Informations segmentales pour la caractérisation phonétique du locuteur : variabilité inter- et intra-locuteurs

Cédric Gendrot, Emmanuel Ferragne, Thomas Pellegrini

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Informations segmentales pour la caractérisation phonétique du locuteur : variabilité inter- et intra-locuteurs

Cédric Gendrot[1] Emmanuel Ferragne[1] Thomas Pellegrini[2]
(1) LPP, UMR 7018 CNRS-Sorbonne Nouvelle, 19 rue des Bernardins, 75005 Paris, France
(2) IRIT, UMR 5505 CNRS-INP, UT1, UT3, UT2J, 118 Route de Narbonne, F-31062 Toulouse Cedex 9
cedric.gendrot@cnrs.fr, emmanuel.ferragne@u-paris.fr,
thomas.pellegrini@irit.fr

RÉSUMÉ

Nous avons effectué une classification automatique de 44 locuteurs à partir de réseaux de neurones convolutifs (CNN) sur la base de spectrogrammes à bandes larges calculés sur des séquences de 2 secondes extraites d'un corpus de parole spontanée (NCCFr). Après obtention d'un taux de classification moyen de 93,7 %, les différentes classes phonémiques composant chaque séquence ont été masquées afin de tester leur impact sur le modèle. Les résultats montrent que les voyelles orales influent avant toute autre classe sur le taux de classification, suivies ensuite par les occlusives orales. Ces résultats sont expliqués principalement par la représentation temporelle prédominante des voyelles orales. Une variabilité inter-locuteurs se manifeste par l'existence de locuteurs attracteurs qui attirent un grand nombre de faux positifs et qui ne sont pas sensibles au masquage effectué. Nous mettons en avant dans la discussion des réalisations acoustiques qui pourraient expliquer les spécificités de ces locuteurs.

ABSTRACT

An automatic classification task involving 44 speakers was performed using convolutional neural networks (CNN) on broadband spectrograms extracted from 2-second sequences of a spontaneous speech corpus (NCCFr). We obtained a mean classification rate of 93,7 % and carried out an occlusion experiment afterwards : different phonemic classes were hidden within each sequence so as to test their impact on classification rates. Results show that oral vowels influence the classification much more than the other classes. These results are mainly explained by the prevailing temporal representation of oral vowels. Substantial inter-speaker variability is observed and correlated with the presence of magnet speakers who 'attract' most false positive classifications from other speakers. In the discussion, we display acoustic measurements that may be relevant to explain these speakers' behaviour.

MOTS-CLÉS : caractérisation du locuteur, deep learning, segmental.

KEYWORDS: speaker characterization, deep learning, segmental.

1 Introduction

Il est fréquent de constater que des locuteurs de notre entourage présentent une production particulière de certains phonèmes : on pense par exemple à la réalisation de /ʃ-s/ et /ʒ-z/ en fricatives latérales

Volume 1 : Journées d'Études sur la Parole, pages 262–270.
HAL : hal-02798547.

[ɬ] et [ɮ]. Il est donc légitime de se poser la question de l'utilité potentielle de la prononciation spécifique de certains phonèmes dans la caractérisation du locuteur sur la base d'indices acoustiques. La caractérisation du locuteur peut avoir plusieurs objectifs : elle peut se faire dans une optique criminalistique pour la comparaison de voix (Morrison & Thompson, 2017; Ajili *et al.*, 2018). Elle peut également servir la recherche fondamentale en phonétique, qui vise à mieux comprendre la variabilité que l'on observe pour la description de phénomènes tels que la coarticulation, la réalisation des groupes prosodiques, etc. (Keating *et al.*, 2017, par exemple).

Notre travail se concentre sur l'influence du segmental pour la caractérisation du locuteur : certains phonèmes ou certaines classes de phonèmes sont-ils plus pertinents que d'autres ? Et est-ce que ces éventuels phonèmes discriminants sont partagés par l'ensemble des locuteurs ? Si nous nous focalisons sur le segmental, il ne faut pas oublier non plus que les informations prosodiques (valeurs de f0, d'intensité, etc.) peuvent également permettre de caractériser certains locuteurs ; citons notamment Dellwo *et al.* (2015) et Keating *et al.* (2017). Plusieurs études ont démontré que les caractéristiques segmentales des locuteurs peuvent être pertinentes pour la classification de ces derniers (Kahn, 2014; Amino *et al.*, 2006). Dans ce cadre, il est fréquent de trouver dans la littérature que les voyelles nasales ont un poids plus important dû à l'ajout de la cavité nasale dans la production (Shriberg & Stolcke, 2008). Mais Ajili *et al.* (2018), dans une étude plus récente, ont montré que les voyelles orales étaient les plus utiles dans une tâche de reconnaissance du locuteur pour une approche criminalistique, ces résultats contradictoires pouvant être interprétés comme une forte variabilité inhérente à la parole. Les auteurs notent également une forte variabilité inter-locuteurs, qu'il est nécessaire d'approfondir selon eux, et nous tentons ici d'analyser cette variabilité.

Dans une expérience s'appuyant sur l'apprentissage automatique comme pour la présente étude, il existe différents procédés pour déterminer quelles sont les informations mises à profit par le modèle afin de classifier les locuteurs (Ferragne *et al.*, 2019). Sur le même principe que Ajili *et al.* (2018), nous avons réalisé une expérience de classification avec masquage (occlusion), où une partie de l'information contenue dans le signal acoustique est cachée afin de comparer la classification avant et après masquage. Cette étude se rapproche également des travaux effectuées par Besacier & Bonastre (1998) dans lesquels des blocs temporels de signal sont sélectionnés pour améliorer les taux d'identification du locuteur. Cependant, contrairement aux études citées ci-dessus, nous utilisons des spectrogrammes à bandes larges en entrée, car notre objectif de phonéticien sera à terme de retracer la correspondance acoustique - articulatoire. Après une présentation de la méthode employée, et des résultats de l'influence du masquage, nous nous concentrerons sur la variabilité observée, en insistant notamment sur quelques locuteurs caractéristiques.

2 Méthode

Nous avons dans un premier temps mené une tâche de classification de segments de paroles en 44 classes de locuteurs en utilisant un réseau de neurones convolutif (CNN) qui prenait en entrée des spectrogrammes. Nous avons ensuite procédé à une occlusion partielle des spectrogrammes, d'abord par phonème (e.g. tous les phonèmes étaient remplacés tour à tour par un masque noir), puis par classe de phonèmes (e.g. toutes les voyelles orales étaient masquées simultanément) afin d'observer la dégradation du taux de classification conséquente au masquage de l'information.

2.1 Corpus et prétraitement

Des séquences de 2 secondes ont été utilisées pour la classification ; elles ont été extraites de 44 locuteurs du corpus NCCFr (Torreira *et al.*, 2010). Ce corpus est constitué de conversations spontanées d'environ une heure entre deux (voire trois) amis ; il a été annoté par des transcripteurs professionnels, puis aligné phonémiquement par le système du LIMSI. Le corpus est composé d'enregistrements de 23 hommes et 21 femmes ; les séquences contenant un minimum de 18 et un maximum de 43 phonèmes ont été retenues, sans autre type de contrainte. Ces séquences, échantillonnées à 16 kHz, ont été converties en spectrogrammes à bandes larges avec des trames de 5 ms, un chevauchement de 90 % et une taille de FFT de 512 points. La dynamique a été fixée à 70 dB et quantifiée sur 8 bits de niveaux de gris dans les images finales. La résolution en fréquence, 257 points pour 8 kHz, a été laissée telle quelle dans les images fournies en entrée du modèle. En revanche, nous avons réduit la résolution temporelle des spectrogrammes (de 3 991 à 400 points) pour des raisons évidentes de mémoire. Nous disposons donc de 15 400 images de spectrogrammes : 350 pour chacun des 44 locuteurs.

2.2 Modèle initial

Un réseau de neurones profond de type ResNet-18 (He *et al.*, 2016) a été utilisé pour la classification automatique des spectrogrammes en 44 classes de locuteurs. L'ensemble d'apprentissage contenait 70 % des données ; 10 % servaient pour la validation et 20 % pour l'évaluation. Nous avons utilisé l'optimiseur Adam (Kingma & Ba, 2014) avec une valeur initiale du taux d'apprentissage de $1e$-4, divisé par deux après huit itérations complètes sur les données d'apprentissage. Un maximum de 10 itérations en tout a été effectué avec des mini-lots (*mini-batchs*) de 32 exemples, ce qui fait que le modèle a convergé en 28 minutes sur une carte GPU NVIDIA GTX 1080.

2.3 Protocole d'occlusion

Dans un premier temps, afin de quantifier la pertinence des phonèmes pour la caractérisation du locuteur, nous avons effectué le masquage phonème par phonème tout au long de chaque séquence de 2 secondes. L'objectif était d'identifier un ou plusieurs phonèmes susceptibles de faire basculer l'identification du locuteur (i.e. engendrer une classification erronée). Ce type de changement dans la classification n'a été obtenu que pour les séquences dont la probabilité de classification dans la classe correcte avant masquage était faible, inférieure à 50 %. Les taux d'identification étant supérieurs à 90 % avec des probabilités d'identification très élevées, le masquage d'un seul phonème pouvait suffire que très rarement à engendrer une erreur de classification. Au total, à l'issue du masquage par phonème, seules 2.5 % des séquences présentaient un changement de classe de locuteur, ce qui est insuffisant pour effectuer une analyse quantitative. Notons tout de même que les phonèmes /s/ et /ʃ/ ont été identifiés pour deux locuteurs comme particulièrement pertinents, notamment parce que -après écoute des séquences concernées- ceux-ci réalisés avec un chuintement. Pour quelques autres locuteurs, le masquage d'une hésitation longue (>300 ms) pouvait faire basculer la classification en locuteurs sur la séquence testée. Mais ces cas étaient par trop rares et nous avons donc procédé dans un second temps à une occlusion par classes phonémiques, où tous les phones correspondant à une classe phonémique ont été masqués simultanément. Cette occlusion représente ainsi une durée plus importante (293 ms en moyenne, toutes classes confondues) et nous espérions voir augmenter le

nombre de changements d'identifications du locuteur après occlusion.

Les classes phonémiques ont été regroupées de la façon suivante : voyelles orales (ORAVO), consonnes/voyelles nasales (NASAL), occlusives (OCCLU), fricatives (FRICA) et sonantes (SONOR). Les consonnes et voyelles nasales ont été regroupées puisque la nasalité est estimée comme un critère prépondérant mais également afin d'obtenir des groupes plus équilibrés en termes de fréquences pour chaque catégorie.

Dans l'étude de Ajili *et al.* (2018), un masque est également appliqué de façon aléatoire sur un échantillon de signal de durée équivalente à celle correspondant aux phonèmes de la classe phonémique masquée. Nous avons choisi de ne pas procéder de cette façon mais plutôt de prendre en compte la variation de durée a posteriori dans nos analyses statistiques. Les résultats que nous présenterons dans les sections suivantes seront basés sur les séquences dont la classification du locuteur passe de correcte à incorrecte (3 821 sur 15 029).

3 Résultats

Les résultats présentés ci-dessous ont été calculés sur les séquences d'évaluation exclusivement. Le taux moyen des bonnes classifications avant masquage est de 93.7 % (14100/15029). Après masquage, il passe à 68.3 % (10279 / 15029) : pour les occlusives 78.0 % (2236/2867), les fricatives 83.6 % (2347/2805), les nasales 83.9 % (2291/2729), les voyelles orales 36.2 % (1046/2887) et les sonantes 83.9 % (2359/2812). Ces confusions vont vers un autre locuteur du même sexe dans 96.6 % des cas pour les femmes et seulement 61.5 % des cas pour les hommes.

TABLE 1 – Résumé des résultats de classification (taux de bonne classification en %, score de probabilité de classification dans la classe correcte en %, taille du masque en ms) pour les séquences sans masquage puis en fonction chaque classe phonémique masquée.

	séquence sans masque	ORAVO	OCCLU	NASAL	FRICA	SONOR
classif. correct.	93.8	36.2	78.0	83.9	83.6	83.9
probabilité	84.9	27.7	60.8	65.4	66.4	67.9
durée masque	0	490	304	223	254	190

En analysant les scores d'identification après masquage, nous déduisons que les voyelles orales ont un effet important sur la classification, loin devant les occlusives puis les autres catégories. Notons également que le niveau de probabilité indiqué par le réseau lors de la classification, qui donne un indice de la certitude du résultat, passe de 84.9 % avant masquage à 57 % en moyenne (et 28 % pour les voyelles orales). Pour la suite de cette étude, nous avons cependant décidé de nous concentrer sur les cas de changements de classification (de correcte à erronée). Ce résultat est surprenant au premier abord puisque la classes des nasales n'est que peu pertinente dans la caractérisation du locuteur ici. Cela pourrait être en partie expliqué par le fait que nous avons combiné voyelles et consonnes nasales dans la même classe (pour des raisons d'homogénéité des classes). En effet, les voyelles nasales sont plus particulièrement mentionnées comme pertinentes dans la littérature, et les consonnes nasales dans une moindre mesure seulement. Nous avons comptabilisé le nombre de voyelles nasales présentes dans chaque séquence afin d'estimer si un nombre plus important de

voyelles nasales pouvait être corrélé à des changements de classifications plus fréquents. Les résultats montrent que lorsque le masquage de la classe des nasales implique 2 voyelles nasales et plus (1.78 en moyenne dans l'ensemble des séquences), le taux de classifications correctes chute à 79 %, contre 90 % quand il y a moins de 2 voyelles nasales dans les séquences. Ce résultat corrobore l'idée que les voyelles nasales sont plus pertinentes que les consonnes nasales pour la caractérisation du locuteur.

Il est nécessaire de pondérer ces résultats en considérant les fréquences d'apparition inhérentes à chaque classe, la classe des voyelles orales étant par exemple beaucoup plus fréquente que les autres classes. La durée masquée moyenne – en effectuant le calcul pour toutes les occurrences – est de 300 ms, celle des voyelles orales monte à 490 ms. Les occlusives ont une durée masquée (304 ms) supérieure aux nasales (223 ms), puis pour les plus courtes les fricatives (253 ms) et les sonantes (190 ms).

Pour ce faire, nous avons utilisé un modèle mixte linéaire généralisé (GLMM) afin d'évaluer la probabilité d'un changement de classe en fonction de la classe phonémique masquée et de sa durée. Les variables aléatoires utilisées dans ce modèle sont les locuteurs et les différentes séquences testées. Le modèle indique un effet significatif avec une valeur F de 391.4 ($p=0.0013$) pour la classe phonémique et 1284.6 ($p<0.0001$) pour la durée, l'interaction entre la classe et la durée est quant à elle non significative avec une valeur F de 3.9 ($p=0.11$). Ces résultats indiquent que l'importance de la classe phonémique sur la caractérisation du locuteur est bien pertinente, parallèlement à l'influence de la durée occultée au sein de la séquence.

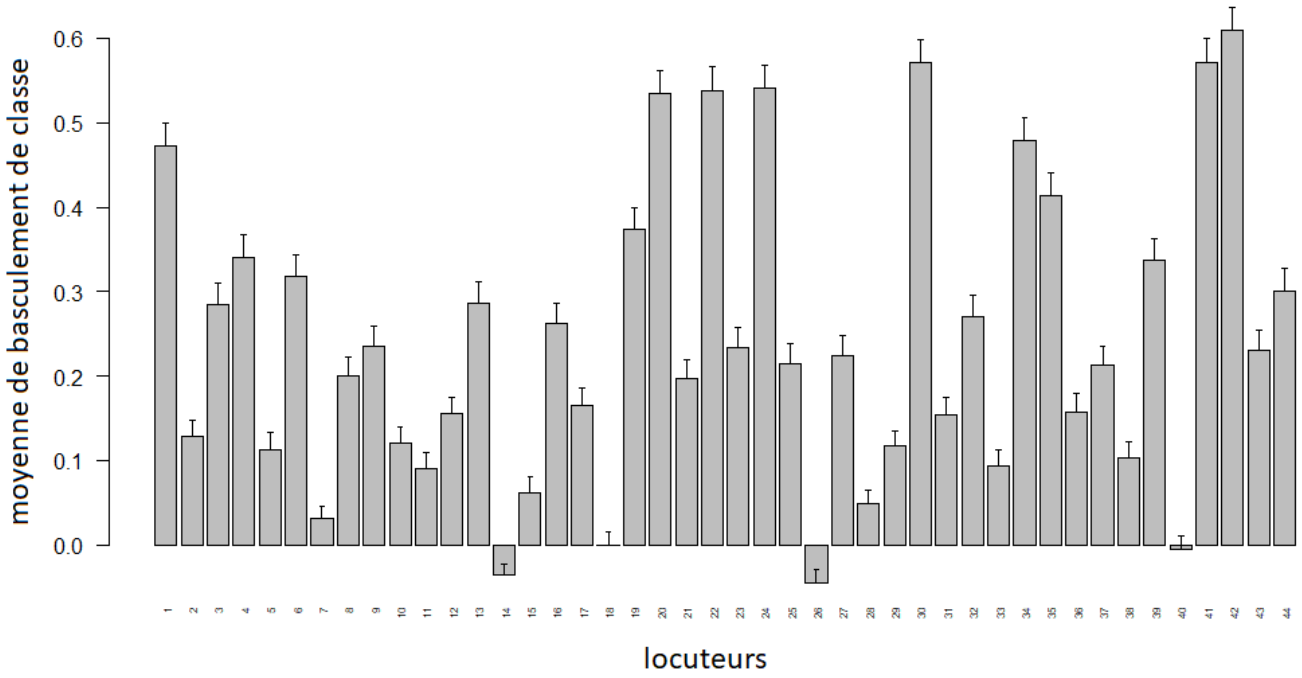

FIGURE 1 – Sensibilité des locuteurs au masquage : en abscisse les identifiants des locuteurs et en ordonnée la moyenne des changements de classification

4 Variation inter-locuteurs

Nous observons que pour environ 20 % des locuteurs, le masquage n'a que peu voire pas d'effet sur le résultat de la classification. Pour quantifier ce phénomène, nous avons calculé la moyenne des occurrences où un changement de classe a lieu (0 si l'identification du locuteur est passée de correcte à incorrecte, 1 si l'identification est restée correcte, et -1 si l'identification passe d'incorrecte à correcte). Comme illustré sur la Figure 1, les locuteurs 14, 18, 26 et 40 ont une moyenne nulle à négative, les locuteurs 7, 15 et 28 ont une moyenne inférieure à 0.05, tandis que les locuteurs 11 et 33 ont une moyenne située entre 0.05 et 0.1. Ces 9 locuteurs ont un score moyen de bonne classification à 92.0 %, contre 94.1 % pour les 35 autres locuteurs, et 93,7 % pour la moyenne, avec des taux de probabilité de 84.9 % identiques à ceux des 35 autres locuteurs. Ces résultats montrent que les locuteurs insensibles au masquage ne sont pas des locuteurs plus difficiles (ou faciles) à classifier, et que la source de cette insensibilité au masquage doit être cherchée ailleurs. En considérant les différentes classes phonémiques dans ce résultat sur la Figure 2, on observe qu'après masquage, le taux moyen des bonnes classifications descend seulement à 88 % lorsque la classe des voyelles orales est masquée et ne bouge pas pour les autres classes phonémiques, ce qui conforte l'insensibilité de ces locuteurs au masquage pour toutes les classes phonémiques, les différences observées entre celles-ci étant considérablement réduites (Figure 3).

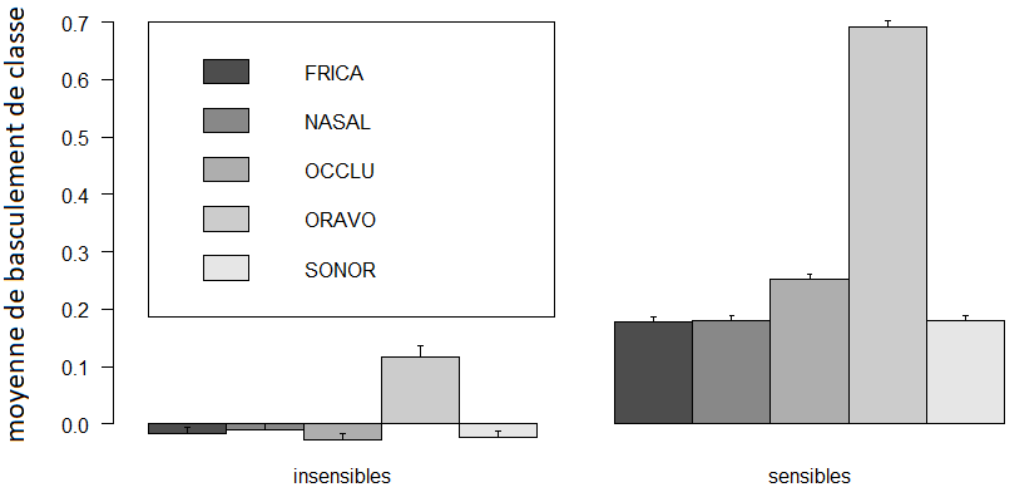

FIGURE 2 – Sensibilité au masquage en fonction de la classe phonémique pour les 9 locuteurs peu sensibles au masquage comparés aux 35 autres locuteurs

Nous tentons ici d'approfondir la compréhension de cette variabilité inter-locuteurs en observant la matrice de confusion. Les lignes de la matrice nous renseignent sur les vrais positifs et faux positifs obtenus pour les 44 locuteurs, et les 9 locuteurs mentionnés plus haut reçoivent un nombre important de faux positifs de la part des autres locuteurs. La non sensibilité au masquage d'un locuteur serait donc liée au nombre de faux positifs reçus, et nous donnerons à ces locuteurs le terme de locuteurs attracteurs.

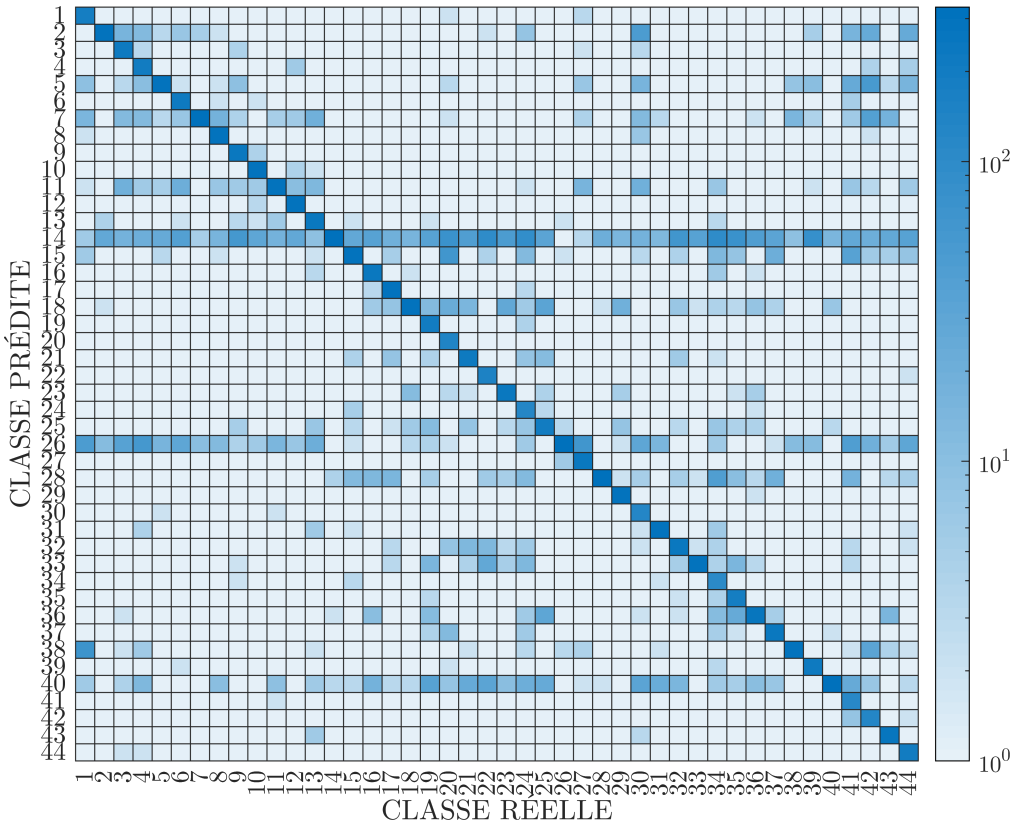

FIGURE 3 – Matrice de confusion de la classification des locuteurs après masquage

Afin de valider ce résultat, nous avons calculé une corrélation de Spearman entre le nombre de faux positifs que chaque locuteur a reçu et sa sensibilité au masquage (calculée comme la moyenne du nombre de changements de classes de correct à incorrect). Le coefficient obtenu est de -0.67 ($t = -5.8966; p = 5.594e - 07$) ce qui montre un lien très net entre ces deux variables. Ce taux est plus élevé si l'on considère uniquement la classe phonémique des occlusives (-0.8), des nasales (-0.8), des fricatives (-0.77) et des sonantes (-0.79), mais il est moins élevé que la moyenne pour les voyelles orales (-0.53). Ces résultats confirment un statut particulier de la classe des voyelles orales, qui lorsqu'on les masque, modifie considérablement la classification des locuteurs, mais les erreurs de classification sont moins focalisées sur certains locuteurs que pour les autres classes. Enfin, si l'on effectue une corrélation entre la sensibilité au masquage de chaque locuteur et le nombre de faux positifs qu'il a reçus avant la procédure de masquage, celle-ci tombe à 0.088 ($t = -0.57278; p = 0.5698$). Cette dernière corrélation montre que le masquage permet de mettre en avant ces locuteurs attracteurs qui n'apparaissent pas lorsque le spectrogramme entier est utilisé (avant la procédure de masquage). Les locuteurs qui au contraire sont très sensibles à l'occlusion ne reçoivent qu'un nombre très limité de faux positifs quelle que soit la classe phonémique masquée.

5 Discussion et conclusion

Notre étude a montré que lorsque le masquage est effectué phonème par phonème, certains segments très spécifiques s'avèrent être pertinents pour la caractérisation du locuteur.Il ne s'agit cependant en général que de phonèmes très spécifiques avec des prononciations atypiques tels que /s/ et /ʃ/ ou des

hésitations.

Lorsqu'on procède au masquage par classes de phonèmes, les résultats montrent que les voyelles orales, notamment du fait de leur durée plus importante, jouent un rôle dans la bonne classification des locuteurs puisque leur absence détériore considérablement les résultats. Mais environ 20 % des locuteurs ne sont pas sensibles au masquage, ces locuteurs attirant à eux les prédictions dont le score de probabilité est plus faible. Ces locuteurs que nous avons qualifiés d'attracteurs pourraient être considérés comme les agneaux (lambs) selon la terminologie de Doddington et al. George Doddington & Reynolds (1998) car ces locuteurs pourraient apparaitre comme faciles à imiter. Afin de comprendre pourquoi ces locuteurs recueillent un nombre important de faux positifs, nous avons effectué des mesures acoustiques sur les différentes séquences testées de ces locuteurs et avons pu constater qu'ils étaient caractérisés par une variation acoustique plus importante que les autres locuteurs, notamment pour leurs valeurs de f0 et d'intensité. Nous avons également pu faire ressortir des locuteurs qui se distinguent par leur caractère moyen sur l'ensemble des mesures acoustiques, plutôt qu'extrêmes sur une seule. Une classification de ces 44 mêmes locuteurs a été réalisée sur la base d'indices prosodiques seuls par Chignoli *et al.* (2020) (soumis à cette conférence) et montre que les informations de f0 et d'intensité peuvent être complémentaires au spectrogramme dans près d'un tiers des classifications des locuteurs.

Il est à noter que lorsqu'une classe phonémique permet de faire basculer la classification du locuteur de correcte à erronée, il est très fréquent que les autres classes phonémiques testées fassent également basculer la classification (25 % de cas où une seule classe phonémique est impliquée dans un changement de catégorie pour une séquence, 24 % de cas où il y a 2 classes, 51 % de cas où il y a entre 3 et 5 classes) Ce résultat indique qu'au delà de la pertinence de la classe phonémique pour la classification du locuteur, c'est la séquence dans son ensemble qui joue un rôle dans le résultat de la classification. Nous avons remarqué après écoute des séquences mal identifiées que celles-ci contenaient des rires, chuchotements, éclats de voix, etc., ce qui rend plus complexe la classification par le réseau dans une expérience de masquage qui a pour effet de faire baisser la probabilité d'appartenance à la classe (certitude du résultat), et donc d'atteindre plus facilement un seuil critique d'identification.

Pour la suite de cette étude, nous envisageons de prendre en compte le vecteur de probabilité d'appartenance à chaque classe afin d'obtenir une analyse plus fine de l'évaluation du masquage en calculant par exemple un classement entre locuteurs. Il est également envisagé de travailler sur la base d'une tâche de discrimination (et non plus de classification) dans un ensemble ouvert de locuteurs afin de pouvoir généraliser les conclusions proposées dans ce travail.

Références

AJILI M., BONASTRE J.-F., BEN KHEDER W., ROSSATO S. & KAHN J. (2018). Comparaison des voix dans le cadre judiciaire : influence du contenu phonétique. In *Proc. XXXIIe Journées d'Études sur la Parole*, p. 28–36. DOI : 10.21437/JEP.2018-4.

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

BESACIER L. & BONASTRE J.-F. (1998). Time and frequency pruning for speaker identification. In *Proc. on Speaker Recognition and its Commercial and Forensic Applications (RLA2C)*.

CHIGNOLI G., GENDROT C. & FERRAGNE E. (2020). Caractérisation du locuteur par CNN à l'aide des contours d'intensité et d'intonation : comparaison avec le spectrogramme. In *soumis aux Journées d'Etude de la Parole 2020*.

DELLWO V., LEEMANN A. & KOLLY M.-J. (2015). Rhythmic variability between speakers : Articulatory, prosodic, and linguistic factors. *The Journal of the Acoustical Society of America*, **137**(3), 1513–1528.

FERRAGNE E., GENDROT C. & PELLEGRINI T. (2019). Towards phonetic interpretability in deep learning applied to voice comparison. In *Proc. ICPhS*.

GEORGE DODDINGTON, WALTER LIGGETT A. M. M. P. & REYNOLDS D. (1998). Sheep, goats, lambs and wolves : a statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. In *NIST 1998 Speaker Recognition Evaluation*.

HE K., ZHANG X., REN S. & SUN J. (2016). Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 770–778, Las Vegas, NV, USA : IEEE. DOI : 10.1109/CVPR.2016.90.

KAHN J. (2014). *Parole de locuteur : performance et confiance en identification biométrique vocale*. Thèse de doctorat, Avignon.

KEATING P., KREIMAN J. & VASSELINOVA N. (2017). Acoustic similarities among voices. part 2 : Male speakers. *The Journal of Acoustic Society of America*, **142**.

KINGMA D. P. & BA J. (2014). Adam : A method for stochastic optimization. *arXiv preprint arXiv :1412.6980*.

MORRISON G. S. & THOMPSON, C. W. (2017). Assessing the admissibility of a new generation of forensic voice comparison testimony. *Columbia Science and Technology Law Review*, **18**, 326–434.

SHRIBERG E. & STOLCKE A. (2008). The case for automatic Higher-Level features in forensic speaker recognition. In *Proc. International Conference on Speech Communication and Technology (Interspeech)*, p. 1509–1512.

TORREIRA F., ADDA-DECKER M. & ERNESTUS M. (2010). The Nijmegen Corpus of Casual French. *Speech Communication*, **52**(3), 201–212. DOI : 10.1016/j.specom.2009.10.004.