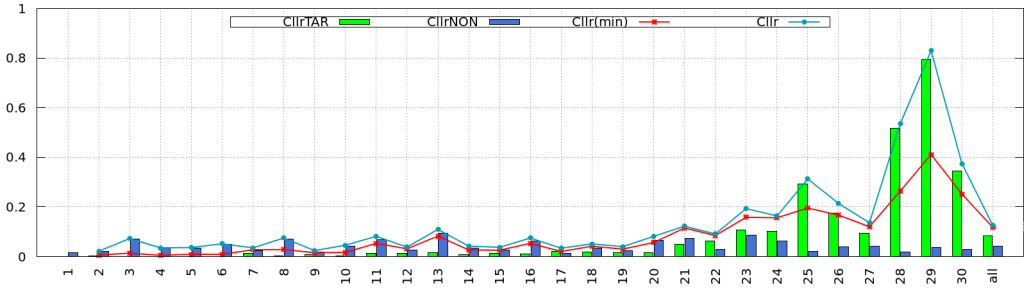

FIGURE 1 – C_{llr}, C_{llr}^{min}, C_{llr}^{TAR}, C_{llr}^{NON} *par locuteur et pour “all” (les données de tous les locuteurs sont utilisées).*

Nous poursuivons notre analyse en réalisant un focus par locuteur et en fonction de la catégorie des comparaisons, cible ou imposteur. La figure 1 résume les résultats.Elle montre que la perte d’information liée aux comparaisons non-cibles (mesurée par C_{llr}^{NON}) présente une variation assez faible en fonction du locuteur alors que la variation est importante pour les comparaisons cibles (mesurée par C_{llr}^{TAR}). Notons également que la perte d’information provenant des essais cibles (calculée par C_{llr}^{TAR}) est principalement responsable des coûts élevés attachés à certains locuteurs.

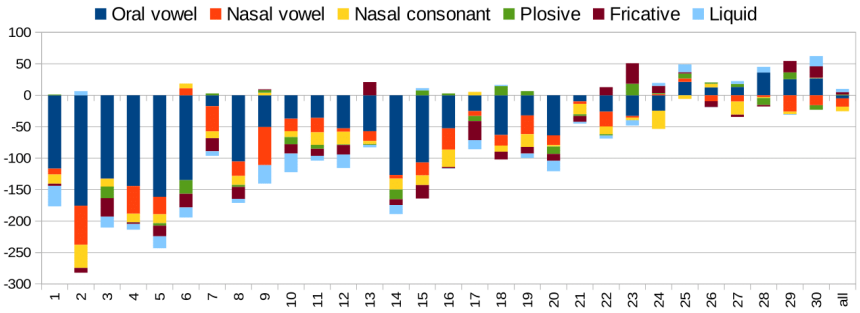

FIGURE 2 – C_{llr}^{R} *par locuteur et “all”.*

La figure 2 présente la contribution de chaque classe phonétique au C_{llr}^{R} en fonction du locuteur. La tendance générale rejoint les résultats présentés dans la table 1 mais une grande variabilité des résultats par classes phonétiques en fonction du locuteur considéré est à noter. Par exemple, le locuteur 2 montre une perte relative de 175% lorsque les voyelles orales sont supprimées quand le 28 accepte un gain de 40% dans la même situation.

La figure 3 montre l’impact de chaque classe phonétique en termes de C_{llr}^{NON}, le C_{llr}^{R} relatif calculé sur C_{llr}^{NON} en utilisant les seules paires non-target. Cet indice est supposé être lié principalement au pouvoir de discrimination entre les locuteurs. Les 6 classes phonétiques apparaissent porteuses de pouvoir de discrimination. Les voyelles orales arrivent en premier en termes de pouvoir de discrimination, avec une large avance sur les classes suivantes. Les nasales, voyelles en tête suivies par les consonnes, prennent les places suivantes. Les classes restantes ont un comportement similaire bien que porteuses de moins d’information discriminante du locuteur. Les résultats obtenus sont plutôt consistent pour les 30 locuteurs étudiés.

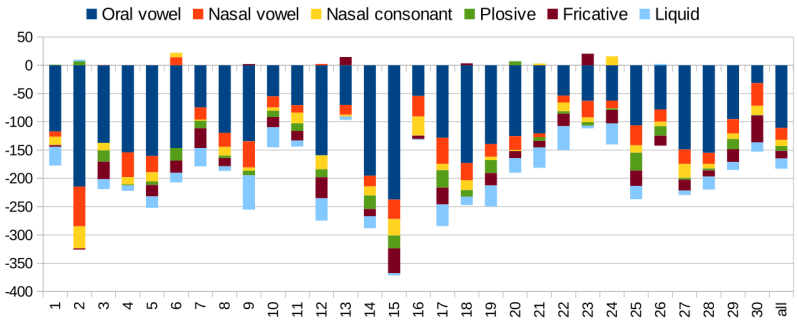

FIGURE 3 – $C_{\text{llr}}^{\text{R}}$ *calculé sur* $C_{\text{llr}}^{\text{NON}}$ *par locuteur et "all".*

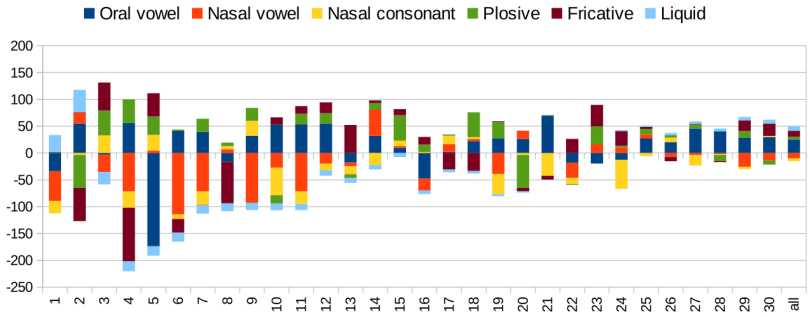

FIGURE 4 – $\text{C}_{\text{llr}}^{\text{R}}$ calculé sur $\text{C}_{\text{llr}}^{\text{TAR}}$ par locuteur et "all".

La figure 4 utilise un principe similaire à la figure 3. Elle présente l'impact des classes phonémiques par locuteur, en termes de $\text{C}_{\text{llr}}^{\text{TAR}}$. $\text{C}_{\text{llr}}^{\text{TAR}}$ est calculé uniquement en utilisant des comparaisons cibles et, grâce à notre protocole offrant un grand nombre par locuteur de ces tests, traduit directement les effets de la variabilité intra-locuteur. En opposition avec les résultats précédents, pour ce $\text{C}_{\text{llr}}^{\text{TAR}}$, retirer les voyelles orales des enregistrements conduit à une amélioration de C_{llr} pour environ 70% des locuteurs. Ces sons semblent donc, ici, apporter peu en termes de discrimination du locuteur ou même perturber la discrimination. Les classes fricatives, liquides et plosives ont le même comportement que les voyelles orales. Au contraire, les nasales (et en particulier les voyelles nasales) jouent un rôle positif : retirer ces phonèmes augmente le C_{llr}.

Le rôle positif des nasales pour la comparaison des locuteurs pourrait s'expliquer par la contribution importante des cavités nasales et para-nasales. Cet aspect morphologique propre à ces phonèmes constitue un élément que les locuteurs peuvent difficilement contrôler (volontairement ou involontairement). Cela induit une faible variabilité intra-locuteur, pour une variabilité inter-locuteur significative. (Stevens, 1999; Schindler & Draxler, 2013).

Les voyelles orales apportent la plus grande partie en termes de pouvoir de discrimination des locuteurs mais présente en même temps une grande variabilité intra-locuteur, liée à une part significative des pertes de performance. La variabilité intra-locuteur apporte en général environ deux tiers des pertes de C_{llr} (0,66 contre 0,33 pour les pertes portées par la variabilité inter-locuteur). Cette proportion est significativement plus élevée (jusqu'à 0,94 vs 0,06) pour les locuteurs qui présentent la plus grande contribution à la perte de C_{llr}. Il est intéressant de lier cette constatation à deux faits : (1) presque toutes les classes de phonèmes étudiées aident à la discrimination des locuteurs pour tous les locuteurs ; (2) certaines classes de phonèmes dégradent la partie cible de C_{llr} lorsque d'autres classes offrent un comportement positif. Il est intéressant de remarquer que la même classe de phonèmes peut avoir un comportement très différent selon le locuteur, ce qui renforce la nécessité d'une prise en compte fine de l'effet locuteur.

5 Conclusion

Cet article est consacré à l'étude de impact du contenu phonémique sur le processus de comparaison vocale. Pour cela, il utilise un système automatique de reconnaissance du locuteur comme instrument de mesure et un protocole de "knock-out" et une bande téléphonique appropriée au contexte criminalistique de cette étude.

Nous avons étudié l'impact de chaque classe phonémique sur la performance de la comparaison vocale mesurée avec le critère C_{llr}. Les résultats ont montré que toutes les classes phonémiques jouent un rôle en termes de pouvoir de discrimination du locuteur. Les voyelles orales, les voyelles nasales et les consonnes nasales, dans cet ordre, sont meilleures que le contenu phonémique moyen en termes de performance de comparaison vocale, rejoignant des constats précédents. Les fricatives, par contre, n'apportent pas plus qu'un contenu moyen. Ce résultat surprenant par rapport à la littérature a été expliqué par le choix d'une bande passante étroite : en bande large, cette catégorie retrouve sa pertinence connue en termes de discrimination des locuteurs.

Lorsque nous nous sommes concentrés sur la variabilité intra-locuteur, les voyelles orales sont apparues liées à un niveau élevé de C_{llr} intra-locuteur. Nous avons vu précédemment que cette classe phonémique apportait une grande partie du pouvoir de discrimination du locuteur mais elle apparaît également très sensible à la variabilité intra-locuteur. En revanche, les nasales ont montré une bonne capacité de discrimination et, en même temps, apparaissent robustes en ce qui concerne la variabilité intra-locuteur.

Dans cet article, nous avons souligné à plusieurs reprises l'importance du facteur locuteur. Nous avons observé de grandes variations de C_{llr} et de C_{llr}^{TAR} entre nos 30 locuteurs. Nous avons également observé en fonction du locuteur des comportements très différents du système en termes de C_{llr}^{TAR} suivant les classes phonémiques sélectionnées.

La principale conclusion du travail présenté est une remise en cause nécessaire des protocoles d'évaluation habituels en reconnaissance automatique du locuteur (ASpR). Ces protocoles se basent en effet principalement sur la discrimination des locuteurs (C_{llr}^{NON}) en négligeant largement la variabilité intra-locuteur (C_{llr}^{TAR}) Il nous apparaît obligatoire de prendre en considération de manière approfondie la variabilité intra-locuteur ainsi que le facteur locuteur lui-même. Cela est particulièrement important lorsqu'il s'agit d'attester de la fiabilité d'une solution de comparaison de voix dans le domaine criminalistique.

Les résultats présentés dans cet article restent préliminaires (100 extraits de parole par locuteur, peu de variabilité contextuelle et seulement 30 locuteurs masculins, français natifs). Dans nos futurs travaux, nous souhaitons effectuer une analyse similaire sur une base de données supérieure d'un ordre de magnitude (~1000 enregistrements par locuteurs et plusieurs centaines de locuteurs). Une telle base permettrait également d'affiner notre étude en nous intéressant aux phonèmes individuels au lieu de classes phonémiques. Enfin, ce travail nous permet de promouvoir un système automatique de comparaison de voix **explicite** et **transparent** capable d'analyser en profondeur le contenu phonétique des extraits de discours et de détailler ses sorties dans un langage compréhensible par un expert. Loin d'être un simple rêve, de telles caractéristiques sont certainement un élément indispensable pour une utilisation responsable de systèmes de comparaison de voix en milieu criminalistique.

6 Remerciements

La recherche rapportée ici a été soutenue par le projet ANR-12-BS03-0011 FABIOLE.

Références

AITKEN C. G. & TARONI F. (2004). *Statistics and the evaluation of evidence for forensic scientists*, volume 10. Wiley Online Library.

AJILI M., BONASTRE J., KAHN J., ROSSATO S. & BERNARD G. (2016a). 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.*

AJILI M., BONASTRE J.-F., BEN KHEDER W., ROSSATO S. & KAHN J. (2016b). Phonetic content impact on forensic voice comparison. In *Spoken Language Technology Workshop (SLT), 2016 IEEE*, p. 210–217 : IEEE.

AJILI M., F. BONASTRE J., ROSSATTO S. & KAHN J. (2016c). Inter-speaker variability in forensic voice comparison : A preliminary evaluation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 2114–2118.

AMINO K., OSANAI T., KAMADA T., MAKINAE H. & ARAI T. (2012). Effects of the phonological contents and transmission channels on forensic speaker recognition. In *Forensic Speaker Recognition*, p. 275–308. Springer.

AMINO K., SUGAWARA T. & ARAI T. (2006). Idiosyncrasy of nasal sounds in human speaker identification and their acoustic properties. *Acoustical science and technology*, **27**(4), 233–235.

ANTAL M. & TODEREAN G. (2006). Speaker recognition and broad phonetic groups. In *SPPRA*, p. 155–159.

BESACIER L., BONASTRE J.-F. & FREDOUILLE C. (2000). Localization and selection of speaker-specific information with statistical modeling. *Speech Communication*, **31**(2), 89–106.

BONASTRE J.-F., SCHEFFER N., MATROUF D., FREDOUILLE C., LARCHER A., PRETI A., POUCHOULIN G., EVANS N. W., FAUVE B. G. & MASON J. S. (2008). Alize/spkdet : a state-of-the-art open source software for speaker recognition. In *Odyssey*, p.20.

BONASTRE J.-F., WILS F. & MEIGNIER S. (2005). Alize, a free toolkit for speaker recognition. In *ICASSP (1)*, p. 737–740.

BRÜMMER N., BURGET L., ČERNOCKỲ J. H., GLEMBEK O., GREZL F., KARAFIAT M., VAN LEEUWEN D. A., MATĚ P., SCHWARZ P. & STRASHEIM A. (2007). 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*, **15**(7), 2072–2084.

BRÜMMER N. & DU PREEZ J. (2006). Application-independent evaluation of speaker detection. *Computer Speech & Language*, **20**(2), 230–275.

CASTRO D. R. (2007). *Forensic evaluation of the evidence using automatic speaker recognition systems*. PhD thesis, Universidad autónoma de Madrid.

CHAMPOD C. & MEUWLY D. (2000). The inference of identity in forensic speaker recognition. *Speech Communication*, **31**(2), 193–203.

DEHAK N., KENNY P., DEHAK R., DUMOUCHEL P. & OUELLET P. (2011). Front-end factor analysis for speaker verification. *Audio, Speech, and Language Processing, IEEE Transactions on*, **19**(4), 788–798.

EATOCK J. P. & MASON J. S. (1994). 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*, volume 1, p. I–133 : IEEE.

GALLARDO L. F., WAGNER M. & MÖLLER S. (2014). I-vector speaker verification based on phonetic information under transmission channel effects. In *INTERSPEECH*, p. 696–700.

GALLIANO S., GEOFFROIS E., MOSTEFA D., CHOUKRI K., BONASTRE J.-F. & GRAVIER G. (2005). The ester phase ii evaluation campaign for the rich transcription of french broadcast news. In *European Conference on Speech Communication and Technology*, p. 1149–1152.

GIRAUDEL A., CARRÉ M., MAPELLI V., KAHN J., GALIBERT O. & QUINTARD L. (2012). The repere corpus : a multimodal corpus for person recognition. In *LREC*, p. 1102–1107.

GONZALEZ-RODRIGUEZ J. & RAMOS D. (2007). Forensic automatic speaker classification in the "coming paradigm shift". In *Speaker Classification I*, p. 205–217. Springer.

GRAVIER G., ADDA G., PAULSON N., CARRÉ M., GIRAUDEL A., GALIBERT O. *et al.* (2012). The etape corpus for the evaluation of speech-based tv content processing in the french language. *International Conference on Language Resources, Evaluation and Corpora*.

HOFKER U. (1977). Auros-automatic recognition of speakers by computers : phoneme ordering for speaker recognition. In *Proc. 9th International Congress on'Acoustics, Madrid*, p. 506–507.

KASHYAP R. (1976). Speaker recognition from an unknown utterance and speaker-speech interaction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **24**(6), 481–488.

LARCHER A., BONASTRE J.-F., FAUVE B. G., LEE K.-A., LÉVY C., LI H., MASON J. S. & PARFAIT J.-Y. (2013). Alize 3.0-open source toolkit for state-of-the-art speaker recognition. In *INTERSPEECH*, p. 2768–2772.

LINARES G., NOCÉRA P., MASSONIE D. & MATROUF D. (2007). The lia speech recognition system : from 10xrt to 1xrt. In *International Conference on Text, Speech and Dialogue*, p. 302–308 : Springer.

MAGRIN-CHAGNOLLEAU I., BONASTRE J.-F. & BIMBOT F. (1995). Effect of utterance duration and phonetic content on speaker identification usind second order statistical methods. In *Proceedings of EUROSPEECH*.

MATROUF D., SCHEFFER N., FAUVE B. G. & BONASTRE J.-F. (2007). A straightforward and efficient implementation of the factor analysis model for speaker verification. In *INTERSPEECH*, p. 1242–1245.

MORRISON G. S. (2009). Forensic voice comparison and the paradigm shift. *Science & Justice*, **49**(4), 298–308.

PRINCE S. J. & ELDER J. H. (2007). Probabilistic linear discriminant analysis for inferences about identity. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, p. 1–8 : IEEE.

PROVIDERS A. (2009). Standards for the formulation of evaluative forensic science expert opinion. *Sci. Justice*, **49**, 161–164.

SAMBUR M. (1975). Selection of acoustic features for speaker identification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **23**(2), 176–182.

SCHINDLER C. & DRAXLER C. (2013). The influence of bandwidth limitation on the speaker discriminating potential of nasals and fricatives. *International Association for Forensic Phonetics and Acoustics (IAFPA)*.

STEVENS K. (1999). Acoustic phonetics. 1998.

WOLF J. J. (1972). Efficient acoustic parameters for speaker recognition. *The Journal of the Acoustical Society of America*, **51**(6B), 2044–2056.