Sorbonne Nouvelle ED 622 sciences du langage

Université Sorbonne Nouvelle - CNRS

Ecole Dotorale 622 : Sciences du Langage
UMR7018 Laboratoire de Phonétique et Phonologie

Thèse en vue de l'obtention du grade de
Docteur en Phonétique:

RÉSUMÉ SUBSTANTIEL

Les composantes de la parole dans la caractérisation phonétique du locuteur

étude sur la complémentarité et la redondance véhiculées des informations

Presentée par:

Gabriele CHIGNOLI

gabriele.chignoli@sorbonne-nouvelle.fr

Sous la direction de:

Cédric GENDROT

Rapporteur: **Damien LOLIVE** HDR, CNRS - IRISA - Université de Rennes 1
Rapporteure: **Ioana VASILESCU** HDR, CNRS - LISN - Université Paris-Saclay
Examinatrice: **Cécile FOUGERON** DR, CNRS - LPP - Université Sorbonne Nouvelle
Examinateur: **Jean-François BONASTRE** PR, CNRS - LIA - Université d'Avignon
Examinatrice: **Christine MEUNIER** DR, CNRS - LPL - Université Aix-Marseille
Directeur de thèse: **Cédric GENDROT** HDR, CNRS - LPP - Université Sorbonne Nouvelle

Paris, September 15, 2022

Résumé substantiel

La parole que nous produisons et percevons chaque jour est le plus souvent spontanée. Elle ne nécessite aucun entraînement particulier, hormis le processus d'acquisition qui a lieu au cours des premières années de croissance, ou certaines situations spécifiques telles que les cours de langues étrangères. De plus, l'efficacité de la parole est un trait remarquable compte tenu à la fois de la charge cognitive minimale qu'elle requiert et du large éventail d'informations qu'elle peut véhiculer. Pendant la production et la perception de la parole, en tant qu'êtres humains, nous sommes capables d'opérations telles que la localisation de ponctuations cachées, la gestion des disfluences ou la récupération au-delà du sens des mots. Ce ne sont là que quelques conséquences des mécanismes intrinsèques hautement sélectifs et optimisés qui s'activent dans notre cerveau.

Le processus de caractérisation d'une voix est l'objet d'étude de cette thèse. Afin de caractériser un objet fortement variable comme la voix, un processus en deux étapes est nécessaire. D'abord, la décomposition du signal vocal, ensuite, l'analyse de ses différentes composantes afin de les associer à l'information qu'elles portent. Les composantes de la voix sont communément étudiées au moyen de mesures phonétiques. Comment définir les composantes à analyser ? Comment interagissent-elles entre elles ? Quels sont les facteurs qui influencent le plus leurs variations ? Ce sont là certaines questions auxquelles la littérature phonétique sur la caractérisation de la voix a tenté de répondre. Nous en ajoutons d'autres auxquelles nous tenterons de répondre, telles que : L'apprentissage automatique et la phonétique peuvent-elles être des approches complémentaires ? Quelles sont les implications de la perception humaine dans les domaines de la caractérisation des locuteurs ?

L'objectif principal de cette thèse est donc d'analyser comment les caractéristiques individuelles peuvent être récupérées par différentes composantes de la parole et leurs interactions, en ajoutant des connaissances sur la distribution de l'information du locuteur par l'étude des productions vocales en français. Pour ce faire, nous effectuons trois analyses dans le présent travail.

La première partie de ce résumé est une introduction générale à l'idée de caractérisation de la voix et ses implications. Nous nous penchons également sur des travaux sur les domaines de la caractérisation de la voix qui ont influencé cette thèse. Ces travaux fixent les concepts de base de cette thèse et des trois domaines qu'elle aborde : la phonétique, l'apprentissage automatique, et la perception.

Dans la deuxième partie, nous présentons nos méthodes et nos résultats. D'abord, une étude perceptive est présentée, sous-section 2.1. Cette dernière, relie les résultats obtenus par les différentes approches de caractérisation du locuteur, qui seront présentées par la suite. Ainsi, les ressemblances et les différences sont soulignées en offrant des pers-

pectives pour une analyse plus approfondie. Ensuite, nous présentons nos études sur les caractéristiques des locuteurs à travers des approches phonétiques classiques. Des mesures acoustiques et des corrélats rythmiques sont testés, montrant des résultats en accord avec d'autres études phonétiques. Nous appliquons également des méthodes plus proches de celles de la phonétique légale afin de modéliser la dynamique des caractéristiques acoustiques de la parole lue. Enfin, dans la sous-section 2.3, nous fournissons de nouvelles méthodes de représentation de différentes composantes de la parole en utilisant des techniques d'apprentissage profond, par le biais de réseaux de neurones à convolution (CNN).

Enfin, la troisième et dernière partie discute de nos analyses en reliant la revue de littérature et nos résultats. Les principales contributions de cette thèse nous permettent d'expliquer comment l'information des locuteurs est distribuée dans les différentes composantes de la parole pour des locuteurs français. En outre, la fusion de plusieurs approches est discutée, montrant l'importance, du point de vue du traitement automatique du langage naturel (TAL), d'appliquer les théories de la linguistique classique de manière complémentaire avec les méthodes modernes. Certaines questions non résolues et des propositions pour d'éventuelles recherches futures sont également suggérées.

1 État de l'art de la caractérisation phonétique du locuteur

Premièrement, nous regardons la manière dont la littérature phonétique a défini les composantes de la parole par le biais de mesures phonétiques. La composante appelée source et filtre est sans doute la plus étudiée. La prosodie et la qualité de la voix représentent des ensembles plus hétérogènes dans lesquels de multiples indices ont été mis en relation. La prosodie se concentre sur l'étude d'indices temporels et des changements d'intonation, tandis que la qualité de la voix se concentre sur les paramètres laryngés et supra-laryngés qui influencent à la fois la partie harmonique et le bruit des productions de la parole. Les mesures articulatoires font également partie d'un autre vaste ensemble qui a donné de nombreux résultats concernant l'analyse entre locuteurs. Les corpus sur lesquels nous nous concentrons ne présentant pas de données articulatoires, ce dernier ensemble de composantes ne fait pas partie de nos recherches. Une liste complète des mesures phonétiques et des relatives composantes que nous avons sélectionnées pour notre étude est présentée dans le Tableau 1.

Lors de la communication orale, différentes sources fonctionnent simultanément pour créer une onde complexe de signal vocal. La première composante que nous considérons est dite *source et filtre*, dont les formants et le f0 sont les contributeurs les plus importants. Dans cette composante, nous retrouvons principalement des caractéristiques biophysiques qui se rapportent aux mesures phonétiques des sources de la parole telles que la glotte ou le bruit et les filtres qui, à travers les cavités vocales, modifient l'onde vocale. Dans des études telles que [Hudson et al., 2007; Boë et al., 1975; Rose and Wang, 2016; Lindh, 2006; Eriksson and Wretling, 1997], f0 est décrit comme la caractéristique la plus efficace pour reconnaître les locuteurs dans plusieurs langues. L'étude des résonances est mise en évidence dans [Zuo and Mok, 2012; McDougall and Nolan, 2007], et montre que les voyelles nasales sont les segments portant le pouvoir discriminant le plus important pour la reconnaissance des locuteurs comme discuté dans [Kahn, 2011; Ajili et al., 2016; Gendrot

et al., 2020]. Cependant, de nombreux autres corrélats de la composante source et du filtre sont présents dans la littérature, tels que les indices d'enveloppe ou les moments spectraux [Culling and Darwin, 1993; Ardoint and Lorenzi, 2009; He and Dellwo, 2017; Niebuhr and Skarnitzl, 2019], et étudiés en tant que caractéristiques du locuteur.

D'autres composantes qui ont fait l'objet d'études sur la caractérisation du locuteur et nous intéressent dans cette thèse sont la **temporelle** et la **ligne prosodique**, en suivant la nomenclature donnée par [Kreiman and Stidtis, 2011]. Ces dernières sont communément connues sous la définition plus large de *prosodie*. En phonétique, le terme prosodie désigne largement tout aspect supérieur au plan segmental, sans se référer à une seule composante ou mesure phonétique. Les études sur la caractérisation des locuteurs se concentrent principalement sur les indices temporels [Mary and Yegnanarayana, 2008; Künzel, 2013; Leemann and Kolly, 2015; Kolly et al., 2015] et les changements d'intonation [Barlow and Wagner, 1988; Cangemi, 2009; Ouyang and Kaiser, 2015], pour lesquels la dynamique de f0 est étudiée.

En outre, comme mentionné ci-dessus, une mesure phonétique ne peut pas être considérée comme strictement associée à une seule composante, en raison des multiples mesures qui peuvent être examinées dans le cadre de différentes composantes. Par exemple, la ligne prosodique, la structure syllabique et les indices temporels peuvent être considérés comme représentant la Prosodie, l'articulation et le cadre articulatoire peuvent aussi être regroupés dans une composante articulatoire plus large et ainsi de suite. En considérant les mesures phonétiques, nous observons, comme nous l'avons dit précédemment, qu'elles peuvent être utilisées dans l'étude des différents aspects de la parole dépendant de la manière dont le chercheur les utilise.

Étant donné la dichotomie acoustique-articulatoire dans la description des mesures phonétiques, un autre aspect important à prendre en compte est l'étendue de l'étude de ces mesures. Comme mentionné plus haut, la parole est un objet dynamique complexe formé par l'articulation ultérieure d'unités plus petites, qui sont les phonèmes dans une langue choisie, dont les interactions et les influences les unes sur les autres aboutissent à ce que nous ressentons comme des unités de parole. Considérer des unités plus grandes que des segments isolés implique la prise en compte de ce qui a été appelé la dimension prosodique ou suprasegmentale. La prosodie n'est pas une composante unique de la parole, mais représente un large ensemble d'éléments différents permettant la segmentation de la parole continue en unités plus petites, porteuses d'informations sur la langue parlée ou l'organisation du discours, et éventuellement sur les caractéristiques du locuteur. Avant de plonger dans les différents indices prosodiques qui ont été étudiés en relation avec les caractéristiques du locuteur, nous discutons de certains aspects généraux de la prosodie dans la littérature phonétique.

La prosodie implique des mécanismes qui s'étendent au-delà du niveau phonémique de la parole et, pour la plupart, indépendamment des caractéristiques segmentaires, contenues dans des unités linguistiques organisées de manière hiérarchique. Au niveau acoustique, certaines de ces caractéristiques prosodiques sont le f0, l'intensité et la durée. Nous avons déjà abordé ce que représentent f0 et l'intensité lorsqu'ils sont étudiés à un niveau inférieur, mais, dans une perspective prosodique, leurs modulations sont prises en compte. La durée concerne la longueur des unités de parole et des intervalles silencieux.

La discrimination des langues couvre une grande partie des études prosodiques, y compris chez les adultes. Comme décrit par [Arvaniti, 2013; Ramus and Mehler, 1998], la durée, le

rythme, l'intonation ou le débit de parole analysés dans une perspective suprasegmentale peuvent expliquer une très grande partie des différences linguistiques. Ces deux travaux ont utilisé une approche similaire pour étudier le rôle joué par les indices prosodiques et la manière dont les humains les utilisent pour différencier les différentes langues. [Ramus and Mehler, 1998] se sont particulièrement concentrés sur le rythme syllabique qui apparaît comme un excellent, et peut-être le meilleur, indice prosodique pour la discrimination des langues qui seraient différentes en termes de rythme.

La discrimination avec des stimuli appauvris est difficile et les auditeurs profitent de toutes les différences qu'ils peuvent trouver pour les aider dans cette tâche. Ces différences concernent principalement le débit de parole et le f0, mais aussi des différences temporelles localisées, comme l'allongement final, lorsque le débit de parole et le f0 sont absents. Étant donné que tous ces facteurs prosodiques sont normalement présents dans le signal vocal et interagissent dans la perception, les présents résultats mettent en doute l'idée que le timing est primaire et peut être traité par les auditeurs indépendamment de toutes les autres variables prosodiques. La discrimination était possible à la fois entre les classes de rythme et à l'intérieur de celles-ci, lorsque les taux de parole différaient entre le contexte et le test. Cependant, cette dernière est largement impossible une fois les différences de taux de parole éliminées. Les changements dans les réponses associés à la vitesse d'élocution et à f0 indiquent que la discrimination linguistique résulte d'interactions entre les facteurs prosodiques et que le timing y contribue, mais à une échelle moindre.

Étant donné les vastes discussions sur le poids des indices prosodiques dans la production et la perception de la parole, il est évident que les auteurs se sont penchés au fil des années sur les implications possibles pour l'analyse des différences individuelles. Nous avons déjà cité [Barlow and Wagner, 1988] comme un travail précoce sur le sujet. Dans cette étude, les auteurs utilisent des indices tels que les modulations de f0, d'énergie ou de voisement afin de modéliser les modèles correspondant à chacun des cinq locuteurs étudiés dans des phrases lues en anglais. La comparaison des patrons à l'aide de la prédiction linéaire et des distances de clustering montre des résultats très fiables, soulignant le rôle joué par les indices prosodiques dans la caractérisation des habitudes de parole des locuteurs. Les études focalisées sur les indices prosodiques pour analyser les différences individuelles feront l'objet des sections suivantes, en distinguant les indices reposant sur la composante **temporelle** et ceux reposant sur la composante **ligne prosodique**.

Certaines études utilisent les caractéristiques prosodiques en relation avec les composantes du modèle de variabilité entre locuteurs, telles que la durée, l'intensité, les corrélats rythmiques, ou d'autres représentations dynamiques, sans épuiser la compréhension que nous avons des informations qu'elles véhiculent. Toutes incluent, plus ou moins explicitement, l'idée qu'un certain modèle ou événement se répète dans le temps et que ces modèles peuvent être utilisés pour caractériser les résultats de sources très hétérogènes, comme les locuteurs. La place des indices suprasegmentaux dans la caractérisation de la voix et leur relation avec les autres composantes et l'information véhiculée doivent encore être pleinement comprises.

Ainsi, le **mode de vibration des plis vocaux**, ou **qualité de la voix**, représente, comme la prosodie, un ensemble plus hétérogène auquel de multiples indices sont associés. L'accent est mis sur les paramètres laryngés et supra-laryngés qui influencent les sources harmoniques et inharmoniques de la production vocale, par exemple dans [Hughes et al., 2019; Vaňková and Skarnitzl, 2014]. Parmi les études prenant en compte plusieurs

composantes, il nous semble important de citer les travaux de [Lee et al., 2019; Keating and Kreiman, 2016; Keating et al., 2017], qui ont eu une grande influence sur cette thèse. Ces trois études se concentrent sur la même question : l'analyse des similitudes entre les voix du locuteur à travers différentes mesures phonétiques principalement associées à la source et au filtre. La qualité de la voix recouvre aussi une place importante dans ces études qui tentent de définir les rôles que ces composantes jouent dans les matrices d'identité du locuteur. Le même ensemble de données, composé de cinquante femmes et cinquante hommes provenant de la base de données sur la variabilité des locuteurs de l'Université de Californie, Los Angeles [Keating et al., 2019], est utilisé. Dans [Lee et al., 2019], les deux sexes sont analysés et comparés, tandis que [Keating and Kreiman, 2016] examine uniquement les locutrices et [Keating et al., 2017], uniquement les locuteurs.

Les mesures phonétiques utilisées sont les suivantes : (i) la fréquence fondamentale (f0) ; (ii) les formants du premier au quatrième (F1-4) et leur dispersion (FD), calculée comme la différence moyenne de fréquence entre chaque paire adjacente de formants ; (iii) la forme spectrale de la source harmonique comme les amplitudes relatives des première et deuxième harmoniques (H1-H2) et des deuxième et quatrième harmoniques (H2-H4), les pentes spectrales de la quatrième harmonique à l'harmonique la plus proche de 2 kHz en fréquence (H4-H2k) et de l'harmonique la plus proche de 2 kHz à l'harmonique la plus proche de 5 kHz en fréquence (H2k-H5k) ; et (iv) la variabilité de la source harmonique/du bruit spectral la proéminence du pic cepstral (CPP), l'énergie et le rapport d'amplitude entre les sous-harmoniques et les harmoniques (SHR).

Les résultats obtenus dans l'analyse intra-locuteur montrent que H2k-H5k et CPP représentent la plus grande partie de la variabilité, ayant les poids les plus élevés dans l'Analyse en Composantes Principales (ACP), suivis par FD et F4. Cela suggère que, parmi toutes les mesures phonétiques étudiées, ces dernières entraînent une distribution plus stable dans la matrice de variation du locuteur, mais définissent toujours un ensemble limité d'éléments. En outre, le fait que les deux premières composantes principales (CP) représentent moins de la moitié de la variance expliquée (32 % et 34 % respectivement pour les locuteurs féminins et masculins), confirme l'idée de matrices hautement variables dans la description des caractéristiques des locuteurs.

En examinant les résultats de la variabilité entre locuteurs, les auteurs observent des similitudes entre les trois PC des deux sexes. Les mesures phonétiques similaires jouent des rôles similaires dans la description des différences individuelles, sauf pour le coefficient de variation de H1-H2 qui montre un poids conséquent uniquement pour les hommes. Les mesures les plus importantes sont H2k-H5k, CPP et H2-H4. En effet, ces trois mesures sont présentes dans le premier PC et représentent 18 % et 20 % de la variance chez les femmes et les hommes, respectivement. Les fréquences formantiques (F4, FD, F3), correspondent à la deuxième composante pour les femmes et à la troisième pour les hommes, respectivement, représentant 11 % de la variance des voix féminines et 9 % pour les voix masculines. De même que la deuxième PC pour les femmes correspond à la PC3 pour les hommes, la PC3 des locuteurs féminins correspond à la PC2 des locuteurs masculins. Pour les deux sexes, le PC correspond à la pente spectrale dans les hautes fréquences, H2k-H5k, ainsi que F2. Ce PC représente 10 % de la variance chez les locuteurs féminins et masculins.

Le deuxième ensemble de notions fixées dans cette première partie de notre travail concerne l'approche multidisciplinaire que cette thèse adopte. Suivant l'idée, énoncée dans [Nolan, 2001; Bonastre et al., 2003; Morrison and Thompson, 2017], qu'il n'existe aucun pro-

cessus scientifique permettant de caractériser de manière unique la voix d'une personne, différents types de reconnaissance ont été associés à la reconnaissance du locuteur. Ces quatre processus définissent diverses applications qui influencent le présent travail. La **reconnaissance auditive** est définie comme la capacité humaine à écouter et à reconnaître les locuteurs sur la base de leur seule voix, avec plus ou moins de succès. Elle constitue la base de la reconnaissance du locuteur. Les auditeurs présentent différents niveaux de capacité à reconnaître des locuteurs et divers facteurs peuvent affecter la fiabilité de cette méthode. En effet, des facteurs supplémentaires peuvent augmenter la fiabilité de la reconnaissance auditive, tels que la familiarité avec le locuteur, la durée ou le contexte de l'échantillon audio, l'absence de bruit de fond ou le stress/déguisement vocal.

La **reconnaissance spectrographique** implique, comme son nom l'indique, l'utilisation d'une représentation spectrographique pour effectuer l'identification du locuteur. Le sous-comité d'identification de la voix et d'analyse acoustique de l'Association internationale pour l'identification[1] fournit des méthodes pour effectuer des comparaisons spectrographiques de la voix fiables et uniformes dans le cadre d'une enquête médico-légale. Conformément aux dites normes, les échantillons connus et inconnus doivent contenir des mots comparables. Un examen ne peut produire qu'une seule des sept décisions suivantes : 1-Identification, 2-Identification probable, 3-Identification possible, 4-Inconclusion, 5-Élimination possible, 6-Élimination probable, 7-Élimination. Les spectrogrammes ont été communément appelés "empreintes vocales", ce qui conduit à une comparaison erronée avec les empreintes digitales, indices biométriques invariables d'une personne et très difficilement falsifiables. Les spectrogrammes affichent les signaux vocaux dans une représentation $temps \times fréquence \times intensité$, où les axes horizontal et vertical représentent respectivement les variations de temps et de fréquence, tandis que la profondeur de couleur représente l'intensité. Ils constituent des outils d'ingénierie et d'analyse de la voix utiles pour son étude. L'impression d'un spectrogramme peut être considérée comme la seule raison de l'utilisation du terme "empreinte vocale".

Dans la **reconnaissance phonétique légale du locuteur**, les informations du locuteur sont analysées selon une approche linguistique, l'expert se concentrant sur l'extraction de traits ou de caractéristiques spécifiques. Cependant, contrairement aux deux approches citées, elle donne une estimation numérique de la probabilité que l'échantillon de voix examiné soit produit par un locuteur désigné, selon le rapport de vraisemblance de la théorie bayésienne. Une autre définition pourrait être d'estimer combien de fois il est plus probable d'observer les différences entre les échantillons en supposant qu'ils proviennent du même locuteur, plutôt que de locuteurs différents [Rose, 2002]. Enfin, nous avons la **reconnaissance automatique du locuteur (RAL)**. Son idée de base est d'effectuer une reconnaissance via une machine et en utilisant des procédures automatiques. Les techniques de RAL les plus récentes reposent sur des mesures de similarité entre les paramètres acoustiques extraits d'un ensemble d'enregistrements. Ces mesures peuvent prendre en compte les distributions statistiques pour un locuteur particulier, le contenu du message et/ou des informations sur l'environnement et le support d'enregistrement. Cependant, l'utilisation d'un système entièrement automatisé peut s'avérer difficile et est limitée par les contraintes de progression du domaine.

Dans une perspective similaire, la mise en parallèle de la reconnaissance du locuteur par un système automatique avec la perception humaine montre des variations de performance

1. Voice Identification and Acoustic Analysis Subcommittee of the International Association for Identificaiton.

qui peuvent être attribuées à des facteurs liés aux locuteurs. Dans ce sens, les différences dans la reconnaissabilité des locuteurs par les systèmes automatiques ont été définies en utilisant des animaux dans [Doddington et al., 1998; Stoll and Doddington, 2010] : Les moutons font référence aux locuteurs pour lesquels de faibles performances sont observées, et dont les caractéristiques ne présentent pas de variation importante par rapport à la moyenne de la population observée ; en revanche, les chèvres représentent les locuteurs dont les caractéristiques sont très visibles, et donc plus faciles à identifier ; les agneaux et les loups sont les locuteurs dont les caractéristiques permettent respectivement une imitation facile et une imitation particulièrement réussie d'autres locuteurs. Ainsi, la capacité à décoder les caractéristiques des locuteurs montre une influence importante de la comparaison avec la population générale plutôt qu'avec le locuteur isolé. Comme déjà mentionné, l'importance de la perception humaine doit être aussi considérée. La fiabilité d'auditeurs humains est une idée qui traverse plusieurs domaines, car elle est importante à la fois dans les perspectives automatiques et dans la perception afin de valider les informations phonétiques que nous analysons. Afin de comprendre ce dont les machines sont capables, la comparaison avec les capacités humaines est fondamentale. De plus, les informations que les humains peuvent réellement extraire du signal vocal et la manière de les représenter sont des questions importantes auxquelles il faut répondre afin d'accroître la fiabilité de la recherche sur la parole. L'interprétation des résultats et des composantes sélectionnées est impossible sans l'ajout de résultats liés à l'humain, ce qui permet d'établir un lien entre tous les domaines étudiés.

Dans ce sens, la perception humaine étant un aspect aussi important à considérer, [Latinus and Belin, 2011; Belin and Grosbras, 2010; Kreiman, 1997] ont discuté de la façon dont le cortex auditif de l'être humain présente un traitement différent de la voix et des niveaux intermédiaires de stimuli. L'activité cérébrale observée par imagerie par résonance magnétique fonctionnelle chez des sujets normo-entendant a montré des activations le long des parties antérieure et moyenne du sillon temporal supérieur (STS). Ces activations restent inchangées lorsque les stimuli vocaux sont joués à l'envers, ce qui supprime la majeure partie du contenu linguistique mais laisse intact le timbre de la voix. Ces zones dites "zones temporales de la voix" [2] sont sensibles aux voix, contenant ou non de la parole, et apparaissent quelques mois après la naissance. En outre, la sensibilité humaine à des changements spécifiques des caractéristiques de la parole a été montrée, des caractéristiques de la source et du filtre [Gamal, 2015; Sorin, 1981] jusqu'aux changements temporels [Hollien et al., 1982; van Lancker et al., 1985], en passant par les variations de la pente spectrale harmonique et des niveaux de bruit de la qualité de la voix [Kreiman and Gerratt, 2012].

Une dernière considération importante sur la caractérisation de la voix concerne la capacité elle-même de reconnaître un locuteur et donc de marquer un certain aspect d'une comme caractéristique. Ce mécanisme implique un décodage des composantes de la voix entendue. Le résultat décodé peut être associé soit à une voix familière, rappelant des concepts de perception catégorielle, soit à une voix inconnue qui crée une nouvelle catégorie chez l'auditeur. Le codage et le décodage de diverses informations par le biais de productions vocales ont déjà été étudiés [Ohala, 1983, 1994; Dediu et al., 2017; **?**]. Des études suggèrent que la reconnaissance de voix familières est ancienne dans l'évolution [Grossmann et al., 2010; Belin and Grosbras, 2010] et a émergé par exemple chez les amphibiens [Burke and Murphy, 2007; Bee and Gerhardt, 2002], les oiseaux [Hsu et al., 2004] ou les primates [Furuyama et al., 2016; **?**]. De ce fait, cette capacité précède consi-

2. Temporal Voice Areas.

dérablement le développement évolutif de la parole et du langage dans la communication et la cognition humaines.

2 Résultats

Dans cette deuxième partie nous allons résumer l'ensemble de nos résultats, offrant une vue d'ensemble de l'étude des composantes de la parole à travers différentes approches visant à comprendre plus en profondeur la caractérisation phonétique des voix. Les analyses effectuées contribuent à élargir les aspects fondamentaux de la caractérisation phonétique que nous avons évalués. En examinant les réponses des auditeurs humains sur les clusters de voix, nous les comparons à celles obtenues avec la modélisation phonétique et CNN afin d'étudier leurs éventuelles différences et similitudes. Les résultats des approches phonétique et CNN mettent en évidence la qualité de la voix ou les formants comme la composante de la parole la plus fiable pour caractériser les voix des locuteurs.

Le Tableau 1 récapitule toutes les mesures phonétiques extraites et les noms de groupes correspondants que nous utilisons tout au long de nos investigations. La composante correspondante est indiquée ainsi que l'approche dans laquelle le groupe est utilisé. Nous avons principalement utilisé les logiciels Praat [Boersma, 2001] et VoiceSauce [Shue et al., 2011] pour extraire les mesures sélectionnées. Ce dernier fournit principalement des descripteurs de la qualité de la voix, absents dans Praat, et offre le choix de mesures acoustiques provenant de plusieurs programmes externes pour des paramètres tels que la fréquence fondamentale et les formants. En appliquant la même méthode d'extraction dans les deux logiciels, nous avons obtenu des valeurs toutes les millisecondes afin d'avoir une reconstitution précise des modulations des paramètres pour chaque locuteur cible. Nous avons pris des mesures sur les consonnes voisées pour des paramètres tels que l'énergie et avons inclus les fricatives pour les moments spectraux. Certains groupes utilisent des fenêtres d'analyse plus larges, par exemple 10 ms pour le coefficients cepstraux à fréquence Mel (MFCC).

Les trois approches sont basées sur l'état de l'art décrit ci-dessus. Quatre composantes de la parole sont les objets principaux de notre étude. La première approche concerne la **Perception**, qui repose sur une tâche de clustering perceptif par un groupe hétérogène d'auditeurs. La majorité de nos auditeurs étant des femmes de langue maternelle française, nous avons pu comparer des sous-groupes basés sur le sexe, la langue maternelle (groupe Natif et groupe Non-Natif) et l'expertise phonétique (Experts et Naïfs) comme principales variables de contrôle. Les sujets devaient écouter des séquences de discours spontanés et les regrouper par leur similarité sur la base de ce qu'ils pensaient être la caractéristique la plus pertinente. Nous avons calculé les distances entre les groupements résultants afin d'évaluer la similarité entre les groupements des auditeurs et ceux obtenus par les deux autres approches sur lesquelles nous nous sommes concentrées.

Dans l'approche **Phonétique**, nous avons effectué différentes analyses selon la littérature phonétique classique, à savoir : une ACP, des modèles mixtes linéaires ou une analyse discriminante linéaire (LDA). La comparaison entre les résultats statistiques et la représentation de la dynamique de la parole est un point important de cette approche. Cette dernière est réalisée par des mesures de type quotient ouvert inspirées de [He and Dellwo, 2017; Kreiman and Shue, 2010] et par des coefficients polynomiaux à travers une analyse

du rapport de vraisemblance (LLR) de comparaison de voix comme dans [san Segundo and Yang, 2019; McDougall and Nolan, 2007]. Cependant, une représentation plus précise des changements temporels de la parole est présente dans notre troisième approche. En effet, dans le cadre de l'approche par **TAL** ou automatique, nous avons utilisé des CNN dans trois tâches de reconnaissance de locuteurs : une tâche d'identification (1 sur N), une tâche de vérification (classification binaire entre un locuteur cible et une population), et une tâche de généralisation (similaire à la tâche de vérification mais avec des locuteurs supplémentaires inconnus du CNN pendant la phase de test). Les trois tâches visent à mettre en parallèle les différents mécanismes de perception impliqués dans la caractérisation du locuteur afin d'observer comment un modèle d'apprentissage automatique peut réagir et quelles sont les implications pour la caractérisation du locuteur qui en résulte.

Groupe	**Mesures**	**Composante**	**Approche**
Amp	Amplitudes des harmoniques près des trois premiers formants (A1-3)	Source et filtre	Perception & Phonétique & TAL
f0	f0 et ses harmoniques (H1, H2, H2k, H42k, H5k)	Source et filtre	Perception & Phonétique & TAL
Form	Quatre premiers formants (F1-4)	Source et filtre	Perception & Phonétique & TAL
Acoust	Association de *Amp*, *f0* et *Form*	Source et filtre	Perception & Phonétique & TAL
Int	f0 et intensité	Ligne prosodique	TAL
Env	ENV et TFS	Temporelle	TAL
Pros	Association de *Env* et *Int*	Prosody	TAL
Nrg	RMS, soe, énergie basé sur Praat	Source et filtre | Mode de vibration des plis vocaux	Perception & Phonétique & TAL
Ms	Quatre moments spectraux (centre de gravité, déviation standard, kurtosis, skewness)	Source et filtre | Mode de vibration des plis vocaux	Perception & Phonétique & TAL
Ha	Différences entre harmoniques (H1-H2, H2-H4, H2k-H5k)	Mode de vibration des plis vocaux	TAL
Hh	Différences entre harmoniques et amplitudes (H1-A1, H1-A2, H1-A3)	Mode de vibration des plis vocaux	TAL
Hadiff	Association de *Ha* et *Hh*	Mode de vibration des plis vocaux	Perception & Phonétique & TAL
Hr	HNR à différent gamme de f0 (0-500 Hz HNR05, 0-1500 Hz HNR15, 0-2500 Hz HNR25, 0-3500 Hz HNR35), SHR	Mode de vibration des plis vocaux	Perception & Phonétique & TAL
Ltas	LTAS à quatre largeurs de bande de fréquence entre 1 et 5 kHz	Mode de vibration des plis vocaux	Perception & Phonétique & TAL
Qual	Association de *Ltas*, CPP et *Nrg*	Mode de vibration des plis vocaux	Perception & Phonétique & TAL
MFCC	13 MFCC	?	Perception & Phonétique & TAL
Glob	Association de toutes les mesures phonétique	Toutes	Perception & TAL
Spectros	Spectrogramme à bande large	Toutes	Perception & TAL
MPS	Spectre de puissance de la modulation	Source et filtre	TAL
Rythme	Intensité, ENV et TFS	Temporelle	Perception & Phonétique
Temporelle	Débit, PVI consonantique et vocalique, pauses et durées segmentales	Temporelle	Phonétique
Dynamiques	Mesures basées sur le quotient ouvert sur l'intensité, coefficients polynomiaux pour f0 et les formants	Source et filtre	Phonétique

TABLE 1 – Groupes de mesures phonétiques utilisés tout au long de nos expériences, composants connexes et approche dans laquelle ils sont utilisés.

Pour cela, les deux corpus français utilisés tout au long de nos études sont le Nijmegen Corpus of Casual French (NCCFr) et PTSVOX, documentés respectivement dans [Torreira et al., 2010] et [Chanclu et al., 2020]. Ils présentent certaines similitudes telles que la présence de locuteurs natifs français des deux sexes et des tranches d'âge similaires. Cependant, leurs matériaux linguistiques sont très différents puisque NCCFr est composé d'enregistrements de paroles spontanées, alors que PTSVOX est composé de paroles lues. NCCFr est composé de 44 locuteurs (21 femmes et 23 hommes), les enregistrements ont une durée moyenne de 40 minutes et ont été effectués dans une chambre silencieuse du Laboratoire de Phonétique et Phonologie de Paris. Tous les sujets portaient un micro-casque afin de réduire les variations d'intensité dues aux mouvements de la tête. Les locuteurs sont principalement âgés de 18 à 27 ans, avec deux femmes âgées de 40 et 50 ans. Ils sont tous de langue maternelle française, 34 d'entre eux viennent de la région parisienne, tandis que les autres viennent d'autres régions du centre et du nord de la France. Tous les locuteurs étaient issus du même milieu social et poursuivaient des études similaires. La tâche de conversation associait toujours des locuteurs de même sexe.

Le deuxième corpus, PTSVOX, visait à recréer les principaux défis de la comparaison de voix en médecine légale. Les discours spontanés et lus sont présents dans ce corpus et enregistrés dans des conditions de microphone et de téléphone. Comme le PTSVOX a été créé pendant le travail de cette thèse, les annotations n'étaient pas immédiatement disponibles pour tous les enregistrements. Pour cette raison, nous n'avons considéré que le sous-ensemble de lecture pour nos expériences. Il est composé de deux sessions de trois passages lus par un total de 24 locuteurs natifs français (12 femmes et 12 hommes). Les locuteurs qui ont présenté des sessions incomplètes, c'est-à-dire qu'il leur manquait un ou plusieurs passages lus ou une session d'enregistrement entière, ont été écartés.

2.1 Perception

Ce qui différencie les résultats perceptifs des autres qui suivront est la présence de l'auditeur comme deuxième facteur de variabilité. Effectivement, dans nos études sur les interactions entre les caractéristiques de la voix, le locuteur est l'unique facteur de variabilité. Cependant, dans une étude de la perception, l'auditeur représente un facteur de variabilité fondamental à prendre en compte. La perception humaine, en particulier lors de l'étude de la caractérisation de la voix, peut être influencée par une multitude de variables telles que la familiarité de l'auditeur avec les locuteurs, ou la langue parlée, ainsi que la conception de la tâche en général.

L'âge de nos participants varie de 21 à 77 ans, avec une moyenne de 27, vingt et un d'entre eux sont des femmes contre six hommes. Étant donné le petit échantillon que nous avons pu rassembler, les résultats de ces derniers, ainsi que ceux des auditeurs français non natifs, doivent être pris avec prudence et considérés comme des tendances ou des hypothèses possibles. Pour les regroupements de voix féminines, tous les sous-ensembles d'auditeurs féminins partagent un tiers de l'information, alors que pour les voix masculines, une plus grande variabilité est observée. Un effet du même sexe est observé, puisque pour la tâche des voix féminines, la similarité moyenne entre les regroupements est plus élevée que pour les voix masculines. Les groupes d'auditeurs les moins similaires sont les experts et les experts non natifs. En outre, les naifs et les non natifs ont une approche différente de la tâche par rapport à tous les autres groupes, ce qui suggère que le sexe et la langue

maternelle jouent un rôle majeur dans la caractérisation des voix, même lorsque la langue parlée est bien connue des auditeurs non natifs.

Les résultats de l'approche Perception montrent que, dans les regroupements basés sur la phonétique, les combinaisons de regroupement les plus similaires impliquent celles obtenues à partir des quatre premiers formants pour les voix féminines et des moments spectraux pour les voix masculines. Cela suggère les rôles importants que ces mesures pourraient avoir dans la modélisation des informations partagées avec d'autres composantes. Le spectre moyen à long terme (LTAS) a une plus grande influence sur la modélisation des locuteurs masculins et les MFCC partagent également une grande quantité d'informations sur les voix masculines avec les regroupements résultant de représentations globales, c'est-à-dire les spectrogrammes ou le groupe Glob, toutes les mesures phonétiques ensemble. Les résultats du clustering basé sur les CNN montrent une plus grande redondance de l'information, indiquée par une plus petite gamme de scores de similarité entre clusterings. Pour les deux sexes, les mesures d'énergie et les ensembles de différences harmoniques-amplitudes (Hadiff) présentent les informations les plus redondantes. Les regroupements basés sur le rythme ont le plus d'influence sur la modélisation des voix des femmes, alors que pour les hommes, il s'agit du rapport harmonique sur bruit (HNR). La comparaison entre les trois principaux types de clustering, CNN, humain et phonétique, montre que le clustering des formants est le plus similaire aux réponses des auditeurs pour les voix féminines, tandis que le plus dissemblable est basé sur le LTAS. Les clusters obtenus par les non-natifs sont les plus différents des clusters CNN et phonétiques pour les tâches de voix féminines et masculines.

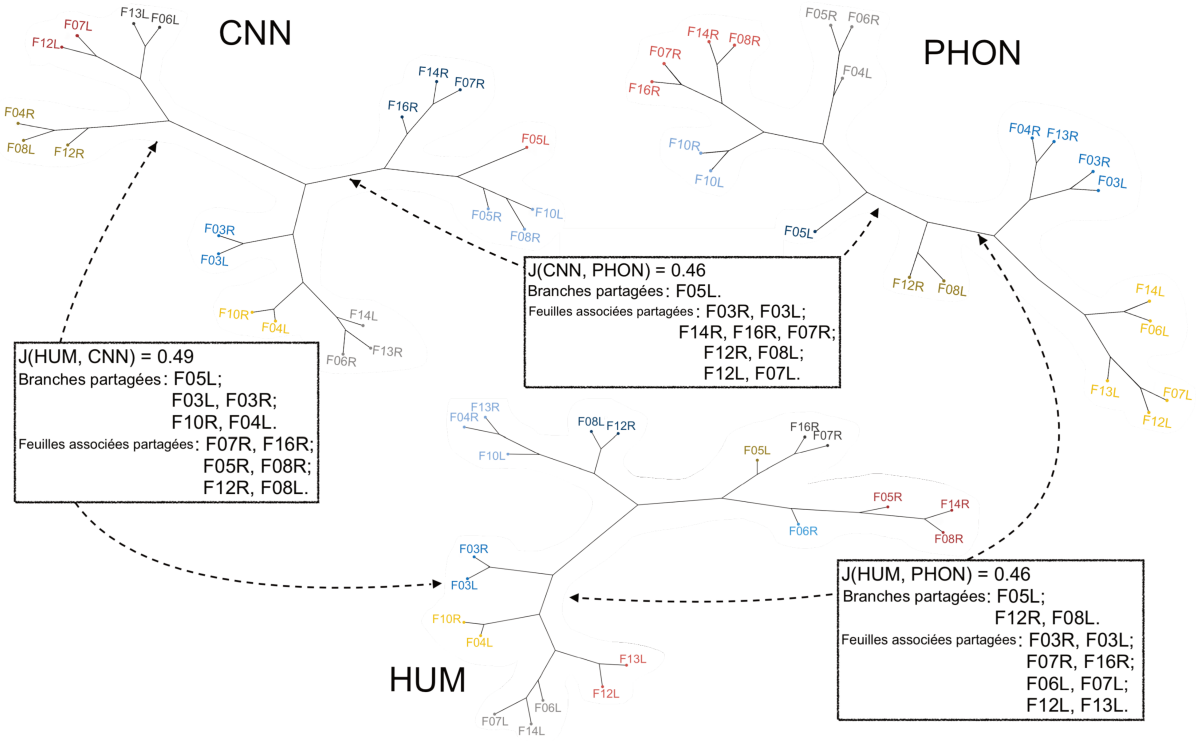

FIGURE 1 – Regroupements similaires entre auditeurs natifs naïfs dans la tâche de perception de voix féminines. CNN et PHON correspondent aux clusterings résultant des formants. Les couleurs représentent les différentes branches.

Plus précisément, les non-natifs semblent se baser davantage sur la qualité de la voix, le HNR et le MFCC pour regrouper les voix féminines. De plus, pour la caractérisation des locuteurs masculins, les clusters des auditeurs experts, comparés aux clusters CNN et phonétiques, sont similaires à ceux obtenus avec les informations harmoniques et énergétiques. Les réponses des naïfs sont parallèles à celles du clustering rythmique pour la caractérisation des voix féminines.

Dans ces résultats, nous observons que les auditeurs des deux sexes sont susceptibles d'utiliser des caractéristiques différentes pour regrouper les voix féminines et masculines. L'importance de la qualité de la voix pour la caractérisation des locuteurs masculins réapparaît dans notre analyse ultérieure, tout comme les ensembles de caractéristiques plus variables utilisés pour les locuteurs féminins. Des différences importantes dans la caractérisation perceptive des locuteurs sont également observées entre les natifs et les non-natifs français, même lorsque ces derniers ont une bonne compréhension de la langue parlée.

2.2 Phonétique

Cette deuxième approche vise à apporter une contribution à la description phonétique des interactions entre les caractéristiques des locuteurs. Nous avons fourni des résultats à partir de la parole spontanée et lue en français. Une ACP a été appliquée pour réaliser la description statistique de plusieurs mesures phonétiques mais la variabilité que les mesures décrites parviennent à fournir n'est pas toujours équilibrée face à des problèmes de classification des locuteurs. Des approches plus complexes ont permis d'augmenter la robustesse des modèles de caractéristiques des locuteurs mais la recherche de représentations capables de décrire la grande variabilité intrinsèque de la parole reste une préoccupation majeure dans la suite de nos études.

L'influence du sexe sur la majorité des mesures phonétiques est un point majeur dans toutes nos expériences. Cependant, cela ne signifie pas que les caractéristiques des locuteurs sont distribuées de manière égale pour tous les sous-ensembles de locuteurs. Nous observons que certaines caractéristiques sont plus importantes dans certains sous-ensembles mais non pertinentes pour d'autres, par exemple c'est le cas pour les informations de variabilité de la forme spectrale. De même, l'influence du support d'enregistrement et les différences entre les types de discours sont évaluées, ce qui prouve la capacité de modéliser les mêmes informations à partir de signaux moins clairs. Dans l'ensemble, nous observons une dégradation importante de toutes les mesures phonétiques lors de la comparaison des enregistrements microphoniques et téléphoniques. Cependant, l'information sur le locuteur masculin est plus cohérente entre les deux supports. Les valeurs des fréquences des formants sont conformes à la référence de la littérature en français [Georgeton et al., 2012; Gendrot and Adda-Decker, 2005]. En considérant les différences d'aires de l'espace vocalique entre les locuteurs, nous n'observons pas de résultats significatifs. La comparaison entre le discours spontané et le discours lu met en évidence une plus grande variabilité de /y/, /u/ et /o/ dans le discours spontané.

Au-delà du niveau phonémique de l'analyse, le débit de la parole n'apparaît pas significatif dans les comparaisons intra- et inter-locuteurs, la durée des pauses, et des phonèmes en début et fin de mot montrent une grande cohérence intra-locuteur. Dans l'ACP, la variance expliquée entre les locuteurs est plus élevée pour la parole lue que pour la parole spontanée,

confirmant la nature plus instable de cette dernière. Les résultats confirment que les formants sont plus caractéristiques pour les femmes que pour les hommes. De plus, pour les locuteurs féminins, les formants inférieurs et l'information f0 sont utilisés de manière plus cohérente pour caractériser les productions d'un locuteur, tandis que la qualité de la voix, c'est-à-dire les indicateurs de souffle et d'enrouement, ont un poids beaucoup plus important dans la caractérisation des locuteurs masculins. Les MFCC montrent un pouvoir statistique plus faible que les mesures phonétiques classiques. La combinaison même des deux n'influence pas les distributions des mesures phonétiques dans les PC, et n'augmente pas la variance expliquée globale. Il est intéressant de noter que les MFCC de rang inférieur sont apparus avec l'intensité et le f0 pour les locuteurs féminins, tandis que les MFCC de rang moyen suivent la distribution de l'énergie et des informations de forme spectrale de bas niveau pour les hommes. Comme nous l'avons mentionné, les descriptions statistiques ne permettent pas d'obtenir des modèles suffisamment robustes à l'aide de la LDA. La modélisation par machine à vecteurs de support montre des résultats plus prometteurs pour les séquences de parole lue, mais les résultats de classification restent très faibles pour la parole spontanée.

L'analyse et la représentation de la dynamique de la parole a été effectuée par des mesures d'intensité de type quotient ouvert et par des coefficients polynomiaux sur la f0 et les formants. Pour la parole spontanée, des résultats prometteurs sont montrés en utilisant l'approche de type quotient ouvert. Cependant, le problème d'une variance expliquée inter-locuteurs élevée non équilibrée par les scores dans les tâches de classification demeure. Le temps d'ouverture est significatif pour les locuteurs féminins et le quotient ouvert pour les hommes, ce qui indique que le soulèvement de la production vocale est plus caractéristique pour les locuteurs féminins, tandis que pour les hommes les trajectoires de fermeture ont une plus grande influence. L'utilisation de fonctions polynomiales pour recréer les trajectoires de f0 et de formants est très exigeante en termes de cohérence du contenu linguistique, de prétraitement et de traitement effectif de l'analyse. Cependant, des résultats prometteurs sont obtenus avec des différences claires entre les locuteurs féminins et masculins où f0 est la mesure la plus performante pour les premiers et F3 pour les seconds. Les coefficients polynomiaux semblent véhiculer des informations redondantes, puisqu'aucune amélioration n'est montrée par la combinaison de plusieurs paramètres.

2.3 TAL

L'idée sous-jacente de l'approche TAL utilisée est de combiner des données d'entrée interprétables issues de la littérature phonétique avec une approche moderne capable de modéliser efficacement les modulations de telles données. Pour ce faire, nous avons extrait des mesures phonétiques et les avons étudiées afin d'avoir une connaissance phonétique de base sur les locuteurs que nous utilisons dans les enquêtes de caractérisation. L'utilisation de CNN fournit une modélisation plus complexe des données qui manquait à l'approche précédente. Dans chacune des trois composantes représentées, nous effectuons des tests de sous-ensembles afin d'avoir une vision approfondie des interactions entre les informations véhiculées par les locuteurs. Cela permet une compréhension plus profonde des résultats, au-delà des simples scores de performance, qui restent utiles pour la fiabilité de la représentation étudiée. La Figure 2 montre un aperçu des représentations fournies en entrée aux CNN.

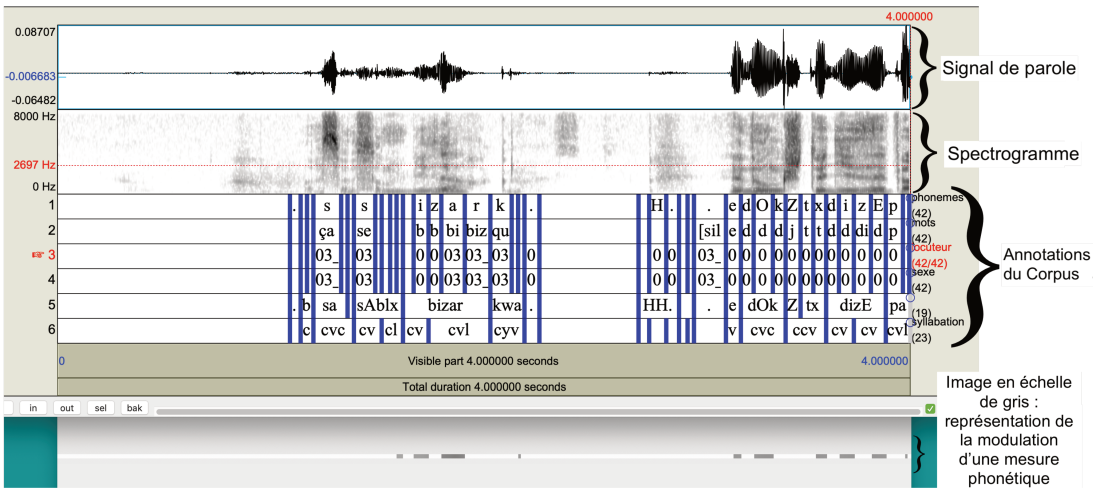

FIGURE 2 – Exemple d'un spectrogramme et d'une représentation de la modulation d'une mesure phonétique pour une séquence de 4 s du corpus NCCFr.

Dans les résultats, pour les deux sexes, les spectrogrammes sont les plus performants dans la tâche d'identification avec les F1-scores les plus élevés (moyenne harmonique de précision et rappel) et des faux positifs très faibles. Les informations sur les locuteurs ne sont pas partagées de manière égale entre les spectrogrammes et l'utilisation de toutes les mesures phonétiques, mais les MFCC combinent les résultats des deux. Cependant, les MFCC donnent de très mauvais résultats dans la tâche de généralisation, où des locuteurs inconnus par le CNN sont introduits dans la phase de test, les meilleurs résultats étant obtenus par l'utilisation de Glob.

Les résultats montrent que le groupe de mesures liées à l'HNR est le plus performant pour identifier les locutrices avec un F1-score de 0.76. Le rapport subharmonique sur harmonique (SHR) montre une redondance importante avec les autres mesures du HNR, tandis que le calcul du HNR avec la plage 0-500 Hz a la plus grande influence sur les résultats. Les locuteurs sont identifiés avec le plus de précision par Hadiff avec un score de 0.81, H1-A1 et H1-H2 ayant la plus grande influence sur les résultats. Parallèlement, la prosodie est le groupe qui obtient les scores globaux les plus bas avec une redondance élevée des informations extraites pour les locuteurs féminins et masculins. En effet, un locuteur féminin marqué est partagé dans la plupart des résultats, mais les autres femmes marquées varient constamment. La composante source et filtre ainsi que la qualité de la voix partagent des locuteurs non marqués communs. De plus, dans la tâche de vérification, pour les locuteurs féminins, la qualité vocale obtient les meilleurs résultats, suivie par les moments spectraux et le f0 avec les quatre premiers formants. Pour les locuteurs masculins, la modulation de Hadiff reste la meilleure représentation. La tâche de généralisation met en évidence une plus grande variabilité en termes de caractérisation des locuteurs masculins. Ceci suggère que l'influence des modèles de référence est plus importante pour la caractérisation des locuteurs masculins, alors que pour les locuteurs féminins, l'utilisation de locuteurs inconnus dans la phase de test n'affecte pas significativement les résultats du CNN.

3 Discussion et conclusion

Dans cette thèse, nous avons exploré la caractérisation phonétique des locuteurs dans le but de combiner les connaissances phonétiques sur les interactions des composants de la parole avec des techniques avancées de modélisation CNN. La comparaison des connaissances phonétiques et basées sur les CNN avec les résultats perceptifs fait partie de la recherche continue d'une évaluation plus poussée des modèles existants des caractéristiques du locuteur. Nos résultats confirment l'idée que, dans la description des distributions d'informations sur les locuteurs, il n'y a pas qu'une seule composante qui caractérise facilement les locuteurs. Au contraire, l'identification des interactions entre les multiples aspects qui sont importants pour la caractérisation des locuteurs est une étape fondamentale pour comprendre la distribution des informations sur les locuteurs dans une matrice de variabilité complexe. Une meilleure compréhension de la façon dont la variabilité des locuteurs est structurée est une étape importante pour décrire les variations des caractéristiques phonétiques au sein d'un groupe du locuteur [Tanner et al., 2020]. La modélisation des interactions doit faire partie de la métrique utilisée dans l'analyse plutôt que d'être considérée comme un facteur de confusion. Dans l'ensemble, les caractéristiques des locuteurs représentent une combinaison unique de caractéristiques segmentaires, suprasegmentales et paralinguistiques.

Comment représenter efficacement la matrice de variabilité associée aux caractéristiques des locuteurs a été l'une des principales questions de cette thèse. En effet, la nature variable de la parole la rend susceptible d'être représentée de multiples façons, comme nous l'avons montré tout au long des recherches en phonétique et en linguistique. L'utilisation de valeurs moyennes pour représenter f0 ou la moyenne de trois points dans l'analyse des formants sont quelques exemples courants du point de vue phonétique. Une analyse moyenne sur une description fine des variations temporelles peut suffire à intégrer la variabilité intrinsèque, alors que des représentations complexes de la dynamique de la parole peuvent aboutir à des rendus médiocres. Les voix des locuteurs ne sont pas une simple addition de variations autour d'un prototype. Les distributions des composantes que nous analysons semblent plus importantes que les valeurs qu'elles prennent. Les productions vocales sont des systèmes complexes résultant des interactions des composantes de la parole, qui sont à leur tour influencées par des facteurs provenant de sources multiples.

La variation est essentielle pour comprendre et expliquer le fonctionnement de la parole. En phonétique, il est courant de faire référence au simple fait qu'il n'y a jamais deux énoncés identiques, même s'ils sont produits par le même locuteur. Ce fait est lié à l'idée de variation intra-locuteur, qui représente le premier grand problème lors de l'étude de la caractérisation du locuteur. Surmonter la variation intra-locuteur peut représenter une tâche difficile selon ce sur quoi nous nous concentrons et quels facteurs d'influence sont impliqués. Par exemple, c'est la principale raison qui distingue la science médico-légale du langage des autres disciplines médico-légales. Pour les preuves médico-légales telles que l'ADN, les échantillons du criminel et du suspect peuvent être identiques, et le criminel peut donc être identifié. Dans la comparaison de la parole, le niveau de confiance dans les résultats reste une question importante.

Pour cela, considérer la dichotomie inné-appris pour décrire la multitude de sources de variation dans les informations que les productions vocales peuvent transmettre sur un locuteur, semble réducteur. L'utilisation d'une trichotomie physique-psychologique-sociale

semble plus appropriée. Par exemple, les différentes composantes de la parole qui peuvent être associées aux caractéristiques du locuteur ne doivent pas être considérées comme liées à un seul type d'information, mais doivent être étudiées dans une perspective multidimensionnelle. Tout au long de cette thèse, nous avons souligné l'importance de comprendre les interactions entre les différentes composantes et les facteurs d'influence afin de mieux anticiper les expressions possibles des caractéristiques des locuteurs.

Les objectifs de recherche de cette thèse consistaient à donner un large aperçu des interactions des composantes de la parole et de leurs rôles dans la caractérisation des locuteurs, afin de mieux comprendre leur contribution réelle à la caractérisation des locuteurs. La comparaison avec la perception humaine couvrait le troisième objectif de recherche afin de questionner la validité des résultats obtenus. Même si nous avons en grande partie atteint nos objectifs, leur réalisation complète nécessite des réponses qui ne seront peut-être jamais obtenues, puisque chaque nouvelle explication amène une nouvelle question. Cependant, certains résultats qui ont émergé dans cette thèse, peuvent être considérés comme importants pour notre contribution aux domaines de la caractérisation du locuteur. (i) La description de la variation intra-locuteur est fondamentale pour comprendre la variabilité inter-locuteur. Les composantes reliées à la variation des stimuli d'un seul locuteur sont cohérentes avec celles qui jouent un rôle dans la séparation des stimuli de différents locuteurs. Cela permet de créer de multiples groupes du locuteur au sein de la population étudiée qui sont caractérisés par des distributions similaires des composantes de la parole. Dans ce sens, nous observons que (ii) les caractéristiques de la source et du filtre sont plus importantes dans la description de la variation des locuteurs féminins, tandis que (iii) les caractéristiques de la qualité de la voix telles que le souffle et la raucité ont un impact plus important sur les locuteurs masculins. Ces résultats suggèrent que la caractérisation des locuteurs féminins est plus liée aux aspects linguistiques, alors que pour les locuteurs masculins il y a une part importante des aspects paralinguistiques. (iv) Les réponses perceptives confirment également ces tendances, les regroupements basés sur l'humain montrant une cohérence avec les résultats basés sur le CNN et la phonétique. En particulier, les sous-ensembles liés aux composantes mentionnées présentent de plus grandes similitudes avec les groupements perceptifs correspondant aux tâches de voix féminines et masculines. L'analyse de clustering souligne en outre (v) la cohérence des résultats CNN avec l'analyse statistique des composantes de la parole, ce qui soutient l'application ultérieure de ces méthodes pour les études phonétiques. Le dernier apport important de cette thèse concerne le rôle des MFCC par rapport aux mesures phonétiques classiques. (vi) Ils montrent une grande adaptation aux caractéristiques des locuteurs dans les données observées, concernant des aspects différents pour les locuteurs féminins et masculins, et pour les multiples groupes du locuteur présents dans notre population, plutôt que d'être représentatifs d'un trait spécifique.

Notre enquête a apporté certaines réponses aux questions concernant la caractérisation des locuteurs, même s'il existe toujours un potentiel d'amélioration et d'ajustement pour les recherches futures. En outre, ce travail aurait pu bénéficier de l'utilisation de mesures articulatoires, afin de fournir une compréhension plus approfondie de la relation entre l'anatomie des locuteurs et leurs productions orales. Cependant, comme ce travail a été mené sur des ensembles de données déjà existants où de telles mesures sont absentes, leur intégration était impossible. De même, les études de perception impliquant l'isolation de composantes individuelles auraient pu fournir des informations supplémentaires sur le poids qu'ils ont dans la caractérisation perceptive. Le nombre et la variété des auditeurs est l'autre question importante concernant la tâche de perception de cette thèse.

Les auditeurs masculins représentent un très petit groupe par rapport aux auditeurs féminins, six contre 21. Un plus grand nombre de locuteurs non-natifs aurait pu montrer des tendances plus cohérentes dans la perception des caractéristiques des locuteurs dans une perspective multilingue. C'est pourquoi nos résultats de perception doivent être considérés comme montrant des tendances qui nécessitent des investigations supplémentaires pour des confirmations significatives. Au cours de la phase exploratoire pour la création de l'expérience de perception, nous avons émis l'hypothèse de tester un cadre de perception ressemblant à l'évaluation des pathologies de la voix. Cependant, l'amélioration de la notation classique de la voix par l'incorporation d'une classe abstraite peut avoir aidé les auditeurs à classer des voix similaires sur la base de leur propre perception symbolique. Le choix de cette classe supplémentaire nécessite une analyse fine des capacités perceptives préalables des auditeurs afin d'évaluer leur fiabilité. Ainsi, l'utilisation d'une tâche de clustering sans instructions strictes sur les caractéristiques à privilégier semble être un compromis acceptable.

En suivant cette idée, l'intégration de tâches de clustering directement sur le CNN aurait pu être explorée. En particulier, cela pourrait être fait en améliorant la tâche de généralisation que nous avons présentée. Les CNN ont déjà montré leur potentiel d'application dans les problèmes de classes ouvertes [Shu et al., 2018]. L'intégration d'architectures de mémoire à long-court terme (LSTM) [3] et de mécanismes d'attention plus avancés a été partiellement explorée et pourrait représenter une réelle amélioration pour la modélisation des modulations des composantes de la parole. Dans cette perspective, des résultats encore plus comparables entre l'homme et la machine pourraient être obtenus pour trouver des techniques de modélisation plus efficaces. Concernant la tâche de généralisation, elle n'utilise qu'un ensemble partiel de toutes nos composantes sélectionnées, afin de réduire la redondance des informations que pourraient provoquer toutes les itérations des différents sous-ensembles. En perspective, l'idée d'une représentation faite par les composantes les plus performantes pour la tâche de généralisation peut améliorer l'efficacité de la caractérisation. De plus, l'absence de cohérence entre entraînement et test, comme le fait de réaliser l'entraînement sur une composante complète et de ne tester qu'un sous-ensemble, pourrait représenter un moyen de mieux comprendre le poids de chaque mesure phonétique.

Cette thèse a exploré la caractérisation phonétique du locuteur dans le but de combiner les connaissances phonétiques sur les interactions des composantes de la parole avec des techniques de modélisation avancées par CNN, afin de créer une analyse mutuellement bénéfique. La comparaison des connaissances phonétiques et automatiques avec les résultats perceptifs fait partie de l'ambition d'étendre les modèles existants des caractéristiques du locuteur par des caractéristiques supplémentaires du monde environnant que nous percevons comme humain. Notre recherche a contribué aux domaines de la caractérisation du locuteur et a confirmé un certain nombre de défis auxquels les chercheurs sont confrontés dans leurs efforts pour faire avancer ces disciplines. En abordant les principaux objectifs de cette thèse, des résultats supplémentaires ont été présentés. Nous espérons qu'ils encouragent la discussion menant à la solution des problèmes impliqués dans la description des caractéristiques du locuteur.

Ainsi, nous sommes confrontés à des défis permanents dans la description de la variation des caractéristiques de la parole et du locuteur. Dans cette thèse, nous avons répondu à certaines questions et fourni des résultats sur la variation entre locuteurs en français, en

3. Long-Short Term Memory

soulignant les tendances individuelles qui sont portées par les interactions du groupe et en espérant encourager d'autres recherches suivant des approches multidisciplinaires.

Bibliographie

Ajili, M., Bonastre, J.-F., Kheder, W. B., Rossato, S., and Kahn, J. (2016). Comparaison des voix dans le cadre judiciaire : influence du contenu phonétique. *JEP (Journées d'Etudes sur la Parole)*.

Ardoint, M. and Lorenzi, C. (2009). Effects of lowpass and highpass filtering on the intelligibility of speech based on temporal fine structure or envelope cues. *Hearing Research*.

Arvaniti, A. (2013). The role of rhythm class, speaking rate and f0 in language discrimination. *Laboratory Phonology*.

Barlow, M. G. and Wagner, M. (1988). Prosody as a basis for determining speaker characteristics. *Proceedings of the Second Australian International Conference on Speech Science and Technology*.

Bee, M. and Gerhardt, H. (2002). Individual voice recognition in a territorial frog (rana catesbeiana). *Proceedings of the Royal Society of London B : Biological Science*, 269.

Belin, P. and Grosbras, M. (2010). Before speech : cerebral voice processing in infants. *Neuron*, 65.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, 5 :9/10 :341–345.

Bonastre, J.-F., Bimbot, F., Boë, L.-J., Campbell, J., Reynolds, D., and Magrin-Chagnolleau, I. (2003). Person authentication by voice : A need for caution. *Proceedings of EUROSPEECH*.

Boë, L.-J., Contini, M., and Rakotofiringa, H. (1975). Etude statistique de la fréquence laryngienne. *Phonetica*, 32 :1–23.

Burke, E. and Murphy, C. (2007). How female barking tree frogs, hyla gratiosa, use multiple call characteristics to select a mate. *Animal Behaviour*, 74.

Cangemi, F. (2009). Phonetic detail in intonation contour dynamics. *AISV (Associazione Italiana di Scienze della Voce)*.

Chanclu, A., Georgeton, L., Fredouille, C., and Bonastre, J.-F. (2020). Ptsvox : une base de données pour la comparaison de voix dans le cadre judiciaire (ptsvox : a speech database for forensic voice comparison). *JEP-TALN-RECITAL (Conférence conjointe Journées d'Etudes sur la Parole - Conférence sur le Traitement Automatique des Langues Naturelles - Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues)*.

Culling, J. and Darwin, C. (1993). The role of timbre in the segregation of simultaneous voices with intersecting f0 contours. *Perception Psychophysics*, 54 :303–309.

Dediu, D., Janssen, R., and Moisik, S. (2017). Language is not isolated from its wider environment : Vocal tract influences on the evolution of speech and language. *Language Communication*, 54 :9–20.

Doddington, G., Liggett, W., Martin, A., Przybocki, M., and Reynolds, D. (1998). Sheep, goats, lambs and wolves : A statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation. *Proceedings of ICSLP (International Conference on Speech and Language Processing).*

Eriksson, A. and Wretling, P. (1997). How flexible is human voice ? - a case study of mimicry. *Proceedings of EUROSPEECH.*

Furuyama, T., Kobayasi, K., and Riquimaroux, H. (2016). Role of vocal tract characteristics in individual discrimination by japanese macaques (macaca fuscata). *Nature Scientific reports*, 6.

Gamal, E. E. (2015). *Speaker identification based on temporal parameters.* PhD thesis, Phonetics and Linguistics Department of the University of Alexandria.

Gendrot, C. and Adda-Decker, M. (2005). Impact of duration on f1/f2 formant values of oral vowels : an automatic analysis of large broadcast news corpora in french and german. *Proceedings of INTERSPEECH.*

Gendrot, C., Ferragne, E., and Pellegrini, T. (2020). Information segmentale pour la caractérisation phonétique du locuteur : variabilité inter- et intra- locuteurs. *JEP-TALN-RECITAL (Conférence conjointe Journées d'Etudes sur la Parole - Conférence sur le Traitement Automatique des Langues Naturelles - Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues).*

Georgeton, L., Paillereau, N., Landron, S., and Kamiyama, T. (2012). Analyse form antique des voyelles orales du français en contexte isolé : à la recherche d'une référence pour les apprenants du fle. *JEP-TALN-RECITAL (Conférence conjointe Journées d'Etudes sur la Parole - Conférence sur le Traitement Automatique des Langues Naturelles - Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues)*, pages 145–152.

Grossmann, T., Oberecker, R., Koch, S., and Friederici (2010). The developmental origins of voice processing in the human brain. *Neuron*, 65.

He, L. and Dellwo, V. (2017). Between-speaker variability in temporal organizations of intensity contours. *Journal of Acoustical Society of America*, 141 :488–494.

Hollien, H., Majewski, W., and Doherty, E. (1982). Perceptual identification of voices under normal, stress and disguise speaker conditions. *Journal of Phonetics*, 10.

Hsu, A., Woolley, S., Fremouw, T., and Theunissen, F. (2004). Modulation power and phase spectrum of natural sounds enhance neural encoding performed by single auditory neurons. *The Journal of Neuroscience*, 24 :9201–9211.

Hudson, T., de Jong, G., McDougall, K., Harrison, P., and Nolan, F. (2007). F0 statistics for 100 young male speakers of standard southern british english. *Proceedings of the ICPhS (International Congress of Phonetic Sciences)*, pages 1809–1812.

Hughes, V., Cardoso, A., Harrison, P., Foulkes, P., French, P., and Gully, A. (2019). Forensic voice comparison using long-term acoustic measures of laryngeal voice quality. *Proceedings of the ICPhS (International Congress of Phonetic Sciences)*.

Kahn, J. (2011). *Parole de locuteur : performance et confiance en identification biométrique vocale*. PhD thesis, ED 536.

Keating, P. and Kreiman, J. (2016). Acoustic similarity among female voices. *The Journal of Acoustical Society of America*, 140.

Keating, P., Kreiman, J., and Alwan, A. (2019). A new speech database for within- and between-speaker variability. *Proceedings of the ICPhS (International Congress of Phonetic Sciences)*.

Keating, P., Kreiman, J., and Vasselinova, N. (2017). Acoustic similarities among voices. part 2 : Male speakers. *The Journal of Acoustical Society of America*, 142.

Kolly, M.-J., Leemann, A., de Mareüil, P. B., and Dellwo, V. (2015). Speaker-idiosyncrasy in pausing behavior : evidence from a cross-linguistic study. *Proceedings of the ICPhS (International Congress of Phonetic Sciences)*.

Kreiman, J. (1997). Listening to voices : theory and practice in voice perception research. *Johnson K. Mullenix J. Talker Variability in Speech Research. Academic Press*.

Kreiman, J. and Gerratt, B. (2012). Perceptual interaction of the harmonic source and noise in voice. *Journal of Acoustical Society of America*, 131.

Kreiman, J. and Shue, Y.-L. (2010). Variability in the relationships among voice quality, harmonic amplitudes, open quotient, and glottal area waveform shape in sustained phonation. *Journal of Acoustical Society of America*, 132 :2625–2632.

Kreiman, J. and Stidtis, D. (2011). Voices and listeners : toward a model of voice perception. *Acoustics Today*, 7 :7–15.

Künzel, H. (2013). Some general phonetic and forensic aspects of speaking tempo. *International Journal of Speech Language and the Law*, 4 :48–83.

Latinus, M. and Belin, P. (2011). Human voice perception. *Current Biology*, 21.

Lee, Y., Keating, P., and Kreiman, J. (2019). Acoustic voice variation within and between speakers. *The Journal of Acoustical Society of America*, 146 :1569–1579.

Leemann, A. and Kolly, M.-J. (2015). Speaker-invariant suprasegmental temporal features in normal and disguised speech. *Speech Communication*, 75 :97–122.

Lindh, J. (2006). Preliminary descriptive f0-statistics for young male speakers. *Working Papers*, 52.

Mary, L. and Yegnanarayana, B. (2008). Extraction and representation of prosodic features for language and speaker recognition. *Speech Communication*, 50 :782–796.

McDougall, K. and Nolan, F. (2007). Discrimination of speakers using the formant dynamics of /u/ in british english. *Proceedings of the ICPhS (International Congress of Phonetic Sciences)*, pages 1825–1828.

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

Niebuhr, O. and Skarnitzl, R. (2019). Measuring a speaker's acoustic correlates of pitch - but which ? a constrastive analysis based on perceived speaker charisma. *Proceedings of the ICPhS (International Congress of Phonetic Sciences)*, pages 1774–1778.

Nolan, F. (2001). Speaker identification evidence : Its forms, limitations, and roles. *Proceedings of the conference Law and Language : Prospect and Retrospect*.

Ohala, J. (1983). Cross-language use of pitch : an ethological view. *Phonetica*, 40 :1–18.

Ohala, J. J. (1994). The frequency codes underlies the sound symbolic use of voice pitch. *Sound symbolism*, pages 325–347.

Ouyang, I. and Kaiser, E. (2015). Individual differences in the prosodic encoding of informativity. *Individual differences in speech production and perception*.

Ramus, F. and Mehler, J. (1998). Language identification with suprasegmental cues : a study based on speech resynthesis. *Journal of Acoustical Society of America*, 105 :512–521.

Rose, P. (2002). *Forensic Speaker Identification*. Taylor Francis Forensic Science Series.

Rose, P. and Wang, X. (2016). Cantonese forensic voice comparison with higher-level features : likelihood ratio-based validation using f-pattern and tonal f0 trajectories over a disyllabic hexaphone. *Proceedings of Odyssey*, pages 326–333.

san Segundo, E. and Yang, J. (2019). Formant dynamics of spanish vocalic sequences in related speakers : A forensic-voice-comparison investigation. *Journal of Phonetics*, 75.

Shu, L., Xu, H., and Liu, B. (2018). Unseen class discovery in open-world classification. *arXiv e-prints*.

Shue, Y.-L., Keating, P., Vicenik, C., and Yu, K. (2011). Voicesauce : A program for voice analysis. *Proceedings of the ICPhS (International Congress of Phonetic Sciences)*, pages 1846–1849.

Sorin, C. (1981). Functions, roles and treatments of intensity in speech. *Journal of Phonetics*, 9 :359–374.

Stoll, L. and Doddington, G. (2010). Hunting for wolves in speaker recognition. *Proceedings of Odyssey*, pages 159–164.

Tanner, J., Sonderegger, M., and Stuart-Smith, J. (2020). Structured speaker variability in japanese stops : relationships within versus across cues to stop voicing. *Journal of the Acoustical Society of America*, 148.

Torreira, F., Adda-Decker, M., and Ernestus, M. (2010). The nijmegen corpus of casual french. *Speech Communication*, 52 :201–212.

van Lancker, D., Kreiman, J., and Emmorey, K. (1985). Familiar voice recognition : patterns and parameters. part i : recognition of backward voices. *Journal of Phonetics*, 13.

Vaňková, J. and Skarnitzl, R. (2014). Within- and between-speaker variability of parameters expressing short-term voice quality. *Speech Prosody*, pages 1081—-1085.

Zuo, D. and Mok, P. (2012). Formant dynamics of bilingual identical twins in non-contemporaneous speech. *Proceedings of SST (Austrasian Speech Science and Technology)*.

Speech components in phonetic characterisation of speakers

a study on complementarity and redundancy of conveyed information

Abstract

The decomposition of the speech signal into phonetically meaningful units allows the analysis of between- and within- speaker variations. These are components associated with characteristics whose nature relates to the physical, psychological and social aspects of a speaker. In this thesis, we compare perceptual characterisation results with a phonetic analysis and advanced modelling techniques through Convolutional Neural Networks (CNN).
Clusterings' analysis shows that the perceptual results are coherent with those obtained by the CNN and phonetic approaches, which supports the application of these methods in Phonetics. Our results highlight that spectrograms are the most accurate speech representation for speaker identification (96 % correct answers on average). Higher formants and harmonics are more important in the characterisation of female voices. Whereas, voice quality characteristics, such as breathiness and hoarseness, play a major role in the characterisation of male speakers. The comparison between Mel Frequency Cepstral Coefficients (MFCC) and classical phonetic measurements is also examined. The MFCC are mainly linked to intensity and f0 in the characterisation of female speakers, while to the distributions of energy and low level spectral shape for male speakers.
Our findings confirm the importance of describing the within-speaker variation for a more complete understanding of between-speakers differences.
Keywords : speaker, characteristics, components, CNN, spontaneous, clustering, informedness, comparison

Les composantes de la parole dans la caractérisation phonétique du locuteur

étude sur la complémentarité et la redondance véhiculées des informations

Résumé

La décomposition du signal vocal en unités phonétiquement significatives permet d'analyser les variations inter- et intra- locuteur. Ces unités sont des composantes associées à des caractéristiques dont la nature est liée aux aspects physiques, psychologiques et sociaux d'un locuteur. Dans cette thèse, nous comparons une caractérisation perceptive, une analyse phonétique et des techniques de modélisation avancées par des réseaux de neurones à convolution (CNN).
L'analyse des clusterings montre que les résultats perceptifs sont cohérents avec ceux obtenus par les approches CNN et phonétique, ce qui soutient leurs applications en phonétique. Nos résultats mettent en évidence que les spectrogrammes sont la représentation de la parole la plus précise pour l'identification des locuteurs (96 % de bonnes réponses en moyenne). Les formants et des harmoniques plus élevés sont plus importants dans la caractérisation des voix féminines. En revanche, les caractéristiques de la qualité de la voix, telles que le souffle et la raucité, jouent un rôle majeur dans la caractérisation des voix masculines. Le lien entre les coefficients cepstraux à fréquence Mel (MFCC) et les mesures phonétiques classiques est également examiné. Les MFCC sont principalement liés à l'intensité et à f0 dans la caractérisation des voix féminines, tandis qu'aux distributions d'énergie et à la forme spectrale de bas niveau pour celle des voix masculines.
Nos résultats confirment l'importance de la description de la variation intra-locuteur pour une compréhension plus complète des différences entre locuteurs.
Mots-clés : locuteur, caractéristiques, composantes, CNN, spontané, clustering, informativité, comparaison

Université Sorbonne Nouvelle
Ecole Doctorale 622 - Sciences du langage
4, rue des Irlandais 75005 Paris