Identification du locuteur : ouvrir la boîte noire

Carole Millot, Cédric Gendrot, Jean-Francois Bonastre

Identification du locuteur : ouvrir la boîte noire

Carole Millot[1, 2*] Cédric Gendrot[2] Jean-François Bonastre[1, 3]
(1) Inria, Domaine de Voluceau, 78153 Le Chesnay, France
(2) Laboratoire de Phonétique et Phonologie (CNRS & Sorbonne Nouvelle), 4 rue des Irlandais, 75005 Paris, France
(3) Laboratoire Informatique d'Avignon, EA 4128, Avignon Université, Avignon, France
{carole.millot, cedric.gendrot}@sorbonne-nouvelle.fr,
jean-francois.bonastre@inria.fr

RÉSUMÉ

L'explicabilité des systèmes relevant du *deep learning* est devenue un enjeu central ces dernières années, dans le droit européen comme le domaine criminalistique. L'approche BA-LR introduit en identification du locuteur un nouveau paradigme de modélisation : elle fait émerger automatiquement les attributs partagés par un groupe de locuteurs et qui sous-entendent la discrimination de ceux-ci. Le score produit est décomposable au niveau des attributs, ce qui augmente significativement l'explicabilité de la méthode. Cette étude propose de compléter la caractérisation des attributs obtenus par le BA-LR, à l'aide de paramètres de qualité de voix. L'analyse suggère que plusieurs attributs utilisent les types de phonation pour regrouper les locuteurs, ceux-ci encodant des informations humainement perceptibles. Cet article pose ainsi des bases pour l'analyse acoustique des attributs, qui permettra à terme d'utiliser le BA-LR dans le cadre du profilage vocal.

ABSTRACT

Opening the black box in speaker recognition

Explaining how deep-learning-based systems work has recently become a central issue, as seen in European legislation and forensic research. The BA-LR approach introduces a new paradigm for speaker recognition, discriminating speakers by bringing out binary attributes shared between them. The obtained score can be broken down at the attribute level, augmenting significantly the BA-LR explainability. This study aims to characterize the attributes proposed by BA-LR, with the help of voice quality parameters. Analysis suggests that multiple attributes use phonation types to group speakers : this shows attributes can be phonetically characterized, and encode humanly perceptible informations. This paper lays foundations for acoustic analysis of binary attributes, which may eventually permit using BA-LR for voice profiling.

MOTS-CLÉS : qualité de voix, traitement automatique de la parole, explicabilité, perception de la parole, reconnaissance du locuteur.

KEYWORDS: voice quality, speech processing, explainability, speech perception, speaker recognition.

*. Thèse co-financée par la Direction générale de l'armement à travers l'Agence de l'innovation de défense.

1 Introduction

La reconnaissance automatique du locuteur consiste à reconnaître ou vérifier l'identité d'une personne à partir d'un échantillon de sa voix. La comparaison de voix s'inscrit dans ce champ et détermine si deux enregistrements de parole ont été produits par le même locuteur, ou deux locuteurs différents. Les systèmes *state of the art* (état de l'art) de reconnaissance du locuteur sont basés sur l'apprentissage d'un modèle *deep learning* (apprentissage profond), appris sur de grandes bases données de locuteurs (Bai & Zhang (2021), Kwon *et al.* (2021)). Leurs performances sont excellentes (Sarni *et al.*, 2023), mais ils ne fournissent aucun élément d'information permettant d'expliquer leur score (Campbell *et al.*, 2009). L'explicabilité est pourtant un enjeu central pour la vérification du locuteur, par exemple dans une optique criminalistique pour la vérification du locuteur (Ben Amor & Bonastre (2022b), Ben Amor *et al.* (2023)) ou de manière générale, pour toutes les activités dîtes « *High Risk* » dans le cadre de l'*AI Act* (Sovrano *et al.*, 2022). En réponse à cette limitation, l'approche BA-LR, a été récemment proposée (Ben Amor & Bonastre, 2022a). Elle représente un enregistrement audio par la présence ou l'absence d'attributs de voix dans celui-ci. Les attributs sont issus d'un ensemble fermé déterminé automatiquement (*bottom-up*) à partir d'une approche de *deep learning* appliquée sur une base de données de plus d'un million d'enregistrements. BA-LR propose comme score un Rapport de Vraisemblance (LR) entre la probabilité pour qu'un seul locuteur ait prononcé les deux enregistrements, versus l'hypothèse inverse. Ce score n'est basé que sur la présence (activation) ou l'absence des attributs dans les deux fichiers et sur les caractéristiques des attributs. Ce paradigme favorise l'explicabilité intrinsèque car la participation de chaque attribut à la décision est connue et est issue des caractéristiques de celui-ci, apprises durant l'entraînement (rareté et fiabilité d'extraction).

Caractériser la nature des informations encodées par ces attributs découverts par un système automatique est important dans le cadre de cette démarche. Nous conjecturons que la qualité de voix fait partie des paramètres pris en compte par le système pour définir les attributs. En effet, Kreiman (Kreiman *et al.* (2003), Lee & Kreiman (2019)) définit la qualité de voix comme la façon « dont les locuteurs projettent leur identité — leurs caractéristiques physiques, psychologiques, et sociales — au monde ». La qualité de voix peut être décomposée en différents corrélats acoustiques et perceptuels (Barsties & De Bodt, 2015), et est liée à des paramètres linguistiques tels que la nasalité ou encore le type de phonation (Lee & Kreiman, 2022).
Les types de phonation qualifient les différents positionnements possibles de la glotte pendant la phonation. On y compte la voix modale, mais aussi les voix craquée (présence de vibrations voisées irrégulières) et soufflée (présence importante de bruit dans le signal), ainsi que tendue et relâchée (Gordon & Ladefoged, 2001). Les types de phonation peuvent être liés à des variations d'ordres divers : le sexe est un facteur important — on retrouve souvent plus de souffle dans la voix des femmes du fait de la fermeture incomplète (*glottal chink*) de leur plis vocaux (Hanson & Chuang, 1999). La langue parlée par un locuteur (Benoist-Lucy & Pillot-Loiseau, 2013) et son appartenance à une communauté sociale ou géographique sont d'autres influences impactant le type de phonation, comme le cas de jeunes femmes étasuniennes utilisant la voix craquée (Greer & Winters, 2015).

Cette étude s'axe autour de deux enjeux. Le premier est la caractérisation d'attributs de la voix discriminants au sens du locuteur, par des paramètres de la qualité de la voix, ici les types de phonation. Le second est le développement d'une méthodologie cohérente pour l'étude des attributs découverts par un processus automatique.
La corrélation entre les différents attributs extraits par le BA-LR et les types de phonation est étudiée, suivie d'une analyse révélant les paramètres acoustiques pris en compte par le système automatique. Les liens avec d'autres attributs et le sexe des locuteurs sont également analysés.

2 Méthode

2.1 Annotation du corpus

Afin d'extraire les attributs de chaque extrait de parole, le système automatique BA-LR « standard » (Ben Amor & Bonastre, 2022a) est utilisé. Le modèle de celui-ci a été appris à partir de données du corpus anglophone VoxCeleb2 (Nagrani *et al.*, 2017), une base de plus d'un million d'enregistrements produits par plus de 6000 locuteurs.
L'étude est réalisée à partir du corpus francophone PTSVOX (Chanclu *et al.*, 2020), sélectionné en raison de son nombre de locuteurs, 369, permettant d'avoir un vaste éventail de profils vocaux. Chaque locuteur est enregistré pendant deux à quatre minutes, avec une à quatre sessions par locuteur (minimum deux pour les 24 premiers). Comme évoqué précédemment, le modèle BA-LR a été appris sur de l'anglais et est appliqué sur des données francophones. Bien que constituant une limitation potentielle, cette utilisation est justifiée par plusieurs expériences précédentes.
Dans le cadre de ce travail, le corpus a été annoté selon les types de phonation présents dans les enregistrements pour établir les profils vocaux des locuteurs. Pour chaque enregistrement, l'annotateur [1] a attribué une étiquette de profil (craqué, soufflé, modal) perceptuellement pour chaque locuteur. Pour attribuer une étiquette craquée ou soufflée, le type de phonation doit être présent pendant environ les deux tiers des données du locuteur. Un groupe témoin composé d'extraits de 100 locuteurs sélectionnés aléatoirement est également composé.
Dans un deuxième temps, les enregistrements sont décomposés en extraits de trois secondes à l'aide d'un script Praat (Boersma, 2001) qui supprime les pauses dans l'enregistrement. Chaque extrait est ré-annoté selon le type de phonation présent dans celui-ci. Un accord inter-annotateurs est calculé à l'aide d'un autre évaluateur sur 10% des données à l'aide du *package* `psych` dans R (R Core Team, 2023; Revelle, 2024), dont le Kappa de Cohen résultant est de $\kappa = 0,79$. Les extraits présentant le type de phonation renseigné dans le profil du locuteur sont alors sélectionnés afin de réduire le bruit dans les données et la variation intra-locuteur.
La Table 1 montre la répartition des locuteurs en groupes de types de phonation, le nombre d'extraits vocaux par groupe, ainsi que le pourcentage de femmes pour chaque groupe.

Profil vocal du locuteur	Nombre de locuteurs	Nombre d'extraits	% femmes
Profil craqué	38 locuteurs	4227 extraits	16%
Profil soufflé	31 locuteurs	3763 extraits	58%
Profil modal	32 locuteurs	4510 extraits	40%
Groupe témoin	100 locuteurs	8332 extraits	40%
Total	201 locuteurs	20832 extraits	

TABLE 1 – Tableau montrant la répartition des locuteurs en groupes de type de phonation. Le nombre d'extraits correspondants et le pourcentage de femmes dans chaque groupe sont également présentés.

2.2 Extraction des attributs

Le système BA-LR, dérivé des x-vecteurs, représente un extrait de parole (de trois secondes ici) par un *embedding* neuronal de 256 coefficients, ensuite binarisés. Seuls 206 coefficients sont conservés après suppression des coefficients inactifs. Un 1 indique la présence d'un attribut, un 0 son absence.

1. L'annotation a été réalisée par Carole Millot.

3 Résultats

Afin de comparer le comportement des attributs en fonction des groupes de types de phonation, des moyennes d'activation pour chaque attribut sont calculées par groupe de locuteurs. Le groupe témoin permet d'établir des taux d'activation étalons pour chaque attribut. Un aperçu de l'ensemble des taux d'activation des attributs pour chaque groupe de locuteurs est visible Figure 1.

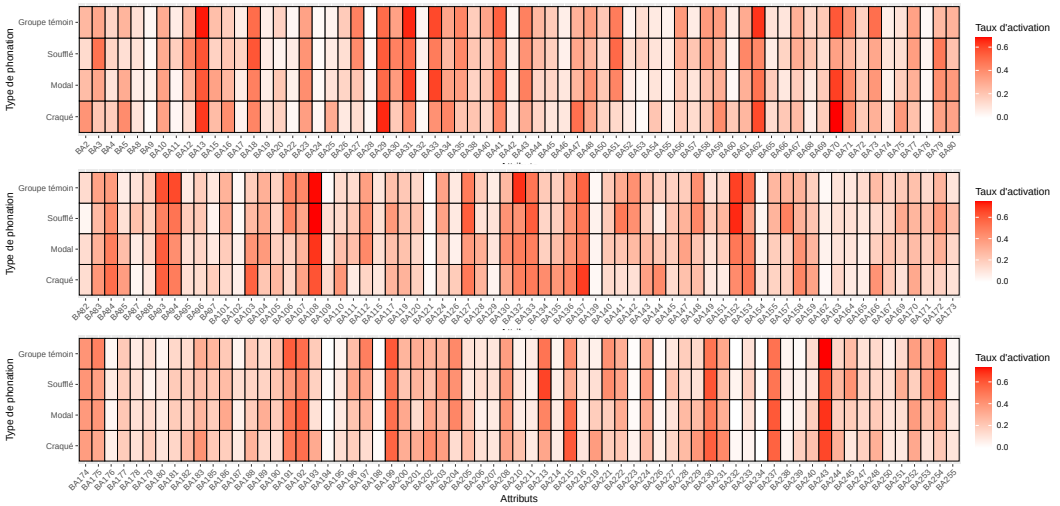


FIGURE 1 – Taux d'activation moyen des 206 attributs pour chaque groupe de locuteurs.

Les moyennes d'activation de chaque attribut pour les trois groupes de type de phonation sont ensuite divisées par la moyenne d'activation du groupe témoin. Les attributs avec les plus forts écarts entre les ratios pour le groupe craqué/témoin et le groupe soufflé/témoin sont sélectionnés pour la suite de cette étude. Afin d'établir un seuil, un critère perceptuel est appliqué : pour les attributs avec les différences de ratio les plus importantes, une écoute de chaque extrait est réalisée pour vérifier si le type de phonation (voix craquée ou de la voix soufflée) est perçu dans l'ensemble des extraits du groupe, afin d'établir un lien perceptuel entre attribut et type de phonation. Ainsi, seuls les huit attributs comportant les différences de ratio les plus importantes (au moins neuf dixièmes de points) sont retenus pour la suite de l'analyse : BA3, 5, 27, 51, 85, 141, 144 et 157. Ils sont renseignés dans la Table 2.

Attribut	Ratio craqué/témoin	Ratio modal/témoin	Ratio soufflé/témoin
BA3	$0,65$	1	$1,55$
BA5	$1,53$	$1,12$	$0,42$
BA27	$0,48$	$0,78$	$1,48$
BA51	$0,33$	1	$1,21$
BA85	$4,63$	$3,13$	$1,50$
BA141	$0,41$	$0,72$	$1,56$
BA144	$2,50$	$1,50$	$0,56$
BA157	$0,62$	$0,59$	$1,66$

TABLE 2 – Tableau représentant les huit attributs sélectionnés pour l'analyse, et le ratio entre les taux d'activation moyens pour les groupes de type de phonation et le groupe témoin.

3.1 Interactions par attribut

Pour chacun des extraits de trois secondes, des mesures acoustiques sont extraites par openSMILE via l'ensemble de paramètres eGeMAPS (Eyben *et al.*, 2010) qui contient 88 paramètres acoustiques. La mesure de la fréquence fondamentale f_0 en demi-tons, le Ratio Harmoniques/Bruit (HNR), ou encore la différence $h_1 - h_2$ font partie de ces paramètres. $h_1 - h_2$ est utilisée pour évaluer le type de phonation d'un extrait et calcule la différence entre la première et la seconde harmonique d'un spectre (Keating *et al.*, 2010). Le HNR permet d'estimer le taux de bruit dans l'extrait, élevé pour les voix craquées et soufflées, par rapport au taux d'harmoniques (Davidson, 2019). Ces mesures sont exploitées à l'aide du *package* `lme4` (Bates *et al.*, 2015) afin de construire les modèles mixtes des interactions entre taux d'activation des attributs et mesures acoustiques. Les modèles mixtes permettent de contrôler la significativité des interactions calculées dans de grands corpus, grâce à l'inclusion d'effets aléatoires et de prédicteurs [2]. Les résultats sont regroupés dans la Table 3 et décrits dans les paragraphes ci-après. L'interaction est considérée significative à partir de $p < 0.005$ **.

	p-value			
Attribut	**Sexe**	f_0	$h_1 - h_2$	**HNR**
BA3	0.043 *	0.101	0.153	0.064
BA5	0.413	0.102	0.000259 ***	0.0015 **
BA27	0.031 *	0.025 *	0.882	0.000483 ***
BA51	0.362	0.164	0.000207 ***	2e-16 ***
BA85	0.039 *	0.561	0.00017 ***	1.05e-07 ***
BA141	6.31e-05 ***	6.31e-05 ***	0.319	0.129
BA144	6.31e-05 ***	6.31e-05 ***	0.055	0.052
BA157	6.31e-05 ***	6.31e-05 ***	0.056	0.021 *
206 attributs	0.810	0.085	0.975	0.020 *

TABLE 3 – Tableau présentant les p-values obtenues à partir des modèles mixtes calculés pour chacun des huit attributs étudiés. La significativité des interactions est évaluée avec le *package* `lmerTest` (Kuznetsova *et al.*, 2017) à l'aide de l'approximation de Satterthwaite.

Le sexe est un paramètre pris en compte de manière explicite ou implicite par les systèmes de reconnaissance de locuteurs du fait de son haut potentiel de discrimination (Jacquelin *et al.*, 2023), ce qui peut introduire un biais dans l'analyse présentée. De plus, les groupes de locuteurs étudiés ne contiennent pas le même ratio d'hommes et de femmes (Table 1). Afin de vérifier que l'information retenue par les huit attributs ne dépend pas du sexe, un modèle mixte est utilisé pour calculer l'interaction entre les attributs et le sexe, avec le type de phonation en prédicteur fixe et le locuteur en variable aléatoire. L'interaction entre les huit attributs et le sexe est significative avec $p < 0.001$ pour trois attributs (Table 3) : BA141, BA144 et BA157.

La fréquence fondamentale est le paramètre acoustique le plus proéminent pour la prédiction du sexe d'un locuteur (Jacquelin *et al.*, 2023). L'interaction entre la f_0 (en demi-tons) et les trois attributs corrélés au sexe du locuteur est calculée, avec le type de phonation et le sexe en prédicteurs fixes et le locuteur en variable aléatoire. Pour les trois attributs ayant une interaction significative avec le sexe du locuteur, les interactions avec la f_0 sont également significatives ($p < 0.001$), voir Table 3.

Pour vérifier les informations utilisées par les cinq autres attributs, un modèle mixte est utilisé pour calculer l'interaction entre les attributs et $h_1 - h_2$, avec le sexe et le type de phonation des extraits

2. Une interaction significative entre deux variables indique l'influence d'une variable sur l'effet de la deuxième.

en prédicteurs fixes, et le locuteur en variable aléatoire. L'interaction entre BA5 et $h_1 - h_2$ est significative ($p < 0.001$), ainsi que pour BA51 et BA85 (voir Table 3). Cependant, pour BA3 et BA27 il n'y a pas d'interaction significative ($p > 0.05$).

Enfin, un modèle mixte similaire est utilisé pour les attributs et le Ratio Harmoniques/Bruit en variables. L'interaction entre les attributs et la mesure est significative ($p < 0.001$) pour BA27, BA51 et BA85, ainsi que pour BA5 ($p < 0.005$), voir Table 3.

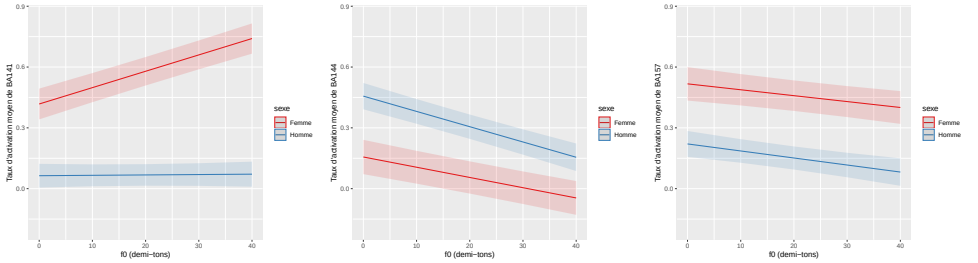
Les interactions calculées précédemment sont comparées avec celles obtenues entre les moyennes combinées des activations de l'entièreté des attributs (206) et les différentes mesures (sexe, f_0, $h_1 - h_2$, HNR), en conservant les paramètres des modèles mixtes. Cette comparaison permet de vérifier si les sept attributs analysés ci-avant contribuent significativement à l'encodage du sexe ou du type de phonation par le système. Les interactions utilisant les 206 attributs ne sont pas significatives pour le sexe, la f_0 et $h_1 - h_2$ ($p > 0.05$). L'interaction avec le HNR est significative ($p = 0.020$).

La vérification des informations retenues par les attributs a apporté les informations suivantes : le type de phonation est encodé par quatre attributs, pour lesquels des interactions significatives ont été relevées pour les mesures $h_1 - h_2$ et le Ratio Harmoniques/Bruit. Trois autres attributs encodent le sexe du locuteur, présentant une interaction significative avec la f_0.
BA3 a un comportement atypique : bien qu'il ait été sélectionné pour sa différence de taux d'activation moyens entre les groupes craqué et soufflé, il ne présente pas d'interaction significative avec la fréquence fondamentale ou le type de phonation. La présence d'autres biais tels que l'âge, des pathologies de la voix ou l'accent régional peuvent expliquer son comportement.

3.2 Interactions entre attributs

D'après les résultats précédents, le type de phonation et le sexe font partie de l'information retenue par sept attributs analysés. Certains attributs ont leurs taux d'activation moyens en interaction significative avec une mesure acoustique commune, comme BA141, 144 et 157 avec f_0. Cette sous-section détaille les différentes conditions d'activation des attributs en interaction avec une mesure acoustique commune.

La Figure 2 représente la régression logistique entre les trois attributs et la f_0 : on observe qu'ils n'encodent pas tous la même information concernant la f_0 . BA141 est activé pour les femmes à la voix aiguë, BA144 est activé pour les hommes à la voix grave, et BA157 est activé pour les femmes à la voix grave. Ces attributs permettent au système d'identifier les principaux contrastes de hauteur de voix selon le sexe : femme à la voix aiguë (BA141 activé), homme à la voix grave (BA144 activé), femme à la voix grave (BA157 activé) et homme à la voix aiguë (aucun des trois attributs activés).



(a) (b) (c)

FIGURE 2 – Modélisation de l'interaction significative entre BA141 (a), BA144 (b) et BA157 (c), et la f_0 en demi-tons et le sexe du locuteur dans chaque extrait, avec le type de phonation en prédicteur fixe et le locuteur en variable aléatoire.

L'écoute des extraits qui activent les quatre attributs liés au type de phonation permet d'obtenir des renseignements supplémentaires sur leurs conditions d'activation.
Par exemple, 80% des extraits activés pour BA85 contiennent des voyelles allongées qui peuvent provoquer le craquement, comme les disfluences verbales. La durée d'allongement de la voyelle requise pour l'activation de l'attribut est de minimum 500ms, c'est-à-dire un sixième de l'extrait. Un exemple avec l'hésitation « euh » est visible sur la Figure 3, où l'on constate l'irrégularité du signal et du spectrogramme lors de sa production. Les extraits activés pour BA5 ont en commun la présence de voix craquée pendant au moins 700ms, et ce peu importe les phonèmes présents.
BA51 est présent lors de la présence de souffle ou d'écho dans l'enregistrement, notamment lorsque des locuteurs parlent près du micro. Enfin, BA27 peut être activé lors de la production de voix dîte « peu efficace », pendant laquelle les locuteurs sont souvent à court de souffle.

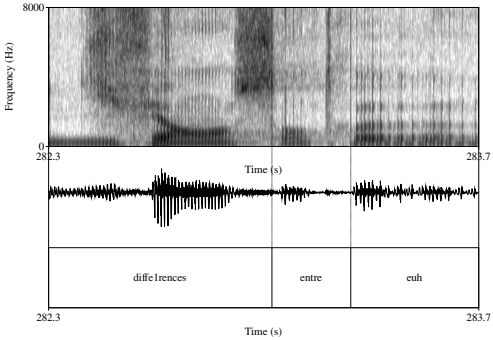


FIGURE 3 – Exemple de la disfluence craquée « euh » présente dans un extrait activant BA85 chez le locuteur LL054.

3.3 Encodage de la prototypicalité de sexe

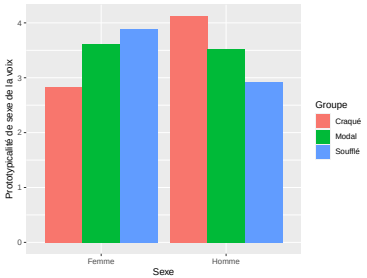
Le type de phonation produit par un locuteur est influencé par son sexe (voir Section 1). Il a été montré en Section 3 que les interactions avec le sexe ne sont pas significatives pour les attributs BA5, BA27, BA51 et BA85 ($p > 0.005$). On étudie dans cette sous-section l'hypothèse suivante : l'activation des attributs BA5, BA27, BA51 et BA85 est corrélée aux caractéristiques perceptuelles des voix prototypiques d'un sexe.
Une voix perceptuellement prototypique d'un sexe possède des caractéristiques associées stéréotypiquement à ce sexe. Cela est dû à des différences biologiques récurrentes entre les deux sexes, qui sont intériorisées par les auditeurs et sont utilisées lors de la perception d'une voix (Latinus & Belin, 2011). Par exemple, des voix prototypiques féminines peuvent comporter un type de phonation soufflé et des contours intonatifs importants (Avery & Liss, 1996), tandis que des voix prototypiques masculines peuvent comporter un type de phonation tendu et une f_0 basse (Baumann & Belin, 2010).
Si cette hypothèse est vérifiée, alors une voix prototypiquement féminine active plus fréquemment les attributs liés à la voix soufflée, et inversement pour une voix prototypiquement masculine.

Une annotation perceptuelle en prototypicalité de voix selon le sexe est effectuée pour les locuteurs des différents groupes de types de phonation. L'évaluation se fait sur chaque enregistrement des différents groupes, à l'aide une échelle de 1 à 5 avec 1 la note la plus basse (locuteur à la voix non prototypique) et 5 la note la plus haute (locuteur à la voix prototypique). L'accord inter-annotateur est calculé selon le Kappa de Cohen, dont le résultat est $\kappa = 0,60$. Ainsi, pour prendre en compte les avis de chaque annotateur, l'annotation finale est la moyenne des annotations précédentes [3]. La Figure 4a

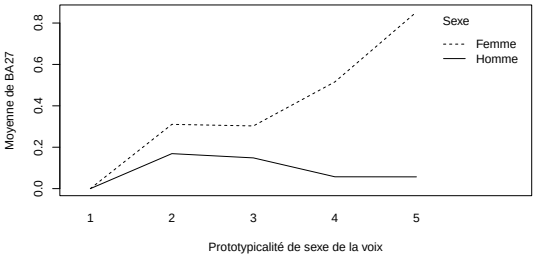
3. Les annotations ont été réalisées par les deux premiers auteurs, en se basant sur l'ampleur des contours intonatifs, l'intensité de la voix et la hauteur de la f_0.

montre la répartition des voix prototypiquement féminines et masculines d'après l'annotation finale, selon le groupe de locuteurs. Un calcul de l'interaction entre les scores de prototypicalité et la f_0 est effectué à l'aide d'un modèle mixte. L'interaction est significative ($p < 0,001$).

L'interaction entre les quatre attributs et l'annotation en prototypicalité de sexe est calculée avec un modèle mixte, dont le prédicteur fixe est le sexe et la variable aléatoire est le type de phonation. Le modèle montre que BA27 est le seul attribut possédant une interaction significative ($p < 0,001$). La Figure 4b représente l'interaction entre BA27 et la prototypicalité de sexe des voix avec le sexe comme prédicteur fixe. Le taux d'activation moyen de BA27 croît selon le score de prototypicalité féminine (80% d'activations pour une note de 5, aucune activation pour une note de 1). BA27 encode donc le principe de prototypicalité de la voix féminine. Les trois autres attributs ne retiennent pas cette information ($p > 0,05$).



(a) Diagramme en barres de l'annotation des locuteurs selon la prototypicalité de leur voix en termes de sexe.



(b) Interaction entre la prototypicalité de sexe des voix des locuteurs et le taux d'activation moyen de l'attribut BA27, avec le sexe comme facteur.

FIGURE 4 – Résultats de l'analyse des attributs selon la prototypicalité de sexe des voix des locuteurs.

4 Conclusion

Les résultats de cette étude montrent que plusieurs attributs du BA-LR sont corrélés à des paramètres de qualité de voix, ici les types de phonation craqué et soufflé (Section 3). Le sexe est également une information discriminante pour trois des huit attributs étudiés, ainsi que la prototypicalité homme/femme des voix pour un d'entre eux (Sous-section 3.3).
Les attributs interagissent entre eux, et de nombreux paramètres sont à prendre en compte afin de comprendre leurs conditions d'activation. Ces résultats encouragent à procéder à l'annotation d'autres caractéristiques perceptibles dans les enregistrements, notamment au niveau de la qualité de voix, afin de caractériser d'autres attributs. Utiliser le BA-LR sur un autre corpus, multi-sessions, afin d'en comparer les résultats avec PTSVOX, est la prochaine piste d'étude.

La méthodologie suivie dans cet article a permis d'établir une interaction entre des attributs extraits à partir d'un système automatique et des paramètres de qualité de voix. Cette démarche confère une plus grande explicabilité au système, utile dans un cadre judiciaire.
La meilleure compréhension des attributs sous un angle perceptuel permet à la fois d'évaluer la proximité entre la perception humaine et le réseau de neurones derrière le BA-LR, mais aussi de documenter les paramètres de qualité de voix utilisés par cet outil : cela le rend utilisable dans des tâches de profilage du locuteur, la qualité de voix pouvant apporter des renseignements physiques et culturelles sur le locuteur étudié. Cela permettrait à terme l'extraction automatique d'un profil de locuteur à partir d'un simple enregistrement sonore.

Références

AVERY J. D. & LISS J. M. (1996). Acoustic characteristics of less-masculine-sounding male speech. *The Journal of the Acoustical Society of America*, **99**(6), 3738–3748. DOI : 10.1121/1.414970.

BAI Z. & ZHANG X.-L. (2021). Speaker recognition based on deep learning : An overview. *Neural Networks*, **140**, 65–99.

BARSTIES B. & DE BODT M. (2015). Assessment of voice quality : current state-of-the-art. *Auris Nasus Larynx*, **42**(3), 183–188. DOI : 10.1016/j.anl.2014.11.001.

BATES D., M
ÄCHLER M., BOLKER B. & WALKER S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**(1), 1–48. DOI : 10.18637/jss.v067.i01.

BAUMANN O. & BELIN P. (2010). Perceptual scaling of voice identity : common dimensions for different vowels and speakers. *Psychological Research PRPF*, **74**(1), 110–120. DOI : 10.1007/s00426-008-0185-z.

BEN AMOR I. & BONASTRE J.-F. (2022a). Ba-lr : Binary-attribute-based likelihood ratio estimation for forensic voice comparison. In *2022 International workshop on biometrics and forensics (IWBF)*, p. 1–6 : IEEE. DOI : 10.1109/IWBF55382.2022.9794542.

BEN AMOR I. & BONASTRE J.-F. (2022b). Ba-lr : une approche transparente de comparaison de voix en criminalistique. In *Actes de la 7e conférence conjointe Journées d'Études sur la Parole (JEP, 34e édition), Traitement Automatique des Langues Naturelles (TALN, 29e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 24e édition). Volume 1 : Journées d'Études sur la Parole*, p. 646–654. DOI : 10.21437/JEP.2022-68.

BEN AMOR I., BONASTRE J.-F., O'BRIEN B. & BOUSQUET P.-M. (2023). Describing the phonetics in the underlying speech attributes for deep and interpretable speaker recognition. In *Proceedings of Interspeech 2023*. HAL : hal-04155146.

BENOIST-LUCY A. & PILLOT-LOISEAU C. (2013). The influence of language and speech task upon creaky voice use among six young american women learning french. In *Proceedings of Interspeech 2013*, p. 2395–2399. HAL : hal-00862349.

BOERSMA P. (2001). Praat, a system for doing phonetics by computer. *Glot. Int.*, **5**(9), 341–345.

CAMPBELL J. P., SHEN W., CAMPBELL W. M., SCHWARTZ R., BONASTRE J.-F. & MATROUF D. (2009). Forensic speaker recognition. *IEEE Signal Processing Magazine*, **26**(2), 95–103.

CHANCLU A., GEORGETON L., FREDOUILLE C. & BONASTRE J.-F. (2020). Ptsvox : une base de données pour la comparaison de voix dans le cadre judiciaire. In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 1 : Journées d'Études sur la Parole*, p. 73–81. HAL : hal-02798519.

DAVIDSON L. (2019). Perceptual coherence of creaky voice qualities. In *Proceedings of the 19th International Congress of Phonetic Sciences. Canberra, Australia : Australasian Speech Science and Technology Association Inc*, p. 147–151.

EYBEN F., WÖLLMER M. & SCHULLER B. (2010). opensmile : the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, p. 1459–1462.

GORDON M. & LADEFOGED P. (2001). Phonation types : a cross-linguistic overview. *Journal of phonetics*, **29**(4), 383–406. DOI : 10.006/jpho.2001.0147.

GREER S. D. & WINTERS S. J. (2015). The perception of coolness : Differences in evaluating voice quality in male and female speakers. In *Proceedings of ICPhS 2015*.

HANSON H. & CHUANG E. (1999). Glottal characteristics of male speakers : Acoustic correlates and comparison with female data. *The Journal of the Acoustical Society of America*, **106**, 1064–77. DOI : 10.1121/1.427116.

JACQUELIN M., GARNIER M., GIRIN L., VINCENT R. & PERROTIN O. (2023). Exploring the multidimensional representation of individual speech acoustic parameters extracted by deep unsupervised models. In *12th ISCA Speech Synthesis Workshop (SSW2023)*, p. 240–241. HAL : hal-04274170.

KEATING P., ESPOSITO C., GARELLEK M., KHAN S. & KUANG J. (2010). Phonation contrasts across languages. In *Poster presented at the 12th Conference on Laboratory Phonology*.

KREIMAN J., VANLANCKER-SIDTIS D. & GERRATT B. R. (2003). Defining and measuring voice quality. In *ISCA Tutorial and Research Workshop on Voice Quality : Functions, Analysis and Synthesis*.

KUZNETSOVA A., BROCKHOFF P. B. & CHRISTENSEN R. H. B. (2017). lmerTest package : Tests in linear mixed effects models. *Journal of Statistical Software*, **82**(13), 1–26. DOI : 10.18637/jss.v082.i13.

KWON Y., HEO H.-S., LEE B.-J. & CHUNG J. S. (2021). The ins and outs of speaker recognition : lessons from voxsrc 2020. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 5809–5813 : IEEE.

LATINUS M. & BELIN P. (2011). Human voice perception. *Current Biology*, **21**(4), R143–R145.

LEE Y. & KREIMAN J. (2019). Within and between speaker variation in voices. In *International Congress of Phonetic Sciences*, p. 1460–1464.

LEE Y. & KREIMAN J. (2022). Linguistic versus biological factors governing acoustic voice variation. In *Proceedings of Interspeech 2022*, p. 640–643.

NAGRANI A., CHUNG J. S. & ZISSERMAN A. (2017). Voxceleb : a large-scale speaker identification dataset. In *Proceedings of Interspeech 2017*, p. 2616–2620.

R CORE TEAM (2023). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. R version 4.3.2 (2023-10-31).

REVELLE W. (2024). *psych : Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package version 2.4.1.

SARNI S., CUMANI S., SINISCALCHI S. M., BOTTINO A. *et al.* (2023). Description and analysis of the kpt system for nist language recognition evaluation 2022. In *Proceedings of 24th INTERSPEECH Conference*, p. 1–5 : ISCA.

SOVRANO F., SAPIENZA S., PALMIRANI M. & VITALI F. (2022). Metrics, explainability and the european ai act proposal. *J*, **5**(1), 126–138.