Analyse phonétique de la variation inter-locuteurs au moyen de réseaux de neurones convolutifs : voyelles seules et séquences courtes de parole

Cédric Gendrot, Emmanuel Ferragne, Anaïs Chanclu

Analyse phonétique de la variation inter-locuteurs au moyen de réseaux de neurones convolutifs : voyelles seules et séquences courtes de parole

Cédric Gendrot[1] Emmanuel Ferragne[2] Anaïs Chanclu[3]
(1) Laboratoire de Phonétique et Phonologie, UMR 7018, Université Sorbonne Nouvelle 19, rue des Bernardins 75005 Paris
(2) CLILLAC-ARP UR 3967, Université de Paris. 8, place Paul Ricœur 75013 Paris
(3) Laboratoire Informatique d'Avignon, EA 4128, 339 chemin des Meinajaries, 84911 Avignon
cedric.gendrot@sorbonne-nouvelle.fr, emmanuel.ferragne@u-paris.fr, anais.chanclu@univ-avignon.fr

RÉSUMÉ

Des réseaux de neurones convolutifs ont été entraînés sur des spectrogrammes de voyelles /ɑ̃/ et de séquences aléatoires de 2 secondes extraites de 44 locuteurs du corpus NCCFr afin d'obtenir une classification de ces derniers. Ces deux modèles présentent une répartition équivalente des locuteurs dans l'espace acoustique, ce qui suggère que la classification a été faite sur des critères indépendants des phonèmes précis extraits. De multiples mesures phonétiques ont été effectuées afin de tester leur corrélation avec les représentations obtenues : la f0 apparait comme le paramètre le plus pertinent, suivie par plusieurs paramètres liés à la qualité de la voix. Des zones d'activation (Grad-CAM : Gradient-weighted Class Activation Mapping) ont été calculées a posteriori afin de montrer les zones spectrales et temporelles utilisées par le réseau. Une analyse quantitative de ces cartes d'activation a donné lieu à des représentations des locuteurs qui ne sont pas corrélées aux mesures phonétiques.

ABSTRACT

Phonetic analysis of individual variation with convolutional neural networks: single vowels and short speech chunks.

Convolutional neural networks were trained on /ɑ̃/ vowels and 2-second random sequences extracted from 44 speakers of the NCCFr corpus in order to obtain a classification of the latter. Both models show a similar organization of the speakers in the acoustic space, which suggests that the classification was learned on criteria phoneme-independent criteria. Multiple phonetic parameters were extracted in order to correlate them with the representations from the classifiers : f0 appears to be the most relevant parameter, followed by several parameters related to voice quality. Salient regions (Grad-CAM : Gradient-weighted Class Activation Mapping) were computed to highlight the spectral and temporal features used by the models. They allowed to classify around 50% of the speakers without being correlated to the phonetic measures, which suggests their independent character.

MOTS-CLÉS : Réseaux de Neurones Convolutifs (CNN), paramètres phonétiques, qualité de voix, classification du locuteur.

KEYWORDS: Convolutional Neural Networks (CNN), phonetic parameters, voice quality, speaker classification.

1 Introduction

Ce travail s'inscrit dans le cadre du projet ANR Voxcrim dont le but est d'améliorer la « comparaison de voix » pour la sécurité nationale et l'expertise judiciaire. Les méthodes employées peuvent être automatiques ou phonétiques, et cette étude emprunte aux deux approches dans le but de renforcer l'interprétabilité phonétique des caractéristiques propres aux locuteurs.

1.1 État de l'art de l'analyse phonétique

Dans les sciences phonétiques, l'utilisation récente de grands corpus de parole non préparée (plusieurs dizaines d'heures), alignés au niveau du phonème, a contribué à des avancées significatives dans la compréhension de la variation. Par exemple, les effets de fréquence des distributions de phonèmes et de mots ont permis d'expliquer leur réalisation acoustique d'une manière qui n'était pas possible avec de petits corpus contrôlés (Bybee, 2001; Gahl *et al.*, 2012). Ces corpus sont pour la plupart analysés avec des outils automatiques permettant une extraction de paramètres phonétiques tels que les formants, la f0, la durée, etc. Ces paramètres et leurs relations avec l'articulation ont été modélisés il y a longtemps par des auteurs tels que Delattre (Delattre, 1951), Fant (Fant, 1960), Stevens (Stevens, 1968) après l'invention des spectrogrammes aux Bell Labs au début des années 1940. Bien sûr, d'autres mesures acoustiques ont été explorées depuis, telles que le rapport harmonique sur bruit (Cardoso *et al.*, 2019), la proéminence du pic cepstral (Lee *et al.*, 2019), le centre de gravité spectral (Jongman *et al.*, 2000), les mesures de pente spectrale (Titze & Palaparthi, 2020), mais toutes ces mesures découlent des travaux antérieurs mentionnés ci-dessus.

L'avènement de l'apprentissage profond permet désormais aux phonéticiens d'utiliser des outils puissants pour réaliser la tâche d'analyse des spectrogrammes de manière automatique sans a priori théorique susceptible de biaiser cette analyse. Un avantage bien connu des réseaux de neurones profonds par rapport aux algorithmes d'apprentissage automatique plus conventionnels tient au fait qu'ils peuvent extraire des caractéristiques des données brutes sans qu'un expert humain ait besoin de les fournir explicitement au modèle (Goodfellow *et al.*, 2016). Reste la question de l'interprétabilité des résultats qui intéresse les phonéticiens : quelles sont les informations qui ont permis d'aboutir à une classification ou une modélisation ? Peuvent-elles être interprétées d'un point de vue articulatoire et/ou perceptif ? Afin de pallier le manque de transparence de nombreuses méthodes d'apprentissage automatique « traditionnelles » ainsi que la composante intrinsèquement subjective de la lecture de spectogramme, l'application de réseaux de neurones convolutifs à l'analyse de spectrogrammes comme cas particulier de reconnaissance d'images nous permet d'utiliser des techniques de visualisation montrant comment les modèles parviennent à leurs décisions (Selvaraju *et al.*, 2017). Comme les modèles nécessitent un grand nombre d'occurrences pour l'entraînement, la disponibilité de très grands corpus de parole permet aujourd'hui de répondre à ce critère. Ce travail se veut dans un premier temps exploratoire afin d'expliquer notre démarche d'utilisation de CNN pour la caractérisation de la variation inter-individuelle, et de montrer des premiers résultats encourageants dans ce cadre.

2 Méthodes d'analyse

Dans des travaux précédents (Ferragne *et al.*, 2019; Gendrot *et al.*, 2019) utilisant une sous-partie du corpus ESTER, nous avons montré que les réseaux de neurones convolutifs (CNN) surpassent significativement les paramètres phonétiques classiques lors de la discrimination acoustique entre 45 locuteurs. Ce type de tâches nécessite de démêler la quantité de variation au sein du locuteur pour la variation linguistique et la quantité d'informations qui sont caractéristiques du locuteur (c'est-à-dire qui le distinguent d'un autre locuteur). Les corpus de parole non préparés sont des outils utiles pour ces tâches dans le sens où ils apportent une variabilité non contrôlée. Dans cette présentation, nous étudions plus avant la visualisation des régions spectrales que les CNN utilisent pour prendre leurs décisions.

Deux expériences ont été mises en place en utilisant des spectrogrammes, l'une avec la voyelle isolée /ɑ̃/ et l'autre avec des extraits de discours de 2 secondes, toutes deux issues des mêmes 44 locuteurs du corpus NCCFr. Dans les deux expériences, environ 350 stimuli pour chacun des 44 locuteurs ont été utilisés dans le but de discriminer les locuteurs en utilisant des CNN. L'objectif est de confronter pour un même groupe de locuteurs des jeux de données différents afin d'en comparer les représentations construites par le CNN. Des cartes d'activation (Grad-CAM) ont également été réalisées pour la visualisation des régions spectrales et temporelles pertinentes. En parallèle, une série de mesures acoustiques a été effectuée sur ces données afin d'analyser leur corrélation avec les résultats du CNN, l'interprétabilité des résultats étant ici notre objectif.

2.1 Données et structure du CNN

Nous avons utilisé le corpus NCCFr, qui regroupe 45 locuteurs (Torreira *et al.*, 2010) engagées dans des conversations spontanées d'environ 1h en binômes. Un des locuteurs ayant une transcription défectueuse, celui-ci a été retiré, et restaient donc 20 femmes et 24 hommes. Pour les voyelles /ɑ̃/, 350 occurrences ont été extraites aléatoirement par locuteur (350 représentait le nombre d'occurrences le plus faible trouvé chez un locuteur). La seule restriction pour prendre en compte un item était que celui-ci soit d'une durée comprise entre 50 et 250ms. Dans un second temps, des séquences de 2 secondes ont été utilisées pour la classification ; seules les séquences contenant entre 18 et 43 phonèmes ont été retenues, sans autre type de contrainte. Les voyelles et les 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 2 secondes (de 3991 à 800 points) pour des raisons évidentes de mémoire, la résolution temporelle des voyelles /ɑ̃/ reste inchangée (257 points en abscisses et 257 points en ordonnées).

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 les 20% restants sont dédiés àr l'évaluation. Nous avons utilisé l'optimiseur Adam (Kingma & Ba, 2014) avec une valeur initiale du taux d'apprentissage de 1e-4. Cette valeur a été divisée par 2 après 8 itérations complètes sur les données d'apprentissage. Un maximum de 10 itérations en tout a été effectué avec des mini-lots (mini-batches) de 32 exemples, ce qui fait que le modèle a convergé en 28 minutes sur une carte GPU NVIDIA GTX

1080.

3 Résultats

3.1 Représentation des classifications

Les résultats présentent des taux de bonne classification de 85% et 95% pour les voyelles /ɑ̃/ et pour les séquences de 2 secondes respectivement. Dans le cadre de cette étude, notre intérêt se porte plus sur les représentations ainsi obtenues plutôt que sur les taux de classifications, qui sont quoi qu'il en soit, supérieurs à ceux obtenus par des paramètres exclusivement phonétiques (Gendrot et al., 2019).

L'analyse linéaire discriminante (LDA) est souvent utilisée comme un classifieur mais elle peut également être utilisée comme une technique de réduction de dimensionalité en exploration de données de grandes dimensions. Elle modélise chaque objet caractérisé par un grand nombre de paramètres par un point dans un espace à deux ou trois dimensions de telle sorte que les objets similaires sont proches et que les objets dissemblables sont éloignés dans ce nouvel espace. Par rapport à d'autres techniques comme la t-SNE ou UMAP, elle a l'avantage de proposer des distances comparables entre représentations.

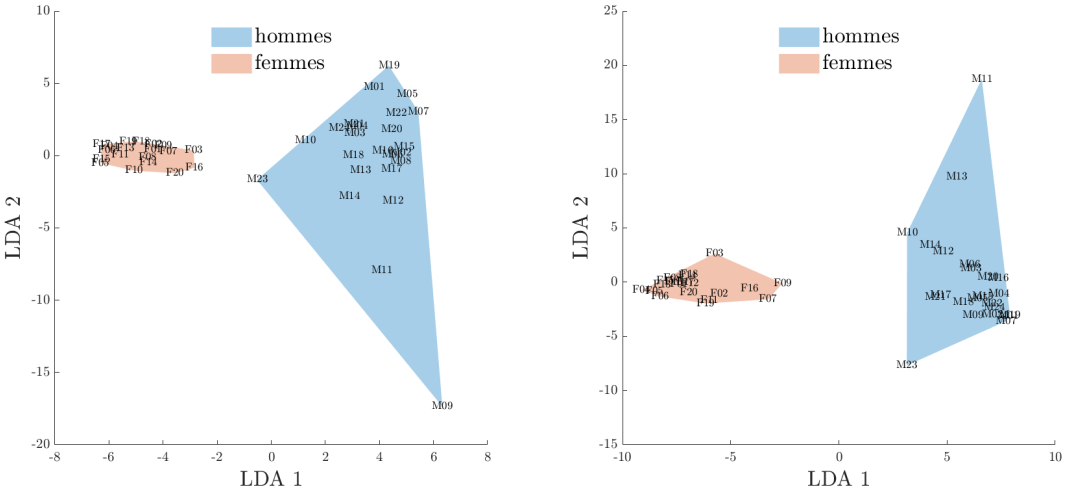

FIGURE 1 – LDA effectuée sur les voyelles /ɑ̃/ (gauche) et les séquences de 2 secondes (droite)

La figure 1 présente les LDA obtenues à partir des classifications des voyelles /ɑ̃/ (à gauche) et des séquences de 2 secondes (à droite) respectivement et sont obtenues à partir des activations de la dernière couche de pooling. Les 2 graphes représentent les 2 jeux de données et montrent une séparation nette entre les sexes des locuteurs avec pour chaque graphe les femmes à gauche et les hommes à droite. La classification effectuée par le CNN pourrait donc s'appuyer sur la f0 des locuteurs ou sur les résonances de leurs cavités, et nous approfondirons ce point dans la section suivante. Les rapprochements entre les locuteurs sont également semblables d'un jeu de données à l'autre : les

locutrices F04, F05, F06, F15 et F13 sont les plus à gauche (sur les 2 figures), alors que F07, F09, F16, F03 et F02 sont les plus à droite pour le groupe des femmes. En ce qui concerne les hommes, M10, M23, M14 et M21 sont les plus à gauche du groupe masculin, alors que M19, M07, M22, M02, M20 et M16 sont les plus à droite, et ce pour les 2 jeux de données. Il conviendra de vérifier par des mesures de f0 si les locuteurs à gauche sont bien caractérisés par une f0 élevée et les locuteurs à droite des 2 figures caractérisés par une f0 basse, et de vérifier s'il existe une progression dans les valeurs entre les positions intermédiaires et les extrêmes.

3.2 Corrélation avec mesures phonétiques et mesures de qualité de voix

Dans cette étude, notre objectif est de comprendre quels paramètres sont utilisés par le CNN pour la classification des locuteurs. Nous avons dans un premier temps repris les coefficients des 512 filtres utilisés dans la précédente section. À ces coefficients, nous avons concaténé une série de mesures acoustiques et nous avons observé leur contribution dans une analyse en composantes principales (ACP). Pour les voyelles /ɑ̃/, nous avons eu recours à openSMILE (Eyben *et al.*, 2010) en utilisant les fichiers de configuration GeMAPS et IS12 utilisant respectivement 18 et 128 mesures recouvrant non seulement des mesures phonétiques telles que les formants, la f0 ou la pente spectrale, mais aussi mesures plus *globales* comme les MFCC ou les moments spectraux. Nous avons également eu recours aux probabilités de type de phonation (modal vs. craqué vs. soufflé) déterminés dans une précédente étude (Chanclu *et al.*, 2021) : en effet, à l'écoute des différents signaux et en comparant les représentations des LDA détaillées ci-dessus, il semblait que le deuxième axe (LD2) correspondait à une variation de la qualité de la voix. Pour finir, puisqu'il s'agissait de voyelles nasales, nous avons appliqué le script scpraat de mesure de nasalité (Styler, NasalityAutomeasure.praat) qui permet d'obtenir des mesures spectrales fines (comme A1-p0) connues comme étant révélatrices de la nasalité.

Pour les séquences de 2 secondes, ces mesures étaient rendues complexes par la combinaison de phonèmes très différents, voisés et non-voisés, la présence de pauses, etc. Nous avons utilisé à nouveau openSMILE avec les 2 fichiers de configuration mentionnés précédemment. Les valeurs réduites à zéro ainsi que les valeurs indéfinies ont été retirées et nous n'avons conservé que les valeurs moyennées sur l'ensemble des 2 secondes pour ne fournir qu'une valeur. Conscients des limites induites par cette méthode, nous devrons trouver par la suite des moyens d'analyse plus adaptés ; mais à ce stade exploratoire de notre travail, nous allons nous en contenter.

L'ACP (figure 2) effectuée pour les voyelles /ɑ̃/ a montré que la f0 était le paramètre acoustique qui contribue le plus (à 97%) à la variabilité observé sur l'ensemble des données mesurées, et qui correspond à la première composante de l'ACP (semblable à l'axe horizontal détaillé sur la figure 1). Les 3 paramètres suivants contribuent à 40% en moyenne à la variabilité et à la deuxième composante de l'ACP et correspondent dans l'ordre à l'indice de Hammarberg, la probabilité de voix craquée/soufflée déterminée issue d'une précédente étude et la pente spectrale sur la bande 0-500Hz. L'indice de Hammarberg est calculé comme la différence d'intensité entre l'intensité maximale dans une bande de fréquence inférieure [0-2000 Hz] et une bande de fréquence supérieure [2000-5000 Hz]. Il est considéré comme un indice de la tension vocale. Ces 3 paramètres sont des indices liés au type de phonation sur le continuum déterminé par Ladefoged et Maddieson (1995) entre glotte fermée et glotte ouverte (en passant de craquée à tendu, puis modal, vers relâché et soufflé) et confirment que la qualité de voix est un paramètre qui entre en compte dans la classification effectuée par le CNN utilisé ici, et que l'on retrouve dans la représentation des données présentée en figure 1.

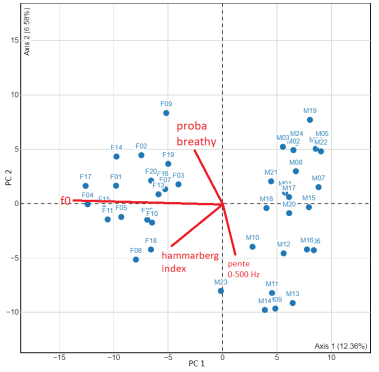

FIGURE 2 – ACP montrant la contribution des facteurs phonétiques à la variabilité des coefficients du CNN pour les voyelles /ɑ̃/

L'ACP effectuée pour les séquences de 2 secondes a montré des résultats semblables bien que les corrélations soient plus faibles, c'est-à-dire la f0 comme la mesure la plus corrélée et ensuite la probabilité de voix craquée/soufflée déterminée et la pente spectrale sur la bande 0-500Hz. Nous avons également d'après les mesures de f0 obtenues un classement hiérarchique (du plus petit au plus grand) des locuteurs et pu observer une tendance très proche de celle observée sur la figure 1, avec des chevauchements plus importants malgré tout.

3.3 Grad-CAM

Dans une précédente étude (Gendrot *et al.*, 2020), afin de quantifier la pertinence des phonèmes pour la caractérisation du locuteur, nous avions effectué un 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'avait 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 ne pouvait suffire que rarement à engendrer une erreur de classification. L'hypothèse que nous avions émise était que le CNN parvenait à se concentrer sur d'autres zones dans la séquence afin de compenser l'absence de la zone masquée.

Ici les zones d'activation obtenues par Grad-CAM permettent d'afficher les zones spectrales et temporelles utilisées par le CNN dans sa classification, leur valeur est contenue entre 0 et 1. Il reste possible d'envisager que d'autres zones que celles affichées soient également partiellement utilisées, mais dans des proportions considérablement réduites.

Pour les voyelles /ɑ̃/, seule la précision fréquentielle des zones d'activation a été prise en compte pour le moment. Nous avons pu mettre en évidence des groupes de locuteurs ayant différents schémas d'activation : une activation simple dans les basses, moyennes ou hautes fréquences, plusieurs activations dans les basses ou moyennes fréquences (voir figure 3). Les fréquences les plus activées

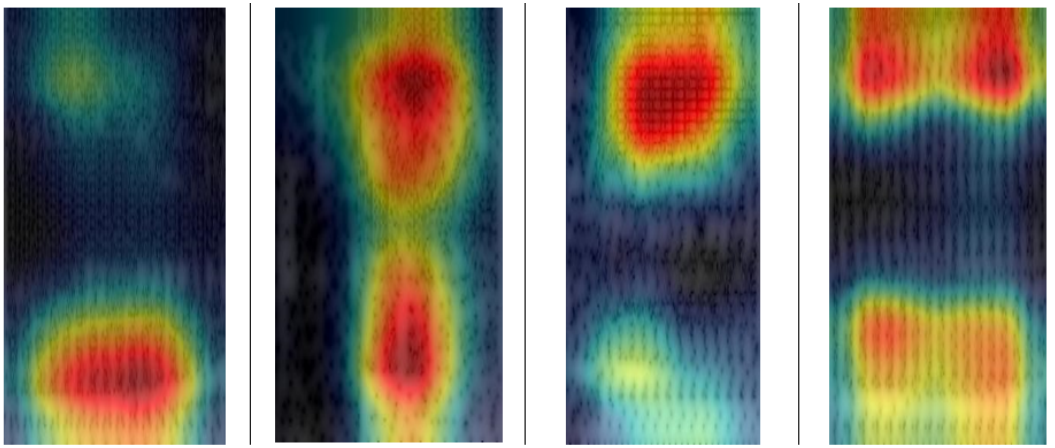

FIGURE 3 – Exemples de Grad-CAM sur les spectrogrammes des voyelles /ɑ̃/ des locuteurs M1, M10, F1, F2. Les zones en rouge foncé indiquent les zones les plus activées lors de la classification

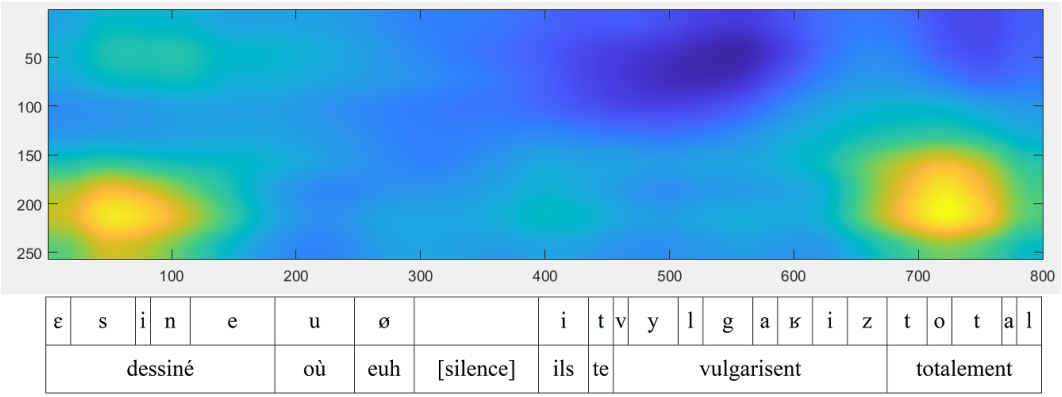

FIGURE 4 – Exemples de Grad-CAM sur une séquence de 2 secondes du locuteur M1

sont aux environs de 2000 Hz et 6500 Hz. La zone d'activation autour de 2000 Hz se situe systématiquement au dessus d'un formant, et il est probable que cette zone corresponde aux anti-formants des nasales (Maeda, 1995). Nous avons pu montrer, sur la base de ces images Grad-CAM (en utilisant les valeurs des pixels des cartes obtenues par Grad-CAM comme paramètre d'entrée dans un classifieur), qu'environ 40% des locuteurs ont été correctement classés par une LDA ce qui montre la pertinence de ces zones pour représenter des locuteurs. Il est à noter également que les mesures phonétiques précédemment citées ne sont pas corrélées à ces zones d'activation.

Quant aux séquences de 2 secondes, nous avons pris en compte à la fois les informations temporelles et fréquentielles étant donné la durée de la séquence. Les phonèmes les plus activés sont /a/, /ɛ/, /ʁ/, /s/, /e/ et /l/, ce qui correspond aux phonèmes les plus fréquents dans notre corpus. La figure 4 montre un exemple d'activation pour une séquence de 2 secondes du locuteur M01. On peut observer que 2 zones sont activées : la première sur le début de la séquence qui correspond à un /s/ (chevauchant le /i/ suivant) et la deuxième vers la fin qui correspond à un /o/ et qui recouvre le début du /t/ qui le suit. Dans les 2 cas, la zone activée reste dans les basses fréquences alors qu'on aurait pu s'attendre à ce que la fréquence activée du /s/ soit plus élevée que la fréquence activée du /o/. Cette illustration se trouve confirmée par l'ensemble des données où la variation inter-locuteurs des zones fréquentielles activées est plus grande que la variation inter-phonèmes ($F=47.2$, $p<0.0001$). Ce résultat montre que les zones d'activation spectrale sont moins spécifiques aux phonèmes qu'aux locuteurs.

4 Discussion et conclusion

La présente étude a permis de montrer que les classifications obtenues sur des voyelles seules et sur des courtes séquences aléatoires des mêmes locuteurs ont des représentations similaires, ce qui laisse à penser que l'aspect phonémique n'est que peu pertinent dans ces résultats. Les mesures phonétiques effectuées en parallèle et corrélées aux coefficients qui ont abouti aux classifications révèlent que la f0 apporte une contribution élevée, suivie par des mesures relatives à la qualité de voix telles que la probabilité de voix craquée/soufflée, l'indice de Hammarberg, et la pente spectrale. Ces résultats démontrent que le CNN utilise la voix dans sa globalité pour catégoriser les locuteurs, sans prendre en compte les phonèmes. L'axe 1 de la LDA (figure 1) apparait par ailleurs beaucoup plus stable et robuste qu'une mesure de f0 acoustique classique car la variation est beaucoup plus faible comparativement aux mesures de f0. Elle s'apparente ainsi à une mesure de hauteur de la voix et laisse envisager des perspectives nouvelles.

Les zones d'activation observées ici sont intéressantes puisqu'elles ne sont pas corrélées aux coefficients du CNN, ni aux mesures phonétiques. Pourtant elles permettent d'obtenir une classification de 40% sur 4 locuteurs. Ces zones spectrales repérées à la fois sur les voyelles /ɑ̃/ et sur les séquences de 2 secondes montrent que les locuteurs ont des zones spectrales caractéristiques qui leur sont propres, même si elles ne permettent pas de les distinguer complètement des autres locuteurs.

Ce travail reste exploratoire et doit être approfondi, notamment pour mesurer plus avant les caractéristiques des zones d'activation. Par exemple, nous ne sommes pas encore en mesure d'expliquer en quoi les zone identifiées par les Grad-CAM influent sur la prise de décision du CNN. La suite de cette étude impliquera des tests de perception afin de mieux identifier ce qui disingue les locuteurs sur le 2ème axe de la LDA.

Références

BYBEE J. (2001). Frequency effects on french liaison. *Typological studies in language*, **45**, 337–360.

CARDOSO A., FOULKES P., FRENCH J. P., GULLY A. J., HARRISON P. T. & HUGHES V. (2019). Forensic voice comparison using long-term acoustic measures of voice quality. In *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS)* : York.

CHANCLU A., AMOR I. B., GENDROT C., FERRAGNE E. & BONASTRE J.-F. (2021). Automatic classification of phonation types in spontaneous speech : towards a new workflow for the characterization of speakers' voice quality. In *Interspeech 2021*, p. 1015–1018 : ISCA.

DELATTRE P. (1951). The physiological interpretation of sound spectrograms. *Pmla*, **66**(5), 864–875.

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.

FANT G. (1960). *Acoustic theory of speech production*. The Hague, The Netherlands, Mouton.

FERRAGNE E., GENDROT C. & PELLEGRINI T. (2019). Towards phonetic interpretability in deep learning applied to voice comparison. In *ICPhS*, p. ISBN–978.

GAHL S., YAO Y. & JOHNSON K. (2012). Why reduce ? phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of memory and language*, **66**(4), 789–806.

GENDROT C., FERRAGNE E. & PELLEGRINI T. (2019). Deep learning and voice comparison : phonetically-motivated vs. automatically-learned features. In *ICPhS*.

GENDROT C., FERRAGNE E. & PELLEGRINI T. (2020). Informations segmentales pour la caractérisation phonétique du locuteur : variabilité inter-et intra-locuteurs. In *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition)*, volume 1.

GOODFELLOW I., BENGIO Y. & COURVILLE A. (2016). Deep feedforward networks. *Deep learning*, (1).

HE K., ZHANG X., REN S. & SUN J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 770–778.

JONGMAN A., WAYLAND R. & WONG S. (2000). Acoustic characteristics of english fricatives. *The Journal of the Acoustical Society of America*, **108**(3), 1252–1263.

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

LEE Y., KEATING P. & KREIMAN J. (2019). Acoustic voice variation within and between speakers. *The Journal of the Acoustical Society of America*, **146**(3), 1568–1579.

SELVARAJU R. R., COGSWELL M., DAS A., VEDANTAM R., PARIKH D. & BATRA D. (2017). Grad-cam : Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, p. 618–626.

STEVENS K. N. (1968). *Acoustic phonetics*. MIT Press.

TITZE I. R. & PALAPARTHI A. (2020). Vocal loudness variation with spectral slope. *Journal of Speech, Language, and Hearing Research*, **63**(1), 74–82.

TORREIRA F., ADDA-DECKER M. & ERNESTUS M. (2010). The nijmegen corpus of casual french. *Speech Communication*, **52**(3), 201–212.