# Du groupe à l'individu, du corpus à l'expérimentation, du spectrogramme au deep learning pour la phonétique

Emmanuel Ferragne

Aix*Marseille université

**ED 356 : Cognition, langage, éducation**

# Du groupe à l'individu, du corpus à l'expérimentation, du spectrogramme au deep learning pour la phonétique

## DOCUMENT DE SYNTHÈSE

présenté et soutenu publiquement le 25 juin 2021

pour l'obtention d'une

**Habilitation à Diriger des Recherches d'Aix-Marseille Université**

**(Études Anglophones)**

par

Emmanuel Ferragne

**Composition du jury**

| | |
|---|---|
| *Présidente :* | Mme Christine MEUNIER, Aix-Marseille Université, CNRS |
| *Rapporteurs :* | Mme Sylvie HANOTE, Université de Poitiers |
| | M. Noël NGUYEN, Aix-Marseille Université |
| | Mme Anne PRZEWOZNY, Université Toulouse Jean-Jaurès |
| *Examinateurs :* | Mme Sophie HERMENT, Aix-Marseille Université (garante) |
| | M. François PELLEGRINO, Université Lyon 2, CNRS |

**Laboratoire Parole et Langage - UMR 7309**

# Remerciements

Je tiens en premier lieu à exprimer toute ma gratitude à Sophie Herment pour avoir dirigé ce travail et m'avoir constamment encouragé.

Il m'est agréable de pouvoir partager ce document avec les membres d'un jury aux compétences aussi variées ; je remercie par avance, pour leur travail d'évaluation et leurs conseils avisés, Sylvie Hanote, Christine Meunier, Noël Nguyen, François Pellegrino et Anne Przewozny.

À toute l'équipe de relecture de dernière minute, Anne Guyot-Talbot, Hannah King, Sylvain Navarro et Maud Pélissier, un grand merci !

Cette synthèse marque symboliquement la douzaine d'années que j'ai passées en tant que maître de conférences, et la vingtaine d'années pendant lesquelles j'ai consacré la majorité de mon temps à faire de la recherche. Il va donc de soi qu'un nombre important de personnes m'ont apporté le soutien et l'inspiration nécessaires. Je n'ose cependant pas ébaucher la moindre liste de peur d'omettre par inadvertance certains personnages pourtant incontournables. Et puis, un nom dans une liste, cela manque de contexte ! J'ai donc pris le parti de mentionner dans le corps du texte le nom de certains collègues, parfois devenus amis, à qui je dois beaucoup.

Ces collègues, ces étudiants ou étudiantes, et ces relations personnelles et membres de ma famille, je ne peux bien évidemment pas les énumérer ; mais à défaut, j'ai une pensée reconnaissante pour chacun et chacune au moment d'écrire ces lignes.

*Lou grond mourió*
*E toutjour n'aprenió*
— Proverbe lozérien[1]

1. De l'occitan : « le grand-père mourait et il en apprenait toujours ».

# Table des matières

# Table des figures

# Liste des tableaux

# 1

# Introduction

## Sommaire



## 1.1 Objectifs

J'aborde la rédaction de ce document de synthèse comme une excellente opportunité de faire le point sur mon parcours scientifique. Quoique a priori circonspect quant aux bénéfices éventuels que j'allais tirer d'un tel exercice, je dois bien reconnaître aujourd'hui qu'il n'était pas inutile. Cela peut sembler paradoxal, mais malgré le temps démesuré que nous consacrons à écrire *à propos* de notre recherche — dans le cadre des soumissions de projets, des évaluations d'équipes ou des divers rapports justifiant nos dépenses, force est de constater qu'il est bien rare de pouvoir s'accorder un temps dédié à une prise de recul structurée et sereine. Je propose donc ici ma lecture de la douzaine d'années passées à exercer le métier de maître de conférences. Mon seul regret, c'est d'avoir tardé à finaliser la rédaction de cette synthèse ; cela m'a en effet conduit à ré-écrire une proportion importante du document.

## 1.2 Trajectoire

Afin de bien cerner les orientations épistémologiques dans lesquelles s'inscrivent mes travaux, il faut revenir un instant sur mon parcours. Certains points mentionnés ici seront développés au Chapitre 2. C'est lors de ma troisième année de Licence à l'Université Lyon 2 que je découvre la phonétique et me prends de passion pour cette discipline. À travers la lecture de O'Connor (1991), j'entrevois à quel point il va m'être possible de combiner mon intérêt prononcé pour la musique et le chant avec mon parcours professionnel. Par ailleurs, en tant que locuteur natif de français méridional évoluant désormais dans un milieu non méridional, et ayant grandi dans un environnement où les anciens parlaient encore couramment l'occitan gévaudanais, j'ai été très tôt sensible à la dimension sociologique de la langue et de la prononciation.

Mes débuts en phonétique doivent beaucoup aux Professeurs Pierre Arnaud, Claude Boisson et Henry Daniels. Grâce à leurs conseils, leur confiance et leur bienveillance, mes premiers pas dans le domaine furent très épanouissants. Après avoir observé une camarade de classe qui constituait son corpus en surlignant au feutre dans des magazines des occurrences du phénomène linguistique qui l'intéressait, ma conviction était faite : je ne resterais pas en marge de la révolution numérique, et je m'investirais dans l'apprentissage de technologies plus avancées que le repérage manuel sur un corpus papier. Mon goût marqué pour la technologie constitue aujourd'hui encore un moteur puissant. Il me semble d'ailleurs peu souhaitable de trop vouloir distinguer un objet d'étude des méthodes employées pour l'analyser. Ces dernières contribuent en effet grandement à le rendre intelligible, et conditionnent dans une large mesure nos résultats et nos conclusions (voir le Chapitre 3).

Afin de commencer mon Master dans les meilleures conditions, j'économise pour acheter mon premier ordinateur. Ceci n'est pas anodin : l'informatique m'est totalement étrangère. Le long apprentissage en autodidacte qui a suivi représentait un défi majeur. Vient ensuite une étape dont les conséquences se sont révélées essentielles dans ma vie de chercheur : le choix d'un logiciel pour l'analyse du signal acoustique. Mon choix s'est fixé sur un logiciel à l'ergonomie rebutante de prime abord, mais qui, grâce à son langage de script, promettait de satisfaire mon aspiration à traiter automatiquement de grandes quantités de données.

J'ai donc décidé d'investir dans l'apprentissage du logiciel `Praat`, à une époque où il n'était pas flagrant que `Praat` deviendrait plus tard l'outil incontournable du phonéticien. Les semaines passées, à force de tâtonnements, à apprendre seul le langage de script

ont vite été rentabilisées. Vingt ans plus tard, c'est une compétence qui est encore très valorisée. Si aujourd'hui un nombre croissant de chercheurs savent coder avec `Praat`, nous étions bien moins nombreux au début des années 2000. Cette situation m'a donné l'opportunité de dispenser plusieurs formations et d'initier de nombreux collègues. En particulier, lors de mon séjour de 6 mois en tant que *visiting scholar* à l'Université de Cambridge, j'ai formé les chercheurs du *Phonetics Laboratory* qui, à cette occasion, ont abandonné leur ancien système pour migrer définitivement vers `Praat`.

C'est pendant mes deux années de Master, sous la direction des Professeurs Boisson et Daniels, que je fais mes premières armes en phonétique. Pendant la première année, je me consacre à l'analyse des formants des monophtongues anglaises de deux étudiantes préparant l'agrégation. L'année suivante, j'enregistre deux anglophones, et tente de déterminer, en m'appuyant sur des paramètres objectifs tels que l'intensité, f0 et la durée, combien de niveaux d'accent lexical il est possible de faire émerger empiriquement en anglais. Déjà, ma préférence pour les corpus *ad hoc*, constitués spécialement pour répondre à une question précise, se manifeste. Corrélativement, j'acquiers peu à peu la « fibre expérimentale » et la conviction que l'observation n'est qu'une étape préliminaire à la mise en place d'un protocole contrôlé.

Lors de ma première année de Master, un cours singulier pour un cursus littéraire attire mon attention : le cours de statistiques du Professeur Philippe Thoiron. C'est, là encore, un jalon fondamental de mon parcours. J'y apprends les tests d'inférence statistique les plus courants, et je consacre une partie de mon temps libre à l'étude de tests plus avancés.

Curieusement, mes premiers articles de conférence (Dumas et Ferragne, 2001, 2003) ne portent pas sur mes propres thématiques de recherche. En effet, c'est en quelque sorte en « mercenaire » que j'interviens pour apporter mes connaissances en phonétique et en traitement de données. Cette façon particulière de commencer une carrière a été prémonitoire car j'ai été très souvent amené par la suite à remplir cette fonction de référent dans le domaine du traitement des données auprès des collègues et des étudiants. Et là où d'autres creusent patiemment un sillon déterminé par leur sujet de thèse, j'ai au contraire une forte propension à explorer sans cesse des domaines qui me sont inconnus.

J'arrive donc en 2003 au seuil de la thèse avec un bagage honorable en phonétique acoustique et en traitement automatique de données, et de bonnes notions de statistiques. Sur les conseils de Henry Daniels, je prends contact avec les chercheurs du laboratoire Dynamique Du Langage (DDL). L'euphorie de l'arrivée des sciences cognitives en France est encore bien palpable. La pluridisciplinarité propre à ce genre de contexte m'interpelle dès

ma première visite à DDL. L'approche holistique du langage mise en avant dans ce paradigme est bien éloignée de la linguistique traditionnelle, en particulier celle pratiquée dans les départements d'anglais, et je suis instantanément séduit. C'est pourtant sans véritablement tenir compte de ce contexte qu'un projet de thèse est élaboré. Après l'obtention de l'allocation doctorale, j'entame donc mes travaux sur la phonétique des variétés de l'anglais sous la direction de Claude Boisson.

C'est à DDL que je fais une autre rencontre qui sera déterminante. François Pellegrino va en effet diriger mes travaux au quotidien. Docteur en informatique, spécialiste de l'identification automatique des langues, il sera bien plus que la caution « sciences dures » qui me permettra de dépasser mes limites. Son style de direction favorise ma créativité. Sa fiabilité, sa constance et sa bienveillance entretiennent ma détermination. Les premières lignes du manuscrit, dans un style qui m'avait pourtant valu d'excellentes notes dans mon cursus de langues, sont rapidement censurées : trop long, trop sinueux, trop littéraire ! J'obtempère, je retravaille ma prose sans cesse en visant la logique et la concision. L'amélioration est flagrante : c'est un tournant majeur dans ma formation.

Je dois également en grande partie à François la sérénité qui caractérise mon approche des méthodes quantitatives. Il n'est en effet pas rare que ces dernières inspirent aux chercheurs en Sciences Humaines et Sociales (SHS) soit une crainte injustifiée, soit une fascination excessive. Je n'ai pas fait exception à la règle en cédant, au début de mon parcours, à la seconde option. Néanmoins, François ne s'est jamais laissé impressionner : avec lui, utiliser avec succès le dernier algorithme à la mode était tout à fait trivial. Dans ces circonstances, l'orgueil laisse rapidement place à l'humilité. Cette désacralisation salvatrice du quantitatif, héritée de mes années de thèse, me procure aujourd'hui un confort certain. Si je consacre encore beaucoup de temps à la recherche et à la mise au point de nouvelles techniques, c'est toujours au service de mon objet d'étude. Mes lacunes en la matière, je les avoue sans honte ; mes connaissances, je les partage sans arrogance. Je reviens sur ces questions dans la Section 2.4.

Après deux ans de réclusion totale pendant mon Master, je bénéficie désormais à DDL d'un environnement de travail collectif très favorable. J'y apprends non seulement la recherche mais également les codes du monde de la recherche : je réalise que nos activités quotidiennes ne s'effectuent pas hors contexte, et que ce contexte conditionne le choix de nos objets d'étude, les méthodes que nous utilisons, et souvent même, nos résultats.

Le corpus *Accents of the British Isles* (ABI), d'où j'ai tiré l'essentiel de mes données de thèse, comporte des enregistrements de 284 locuteurs de 14 accents différents sur les

Îles Britanniques. Ce corpus est constitué de listes de mots à structure /hVd/, ainsi que d'un passage lu. Il présente l'avantage d'être volumineux pour l'époque, mais l'absence de métadonnées précises sur les locuteurs enregistrés entraîne quelques frustrations. Ces frustrations me conduiront entre autres à privilégier une approche bien plus expérimentale dans les travaux qui ont suivi la thèse, et à lancer le développement du logiciel `ROCme!` (voir Section 3.3.1), qui permet de collecter et de stocker des données orales et les métadonnées correspondantes de façon rationnelle et dématérialisée.

La période après la thèse fut très intense : il s'agissait de valoriser le travail effectué à travers des publications, de préparer la recherche future avec des demandes de financement, de continuer de me former à de nouveaux outils et de trouver un travail. J'ai renoncé à partir en post-doctorat : à 31 ans, le besoin de stabilité professionnelle a été plus fort. Au titre de la valorisation de mes travaux de thèse, l'année 2010 est marquée par la publication de deux articles ; l'un dans le *Journal of the International Phonetic Association* (JIPA), l'autre dans *Journal of Phonetics*. Il est intéressant de noter que, à l'heure où j'écris ces lignes, le premier a été cité plus de cent fois et le second, à peine 20 fois[2]. Ces deux articles racontent pourtant une histoire similaire, quoique vue à travers le prisme de deux traditions épistémologiques différentes.

Après ma thèse, qui reçoit le prix de l'Association Francophone de la Communication Parlée (AFCP), j'opère délibérément un changement de paradigme. Condamné à l'observation, de par la nature de mon corpus, je souhaite dorénavant manipuler moi-même certains facteurs « toutes choses égales par ailleurs » ; je m'oriente donc vers des méthodes expérimentales. Et afin de compléter ma panoplie méthodologique, je m'intéresse aux outils employés en sciences cognitives et neurosciences ; en particulier, les expériences de perception auditive et l'électroencéphalographie. Je suis aidé dans cette démarche par Nathalie Bedoin, avec qui j'entame une longue collaboration inaugurée par un article, publié dans *Brain and Language* en 2010, qui rapporte une étude d'écoute dichotique sur les consonnes du français. Cette ouverture me permet également de dispenser un enseignement en sciences cognitives, et surtout, d'encadrer des mémoires de Master dans cette discipline à l'Université Lyon 2 et, plus tard, dans le cadre du Cogmaster[3]. L'expérience est très formatrice car je découvre un niveau d'exigence envers les étudiants particulièrement élevé. J'adopte cette culture dans ma pratique de l'encadrement de mémoire en Études Anglophones.

Je suis nommé maître de conférences à l'UFR d'Études Anglophones de l'Université

2. D'après Google Scholar interrogé le 8 avril 2021.
3. https://cogmaster.ens.psl.eu/fr.

Paris Diderot en septembre 2009. Toujours passionné par la recherche, et héritier d'une culture d'UMR, je me lance dans la rédaction de projets en vue d'obtenir des financements. Mes efforts ne tardent pas à porter leurs fruits : dès 2010, je décroche un Projet Exploratoire Premier Soutien (PEPS) du CNRS et une subvention de recherche de la Fondation Fyssen. Cette enveloppe de 30 000 euros me permet notamment de mettre en place mes premières expériences d'électroencéphalographie, dont le but est de caractériser le statut de contrastes phonologiques marginaux. Je dois cette idée en particulier aux travaux de Kathleen Currie Hall. L'année suivante, j'obtiens un financement ANR JCJC (125 000 euros), qui permettra d'explorer plus avant l'hypothèse de la phonémicité gradiente. Je deviens alors membre junior de l'Institut Universitaire de France (IUF) en 2012.

Ces divers projets, que je résume au Chapitre 4, s'articulent autour de la recherche des corrélats acoustiques et électrophysiologiques des contrastes marginaux, i.e. de différences phoniques « inclassables » qui se comportent tantôt comme des phonèmes, tantôt comme des allophones. La question est audacieuse, mais les possibilités d'y répondre se révèlent limitées : en effet, ces projets portent majoritairement sur l'anglais du fait de mon rattachement aux Études Anglophones, et l'accès à des anglophones est restreint quand on travaille en France. Certes, comme j'ai eu l'occasion de le faire, il est encore partiellement possible de procéder à des enregistrements sonores et des tests de perception à l'étranger. Mais l'accès à des équipements plus lourds est très encadré, au point de devenir dissuasif. C'est donc sous le poids de la contrainte de l'accessibilité à la population que j'étudie que je me tourne vers l'acquisition de l'anglais L2. Il faut y voir un prétexte me permettant de continuer mon travail de phonéticien (avec une coloration « sciences cognitives ») dans de meilleures conditions. Après ce tournant, je deviens porteur principal d'un projet dans le cadre de l'Idex Université Sorbonne Paris Cité (USPC) qui aborde précisément ces thématiques.

Pour ma thèse, j'avais un temps envisagé de travailler sur les aspects phonétiques de la comparaison de voix dans le domaine judiciaire. Mes discussions avec certains des acteurs principaux de cette discipline à l'époque m'avaient dissuadé de m'engager dans cette voie. Je n'ai pourtant jamais totalement abandonné l'idée, et j'ai milité en m'engageant au sein de l'AFCP, pour que ce domaine soit mis en avant. L'opportunité a fini par se présenter quinze ans plus tard : je suis aujourd'hui membre du projet ANR VoxCrim, qui réunit les principaux laboratoires, dont ceux de la Police et de la Gendarmerie, autour de cette thématique.

C'est à l'occasion de ce récent projet, et en partie grâce à mon accueil en délégation CNRS au Laboratoire de Phonétique et Phonologie (LPP) de l'Université Sorbonne Nouvelle que je décide d'investir dans l'apprentissage de certaines techniques de *deep learning*, encouragé par la confiance que me témoignent les collègues du LPP. Je découvre enfin des techniques qui dépassent les clivages entre scientifiques et littéraires, quantitatif et qualitatif, expérience et observation, etc. C'est un saut épistémologique à venir certain pour de nombreuses disciplines. Et c'est un aboutissement incontournable dans mon parcours, qui vient sonner l'heure de rédiger le présent document de synthèse.

## 1.3 Notes sur la forme du document

Sur le plan terminologique, j'ai été contraint d'opérer certains choix. Quoique généralement peu enclin à employer des anglicismes lorsque l'équivalent strict français existe (je n'écris jamais « digital » ou « perceptuel », mais bien « numérique » et « perceptif »), j'ai fait le choix de garder certains acronymes en anglais quand bien même le terme en question aurait été introduit en français dans ce document. À titre d'exemple, j'emploie l'acronyme GPU (*Graphical Processing Unit*) pour désigner le processeur de la carte graphique d'un ordinateur. De la même manière, je parle plus volontiers de *deep learning* que d'apprentissage profond. Ces préférences peuvent paraître arbitraires mais elles reflètent en réalité la terminologie utilisée au quotidien dans mes collaborations et qui, par conséquent, me paraît plus naturelle. Je sollicite par avance l'indulgence des lecteurs sur ce point.

Le style est souvent autobiographique ; le ton, parfois légèrement militant. Il s'agit d'un choix éditorial délibéré et mûrement réfléchi. Là où ma thèse était un exercice scientifique pour l'essentiel dépourvu de prise de position forte, le présent document allie l'exercice scientifique à la prise de recul épistémologique, d'où la nécessité d'une coloration plus « personnelle ».

Les auteurs dont je m'inspire sont dûment cités ; néanmoins, pour les aspects plus techniques, il m'est évidemment impossible de citer explicitement la multitude de ressources que j'ai consultées, qu'il s'agisse de forums, blogs, notices techniques ou de documentations de logiciels. Par exemple, la documentation de `Matlab` — mon environnement de travail de prédilection depuis une quinzaine d'années — a fortement contribué à ma formation, mais je ne peux décemment pas y faire référence à chaque fois qu'une technique « standard » de traitement de données est employée.

Par souci de lisibilité, j'ai parfois eu recours à des URL réduites. Afin d'anticiper d'éventuels problèmes de compatibilité ou de sécurité, l'Annexe C présente un tableau de correspondance entre ces URL réduites et les URL originales.

La version numérique de ce document comporte des liens hypertextes, qui apparaissent dans une police de couleur différente selon la nature du lien (référence bibliographique, note de bas de page, etc.). Les logiciels `Acrobat Reader` et `Acrobat Pro` possèdent de nombreux raccourcis clavier pour naviguer dans ce type de documents, l'un des plus utiles étant probablement la combinaison Alt + ← (ou command + ← sous Mac), qui permet par exemple de revenir au lien sur lequel on a cliqué après avoir examiné le contenu vers lequel il pointait.

## 1.4 Organisation du document

Le Chapitre 2 propose une esquisse d'analyse de mon positionnement épistémologique. J'y commente des concepts récurrents dans notre activité quotidienne comme, par exemple, la notion de corpus ou encore l'utilisation qui est faite des méthodes quantitatives dans nos domaines.

Au Chapitre 3, j'énumère et analyse quelques-unes de mes réalisations techniques. J'en profite également pour mettre en lumière cette partie très chronophage, mais néanmoins très gratifiante, de la façon dont j'ai choisi d'exercer mon métier : je fais référence à l'autonomie technologique, vers laquelle je tends depuis toujours. La phonétique que je pratique, instrumentale et expérimentale, est très singulière au sein des Études Anglophones ; elle nécessite donc, loin des grands laboratoires de sciences du langage ou d'informatique, un investissement personnel très marqué.

Les trois chapitres suivants synthétisent les travaux que j'ai conduits en phonétique depuis ma thèse. Le Chapitre 4 inventorie l'essentiel des recherches que j'ai menées en lien avec la notion de variétés d'une langue et s'intéresse, en particulier, aux accents de l'anglais. Je prends pour point de départ ce qui a immédiatement suivi ma thèse pour aboutir au test de l'hypothèse de la phonémicité gradiente.

Au Chapitre 5, j'offre un aperçu des domaines de recherches que j'ai explorés jusqu'ici. Il n'était probablement pas souhaitable de consacrer un chapitre à chaque thématique, ce qui aurait considérablement allongé le document. C'est donc à un panorama des travaux qui n'apparaissent pas au Chapitre 4 que j'invite les lecteurs pour le Chapitre 5.

C'est le Chapitre 6, sur le *deep learning*, qui clôt cet exposé (avant un tout dernier

chapitre, 7, consacré à mes perspectives) ; et ceci pour deux raisons. D'abord, parce que la modélisation s'appuyant sur le *deep learning* coïncide avec les travaux que je suis en train de réaliser. Ensuite, et surtout, parce que les méthodes de *deep learning* représentent dans mon parcours un aboutissement scientifique qui, en quelque sorte, fait la synthèse de nombreux aspects évoqués dans les chapitres précédents.

# 2

# Positionnement épistémologique

## Sommaire



## 2.1 Introduction

Dans ce chapitre, je tente de caractériser mon identité de chercheur à travers une discussion de certaines préférences méthodologiques et épistémologiques héritées de mes lectures et de mon expérience. Au fil des ans et de nombreuses remises en question, ma conviction sur les aspects que j'aborde ici s'est renforcée, étayée par de nombreuses observations. J'espère démontrer en particulier que les notions de corpus et d'analyse quantitative doivent être désacralisées si l'on souhaite pratiquer une science sereine. Je

conclus ce chapitre en soulignant la nécessité de répliquer les études et de tendre vers une véritable culture du graphique. Je n'esquisse naturellement ici que quelques pistes embryonnaires qui mériteraient des développements bien plus riches.

## 2.2 Contextualisme

L'impact du contexte est déterminant dans la construction de toute connaissance scientifique. L'argument contextualiste (voir par exemple Longino, 1990) pose que, la science étant une entreprise sociale, un protocole ou un résultat ne sont déclarés objectifs que s'ils sont conformes aux valeurs partagées par une communauté scientifique. Il n'y a donc pas de vérité scientifique en dehors d'un environnement précis. Cette position tranche très nettement avec la conception un peu naïve de la science que j'avais avant d'exercer ce métier, qui faisait de l'objectivité scientifique en dehors de tout contexte un but réaliste.

Dans la continuité de Longino (1990), Oreskes (2019) développe un argumentaire très convaincant dans son ouvrage au titre audacieux : *Why Trust Science?*Pour cette dernière, l'Histoire nous apprend que la connaissance scientifique est périssable. Elle l'est, entre autres, parce que, malgré la constance d'un phénomène observé, les valeurs de la société évoluent. Or les hommes et les femmes qui pratiquent la science possèdent tous leurs propres valeurs, qu'elles soient religieuses, politiques ou culturelles. Et bien que la formation à la recherche espère nous conduire à réprimer nos propres biais, il est évident que cet objectif ne peut être que très partiellement atteint. Pour Oreskes (2019), le seul moyen de rendre la science plus fiable, c'est donc de veiller à ce qu'il y ait parmi les personnes qui la pratiquent une diversité démographique suffisante pour garantir une diversité des points de vue qui, ensuite, va donner lieu à un consensus. Idéalement pour Oreskes (2019), cette communauté scientifique très inclusive devrait aller encore plus loin : puisque nos valeurs individuelles forgent malgré nous[4] notre grille de lecture, pourquoi ne pas rendre nos valeurs publiques au moment où nous communiquons nos résultats ? Si cette proposition conduirait très vraisemblablement à une science plus fiable, je doute cependant que beaucoup s'y plient.

Sans aller jusqu'à de telles extrémités, Thomas Kuhn défendait une position similaire déjà dans les années 1960 : c'est un cadre particulier qui détermine quel problème doit être résolu, quelles sont les technologies nécessaires à sa résolution, et comment la pensée doit être orientée pour parvenir au but pré-défini. L'affirmation qui suit est emblématique

4. Ce n'est pas toujours involontaire : par exemple la (socio-)linguistique dite « interventionniste » semble pleinement assumer le mariage entre « objectivité » scientifique et militantisme politique.

de la pensée de Kuhn, et j'y souscris très volontiers : « something like a paradigm is prerequisite to perception itself » (Kuhn, 1996, p. 113). Dans le langage courant, il n'est pas rare qu'on reproche à quelqu'un d'avoir des a priori ; or en science comme dans le quotidien, la construction d'un savoir et la plausibilité d'un état des choses ne peuvent être élaborés qu'à partir d'a priori, d'attentes, qui seront soit corroborées soit remises en question.

Dans mon parcours, les interrogations liées à la philosophie des sciences ont été omniprésentes. Du fait de mon intérêt particulier pour la comparaison de voix en criminalistique depuis plus de vingt ans, et encore davantage depuis ma participation au projet VoxCrim et mon rôle de coordinateur dans le projet VoCSI-Telly, je m'interroge aussi sur la réception de la science par le grand public. Symboliquement, le point de départ de ces questionnements remonte au début des années 2000, lorsque j'avais assisté à un procès aux assises de Lyon impliquant l'identification d'un suspect par la voix. L'auteure du rapport d'expertise vantait la grande fiabilité de sa méthode. En face d'elle témoignait Louis-Jean Boë, linguiste de Grenoble, qui mettait en garde contre l'absence de validation scientifique des méthodes proposées dans le rapport. J'avais tout juste ma Maîtrise en poche, mais cela ne m'avait pas empêché de détecter quelques faiblesses dans l'argumentation de l'experte. Mais qu'importe... aux yeux du jury, l'experte, ses méthodes pourtant non validées par des pairs et son matériel identique, avait-t-elle déclaré, à celui du FBI, avaient gagné la bataille rhétorique. Car au final, c'est bien de cela qu'il s'agissait : convaincre avec des phrases habiles, séduire le public et le jury. De mon humble point de vue, que Louis-Jean Boë fût là pour représenter toute la communauté scientifique impliquée dans l'identification du locuteur, cela n'a pas pesé bien lourd dans le débat. Là encore, le contexte a eu une influcence déterminante sur la crédibilité perçue.

Ce que rapporte Jean-François Bonastre de son expérience de témoin dans les tribunaux, dont l'essence est consignée dans Bonastre (2020), est particulièrement éclairant à ce sujet. Il est contreproductif de dévaloriser l'intime conviction d'un expert (ou même d'un témoin) car cette personne est tout de même parvenue à une forme de connaissance (sur la base de son expérience personnelle, de sa pratique) si subjective et biaisée soit-elle. Il faut se contenter d'objecter que seule la méthode scientifique offre les garanties nécessaires (reproductibilité, explicabilité, traçabilité, revue par les pairs, etc.) pour répondre de la manière la plus juste possible à une question donnée. Oreskes (2019, p. 62) généralise ce point de vue à tous les types de connaissances : « Where lay knowledge overlaps with scientific knowledge, one should not assume that the latter is necessarily superior

to the former. ». Le débat sur le critère de démarcation qui sépare les sciences véritables des pseudo-sciences est central en philosophie des sciences (Curd et Cover, 1998), et son intérêt sociétal n'a jamais été aussi évident qu'en cette période de pandémie.

Dans cette première section, j'ai très brièvement mis en avant le poids du contexte dans toute connaissance scientifique et l'impact de la rhétorique dans la communication des chercheurs. Pour compléter cet arrière-plan épistémologique général qui éclaire en permanence ma réflexion, j'aborde à présent des considérations plus spécifiques aux domaines liés à l'étude du langage.

## 2.3 Linguistique de corpus ?

Je souhaiterais tempérer l'enthousiasme que suscite le recours à des corpus en phonologie. Loin de remettre en question leur utilité, je suggère que leur pertinence est souvent exagérée, que l'approche expérimentale est souvent mieux adaptée, et je défends en parallèle l'idée qu'il est temps d'opérer un changement d'échelle par rapport aux standards qui ont prévalu ces vingt dernières années.

L'expression « linguistique de corpus » me met mal à l'aise depuis toujours. En effet, la plupart des collègues travaillent avec des données empiriques (parfois abondantes), il me semble donc bien souvent superflu de mettre en avant ce trait qui n'a aujourd'hui plus rien de particulier. Pourtant, les occasions où cette expression est brandie comme un gage de qualité sont nombreuses. Incidemment, je dois indiquer en préambule que « corpus » est un terme que je n'utilise que très rarement pour mes propres données tellement il me paraît connoté : je parle plus volontiers de « base de données » ou tout simplement de « mes données ».

Avec le recul, j'identifie trois raisons expliquant mon malaise. D'abord, j'arrive probablement trop tard pour mesurer pleinement l'ampleur du changement intervenu en linguistique après une suprématie des méthodes introspectives et des jugements de grammaticalité dans les années 1950-1960 (Gilquin et Gries, 2009). Ensuite, je rejoins certains auteurs (Meyer, 2002 ; Gilquin et Gries, 2009) qui considèrent que la linguistique de corpus n'est souvent qu'une façon de faire de la linguistique ; pas un objet ou une thématique. Revendiquer cette étiquette est donc souvent injustifié[5], a fortiori quand, comme cela se

5. Dans certains cas pourtant, l'appellation n'est pas usurpée : je fais référence à la réflexion sur les stratégies d'annotation, de stockage, de pérennisation, d'interopérabilité, de normalisation des corpus. Tout ce travail correspond davantage à des profils d'ingénieurs en base de données. En dehors de ces cas précis, il me paraît inutile de se présenter comme linguiste de corpus quand on fait de la linguistique qui, accessoirement, s'appuie sur des données.

fait encore parfois, les travaux en question se contentent d'extraire quelques exemples d'un corpus pour les commenter. Je crois en effet que, dans ces cas précis, collecter spontanément des exemples linguistiques authentiques (à la radio, dans les journaux, ou même dans les productions de proches) n'est pas moins noble ou moins légitime ; sans compter que cette dernière méthode garantit la « fraîcheur » des exemples choisis ![6]

La troisième raison qui a renforcé chez moi l'idée que le concept de corpus n'était pas central me vient probablement du milieu très interdisciplinaire dans lequel j'ai appris la recherche, qui comportait, outre des linguistes, des informaticiens et des spécialistes de sciences cognitives et neurosciences. Par exemple, de la même manière que, comme je le notais dans l'Introduction de ce document, le pragmatisme de François Pellegrino m'a aidé à démystifier les méthodes quantitatives, c'est encore cette même approche très terre-à-terre qui m'a conduit à ne rien voir de si spécial dans l'idée de corpus.

### 2.3.1 Corpus et méthode hypothético-déductive

Mon expérience et mes lectures ont conditionné ma très nette préférence pour les corpus *ad hoc* par rapport aux corpus « tout-venant ». Le corpus *ad hoc* est recueilli pour répondre à une hypothèse précise en s'appliquant à réduire le bruit induit par des facteurs non pertinents au moment de la collecte ; le corpus tout-venant le plus caractéristique consiste, lui, en une collection de données sans contrôle systématique du bruit, et sans hypothèse très précise a priori. Gilquin et Gries (2009, p. 5) proposent, quant à eux, une échelle allant du moins naturel au plus naturel. En bas de l'échelle, ils citent l'exemple de l'échographie de la langue ; en haut, on retrouve la compilation de documents écrits préexistants. Une des raisons qui me fait privilégier le corpus *ad hoc* est fournie par Harrington (2010, p. 6) :

> Unfortunately, most kinds of phonetic analysis still require building a speech corpus that is designed to address a specific research question. In fact, existing large-scale corpora of the kind sketched above are very rarely used in basic phonetic research, partly because, no matter how extensive they are, a researcher inevitably finds that one or more aspects of the speech corpus in terms of speakers, types of materials, speaking styles, are insufficiently covered for the research question to be completed.

Une dizaine d'années après, je pense que cette citation est toujours d'actualité. Il est bien évident que certains grands corpus tout-venant peuvent apporter des réponses à

6. D'ailleurs, même les plus fervents tenants des corpus n'hésitent pas à avoir recours à ces observations fortuites comme par exemple Durand et Lyche (2016, p. 374) ; cela prouve que ces approches ne s'excluent pas mutuellement.

des questions très générales comme l'effet du voisement des occlusives sur la durée de la voyelle qui précède (Tanner *et al.*, 2020) ou encore le comportement de la liaison en français (Adda-Decker *et al.*, 2012 ; Durand et Lyche, 2016). Néanmoins, je rejoins Harrington sur le fait qu'il est nécessaire de collecter soi-même ses propres données, mais je m'oppose très franchement à considérer cela comme un pis-aller (« Unfortunately… »). J'espère, au contraire, que cela continuera d'être le cas ; les raisons qui motivent ma position devraient apparaître clairement à la fin de ce chapitre.

La lecture de Popper a eu un sérieux impact sur mes choix épistémologiques. Dans sa méthode déductiviste, Popper insiste sur le point suivant : « a hypothesis can only be empirically *tested* — and only *after* it has been advanced » (Popper, 2002b, p. 7). Or l'étude de Gilquin et Gries (2009), qui se fonde sur un examen approfondi de 81 articles publiés dans des revues de linguistique de corpus, démontre que 72 % des auteurs adoptent un angle exploratoire sans même formuler d'hypothèse explicitement. Si on adhère à l'idée que la science consiste à tester des hypothèses, et puisqu'une hypothèse ne peut être testée qu'une fois qu'elle a été formulée, alors force est de constater que tout le monde ne partage pas l'idéal popperien. En réalité, ces pratiques sont de plus en plus considérées comme scientifiquement « douteuses » (Warren, 2018), et devraient donc tendre à disparaître.

Le constat de Gilquin et Gries (2009) est tout à fait cohérent avec mon ressenti au quotidien : quand j'ai commencé la recherche au début des années 2000, une sorte d'injonction collective poussait à une utilisation immodérée du terme « corpus » au point où la question de recherche passait parfois au second plan. Il est indéniable, je le répète, qu'utilisés avec discernement, les corpus renseignent la recherche en linguistique et en phonologie. Mais puisque les mérites du corpus font l'objet d'un quasi consensus, il est bien plus utile ici d'en rappeler les limites et d'évoquer quelques alternatives qui sont au moins aussi légitimes.

Dans *A Scandal in Bohemia*, la première nouvelle de *The Adventures of Sherlock Holmes* d'Arthur Conan Doyle, le célèbre détective déclare : « It is a capital mistake to theorise before one has data ». Cette phrase est restée affichée pendant des années sur la porte d'un bureau du Laboratoire Dynamique Du Langage à l'époque de ma thèse ; et plus je passais devant cette porte, moins je souscrivais à ce qui était écrit dessus. En effet, pour faire écho à ce qui a été dit jusqu'ici, il est capital d'avoir des a priori auxquels des données recueillies pour l'occasion seront confrontées, ce qui entache sévèrement le bien-fondé de la maxime holmesienne. Mais alors, comment éviter un biais de confirmation qui conduirait à ne prêter attention qu'aux données qui vont dans le sens de nos idées

préconçues ? Popper (2002a,b) apporte une réponse claire : le seul moyen de tester une théorie, c'est de tenter de la réfuter, de la falsifier[7].

### 2.3.2 Le problème de l'induction et le surajustement

Dans la scène VII de *La Cantatrice chauve* d'Eugène Ionesco, on sonne à la porte ; Mme Smith va ouvrir et, trois fois d'affilée, elle ne trouve personne. Elle en conclut que l' « expérience nous apprend que lorsqu'on entend sonner à la porte, c'est qu'il n'y a jamais personne ». Mme Smith — pour faire écho à la citation de Sherlock Holmes *supra* — a pourtant bel et bien attendu d'avoir des données avant de théoriser... Popper résume le problème de l'induction avec un exemple célèbre : « no matter how many instances of white swans we may have observed, this does not justify the conclusion that *all* swans are white. » (Popper, 2002b, p. 4)[8]. Le problème de l'induction, illustré dans ces deux exemples, est bien connu en philosophie des sciences (Curd et Cover, 1998). Il me paraît particulièrement marqué dans la linguistique (et donc en phonologie) de corpus lorsque la démarche est exploratoire et qu'elle s'appuie sur un ensemble de données qui n'a pas été collecté pour l'occasion. En effet :

1. Comment généraliser à partir de données qui sont intrinsèquement biaisées ?
2. Comment généraliser à des données absentes du corpus ?

Concernant le premier point, il n'est probablement pas utile de rappeler dans le détail le décalage bien connu entre les seules techniques d'échantillonnage permettant de généraliser des observations à l'ensemble d'une population (c-à-d. celles s'appuyant sur un tirage aléatoire strict) et l'échantillonnage opportun (et donc sous-optimal) dont nous devons nous contenter lorsque nous collectons nous-mêmes nos données. Quiconque a été amené à réaliser ce type de travail sait, par exemple, que les premiers volontaires contactent ensuite leur famille et leurs amis, ce qui induit un biais conséquent !

Et quand bien même ce ne serait pas le cas, il est bien difficile d'affirmer que les 15 ou 20 locuteurs enregistrés sur tel ou tel site représentent la tendance phonologique générale du site en question. Sans oublier que cette représentativité peut varier d'un site à l'autre comme cela semble avoir été le cas dans le corpus *Accents of the British Isles* (ABI), que j'ai utilisé pour ma thèse. D'ailleurs, 20 locuteurs pour chacun des 14 sites d'enregistrement, cela me paraissait, déjà à l'époque, un peu sous-dimensionné. Presque

7. Pour un aperçu très synthétique de la position de Popper à ce sujet, voir Popper (2002a, pp. 36-37).
8. On sait grâce à cet exemple très connu qu'il existe des cygnes noirs en Australie. Brassens aurait probablement plus volontiers parlé de merle blanc ; les références aviaires semblent opportunes.

20 ans plus tard, à l'heure où le corpus anglais du projet *Common Voice* regroupe 66 173[9] locuteurs, on ne peut plus se contenter de si peu.

Pour répondre à la seconde question, qui concerne, entre autres, la capacité des modèles à généraliser à des données qui n'ont pas été vues lors de la phase d'apprentissage, j'ai choisi une illustration ludique issue du domaine de la vision par ordinateur. Dans la Figure 2.1, on aperçoit en haut à gauche le Ohbot Pi, petit robot qui est — très partiellement, j'en conviens — anthropoïde. Pour une première expérience, on tente de classer l'image du Ohbot à partir du modèle Alexnet, un réseau de neurones à convolution entraîné à classifier plus d'un million d'images en 1 000 types différents (objets, animaux, etc.)[10]. La classe affichant la probabilité la plus forte est celle des téléphones publics ; autrement dit, Alexnet identifie le Ohbot comme un téléphone public. Parmi les autres classes avec une probabilité relativement forte, on trouve une pompe à essence, une tapette à souris et une machine à coudre. La proximité avec le téléphone public et la pompe à essence s'explique probablement par la présence du câble à droite de l'image. Cet exemple souligne donc que, comme tout modèle entrainé à classifier, Alexnet ne peut faire de prédictions qu'à partir de ce qu'il a vu, négligeant ainsi le fait que le Ohbot partage de nombreux attributs avec des êtres animés : il a des yeux réalistes, il bouge et il parle (ces deux derniers traits ne se voient évidemment pas sur l'image). Les yeux du Ohbot auraient pu conduire Alexnet à le prendre pour l'un des nombreux animaux qu'il a appris, mais ce n'est pas le cas.

Dans une seconde petite expérience, en postulant que les yeux du Ohbot présentent un anthropomorphisme convaincant, j'ai soumis son image, ainsi que trois autres, à l'algorithme de détection des yeux livré avec la Computer Vision Toolbox de `Matlab`. On constate à la Figure 2.1 que l'algorithme fait parfaitement son travail à partir d'une photographie, d'un portrait[11], et même avec l'image d'un monstre issu d'un dessin animé ; mais les yeux du Ohbot ne sont pas détectés. L'apprentissage statistique à partir de regards humains ne permet donc pas de détecter les yeux du Ohbot.

Il serait informatif de savoir si un enfant qui n'aurait jamais vu d'autres yeux que ceux d'êtres humains bien en place dans leurs orbites serait capable, en voyant le Ohbot, de considérer ces deux billes comme des yeux. Si oui, cela démontrerait que malgré la sophistication de l'apprentissage statistique effectué par la machine, elle n'a pas été en

9. Dans sa version en_2181h_2020-12-11. Voir : https://commonvoice.mozilla.org/fr/datasets.
10. On retrouvera Alexnet plus en détail au Chapitre 6.
11. Les deux humains dans cette figure sont Yngwie Malmsteen, guitariste électrique virtuose emblématique, et Gilbert-Louis Duprez, ténor d'opéra français ayant importé d'Italie au début du XIX$^{e}$ siècle la technique du contre-ut de poitrine.

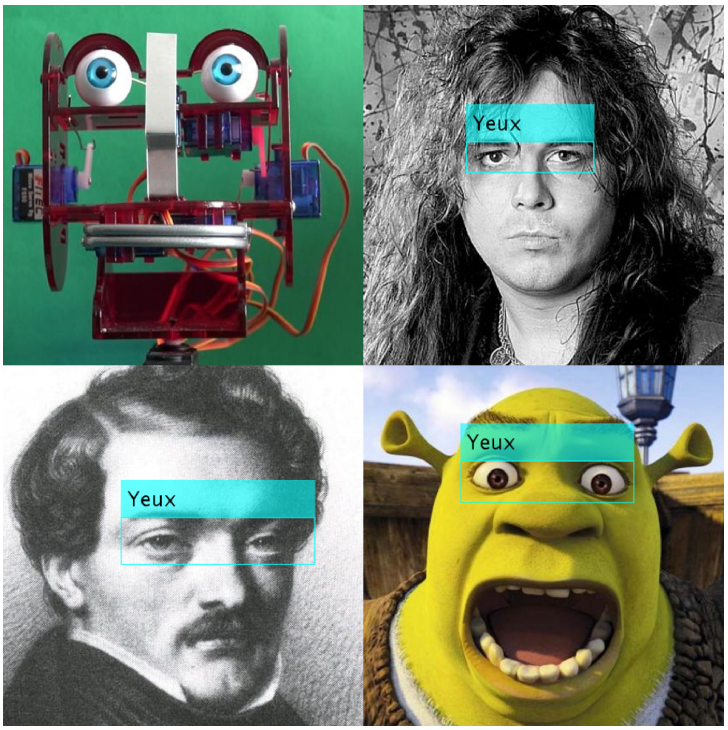



FIGURE 2.1 – Détection automatique des yeux pour illustrer des capacités de généralisation lacunaires.

mesure de capturer l'abstraction caractérisant l'essence d'un œil[12]. En d'autres termes, le modèle a été surajusté à ses données d'apprentissage, et c'est exactement ce qui est susceptible de se produire quand on généralise à partir d'un corpus.

Ces deux exemples ne sont à mon avis pas de simples analogies avec la linguistique de corpus. Ils me paraissent au contraire parfaitement illustrer le biais de l'induction, celui-là même dont souffre la linguistique de corpus, en particulier quand elle est pratiquée sans une question de recherche claire présidant à la constitution du corpus. Comment, par exemple, étudier les phénomènes phonologiques marginaux qui m'ont conduit à développer la notion de phonémicité gradiente (voir Section 4.3) à partir de corpus préexistants ? En effet, en raison de la rareté de certains de ces phénomènes, il est plus réaliste de recourir à des méthodes d'élicitation plutôt qu'espérer en trouver en nombre suffisant (qui plus est dans des contextes phonologiques contrôlés) dans un corpus tout-venant.

L'absence générale de remise en question forte des études s'appuyant sur les corpus en phonétique semble d'autant plus paradoxale que les critiques pleuvent sur l'intelligence artificielle et ses biais (Zou et Schiebinger, 2018 ; Mehrabi *et al.*, 2019). On accepte

12. On peut supposer ici que le modèle a appris la classe « yeux » à partir d'un attribut accidentel : la présence de paupières.

donc avec les corpus des généralisations qui découlent de l'observation d'une dizaine ou centaine de locuteurs par une subjectivité humaine et son filtre de valeurs, mais on se presse aux nombreuses tables rondes où l'on fustige les carences d'un algorithme pourtant reproductible qui a généralisé à partir de millions d'exemples.

Pourtant il est certain que les biais d'échantillonnage, connus depuis longtemps par les instituts de sondage[13] sont un problème central dans la collecte de corpus phonétiques. Outre le faible nombre de locuteurs, ce sont prioritairement les personnes disposées à se laisser enregistrer — parce qu'elles ont le temps, ou parce qu'elles ont une propension naturelle à se prêter à ce genre d'exercice, parfois se faisant un devoir de fournir aux linguistes ce qu'elles pensent qu'ils sont venus chercher (le paradoxe de l'observateur, Labov, 1972) — qui vont être sur-représentées dans des données. Les généralisations qui en découlent ne s'appliquent donc qu'à la population des personnes qui sont disposées à enregistrer des corpus... Mais, bien évidemment, les corpus phonétiques n'ont pas l'apanage de ce type de limitation ; il en va de même pour tout type d'expérience psycholinguistique par exemple. Néanmoins, l'approche expérimentale comporte quelques « astuces » visant à court-circuiter des réponses explicites (sans parler des mesures électrophysiologiques qui sont difficilement manipulables par les participants).

### 2.3.3 Corpus mental et réhabilitation de l'intuition

Malgré ma réticence certaine concernant les corpus tout venant et ma préférence marquée pour les corpus *ad hoc*, il existe néanmoins un cadre théorique dans lequel le corpus généraliste a un intérêt ; il s'agit des modèles à exemplaires (Nosofsky, 1988) et du courant linguistique associé, les grammaires *usage-based* (Bybee, 2010). Ces modèles mettent en avant le rôle de la fréquence dans la formation de catégories linguistiques, dans le processus de catégorisation lui-même, ainsi que dans certains phénomènes de production (e.g. lénition). Pour prendre l'exemple de la phonologie (Pierrehumbert, 2001), chaque occurrence d'une voyelle est stockée dans un espace multidimensionnel qu'on pourrait qualifier de psycho-acoustique. Il s'agit donc d'un apprentissage statistique donnant lieu à l'estimation de distributions de probabilités dont les paramètres sont constamment mis à jour en fonction de chaque nouvelle occurrence rencontrée. On peut donc (provisoirement) en déduire que l'estimation la moins mauvaise qu'on puisse faire des paramètres de

13. L'exemple célèbre du sondage du *Literary Digest* pour l'élection présidentielle américaine de 1936 démontre que si l'on pose une question aux mauvaises personnes, quand bien même la taille de l'échantillon serait gigantesque (en l'occurrence plus de deux millions d'individus), une prédiction peut se révéler fausse (Squire, 1988).

ces distributions peut être obtenue à partir de corpus.

Pour résumer, nos représentations mentales, d'après les modèles *usage-based*, auxquels j'adhère totalement, accordent une place prépondérante aux fréquences d'occurrences. Si l'on souhaite une estimation quantifiée de ces fréquences, seul un corpus de très grande taille peut fournir une réponse au moins partiellement satisfaisante. Est-ce à dire que les usagers d'une langue ne sont pas capables de nous renseigner sur les fréquences qui pourtant forment le socle de leurs connaissances linguistiques ? Autrement dit : faut-il vraiment renier l'introspection et l'intuition ? Sur ce point précis, je réponds avec Taylor (2012, p. 176) par la négative :

> Half a century of research has confirmed that speakers' subjective judgements do indeed correspond, by and large, to objective measures of frequency, as established from corpora, or, in more recent research, Internet searches...

Il existe donc un corpus mental, pour reprendre le titre de l'ouvrage de Taylor, pour lequel la représentation des fréquences est, en général, corrélée à celle qu'on peut estimer à partir de données linguistiques. Et c'est notamment sur ce point que je souhaiterais insister : nos représentations linguistiques sont formées par l'accumulation de traces mnésiques multidimensionnelles, et nous avons encore accès à la fréquence à laquelle nous avons été exposés à tel ou tel point dans cet espace multidimensionnel. En d'autres termes, pour paraphraser la célèbre locution cartésienne, on pourrait proclamer : j'entends, donc je suis un corpus ; *audio, ergo corpus sum.*

Bien sûr, accéder à ce corpus mental doit se faire avec méthode. Par exemple, en se demandant si on emploierait telle forme plutôt que telle autre, des considérations normatives inspirées de l'orthographe ou héritées de l'école risquent de biaiser notre jugement (Durand, 2009). Par ailleurs, notre perception de la fréquence n'est pas toujours objective (on pense à l'exemple célèbre de *kick the bucket*, perçu comme fréquent mais relativement rare dans les corpus comme le montrent Popiel et McRae, 1988), et pour des raisons d'optimisation de l'utilisation de nos ressources cognitives, cette perception de la fréquence, comme un grand nombre de quantité psychophysiques, est logarithmique (Varshney et Sun, 2013). Néanmoins, le fait-même qu'on continue de mener des expériences de perception sur le langage prouve d'une part qu'on accepte implicitement l'intérêt de questionner directement le corpus mental, et d'autre part, que nous accédons par ce biais à des renseignements que les corpus plus « matériels » ne nous fournissent pas.

Dans le domaine des Études Anglophones en France, on parle plus volontiers d'« oralistes » que de « phonéticiens » ou « phonologues ». L'étiquette est pour le moins paradoxale puisque historiquement, dans la lignée des travaux fondateurs de Lionel Guierre,

les oralistes prototypiques n'analysent pas de données orales à proprement parler, mais plutôt des transcriptions phonétiques issues de dictionnaires. Il est facile — et légitime — d'objecter à cette démarche l'authenticité plus que douteuse du corpus. En effet, les transcriptions de ces dictionnaires ne peuvent refléter que la (ou l'absence de) fantaisie de leurs auteurs ainsi qu'un prescriptivisme suranné. J'ai longtemps fait partie des détracteurs de cette approche mais mon opinion s'est nuancée avec le temps à la lumière de ce qui a été dit plus haut concernant le biais de l'induction.

En effet, qu'étudie-t-on réellement en linguistique et en phonologie ? S'agit-il de l'ensemble restreint des formes que le hasard a bien voulu nous offrir pour notre corpus lors d'une collecte de données ponctuelle et hautement tributaire du contexte ? Ou s'agit-il plus généralement de la faculté de langage, ce qui inclut, en plus du modeste sous-ensemble que représente notre corpus, le linguistiquement plausible et le phonologiquement vraisemblable ? Je crois que c'est ce second aspect qui nous intéresse.

C'est également ce que semble penser Bloomfield lorsqu'il écrit « The totality of utterances that can be made in a speech community is the language of that speech-community. We are obliged to predict; hence the words 'can be made'. » (Bloomfield, 1926, p. 155). C'est bien un potentiel, comme le relève Taylor (2012) en commentant cette citation, qui semble intéresser Bloomfield. Et je crois que l'approche de la linguistique qui s'appuie sur des corpus se trouve parfois dans une forme d'impasse puisqu'elle ne parvient pas à révéler ce potentiel. Un corpus, en cela qu'il n'est qu'un échantillon, enregistre déjà une version simplifiée et biaisée de la réalité. Mais quand bien même ce corpus contiendrait tous les mots et les phrases qui ont été prononcés dans une langue depuis sa création, il ne permettrait pas de cerner complètement le potentiel de cette langue.

D'ailleurs, dans leur ouvrage *Comprendre la Phonologie*, de Carvalho *et al.* (2010, p. 74) ne s'y trompent pas lorqu'ils écrivent que l' « existence, fortuite, de paires minimales suffit, en général, à prouver celle d'une opposition, mais l'absence de telles paires n'implique absolument pas qu'il n'y ait pas d'opposition. » C'est donc vers le phonologiquement et le phonotactiquement plausible qu'il faut se tourner. Il convient ainsi de combler les lacunes du corpus par le recours à l'intuition et à l'expérimentation.

Voici une anecdote opportune qui illustre bien mon propos concernant la dichotomie entre faculté de langage et occurrences attestées. À l'occasion de la réalisation d'examens de transcriptions phonétiques contraints au distanciel avec mes collègues Anne Guyot-Talbot et Sylvain Navarro, nous avons dû adapter notre pratique. Comment tester à distance la transcription phonologique des étudiants sachant qu'il leur est bien facile de

rechercher les mots à transcrire en temps réel dans un dictionnaire ou en ligne ? Nous avons tout simplement décidé de créer des pseudo-mots.

En connaissant les règles d'accentuation et de correspondance graphème-phonème ou, à défaut, en mettant à contribution des connaissances implicites et le principe d'analogie, la transcription des mots en question était totalement prévisible. Ainsi, *plotation, crocodonia* ou encore *detesticate* sont autant de mots plausibles dont le schéma accentuel et la nature des segments sont prévisibles, et qui pourraient même, dans un contexte précis, avoir une signification. Dans la consigne, il était précisé que la prononciation des mots du test était régulière.

Cette expérience, qui a valu aux étudiants des résultats cohérents avec ceux obtenus à partir d'exercices plus classiques les autres années, nous apprend aux moins deux choses. D'abord, bien sûr, cela confirme que les principes phonographématiques compilés dans la tradition de Lionel Guierre (voir Ballier, 2016, pour une vue détaillée de cette tradition) sont extrêmement puissants. Ensuite et surtout, le fait qu'il n'y ait pas eu la moindre réclamation de la part des étudiants tend à prouver qu'ils considèrent eux-mêmes qu'analyser la forme phonologique de mots qui n'existent pas n'a rien de surprenant.

Un autre aspect qui échappe en grande partie au corpus (par rapport au corpus mental), c'est la richesse de la mémoire épisodique (voir par exemple Pierrehumbert, 2016). Les mots (et leurs variantes phonologiques) sont entendus dans un contexte social et émotionnel précis, et les théories épisodiques considèrent que ce contexte est stocké en mémoire avec le mot correspondant. Comment expliquer les connotations et tous les aspects pragmatiques de la communication sans poser l'existence d'une mémoire extrêmement détaillée ? Par exemple, comment expliquer la différence de connotation entre *a real spinster* et *a real bachelor* (Taylor, 1995, p. 97) si on n'a pas entendu ces mots dans des contextes précis, avec, par exemple, un ton sarcastique pour le premier et un ton admiratif ou envieux pour le second ?

Dans le même ordre d'idée, comment savoir à partir d'un corpus que, pour des apprenants francophones tardifs de l'anglais, *slut* n'est pas équivalent à *salope* ? En effet, quoique sémantiquement comparables, ces deux mots ont pour eux une valence émotionnelle très différente — comme nous l'avons montré dans Rastovic *et al.* (2019) en utilisant la réponse électrodermale — très probablement parce que *slut* a été appris dans un contexte émotionnel relativement neutre ou plutôt, en dehors de tout contexte de communication réaliste.

Pour prendre un exemple plus proche de l'étude de la phonétique : un corpus ne saurait

nous dire qu'entendre « Je passe des heures dans le hall de l'immeuble » prononcé avec un accent marqué typique des classes supérieures génère une onde électroencéphalographique spécifique reflétant l'incongruence entre les stéréotypes liés à l'accent en question et le contenu sémantique de la phrase (Pélissier et Ferragne, 2021). Ces derniers exemples illustrent le fait que les informations extraites de corpus sont avantageusement complétées par d'autres méthodes que je valorise dans ma pratique.

### 2.3.4 Un statut spécial pour le corpus en phonologie ?

Le corpus d'enregistrements audio jouit de fait d'un statut particulier : les difficultés liées au recueil de données (recrutement de participants, disponibilité d'un lieu calme, etc.) et les contraintes liées à l'exploitation de ces données — en particulier l'annotation ; on compte généralement une heure de travail pour transcrire phonologiquement une minute de parole (Gut et Voormann, 2014) — sont autant d'éléments qui limitent la taille du produit fini[14]. Cependant, je ne crois pas que ces corpus puissent bénéficier d'une dispense quant à leur taille simplement à cause de contraintes matérielles. Et c'est encore plus vrai pour une langue aussi répandue que l'anglais quelque 70 ans après Peterson et Barney (1952) et leurs 76 locuteurs.

À ce sujet, Viollain et Chatellier (2018) émettent l'idée qu'on peut se contenter de petits corpus en phonologie puisque l'inventaire phonémique d'une langue est fini, par opposition au caractère potentiellement infini de l'inventaire de ses phrases. Il m'est difficile d'adhérer à ce point de vue. D'abord, parce que, comme déjà évoqué dans la Section 2.3.3, l'absence de paires minimales attestées ne suffit pas à établir une absence de distinction phonémique (de Carvalho *et al.*, 2010, p. 74) — un corpus, a fortiori petit, ne suffit donc pas à établir un inventaire phonologique exhaustif — et ensuite parce que dresser la liste des oppositions au sens le plus structuraliste du terme ne représente qu'une fraction de ce que l'on cherche à savoir. En effet, pour prendre l'exemple de la variation dans les accents de l'anglais, la variabilité sur le plan de l'incidence lexicale (certains mots présentent un phonème différent dans certaines variétés) ou encore le caractère saillant de certains phénomènes pourtant rares invitent, a minima, à tendre vers des corpus de très grande taille.

14. Nuançons cependant : Cole et Hasegawa-Johnson (2012, p. 436) estiment qu'une minute de parole à transcrire orthographiquement en prend quatre ; et ils affirment que (comme ce fut d'ailleurs le cas pour ma thèse) les chercheurs procèdent souvent d'abord à une transcription orthographique pour ensuite obtenir automatiquement une transcription phonologique par alignement forcé (voir aussi Liberman, 2019, à ce sujet).

Viollain et Chatellier (2018) ajoutent que puisque la représentativité dans un corpus ne peut être atteinte, et puisque des contraintes matérielles (comme la durée d'une thèse ou la disponibilité de financements) conduisent à modérer nos ambitions, le choix des petits corpus — 10 à 20 locuteurs comme fourchette « pragmatique » dans le projet *Phonologie de l'Anglais Contemporain* (Viollain et Chatellier, 2018) comme pour son aîné, *Phonologie du Français Contemporain* (Detey *et al.*, 2016) — s'impose comme une solution optimale. Concernant la question de la représentativité, si on ne peut, certes, pas atteindre l'exhaustivité, on peut néanmoins améliorer la précision de notre estimation en particulier en augmentant le nombre de locuteurs. D'après les règles élémentaires de l'inférence statistique[15], pour prendre un exemple fictif, l'affirmation « 70 % des locuteurs sondés présentent une opposition FOOT-STRUT » a un degré de fiabilité différent selon que l'effectif des locuteurs en question est de 10, 100 ou 1000.

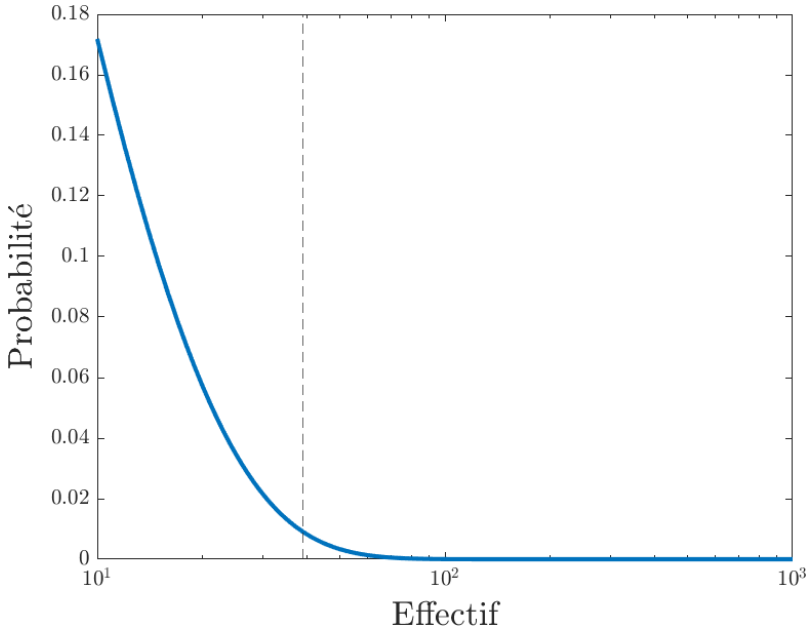



FIGURE 2.2 – Probabilité qu'un événement avec une probabilité de 0,5 se produise au moins 7 fois sur 10 en fonction de l'effectif.

La Figure 2.2 présente l'évolution de la probabilité qu'au moins 70 % des locuteurs aient une opposition FOOT-STRUT en fonction du nombre de locuteurs si la probabilité a priori était de 0,5. Autrement dit (j'utilise ici la loi binomiale), on cherche à savoir à

15. Voir les introductions aux statistiques comme par exemple Wonnacott et Wonnacott (1991) ou Baillargeon (1982).

partir de quand un déséquilibre de 70 % constitue un écart fiable par rapport à 0,5[16]. En prenant le critère, courant en statistiques, d'une probabilité inférieure ou égale à 0,01, on voit (ligne verticale en pointillés) qu'il faut au moins 39 locuteurs. La probabilité continue de décroitre (et donc la fiabilité, de croitre) au fur et à mesure que l'effectif augmente, mais de moins en moins rapidement. Ces observations prouvent bien, d'une part, qu'il existe un nombre minimal de locuteurs requis pour qu'une estimation soit relativement robuste. D'autre part, le ralentissement de la courbe et le « coude » qu'on observe montrent qu'il n'est en pratique pas avantageux d'aller au-delà d'une cinquantaine de locuteurs dans ce cas précis[17]. Notons au passage que puisque cette estimation s'appuie sur une loi de probabilité qui requiert des événements indépendants et qu'en pratique ce pré-requis est compromis (si on enregistre des locuteurs d'une même famille, ou d'un même quartier, on peut imaginer que leurs systèmes phonologiques ne sont pas « indépendants »), mon estimation sommaire de 39 locuteurs est très optimiste.

Au sujet de la taille du corpus, je rejoins Liberman (2019), qui remarque que dans un contexte d'élicitation contrôlée, c'est déjà une grande quantité de données qu'il est nécessaire de collecter pour couvrir une partie raisonnable de la variation ; et c'est donc sur une échelle bien plus grande qu'il faut se placer pour espérer recueillir des échantillons qui soient représentatifs de la parole naturelle.

Si, donc, les corpus tout-venant déjà disponibles ne répondent aux questions qu'on se pose que de manière très imparfaite, et si constituer un corpus *ad hoc* est soumis à des contraintes matérielles qui nous condamnent à une taille minimale et de nombreux biais, que nous reste-t-il ? C'est une question que je me suis posée en particulier au milieu de ma thèse, et la réponse m'a conduit à envisager une suite des événements plus expérimentale et instrumentée. Les expériences de perception sont complémentaires (voire une alternative) aux grands corpus oraux d'abord parce qu'elles présentent l'avantage du gain de temps au moins au moment de l'analyse (Nguyen, 2016). Ensuite, en permettant de cibler une question scientifique précise à laquelle on tente de répondre avec un protocole extrêmement contrôlé, les expériences de perception constituent un choix scientifique avantageux qui évite les travers de l'inductivisme. Quand cela est nécessaire, je les complète avec l'enregistrement d'un bref corpus audio *ad hoc* lui aussi destiné à n'enregistrer

16. C'est une question qu'on pourrait légitimement poser pour décider par exemple, si, sur la base du FOOT-STRUT Split, l'accent de Birmingham, connu pour être intermédiaire entre nord et sud linguistique en Angleterre (Ferragne et Pellegrino, 2010b), devrait être plus naturellement considéré comme méridional ou septentrional.

17. Je me contente d'illustrer de manière intuitive le concept de puissance statistique. Voir par ex. Cohen (1988), Dattalo (2008) ou Kraemer et Blasey (2016).

que le matériel linguistique sur lequel portent mes hypothèses (voir par exemple le court texte dans la note de la page 68).

Dans les expériences de perception, il est en partie possible de contourner le paradoxe de l'observateur et la mise en place consciente de stratégies des participants, alors qu'il est bien malaisé d'obtenir de la parole naturelle dès lors qu'on place un microphone devant quelqu'un. En effet, dans une tâche chronométrée ou dont l'objectif réel n'est pas connu des participants, il est difficile d'élaborer une quelconque stratégie de réponse.

Et afin d'éviter encore davantage de questionner explicitement les participants, le recours à des mesures électrophysiologiques est une solution intéressante ; je pense en particulier à la conductance électrodermale, que nous avons utilisée récemment (Rastovic *et al.*, 2019, dont je décris l'expérience à la Section 5.2.5) ou encore aux potentiels évoqués en électroencéphalographie (EEG), que j'ai employés à plusieurs reprises depuis la fin de ma thèse (Boulenger *et al.*, 2011 ; Pota *et al.*, 2012 ; Bedoin *et al.*, 2019 ; Heidlmayr *et al.*, 2021 ; Pélissier et Ferragne, 2021). Les réponses EEG précoces impliquées dans les traitements phonologiques sont réputées automatiques (Näätänen *et al.*, 2007) ; une quelconque stratégie de la part des participants semble donc peu probable. On peut bien sûr objecter le caractère très artificiel du paradigme expérimental *oddball*[18], mais sur cet aspect, je ferai deux commentaires. D'abord, nous avons pu récemment mettre en lumière les corrélats EEG du traitement de certains contrastes de l'anglais par des apprenants francophones en utilisant une tâche plus « écologique » (Heidlmayr *et al.*, 2021). Plutôt qu'une suite de syllabes, les participants étaient invités à écouter des phrases entières et c'est entre autres par le biais de violations lexico-sémantiques « portées » par un phonème — *The anchor of the ship/*sheep was let down* — que nous avons pu accéder au traitement phonologique. Ensuite, si vraiment on souhaite opérer avec de la parole naturelle, il faut d'abord de très nombreuses données à la fois pour atténuer les biais inhérents aux situations non expérimentales, pour espérer enregistrer assez longtemps pour que les locuteurs « oublient » le microphone et pour arriver à une taille qui puisse constituer un échantillon crédible. Il faut ensuite une grande équipe de transcripteurs. Tout ceci me porte à croire qu'alors que la constitution de corpus *ad hoc* peut encore être prise en charge par un doctorant ou une petite équipe enthousiaste, la collecte et la gestion de grands corpus oraux devrait être confiée à des sous-traitants spécialisés.

18. Par exemple deux phonèmes sont présentées auditivement, l'un, fréquemment (le « standard ») et l'autre, beaucoup plus rarement (le « déviant »). La présentation du déviant génère une onde particulière, la *Mismatch Negativity*, sans que l'auditeur prête attention aux stimuli (Aaltonen *et al.*, 1987 ; Näätänen *et al.*, 2007).

Ma préférence va donc aux approches expérimentales, aux corpus dont le protocole a été élaboré pour répondre à une question ponctuelle et ciblée, et bien sûr, comme cela apparaîtra encore plus clairement dans le reste de ce document, à l'emploi de techniques instrumentales et d'analyses quantitatives variées et complémentaires. Je présente dans la prochaine section une note concernant ce dernier point.

## 2.4 Note sur les méthodes quantitatives

Dans la Section 1.2, je mentionnais la sérénité que j'éprouve vis-à-vis des méthodes quantitatives. Je constate que ces méthodes exercent une grande fascination auprès des chercheurs SHS, aussi, je souhaiterais livrer ici ma vision de la place qu'il est tolérable de leur réserver. J'espère démontrer que notre pratique du quantitatif (en particulier du test de l'hypothèse nulle) est en partie conditionnée par notre volonté de crédibilité scientifique et que cela conduit parfois — et je ne fais pas exception à la règle — à détourner notre attention de notre question de recherche.

En 2008-2009, j'ai tenté de prendre du recul par rapport aux tests d'inférence statistique et à la notion de résultat « significatif ». Ceci m'a amené à exhumer une littérature particulièrement fournie, pourtant largement ignorée dans notre pratique quotidienne[19]. Cette expérience m'a conduit à pendre la juste mesure de facteurs sociologiques (plutôt que strictement scientifiques) qui conditionnent nos habitudes méthodologiques et nos résultats, et à exercer un regard critique à l'égard des fameuses *p-values*. Depuis, je n'omets jamais de sensibiliser les étudiants de Master aux dérives de telles pratiques, en leur distribuant notamment l'article représentatif de Cohen (1994) qui, entre autres, proposait de rebaptiser les tests statistiques « statistical hypothesis inference testing » en vue d'obtenir un acronyme éloquent...

Aborder ce sujet ici est d'autant plus justifié que pour la première fois en 177 ans d'existence, l'*American Statistical Association* a publié assez récemment une déclaration motivée par l'accumulation de controverses au sujet des *p-values* (Wasserstein et Lazar, 2016). Puisque l'inférence statistique est largement utilisée dans nos domaines, un rappel des faits essentiels ne paraît pas superflu.

Ces controverses me touchent particulièrement car j'évolue dans un contexte pluridisciplinaire qui génère des signaux discordants : d'une part, les Études Anglophones, où les

19. Dans le domaine de la linguistique, je note cependant que le manuel d'introduction récent de Winter (2019, pp. 171-178) passe davantage de temps sur les problèmes liés à l'inférence statistique que son aîné de plus de 10 ans, Baayen (2008).

méthodes quantitatives restent souvent élémentaires, peu utilisées et parfois mal perçues. D'autre part, la culture scientifique en psychologie et neurosciences, où les *p-values* sont des critères de décision jugés indispensables ; je ne crois pas exagérer en parlant de rituel (Gigerenzer, 2004) ou de véritable culte (pour reprendre le titre de Ziliak et McCloskey, 2008). Et enfin, le domaine des sciences et technologie de l'information et de la communication, où on se concentre davantage sur l'estimation des paramètres et la capacité de généralisation d'un modèle, plutôt que sur des questions de « significativité ».

### 2.4.1 Une routine méthodologique surestimée

La valeur de probabilité chiffrée, précise, renvoyée par un test d'inférence statistique véhicule une impression de scientificité qui, historiquement, a été valorisée pour légitimer le statut de certaines sciences sociales (Ziliak et McCloskey, 2008), et asseoir la crédibilité de leurs résultats. Ce schéma s'applique désormais aux domaines des sciences du langage. Ce n'est pas tant la méthode elle-même qui est critiquable que l'utilisation qui en est souvent faite. Il serait trop long de récapituler tous les arguments soutenus par les détracteurs du test d'hypothèse — un bon nombre est compilé dans Nickerson (2000) ; voir aussi par ex. Amrhein *et al.* (2019) pour un rappel condensé, et Roettger (2019) et Roettger *et al.* (2019)[20] pour une application au domaine de la phonétique — mais je souhaiterais m'arrêter sur trois aspects particuliers :

1. Ce qu'on nomme « fallacy of the transposed conditional », le fait de confondre $P(D|H)$, $D$ sachant $H$, avec $P(H|D)$, $H$ sachant $D$ ;
2. Le caractère binaire, automatique et institutionnalisé des décisions qui découlent de cette routine et leurs conséquences sur nos pratiques ;
3. La question de la reproductibilité, sur laquelle un test d'hypothèse nulle ne nous dit rien.

Le principe de ce test consiste à poser une hypothèse dite « nulle » ($H_0$) stipulant qu'il n'y a pas de différence (ou pas de corrélation) entre plusieurs groupes pour une variable donnée, et à déterminer la probabilité d'obtenir des données au moins aussi extrêmes que celles dont on dispose si cette hypothèse nulle est vraie. Une valeur de probabilité très faible conduit à rejeter l'hypothèse nulle. Ce test renvoie donc $P(D|H_0)$, la probabilité d'obtenir $D$ (nos données) étant donné que $H_0$ est vraie. Or ce que nous

20. Je découvre ces deux articles bien après avoir écrit l'essentiel de cette sous-section. Je me réjouis de voir que les mises en gardes méthodologiques sur l'usage des tests statistiques sont désormais disponibles pour un lectorat de phonologues et phonéticiens.

voudrions véritablement connaître, c'est la probabilité que notre hypothèse soit vraie étant donné nos données : $P(H_0|D)$. Passer de l'une à l'autre de ces deux probabilités conditionnelles n'est pas chose aisée, puisque cela implique, d'après le théorème de Bayes (Équation 2.1), de connaître $P(H_0)$, la probabilité que l'hypothèse nulle soit vraie, qui n'est généralement pas connue, et très souvent proche de zéro. L'une des critiques les plus fréquentes consiste donc à dire que le test ne répond pas à la question que l'on se pose ; et l'une des erreurs les plus fréquentes consiste à penser que $P(D|H_0) = P(H_0|D)$.

$$P(H_0|D) = \frac{P(D|H_0)P(H_0)}{P(D|H_0)P(H_0) + P(D|\neg H_0)P(\neg H_0)} \tag{2.1}$$

Une autre critique récurrente du test de l'hypothèse nulle consiste à pointer du doigt le caractère binaire de la décision qu'il occasionne (Nickerson, 2000 ; Wasserstein et Lazar, 2016 ; McShane *et al.*, 2019). En effet, si l'on prend par exemple le seuil de signification conventionnel de 5 %, la doctrine nous enjoint à ignorer $p = 0,051$ et à nous réjouir du résultat significatif reflété dans $p = 0,049$. Bien souvent, une différence aussi ténue ne justifie pas le caractère dichotomique de la décision qu'on en tire. Mais il faut tout de même reconnaître qu'il est plus facile de raisonner avec des seuils fermes ; et certains auteurs voient la logique binaire du test d'hypothèse comme un avantage (Greenwald *et al.*, 1996). Si je condamne ce binarisme, il est cependant bien difficile d'aller à l'encontre d'une procédure institutionnalisée, comme le montre le Tableau 2.1 issu de Ferragne *et al.* (2010). En effet, à l'époque-même où ma position vis-à-vis du test d'hypothèse était la plus virulente, je publiais pourtant ce tableau récapitulatif avec des 1 et des 0 matérialisant le caractère binairement significatif du test en question (appliqué aux phénomènes de *Scottish Vowel Length Rule* que je détaille à la Section 4.4.1).

Au-delà de la mise en évidence du triomphe de la routine institutionnalisée sur un individu en désaccord avec elle, le Tableau 2.1 illustre une partie du titre de ce document de synthèse : « Du groupe à l'individu ». La nécessité de constituer des groupes, des classes, dans le but de faire des analyses quantitatives est indéniable. Néanmoins, je crois qu'il faut toujours se laisser la possibilité de revenir à l'individu et de questionner son appartenance au groupe. S'il y a bien une constante dans mes travaux, c'est ce besoin d'estimer dans quelle proportion un individu représente le groupe auquel il est censé appartenir. Dans ma thèse (Ferragne, 2008, p. 338 *sqq.*), j'avais défini une fonction d'appartenance d'un locuteur à l'accent dont il constituait un « exemplaire » en m'inspirant de la théorie des ensembles flous. Je n'ai jamais cessé d'utiliser ce concept en filigrane de tous

TABLEAU 2.1 – Résultats des t-tests pour chaque individu. 1 : on rejette l'hypothèse nulle d'égalité des moyennes entre la version suffixée et non suffixée du timbre en question (/u/ ou /ai/) sur le paramètre concerné (durée ou variation de timbre). 0 : on ne peut pas rejeter l'hypothèse nulle. Alpha est fixé à $\alpha = 0,01$.

| Sujet | /u/ | | /ai/ | |
|---|---|---|---|---|
| | durée | timbre | durée | timbre |
| gla01 | 1 | 0 | 1 | 0 |
| gla03 | 1 | 0 | 1 | 1 |
| gla04 | 1 | 0 | 1 | 1 |
| gla05 | 1 | 0 | 1 | 1 |
| gla06 | 1 | 0 | 0 | 0 |
| gla07 | 1 | 0 | 1 | 0 |
| gla08 | 1 | 0 | 1 | 0 |
| gla10 | 1 | 0 | 1 | 0 |
| gla11 | 1 | 0 | 1 | 0 |
| gla12 | 1 | 0 | 1 | 1 |
| gla13 | 1 | 0 | 1 | 0 |
| gla14 | 1 | 0 | 1 | 0 |

mes travaux. Le tirage aléatoire strict, le seul qui garantirait une estimation statistique totalement viable, est rarement possible lorsqu'on recueille des données linguistiques. Il faut donc parfois mobiliser les connaissances de l'expert en amont pour procéder à un échantillonnage « dirigé », et ainsi substituer au descriptivisme biaisé, un prescriptivisme nécessaire et assumé.

Pour en revenir au caractère binaire du test d'hypothèse, si l'obtention de valeurs significatives constitue le critère déterminant pour publier une étude, et puisque certains avancements de carrière et financements sont très largement conditionnés par le nombre de publications, les chercheurs ont tout intérêt à obtenir des résultats significatifs. Cette pression donne lieu à la pratique du *p-hacking* (Head *et al.*, 2015), qui consiste à maximiser ses chances d'obtenir des valeurs significatives, par exemple en menant des analyses statistiques tout en collectant des données avec l'intention de continuer la collecte jusqu'à ce que le seuil de signification des tests soit atteint. Il ne s'agit bien souvent pas d'une démarche délibérément malhonnête. Simonsohn *et al.* (2014) expliquent que si les décisions qui concernent l'analyse (données supplémentaires, rejet de valeurs déviantes, etc.) sont prises pendant l'analyse, et non avant, il y a un risque pour que les chercheurs suivent des options qui favorisent leur probabilité de publier, même inconsciemment (Nuzzo, 2014).

### 2.4.2 Quelles solutions ?

Les débats entre statisticiens font apparaître de nombreux désaccords, mais une solution semble faire l'unanimité : pour savoir si un résultat n'est pas simplement un faux-positif, il faut répliquer l'étude (Cohen, 1994 ; Nickerson, 2000). Cohen (1994, p. 1002) est d'ailleurs particulièrement cynique :

> we have a body of statistical techniques, that, used intelligently, can facilitate our efforts. But given the problems of statistical induction, we must finally rely, as have the older sciences, on replication.

Cet avis est partagé par des auteurs d'autres domaines ; par exemple, pour l'analyse des potentiels évoqués en électroencéphalographie, Luck (2014, p. 310) pense que répliquer une étude constitue la meilleure méthode statistique existante. Popper (2002b, p. 66) insiste lui aussi sur cet aspect :

> a few stray basic statements contradicting a theory will hardly induce us to reject it as falsified. We shall take it as falsified only if we discover a *reproducible effect* which refutes the theory.

Malheureusement, les réplications à l'identique sont infiniment rares, et un article proposant une telle étude présenterait un risque important de susciter des commentaires négatifs de la part des relecteurs et des éditeurs (Nickerson, 2000). Un consortium de chercheurs a néanmoins eu l'idée de lancer un projet de grande envergure qui consistait à tenter de répliquer 100 études publiées en 2008 dans trois revues de psychologie expérimentale (Open Science Collaboration, 2015). Cette démarche s'est faite en lien étroit avec les auteurs originaux, afin de garantir une similitude maximale entre les deux versions de la même étude. Les résultats font apparaître que 97 % des études originales affichaient un résultat significatif au seuil $p < 0,05$. Après réplication, seules 35 études ont atteint ce seuil. De plus, la taille moyenne des effets entre étude originale et réplication a été divisée par deux. Ce projet souligne donc clairement qu'une étude non répliquée, quel que soit le niveau de significativité qu'elle affiche, n'apporte finalement qu'une preuve faible et éphémère, ce qui devrait inciter à davantage de discernement, mais aussi à valoriser les réplications.

Afin d'éviter le *p-hacking*, une solution, qui pourrait se développer dans les années à venir à condition que la surcharge bureaucratique occasionnée ne soit pas dissuasive, viendra peut-être de la pratique encore peu utilisée dans nos domaines du pré-enregistrement (Kupferschmidt, 2018 ; Warren, 2018). Cela consiste à enregistrer très précisément sur des

sites dédiés[21] le détail de l'étude que l'on souhaite mener en renseignant les hypothèses, le nombre de participants, les variables étudiées, etc.

À titre personnel, le mode opératoire propre au *machine learning* (Duda *et al.*, 2001), qui valorise plutôt le potentiel prédictif des modèles, me convainc davantage que le test de l'hypothèse nulle. En ajustant dans un premier temps un modèle à des données, et en le validant ensuite sur des données qui n'ont pas été vues dans la première phase, j'ai la sensation que les généralisations obtenues sont plus « tangibles ». J'ai particulièrement ressenti cet effet dans ma pratique intensive du *deep learning* ces trois dernières années.

### 2.4.3 Pour une culture du graphique

« A picture is worth a thousand $p$ values », nous dit le titre de Loftus (1993). À l'époque de ma thèse, j'avais déjà la conviction que de bonnes représentations graphiques valaient mieux que de longs discours et, dans de nombreux cas, bien mieux que des équations ou des valeurs de probabilité. Le manuscrit était donc abondamment illustré, et j'avais apporté un soin particulier à la réalisation des figures. Dans les années qui ont suivi, de nombreux étudiants ont pu constater mon attitude quasi obsessionnelle à l'égard de la qualité des figures, et ce trait de caractère perdure.

Le graphique est un élément central de la communication scientifique qui devrait faire l'objet d'un enseignement spécifique auprès des étudiants. C'est par ailleurs un formidable outil d'émancipation pour les chercheurs car il repose sur un langage universel, certes codifié, mais dont la sémantique ne requiert pas la sophistication, parfois cabalistique et donc ostracisante, de certaines méthodes quantitatives. J'ai en outre acquis la conviction qu'un graphique bien réalisé est plus informatif que n'importe quel autre outil statistique. C'est une nouvelle fois le psychologue-statisticien Jacob Cohen qui résume bien ma position (Cohen, 1990, p. 1305) :

> We sometimes learn more from what we see than from what we compute; sometimes what we learn from what we see is that we shouldn't compute, at least not on those data as they stand.

Les principes d'excellence graphique de Tufte (2001) ont attiré mon attention. Je n'adhère pas à l'intégralité de ses préceptes, où le principe d'économie atteint parfois des extrêmes, mais je salue le raisonnement de l'auteur, qui conçoit l'art de créer des graphiques comme une discipline académique à part entière. Mon expérience de l'encadrement

21. Par exemple https://osf.io/.

d'étudiants m'a fait réaliser que ces derniers n'accordent souvent pas aux graphiques la réflexion qu'ils méritent. C'est dommageable à double titre : d'abord, parce qu'ils se privent d'un outil de communication inégalable, et ensuite, parce qu'ils font l'impasse sur un pan de notre métier où la créativité peut s'exprimer librement. J'espère que les graphiques qui agrémentent ce document reflètent fidèlement ces principes.

## 2.5 Conclusion

Dans ce chapitre, j'ai ébauché une discussion de divers points épistémologiques et méthodologiques. Je souhaitais mettre en avant le fait que toute connaissance scientifique se construit à l'intérieur d'un contexte précis et qu'une partie de ce contexte est déterminée par les préférences de l'individu qui mène la recherche en question. Ce chapitre et le suivant (Chapitre 3) ont donc pour vocation d'expliciter ces préférences afin que la revue des travaux présentée dans les Chapitres 4 à 6 soit précisément contextualisée. J'ai voulu tempérer l'enthousiasme qui entoure le concept de « corpus » en montrant ses limites et en illustrant rapidement des alternatives qui me semblent souvent mieux adaptées. J'ai souhaité également rappeler que la taille des données et le tirage aléatoire des participants continuent d'être les meilleurs garants de généralisations valides. Puisque le temps et l'argent sont des contraintes parfois strictes, un corpus constitué pour l'occasion en maximisant les occurrences du phénomène étudié et/ou une expérience de perception (qui d'ailleurs peut être réalisée à distance) sont des solutions optimales. L'introspection et l'intuition sont des formes de connaissance qui ont toute leur place dans le processus de recherche, à condition de pouvoir les questionner avec méthode, ce que s'emploie à faire la psycholinguistique. Ce chapitre se conclut par un rappel des risques liés à l'usage institutionnalisé souvent déraisonnable des tests statistiques qui conduisent à décider automatiquement, et de façon binaire, si des résultats méritent d'être publiés. Dans un monde où l'investissement scientifique est rétribué en monnaie de *p-values*, la prise de risque et la créativité pourraient ne plus avoir leur place. Une culture de la prédiction, de la réplication et du graphique pourrait néanmoins compenser les abus liés à la religion (Salsburg, 1985) du test d'hypothèse.

# 3

# La technologie comme moteur scientifique

## Sommaire



## 3.1 Introduction

Ce chapitre retrace quelques éléments représentatifs de ma démarche vis-à-vis des aspects technologiques impliqués dans ma recherche. Je suis convaincu qu'une compréhension détaillée du volet technique est absolument fondamentale dans notre activité quotidienne ; lui consacrer un espace dédié n'est donc pas superflu. Je trouve ce chapitre utile car si le cadre épistémologique, comme mentionné au Chapitre 2, définit en grande partie l'identité d'un chercheur, je souhaite mettre en avant que les outils qu'il utilise contribuent également à mieux cerner cette identité.

Je crois que le succès d'un projet est en partie lié à notre connaissance des aspects technologiques impliqués. Une caractéristique singulière de ma pratique tient au fait que je suis en quelque sorte mon propre ingénieur. C'est arrivé par nécessité, certes, mais également par goût. Et puisque ma formation initiale est « littéraire », et donc mal adaptée à ce type d'ambition, cet investissement personnel a été très chronophage, et a connu des fortunes diverses. Mais certains succès gratifiants, et notamment ma capacité à épauler mes étudiants sur une palette très variée de méthodologies et de techniques, m'encouragent à maintenir le cap.

Avec vingt années de pratique de la recherche, ma démarche, qui ne conçoit pas de savoir sans savoir-faire, est restée constante. L'abbé Rousselot (Rousselot, 1897) était un modèle du genre, faisant cohabiter, dans un contraste saisissant, des objets d'étude qu'on n'associe pas à la modernité (le patois de sa famille) avec les dispositifs expérimentaux les plus innovants. Plus près de nous, je trouvais des profils comme ceux de Mark Huckvale et René Carré très inspirants : travailler à partir d'interfaces logicielles qu'on développe soi-même (ce que René a fait jusqu'à la fin de sa carrière), voilà des exemples qui m'ont stimulé. Et je dois ajouter à cette liste (qui restera très incomplète) un de mes contemporains, Volker Dellwo, avec son corpus BonnTempo (Dellwo *et al.*, 2004). Contrairement aux autres noms énumérés, quand je l'ai rencontré dans la première moitié des années 2000, moi qui travaillais également sur le rythme à cette époque, j'avais toutes les compétences techniques pour faire la même chose : un corpus constitué pour répondre à une question précise (l'interaction rythme-débit), dans plusieurs langues, parfaitement intégré à `Praat` et interrogeable via une interface graphique de `Praat` développée pour l'occasion. Ce que j'admirais là, en plus du caractère *ad hoc*, c'était l'aspect « produit fini » ; on pouvait télécharger la base de données et `Praat`, et répliquer les calculs. C'était de la science ouverte bien avant que l'idée ne s'impose, quelque vingt ans plus tard, comme un critère de qualité incontournable.

Une anecdote célèbre raconte que Daniel Jones s'apprêtant à effectuer un travail de terrain avait déclaré n'avoir besoin d'emporter que ses oreilles (Ladefoged, 2003, p. 27). Rousselot disait au sujet de l'oreille (Rousselot, 1897, pp. 34-35) :

> nous rechercherons les moyens de corriger, de compléter les données qu'elle fournit, mais nous ne trouverons point celui de nous en passer. Quand l'oreille se reconnaît impuissante, il faut bien la suppléer ; mais, dans les cas où elle suffit (et ils sont nombreux), nul moyen d'expérimentation n'est aussi rapide ni aussi commode.

Il y a certainement un peu de vrai dans ce que disent Jones et Rousselot, mais force est de constater que l'un comme l'autre sont bien loin de s'être contentés de ce qu'ils

entendaient[22]. D'ailleurs, l'histoire de la phonétique est très intimement liée aux développements technologiques qui l'ont jalonnée (Boë et Vilain, 2010). Par ailleurs, pour des raisons évidentes de reproductibilité, il serait bien délicat aujourd'hui de présenter les résultats d'une recherche ne s'appuyant que sur l'acuité auditive du chercheur qui l'a produite.

Ce chapitre s'articule en deux temps : je propose d'abord une esquisse de réflexion sur la technologie dans la recherche en présentant ce que j'appellerai mon « écosystème ». Ensuite, je mentionne quelques exemples choisis de réalisations techniques. Outre la démonstration d'un investissement marqué dans les technologies pour la recherche, les exemples de cette seconde partie de chapitre visent à illustrer mon implication dans la formation *par* la recherche. En effet, dès mon arrivée à l'UFR d'Études Anglophones en 2009, j'ai intégré dans mes cours de Master et de L3 des éléments scientifiques et techniques issus de mes propres travaux, sans exclure les aspects les plus techniques. J'ai également, dès mon recrutement, « réorienté » un séminaire de Master pour en faire le premier cours d'analyse quantitative à l'UFR en m'appuyant sur le logiciel `R`. Je renforce donc ma pratique pédagogique avec un savoir-faire technique destiné à la fois à mieux préparer les étudiants à la recherche, et à les intéresser à mon domaine à partir de manipulations très concrètes et ludiques.

## 3.2 Mon écosystème

L'autonomie technologique est un objectif très utile dans notre métier. Tendre vers cet idéal implique, je crois, d'investir dans des technologies qui permettent une personnalisation avancée des méthodes que nous employons. C'est pour cela que j'ai toujours milité auprès des étudiants et des collègues en faveur de solutions logicielles comportant un langage de script plutôt qu'une interface graphique seule. Concrètement, je privilégie en particulier `Praat` et `R` au détriment des autres solutions équivalentes en apparence. Mon environnement de prédilection reste le logiciel `Matlab`. J'ajoute que, `Python` étant devenu le langage de référence dans le domaine du *deep learning*, j'y ai de plus en plus souvent recours bien que mes connaissances en la matière soient plus limitées.

Il m'arrive ponctuellement d'aborder par obligation des langages de plus bas niveau (`C`, `C++`, `C#`, etc.) ; mais je dois bien reconnaître que ces épisodes, qui me font basculer

22. La biographie de Jones (Collins et Mees, 1999) fait clairement apparaître son intérêt pour l'instrumentation ; quant à Rousselot, l'instrumentation est totalement indissociable de son parcours (Rousselot, 1897).

dans le domaine du développement et du déploiement d'applications, font rapidement apparaître les limites de ce qui me paraît être mon métier. Néanmoins, s'intéresser à des technologies plus proches du niveau de la machine présente un intérêt pédagogique indéniable : mes expériences occasionnelles avec des langages de bas niveau, ou encore la réalisation de quelques montages électroniques, m'ont fait prendre conscience de la complexité des tâches sous-jacentes que les langages de haut niveau (ou a fortiori les programmes compilés) nous épargnent.

Pour ce qui est du choix du matériel, je privilégie également les solutions autorisant une personnalisation avancée en anticipant par exemple, autant que faire se peut, la possibilité d'intégrer le matériel en question à un système d'acquisition (ou de stimulation) plus vaste. À titre d'exemple, le système d'électroglottographie EG2-PCX de Glottal Enterprises comporte des sorties analogiques ; il est donc peu coûteux de relier ces sorties à une carte d'acquisition pour obtenir des signaux synchronisés avec ceux d'un autre appareil. En plus du choix de la personnalisation — *hackability* est le terme anglais qui conviendrait, j'essaie d'avoir une démarche privilégiant le *rightsizing* ; quand cela est possible, je me tourne vers des équipements à bas coût, comme par exemple la plateforme d'acquisition électrophysiologique BITalino[23] ou encore l'ordinateur low-cost Raspberry Pi[24].

Dans la période qui a suivi la thèse, j'ai particulièrement investi dans l'apprentissage et dans l'acquisition de nouveaux matériels. Cet élan était probablement imputable à mon envie déjà quasi obsessionnelle de travailler avec les technologies les plus récentes, à la frustration d'avoir dû me plier aux imperfections d'un corpus audio que je n'avais pas recueilli moi-même, et à cette fameuse autonomie vers laquelle je tends. Et ce besoin de technologies de pointe est non seulement motivé par un goût personnel, mais également par mon désir affirmé de pouvoir proposer à mes étudiants les outils les plus actuels. J'espère aussi, ce faisant, les motiver à faire de la phonétique grâce au caractère concret et ludique associé à l'utilisation de matériel.

Ainsi, au fil des ans, en plus du signal audio et des réponses comportementales à des expériences de perception, la diversité des données que j'ai pu collecter et analyser n'a fait que croître. Cela a commencé par l'électroencéphalographie (Boulenger *et al.*, 2011 ; Pota *et al.*, 2012 ; Bedoin *et al.*, 2019 ; Heidlmayr *et al.*, 2021 ; Pélissier et Ferragne, 2021) et l'échographie de la langue (King et Ferragne, 2020) dès la fin des années 2000. Ce sont ensuite des techniques comme l'électroglottographie[25], l'analyse d'images (King

23. https://bitalino.com/.
24. https://www.raspberrypi.org.
25. Employée actuellement dans la recherche de Léa Burin et de Coline Caillol.

et Ferragne, 2021) ou de spectrogrammes (Ferragne *et al.*, 2019) via le *deep learning* ou encore la conductance électrodermale (Rastovic *et al.*, 2019) que j'ai d'abord utilisées moi-même pour ensuite les mettre à la disposition des étudiants.

## 3.3 Quelques exemples de réalisations techniques

### 3.3.1 Le logiciel `ROCme!`

La collecte d'un corpus audio comporte une phase de stimulation — présentation d'un texte, d'une image, d'une consigne, etc. — et une phase d'acquisition. L'acquisition concerne le recueil de données audio brutes ainsi que de métadonnées sur les locuteurs. Dans la pratique en linguistique, ce sont souvent trois supports différents qui gèrent la stimulation, l'acquisition audio et la collecte de métadonnées ; par exemple, respectivement, un texte imprimé sur papier, un enregistreur et un fichier (tangible ou numérique). La pérennité du lien entre ces trois supports est conditionnée en grande partie par la méticulosité de l'investigateur. De plus, l'homogénéité des normes de stockage n'est pas garantie : des enregistrements successifs peuvent être effectués par mégarde en mono, puis en stéréo, avec des taux d'échantillonnages différents ; les métadonnées sont susceptibles d'être dans des formats différents et nécessitent un traitement a posteriori pour être exploitables par un ordinateur.

Afin de rationaliser la collecte de corpus audio, dans un environnement totalement dématérialisé, j'ai lancé en 2010 le développement du logiciel *Recording of Oral Corpora Made Easy* (`ROCme!`). Deux financements (symboliques) du bonus qualité recherche de l'Université Paris Diderot ont permis d'initier le processus. J'ai constitué une équipe avec deux informaticiens du laboratoire Dynamique Du Langage. C'est une aventure particulièrement valorisante, au cours de laquelle j'ai acquis une somme de connaissances que je ne soupçonnais pas au départ.

Dans le cahier des charges, je tenais absolument à offrir à l'utilisateur une interface conviviale, différente de nombreux logiciels universitaires caractérisés par une austérité décourageante. Toujours pour des raisons ergonomiques, je tenais également à ce que l'installation soit la plus transparente possible pour l'utilisateur. Enfin, la portabilité du logiciel (initialement entre Windows, Apple et Linux) était un critère important. J'ai en effet plusieurs expériences de partages de scripts, commandes et interfaces graphiques développées surtout avec `Praat` ou `Matlab`, qui posaient souvent problème au moment du déploiement. Par exemple, le partage d'une interface graphique écrite avec `Matlab`

implique l'installation sur la machine cible du moteur d'exécution de `Matlab`. Or celui-ci est très lent au démarrage et consomme beaucoup de ressources. De plus, les exécutables développés dans ce contexte nécessitent une parfaite correspondance entre la version du moteur d'exécution et la version de `Matlab` qui a servi à compiler le programme. La portabilité est également très limitée. Afin de nous affranchir de ces contraintes, nous nous sommes tournés vers les technologies s'appuyant sur le moteur d'exécution `Adobe Air`.

Le logiciel `ROCme!` permet donc de présenter différents types de stimuli — texte, images, vidéos, page HTML — disposés sur des diapositives différentes, et de recueillir les données audio élicitées par ces différents médias. Un nouveau fichier audio est enregistré à chaque changement de diapositive, ce qui permet déjà une première segmentation du corpus. Il est possible d'inclure dans le logiciel un questionnaire visant à recueillir des métadonnées, qui seront archivées au format XML. `ROCme!` dispose en outre d'un module permettant de visualiser sous forme graphique les métadonnées de plusieurs participants, et de les exporter pour un traitement ultérieur dans un tableur, par exemple. Les locuteurs peuvent gérer leur enregistrement de façon totalement autonome.

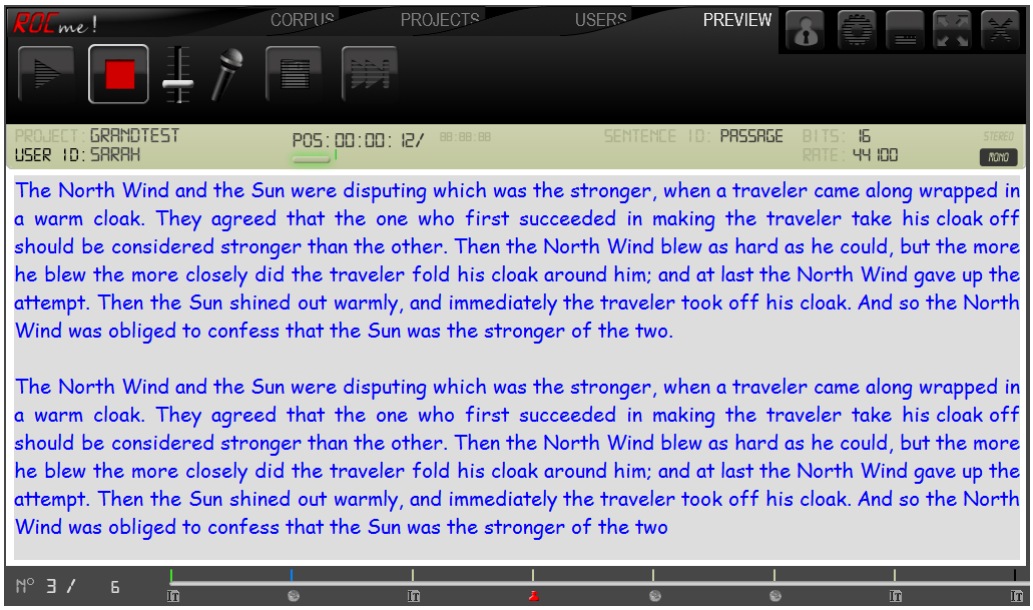



FIGURE 3.1 – Le logiciel `ROCme!` en action.

La philosophie du logiciel gravite autour de la notion de projet : l'utilisateur définit une fois pour toutes un ensemble de contraintes qui seront appliquées à tous les enregistrements d'un projet. Ces paramètres incluent le format audio, le format des noms de fichiers, l'ordre de présentation des stimuli, la présence d'un questionnaire (avec champs obligatoires ou

non), etc. Le projet garantit une cohérence parfaite.

Il serait fastidieux de dresser ici la liste des possibilités du logiciel, qui n'a cessé de croître, en s'adaptant autant que faire se peut aux demandes spécifiques d'utilisateurs, et qui est disponible dans l'aide.

Après une période active de développement et de maintenance dans la première moitié des années 2010, mais en l'absence de volonté politique forte, le projet a fini par être abandonné et le logiciel n'est plus maintenu. À l'heure où j'écris ces lignes, des étudiants continuent pourtant de l'utiliser, et la crise sanitaire l'a rendu plus utile que jamais. La pérennisation de ce type d'outil est délicate puisqu'elle dépend en grande partie de la bonne volonté de l'équipe. En cela, j'admire la longévité du logiciel `Praat`.

### 3.3.2 CminR Praatik

L'optimisation méthodologique est une constante de mon parcours. À chaque nouveau projet, j'accorde en effet un temps considérable à la mise en place d'un scénario d'analyses de données qui soit ergonomique, reproductible, et qui permette un gain de temps tangible. Ces objectifs sont atteints grâce à l'automatisation de certaines procédures et au développement d'interfaces graphiques à destination des étudiants ou des collègues. Ainsi, chaque travail d'analyse nécessite l'élaboration minutieuse d'un *workflow* adapté.

J'ai pu constater que les étudiants en SHS, quand ils sont confrontés au traitement de données quantitatives, pâtissent d'un manque de formation. Ainsi, à chaque nouveau mémoire de Master en phonétique, les étudiants sont souvent contraints d'improviser leur propre *workflow*, de ré-inventer ce que d'autres ont déjà mis en place, parfois au prix d'une perte de temps démesurée, et pour un résultat qui ne reflète pas leur investissement. J'ai donc pensé que je pouvais modestement pallier quelques lacunes dans la formation typique des étudiants en proposant un nouveau séminaire. Le CminR Praatik a consisté en une série de 6 séances de 2 heures consacrées au *formant workflow*, c'est-à-dire à un scénario unifié pour la mesure, la représentation graphique et l'analyse des formants vocaliques.

La première séance a été dédiée à l'estimation de formants au moyen d'interfaces graphiques que j'ai créées pour l'occasion avec le `Demo Window` du logiciel `Praat`[26]. Le défi majeur dans le développement de ces interfaces réside davantage dans la prise de recul sur le plan ergonomique que dans la difficulté de programmation.

Les trois séances suivantes ont été consacrées à l'analyse et la représentation graphique des valeurs formantiques avec le logiciel `R`. Diverses variantes des représentations

26. Programmes disponibles à l'adresse : https://tinyurl.com/snvz387u.

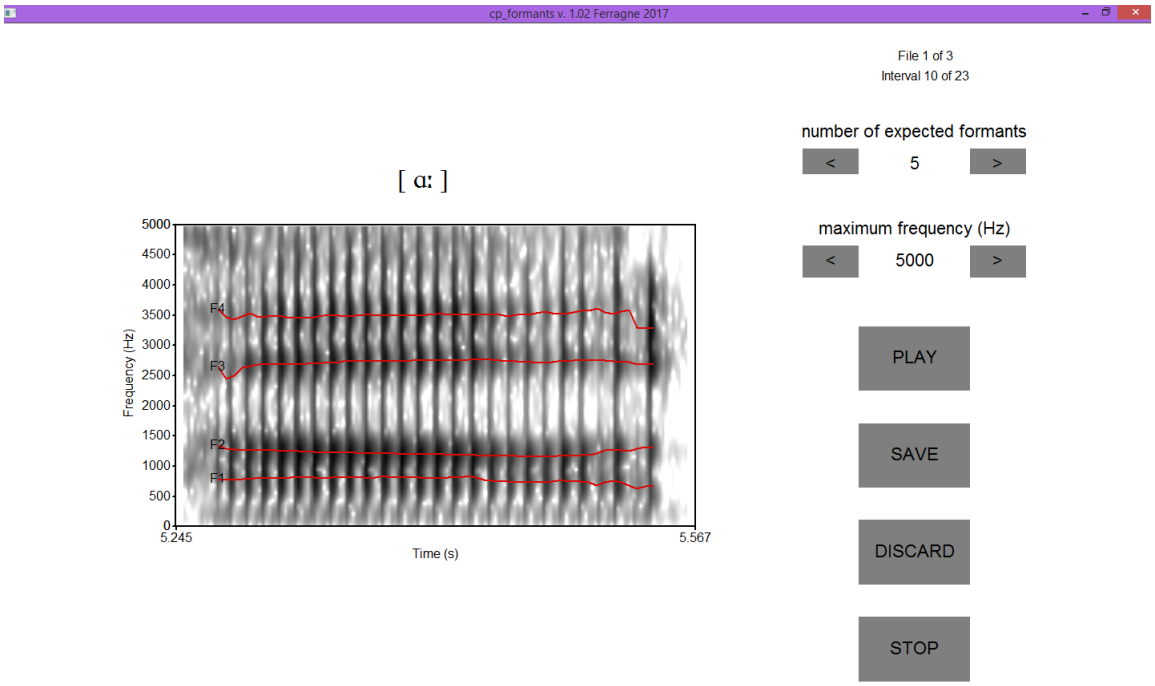



FIGURE 3.2 – Interface principale du programme `cp_formants`.

bidimensionnelles classiques ont été présentées, ainsi que des représentations dynamiques. Ces dernières ont nécessité une préparation très longue : en effet, dans un souci de lisibilité du code et d'élégance algorithmique, je me suis interdit, autant que possible, de présenter à mon auditoire certaines recettes « bricolées » que j'utilise moi-même. Par exemple, j'ai opté pour le stockage sous forme de listes dans `R` des trajectoires formantiques, et je me suis efforcé d'appliquer à ces listes des fonctions de pseudo-vectorisation plutôt que des boucles.

Ces trois séances ont également été ponctuées par des démonstrations au moyen d'interfaces graphiques créées avec `Matlab` pour l'occasion, qui visaient à approfondir certains aspects comme la transformée en cosinus discrète ou le *dynamic time warping*. Les deux dernières séances ont été consacrées à la programmation avec `Praat`. Connaissant très bien le langage de script et la métalangue pour décrire son fonctionnement, c'est plutôt l'angle pédagogique qui a nécessité un investissement particulier.

Le CminR Praatik a donc nécessité plusieurs semaines de préparation, ce qui n'aurait sans doute pas été possible sans ma délégation IUF. Les retours anonymes que j'ai systématiquement sollicités après chaque séance font apparaître un enthousiasme marqué pour cette formation, et m'ont permis d'ajuster au fur et à mesure ma pratique pédagogique. Les programmes et supports mis au point pour cette formation ont été progressivement

inclus dans les séminaires de Master que je dispense, et sont mis à la disposition des masterants, doctorants et collègues enseignants-chercheurs au gré des diverses formations ponctuelles qu'il m'arrive de donner à la demande ou de services que je rends très volontiers depuis toujours.

Comme en témoignent les liens hypertextes vers mon Github, qu'on trouve entre autres dans les notes de bas de page de cette section, je partage volontiers mes réalisations. Néanmoins, les 18 ans que je viens de passer à former des collègues et des étudiants à l'utilisation d'outils comme `Praat`, `R` ou `Matlab` m'ont appris qu'il ne fallait partager que ce qui était en état de l'être. Par exemple, j'ai écrit des centaines de scripts dans tous ces langages de programmation, mais l'immense majorité ne sont pas assez génériques pour être partagés sans que cela n'occasionne un travail important de la part de l'utilisateur final (par ex. pour comprendre la structure des données, pour passer outre les sempiternelles erreurs liées au mauvais noms de répertoires, etc.). C'est d'ailleurs dans la nature même d'un « script » d'être un brouillon, par rapport à une « fonction », dans la terminologie de `Matlab` : la seconde est une version générique du premier, mise au propre et commentée. Cela peut paraître anodin, mais j'ai constamment cette dichotomie à l'esprit lorsque je m'apprête à partager un programme ; et l'expérience m'a donc conduit à éviter, si je n'ai pas la possibilité d'interagir en personne avec l'utilisateur final, de mettre en téléchargement libre toutes mes réalisations informatiques.

### 3.3.3 `ET VOYLA !`

À l'occasion du CminR Praatik, j'ai développé plusieurs petits programmes ; en voici un à qui j'ai donné récemment une nouvelle vie à la suite des cours et séminaires que j'ai dispensés sur l'analyse de voyelles, et plus généralement sur le traitement de données et la programmation informatique. En effet, tous les étudiants de Master de linguistique anglaise ne vont pas faire de la recherche leur métier, et tous ne vont pas faire de la phonétique. Certains voudront peut-être ponctuellement être en mesure de représenter des voyelles dans l'espace des 2 premiers formants en évitant les quelques lignes de commande de rigueur, et l'inévitable fichier qu'on va chercher là où il n'est pas.

C'est pour laisser à ces étudiants la possibilité de réaliser simplement des espaces vocaliques que j'ai développé `ET VOYLA !`, une petite interface graphique compilée avec `Matlab`[27]. Elle souffre, comme tous ces exécutables autonomes créés avec `Matlab`, de lenteurs très caractéristiques au démarrage. Mais elle permet, en un clic, de réaliser entre

27. https://tinyurl.com/2jk8y3za.

autres les illustrations présentées dans les Figures 3.3 et 3.4, qui, d'expérience, suscitent l'enthousiasme des étudiants. Ces illustrations montrent les voyelles de l'étude de Peterson et Barney (1952) représentées dans l'espace des deux premiers formants vocaliques, d'abord sous forme de nuages de points (panneau gauche de la Figure 3.3), puis de valeurs moyennes par phonème avec carte de chaleur superposée représentant, en rouge, les zones les plus densément peuplées (panneau droit de la Figure 3.3). La Figure 3.4 montre à gauche le même ensemble de données sous forme d'histogramme puis, à droite, en version lissée, sous forme de densités de probabilité.

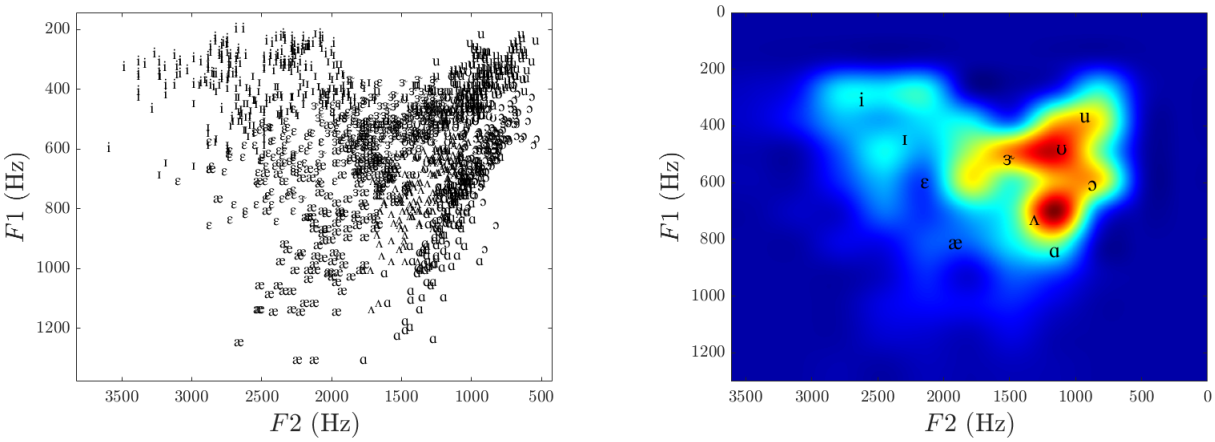



FIGURE 3.3 – Système vocalique de l'anglais américain vu par `ET VOYLA !`.

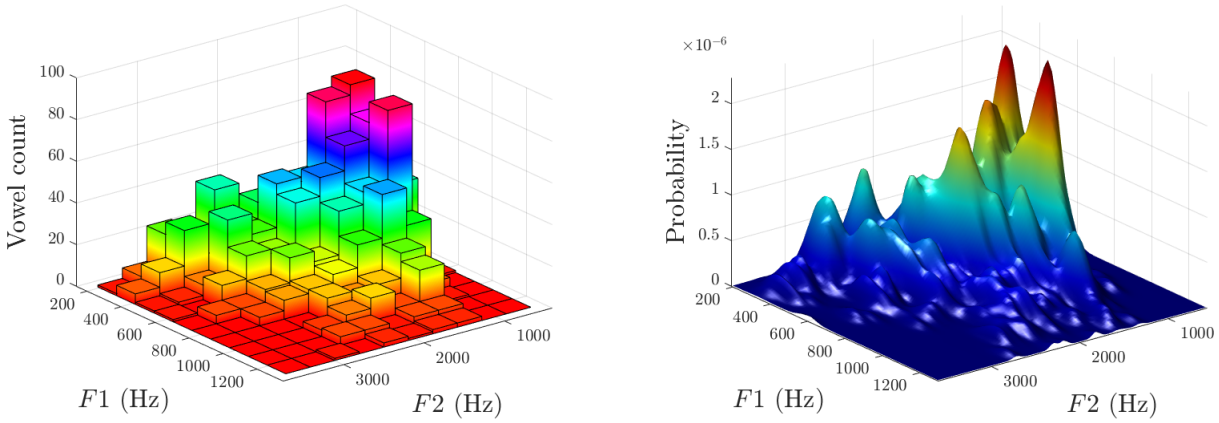



FIGURE 3.4 – Système vocalique de l'anglais américain en 3D vu par `ET VOYLA !`.

### 3.3.4 Le cartogramme vocalique

Parmi les réalisations techniques qui ont demandé un investissement colossal, je dois évoquer mon cartogramme vocalique. Je mentionne brièvement le principe ici en laissant une explication détaillée pour une éventuelle publication ultérieure.

Les cartogrammes sont des cartes géographiques dans lesquelles l'aire des différents pays (ou régions, départements, etc.) a été modifiée proportionnellement à une variable non spatiale comme par exemple, la densité de population, le taux de précipitation ou encore le revenu. Un cartogramme typique est présenté à la Figure 3.5, où j'ai déformé les régions de France métropolitaine de sorte que leur aire reflète le nombre d'hospitalisations enregistrées au 7 avril 2020[28].

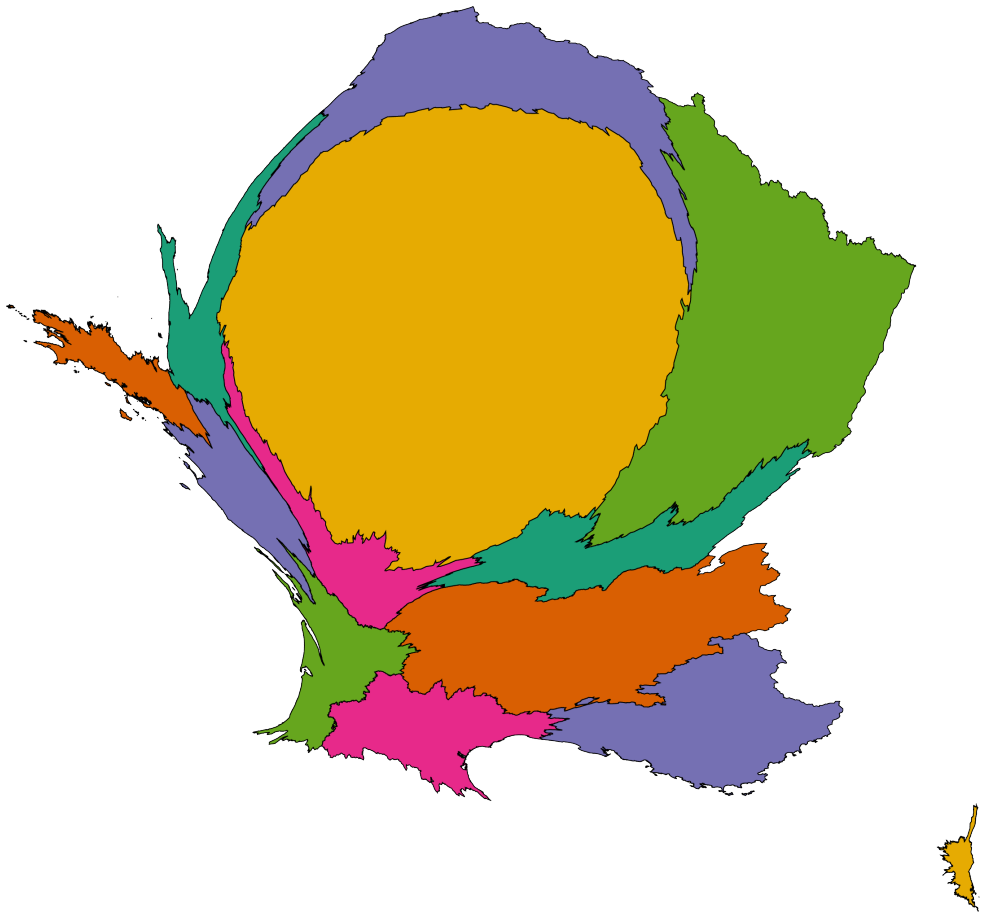


FIGURE 3.5 – Cartogramme du nombre d'hospitalisations dans les régions de France métropolitaine au 7 avril 2020.

De nombreux algorithmes ont été proposés pour générer ce type de carte. Ils diffèrent selon que les régions obtenues après transformation restent contiguës ou pas, que l'aire

28. Voir mon article de blog ici : https://tinyurl.com/yx4mf4ys.

modifiée d'une région reflète fidèlement la variable qu'elle représente, que la forme des régions originales et leur localisation restent identifiables, et que la topologie de la carte originale est préservée (Nusrat et Kobourov, 2016).

Pour cet exemple, j'ai choisi l'algorithme de Gastner et Newman (2004). Ce choix est motivé par trois raisons. D'abord, cet algorithme offre une représentation contiguë qui affiche une erreur cartographique très faible (les aires correspondent fidèlement à la variable représentée) et préserve la topologie initiale (Nusrat et Kobourov, 2016). Ensuite, la disponibilité du code source en a fait une représentation très populaire. Enfin, il a été utilisé dans un article récent pour représenter des données phonologiques (Mielke, 2017).

Dans l'article de Mielke (2017), des tableaux contenant des symboles de l'Alphabet Phonétique International ont été transformés en cartogrammes. Par exemple, des voyelles ont été représentées selon les deux dimensions traditionnelles d'ouverture et d'antériorité, chaque cellule comportant une voyelle et présentant une aire constante dans le tableau initial. L'image du tableau a ensuite été déformée de sorte que l'aire de la cellule contenant un symbole dans le tableau final reflète la fréquence d'occurrence du phonème en question dans certaines langues. Le processus implique un mécanisme de diffusion classique selon lequel les zones à forte densité augmentent et celles à faible densité rapetissent jusqu'à ce que la densité devienne uniforme sur toute la carte (Gastner et Newman, 2004).

Pour mon cartogramme vocalique, les données comportaient 10 locuteurs masculins de l'accent *Standard Southern British English* (SSBE) issus de l'étude de Ferragne et Pellegrino (2010b). Pour chaque locuteur, j'ai analysé les 11 monophtongues nominales[29] de SSBE présentées dans des mots de type /hVd/ : *heed, hid, head, had, hard, Hudd, hard, hod, hoard, heard, hood, who'd*. Les 110 voyelles ont été segmentées manuellement dans `Praat` et le programme `cp_formants` a été utilisé pour extraire les fréquences de F1 et F2. Dans un premier temps, les fréquences des deux premiers formants ont été mesurées au milieu temporel. Ensuite, en suivant une procédure proche de celle décrite dans Williams et Escudero (2014) et dans la Section 4.4.3, les tracés formantiques ont été interpolés de sorte à obtenir 30 points pour chaque courbe. Les trajectoires formantiques ont été analysées au moyen d'une transformée en cosinus discrète (DCT) calculée sur chaque contour. Le coefficient 1 — le deuxième coefficient de la DCT, reflétant l'énergie contenue dans une demi période de fonction cosinus — a été choisi comme corrélat pertinent de la magnitude de la variation de fréquence du formant en question sur toute la voyelle.

Afin que l'interprétation du cartogramme soit facile, il est important que la carte

29. Cet adjectif fait référence au fait que certaines voyelles sont considérées comme des monophtongues par convention bien que leur réalisation puisse comporter un certain degré de mouvements formantiques.

initiale soit issue d'une représentation connue. J'ai opté pour le plan F1-F2, avec F2 sur l'axe horizontal, F1 sur l'axe vertical, et les directions des axes inversées. Il s'agit là en effet de la représentation la plus familière pour les phonéticiens et les linguistes.

L'espace a ensuite été partitionné en 11 régions contiguës, une par voyelle. Ce pavage a été réalisé par le biais d'un diagramme de Voronoi dont les germes étaient les coordonnées moyennes des voyelles dans F1-F2 calculées avec tous les locuteurs. Afin de réduire la variation inter-individuelle de fréquence de formants induite par des différences physiologiques, ainsi que pour offrir une représentation plus psycho-acoustique, les fréquences de formants ont été converties en Bark et centrées-réduites par formant et par locuteur avant de calculer les moyennes pour construire le graphe de Voronoi. Cette transformation fait passer le pourcentage de voyelles mal classées — les voyelles qui tombent dans la mauvaise cellule du graphe — de 23 % à seulement 5 %. Cet espace initial apparaît à la Figure 3.6.

Ce sont les polygones ainsi définis qui vont être déformés afin de représenter deux dimensions supplémentaires : les mouvements formantiques et la durée. Les mouvements formantiques vont être utilisés pour le cartogramme de diffusion, et la représentation de la durée s'inspire d'un autre type de cartogramme, non contigu (Olson, 1976). Des images ont été construites à partir de 4 variables : F1 et F2 estimés au milieu temporel de la voyelle, la valeur absolue du coefficient DCT 1 mesurée à partir de la courbe de F1, et la durée totale de la voyelle. Les cartogrammes sont représentés dans les Figures 3.7 et 3.8 pour les locuteurs *dme* et *hak*.

L'aire d'une zone vocalique après altération de la carte initiale représente la magnitude des variations de F1 pour chaque voyelle. La durée des voyelles a été incluse en reprenant le principe du cartogramme non contigu, où les régions sont simplement rétrécies proportionnellement à la statistique à représenter ; la forme est préservée, mais pas la contiguïté (Olson, 1976). Dans chaque cartogramme individuel, les durées des voyelles ont été re-dimensionnées entre 0,05 et 0,95. Cette valeur a été ensuite utilisée comme facteur de réduction pour construire une forme identique à la région vocalique concernée, mais dont l'aire est égale à l'aire de la région vocalique multipliée par le facteur de réduction. Le re-dimensionnement entre 0,05 et 0,95 est un choix arbitraire motivé par des raisons esthétiques.

Voici quelques commentaires pour illustrer comment interpréter les Figures 3.7 et 3.8. La forme en haut de la Figure 3.7, qui représente la région du /uː/, occupe 31 % de la surface du cartogramme. Ceci reflète un mouvement ample de F1 pour /uː/ chez le locuteur *dme*, suggérant un fort degré de diphtongaison. Cette région du /uː/ est presque

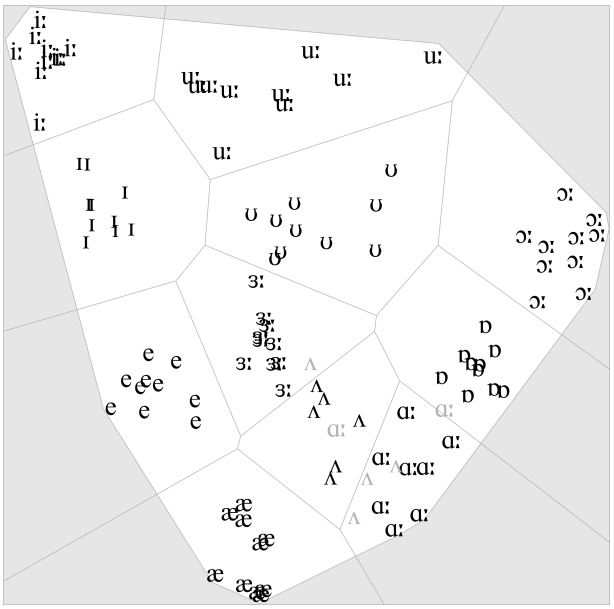


FIGURE 3.6 – Monophtongues dans l'espace F1-F2 avec graphe de Voronoi.

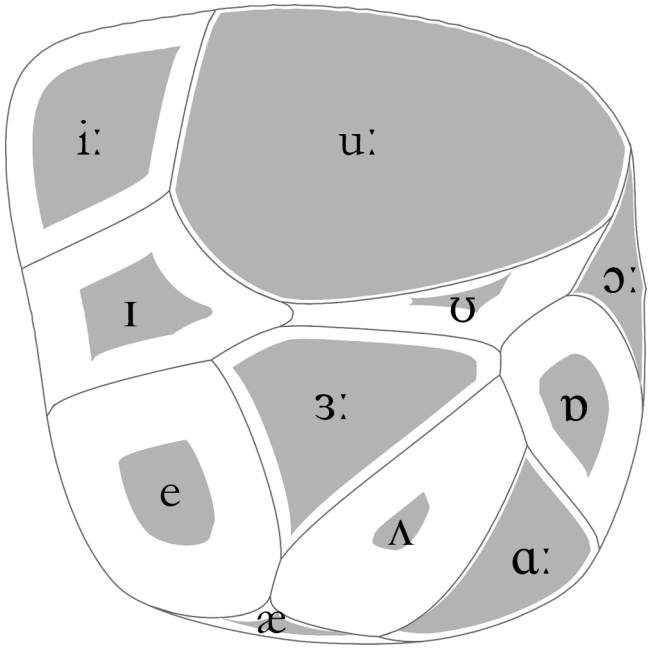



FIGURE 3.7 – Cartogramme du locuteur *dme*.

entièrement remplie de gris, ce qui matérialise le fait que cette voyelle est la plus longue chez ce locuteur (360 ms). Cette durée a donc été re-dimensionnée à 0,95, ainsi, le polygone gris couvre 95 % de l'aire de la voyelle /uː/. Le locuteur *hak* dans la Figure 3.8 présente un schéma différent. En effet, pour lui, l'aire de la région du /uː/ n'occupe que 5 % de l'espace vocalique, ce qui traduit le fait que les mouvements de F1 de cette voyelle sont parmi les

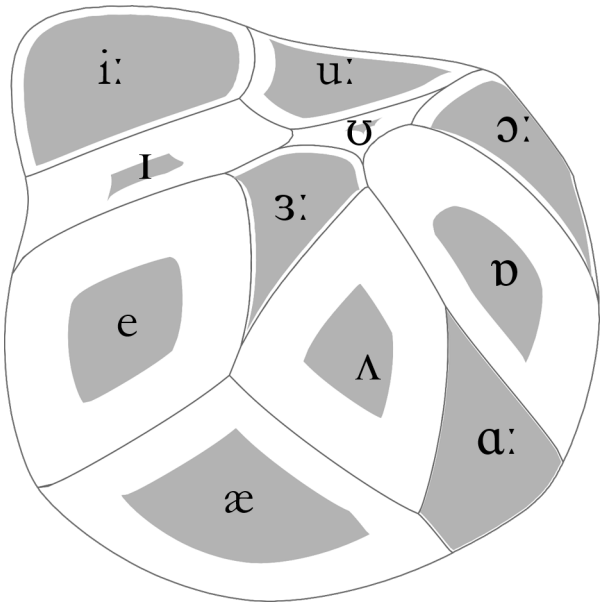



FIGURE 3.8 – Cartogramme du locuteur *hak*.

plus faibles dans le système de ce locuteur. Le fait que la forme grise à l'intérieur du /uː/ occupe une plus petite fraction de cette région par rapport à ce qu'on a pu constater dans la Figure 3.7 indique que cette voyelle, quoique figurant parmi les voyelles longues du locuteur *hak*, n'est pas la plus longue de son système.

Ces images ne sont bien sûr que des illustrations du potentiel de ce type de représentation. J'ai par exemple fait le choix de représenter à la fois la dynamique des formants et la durée ; il est possible que trop d'informations nuisent à l'intelligibilité de la figure. La représentation de la durée pour laquelle j'ai opté ici ne constitue probablement pas ce qu'il y a de plus immédiatement accessible ; une option peut-être plus lisible consisterait à utiliser une carte de chaleur. Quoiqu'il en soit, le cartogramme vocalique traduit mon souci constant de rendre ma recherche accessible, en particulier aux étudiants, par le biais de graphiques.

### 3.3.5 Lab Monitor

Je présente enfin un projet très personnel, que je crois profondément utile en termes d'ergonomie de la recherche mais qui s'éloigne quelque peu du travail de chercheur si on s'en tient à une définition trop orthodoxe. Ce projet illustre bien ma démarche de veille technologique, qui implique de sortir de ma zone de confort et d'oublier toute utilité immédiate pour disposer d'une technologie qui, un jour potentiellement, sera directement utile à ma recherche.

J'ai eu l'opportunité de mettre à disposition des étudiants, dans une salle réservée à cet usage, l'ensemble du matériel acquis au gré des divers financements obtenus depuis le début de ma carrière. Plutôt que d'imposer a priori des règles strictes de fonctionnement de cet espace, j'ai souhaité comprendre, par une démarche qu'on pourrait qualifier de « bottom up », comment les étudiants utilisaient cette salle. L'objectif consistait à rationaliser l'utilisation du matériel afin que chacun puisse travailler dans des conditions optimales. Il fallait donc être en mesure de collecter des données pertinentes et de les stocker afin d'analyser la fréquentation de cet espace et mettre en évidence une éventuelle sous-/sur-utilisation des ressources disponibles. J'ai donc mis en place un système s'appuyant sur l'Internet des Objets : il s'agit d'un prototype impliquant divers capteurs et un nano-ordinateur Raspberry Pi que j'ai appelé « Lab Monitor ». Ce prototype, qui m'a valu une interview dans le magazine officiel du Raspberry Pi[30], permet notamment de savoir quels ordinateurs sont utilisés à un instant donné, ou encore, comme l'illustre la Figure 3.9, connaître le taux de fréquentation de la salle en fonction des créneaux horaires d'une journée de travail. Bien qu'une description précise du prototype ne soit pas essentielle à la compréhension de ce document de synthèse, je l'ai néanmoins incluse dans l'Annexe B car le degré de sophistication est assez emblématique du travail technique que je mène constamment et qui n'est que rarement valorisé.

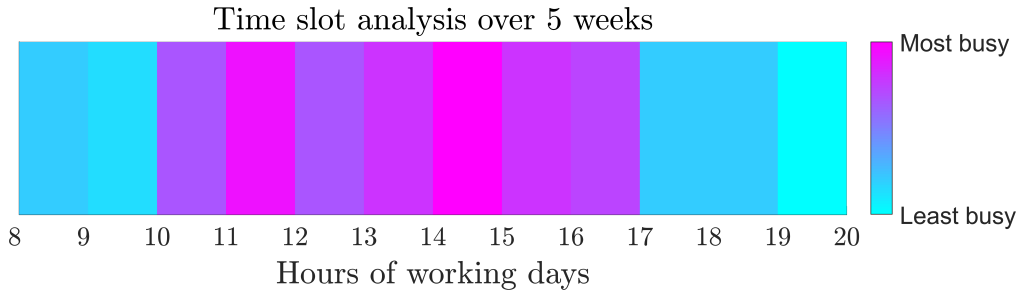



FIGURE 3.9 – Analyse de la fréquentation de la salle.

## 3.4 Réalisations diverses

Si tous les dispositifs expérimentaux et programmes mis au point ne peuvent faire l'objet d'une section à part entière, ni même figurer dans ce document de synthèse, il serait néanmoins dommage de passer à côté de quelques réalisations, dont certaines pourraient être qualifiées d'« épiques ». Il suffit de mentionner par exemple le dispositif créé à

30. https://magpi.raspberrypi.org/articles/monitoring-lab-raspberry-pi.

l'occasion de la thèse de Hannah King présenté à gauche dans la Figure 3.10, ou encore le détournement (à droite dans cette même figure) du jeu du mâche-mots dans King et Ferragne (2018) pour se convaincre que l'imagination collective dans notre équipe a su laisser quelques souvenirs marquants. Le dispositif d'acquisition inventé pour l'occasion permettait de synchroniser l'audio et le signal d'un accéléromètre fixé sur le menton des participants au moyen d'un bouton grossièrement soudé à des fils de récupération...

FIGURE 3.10 – À gauche : système d'acquisition audio-visuel « fait maison » porté par Hannah King. À droite : dispositif articulatoire combinant un « écarteur de lèvres », un accéléromètre et une sonde échographique maintenue par un casque.

Pour les travaux de thèse de Hannah King , nous avions imaginé l'interface graphique factice d'un logiciel supposé enregistrer tantôt le signal audio, vidéo, ou échographique. J'avais codé cette interface en `C#` (Figure 3.11) et avait poussé la supercherie jusqu'à créer un programme d'installation pour faciliter son déploiement.

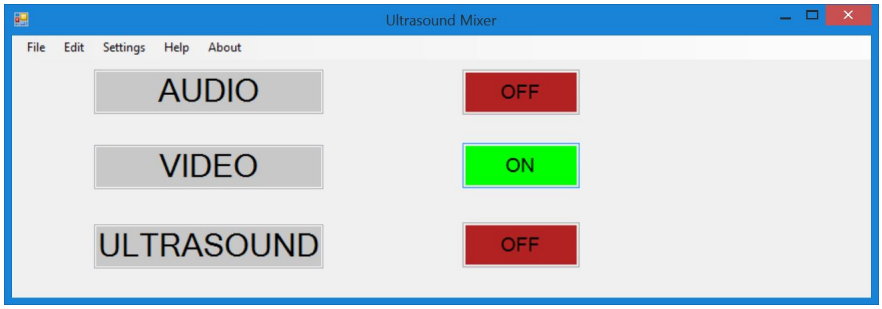



FIGURE 3.11 – Interface factice pour les besoins d'une expérience.

Dans le cadre de la thèse de Jennifer Krzonowski, qui consistait à mettre en œuvre des entraînements à la production et à la perception des voyelles de l'anglais pour des apprenants francophones, nous avions développé une interface avec `Praat` qui permet de calculer automatiquement et très rapidement les valeurs des deux premiers formants

vocaliques ainsi que la durée de la voyelle produite, et de comparer ces valeurs à des références enregistrées chez des natifs. Le tout était présenté sous la forme d'un *feedback* visuel comme l'illustre la Figure 3.12, où le point bleu matérialise la voyelle produite par l'étudiant dans l'espace F1-F2. Le point rouge représente, dans ce même espace, la cible dont il faut se rapprocher. Le cercle gris matérialise un rayon de 200 Hz, seuil au-delà duquel nous considérions, arbitrairement, que la cible acoustique n'était pas atteinte. La barre bleue en bas de la figure représente la durée de la voyelle de l'étudiant ; et la rouge marque la durée de la voyelle cible[31].

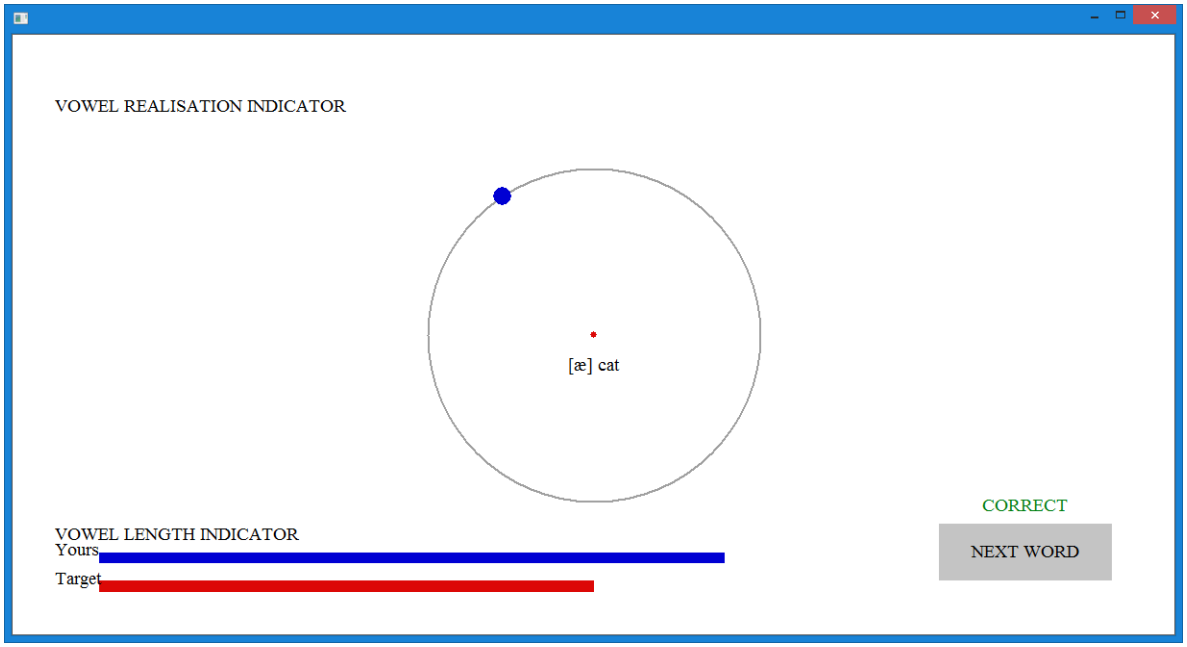



FIGURE 3.12 – Interface pour l'entraînement à la prononciation des voyelles de l'anglais.

Le Ohbot chantant est une de mes réalisations qui comporte un intérêt pédagogique pour l'initiation à la recherche en phonétique[32]. En effet, le système que j'ai mis au point pour faire « chanter » cette petite tête parlante (que nous avons déjà rencontrée dans la Figure 2.1 à la page 19) permet d'introduire la notion d'aperture des voyelles, de se familiariser avec des environnements logiciels tels que `Praat` et `Python`, et de découvrir quelques aspects élémentaires du monde de la robotique. Rendre notre discipline ludique pour des étudiants qui, en Études Anglophones, abordent la phonétique sous l'angle fastidieux de la transcription, n'est pas inutile. En effet, dans une filière constituée de disciplines qui suscitent un intérêt plus immédiat chez les étudiants, il faut bien rivaliser de créativité pour

31. Scripts et explications détaillées sont disponibles à l'adresse https://tinyurl.com/ys2urfu6.
32. Le code est disponible à cette adresse : https://tinyurl.com/3h94ejue.

susciter de nouvelles vocations[33]. Pour être complet, je dois ajouter que puisqu'avec mon système le Ohbot se contente d'ouvrir et de fermer la bouche (avec quelques mouvements de tête, de paupières et d'yeux générés aléatoirement) en synchronie avec des intervalles étiquetés dans un fichier TextGrid de `Praat`, il peut bien entendu chanter n'importe quel type de musique, ou même se contenter de parler !

## 3.5 Conclusion

Je souhaitais mettre en avant le fait que, d'une part, une bonne compréhension du domaine de la phonétique passe par la connaissance des technologies qui permettent son étude, et, d'autre part, qu'une forte implication dans la veille technologique, l'appropriation de nouvelles méthodes et matériels, et la personnalisation de ces techniques constituent un trait caractéristique de mon identité. Ainsi, le panorama des travaux que je présente au Chapitre 5 s'explique non seulement en grande partie par la variété des intérêts scientifiques qui m'animent, mais également par la diversité des techniques dans lesquelles j'ai investi.

33. J'ai réalisé deux vidéos montrant le Ohbot chantant ; ici https://tinyurl.com/nsupwav4 et là https://tinyurl.com/ruzpjt69.

# 4

# Des accents de l'anglais à la phonémicité gradiente

## Sommaire

## 4.1 Motivations

À l'issue de mon doctorat, j'étais perçu comme spécialiste des accents de l'anglais. Bien avant ma soutenance, je savais que je ne voulais pas d'une carrière qui soit fondée sur l'accumulation de connaissances descriptives concernant une langue ou les variétés d'une langue. La documentation et l'archivage, ce à quoi se résume parfois la collecte de corpus en phonologie, représentent certes des activités pertinentes pour la recherche ; mais si les données ne servent pas à éclairer une question théorique formulée en amont, elles courent le risque de ne pas être beaucoup plus utiles qu'une collection de timbres. L'article de référence des formants vocaliques de 13 accents de l'anglais (Ferragne et Pellegrino, 2010b) était en cours de rédaction dès la fin de mon doctorat ; il me fallait au plus vite compléter cette phonétique un peu trop descriptive par une grande question théorique.

À travers la description de l'anglais d'Écosse dans ma thèse, j'avais déjà rencontré le cas des contrastes dérivés, ces différences (quasi) phonémiques qui apparaissent pour distinguer deux séquences apparemment identiques de segments, mais dont l'une contient une frontière morphémique. Ces phénomènes marginaux ayant piqué ma curiosité, j'ai été conduit à lire la thèse de Kathleen Currie Hall (Hall, 2009). En tentant d'imaginer comment je pouvais allier ce type de recherche avec une démarche pluridisciplinaire, j'ai été amené à formuler l'hypothèse de la phonémicité gradiente (HPG), que je développe à la Section 4.3. Ma recherche s'est véritablement concentrée autour de cette question pendant mes quatre premières années en tant que maître de conférences. Elle m'a valu l'obtention de plusieurs financements, et reste donc très marquante dans mon parcours scientifique. Par la suite, si je m'y suis intéressé un peu plus sporadiquement, j'ai néanmoins conservé cette thématique en arrière-plan, et je n'hésite pas dans mes projets actuels à mobiliser des concepts ou des techniques que j'ai pu explorer avec HPG. HPG est intellectuellement très stimulante ; il y a cependant trois raisons qui me conduisent à lui accorder moins de place aujourd'hui. D'abord, j'ai une tendance naturelle à diversifier mes travaux, comme cela sera particulièrement visible au Chapitre 5. Ensuite, imaginer des expériences contrôlées qui comparent des phonèmes et des quasi-phonèmes sans qu'interviennent des différences acoustiques de bas niveau est malaisé ; c'est pour cela que les approches que j'ai pu proposer ne sont finalement que partiellement expérimentales car il est bien rare que le contenu acoustique ne co-varie pas avec les niveaux phonémique, quasi-phonémique et allophonique. Enfin, HPG constitue un prototype de question de recherche fondamentale : il est bien difficile d'en vanter les retombées sociétales lors de soumissions de projets, et motiver des étudiants de Master à se passionner pour un concept aussi abstrait est une

tâche délicate.

Dans ce chapitre, j'analyse le cheminement qui m'a conduit, à partir de mes travaux de thèse, vers HPG. Je passe ensuite en revue quelques-uns des résultats scientifiques les plus marquants. Mais avant cela, j'aborde deux points méthodologiques qui s'inscrivent en filigrane dans le déroulé de ma carrière : la représentation acoustique des voyelles (Section 4.2.1) et la modélisation des trajectoires formantiques (Section 4.2.2).

## 4.2 Les systèmes vocaliques des accents de l'anglais

Si certains de mes travaux concernent l'étude des consonnes (Bedoin *et al.*, 2010) ou de paramètres suprasegmentaux (Ferragne, 2013), force est de reconnaître que la plupart d'entre eux a trait aux voyelles. Je ne crois pas avoir opéré un choix délibéré ; mais je soupçonne certains facteurs d'avoir conditionné cette apparente préférence. En premier lieu, le caractère attractif de la représentation des voyelles dans le plan F1-F2 autorise la comparaison de toutes les voyelles avec une méthodologie identique. Ce modèle bi- ou tri-dimensionnel est en outre fermement ancré dans la tradition phonétique, et il est « intellectuellement » très accessible. Ensuite, les années de thèse passées à étudier 13 accents des Îles Britanniques — dans lesquels la majorité des traits discriminants étaient vocaliques — a certainement eu une influence. Enfin, le choix d'une approche automatique favorise les voyelles pour des raisons pratiques : l'alignement forcé (ou la détection automatique de f0 pour des mots hVd) fournit des frontières entre segments approximatives, mais le poids de cette approximation est bien moindre pour l'étude de valeurs formantiques prises au milieu temporel d'une voyelle, que, par exemple, pour déterminer l'emplacement d'une occlusive.

### 4.2.1 Deux approches méthodologiques concurrentes

La description des systèmes vocaliques des accents de l'anglais était un des objectifs de ma thèse, qui a ensuite été valorisé dans deux articles publiés dans des revues reconnues (Ferragne et Pellegrino, 2010b,a). Le premier article (Ferragne et Pellegrino, 2010b) est devenu avec le temps ma publication la plus emblématique puisque c'est de loin celle qui a été citée le plus grand nombre de fois (une centaine). Il contient une description des valeurs de formants des voyelles de 13 accents des Îles Britanniques. Ce travail est cité car il constitue en quelque sorte un jeu de données typique auquel les autres spécialistes peuvent se référer dans leurs publications. Il s'agit d'un article qui est centré sur la des-

cription et qui fait quelques allusions assez sommaires aux implications théoriques qui en découlent. La méthodologie s'inspire de la phonétique acoustique désormais traditionnelle (la description des voyelles par le biais des deux premiers formants vocaliques). Il n'y a pas d'innovation technique marquante ; c'est peut-être la raison pour laquelle j'ai attendu la fin de ma thèse pour me livrer à ce type de travail alors qu'il aurait probablement dû constituer la première analyse de mon doctorat !

L'autre article de description phonétique des accents publié la même année (Ferragne et Pellegrino, 2010a) promettait, quant à lui, une histoire bien plus stimulante. En s'appuyant sur les distances entre voyelles calculées dans l'espace de paramètres typiques en technologies de la parole, les *Mel-Frequency Cepstral Coefficients*, MFCC, cet article tentait de présenter une nouvelle façon de faire de la description phonétique des voyelles. Ma démarche dans ce cadre-là, qui m'a accompagné pendant toute ma thèse, présentait plusieurs avantages sur la représentation classique des formants. D'abord, elle évitait les erreurs — parfois démesurées, comme les « sauts » de formants, et nombreuses pour certaines voyelles — qui sont le lot de l'estimation formantique, en particulier quand elle est faite automatiquement. Ensuite, je tendais vers une représentation plus exhaustive, qui aurait pu faire émerger des subtilités que les formants ne codent pas. Ce n'est en effet pas un hasard si les systèmes de reconnaissance de la parole n'utilisent pas les formants. Enfin, l'idée de travailler à partir d'une distance comportait intrinsèquement une forme de normalisation du locuteur. Mais cette approche comportait malheureusement deux faiblesses qui l'ont définitivement empêchée de se développer davantage : la représentation historique des voyelles dans F1-F2 est beaucoup trop ancrée dans notre pratique, et surtout, on perdait (au moins en attendant des travaux plus aboutis) le lien entre indice acoustique et geste articulatoire.

L'article sur les distances (Ferragne et Pellegrino, 2010a) dans l'espace MFCC n'a pas eu l'impact que j'avais escompté. Il a évidemment souffert des faiblesses que je viens de mentionner. Avec le recul, je constate qu'il a également souffert des remaniements multiples dont il a fait l'objet pendant le processus de relecture : la première version des représentations graphiques (proche de ce qu'il y a dans ma thèse) était plus directement interprétable. J'ai eu l'occasion de défendre mon approche à quelques reprises, en particulier lors de la première édition des Journées d'Études et de Formation sur la Parole (JEFP) en 2010. C'est lors d'une confrontation amicale avec Cédric Gendrot, où je vantais les mérites de mon approche, que Laurianne Georgeton et Juliette Kahn ont eu l'idée de créer les JEFP. Pour la première édition, sous l'égide de l'Association Francophone

de la Communication Parlée (AFCP), sobrement intitulée « MFCC vs formants », j'étais « opposé » à Cédric Gendrot et Jacqueline Vaissière. Bien que ma méthode innovante n'ait pas remporté l'adhésion de l'auditoire, ce type de confrontation, très stimulant, a un écho particulier dans ma recherche actuelle. En effet, quand je rencontre une méthode capable de combler certaines lacunes des procédures plus traditionnelles du domaine, je m'y intéresse bien souvent. Et les techniques de *deep learning* que j'emploie désormais couramment pour toutes mes questions phonétiques, et que je décris au Chapitre 6, me remettent dans la position de défendre une approche nouvelle.

La parcimonie du modèle formantique explique en grande partie son succès. De nombreuses études s'appuyant sur la synthèse de formants ont renforcé la légitimié de cette représentation sur le plan de la perception (voir la revue de question de Kiefte *et al.*, 2013). Un facteur déterminant pour expliquer l'omniprésence de la représentation formantique bidimensionnelle provient de sa correspondance avec le triangle pédagogique historique[34], comme le rappellent Kiefte *et al.* (2013, p. 162) :

> Plots of F1×F2 formant frequencies show a relatively direct correlation to traditional descriptors of vowel tongue position in production: F1 is roughly inversely correlated to tongue or jaw height while F2 is roughly correlated with tongue advancement [...] These relationships are very attractive to phonetic theories that attempt to unify vowel production and vowel perception.

On relèvera néanmoins l'emploi des adverbes « relatively » et « roughly », qui viennent tempérer la force du lien entre les triangles pédagogique et formantique. Et j'ajouterai que toutes les études ne souhaitent pas mettre en avant ces relations : par exemple, Hillenbrand (2013) représente F1 en abscisses et F2 en ordonnées, avec une origine dans le coin inférieur gauche[35].

La méthodologie de l'analyse formantique souffre de plusieurs défauts. D'abord, l'estimation manuelle des maxima spectraux d'une voyelle n'est pas une méthode totalement reproductible puisqu'elle fait intervenir des choix et des paramétrages variables, ce qui conduit à une grande variabilité inter-expert (voir les simulations dans l'étude de Kendall et Vaughn, 2015). Ensuite, comme la pratique nous l'apprend assez rapidement, l'estima-

34. C'est à dessein que je qualifie cette représentation de pédagogique car elle n'est ni totalement acoustique ou tout à fait auditive, ni même articulatoire à proprement parler. Elle est pédagogique à deux titres. D'abord, sa version définitive a été mise au point dans un contexte où l'enseignement des langues étrangères était une question centrale pour les phonéticiens. C'est d'ailleurs la représentation privilégiée des manuels universitaires et des enseignants du supérieur encore aujourd'hui. Ensuite, elle est assez plausible sur un plan articulatoire pour que les étudiants s'en contentent, comme ils se contentent d'autres conventions.

35. René Carré, qui suit la même pratique, m'a confié un jour que cette habitude marquait la différence entre les phonéticiens et les acousticiens.

tion automatique est encore moins fiable (voir cependant les améliorations potentielles dans Weenink, 2015). Enfin, comme le notent très justement Shadle *et al.* (2016), s'il est possible de mesurer la cohérence (inter-annotateurs ou inter-logiciels) de plusieurs estimations de formants à partir d'un signal acoustique de parole, il est quasiment impossible d'en évaluer la précision. En effet, il faudrait pour cela connaître de façon fiable — et par un autre moyen que le signal de parole en question — les véritables fréquences de résonance du conduit vocal, qui pourraient alors servir de référence. Or les opportunités de mesurer plus directement la fonction de transfert du conduit vocal sont limitées[36].

En résumé, la représentation des voyelles à partir d'un point dans l'espace F1-F2 présente l'avantage d'être parcimonieuse, intelligible et validée par une longue tradition. Elle souffre cependant de certains défauts qui, schématiquement, sont liés à des problèmes de reproductibilité des mesures et d'exhaustivité. En effet, concernant ce dernier point, l'aspect dynamique est négligé ; je l'aborde à présent dans la Section 4.2.2.

### 4.2.2 Représentations de la dynamique des formants vocaliques

Les études proposant de véritables représentations dynamiques des trajectoires formantiques pour caractériser les voyelles sont nettement plus rares que celles qui se cantonnent au modèle bidimensionnel statique. La puissance de ce dernier détermine en partie cet état des choses, mais je suis convaincu que cette lacune s'explique également par le fait que les outils nécessaires à des analyses plus dynamiques ne sont pas encore tout à fait accessibles aux phonéticiens issus de formations SHS. Car en effet, ces méthodes impliquent une sophistication supérieure, qui oblige à manipuler des outils moins familiers. Dans sa version la plus élémentaire, représenter la dynamique des mouvements de formants consiste souvent à mesurer une valeur vers le début de la voyelle, et une autre, vers la fin. Dans Ferragne et Pellegrino (2010b), nous avions retenu la 2$^{e}$ et la 11$^{e}$ valeur de chaque formant, chacun étant représenté par 13 échantillons. Il est assez fréquent de s'appuyer sur les valeurs relevées à 20 % et 80 % de la durée de la voyelle (Fox et Jacewicz, 2009).

Une grande partie des recherches sur les mouvements formantiques est fédérée dans le cadre du paradigme VISC : *Vowel Inherent Spectral Change* (Morrison et Assmann, 2013). À travers VISC, la nécessité de prendre en compte la dynamique de toutes les

36. Cela est par exemple possible en excitant les résonances avec un signal de source artificiel connu, caractérisé par une résolution fréquentielle très élevée comme pour le dispositif présenté dans Epps *et al.* (1997).

voyelles (monophtongues comme diphtongues) est mise en avant, ce qui offre un cadre unifié pour la représentation acoustique des voyelles. Hillenbrand (2013) fournit quatre bonnes raisons de prendre en compte la dynamique des formants :

1. Même les voyelles que les phonéticiens classent parmi les monophtongues comportent des mouvements formantiques importants ;
2. Les études en classification automatique montrent une meilleure séparabilité des voyelles quand les changements spectraux au fil de l'émission de la voyelle sont pris en compte ;
3. Les expériences sur les *silent centers* prouvent qu'il est possible de remplacer l'état stable d'une voyelle par du silence sans que cela ait un effet marqué sur l'intelligibilité ;
4. Les expériences avec des voyelles synthétiques et naturelles montrent que les voyelles avec un spectre stable sur toute la durée sont mal identifiées.

La pertinence des pentes formantiques pour l'identification des voyelles par des auditeurs est très bien illustrée dans l'étude de Chládková *et al.* (2017). En anglais britannique standard, les voyelles des ensembles lexicaux FLEECE et GOOSE sont plus proches dans le plan F1-F2 qu'elles ne l'étaient auparavant en raison d'une antériorisation de la seconde. Il se trouve par ailleurs que, quel que soit le contexte consonantique, la voyelle de GOOSE comporte une trajectoire de F2 descendante, alors que celle de FLEECE présente une trajectoire montante. Chládková *et al.* (2017) étudient dans quelle mesure les auditeurs utilisent cet indice dans l'identification des voyelles et, en particulier, si un groupe d'auditeurs âgés, habitués à une séparabilité optimale dans le plan F1-F2 statique, l'utilise différemment d'un groupe d'auditeurs plus jeunes, pour qui les deux voyelles sont proches quand on ne prend en compte que l'état stable. Dans une tâche d'identification à partir de voyelles de synthèse avec trois types de pentes différentes (montante, descendante ou plate), et différentes valeurs de F2 au milieu temporel, les auteurs démontrent la pertinence de la pente comme indice secondaire. En effet, pour une valeur centrale de F2 ambiguë, une pente descendante conduit à une probabilité accrue d'identifier la voyelle comme étant celle de GOOSE. La direction de la pente (montante vs descendante) déplace la frontière de F2 d'environ 100 Hz.

Morrison (2013) passe en revue les trois principales hypothèses du modèle VISC concernant les paramètres pertinents sur le plan de la perception. Ces trois hypothèses concordent sur le fait que l'identification des voyelles nécessite la prise en compte des valeurs formantiques en début de voyelle ; mais l'autre paramètre — valeur formantique

finale, vitesse de la pente formantique ou direction de la pente — ne fait pas l'objet d'un consensus. L'hypothèse *onset + offset* implique que seules les valeurs de fréquence de début et de fin (et donc la différence entre ces deux valeurs) sont pertinentes. L'hypothèse *onset + slope* pose, quant à elle, que c'est la vélocité de la pente qui compte, que la cible finale soit atteinte ou non. Enfin, avec l'hypothèse *onset + direction*, on considère que c'est la direction des trajectoires formantiques dans le plan F1-F2 qui permet l'identification de la voyelle en question.

Après un examen critique des études empiriques disponibles, Morrison (2013) arrive à la conclusion que c'est l'hypothèse *onset + offset* qui est la plus plausible. Il ajoute néanmoins que cette représentation semble certes contenir l'information suffisante à la classification des voyelles, mais des représentations plus élaborées sont nécessaires pour obtenir l'information liée aux consonnes adjacentes, ou aux propriétés du locuteur.

### 4.2.3 Techniques de représentation des trajectoires

Les valeurs de formants mesurées sur une voyelle sont des séries temporelles, un cas particulier de données fonctionnelles. Les méthodes qui s'appliquent à ces dernières sont donc tout à fait indiquées, d'autant plus que leur usage commence à se populariser en phonétique (Gubian *et al.*, 2015) et que les outils pour les implémenter sont disponibles dans `R` et `Matlab` (voir par exemple Ramsay *et al.*, 2009).

Une analyse fonctionnelle des trajectoires formantiques consiste non plus à se contenter des valeurs discrètes de fréquence dans le temps, mais revient à estimer les paramètres de la fonction sous-jacente qui a généré ces valeurs. Les avantages sont multiples. D'abord, l'estimation qu'on obtient d'une trajectoire formantique est un signal bruité, et l'approche fonctionnelle peut constituer une forme de débruitage. Car, en effet, derrière les erreurs d'estimation, les variations de débit, les approximations du contrôle moteur des locuteurs, il peut être souhaitable de retrouver l'intention articulatoire première. Ensuite, les paramètres d'une fonction qui stylise un signal sont un moyen de réduire sa dimensionnalité. On pourra ainsi, par exemple, se contenter des 3 ou 4 premiers coefficients d'une transformée en cosinus discrète pour reconstituer l'essentiel d'un tracé comprenant initialement des dizaines de valeurs différentes. Enfin — mais ce n'est pas l'apanage de l'approche fonctionnelle — analyser des contours formantiques dans leur intégralité rend possible une approche cinématique, pour une caractérisation plus exhaustive incluant la vitesse et l'accélération. En somme, si on considère un tracé formantique comme un signal, une multitude de techniques éprouvées (sans oublier le filtrage et l'analyse fréquentielle) s'offre

aux chercheurs.

Dans Ferragne et Pellegrino (2010b), nous avions ajusté des fonctions polynomiales aux contours de formants. Si cette méthode fonctionne, elle est néanmoins contraignante car il faut a priori décider pour chaque voyelle le degré du polynôme à ajuster selon que la trajectoire suit approximativement une ligne droite, une courbe simple ou en forme de S, etc. De plus, si on souhaite comparer toutes les voyelles à partir des paramètres des fonctions ajustées, puisqu'ils sont en nombre différent selon le type de voyelle, la comparaison est impossible.

Parmi les méthodes employées pour modéliser les trajectoires formantiques, on compte la transformée en cosinus discrète. Proche de la transformée de Fourier, cette technique est utilisée en compression d'images car elle permet de réduire l'essentiel de l'énergie d'un signal à un nombre très restreint de coefficients. C'est la méthode que j'utiliserai à la Section 4.4.3 pour analyser les réalisations des voyelles de type *tide* et *tied* en anglais d'Écosse.

## 4.3 L'hypothèse de la phonémicité gradiente

L'hypothèse de la phonémicité gradiente (HPG) a caractérisé une partie des travaux qui ont suivi la thèse ; en particulier à travers l'obtention d'un Projet Exploratoire/Premier Soutien (PEPS) du CNRS, d'une subvention de recherche de la Fondation Fyssen et du projet ANR COREGRAPHY[37]. La formulation de cette hypothèse m'a été largement inspirée par des travaux comme ceux de Hall (2013) ou Scobbie et Stuart-Smith (2008).

### 4.3.1 Définition

La plupart des linguistes qui posent le phonème comme prémisse à leur analyse conçoivent l'opposition phonologique à travers une logique strictement binaire : deux sons du langage sont susceptibles soit de former une opposition phonémique, soit de n'être que des variantes d'une seule et même entité fonctionnelle, à l'exclusion de toute autre possibilité. Ce modèle, s'il permet d'établir un inventaire phonologique à moindre coût, ne rend cependant pas compte de nombreux phénomènes attestés dans les langues du monde. Ces phénomènes — restrictions distributionnelles conditionnées par le contexte phonologique, par la dérivation morphologique, par le type de lexique, nombre de traits

37. *COgnitive REality of the GRAdient Phonemicity HYpothesis.*

spécifiés variable d'un phonème à l'autre, etc. — conduisent à penser que certaines oppositions sont plus typiquement phonémiques que d'autres. La notion de degré de phonémicité — pourtant latente chez les structuralistes — commence tout juste à recevoir l'attention qui lui est due. Si l'on postule que le statut de phonème obéit à une logique plurivalente — voire floue — il convient de soumettre cette hypothèse à des tests empiriques adaptés, en complément d'une démarche de formalisation. Avant de décrire ces tests plus en détail, j'illustre brièvement les quelques phénomènes linguistiques qui autorisent à faire l'hypothèse d'un degré de phonémicité variable.

### 4.3.2 L'argument taxinomique

Une fois établi l'inventaire phonologique d'une langue, l'analyse linguistique peut prendre la forme d'une taxinomie dans laquelle la relation entre phonèmes est exprimée en termes de traits distinctifs. Quelle que soit la méthode utilisée (Dresher, 2009), décider quels traits ont un statut distinctif amène les phonèmes d'une langue à être définis par un nombre de traits variant d'une entité à l'autre. Ou, plus exactement, le nombre de traits pour lesquels une valeur est spécifiée varie entre phonèmes. Ceci conduit certains phonèmes à apparaître comme des entités « par défaut », moins marquées que d'autres (Rice, 2007). Ainsi, l'analyse taxinomique classique laisse entendre qu'il est possible d'aller au-delà de la partition phonème vs non phonème. Cette variation d'un phonème à l'autre quant au nombre de traits ayant une valeur spécifiée reflète d'ailleurs une réalité empirique. À titre d'exemple, les coronales ont tendance à être sous-spécifiées dans les langues du monde, et cette particularité peut être mise en parallèle avec le fait que les coronales sont plus fréquentes, apprises en général plus précocement, enclines à l'assimilation, etc. (Paradis et Prunet, 1991 ; Montreuil, 2001).

### 4.3.3 L'argument distributionnel

La distribution d'un son fait référence à l'ensemble des positions et contextes dans lesquels ce son est attesté pour une langue donnée. Je considère que lorsqu'un phonème présente une distribution restreinte, son comportement n'est pas typiquement phonémique. Si, comme Hall (2013), on applique le critère de prédictibilité d'un son pour décider s'il s'agit d'un phonème ou non (imprévisible : phonème ; prévisible : allophone), alors la notion de degré de phonémicité s'impose. En s'appuyant sur une version simplifiée de la typologie de Hall (2013), qui recense les facteurs contraignant la distribution de phonèmes,

on peut identifier trois grandes familles :

1. Distribution contrainte par des facteurs phonologiques positionnels ;
2. Distribution contrainte par le type de lexique (étranger, spécialisé) ;
3. Distribution contrainte par la complexité morphologique.

Dans la première catégorie, on compte toutes les restrictions d'occurrence liées à la position dans la syllabe, dans le mot, et au type de syllabe. Par exemple, en anglais standard, /ŋ/ n'apparaît que dans la rime alors que /h/ est restreint aux attaques de syllabes. En français standard, la neutralisation de l'opposition /o/-/ɔ/ en syllabe ouverte et de /e/-/ɛ/ en syllabe fermée constitue un autre exemple de cette catégorie. On peut citer encore la neutralisation de l'opposition de voisement en finale de mot en allemand (Wiese, 2006) et en russe (Dmitrieva *et al.*, 2010). La distribution contrainte par le type de lexique fait référence aux phonèmes dont l'utilisation est restreinte à certains mots étrangers, spécialisés ou rares. En japonais, les 4 classes de morphèmes, Yamato (vocabulaire natif), sino-japonais, mimétique et étranger imposent des contraintes bien spécifiques sur la phonologie. Par exemple, seules les obstruentes voisées sont licites après une nasale dans les morphèmes Yamato et mimétique alors que, dans la même position, des obstruentes non voisées sont permises dans les morphèmes sino-japonais et étrangers (Îto et Mester, 1995). Le phonème /x/ en anglais, attesté dans la paire *lock/loch* ne se retrouve que dans des emprunts ou des noms de famille, comme un effort d'imitation d'une prononciation étrangère (comme dans *sheikh* et *Bach*). En anglais standard, /ð/-/θ/ obéissent au même type de contrainte ; à l'initiale de mot, /ð/ n'apparaît que dans les mots grammaticaux, et /θ/, uniquement dans des items lexicaux. Ces exemples conduisent d'ailleurs à penser qu'une langue n'a peut-être pas un seul système phonologique (mono-systémique), mais plusieurs (Lodge, 2009). Enfin, les contraintes induites par la complexité morphologique concernent les contrastes dérivés, qui surviennent lorsque l'inventaire phonologique incluant des items morphologiquement complexes est plus grand que celui obtenu à partir de mots morphologiquement simples. Les exemples donnés ici sont plus détaillés car ces deux phénomènes ont été centraux dans le développement de mes travaux. Si *side* et *sighed* sont de parfaits homophones en anglais britannique standard, il en va différemment pour certaines variétés d'anglais d'Écosse en raison du phénomène de *Scottish Vowel Length Rule* (SVLR — Rathcke et Stuart-Smith, 2016 ; Ferragne, 2020). En effet, relativement à celle de *side*, la voyelle de *sighed* est allongée et présente un timbre distinct. Cette distinction n'est pas répertoriée dans l'inventaire des phonèmes de l'anglais d'Écosse pour la bonne raison qu'elle est totalement prévisible si l'on s'appuie sur le contexte morphologique : l'ajout du

suffixe <ed> conditionne la variation de longueur et de timbre de /ai/[38]. L'allongement de la voyelle avant frontière morphémique concerne, outre /ai/, les voyelles /i/ et /u/. Ainsi, *need/kneed* et *brood/brewed* forment deux autres paires minimales présentant ce que certains auteurs nomment des quasi-phonèmes. Ce type de phénomène se produit également avec des paires de consonnes. Toujours en anglais, la plupart des réalisations de /t/, /d/ et /n/ ont une articulation apico-alvéolaire. Lorsque /t/, /d/ ou /n/ sont suivis d'une consonne dentale, une assimilation de lieu conduit à une réalisation dentale (et non plus alvéolaire) de ces trois consonnes. Chez certains locuteurs de l'anglais en Irlande du Nord, lorsque ces consonnes précèdent <er>, la même réalisation dentale de /t/, /d/ et /n/ survient, sauf — et c'est ce point qui mérite notre attention — lorsque <er> est un morphème. Ainsi, dans le verbe (mono-morphémique) *flatter*, /t/ est dental, alors que dans le comparatif (bi-morphémique) *flatter*, /t/ reste alvéolaire (Wells, 1982 ; Harris, 1990). Selon le test des paires minimales, ces deux phénomènes donnent lieu à deux phonèmes différents qui n'émergent pourtant que dans un contexte morphologique bien précis.

### 4.3.4 L'argument fréquentiste

L'une des versions du test de commutation pose que les paires minimales effectivement attestées dans une langue ne sont qu'une manifestation fortuite de l'inventaire phonologique, et qu'il convient donc d'inclure dans le test des pseudo-mots (conformes aux contraintes phonotactiques de la langue) pour dresser un inventaire phonologique complet (de Carvalho *et al.*, 2010). Cette approche formelle est tout à fait viable empiriquement en cela qu'elle permet de révéler un potentiel phonologique, pas nécessairement réalisé dans une langue donnée. Il est cependant possible d'objecter qu'elle ne tient pas compte des effets de la fréquence pourtant fondamentaux dans les modèles type *usage-based* (Pierrehumbert, 2001), et centraux dans les modèles à exemplaires (Johnson, 2007). À titre d'exemple, les études citées dans Pierrehumbert (2001, p. 89-93) concourent à démontrer que la connaissance d'un locuteur concernant l'acceptabilité d'un son de parole dans une séquence donnée est d'ordre probabiliste. En d'autres termes, l'acceptabilité phonotactique est affaire de degrés : les séquences les plus fréquentes sont jugées comme plus

38. Les anglicistes ont parfois pris l'habitude d'utiliser la convention de l'accent RP pour symboliser les phonèmes des autres variétés et attendraient donc ici /aɪ/ plutôt que /ai/. Tout est affaire de convention : Scobbie *et al.* (1999) utilisent /ai/ ; Wells (1982) oppose /ae/ et /ʌi/ en Écosse. Ces conventions phonologiques mériteraient un chapitre à elles-seules, qui déterminerait le poids à accorder à chacune de ces forces contradictoires conduisant au choix d'un symbole : la tradition, la justesse phonétique, ou encore les biais en faveur des variétés standards.

acceptables.

Les arguments que nous venons brièvement de passer en revue démontrent la nécessité de faire appel à la notion de degrés de phonémicité. Il est désormais évident que, pour les raisons que nous venons d'invoquer, tous les phonèmes d'une langue n'ont pas le même statut : une analyse taxinomique le démontre avec les notions de marque et de sous-spécification, l'analyse distributionnelle confirme qu'il existe des oppositions plus ou moins typiques, et les études menées dans une optique fréquentiste indiquent que la manière dont les entités linguistiques sont stockées est largement tributaire de la fréquence. Il reste donc à envisager comment ces phénomènes peuvent influer sur les représentations phonologiques, et comment obtenir une mesure de cette influence à partir d'expériences de production et de perception.

## 4.4 Tester HPG en production

Les arguments que je viens d'énumérer suffisent à démontrer que tous les éléments d'un système phonologique n'ont pas le même statut. L'entreprise dans laquelle je me suis investi dès 2010 consistait donc non pas à prouver ce qui l'était déjà, mais bien plutôt à mettre en évidence les éventuels corrélats expérimentaux reflétant ces différents degrés de phonémicité. Afin que les expériences menées dans ces divers projets soient intelligibles, il est indispensable de les situer dans un cadre qui présuppose un lien très direct entre la forme acoustique des sons et leur représentation mentale. Les modèles à exemplaires, tels que décrits dans Pierrehumbert (2001) et la modélisation des catégories phonémiques dans Feldman *et al.* (2009) et Kronrod *et al.* (2016) constituent des bases particulièrement pertinentes en la matière.

Si les représentations mentales des sons de la parole se construisent sur la base d'un apprentissage statistique à travers lequel des distributions de probabilité se forment dans un espace psycho-acoustique dérivé de l'espace acoustique initial, une étude de la production doit pouvoir fournir des éléments sur ces représentations. Les résultats empiriques qui suivent se fondent en particulier sur un chapitre que j'ai publié récemment (Ferragne, 2020), et qui fait la synthèse d'une partie des travaux que j'ai menés en rapport avec HPG.

### 4.4.1 Données en production

Un premier ensemble de données est constitué d'enregistrements effectués à Hull en 2011. Dix-huit locuteurs ont été invités à lire cinq paires minimales contenant le contraste TRAP-BATH[39] apparaissant dans la phrase porteuse *He said the word ...* Les enregistrements ont été effectués avec le logiciel `ROCme!` (Ferragne *et al.*, 2013). Cet ensemble contient 540 voyelles : 18 locuteurs × 10 mots cibles × 3 répétitions. Dans ce qui suit, c'est la moyenne des 3 répétitions qui sera utilisée.

L'ensemble GLA1 a été enregistré à Glasgow en 2009 (Ferragne *et al.*, 2010). Les locuteurs ont lu des mots susceptibles de contenir des exemples de la SVLR apparaissant dans la phrase porteuse *He said the word [...] and I didn't know how to spell it.* Les données ont été recueillies via une interface programmée pour l'occasion en Tcl/Tk. Le sous-ensemble utilisé ici contient 432 voyelles illustrant le contraste dérivé GOOSE-BREWED (12 locuteurs × 12 mots-cibles × 3 répétitions) ; là encore, c'est la moyenne des 3 répétitions qui sera utilisée.

L'ensemble GLA2 a été enregistré à Glasgow en 2010. Les locuteurs ont lu un court texte que j'avais spécialement conçu pour cette expérience[40] dont seule l'opposition *crude-crewed* sera analysée ici, produite par 20 locuteurs pour un total, donc, de 40 items.

Les données d'Ulster (Stephan et Ferragne, 2012) ont été enregistrées à Enniskillen en 2011 avec le logiciel `ROCme!`. L'ensemble analysé ici comprend quatre mots test visant à éliciter deux paires minimales du type GOOSE-BREWED présentées dans des phrases porteuses variées[41]. Au total, ce sont 96 voyelles qui seront incluses (24 locuteurs × 4 items) dans l'analyse.

Malgré la diversité des systèmes d'enregistrement, tous les fichiers audio étaient au format LPCM mono échantillonnés à 44,1 kHz avec une profondeur de 16 bits. J'ai personnellement collecté les données pour le corpus GLA2. Pour Hull, GLA1 et Ulster, ce sont trois étudiantes — Séverine Delcourt, Joana Afonso-Santiago et Pauline Stéphan respectivement — qui se sont acquittées de cette tâche dans le cadre de leur mémoire de Master.

39. *back-bark, cad-card, cap-carp, match-march, pat-part.*

40. Les mots test apparaissent en gras : Captain Duncan **crewed** on two ships, one of which remained permanently **tied** to a pier in Ayrshire. It was said that he **wooed** fame and success ; but few people **knew** that he actually **sighed** for early retirement, and had great **need** of rest. When the **tide** was high, the ship would sway from **side** to **side**, and Duncan **would** sit quietly, listening to the endless creak of the **wood**. One day, a fellow mariner made a very **crude** joke and Duncan **kneed** the poor lad overboard.

41. A synonym for 'offspring' is '**brood**'. Coffee is tastier when freshly **brewed**. I can't believe she painted the nursery in magenta. This colour is so **crude**. The captain was ready to cast off. His vessel was well-equipped and **crewed**.

### 4.4.2 Modélisation de la durée

Dans cette section, l'opposition de longueur entre TRAP et START à Hull, dont le statut est phonémique, est comparée aux différences de durées quasi-phonémiques entre les ensembles lexicaux GOOSE et BREWED afin de mettre en évidence une éventuelle différence dans le comportement de l'indice de durée qui découlerait de la différence de statut linguistique.

Une première étape exploratoire consiste à observer dans la Figure 4.1 la variation de durée selon que la voyelle est brève ou longue. On y remarque tout d'abord que les deux catégories supposées semblent bel et bien attestées pour les données de Hull, GLA1 et GLA2. En revanche, l'ensemble de données Ulster ne fait pas apparaître ce schéma ; l'allongement escompté n'a pas pu être mis en évidence. Afin de quantifier le degré de séparabilité entre brèves et longues, un classifieur logistique avec validation croisée a été entrainé sur les données. Les résultats sont présentés dans le Tableau 4.1 ; l'ensemble Ulster n'y figure pas puisque le chevauchement des deux densités de probabilité dans la Figure 4.1 suffit à convaincre de l'absence de séparabilité. Pour les 3 autres ensembles de données, nous allons considérer que la séparabilité est à peu près équivalente. Si l'on se concentre sur les données de Hull et GLA1 dans le Tableau 4.1 et la Figure 4.1, plusieurs commentaires utiles peuvent être formulés. D'abord, les durées de GLA1 sont plus faibles (médiane GLA1 : 131 ms, Hull : 167 ms ; iqr[42] GLA1 : 88 ms, Hull : 113 ms), et la frontière y est plus basse[43]. À ce stade, on peut supposer que cette différence provient de la différence de phrase porteuse, qui a pu induire un allongement à Hull résultant de la position finale du mot cible. Ensuite, alors que les densités sont de forme équivalente à Hull, elles sont dissemblables pour GLA1 : le caractère « pointu » de la densité des brèves contraste avec l'aspect plus « écrasé » des longues. En s'appuyant sur les valeurs du Tableau 4.1 pour ces deux ensembles de données, on constate en effet que les voyelles longues affichent une variation de durée objective plus grande que celle des brèves. Et la Figure 4.1 fait apparaître que ce phénomène est particulièrement marqué pour GLA1.

Si on compare à présent GLA1 et GLA2, on constate que, pour ce dernier, les durées sont plus faibles, les variations intra-catégorielles plus faibles également, et la frontière intervient plus bas dans l'échelle de durée. La très nette différence de forme des densités constatée pour GLA1 n'apparaît pas de façon évidente avec GLA2. La possibilité de comparer ces deux ensembles est néanmoins limitée : les items de GLA1 ont été recueillis

42. J'utilise l'acronyme anglais, *interquartile range* pour l'écart interquartile.
43. La limite supérieure des abscisses a été fixée à 320 ms de façon à préserver une résolution convenable dans les quatre graphiques, quitte à légèrement amputer les valeurs de START pour Hull.

TABLEAU 4.1 – Durée moyenne et empan (ms) des voyelles brèves et longues pour les ensembles de données Hull, GLA1 et GLA2 ; localisation de la frontière (ms) et taux de classification correcte (%).

| données | brève | | longue | | frontière | taux de class. |
|---|---|---|---|---|---|---|
| | moyenne | empan | moyenne | empan | | |
| HULL TRAP-START | 109 | 133 | 221 | 199 | 161 | 95 |
| GLA1 GOOSE-BREWED | 96 | 102 | 179 | 151 | 129 | 91 |
| GLA2 GOOSE-BREWED | 74 | 93 | 129 | 99 | 100 | 93 |

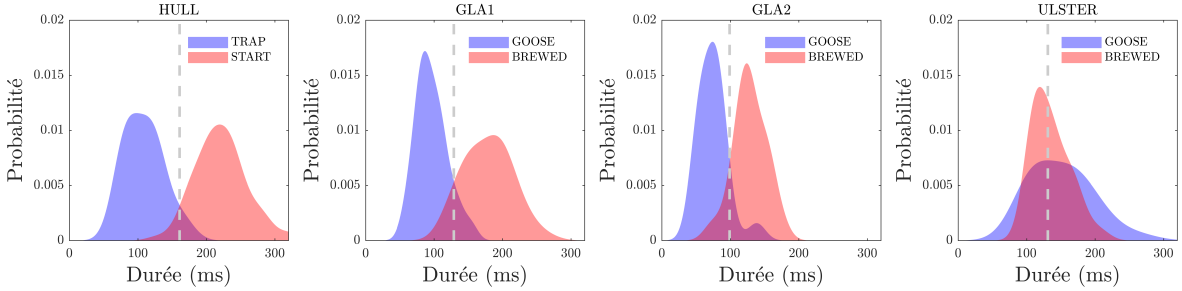



FIGURE 4.1 – Durée des voyelles brèves et longues pour Hull, GLA1, GLA2 et Ulster.

dans un environnement très contrôlé favorisant l'allongement, et identique d'un essai à l'autre, alors que pour GLA2, il s'agit d'un texte lu ; le contrôle de biais éventuels (e.g. position dans le groupe intonatif) est incertain. Malgré ces limites imputables à la manière dont les données ont été collectées, je retiendrai quatre aspects :

1. Une différence robuste de durée a pu être mise en évidence entre les voyelles brèves et longues des ensembles Hull, GLA1 et GLA2 ;
2. Quelles que soient les conditions de recueils des données, le degré de séparabilité est maintenu ;
3. La frontière catégorielle en production n'est pas fixe ;
4. La distribution des voyelles longues affiche une plus grande variance que celle des voyelles brèves ; ceci est particulièrement vrai pour GLA1.

L'analyse des durées brutes qui vient d'être menée omet plusieurs facteurs. En effet, le débit de parole est connu pour influencer la durée des segments, en particulier des voyelles (Gay, 1978), certains contextes phonologiques ont un impact sur la durée vocalique (Klatt, 1976), des indices secondaires, comme des différences de timbre, sont susceptibles de venir favoriser des distinctions que la durée peine à matérialiser, l'utilisation de ces indices peut varier d'un locuteur à l'autre, et, enfin, la représentation des catégories sous forme de

distributions de durées objectives ignore le filtre de la perception. Ces différents éléments sont abordés brièvement dans les Sections 4.4.2.1 à 4.4.2.4.

### 4.4.2.1 Effet du débit

Puisqu'il est établi que le débit de parole a une influence sur la frontière entre deux catégories adjacentes dans un espace acoustique — voir par exemple Miller et Volaitis (1989) pour le cas du VOT — ainsi que sur la forme de ces catégories, il est nécessaire d'étudier cet effet. Afin de déterminer l'impact du débit de parole sur la durée des voyelles, des modèles de régression ont été ajustés aux données des ensembles Hull et GLA1, indépendamment pour les voyelles brèves et les voyelles longues. La phrase porteuse *He said the word ...* étant commune aux deux ensembles, sa durée constitue une mesure de débit — ou plutôt de l'inverse du débit — sinon fiable, au moins, comparable. Les deux premiers graphiques de la Figure 4.2 montrent la droite des moindres carrés ordinaires ajustée aux données ainsi qu'un intervalle de confiance à 95 %. Les valeurs de $R^2$ sont rapportées dans le Tableau 4.2 ; tous les résultats sont significatifs au moins au niveau $p < 0,01$.

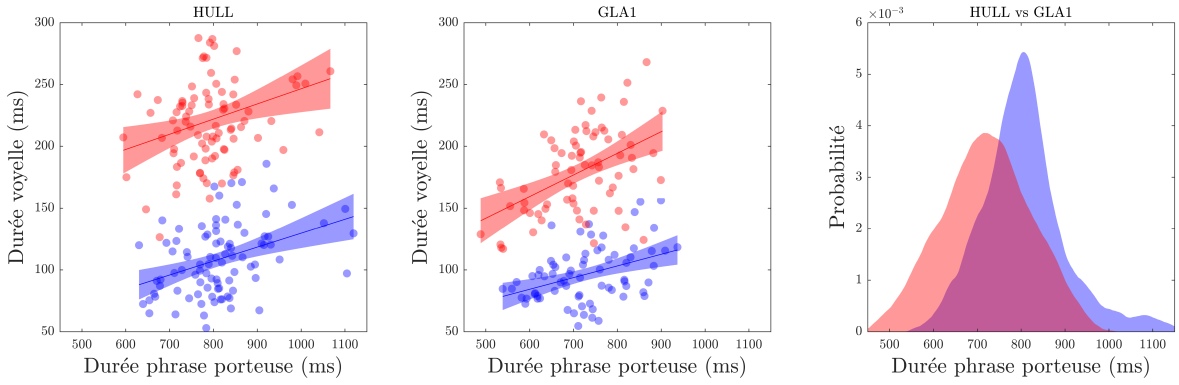



FIGURE 4.2 – Régression linéaire des moindres carrés montrant l'effet de la durée de la phrase porteuse sur la durée des voyelles brèves (bleu) et longues (rouge). Les nuages de points représentent les valeurs brutes et la zone colorée autour de chaque droite matérialise l'intervalle de confiance à 95 %. Le panneau de droite montre la durée des phrases porteuses pour GLA1 (rouge) et Hull (bleu).

La Figure 4.2 et les valeurs du Tableau 4.2 suffisent à convaincre que l'impact du débit sur la durée vocalique, quoique statistiquement significatif et allant dans le sens escompté, n'est pas particulièrement bien décrit par les modèles proposés ici. On peut noter que la qualité de l'ajustement est relativement élevée pour BREWED pour GLA1 et particulièrement faible pour START dans Hull. Les modèles de régression pour les voyelles

TABLEAU 4.2 – Coefficients de détermination $R^2$ des 4 modèles de régression de la Figure 4.2.

| | Brèves | Longues |
|---|---|---|
| HULL | 0,14 | 0,07 |
| GLA1 | 0,15 | 0,23 |

brèves des deux ensembles de données affichent, quant à eux, un comportement similaire. Les modèles sont mal adaptés en particulier en raison d'un certain degré d'hétéroscédasticité : ceci est particulièrement visible dans la Figure 4.2 pour la voyelle de BREWED, où la variance des durées vocaliques augmente en même temps que les valeurs de débit. Par ailleurs, on peut s'interroger concernant la justesse avec laquelle la durée d'une phrase porteuse, répétée de manière incessante tout au long d'une expérience, reflète le débit de parole. On remarque également, en inspectant le troisième graphique de la Figure 4.2, que la durée moyenne des phrases porteuses est plus faible pour GLA1 tout en étant globalement[44] plus variable que pour Hull. Les médianes pour GLA1 et Hull respectivement, sont de 718 et 800 ms, et les écart interquartiles correspondants sont de 132 et 99 ms. On peut supposer intuitivement que cette différence de durée moyenne est imputable à une légère différence de protocole : pour GLA1, la phrase porteuse était *He said the word . . . and I didn't know how to spell it* alors que pour Hull, la partie après le mot cible, lequel est matérialisé ici par des pointillés, n'était pas produite.

#### 4.4.2.2 Contexte phonologique allongeant

Dans les mots cibles de l'expérience de production menée à Hull, le contexte phonologique à droite de la voyelle varie, alors qu'il est constant pour GLA1. On sait par ailleurs que dans des cas similaires, une consonne voisée provoque un allongement de la voyelle qui précède, de 50 à 100 ms d'après Klatt (1976). La Figure 4.3 apporte un élément de réponse : on y constate en effet que les durées moyennes les plus élevées, dans le groupe des voyelles longues comme dans celui des voyelles brèves, correspondent aux mots cibles se terminant par une consonne voisée. L'effet du contexte consonantique qui suit la voyelle est donc bien illustré ici.

44. Comme le suggère la forme et le degré d'aplatissement de la distribution, et non pas en prenant en compte tout l'empan des valeurs puisque, si nous le faisions, les quelques valeurs très élevées de Hull viendraient fausser cette impression.

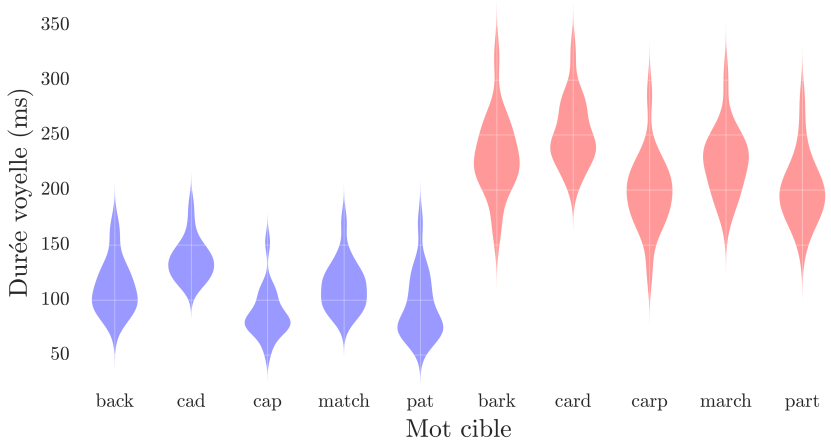



FIGURE 4.3 – Durée des voyelles de Hull en fonction du mot cible.

#### 4.4.2.3 Interaction potentielle entre durée et mesures spectrales

En anglais standard d'Angleterre, les oppositions de longueur phonologique impliquent non seulement des différences de durée, mais également des différences de timbre (Ferragne et Pellegrino, 2010b). Afin de mettre en lumière un éventuel écart de timbre entre les longues et les brèves des quatre corpus, la Figure 4.4 représente les densités de probabilité des voyelles brèves et longues dans l'espace F1-F2[45]. Des classifieurs logistiques visant à prédire la longueur vocalique (brève ou longue) à partir des valeurs des F1 et F2 (le détail des analyses est donné dans Ferragne, 2020) ont été entrainés. Les modèles pour GLA1 et Hull prédisent un effet significatif, quoique modeste, de F1 et F2 respectivement. Il est donc possible que les auditeurs utilisent ces différences comme indices secondaires pour identifier la distinction entre longues et brèves. Cependant, il se peut, d'une part, que cette différence soit trop faible pour être perçue et, d'autre part, que certains individus présentent cette distinction de timbre, alors que d'autres, non.

La Figure 4.5 illustre le degré de séparabilité entre voyelles longues et brèves pour chaque locuteur. J'ai utilisé la valeur de silhouette calculée dans l'espace F1-F2 ; cette valeur reflète la distance entre une occurrence de voyelle et les autres voyelles de sa catégorie de longueur par rapport à la distance de cette même voyelle avec les voyelles de l'autre catégorie de longueur. La valeur moyenne de silhouette par locuteur nous donne une indication de la qualité de la partition entre longues et brèves sur la base des mesures de formants. D'après Everitt (2011), une valeur moyenne de 0,2 dénote une absence

45. Les valeurs de formants ont été extraites manuellement au milieu temporel de chaque voyelle, converties en Bark et centrées-réduites par locuteur, sauf pour GLA2 puiqu'il n'y avait que 2 voyelles par locuteur ; les densités bivariées incluent 99 % de la distribution estimée. Les axes ont été inversés ; cette figure se lit donc comme un espace acoustique phonétique classique.

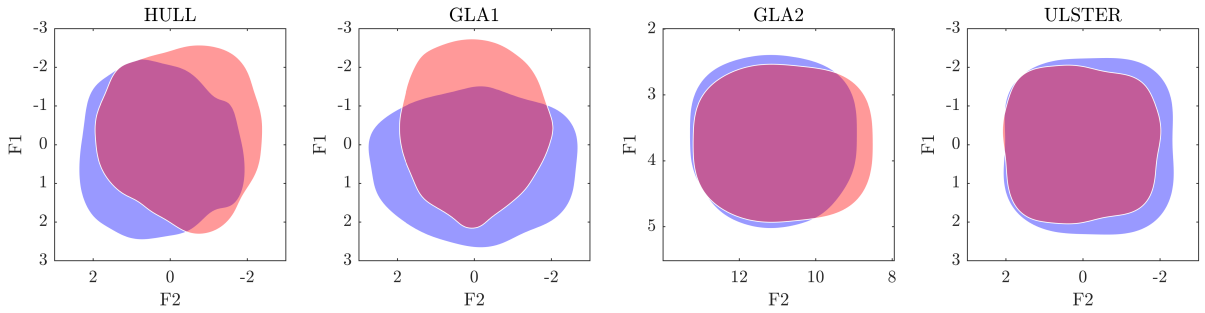



FIGURE 4.4 – Formants des voyelles de TRAP, START, GOOSE et BREWED.

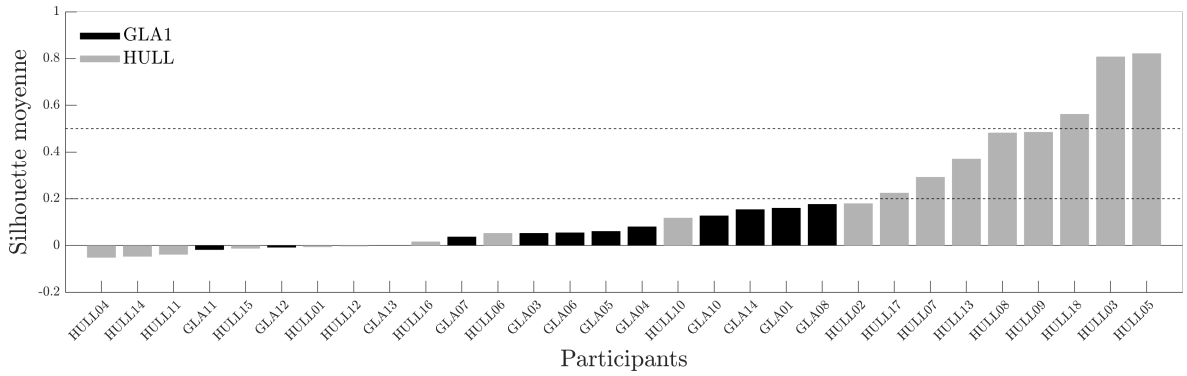



FIGURE 4.5 – Valeurs de silhouette moyennes par locuteur illustrant la séparabilité entre longues et brèves à partir des valeurs de formants.

de partition fiable, alors qu'une valeur supérieure à 0,5 reflète une bonne séparabilité. Les lignes en pointillés dans la Figure 4.5 matérialisent ces deux seuils. On y voit donc qu'aucun des locuteurs de GLA1 ne semble pouvoir prétendre à une distinction de timbre fiable. Quant à Hull, ce sont 3 locuteurs qui affichent des valeurs au-dessus du seuil de 0,5. Cette figure illustre bien la nécessité d'examiner le niveau individuel, en particulier quand on travaille sur les accents. Sur cette analyse en particulier, on pourra consulter Ferragne (2020) pour plus de détails.

#### 4.4.2.4 Note sur la perception de la durée

Revenons à présent sur le lien entre représentations objectives des mesures de durée comme dans la Figure 4.1 et les représentations cognitives des voyelles longues et brèves correspondantes. Si, comme dans le cas du modèle de Pierrehumbert (2001), les représentations mentales sont organisées dans un espace semblable à l'espace acoustique initial, il faut néanmoins tenir compte de la résolution du système perceptif. Pierrehumbert (2001) parle de « granularization », une étape qui consiste à discrétiser l'espace de sorte que deux stimuli qui ne peuvent être distingués soient stockés comme une seule et même entité. Ceci

conduit à un premier décalage avec la Figure 4.1 : le caractère continu des densités dans cette figure n'est pas justifié sur un plan perceptif ; un histogramme serait graphiquement plus juste. Logiquement, la question suivante consiste à se demander schématiquement quelle serait la taille des boîtes de cet histogramme. Intuitivement, s'appuyer sur une mesure de seuil différentiel de perception, comme le suggère Pierrehumbert (2001) fait sens. L'étude de Rossi (1972) sur la durée indique que le seuil de durée est absolu et constant, d'environ 30 ms, pour les voyelles de moins de 130 ms, puis pour les voyelles de 130 à 290 ms, il suit la loi de Weber, avec un rapport constant d'environ 22,5 %. Autrement dit, à partir de 130 ms, les boîtes de notre histogramme sont de taille croissante, ce qui suggère peut-être une perception logarithmique de la durée.

### 4.4.3 Modélisation des trajectoires formantiques

Parmi les contrastes dérivés typiques de l'anglais d'Écosse, celui qui implique les voyelles de *tide* et *tied* se prête particulièrement bien à une modélisation des contours formantiques impliquant des techniques d'analyse de séries temporelles telles que la transformée en cosinus discrète (DCT). Contrairement aux contrastes de type GOOSE-BREWED, pour lesquels il apparaît que la durée constitue l'indice essentiel matérialisant la différence entre les deux voyelles, le contraste qui nous intéresse ici implique à la fois la durée et le spectre. En effet, le membre monomorphémique de la paire, e.g. *tide*, présente une durée inférieure et une réalisation de type [ʌi], alors que l'item morphologiquement complexe, e.g. *tied*, affiche une durée supérieure et une qualité de type [ɑe] (Rathcke et Stuart-Smith, 2016 ; Ferragne, 2020).

Les données pour cette partie de l'étude comportent 12 locuteurs de Glasgow produisant 3 répétitions de 10 mots différents ; au total, donc, 360 voyelles ont été analysées. Les 10 mots représentent en réalité 5 paires impliquant le contraste dérivé qui nous intéresse : *bide-byed, pride-pried, ride-ryed, side-sighed* et *tide-tied*. Les valeurs de formants ont été mesurées avec le script `cp_formants`, que j'ai conçu pour le CminR Praatik (voir Section 3.3.2). Une particularité de la méthodologie que j'ai choisie tient au fait que le pas d'analyse est fixe (une valeurs toutes les 5 ms), ce qui n'est pas le cas de l'analyse par défaut proposée dans `Praat`[46].

La démarche de modélisation des contours formantiques avec la DCT comporte deux

46. Les analyses présentées dans ce document ont été réalisées avec `Matlab` ; je propose cependant l'équivalent avec le logiciel `R` pour que le plus grand nombre puisse en profiter à l'adresse suivante : https://tinyurl.com/xbxxkxbk ; les valeurs de formants mesurées, dont je me sers dans ce chapitre, sont disponibles publiquement à partir de mon Github à l'adresse https://tinyurl.com/4nxzrepp.

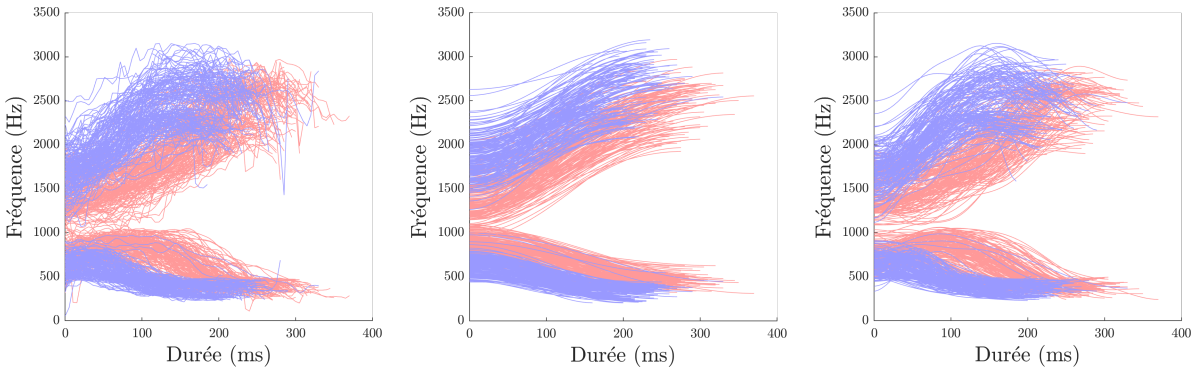



FIGURE 4.6 – Trajectoires formantiques des voyelles de type *tide* (bleu) et *tied* (rouge). À gauche : valeurs brutes. Au milieu : lissage par transformée en cosinus discrète (DCT) avec les 2 premiers coefficients. À droite : lissage DCT avec 6 coefficients.

choix que je vais commenter ici. Le premier concerne le degré de simplification du contour ; le second, les manipulations éventuelles de la base de temps.

Pour illustrer le premier point, je propose une série de graphiques permettant de cerner les enjeux des représentations qui nous intéressent. La Figure 4.6 montre différents degrés de lissage des trajectoires formantiques. Les deux premiers formants des voyelles de type *tide* sont matérialisés par les courbes bleues, ceux des voyelles de type *tied*, par les courbes rouges. Le graphique de gauche affiche les données brutes ; on y décèle les micro-variations inhérentes aux approximations de l'appareil phonatoire et de la méthode d'estimation. Ces fluctuations ne reflètent probablement pas l'intention articulatoire première, et elles sont certainement trop fines pour avoir un impact sur la perception. Les deux autres panneaux présentent un lissage de ces données aux moyens de la transformée en cosinus discrète (DCT). Le graphique du milieu ne retient que les 2 premiers coefficients. La stylisation est évidente ; elle corrige en particulier les mouvements abrupts de certaines courbes de F2 en bleu très visibles dans la première figure. Néanmoins, le lissage est peut-être trop marqué, et le troisième graphe (avec 6 coefficients DCT) propose une alternative peut-être moins restrictive. En résumé, comme on vient de le voir dans la Figure 4.6, la première étape consiste à effectuer un lissage qui constitue une première abstraction par rapport aux données brutes (afin de ne pas « coller » aux moindres micro-variations) tout en offrant une approximation réaliste, quoique parcimonieuse.

La deuxième question qui se pose concerne la base de temps : faut-il garder les durées originales ou bien normaliser la durée ? Dans le premier cas, les contours conservent leurs longueurs variables (comme dans la Figure 4.6), dans le second, on procède à une interpolation conduisant à l'obtention d'un nombre fixe de valeurs par contour. Garder

l'échelle temporelle initiale garantit la possibilité d'analyser le contour comme une série temporelle, c'est-à-dire, par exemple, de procéder à une analyse fréquentielle du contour comme s'il s'agissait d'un signal. Normaliser la durée compromet ce type d'analyse mais autorise des représentations graphiques peut-être plus accessibles, au prix, évidemment, d'une certaine simplification de la réalité ; c'est l'option retenue le plus souvent (Watson et Harrington, 1999 ; Fox et Jacewicz, 2009 ; Williams et Escudero, 2014 ; Elvin *et al.*, 2016). De ce que j'ai pu observer dans la pratique, le choix de normaliser la durée s'opère souvent par défaut à cause de contraintes techniques qu'il est pourtant possible de contourner. En effet, disposer d'un nombre équivalent de valeurs pour chaque contour après normalisation de la durée permet de gérer l'ensemble des données à partir d'un fichier plat, c'est-à-dire d'un simple tableau en 2 dimensions qu'on peut visualiser et qu'un logiciel ne peinera pas à convertir en graphique. L'option la plus adaptée conduit à avoir recours à des objets informatiques plus complexes, par exemple de type `list` dans `R` ou `cell` et `struct` dans `Matlab`, qu'il faut manipuler avec des méthodes moins accessibles[47].

Dans la première vignette de la Figure 4.7, j'illustre ce que font la plupart des études, à savoir, étirer ou compresser tous les contours de sorte qu'ils aient la même durée. C'est une pratique que je n'encourage cependant pas en amont d'une analyse DCT. En effet, la DCT est une analyse fréquentielle : manipuler la durée des contours par commodité méthodologique revient à accélérer ou ralentir la vitesse de lecture d'un disque vinyle pour ensuite s'étonner que la hauteur de la musique ait changé… Dans la deuxième vignette, la durée des longues et des brèves a été ramenée à la durée moyenne de chaque catégorie. On ne trouve généralement pas ce genre de représentation, mais il est peut-être plus fidèle aux données que la première vignette. Enfin, la troisième vignette neutralise la position de départ pour F1 et F2 séparément et permet de se concentrer sur la trajectoire, avec des données représentées sur une échelle psycho-acoustique. Ces illustrations rappellent combien il y est important de représenter graphiquement les données avant toute analyse, et exemplifient encore une fois le principe d'intégrité graphique de Tufte (2001) : même en l'absence de volonté délibérée de manipuler des données, les représentations qu'on élabore comportent toujours des choix qui distordent plus ou moins la réalité.

Au final, l'analyse que je propose est récapitulée dans la Figure 4.8 : les contours originaux de F1 et F2 sont soumis à une DCT et les 3 premiers coefficients de cette analyse sont représentés deux à deux, séparément pour chaque formant. La bonne séparation

47. Pour un code élégant, éviter les boucles conduit à utiliser des fonctions de pseudo vectorisation, type `cellfun` dans `Matlab`, qui, en plus de nécessiter souvent le recours à des fonctions anonymes, peu intuitives, n'occasionnent pas toujours un gain en termes de performances (Altman, 2015).

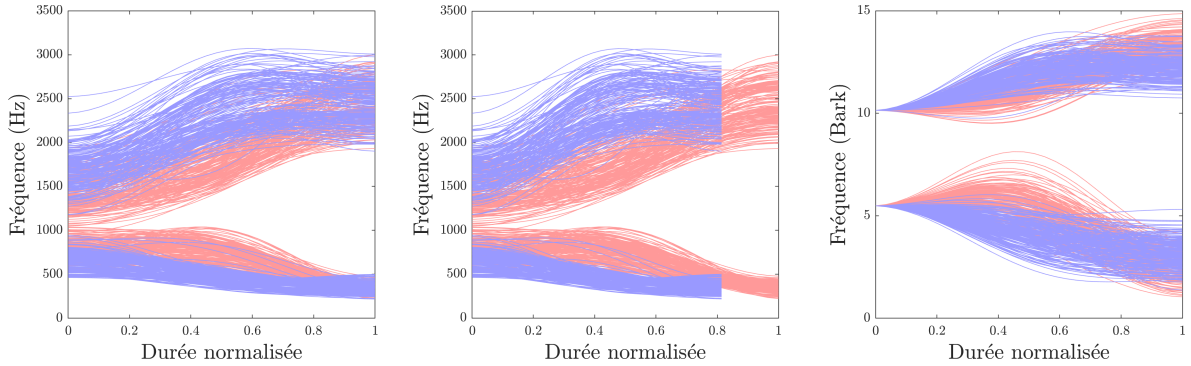



FIGURE 4.7 – Trajectoires formantiques des voyelles de type *tide* (bleu) et *tied* (rouge). À gauche : normalisation temporelle simple. Au milieu : normalisation temporelle par rapport à la durée moyenne des brèves et des longues indépendamment. À droite : normalisation temporelle simple, fréquences en Bark, origines centrées par rapport à la moyenne de chaque formant, brèves et longues confondues.

entre nuages de points des deux couleurs conduit à penser que ces paramètres sont tout à fait pertinents pour caractériser la différence entre les deux types de voyelles qui nous occupent.

L'interprétation de ces coefficients en termes phonétiques est la suivante (voir e.g. Elvin *et al.*, 2016) : le premier coefficient, souvent omis des analyses phonétiques, représente l'*offset*, c'est-à-dire, le décalage moyen du contour sur l'axe vertical. Le deuxième traduit la magnitude et la direction de la pente formantique, et le troisième, la courbure. La Figure 4.8 fait ressortir que le 3$^{e}$ coefficient, à la fois sur F1 et F2, permet globalement de bien distinguer les deux classes.

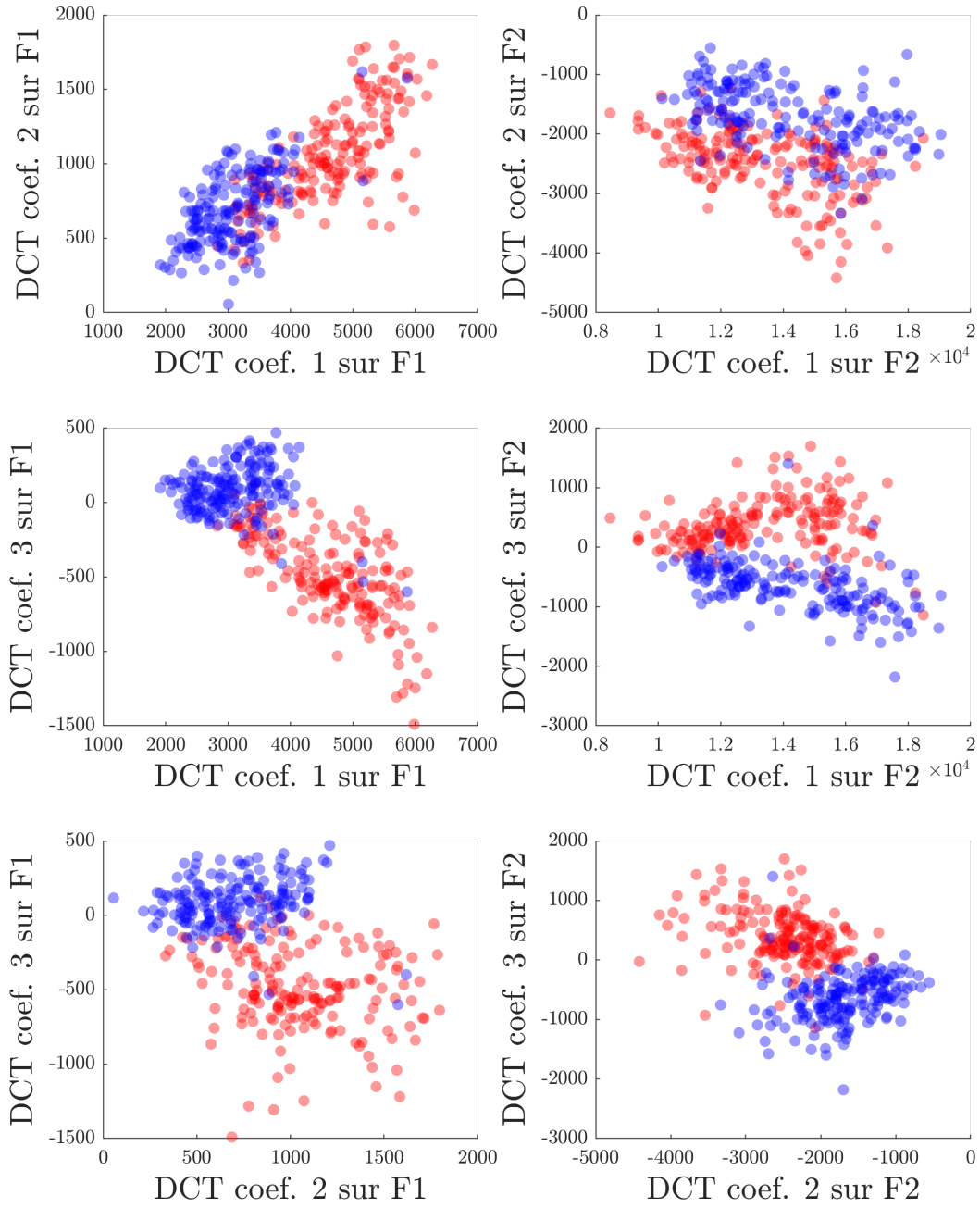



FIGURE 4.8 – Coefficients DCT pour les trajectoires de F1 et F2 des voyelles de type *tide* (bleu) et *tied* (rouge).

## 4.5 HPG en perception

L'essentiel de ce chapitre est consacré à la production. Cela est dû, en premier lieu, au fait que, comme je l'ai noté à la Section 4.1, il est très difficile de proposer des expériences de perception qui soient strictement contrôlées. Ensuite, et par voie de conséquence, les travaux que j'ai réalisés sur HPG en perception sont moins nombreux et plus anciens (Boulenger *et al.*, 2011 ; Ferragne *et al.*, 2011). Il me paraît néanmoins instructif de rappeler ici, fût-ce brièvement, les principaux résultats obtenus en perception.

La logique qui sous-tendait les expériences réalisées posait que les mesures caractérisant la frontière entre deux sons — qu'il s'agisse par exemple d'une courbe d'identification, de temps de réponses lors d'une tâche de discrimination, ou encore de l'amplitude variable d'une composante EEG spécifique — variait selon qu'on était en présence d'une différence allophonique, phonémique ou quasi-phonémique. Pour prendre l'exemple précis de la courbe sigmoïde typique qu'on obtient à l'issue d'une tâche d'identification, on s'attend à observer une pente plus abrupte lorsqu'on est en présence d'une opposition phonémique. Cette idée vient en partie de travaux en acquisition montrant qu'au fur et à mesure qu'une opposition phonémique est acquise chez des enfants au développement phonologique typique, la courbe de la fonction psychométrique d'identification est de plus en plus abrupte alors qu'elle est relativement peu abrupte chez des enfants souffrant de dyslexie, connus pour présenter une perception « allophonique » (Bogliotti *et al.*, 2008).

### 4.5.1 Expériences d'identification à Glasgow et Hull

La voyelle[48] d'une occurrence du mot *brewed* enregistré à Glasgow a été artificiellement étirée ou compressée dans le but de créer douze stimuli suivant un continuum de durée allant de 40 ms à 205 ms par pas de 15 ms. La voyelle d'une occurrence de *carp* à Hull a subi strictement le même type de modifications. Nous avions donc à Glasgow un continuum de durée correspondant à l'espace physique typique du contraste quasi-phonémique *brood-brewed*, et à Hull, un continuum équivalent caractéristique, quant à lui, d'un contraste phonémique.

Les participants écossais ont été invités à écouter les stimuli du continuum *brood-brewed* et à dire s'ils entendaient l'un ou l'autre. La même procédure, cette fois-ci pour *cap-carp*, a été proposée aux participants de Hull. Des fonctions logistiques ont été ajustées aux

48. Les expériences mentionnées dans cette section sont décrites en détail dans Boulenger *et al.* (2011), Ferragne *et al.* (2011) et Ferragne (2020).

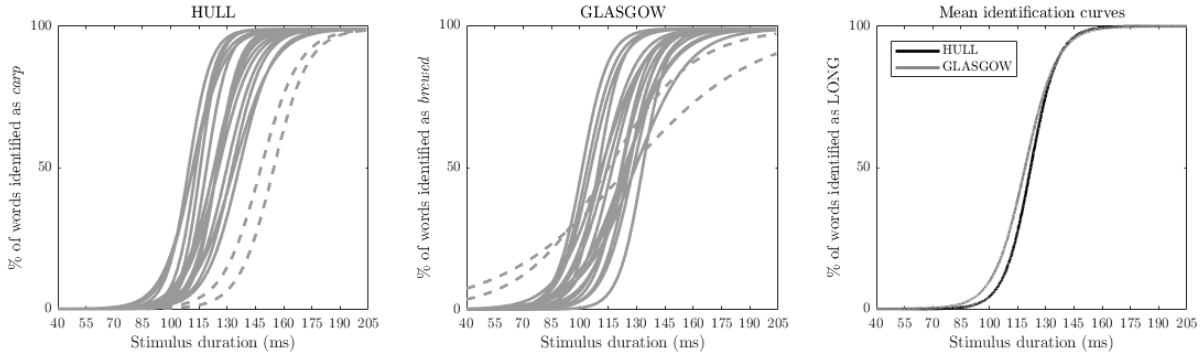



FIGURE 4.9 – Courbes d'identification individuelles (1er et 2e graphe) et moyennes (3e graphe) caractérisant la perception de *brood-brewed* à Glasgow et *cap-carp* à Hull. Les courbes en pointillés, visuellement déviantes en termes de seuil ou de pente, n'ont pas été incluses dans les calculs statistiques.

pourcentages d'identification de la voyelle la plus longue ; elles sont présentées dans la Figure 4.9.

Les résultats montrent une différence significative en termes de pente, mettant en évidence une fonction logistique plus abrupte à Hull, c'est-à-dire dans le cas d'une opposition phonémique par rapport à la perception d'un contraste quasi-phonémique (conformément à nos attentes).

L'expérience comporte évidemment des biais : on ne peut pas exclure que la différence de pente soit imputable à une différence entre les groupes d'auditeurs ; on ne peut pas non plus écarter un effet auditif/acoustique propre à la voyelle. Cet aspect illustre bien d'ailleurs toute la complexité de la mise en place d'expériences de perception contrôlées dans le cadre de HPG. À l'époque, j'étais pleinement conscient de ces limites, mais j'avais choisi l'option de commencer néanmoins à procéder à ces expériences plutôt qu'attendre de réunir les conditions idéales[49].

### 4.5.2 Expériences d'identification à Glasgow et Lyon

L'expérience d'identification menée à Glasgow a également été administrée à des locuteurs francophones (Ferragne *et al.*, 2011). Il s'agissait d'établir avec ces derniers une courbe de référence matérialisant une perception purement allophonique. Leur tâche était en revanche légèrement différente : ils étaient en effet soumis dans un premier temps à une phase d'entraînement pendant laquelle ils étaient invités à distinguer un « mot

49. Avec le recul, il semble évident que, en première approximation, il aurait été profitable de présenter tous les stimuli — *cap-carp* et *brood-brewed* — aux deux groupes. Cependant, les deux expériences n'ont pas été réalisés simultanément ; les enregistrements de Hull ont eu lieu un an après l'expérience de Glasgow.

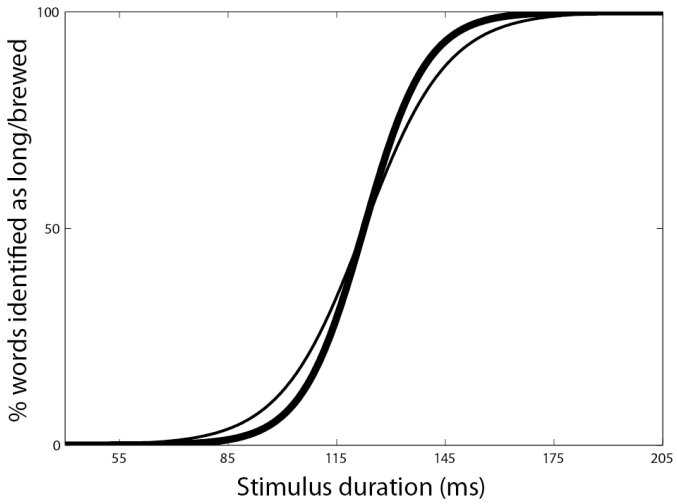



FIGURE 4.10 – Fonctions d'identification moyennes pour la paire *brood-brewed* par des participants écossais (ligne épaisse) et français (ligne fine).

court »— *brewed* dont la voyelle a été raccourcie de sorte que sa durée soit équivalente à la durée moyenne de la voyelle de *brood* (70 ms) — d'un « mot long », *brewed* avec une durée vocalique ramenée à la durée moyenne de ce type de mots (175 ms). Une fois la distinction apprise, ils étaient soumis à la même expérience que leurs homologues de Glasgow, à ceci près qu'ils devaient décider si le stimulus entendu était court ou long. Les fonctions d'identification moyennes sont représentées dans la Figure 4.10. La courbe en trait fin, correspondant au groupe de francophones, affiche une pente statistiquement moins abrupte que la courbe épaisse du groupe écossais.

Les résultats semblent aller dans le sens prédit, mais là encore, ce qui est comparé n'est pas tout à fait comparable. L'acquisition naturelle d'une frontière catégorielle se fait sur la base d'une exposition à des stimuli variant également à l'intérieur de chaque catégorie. La tâche pour le moins artificielle confiée aux francophones a été d'apprendre deux catégories seulement à partir du prototype de chacune. D'ailleurs la durée de ces deux prototypes a été inférée à partir des mesures réalisées sur les données de production de l'ensemble GLA1 (Section 4.4.1) ; or nous n'avons aucune garantie qu'il s'agisse des durées prototypiques authentiques telles qu'elles sont représentées dans l'esprit des participants écossais. En imaginant que ces véritables durées prototypiques soient plus proches l'une de l'autre (par ex. 80 et 160 ms plutôt que 70 et 175 ms), une pente d'identification plus marquée découlerait logiquement.

## 4.6 Conclusion

Ce chapitre débute en évoquant la transition de mes travaux de thèse vers la formulation de l'hypothèse de la phonémicité gradiente (HPG). En parallèle de la présentation de HPG, j'ai rappelé quelques résultats obtenus en production, en mettant un accent particulier sur le raisonnement méthodologique. J'ai délibérément privilégié les représentations graphiques et omis, pour davantage de lisibilité, les sempiternels tests statistiques, qu'on retrouve néanmoins dans mes publications, comme Ferragne (2020). Au Chapitre 2, je mettais en avant mon engagement pour une culture du graphique et pour une utilisation moins systématique des tests d'inférence statistique ; ce que je présente ici est donc conforme à ces ambitions. J'ai ensuite rapidement évoqué les tests de HPG en perception. Ce volet illustre en particulier la grande complexité requise pour la mise en place de tests qui soient expérimentalement valides.

# 5

# La diversité des approches comme credo

## Sommaire



## 5.1 Introduction

On reproche souvent aux étudiants de faire des catalogues ; c'est pourtant ce que je m'apprête à faire. Je voulais en effet recenser ici un certain nombre de mes travaux qui ne méritaient pas à eux-seuls qu'un chapitre entier leur soit consacré. Cette liste, qui va de la collaboration ponctuelle à la thématique malchanceuse vite abandonnée, en passant par des objets qui reviendront probablement dans mes priorités si le contexte s'y prête, présente cependant plusieurs points intéressants. Le premier, c'est que ces projets plus « secondaires » ont contribué au succès des autres. Ce que j'ai pu apprendre ponctuellement d'une collaboration, qu'il s'agisse d'un savoir ou d'un savoir-faire, a eu des retombées

sur le reste de ma recherche. Le second intérêt, c'est que cela permet, encore une fois, de contextualiser ma recherche car sans contexte, certains choix que j'ai opérés dans mon parcours peuvent paraître arbitraires. Enfin, un chapitre complet consacré par exemple à mes recherches sur le rythme serait certes surdimensionné, mais ne pas en parler du tout serait un oubli important dans le cadre de la synthèse de mes recherches.

## 5.2 Acquisition de l'anglais par des francophones

### 5.2.1 Motivations

Comme je l'ai expliqué dans l'Introduction (Chapitre 1), les raisons qui m'ont conduit à travailler sur l'acquisition de l'anglais par les francophones sont avant tout pragmatiques et opportunistes ; il ne s'agit pas d'un goût particulier pour ce domaine précis. Lassé par les innombrables biais que comportent les corpus qui n'ont pas été collectés pour répondre à une question précise, j'avais fait le choix de ne plus recueillir que des données *ad hoc* (voir Section 2.3.1), et il fallait donc que la population de locuteurs ou d'auditeurs de mes études soit facilement accessible. J'aurais pu choisir de consacrer un sous-ensemble de mes travaux au français car il y avait certains phénomènes phonologiques marginaux — je pense en particulier à la gémination, à l'opposition /ɛ̃/-/œ̃/, et au phonème /ɲ/ — qui rentraient dans le cadre de l'hypothèse de la phonémicité gradiente (voir Section 4.3) et pour lesquels j'avais mené des études pilotes. J'ai renoncé à cette option car étant maître de conférences dans le champ des Études Anglophones, on attend de moi que l'essentiel de mes travaux soit consacré à l'anglais.

À cette époque-là, vers 2013, mon choix de travailler sur l'anglais L2 n'a donc pas été guidé par une quelconque mode ou une volonté de me spécialiser dans la didactique des langues, mais bien plutôt par l'envie de confronter mes propres méthodes de recherche au domaine de l'apprentissage de l'anglais. La difficulté de mener des expériences à l'étranger quand elles impliquent un matériel spécifique a constitué un argument supplémentaire en faveur de cette nouvelle orientation.

Un autre événement qui a renforcé mon orientation vers l'acquisition des langues étrangères a eu lieu lorsque, sous l'impulsion de Jacqueline Vaissière, nous avons discuté des projets que nous pourrions élaborer dans ce domaine avec Cédric Gendrot et Frédéric Isel. Nous avons entre autres obtenu un financement pour le projet *Solutions Phonétiques pour les Cours de Langues Étrangères* (SOPHOCLE) en réponse à l'appel à projet de l'IDEX

USPC[50].

Les apprenants de l'anglais ont donc constitué une cible cohérente avec mes nouveaux choix épistémologiques. Deux points très positifs ont naturellement découlé de ce changement : d'abord, étant enseignant d'anglais, ce choix me permettait de créer un lien fort entre ma pratique d'enseignant et ma recherche. Et ensuite, puisque je suis amené à diriger les travaux de recherches de futurs enseignants d'anglais dans le secondaire, il paraissait raisonnable de pouvoir leur proposer des thèmes en rapport avec leur carrière à venir.

L'essentiel des travaux auxquels j'ai participé sur la prononciation de l'anglais par des apprenants francophones comportait des entraînements. Pour que cela soit fait dans un cadre expérimental strictement contrôlé, il faut être en mesure de libérer plusieurs semaines entières pour s'y consacrer. Ce sont donc des travaux qui ne peuvent généralement pas être menés de bout en bout par un enseignant-chercheur, mais plutôt en collaboration avec des (post-) doctorants. Les travaux que je vais brièvement résumer (voire seulement survoler, pour des raisons évidentes de place) concernent donc en particulier les travaux en lien avec la thèse de Jennifer Krzonowski , de Maud Pélissier, ainsi que les travaux menés pendant que Karin Heidlmayr était en post-doctorat dans le cadre du projet SOPHOCLE.

### 5.2.2 Acquisition des voyelles de l'anglais

L'acquisition des voyelles de l'anglais par des apprenants francophones est fortement liée à ma pratique de l'enseignement de la prononciation de l'anglais. La thèse de Jennifer Krzonowski avait pour but d'entrainer des apprenants tardifs à mieux distinguer des oppositions phonologiques de l'anglais dans deux zones de l'espace vocalique connues pour poser problème : d'une part, /iː/ - /ɪ/, et de l'autre, /ʌ/ - /æ/ - /ɑ/. Une première étape a consisté à établir les différences en termes de production des voyelles du français et de l'anglais sur la base de paramètres tels que la valeur des formants dans l'état stable, la trajectoire formantique mesurée par le biais de la transformée en cosinus discrète, et la durée ; tout ceci est décrit dans Krzonowski *et al.* (2018). Après que les productions des hommes et des femmes ont été séparées, la comparaison a été double puisque ce sont d'une part les voyelles du français produites par des francophones qui ont été comparées à des voyelles de l'anglais produites par des anglophones, et d'autre part des voyelles de l'anglais, puis du français par les mêmes locuteurs francophones. Ce type d'analyse

50. Là encore le contexte a été déterminant puisque après un an de financement, le projet s'est arrêté brusquement suite à l'arrêt pour le moins inattendu de l'IDEX USPC.

a pour ambition de déterminer très précisément quelles étaient les stratégies possibles pour des apprenants, en postulant un transfert simple, c'est-à-dire la réutilisation d'un schéma articulatoire de la langue maternelle pour produire les sons de la L2. Autrement dit : les réalisations canoniques de quels phonèmes du français présentaient des valeurs acoustiques assez proches des phonèmes cibles de la L2 pour pouvoir s'y substituer. Certes, ces possibilités de transfert sont intuitivement connues des enseignants et des manuels universitaires depuis longtemps, mais quantifier précisément la dispersion des réalisations dans F1-F2 et le taux de chevauchement entre catégories dans une même langue et d'une langue à l'autre a été très utile.

Dans Krzonowski *et al.* (2016), nous rapportons les effets d'entraînements en production et en perception sur la perception et la production des contrastes vocaliques des régions critiques identifiées au paragraphe précédent par des apprenants francophones. Les participants ont été divisés en 3 groupes. Le groupe PE a suivi un entraînement sur 5 séances s'appuyant sur des tâches d'identification et de discrimination. À chaque essai, un feedback de type « correct » ou « incorrect » était donné à l'écran. Le groupe PR a suivi des séances de répétitions de mots pendant lesquelles un feedback était donné par le biais d'une interface développée pour l'occasion présentée dans la Figure 3.12 à la page 52 du présent document. Sur l'écran, un point matérialisait les coordonnées de la voyelle à imiter dans l'espace F1-F2, et un autre point indiquait la position de la voyelle que le participant venait juste de prononcer. Il y avait également un indicateur de la durée de la voyelle cible et de celle produite par le participant. Les participants avaient ainsi tout le loisir d'ajuster leur production en faisant en sorte que les deux points dans F1-F2 soient le plus proche possible. Il y avait finalement un troisième groupe, C, qui faisait office de groupe contrôle, et qui n'était donc pas entraîné à proprement parler mais était invité à écouter des audio-books en anglais pour une durée équivalente aux entraînements des autres groupes. Les tests, avant et après entraînement, comportaient une tâche d'identification de voyelles, une tâche de discrimination et une tâche de production[51].

Pour être très synthétique, les résultats montrent une amélioration des pourcentages d'identification des voyelles après entraînement dans les groupes PE et PR (excepté pour la voyelle /ɑː/ dans le groupe PR). On note que l'identification s'améliore aussi pour les voyelles /ɑː/ et /ɪ/ dans le groupe C. Le pourcentage de discrimination correcte s'améliore après entraînement dans les groupes PE et PR pour toutes les paires concernées : /æ/-/ʌ/, /ɑː/-/ʌ/ et /iː/-/ɪ/. On note aussi une amélioration pour les paires /æ/-/ʌ/ et /ɑː/-/ʌ/

51. Il y avait aussi des enregistrements de potentiels évoqués en EEG avec un paradigme de type *oddball*. Je n'en parlerai pas car une erreur technique a compromis l'analyse de cette partie de l'expérience.

dans le groupe contrôle, ce qui fait qu'en réalité, si on enregistre bien une différence significative entre PE et C, on ne parvient pas à en trouver une entre PR et C.

Pour l'évaluation de la production, les apprenants étaient invités à prononcer des mots isolés contenant chacune des 5 voyelles concernées, par ex., *bad, bud, bard, bid, bead.* L'évaluation était faite, après avoir mesuré les deux premiers formants et la durée, en comparant les espaces vocaliques des apprenants avant et après entraînement avec ceux de locuteurs natifs. Cette comparaison était faite au moyen de la méthode que j'avais utilisée dans ma thèse : les distances entre les voyelles prises deux à deux de l'apprenant sont comparées aux mêmes distances enregistrées chez des locuteurs natifs par le biais d'un coefficient de corrélation. Ainsi, la corrélation permet une forme de normalisation du locuteur, et le fait de prendre toutes les distances à la fois permet d'évaluer la structure générale du système.

Pour cette partie de l'analyse, nous nous étions heurtés à un problème méthodologique : il y avait des disparités de niveaux d'un groupe à l'autre avant entraînement. Une amélioration a été enregistrée pour les groupes PE et PR, à condition qu'on les analyse séparément (i.e. sans prendre en compte le groupe C), ce qui, évidemment, n'est pas tout à fait conforme au plan expérimental initial puisqu'on ne peut pas choisir a posteriori d'écarter le groupe témoin, qui est précisément là pour nous dire si le succès de notre entraînement n'est pas dû à une simple exposition à de l'anglais.

### 5.2.3 Violations morphosyntaxiques

La thèse de Maud Pélissier, *Effets d'Entraînements Explicites et Implicites sur l'Acquisition de la Syntaxe de l'Anglais par des Apprenants Francophones : Étude en Potentiels Évoqués*, soutenue en 2018, paraît quelque peu détoner dans mon parcours. Mais de nombreux facteurs liés au contexte justifient ce que nous faisons. Dans ce cas précis, ma volonté de bien distinguer ses travaux de ceux de Jennifer Krzonowski, et l'opportunité de disposer d'un système EEG et de financements associés sur des projets en lien avec l'acquisition ont fortement déterminé le sujet. Je ne vais pas synthétiser ce travail, ce qui nécessiterait au moins un chapitre complet, mais simplement illustrer brièvement les deux expériences mises en place, avec l'idée que les lecteurs curieux pourront consulter directement le manuscrit.

Cette thèse s'articule donc autour de deux expériences combinant (de nombreuses) réponses comportementales et EEG, assorties d'une période d'entraînement entre prétest et post-test. Dans une première expérience portant sur la morphologie des verbes en

TABLEAU 5.1 – Types de stimuli dans l'expérience de violation morphosyntaxique impliquant le morphème du passé.

| Similaire à L1 | Correct | Stimulus |
|---|---|---|
| - | + | Did Mary finish our dinner ? |
| - | - | Did Mary finished our dinner ? |
| + | + | Had Mary finished our dinner ? |
| + | - | Had Mary finish our dinner ? |

anglais, des apprenants francophones ont été séparés en deux groupes et soumis à deux types d'entraînement ; l'un, explicite — des « cours » sur la morphologie du passé avec les auxiliaires *had* et *did* ; l'autre, implicite : l'écoute de phrases comportant ces mêmes auxiliaires. Ce premier facteur, le type d'entraînement, a été croisé avec un second : la similarité des structures entre L1 et L2. La construction avec *had*, qui implique l'utilisation du participe passé, est en effet plus proche du français que celle avec *did*, qui implique l'utilisation de la base verbale. Lors des phases de test, la moitié des phrases étaient rendues agrammaticales en supprimant ou en ajoutant le morphème du passé au radical du verbe lexical. Un groupe d'anglophones a également été soumis aux tests afin d'offrir une base de comparaison. Les types de stimuli sont récapitulés dans le Tableau 5.1. Vu l'ampleur des résultats, je ne retiendrai que quelques points-clés. D'abord, les jugements de grammaticalité s'améliorent pour les apprenants : la détection des phrases grammaticales atteint le niveau des locuteurs natifs en post-test alors que la détection des phrases agrammaticales continue d'être en-dessous des performances des anglophones. Cependant, le groupe ayant suivi l'entraînement explicite a mieux détecté les violations que le groupe implicite en post-test. Pour expliquer cette asymétrie entre le traitement des phrases grammaticales et agrammaticales, il est établi dans la littérature que, dans des tests de jugements grammaticaux, les premières testent plutôt des connaissances implicites alors que les secondes mobilisent les connaissances explicites (Pélissier, 2018, p. 177). On retiendra aussi par exemple que les apprenants du groupe explicite étaient plus performants dans le traitement des phrases comportant la structure différente du français (avec *did*) que les apprenants du groupe implicite.

Pour ce qui est de l'analyse des potentiels évoqués, pour ce type de violation, on attend plutôt une réponse biphasique composée d'une LAN (*Left Anterior Negativity*) suivie d'une P600. Or la variabilité a été de mise chez les locuteurs natifs, et c'est davantage un schéma N400 puis P600 qui a été observé. Chez les apprenants, on note que les

schémas évoluent entre pré-test et post-test ; cependant les résultats de cette partie sont particulièrement riches, et il serait fastidieux de les rappeler ici.

La seconde expérience s'est penchée sur la résolution d'ambiguïtés syntaxiques à partir de la prosodie. Les stimuli des pré-test et post-test de cette expérience étaient constitués de phrases comprenant un premier verbe suivi d'un syntagme nominal qui pouvait être soit l'objet de ce verbe, soit le sujet du syntagme verbal suivant. Ainsi dans l'Exemple 1, la position de la frontière prosodique, matérialisée par un dièse, favorise la bonne interprétation, selon laquelle *the ducks* n'est pas objet de *hunting*. À l'inverse, dans l'Exemple 2, la position de la frontière indique que *the ducks* est bien objet de *hunting*. Ces deux conditions avec prosodie congruente étaient accompagnées de deux conditions avec prosodie incongruente (présentées dans les Exemples 3 et 4), obtenues à partir de la combinaison des deux autres types.

(1) As the man is hunting # the ducks fly away.

(2) As the man is hunting the ducks # the pigeons fly away.

(3) As the man is hunting # the ducks # the pigeons fly away.

(4) As the man is hunting the ducks fly away.

Les pré-test et post-test étaient constitués là aussi de jugements d'acceptabilité et d'enregistrements EEG en vue d'analyses de potentiels évoqués. Les composantes ERP pertinentes dans ce type d'études sont en particulier la *Closure Positive Shift* (CPS), qui reflète la détection d'une frontière prosodique et, bien sûr, la P600, corrélat de la réanalyse syntaxique typique de la phrase dans l'Exemple 4 où *the ducks*, d'abord compris comme objet de *hunting* à cause de l'absence de frontière prosodique, est ensuite réinterprété comme sujet. Sans entrer dans la richesse des résultats, on note par exemple que les entraînements ont contribué à une meilleure détection du caractère naturel et non naturel des phrases du type des Exemples 2 et 3 respectivement. Au niveau de l'EEG, l'onde CPS a bel et bien été observée chez les apprenants, ce qui suggère que la détection de frontière prosodique a bien lieu, comme cela se fait d'ailleurs dans leur langue maternelle. Les résultats pour l'onde P600 montrent par exemple le rôle primordial de la compétence des participants : ceux qui ne détectaient pas, d'après leurs réponses comportementales, l'incongruence entre prosodie et syntaxe dans la phrase de l'Exemple 3 n'avaient pas besoin de procéder à une réanalyse, et donc, ne présentaient pas de P600.

### 5.2.4 L'intonation des questions ouvertes et fermées

Dans Guyot-Talbot *et al.* (2016), nous cherchions à aider des apprenants francophones à améliorer leur production de contours intonatifs distincts en anglais, qu'on trouve dans 3 types de phrases : assertions, questions fermées et ouvertes. Les pré-test et post-test consistaient en la lecture d'exemples des 3 types de phrases, et entre les deux figuraient 3 séances d'entraînements où ils devaient répéter des phrases interrogatives prononcées par une anglophone.

Après chaque phrase, leur contour de f0 sur la syllabe portant le noyau intonatif ainsi que celui de la locutrice anglaise étaient affichés à l'écran. Il n'y avait pas de feedback de type correct vs incorrect, mais ces sessions devaient permettre aux apprenants d'inférer une règle du système intonatif de l'anglais : par défaut, les questions fermées ont un contour montant, et les questions ouvertes, un contour descendant. L'apprentissage de cette règle était implicite, dans le sens où aucune phase d'instruction ne présentait explicitement la règle sous-jacente. Puis, une nouvelle séance d'enregistrements permettait de collecter des phrases à comparer au pré-test pour juger l'efficacité de l'entraînement.

Les résultats montrent une corrélation plus grande en post-test entre les contours intonatifs des questions ouvertes des apprenants du groupe test et ceux du modèle anglophone. Le même gain est observé pour les questions fermées. Aucun effet n'est en revanche détecté pour les phrases assertives. Le groupe contrôle, quant à lui, n'affiche pas d'amélioration significative ; tout ceci est résumé dans la Figure 5.1.

Par ailleurs, si l'on écarte le modèle natif pour estimer une éventuelle divergence des contours des questions ouvertes et fermées entre pré-test et post-test, on constate, en s'appuyant sur une distance calculée après déformation temporelle dynamique (*dynamic time warping* : DTW), que les contours de ces deux types de questions divergent davantage en post-test, et pour le groupe test, et pour le groupe contrôle, mais l'interaction significative Séance × Groupe (Figure 5.2) montre que l'amélioration pour le groupe test est supérieure à celle du groupe contrôle.

Si on peut donc considérer que, sur un plan objectif, nos entraînements ont fonctionné, il y a quand même trois limites à cette étude que je dois mentionner. D'abord, on ne sait pas dans quelle mesure cette amélioration est perceptible. Il suffirait pour cela d'inviter un panel d'auditeurs, soit composé de locuteurs naïfs, soit d'enseignants, à juger les productions des apprenants. Ensuite, la dichotomie selon laquelle une question ouverte engendre un contour descendant et une question fermée, un contour montant, est très grossière. Car en effet ces contours peuvent varier en fonction du contexte. D'ailleurs, à

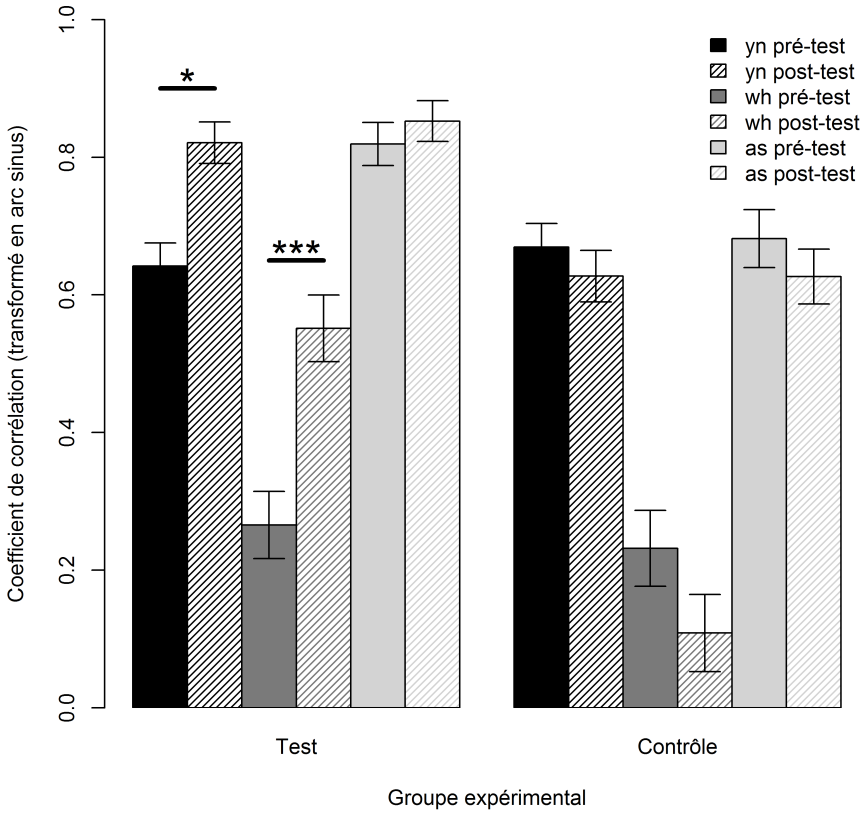



FIGURE 5.1 – Distance entre les contours des apprenants et du modèle natif : *yn* questions fermées, *wh* questions ouvertes, *as* phrases assertives.

l'époque, nous avions dû faire reprendre plusieurs phrases par notre locutrice modèle, qui ne produisait pas toujours spontanément le contour escompté. Enfin, pour pouvoir fournir un retour visuel avec les courbes de f0 qui se superposent, le système s'appuyait sur l'estimation automatique de f0 en temps réel après détection du mot portant le contour intonatif sur la base d'un alignement forcé. Ces deux techniques n'étant pas totalement fiables, l'entraînement lui-même comportait donc quelques faiblesses.

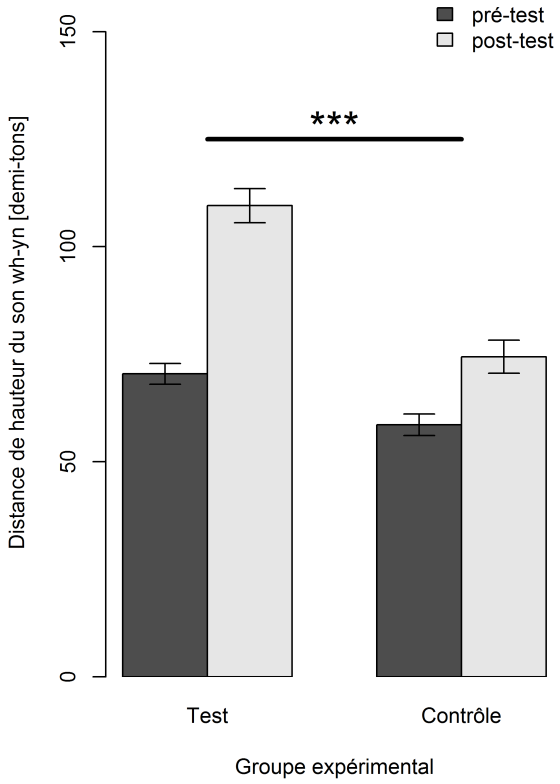



FIGURE 5.2 – Distance DTW entre questions ouvertes et fermées de chaque locuteur.

### 5.2.5 Émotions et langue seconde

L'article intitulé « The perception of swear words by French learners of English : an experiment involving electrodermal activity » (Rastovic *et al.*, 2019) illustre bien ce qui peut se produire quand se rencontrent l'opportunité de disposer d'une technique, la restriction de ne tester que des apprenants et l'envie de développer de nouvelles approches dans l'espoir qu'une étudiante puisse éventuellement se forger sa propre thématique de recherche. Les mots de notre langue maternelle, et a fortiori les mots grossiers, sont acquis dans le cadre d'une expérience multi-sensorielle associant des émotions, positives ou négatives, à des informations visuelles, tactiles, auditives précises. À l'inverse, l'acquisition tardive d'une L2 dans le contexte (ou plutôt l'absence de contexte) de la salle de classe marque une rupture nette avec les mécanismes de la L1. À contenu sémantique équivalent, on peut s'attendre à ce que la valence émotionnelle des mots de notre L2 soit plus neutre que dans notre L1. En d'autres termes, là où la gêne provoquée par un mot grossier peut en-

gendrer des réactions somatiques spécifiques (rougissements, tachycardie, sudation) dans notre L1, il est possible que ce ne soit pas le cas dans notre L2 ; le lexique grossier de notre L2 serait-il donc « désincarné » ? Pour cette expérience, nous avons utilisé des mesures de conductance électrodermale. Il s'agit d'une méthode non invasive qui mesure l'activité des glandes sudoripares censée refléter certains états émotionnels « automatiques ». Dans notre cas, nous nous attendions à observer une réponse électrodermale phasique de grande amplitude lorsque nos participants francophones entendaient des mots grossiers en français, mais pas en anglais. Il y avait également des mots, anglais et français, à valence émotionnelle neutre ; et aucune différence entre les deux langues n'était attendue dans ce cas. Les détails du protocole et de l'analyse sont à consulter dans l'article ; je présente ici l'essentiel des résultats dans la Figure 5.3. On y observe un effet principal du Type de mot ainsi qu'une interaction Type de mot × Langue : comme attendu, la réponse électrodermale de plus grande amplitude pour les mots grossiers est limitée au français. Pour l'anecdote, il s'agit là d'une des rares fois de ma carrière où, une fois le script d'analyse écrit, aucune intervention humaine n'a été nécessaire. En effet, alors que c'est souvent à la main qu'on rejette les artefacts du signal EEG, qu'on segmente du signal audio ou qu'on détermine des régions d'intérêt dans des échographies de la langue — et je ne parle pas du tri manuel et parfois un peu « sauvage » des locuteurs ou sujets d'expériences présentant des valeurs « déviantes » (ou en tout cas, qui contrarient les tests statistiques...) — dans le cas de cette expérience, le traitement des données a été entièrement automatique et donc reproductible. J'ignore encore dans quelle mesure cela est dû à la technique elle-même et non à la grande taille (supposée) de l'effet attendu, mais cela m'encourage à re-proposer des expériences s'appuyant sur cette technique à l'avenir.

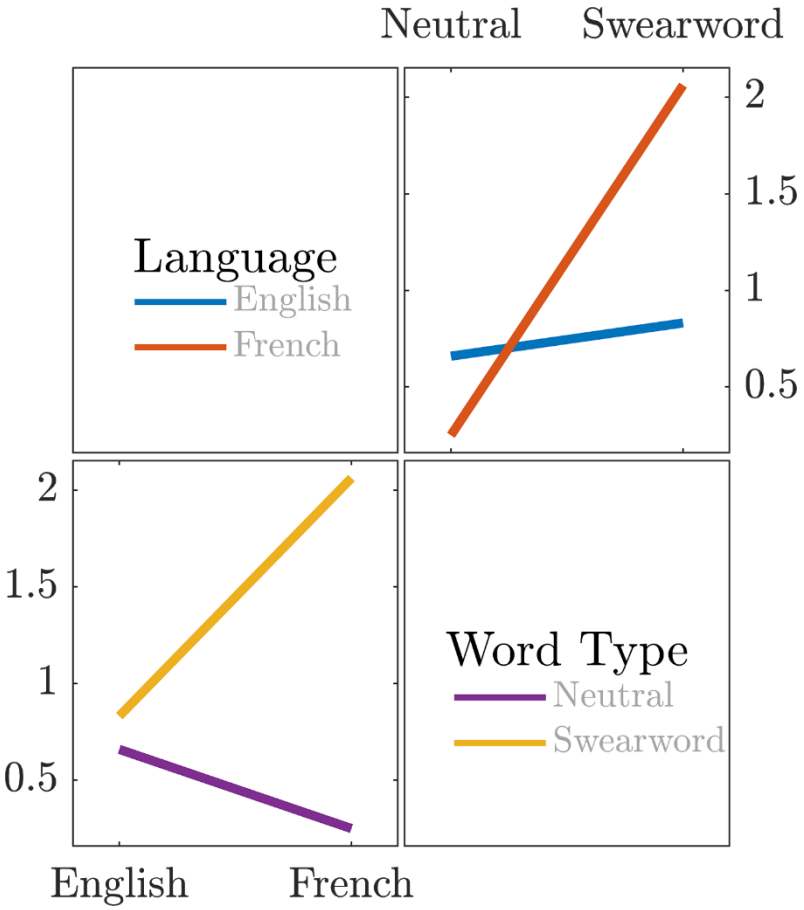



FIGURE 5.3 – Résultats de l'analyse de l'activité électrodermale.

## 5.3 Rythme et tempo

Chronologiquement, les premiers travaux empiriques que j'ai menés lors de ma thèse concernaient le rythme. Là encore, le contexte a été déterminant : au début des années 2000, nous étions particulièrement influencés par les travaux de Franck Ramus et ses collaborateurs (par ex. Ramus *et al.*, 1999), qui fournissaient des mesures précises pour quantifier la notion traditionnelle de classes de rythme. Autrement dit, étudier la notion de rythme dans les langues du monde était à la mode, et cela paraissait un choix logique car François Pellegrino travaillait sur la question et disposait d'un algorithme qui segmentait automatiquement le signal en consonnes et voyelles (ce qui était nécessaire pour calculer les indices pertinents), et d'autres travaux au Laboratoire Dynamique Du Langage portaient sur la question. J'ai donc consacré deux articles de conférences à appliquer les mesures s'appuyant sur la durée aux accents de l'anglais du corpus ABI, et participé à un autre sur les dialectes de l'arabe. Deux ans plus tard, nous avions écrit un article sur une question

indissociable du rythme : la perception du débit en français, anglais et allemand par des francophones (Dellwo *et al.*, 2006). La principale idée sous-jacente était qu'on perçoit le débit intentionnel des locuteurs quel que soit le débit effectif tel que mesuré, par exemple, en phonèmes par seconde. Après la thèse, c'est un article qui utilisait une mesure inédite d'intensité (Ferragne et Pellegrino, 2008) pour quantifier le rythme et montrait que cette mesure permettait de distinguer les accents du corpus ABI mieux que ne le faisaient les mesures établies dans la littérature, qui s'appuyaient sur la durée.

J'ai (un peu trop) rapidement rejeté mes travaux sur le rythme. La théorie qui sous-tendait les mesures de l'époque s'appliquait à des familles de langues, non à des accents d'une même langue. Ces mesures donnaient quelques résultats en arabe, pour des dialectes tellement distants phonétiquement qu'ils ne sont plus mutuellement intelligibles, mais il fallait s'attendre à ne pas avoir la précision nécessaire pour distinguer des accents de l'anglais sur les Îles Britanniques. C'est donc une absence de résultats tranchés, doublée de la frustration de me contenter d'appliquer des méthodes inventées par d'autres, qui m'a conduit à rapidement laisser de côté tous ces travaux. Le chapitre qui portait sur le rythme avait fait office d'intrus dans ma thèse aux dires des membres du jury, et j'avais donc décidé de tourner définitivement la page. D'ailleurs, pour bien illustrer ma lassitude, j'avais « offert » mes dernières données (non publiées) sur le rythme de l'anglais à Aniruddh Patel, professeur de psychologie aux États-Unis rencontré à l'occasion de la conférence de l'*Acoustical Society of America* à San Diego en 2004, et il les avait mentionnées dans son article qui compare le rythme de l'anglais et du français, avec le rythme chez les compositeurs anglais et français (Patel *et al.*, 2006).

Le rythme ne constitue cependant pas un pan négligeable de mes recherches car, pour prendre une mesure grossière de son succès, l'article présenté au colloque *Modélisation pour l'Identification Des Langues* (Ferragne et Pellegrino, 2004) est mon deuxième article le plus cité. Alors il est légitime de se demander ce qu'il reste de cet investissement. La réponse est assez optimiste : il m'arrive ponctuellement, au gré des encadrements d'étudiants ou de diverses collaborations, de retravailler sur le rythme et le débit. Il aura fallu quelques années de « carence » et un chapitre d'ouvrage écrit sur le moment à contre-cœur (Ferragne, 2013), pour comprendre que je pourrai toujours légitimement apporter, et techniquement, et sur le plan de la réflexion, ma modeste contribution à cette thématique. Cela a par exemple été le cas en 2016 dans un article de conférence qui comparait plusieurs mesures plus récentes du rythme dont celle de Tilsen et Arvaniti (2013), et une nouveauté utilisant des distances DTW entre les enveloppes d'amplitude de phrases en français et

en anglais (Michardière *et al.*, 2016).

Pour en revenir au débit, j'ai ponctuellement participé à un projet examinant la perception de la parole rapide. Dans Guiraud *et al.* (2013), nous montrions par exemple qu'il était plus facile pour des enfants de 8 ans de s'adapter à de la parole compressée artificiellement qu'à de la parole naturelle rapide présentant pourtant un débit équivalent. Là où la parole compressée artificiellement se contente de réduire la durée de tous les événements phonétiques dans les mêmes proportions, naturellement, nous réduisons certaines parties (par ex. état stable des voyelles) plus que d'autres (par ex. transitions). Ceci prouve entre autres que cette réduction non linéaire n'est pas une stratégie optimale orientée vers l'auditeur, mais bien le résultat de la contrainte supplémentaire des limites du système de production. L'expérience avait ensuite été étendue à des enfants dysphasiques (Guiraud *et al.*, 2014) et montrait leur difficulté à détecter des incohérences sémantiques à débit rapide (par rapport aux enfants sans pathologie), ce qui suggérait un lien étroit entre systèmes de production et de perception de la parole.

Mes résultats du milieu des années 2000 sur le rythme continuent d'être cités ici et là, et les mesures sur les différences potentielles entre accents de l'anglais, celles-là même que j'avais trouvées si frustrantes sur le moment, ont fait un heureux en la personne de Peter Trudgill, qui écrivait récemment (Trudgill, 2018, p. 136) :

> Happily, my observations have subsequently received strong instrumental phonetic confirmation from the work of Ferragne & Pellegrino (2004) [...] Through measurements of vowel duration, [they] show that East Anglian English is the most stress-timed of all the fourteen British Isles dialects which they investigated.

## 5.4 La perception des occlusives en écoute dichotique

Après ma thèse, Nathalie Bedoin m'a proposé une collaboration visant à déterminer si les deux hémisphères cérébraux prenaient en charge dans les mêmes proportions le traitement des traits de lieu d'articulation et de voisement des occlusives du français. Ces travaux ont donné lieu à deux publications, l'une, orientée recherche fondamentale, dans la revue *Brain and Language* (Bedoin *et al.*, 2010), et l'autre, présentant des applications cliniques, dans *Epilepsy & Behavior* (Bedoin *et al.*, 2011).

Dans Bedoin *et al.* (2010), partant du constat que le langage est traité de préférence par l'hémisphère gauche chez les droitiers, il s'agissait de comprendre s'il fallait privilégier 1) une hypothèse linguistique selon laquelle il y aurait un mode spécial de traitement du langage associé à l'hémisphère gauche, quel que soit le contenu acoustique, ou 2) une

hypothèse auditive selon laquelle certaines régions cérébrales sont spécialisées dans le traitement d'indices temporels ou spectraux, qu'il s'agisse de langage ou non. Et il y avait un aspect de cette seconde hypothèse que nous favorisions en particulier, inspiré du *Asymmetric Sampling in Time* de Poeppel (2003). Ce modèle propose, pour résumer et en m'appuyant sur l'analogie avec la numérisation de données audio, que l'hémisphère gauche présente un taux d'échantillonnage supérieur à celui de l'hémisphère droit. Ainsi, l'hémisphère gauche serait particulièrement adapté pour traiter des événements rapides (dont la parole regorge), alors que lorsque cette résolution temporelle n'est pas requise, il pourrait « déléguer » une partie du traitement à l'hémisphère droit.

Les occlusives constituent des stimuli tout à fait adaptés au test de cette hypothèse puisque le lieu d'articulation est signalé par des changements spectraux rapides (transitions formantiques) alors que le trait de voisement, signalé par le temps d'établissement du voisement (VOT), est plus long. Nous pressentions donc que le lieu d'articulation ne pouvait être traité que par l'hémisphère gauche chez les droitiers, alors que le voisement pouvait donner lieu à une prise en charge au moins partielle par l'hémisphère droit.

Des expériences d'écoute dichotique ont été mises au point. Deux mots différant en terme de lieu d'articulation (*passe-casse*) ou de voisement (*passe-basse*) étaient présentés simultanément à chacune des deux oreilles[52]. Les participants devaient simplement répéter à haute voix le mot qu'ils entendaient (le parfait alignement des deux stimuli sur la barre d'explosion devait donner lieu à la perception d'un seul et unique mot). En calculant la mesure, classique dans ce domaine, de l'avantage de l'oreille droite (*Right Ear Advantage*, REA), il a été possible de déterminer la dominance de l'hémisphère gauche dans les traitements phonétiques qui nous intéressent. Les résultats ont été conformes à nos attentes : un avantage de l'oreille droite/hémisphère gauche est observé dans le traitement du lieu d'articulation ainsi que du voisement, mais cet avantage est bien moindre dans le cas du voisement. Cette baisse de REA est en particulier due aux cas où l'oreille gauche entend un stimulus voisé en concurrence avec, dans l'oreille droite, un stimulus non voisé. Autrement dit, l'hémisphère gauche est capable de « déléguer » une partie du traitement phonologique à l'hémisphère droit lorsque le trait concerné correspond à un événement phonétique long, i.e. la barre de pré-voisement des occlusives sonores du français.

L'étude suivante (Bedoin *et al.*, 2011) utilisant la même technique incluait une dimen-

52. Pour une seconde expérience dans le même article, nous avons utilisé des pseudo-mots de type VCV afin d'éviter une dominance des stimuli voisés qui, commençant plus tôt que leurs homologues non-voisés du fait de la barre de voisement, orientent l'attention des auditeurs. Les effets, quoique moins forts, sont tout de même répliqués dans cette seconde expérience.

sion développementale et une composante pathologique. Elle montrait la mise en place graduelle entre l'âge de 5 et 8 ans de l'avantage de l'oreille droite pour le traitement du lieu d'articulation. L'avantage pour le voisement, quant à lui, restait stable sur cette même période (et jusqu'à l'âge adulte). Sur le plan pathologique, l'analyse d'un groupe d'enfants atteints d'épilepsie rolandique (bénigne) permettait d'établir un déficit de ce type de spécialisation hémisphérique.

Ces deux articles m'ont naturellement conduit à être impliqué dans un travail publié plus récemment (Bedoin *et al.*, 2019), mêlant écoute dichotique, contrôle de l'inhibition et EEG, et découlant logiquement de travaux initiés il y a quelque temps déjà (Bedoin *et al.*, 2013). Cette collaboration m'aura permis d'apprendre ou d'affiner certaines techniques expérimentales.

## 5.5 Collaborations éphémères et « mercenariat »

Après ma thèse, j'ai été brièvement impliqué dans deux projets portés par Fanny Meunier ; l'un concernait la perception de la parole dans le bruit (Boulenger *et al.*, 2010), et l'autre, des expériences en EEG visant à évaluer la sensibilité des auditeurs à des indices phonétiques fins comme ceux qui distinguent *la mie* de *l'amie* (Pota *et al.*, 2012). Si le premier n'a qu'un impact modéré sur la recherche que je pratique aujourd'hui, il en va tout autrement du second. Car en effet, c'est en quelque sorte ce projet qui m'a donné l'opportunité de me lancer dans l'électroencéphalographie et l'étude des potentiels évoqués.

Sous l'impulsion de François Pellegrino, nous avions mené une étude, qui est restée ponctuelle, sur la parole inversée temporellement (Pellegrino *et al.*, 2010). Ce type de transformation du signal, comme lorsque la bande d'un magnétophone est lue à l'envers, est typiquement utilisé dans les études en psycholinguistique expérimentale et en neurosciences comme condition de contrôle, parfois considérée comme de la non parole, parfois comme de la parole « délexicalisée ». La différence n'est pas triviale car, dans le premier cas, on peut envisager que cette transformation engendre un simple traitement auditif de bas niveau, alors que dans le second, ce sont des processus langagiers de plus haut niveau (traitement des phonèmes) qui pourraient être mobilisés. Curieusement, peu d'études s'étaient intéressées à la manière dont ce type de parole était traité. Il faut bien avoir à l'esprit que la parole inversée temporellement présente des caractéristiques bien spécifiques : l'enveloppe d'amplitude est totalement bouleversée puisque, là où la parole

naturelle présente des attaques rapides et des atténuations plus lentes, c'est logiquement l'inverse qui se produit en parole inversée, donnant cette impression de parole « inspirée ». Nous avions fait transcrire des extraits de parole inversée par des phonéticiens. Ces derniers s'étaient montrés capables de transcrire au niveau phonétique une bonne proportion du signal, avec une cohérence inter-juges relativement élevée. Sans trop de surprise, ce sont les sons reposant sur des caractéristiques spectrales instables, par exemple les occlusives, qui ont donné lieu à des divergences entre experts. En effet, les occlusives en parole inversée ont le plus souvent été transcrites comme des fricatives. Il arrive que les transcriptions fassent apparaître des segments épenthétiques ; par exemple, le pseudo-mot [sat], lorsqu'il est retourné, est perçu comme [snas], le [n] perçu étant très vraisemblablement engendré par l'attaque lente du [a] inversé, qui n'existe pas en parole naturelle. Le fait qu'il soit possible d'identifier des « phonèmes » suggère qu'un traitement phonétique menant potentiellement à un accès au lexique se produit dans la perception de la parole inversée. Ce type de parole n'est donc pas aussi « neutre » qu'on pourrait le penser, ce qui doit être pris en compte dans les études qui l'utilisent comme condition de contrôle.

Les travaux auxquels mon nom est associé portant sur le traitement de la prosodie dans la parole affective par des autistes, puis des patients atteints du syndrome de Landau-Kleffner, sont typiques des interventions ponctuelles mobilisant des compétences techniques très précises que j'ai fournies de très nombreuses fois, mais qui atteignent bien rarement ce niveau de valorisation. Deok-Hee Kim-Dufor, alors doctorante de Jean-Luc Nespoulous, m'avait contacté pour voir si j'étais en mesure de manipuler la prosodie de stimuli de parole dont elle disposait de sorte qu'ils reflètent plusieurs émotions prédéfinies. Les programmes écrits pour l'occasion avaient donné satisfaction, et j'avais ainsi co-signé quelques publications (Kim-Dufor *et al.*, 2010, 2012) sans pour autant m'être penché dans le détail sur l'objet étudié.

Les opportunités d'encadrement d'étudiants m'ont également conduit à aborder ponctuellement des thématiques variées. Le mémoire d'orthophonie de Maëlle Le Cerf a été l'occasion d'évaluer à partir de mesures phonétiques acoustiques précises les bénéfices de la méthode LSVT LOUD pour la prise en charge de la dysarthrie hypokinétique, qui est typique de la maladie de Parkinson. Ce travail a donné lieu à une publication dans les actes des *Journées d'Études sur la Parole* (Le Cerf et Ferragne, 2020). Deux ans auparavant, ce sont les travaux du mémoire d'orthophonie d'Anaïs Delhoume qui étaient valorisés dans la même conférence (Delhoume et Ferragne, 2018). Ces collaborations très ponctuelles, dont la liste est longue, m'ont permis de mettre à profit mes compétences en phonétique

et en analyse de données ainsi que d'explorer des domaines très variés.

## 5.6 Conclusion

La liste présentée dans ce chapitre n'est pas exhaustive. Par ailleurs, en plus des travaux énumérés ici, il existe une liste tout aussi longue de projets de recherche qui n'ont pas abouti à des publications mais qui ont néanmoins contribué à ma formation et dont les bénéfices sont, d'une manière ou d'une autre, avantageusement exploités dans ma recherche actuelle. Cette diversification, tant sur le plan des questions scientifiques que sur les méthodes employées pour y répondre, est, je crois, très caractéristique de ma démarche, qui alternativement, teinte la phonétique de psychologie et neurosciences, ou la colore de sciences de l'ingénieur.

# 6

# Deep learning pour la phonétique

## Sommaire



## 6.1 Introduction

Les réseaux de neurones artificiels profonds (DNN pour *deep neural networks*) constituent un ajout relativement récent à mon éventail scientifique[53]. Mais je dois préciser d'emblée qu'il ne s'agit pas d'une simple fantaisie pour être en phase avec la mode actuelle. Au contraire, j'entends démontrer dans ce chapitre que l'adoption de cette famille de méthodes constitue un aboutissement parfaitement cohérent dans l'évolution de mon parcours. Et je dois insister sur le fait qu'il ne s'agit pas (ou très partiellement) d'un

53. Ce chapitre doit beaucoup à mes collègues et amis Cédric Gendrot et Thomas Pellegrini.

aboutissement sur le plan technique — les outils contemporains rendent l'implémentation des DNN presque aussi triviale que d'autres techniques que j'utilise couramment depuis longtemps — mais bien de la synthèse d'une réflexion scientifique qui m'anime depuis le début de ma carrière. Après avoir établi le contexte, je présenterai quelques exemples de questions phonétiques pour lesquelles j'ai eu recours à ces techniques.

Mon goût pour l'intelligence artificielle (IA) et les réseaux de neurones date d'avant ma thèse. Autrement dit, c'est malheureusement lors du dernier « hiver » (puisque c'est le terme consacré) de l'IA que j'ai commencé à m'intéresser à cet ensemble de méthodes. Cet intérêt s'est bien sûr exprimé à travers la lecture de certains ouvrages de vulgarisation scientifique, mais lorsque les choses sont devenues plus sérieuses, c'est vers Negnevitsky (2002) que je me suis tourné pour un panorama des méthodes. Et c'est grâce à Nabney (2002) que j'avais implémenté mes premiers modèles de *machine learning* ; sans oublier le classique Duda *et al.* (2001). À cette époque, j'avais expérimenté avec quelques architectures de réseaux de neurones simples et, comme cela se faisait encore souvent, avec des cartes de Kohonen.

Dans les dernières pages de ma thèse, j'avais esquissé un système expert à base de logique floue qui aurait pu, en quelques règles, permettre à un utilisateur humain connaissant la phonétique (mais pas les accents de l'anglais) de déterminer l'accent d'un extrait de mon corpus. Mais là encore, si l'exercice demeure néanmoins intéressant, ce type de système est à ma connaissance complètement tombé en désuétude. Dans le même ordre d'idées, j'avais milité pour l'utilisation des arbres de classification, par exemple dans Ferragne et Pellegrino (2010c), car ils émulent en quelque sorte un processus de décision par un humain. J'ai donc depuis longtemps le désir de faire accomplir à la machine des tâches qui, quand elles sont réalisées par un être humain, nécessitent de l'intelligence.

Incidemment, j'ai dans mes archives une photographie mise en scène pour le site internet du Laboratoire Dynamique Du Langage à l'époque de ma thèse : j'y apparais de dos, un casque sur les oreilles, avec deux ordinateurs censés représenter mon activité quotidienne. L'un affiche un spectrogramme, l'autre, un graphique en 3 dimensions généré à partir de données aléatoires (c'est dire si la scène était authentique !). Pour l'occasion, j'avais méticuleusement choisi trois ouvrages emblématiques que j'avais disposés dans un désordre élaboré sur le bureau. Parmi les nombreuses références que j'aurais pu choisir, ces trois livres illustraient la vision que j'ai du métier. Il y avait *The Structure of Scientific Revolutions* (Kuhn, 1996), que j'ai eu beaucoup de plaisir à lire plusieurs fois, *The Mathematical Theory of Communication* (Shannon et Weaver, 1975) et *Cybernetics or*

*Control and Communication in the Animal and the Machine* (Wiener, 2007)[54] . Ce dernier ouvrage, qui préfigure ce qui deviendra l'intelligence artificielle, témoigne donc de mon engouement pour l'IA bien avant la mode récente du *deep learning*.

Mais qu'est-ce qui justifie mon implication actuelle dans le *deep learning* au-delà du fait qu'il regroupe des techniques dont les performances dans de nombreux domaines appliqués ont occasionné de véritables révolutions ? En premier lieu, j'ai eu l'occasion tout au long de ce manuscrit de rappeler combien la visualisation représentait dans nos domaines, en particulier pour les collègues des SHS, un besoin crucial pour une recherche sereine. En effet, le formalisme mathématique n'est pas accessible à tout le monde alors que communiquer visuellement sa recherche est une démarche égalitaire : quiconque est en mesure de visualiser un graphique peut émettre un avis sur ce qu'il voit ; il n'en va pas de même pour une équation ou une valeur de probabilité. Les DNN se prêtent particulièrement bien à la visualisation des données, et s'accordent donc parfaitement avec ce credo.

Les techniques actuelles permettent la représentation graphique des diverses tâches accomplies par les DNN et, en particulier, la mise en lumière des paramètres sur lesquels un DNN s'est appuyé pour effectuer une tâche (Zeiler et Fergus, 2014 ; Selvaraju *et al.*, 2017 ; Zhou *et al.*, 2016 ; Chattopadhay *et al.*, 2018 ; Ferragne *et al.*, 2019 ; King et Ferragne, 2019). Il est important d'insister sur ce point car lors de mes premiers essais avec des réseaux de neurones au début de ma thèse en 2003, ils constituaient l'archétype de la boîte noire. De nos jours, la communauté spécialiste du *deep learning* déploie d'intenses efforts dans le domaine de la visualisation et de l' « explicabilité », à travers de nombreux articles sur la question et des conférences entièrement dédiées à ce thème. Je voudrais insister sur le fait que je considère aujourd'hui que les réseaux de neurones profonds n'ont plus de raison d'être considérés comme des boîtes noires.

Pour résumer ces deux premiers points, on retiendra donc que le monde des DNN offre non seulement de nouvelles façons de représenter graphiquement les données mais également des visualisations des différentes étapes (convolution, rectification, etc.) de leur « mécanique » interne ; ce qui a une résonance très forte avec mon engagement dans la communication de la recherche par le graphique. Certes, comme le note très justement Tufte (2001), on peut mentir avec un graphique et, dans ce cas, le fait qu'un graphique soit particulièrement accessible aux non spécialistes fait que le mensonge peut avoir un impact extrêmement fort sur la population. Néanmoins, je remarque qu'il est probablement plus

54. Les dates que je donne en référence pour ces 3 ouvrages sont celles des éditions récentes consultées ; les premières éditions sont de 1962 pour le premier et 1948 pour les deux autres.

facile de détecter un mensonge graphique qu'un mensonge émanant d'une technologie complexe ou encore d'une formule mathématique alambiquée.

L'avantage capital des DNN réside dans le fait qu'ils accomplissent en réalité deux tâches là où la plupart des algorithmes de classification se contentent de classer. En effet, simultanément à la classification, les DNN sont capables d'extraire des descripteurs pertinents, ce qui constitue leur atout majeur. Là où, classiquement, les chercheurs construisaient d'abord des représentations (descripteurs) très « contraintes » de leurs données (formants vocaliques, mesures de durées, etc.), un DNN peut apprendre à partir des représentations les plus brutes. Il est alors légitime de s'interroger sur l'avantage du degré de naïveté important des DNN par rapport aux modèles plus classiques qui reçoivent en entrée des descripteurs plus motivés. J'ai découvert avec la pratique — alors que j'ai pourtant souvent insisté sur la nécessité absolue d'émettre des hypothèses fortes a priori — que se montrer agnostique quant à la nature précise de ces descripteurs en entrée d'un modèle peut ponctuellement présenter quelques avantages.

Que les DNN soient capables d'extraire eux-mêmes les descripteurs (ou caractéristiques) pertinents semble être, de l'aveu même de l'un de leurs inventeurs, la spécificité majeure qui les distingue de toutes les autres techniques depuis les débuts du *machine learning*. Je souscris totalement à ce point de vue. Le Cun (2019, p. 119) écrit en effet :

> Entre les années 1960 et 2015, les chercheurs dépensent une énergie folle à concevoir des extracteurs de caractéristiques pour tel ou tel problème [...] Une de mes idées fixes a été de trouver des méthodes permettant d'entraîner les extracteurs de caractéristiques au lieu de les construire à la main. Mais la communauté, longtemps, n'y a pas cru. Tel est l'enjeu des réseaux de neurones multicouches et du *deep learning*.

À l'occasion d'une réunion de travail, Jean-François Bonastre me faisait remarquer combien mon discours à ce sujet était en parfait décalage avec la position des linguistes dans les années 1990, qui reprochaient justement aux informaticiens d'injecter dans leurs modèles des myriades de paramètres acoustiques dont la motivation phonétique n'était pas évidente. La différence notoire entre ma pratique actuelle et le recours à des paramètres « d'ingénieurs » réside dans le fait que ce que je fournis en entrée de mes modèles reste totalement interprétable par des phonéticiens puisqu'il s'agit de spectrogrammes à bandes larges. Laisser un réseau convolutif se charger de la détection des zones d'intérêts ne revient finalement qu'à remplacer ponctuellement l'œil des phonéticiens par une machine dans l'espoir, d'une part, d'aller plus vite, et d'autre part, d'éventuellement déceler des caractéristiques qui échappent à l'œil humain.

Le réalisme biologique des modèles de *deep learning* n'est pas une préoccupation essen-

tielle des architectures et algorithmes d'apprentissage actuels, contrairement à l'époque (années 1940 à 1960) où ce domaine s'appelait encore « cybernétique », et était très explicitement inspiré par le cerveau humain (Goodfellow *et al.*, 2016). C'est pour cette raison que, d'une part, n'étant pas spécialiste, je ne proposerai pas de discussion de la plausibilité biologique des calculs faits par les DNN et, d'autre part, que je préfère parler de plausibilité cognitive. D'ailleurs, Chollet (2018, p. 8) encourage à :

> forget anything you may have read about hypothetical links between deep learning and biology. For our purposes, deep learning is a mathematical framework for learning representations from data.

Les DNN offrent donc de mon point de vue un scénario plausible d'un processus d'apprentissage par l'humain, ou, au moins, une analogie féconde, qui concerne non seulement le processus d'apprentissage mais également la nature des représentations obtenues. En effet, si on admet que les représentations phonologiques sont riches, comme le proposent notamment les modèles à exemplaires (Pierrehumbert, 2001, 2016), la quête historique d'une représentation guidée uniquement par la parcimonie ne tient plus. Intuitivement au moins, les activations dans les couches successives d'un réseau à convolution, représentées par la méthode des images dites *deep dream* ou encore par les méthodes de type CAM et Grad-CAM (etc.), permettent la visualisation de ce qui pourrait très bien correspondre aux abstractions des catégories apprises par un humain. Si les applications de telles méthodes à la phonétique en sont encore à un stade embryonnaire, les spécialistes de traitement de l'image et de vision par ordinateur développent depuis quelques années des représentations très convaincantes que je m'efforce d'intégrer dans ce nouveau paradigme que je propose.

La thèse de Maud Pélissier et son excellente synthèse sur les deux types d'apprentissage, implicite et explicite, dans l'acquisition de la syntaxe de l'anglais par des apprenants francophones, me fournit une analogie supplémentaire. En effet, il est intéressant d'établir un parallèle entre, d'une part, la nécessité de fournir des paramètres explicites aux techniques de *machine learning* traditionnel et la manière dont l'apprenant tardif d'une langue étrangère apprend des règles explicites et verbalisables ; et d'autre part, la capacité des modèles de *deep learning* à extraire eux-mêmes les paramètres pertinents pour leur apprentissage, comme on peut le faire lorsqu'on apprend sa langue maternelle. Pour résumer cette analogie, le *machine learning* traditionnel serait l'apprenant tardif d'une langue étrangère et le *deep learning*, un apprenant natif.

Enfin, une dernière motivation, qui recouvre une bonne partie de celles que je viens d'énoncer, concerne le caractère ludique du *deep learning*. Le Chapitre 3 témoignait de

mon attirance pour la personnalisation des méthodes, l'expérimentation et la création de nouveaux outils : les techniques de *deep learning* se prêtent particulièrement bien à cet exercice. En effet, avec leurs capacités de visualisation, la liberté qu'ils procurent de créer et d'ajuster sa propre architecture, leurs compétences dans l'extraction de descripteurs, les réseaux profonds constituent un terrain de jeu idéal pour une recherche créative.

## 6.2 De la vision par ordinateur à la phonétique

Le domaine de la reconnaissance automatique d'images a connu un tournant décisif en 2012. En effet, cette année-là, le taux d'erreur des gagnants du *Imagenet Large Scale Visual Recognition Challenge,* compétition dont le but est de développer un algorithme capable de reconnaître automatiquement un millier d'objets à partir de millions d'images, passe d'environ 26 % à 16 %. Au-delà de la performance pure, les évolutions technologique et conceptuelle sont particulièrement marquantes. C'est la première fois qu'un réseau profond à convolution (CNN) est employé dans cette compétition de référence, et la première fois qu'on entraîne un CNN en utilisant les énormes capacités de parallélisation offertes par un processeur graphique, un GPU (*Graphics Processing Unit*). Ce CNN, qui sera baptisé Alexnet (Krizhevsky *et al.*, 2012) marque donc le début de l'utilisation du *deep learning* dans le domaine de la vision par ordinateur, et symbolise bien les révolutions survenues à cette époque non seulement dans des domaines liés à l'image, mais également dans les technologies de la parole.

La phonétique moderne s'est construite sur le recours permanent à la représentation visuelle de phénomènes sonores. Cette synesthésie emblématique de notre discipline, ce *Visible Speech* omniprésent, s'impose à quiconque s'intéresse à la phonétique. Des premiers alphabets phonétiques au spectrographe, en passant par les innombrables inventions dont regorge Rousselot (1897), c'est bien la modalité visuelle qui est mise en avant. Et si on ajoute à cela toutes les analyses s'appuyant sur l'imagerie articulatoire ou cérébrale, il devient évident que comprendre les sons du langage revient à les visualiser.

D'un côté, donc, la phonétique fait la part belle au visuel, et de l'autre, des méthodes très puissantes sont disponibles depuis le milieu des années 2010 dans le domaine de la vision par ordinateur. C'est là le point de départ de mon aventure avec le *deep learning* : je souhaite entraîner les modèles à devenir des phonéticiens. Qu'ils sachent comparer des sons, et qu'ils disent sur quelles régions d'une image d'échographie de la langue, d'une vidéo des lèvres, ou de spectrogramme, ils se fondent pour prendre leurs décisions.

## 6.3 Phonétique pour la comparaison de voix

### 6.3.1 Essais préliminaires

Je présente ici les premiers essais que j'avais réalisés pour la mise en place de réseaux de neurones visant à modéliser les caractéristiques phonétiques inter-individuelles. Je décris rapidement le réseau à convolution, appelé CNN01 pour l'occasion, que j'avais utilisé à l'époque, avant de présenter la méthode de l'occlusion. Ce retour sur mes premières tentatives permet de bien appréhender le cheminement qui a façonné ma réflexion dans le domaine ainsi que le chemin parcouru depuis.

Ces essais préliminaires ont été effectués sur un échantillon de 6 locuteurs du corpus ESTER (Galliano *et al.*, 2005), un corpus radiophonique en langue française. Toutes les occurrences de la voyelle /ɑ̃/ ont été converties en spectrogrammes sur l'échelle des Mel. Les images en couleur codées sur 24 bits (3 canaux de 8 bits chacun) ont été redimensionnées en 40 × 40 pixels. Le nombre de voyelles était variable d'un locuteur à l'autre, allant de 171 à 531, pour une moyenne de 327. L'architecture de CNN01 comprend 15 couches. La couche d'entrée, dont la taille correspond aux images, opère un centrage des données (soustraction de la moyenne). Puis vient la première couche convolutive composée de filtres de taille 5 × 5 avec un pas (*stride*) de 1 et une marge (*padding*) de 2. La couche suivante est une fonction d'activation de type ReLU (*Rectified Linear Unit*) ; elle est suivie d'une couche de sous-échantillonnage utilisant le maximum (*max pooling*). Cette succession de convolution, rectification et sous-échantillonnage se répète encore 2 fois puis vient la première couche entièrement connectée (ou dense), suivie d'une autre couche de type ReLU, suivie d'une autre couche dense. Les deux couches finales sont chargées de la classification à proprement parler : l'avant-dernière applique une fonction *softmax* aux paramètres de sortie de la couche précédente et la couche finale opère la classification à proprement parler. L'architecture du réseau est récapitulée dans le Tableau 6.1.

Une méthode intuitive qui permet de localiser l'information qu'utilise un modèle pour accomplir sa tâche consiste à priver le modèle d'une partie de l'information pour examiner ensuite la dégradation des performances qui résulte de cette privation. Nous avions, avec Cédric Gendrot et Thomas Pellegrini, convenu que ce principe présentait un intérêt, et c'est Thomas qui m'a explicitement suggéré la méthode de sensibilité à l'occlusion (Zeiler et Fergus, 2014) ; il ne me restait plus qu'à écrire le code[55].

55. S'il existe une fonction officielle dans `Matlab` aujourd'hui, `occlusionSensitivity`, ce n'était pas le cas à cette époque-là.

TABLEAU 6.1 – Architecture du réseau de neurones profond CNN01.

| **Name** | **Type** | **Activations** | **Learnables** |
|---|---|---|---|
| input<br>$40 \times 40 \times 3$ images with zerocenter normalization | Image Input | $40 \times 40 \times 3$ | - |
| conv1<br>40 $5 \times 5 \times 3$ convolutions with stride [1 1] and padding [2 2 2 2] | Convolution | $40 \times 40 \times 40$ | Weights $5 \times 5 \times 3 \times 40$<br>Bias $1 \times 1 \times 40$ |
| relu1<br>ReLU | ReLU | $40 \times 40 \times 40$ | - |
| maxpool1<br>$3 \times 3$ max pooling with stride [2 2] and padding [0 0 0 0] | Max Pooling | $19 \times 19 \times 40$ | - |
| conv2<br>40 $5 \times 5 \times 40$ convolutions with stride [1 1] and padding [2 2 2 2] | Convolution | $19 \times 19 \times 40$ | Weights $5 \times 5 \times 40 \times 40$<br>Bias $1 \times 1 \times 40$ |
| relu2<br>ReLU | ReLU | $19 \times 40 \times 40$ | - |
| maxpool2<br>$3 \times 3$ max pooling with stride [2 2] and padding [0 0 0 0] | Max Pooling | $9 \times 9 \times 40$ | - |
| conv3<br>80 $5 \times 5 \times 40$ convolutions with stride [1 1] and padding [2 2 2 2] | Convolution | $9 \times 9 \times 80$ | Weights $5 \times 5 \times 40 \times 80$<br>Bias $1 \times 1 \times 80$ |
| relu3<br>ReLU | ReLU | $9 \times 9 \times 80$ | - |
| maxpool3<br>$3 \times 3$ max pooling with stride [2 2] and padding [0 0 0 0] | Max Pooling | $4 \times 4 \times 80$ | - |
| fc1<br>80 fully connected layer | Fully Connected | $1 \times 1 \times 80$ | Weights $80 \times 1280$<br>Bias $80 \times 1$ |
| relu4<br>ReLU | ReLU | $1 \times 1 \times 80$ | - |
| fc2<br>6 fully connected layer | Fully Connected | $1 \times 1 \times 6$ | Weights $6 \times 80$<br>Bias $6 \times 1$ |
| softmax<br>softmax | Softmax | $1 \times 1 \times 6$ | - |
| classif<br>crossentropyex | Classification Output | - | - |

La phase d'apprentissage du réseau s'effectue sur un ensemble aléatoire constitué de 70 % des voyelles. Le réseau à convolution employé ici atteint une précision de plus de 90 % sur l'ensemble de test. Incidemment, il s'agissait ici, chronologiquement, de l'un des deux premiers modèles de *deep learning* entraînés par mes soins et je dois dire que ce score impressionnant — au regard de la petite taille des données et surtout de ce que j'attendais d'une simple voyelle pour caractériser le locuteur — a suscité un enthousiasme très productif !

Vient ensuite l'étape consacrée plus spécifiquement à l'analyse de l'occlusion. Pour chaque image correctement classifiée de l'ensemble test, une série d'images est générée, chacune comportant un masque de $9 \times 9$ pixels couvrant partiellement l'image originale. Dans chaque image de cette série, le masque apparaît dans des régions différentes, de façon à couvrir l'intégralité de l'espace, comme l'illustre la Figure 6.1. Cette illustration montre 16 images choisies aléatoirement, générées à partir d'un seule voyelle. Ainsi chacune des 533 voyelles correctement classifiées de l'ensemble de test conduit à générer 1 600 images (pour toutes les positions possibles du masque) ; ce sont donc 852 800 images qui sont soumises au modèle pour qu'il estime la dégradation de la probabilité que l'image en cours de traitement soit correctement classifiée.

On reconstruit enfin une image moyenne pour chaque locuteur, reflétant l'impact du masquage. Les représentations pour les six locuteurs apparaissent successivement dans la Figure 6.2. Il est important de noter que, dans chacune des images, l'intégralité de l'échelle de couleurs (allant du bleu foncé au jaune clair) est utilisée. Autrement dit, ces couleurs matérialisent des valeurs qui ne sont pas comparables d'une figure à l'autre puisque la variation de la dégradation est normalisée pour chaque locuteur. Il s'agit d'un aspect qui, en fonction de ce que l'on souhaite montrer, pourra être modifié.

Bien que l'objectif de cette section soit d'illustrer un point méthodologique, un bref commentaire des résultats ne paraît pas superflu. Il serait hasardeux d'émettre un avis tranché quant à ce qui est représenté dans ces figures mais en première approximation, on constate que la localisation de la dégradation induite par l'occlusion n'est pas constante d'un locuteur à l'autre. Plus précisément, il paraît d'une part impossible de localiser une zone fréquentielle qui soit pertinente pour tous les locuteurs et, d'autre part, il semblerait que pour un même locuteur, on ne puisse pas définir des zones critiques sur l'intégralité de la voyelle. À ce stade très spéculatif, on peut donc envisager que les zones de fréquences permettant de caractériser un locuteur ne sont pas les mêmes pour tous les locuteurs et que ces zones varient à l'intérieur d'une seule et même voyelle.

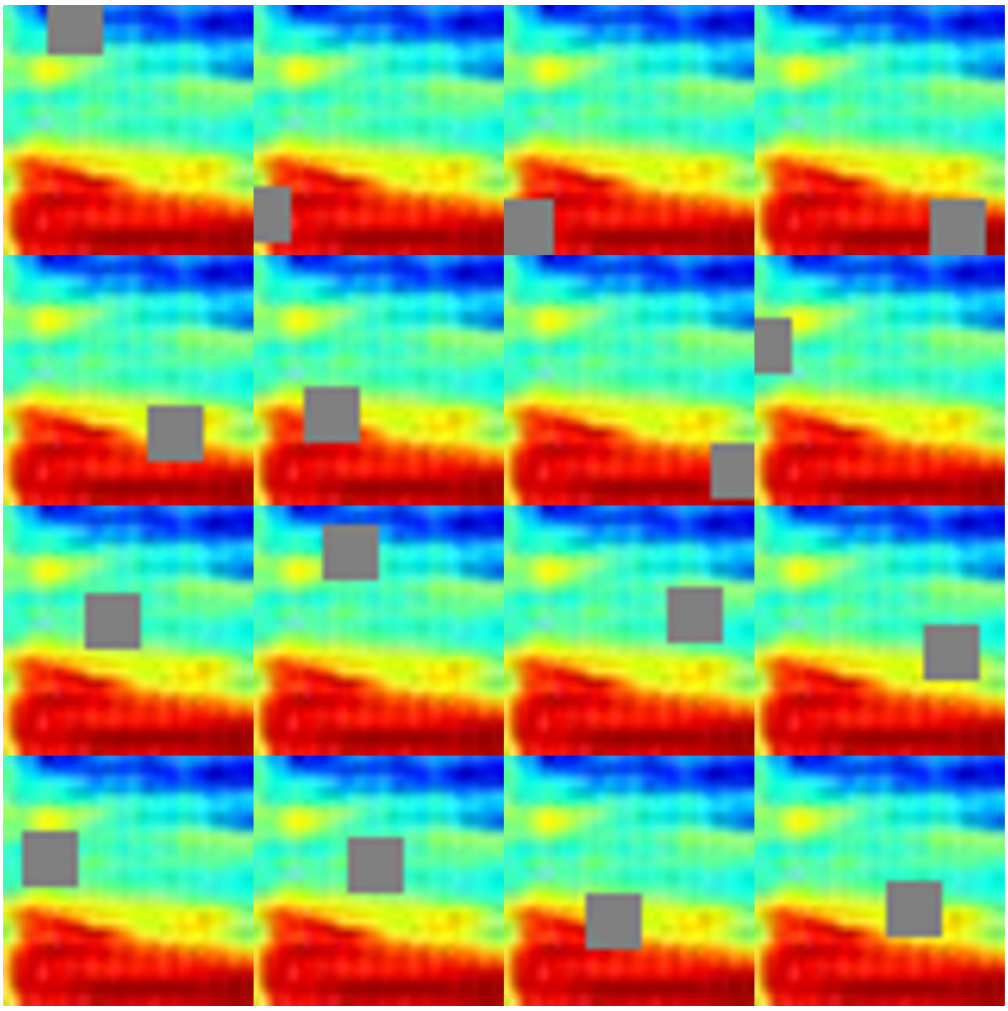
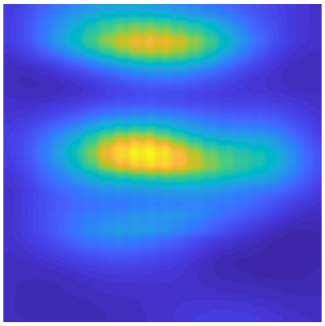


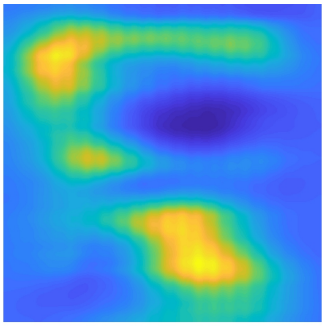
FIGURE 6.1 – Spectrogrammes avec masques illustrant la méthode de l'occlusion.

J'ai pris la peine de résumer ces premiers essais car ils m'ont permis de poser les fondements de ce que j'essaie depuis d'ériger en un programme de recherche nouveau. En premier lieu, réaliser que ce type d'algorithme donne de bons résultats avec des ensembles de taille relativement modeste (comme la phonétique en produit) m'a fortement encouragé à persévérer. Ensuite, j'ai rapidement décidé d'abandonner les spectrogrammes représentant des fréquences sur une échelle psycho-acoustique. En effet, l'alternative pour moi consistait soit à adopter complètement les paramètres opaques utilisés dans les systèmes « état de l'art » (e.g. i-vectors) et me retrouver directement en concurrence avec les collègues dont le traitement automatique de la parole est le métier, soit totalement adapter le *deep learning* aux représentations qu'utilisent historiquement les phonéticiens. Je ne veux pas une intelligence artificielle qui affiche des performances de haut niveau en reconnaissance de la parole ou autres tâches appliquées (les GAFA s'en chargent) ;

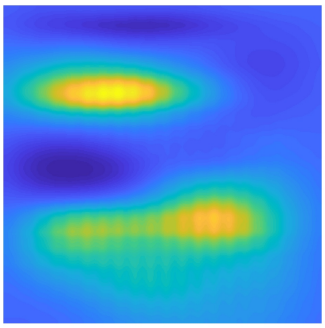
je cherche une méthode qui émule ce que font les phonéticiens. J'ai donc opté pour le spectrogramme classique plutôt que d'autres représentations apparentées : après tout, qui aurait l'idée de donner autre chose qu'une radiographie à une radiologue ?



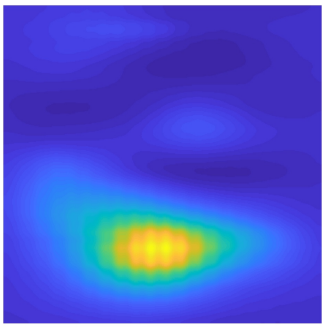
(a) Patrick Boyer



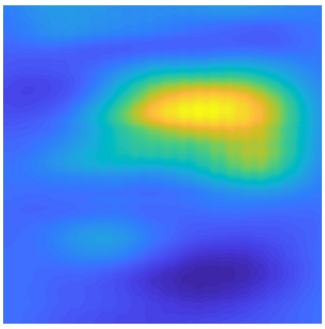
(b) Gérard Courchelle



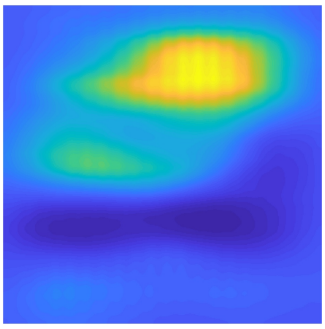
(c) Pascal Le Guern



(d) Alain Passerel



(e) Patrick Roger



(f) Laurent Sadoux

FIGURE 6.2 – Résultats de l'analyse de sensibilité à l'occlusion pour les 6 locuteurs de petit corpus radiophonique. Les zones où le masque présente un impact négatif important apparaissent en clair.

### 6.3.2 Variabilité inter-locuteur et comparaison de voix

Comme le rappellent Morrison et Thompson (2017), il existe historiquement quatre approches de la comparaison de voix dans le domaine judiciaire : auditive, spectrographique, phonétique acoustique et automatique. C'est l'approche automatique qui prédomine aujourd'hui ; mais cela n'empêche pas le recours ponctuel à certains paramètres phonétiques acoustiques, ou encore à des commentaires issus d'impressions auditives. Mon approche allie les aspects spectrographique et automatique. Elle ne se focalise pas sur la performance de la méthode mais plutôt sur son interprétabilité en termes phonétiques. C'est cela qui motive, au moins pour les premiers travaux que j'ai menés dans ce domaine, l'adoption du spectrogramme à bandes larges classique. En effet, le lien entre l'articulation des sons de la parole et un spectrogramme est plus direct qu'avec des paramètres typiquement utilisés en traitement automatique.

Pour Ferragne *et al.* (2019), nous avons travaillé à partir d'ESTER. Nous avons extrait la voyelle /ɑ̃/ pour 45 locuteurs (dont 10 locutrices), et avons entrainé un réseau de neurones à convolution à classer automatiquement les occurrences de /ɑ̃/ en fonction du locuteur.

Le choix de la voyelle /ɑ̃/ répond à deux critères. D'abord, nous avions espéré, comme le laissent entendre certaines sources, que le couplage avec la cavité nasale contiendrait davantage d'informations spécifiques au locuteur. Ensuite, nous avions de toutes façons testé notre modèle sur toutes les voyelles du français, et la voyelle /ɑ̃/ présentait les performances de classification en locuteurs les plus élevées.

Lorsque je parle publiquement de *deep learning* en conférence ou en cours, il m'arrive souvent de mettre en avant deux aspects : la capacité des modèles à extraire eux-mêmes des paramètres à partir des représentations les plus brutes et le fait qu'ils nous font basculer dans une ère scientifique où le principe de parcimonie ne tient plus. Je maintiens ces affirmations, mais la pratique me pousse à apporter quelques nuances. En effet, d'une part, si un spectrogramme est une représentation relativement brute (par comparaison avec, par ex., une extraction de formants), il constitue déjà une construction par rapport au signal initial, et cette construction a donné lieu à certains choix. D'autre part, l'abondance a ses limites : la mémoire des GPU n'étant pas extensible (8 Go pour celui que j'utilise le plus souvent), il faut réapprendre à économiser.

Dans Ferragne *et al.* (2019), les spectrogrammes des voyelles dont la durée était au moins égale à 30 ms et au plus à 250 ms ont été générés avec un script en `Python`. Le signal original échantillonné à 16 kHz a été découpé en trames de 5,0625 ms et l'analyse

avait un pas de 0,5 ms ; il y avait donc un chevauchement de 90 % entre trames successives. Les segments de parole étaient ensuite multipliés par une fenêtre de Hamming, complétés par des zéros de sorte à comporter 512 points, et soumis à une FFT[56].

Pour cet article précisément, il a été décidé de ne pas utiliser de pré-emphase alors que ceci est courant (6 dB par octave) dans les représentations traditionnelles. Ce choix a été motivé par le fait que nous souhaitions effectuer par la suite de la resynthèse après application de masques dans le spectrogramme, ce qui aurait été rendu compliqué par le filtrage de pré-emphase. Incidemment, ce n'est qu'à contre-cœur que j'ai cédé sur ce point car mon objectif initial dans ce volet de ma recherche était de présenter à la machine exactement ce qu'ont vu les phonéticiens pendant des décennies (donc avec pré-emphase).

Les spectrogrammes finaux ont une dynamique de 70 dB. Afin d'être en phase avec les conventions de la machine, pour qui le noir représente une valeur numérique minimale, et le blanc, une valeur maximale, l'échelle des niveaux de gris a été inversée. Autrement dit, les formants apparaissent en blanc ; une telle transformation n'a théoriquement aucune incidence pour nos modèles...je suppose qu'il s'agit là d'une coquetterie de notre part ! Afin de préserver l'intégrité temporelle des voyelles, les spectrogrammes dont la durée était inférieure à 250 ms ont fait l'objet de zero-padding (voir Figure 6.3).

Les limites imposées par la mémoire de la machine conduisent à soigneusement réfléchir à la taille des images de spectrogrammes. Dans Ferragne *et al.* (2019), nous nous sommes cantonnés à des images de 224 × 224 pixels, ce qui correspond à la taille prédéfinie des entrées du réseau « historique » utilisé (VGG16). Les niveaux de gris des spectrogrammes ont été recodés sur 8 bits.

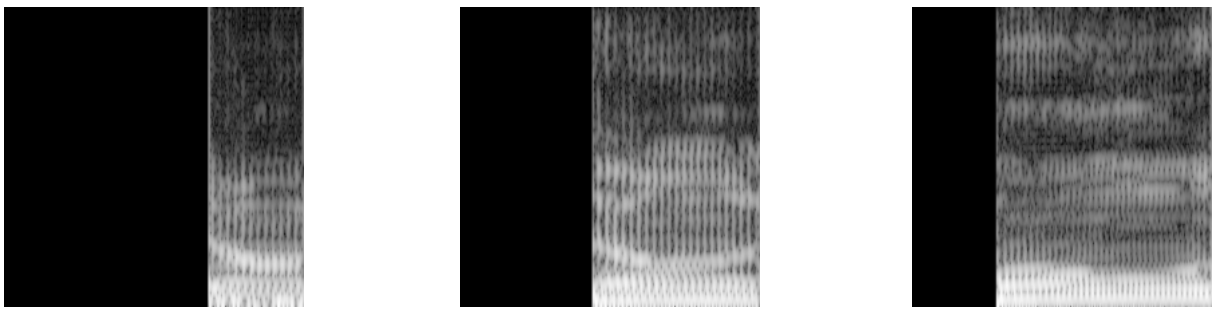


FIGURE 6.3 – Exemples de spectrogrammes.

C'est donc une architecture classique, VGG16 (Simonyan et Zisserman, 2015), qui a été utilisée ici. Sur ce plan-là, mon raisonnement est simple : je traite des images et il est donc naturel que j'utilise des architectures qui se sont illustrées dans des tâches de

56. Fast Fourier Transform : transformation de Fourier rapide.

reconnaissances d'images, comme c'est le cas de VGG16 en 2014 dans le *ImageNet Large Scale Visual Recognition Challenge*. Un autre argument en faveur de ce réseau tient au fait qu'il est largement disponible, non seulement dans la Deep Learning Toolbox de `Matlab`, mais également dans de nombreux exemples de codes sur Internet, en particulier dans le « zoo » de modèles au format interopérable ONNX[57].

Les données, composées donc de 334 spectrogrammes de la voyelle /ɑ̃/ pour chacun des 45 locuteurs, ont été partitionnées de façon aléatoire en ensembles d'apprentissage (70 %), de validation (10 %) et de test (20 %). Le modèle affiche un pourcentage de classification correcte de 85,37 %.

Il serait rébarbatif de rapporter tout ce qui a déjà été écrit dans Ferragne *et al.* (2019) ; l'intérêt essentiel de cet article réside dans la méthode de visualisation que nous avons mise au point. Chacune des 67 voyelles d'un individu de l'ensemble de test a été soumise à la méthode de l'occlusion. Contrairement à la méthode décrite plus haut (Section 6.3.1), le masque recouvrait ici toute la durée de la voyelle. Il avait une hauteur de 15 pixels (environ 536 Hz) et se déplaçait verticalement par pas d'un pixel. Les 67 cartes montrant les zones de dégradation consécutive à l'occlusion ont permis de calculer une carte moyenne ainsi qu'un écart-type. En divisant cette moyenne par l'écart-type des cartes d'un individu, nous avons obtenu une forme de rapport signal/bruit, mettant en avant les zones fréquentielles qui étaient en moyenne très pertinentes pour la bonne classification du locuteur en question et dont la pertinence variait relativement peu entre les 67 /ɑ̃/ de ce locuteur. Deux exemples de ces cartes sont représentés dans la Figure 6.4, avec le locuteur 45 qui, incidemment, présente la zone critique la plus forte de tout le groupe (aux alentours de 1 000 Hz) et le locuteur 09, pour qui certaines zones pertinentes (en jaune) apparaissent. Les 45 cartes SNR sont présentées dans l'Annexe A.

57. https://github.com/onnx/models.

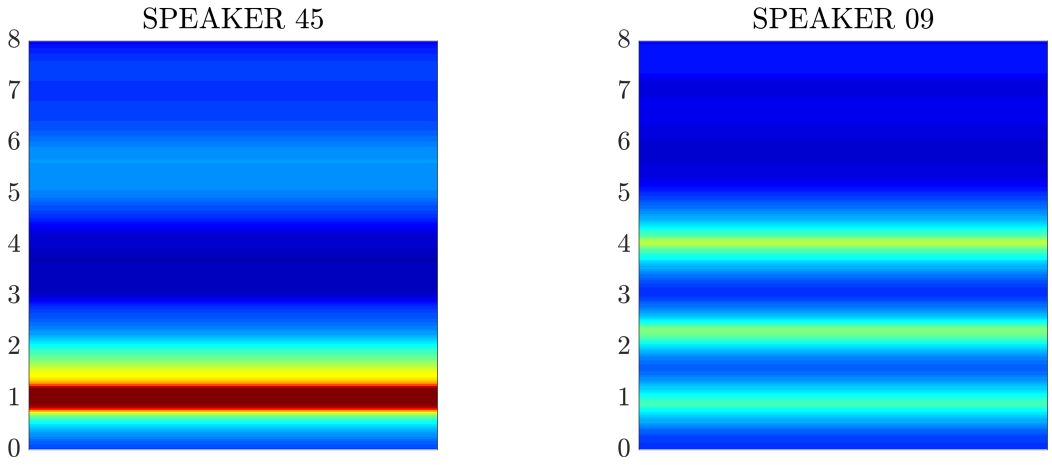



FIGURE 6.4 – Deux exemples de cartes SNR pour les locuteurs 45 et 09.

## 6.4 Retour sur l'opposition *had-hard* à Hull

Une tentative de modélisation des oppositions de type *had-hard* à Hull a déjà été esquissée à la Section 4.4.2. Ici, je souhaitais brièvement montrer une illustration de ce que peut apporter la visualisation d'un réseau convolutif.

Les 1 076 occurrences de voyelles de type *had* et les 1 078 voyelles de type *hard* du corpus de Hull ont été manuellement segmentées et les signaux correspondants ont été convertis en spectrogrammes dans `Praat` avec une fréquence maximale de 4 000 Hz, une fenêtre gaussienne de 5 ms, un pas d'analyse de 1 ms, une résolution de 20 Hz et un filtre de pré-emphase de 6 dB par octave. Les spectrogrammes ont été pré-traités de sorte que le réseau ait en entrée des images en niveau de gris codés sur 8 bits et une taille de $64 \times 64$ pixels. La dynamique a été maximisée sur l'intégralité des 8 bits. On notera au passage deux points cruciaux : la normalisation temporelle implicite — quelle que soit sa durée initiale, la voyelle occupe désormais 64 pixels sur l'axe temporel — et la réduction importante de l'information fréquentielle, puisqu'on passe de 200 bandes (4 000 Hz divisés par 20 Hz) à 64.

Le réseau apprend ses paramètres sur un ensemble issu d'un tirage aléatoire de 80 % des voyelles. Le cycle d'apprentissage est limité à 10 époques de 17 itérations chacune. Une itération correspond dans notre cas au passage de 100 voyelles dans le réseau, suscitant à chaque fois une mise à jour des poids ; quand l'intégralité de l'ensemble d'apprentissage a été vu par le réseau — environ 1 700 items ici — une époque a été complétée. Schématiquement, l'apprentissage atteint 75 % de précision dès les toutes premières itérations, puis progresse un peu plus lentement pour atteindre les 90 % vers la fin de la deuxième

époque, et se « stabilise » à partir de cet instant dans une oscillation entre 90 et 100 % de précision. Dans la phase de test, le modèle affiche une précision de 93,3 % avec une répartition des erreurs équilibrée entre les deux classes.

À ce stade, il est légitime de se demander comment le réseau a pu apprendre avec succès un contraste qui s'appuie très majoritairement sur la durée alors même que ce paramètre est exclus de l'analyse du fait de la normalisation temporelle. Il n'est pas difficile d'anticiper la réponse : la normalisation, qui revient schématiquement à appliquer un zoom d'autant plus grossissant que la voyelle est courte, conduit les stries des spectrogrammes à apparaître comme plus larges et plus espacées pour les voyelles brèves. Ainsi, comme le confirme l'image *deep dream* issue de la dernière couche dense du réseau dans la Figure 6.5, ce que le modèle apprend en réalité, c'est une texture particulière, qui n'est qu'un corrélat de la différence de durée, cette dernière étant le véritable facteur premier qui explique le contraste entre les deux voyelles.

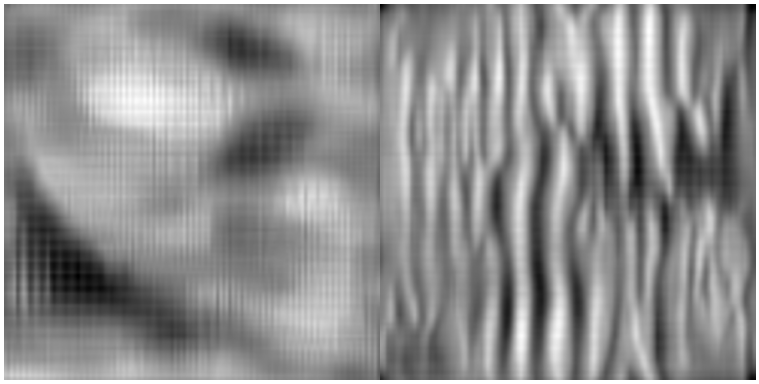


FIGURE 6.5 – Image *deep dream* pour les voyelles de type *hard* (gauche) et *had* (droite).

## 6.5 Visualisations pour la phonétique articulatoire

J'ai eu plusieurs fois l'occasion de dire que si les linguistes étaient médecins, les phonéticiens seraient radiologues (Ferragne, 2019). Ceci est particulièrement vrai pour l'analyse d'images d'échographie de la langue. En effet, la nature des images obtenues rend difficile une analyse automatique des contours de la langue ; l'intervention d'un expert humain est souvent requise. Avec l'avènement du *deep learning*, l'analyse d'images médicales est facilitée, et il arrive même parfois que la machine surpasse le radiologue dans certains tâches (Litjens *et al.*, 2017). Dans le cas précis de l'échographie de la langue pour l'étude de la phonétique, notre travail avec Hannah King, dont une partie est présentée dans King

et Ferragne (2019) et King et Ferragne (2021), m'aura enseigné un bon nombre de points emblématiques du *deep learning*.

### 6.5.1 /r/ hyperarticulé : le rôle des lèvres

Les données de la thèse de Hannah King comportent entre autres des enregistrements d'anglophones produisant des /r/, hyperarticulés ou non, à l'initiale de monosyllabes (e.g. *reed*), et nous disposons de 4 types de signaux synchronisés : l'audio, l'échographie de la langue, une vidéo filmant les lèvres de face, et une autre vidéo capturant les lèvres de profil.

Les images des caméras de face et de profil au point de protrusion maximale ont été extraites des vidéos. Les images d'échographie de la langue montrant le geste lingual antérieur maximal lors de l'articulation du /r/ ont également été extraites. Trois modèles différents (face, profil, échographie) ont ainsi été appris. Nous avons repris l'architecture d'un réseau à convolution pré-existant qui s'est illustré dans le domaine de la reconnaissance d'images, en l'occurrence, ResNet-18 (He *et al.*, 2016)[58].

Chacun des 3 modèles a exécuté deux tâches différentes : la classification des /r/ selon qu'ils étaient hyperarticulés ou non (c'était notre facteur expérimental) et la classification des /r/ en deux configurations linguales : *bunched*[59] ou rétroflexe.

Les images ont été redimensionnées pour correspondre à la taille d'entrée du réseau, 224 × 224 pixels. Les niveaux de gris ont été recodés sur 8 bits de profondeur afin de ne pas surcharger la mémoire de l'ordinateur. Les premiers essais, qui présentaient un surapprentissage assez marqué, nous ont conduits à optimiser les capacités de généralisation de nos modèles par le biais de l'augmentation de données. Ainsi, les images en entrée ont subi des transformations aléatoires (rotations, translation, changements d'échelle), ce qui a eu l'effet escompté. Nous avons en outre utilisé la validation croisée : chaque modèle a été ré-appris 10 fois sur 90 % des données et testé sur les 10 % restants. La méthode CAM (Class Activation Maps, Zhou *et al.*, 2016) a en outre été appliquée pour visualiser les activations au niveau de la dernière couche ReLU du réseau en sortie de la dernière couche convolutive.

La précision moyenne (après validation croisée) et l'écart-type de cette précision pour la classification en hyperarticulés ou non sont de 77,82 % (12,04 %) pour les images ultra-

58. Avec le recul, il s'agit du CNN avec lequel j'ai toujours obtenu les meilleurs résultats ; il présente en plus l'avantage de converger beaucoup plus vite que, par exemple, les réseau de type VGG.

59. À ma connaissance, il n'existe pas de terme en français adapté pour décrire cette articulation caractérisée par une position basse de la pointe de la langue et une élévation du dos de la langue.

son ; 88,44 % (4,90 %) pour la caméra de face et de 70,94 % (16,99 %) pour la caméra de profil. Bien que ces scores soient statistiquement différents du hasard d'après un test binomial, on peut noter une nette supériorité du modèle utilisant l'image des lèvres de face. Cette supériorité s'exprime non seulement à travers une performance moyenne plus élevée, mais également par des scores plus stables d'un modèle à l'autre lors de la validation croisée comme en témoigne l'écart type relativement faible de 4,90 %.

Concernant le modèle bunched vs rétroflexe, on obtient 97,82 % (2,29 %) de précision avec les images échographiques, 97,60 % (2,76 %) avec les images de face, et 96,64 % (3,23 %) avec les images de profil. Ces scores particulièrement élevés et stables d'une itération à l'autre de la validation croisée cachent en réalité un biais méthodologique crucial que la méthode CAM nous a permis de repérer instantanément.

En effet, lorsqu'on regarde la Figure 6.6, où sont superposées, selon la méthode CAM, l'image originale et les zones de cette image sur lesquelles le CNN porte son attention pour effectuer sa tâche, on constate, pour l'image de gauche, que la zone retenue (en rouge) est plausible. À l'inverse, concernant l'item de droite, et bien qu'il soit correctement identifié comme rétroflexe, le CNN s'est appuyé sur une partie de l'image qui ne permet pas d'aboutir à une interprétation phonétique satisfaisante.

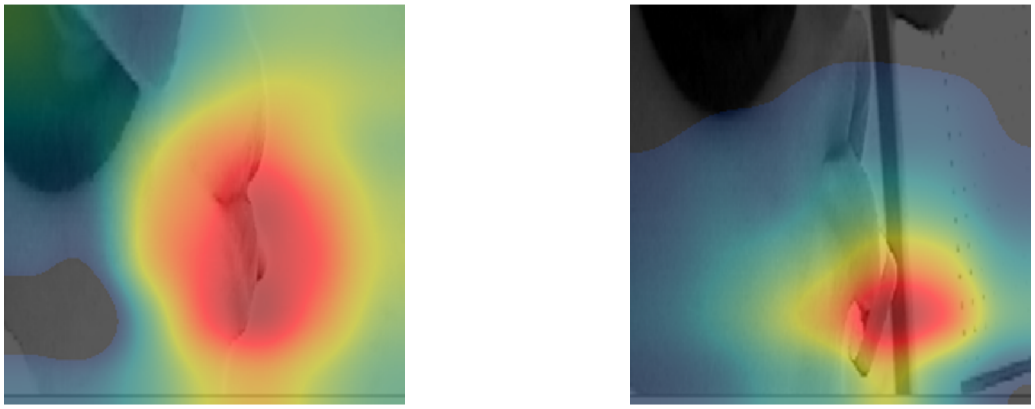


FIGURE 6.6 – Visualisation des activations du CNN pour la classification de /r/ ; bunched à gauche et rétroflexe à droite.

L'image de droite de la Figure 6.6 montre que le CNN s'est focalisé sur une partie du casque pour prendre sa décision. Dans cette étude, notre facteur expérimental opposait /r/ neutres et /r/ hyperarticulés ; ce n'est qu'a posteriori que nous avons souhaité modéliser la différence entre bunched et rétroflexe. Or les données n'ont pas été précisément collectées pour répondre à cette seconde question, et il s'est produit ce qui peut arriver dans ce type de situation (c'est d'ailleurs le type de biais que je reprochais aux corpus tout venant à la

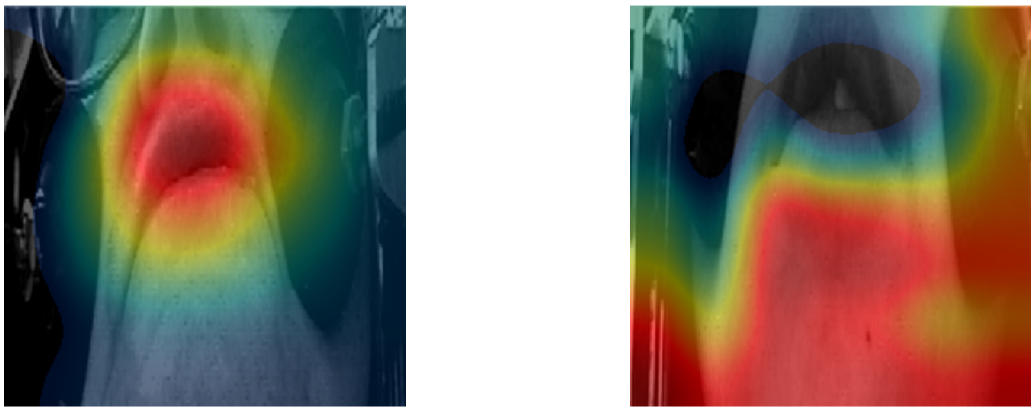


FIGURE 6.7 – Visualisation des activations du CNN pour la classification de /r/ hyperarticulés vs neutres ; les deux exemples sont neutres.

Section 2.3.1) : nous sommes face à un facteur confondant. En effet, puisque la majorité des locuteurs de l'anglais utilise exclusivement soit l'un, soit l'autre type de /r/, et que la position exacte de la caméra et du casque varie d'un locuteur à l'autre, le modèle s'est reposé sur la différence inter-locuteurs la plus évidente pour prendre sa décision. En d'autres termes, la décision était bonne ; mais pour une mauvaise raison !

La Figure 6.7, qui concerne la modélisation de la différence entre /r/ neutre et hyperarticulé, montre, à gauche, une réalisation neutre correctement identifiée. La zone saillante révélée par CAM est tout à fait plausible sur le plan de la phonétique articulatoire. L'image de droite, quant à elle, n'a pas été correctement classifiée. Le caractère très diffus de la région sur laquelle s'est appuyé le modèle et le fait que cette région englobe des parties qui ne présentent que peu d'intérêt phonétique permettent d'entrevoir pourquoi une mauvaise décision a été prise.

Ces illustrations viennent appuyer un élément méthodologique récurrent dans ma recherche actuelle : les réseaux de neurones artificiels profonds ne méritent plus d'être considérés comme des boîtes noires. À des fins pédagogiques, ou encore dans le but de mieux explorer soi-même ses propres résultats, on peut produire des graphiques contenant encore plus d'information utile. La Figure 6.8 représente un « embedding », c'est-à-dire une projection des images dans un espace de dimensionnalité réduite grâce à la méthode t-SNE. Il s'agit ici des images de face de l'ensemble test du modèle ResNet-18 entraîné à distinguer les /r/ hyperarticulés des /r/ neutres. Les cadres jaunes représentent les productions hyperarticulées ; les bleus, les productions neutres. Les images rouges montrent les occurrences mal classées. La méthode CAM, appliquée à la couche res5b_relu permet de localiser, au moyen d'une carte de chaleur superposée à l'image originale, les zones

d'activations pertinentes pour la classification. La méthode t-SNE a été appliquée à la couche pool5 de ResNet-18, réduisant ainsi les 512 dimensions originales des activations de cette couche aux deux dimensions souhaitées pour la représentation graphique.

La Figure 6.8 est bien sûr perfectible. On pourrait la rendre interactive : cliquer sur une image pour l'agrandir, passer le curseur sur l'image et voir les métadonnées correspondantes s'afficher, ou encore obtenir 3 dimensions en sortie de t-SNE et produire une représentation en 3D sur laquelle l'utilisateur puisse opérer des rotations.

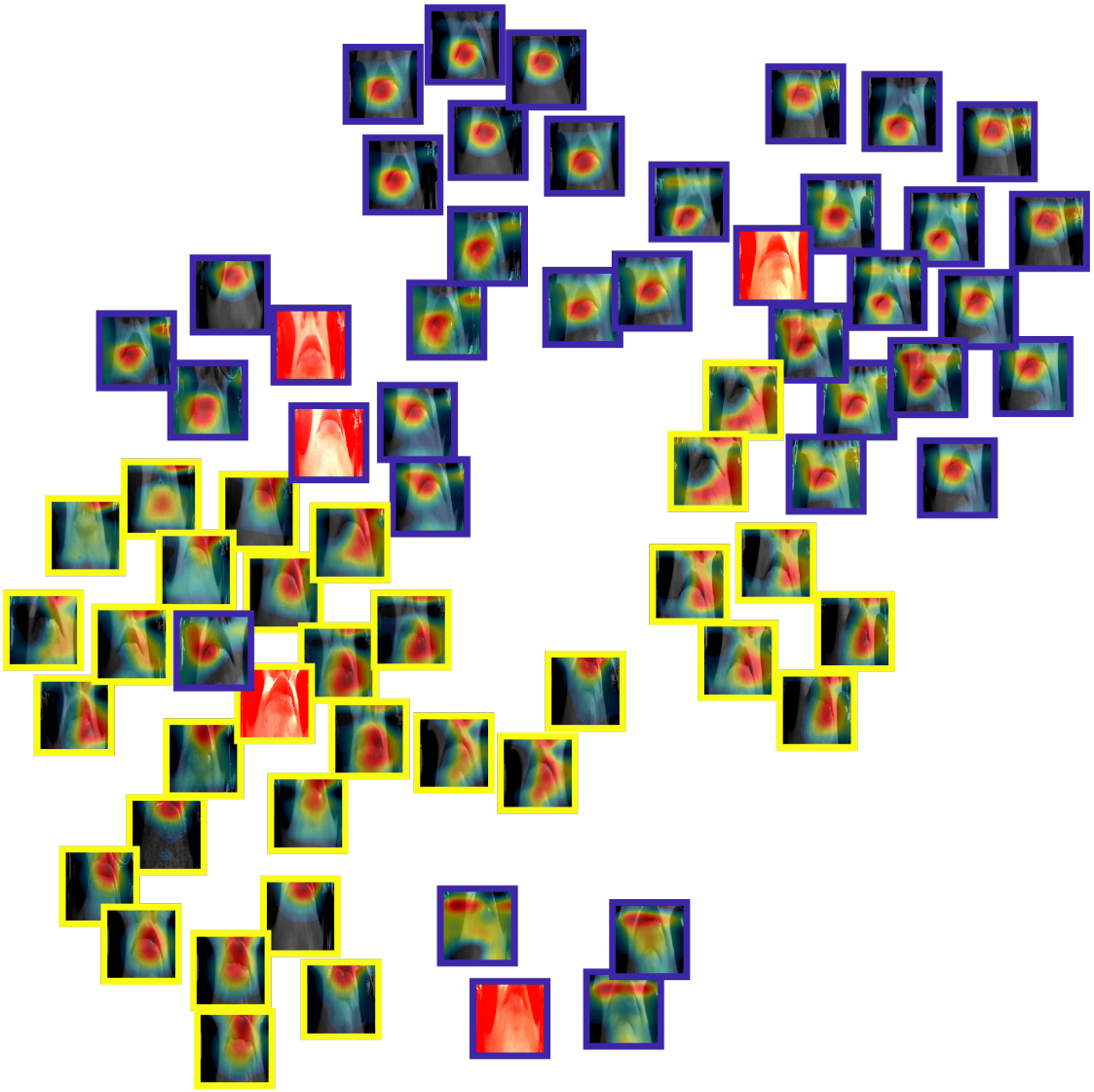


FIGURE 6.8 – Projection des images de l'ensemble test d'un modèle entraîné à distinguer les /r/ hyperarticulés (cadre jaune) des /r/ neutres (cadre bleu) par la méthode t-SNE appliquée à la dernière couche de *pooling* d'un ResNet-18. Une carte de chaleur issue de la méthode CAM est superposée aux images. Les images rouges sont des erreurs de classification.

### 6.5.2 Deux gestes labiaux distincts

Les travaux de thèse de Hannah King ont montré qu'un type particulier de labialisation accompagnait la production du [ɹ] en anglais d'Angleterre. Le développement rapide d'un allophone labiodental ([ʋ]) dans cette variété découle probablement de la préservation de l'articulation secondaire labiale du [ɹ] aux dépens du geste primaire, lingual. Il est donc probable que ce geste labial, y compris quand il se contente d'accompagner le geste lingual, soit déjà labiodental, ce qui le distinguerait du geste labial du [w], qui est unanimement décrit comme arrondi. Dans un King et Ferragne (2021), nous avons souhaité vérifier cette hypothèse de deux gestes labiaux distincts, prévoyant une configuration plus labiodentale pour [ɹ], en utilisant une méthode automatique impliquant le *deep learning*.

Chacun des 23 locuteurs a produit 9 paires minimales du type *reed-weed*, ce qui fait un total de 414 vidéos des lèvres prises de face, dont ont été extraites manuellement les 414 trames correspondant à une constriction labiale maximale. Dans un premier temps, afin de nous assurer de la possibilité de distinguer automatiquement les deux phonèmes initiaux simplement à partir des images des lèvres, nous avons mené une expérience de classification automatique[60]. Nous avons d'abord utilisé une validation croisée en partitionnant aléatoirement les données en 10 sous-ensembles, et avons obtenu un score moyen de classification correcte de 99,52%, avec un faible écart-type de 1,02%. Puisque tous les locuteurs apparaissent potentiellement dans l'ensemble d'apprentissage et dans le test, nous avons, dans un second temps, procédé à une validation croisée de type *leave-one out* afin de mettre à l'épreuve les capacités de généralisation de notre modèle. À chaque étape de validation, c'est l'ensemble de données intégral d'un locuteur qui sert de test alors que le modèle est appris sur tout le reste. Cette méthode permet d'atteindre un score de classification correcte de 92,27%, avec un écart-type de 14,86%. Ces résultats indiquent que la fiabilité de la distinction des gestes labiaux varie d'un locuteur à l'autre.

Dans un second temps, nous avons souhaité caractériser cette différence de configuration labiale en termes articulatoires. Des essais que j'avais menés à d'autres occasions avaient montré que la segmentation automatique des lèvres en utilisant la couleur fonctionnait de manière satisfaisante à condition que l'éclairage permette une reproduction fidèle des couleurs. Ce n'était malheureusement pas les cas dans les données qui nous occupent ici, et nous n'avions pas anticipé qu'un simple rouge à lèvres bleu aurait pu nous faciliter la tâche. Nous nous sommes donc tournés vers les techniques de segmentation sémantique. Cent images parmi les 414 ont été segmentées manuellement afin de délimi-

60. Pour plus de détails techniques, voir King et Ferragne (2021).

ter les lèvres du reste de l'image. Un réseau de neurones profond de type DeepLab v3+ (Chen *et al.*, 2018) a été entrainé à distinguer les lèvres du reste. La Figure 6.9 illustre le résultat de notre segmentation : le haut de l'image montre en bleu les pixels automatiquement identifiés comme faisant partie des lèvres. Au-dessous, on voit l'ellipse qui a été automatiquement ajustée à la zone bleue, avec son axe horizontal qui nous a permis de mesurer la largeur de la bouche, son axe vertical, qui mesure la hauteur, et le point d'intersection des deux axes qui permet de caractériser la position horizontale et verticale de la bouche.

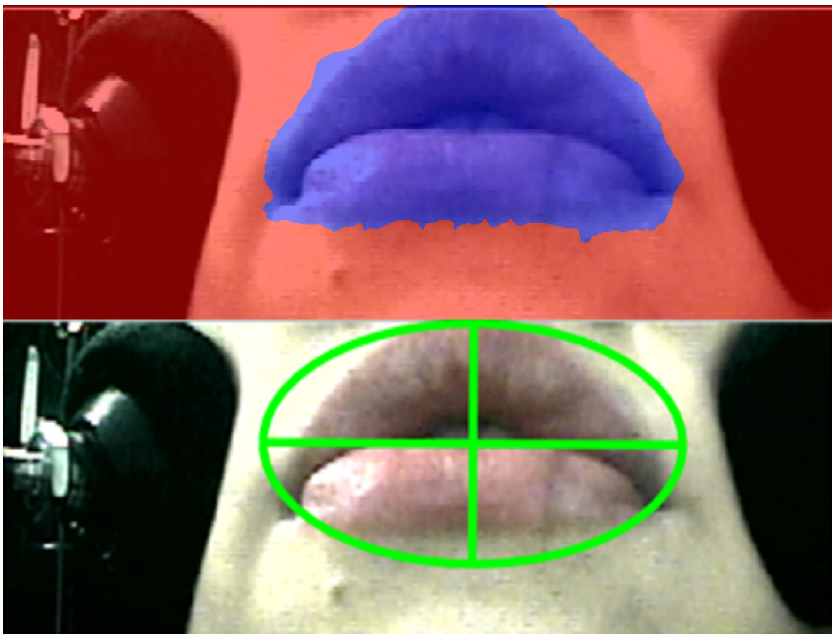


FIGURE 6.9 – Illustration du processus de segmentation sémantique et d'ellipse ajustée aux pixels de la bouche.

Les analyses statistiques pourront être consultées dans l'article. Elles montrent que de ces quatre paramètres, seule la position horizontale des lèvres n'est pas significative. Les [w] sont donc produits avec des lèvres moins larges, recouvrant une zone plus grande sur l'axe vertical et avec une position plus basse par rapport à [ɹ]. La taille des effets montre que la largeur des lèvres et la position verticale sont les prédicteurs les plus fiables de la différence entre les deux phonèmes. Notre hypothèse selon laquelle nous avons affaire à deux gestes labiaux différents, avec le [ɹ] présentant une position de la bouche rappelant une configuration labiodentale, est donc corroborée.

## 6.6 D'autres exemples d'utilisation en phonétique

J'ai connu une période d'environ un an et demi entre 2018 et 2019 où j'ai pratiqué le *deep learning* comme un musicien professionnel ses gammes. J'ai expérimenté avec de nombreux types de modèles différentes — réseaux siamois, réseaux antagonistes génératifs, réseaux récurrents de type *Long Short Term Memory*, etc. — et avec des types de données très variés : signal audio, électroencéphalographique, images médicales, échographies de la langue, spectrogrammes, etc. Pendant cette période, des centaines de modèles ont littéralement fait chauffer mon Alienware jour et nuit.

Parmi les résultats intéressants, je mentionnerai une étude menée avec Jalal Al-Tamimi, où la méthode de l'occlusion a permis de localiser dans des spectrogrammes les zones fréquencielles caractéristiques de différentes classes de consonnes de l'arabe (Al-Tamimi et Ferragne, 2020). L'exemple des consonnes pharyngales et des pharyngalisées est donné dans les Figures 6.10 et 6.11. On y voit par exemple qu'alors que les régions pertinentes pour la classe des pharyngales se concentrent essentiellement dans la consonne (C) présente entre les deux voyelles des mots test, les régions typiques des pharyngalisées sont localisées dans la seconde voyelle.

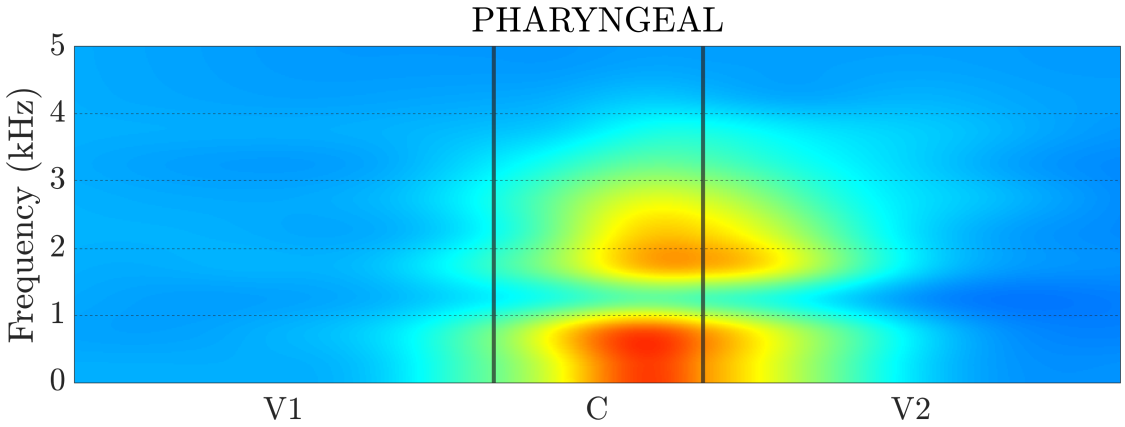



FIGURE 6.10 – Analyse de sensibilité à l'occlusion des consonnes pharyngales de l'arabe après classification par un réseau convolutif en différentes classes naturelles.

Un second exemple illustrant l'intérêt d'utiliser des modèles qui extraient leurs propres paramètres provient des travaux que j'ai effectués avec Anne Guyot-Talbot sur l'accentuation des dissyllabes pluricatégoriels en anglais. Par exemple, selon que *record* est nom ou verbe, le schéma accentuel est différent. On simplifie souvent la règle de la façon suivante : les noms sont accentués sur la première syllabe, et les verbes, sur la seconde. En s'appuyant sur les courbes moyennes d'intensité relevées sur un corpus constitué pour

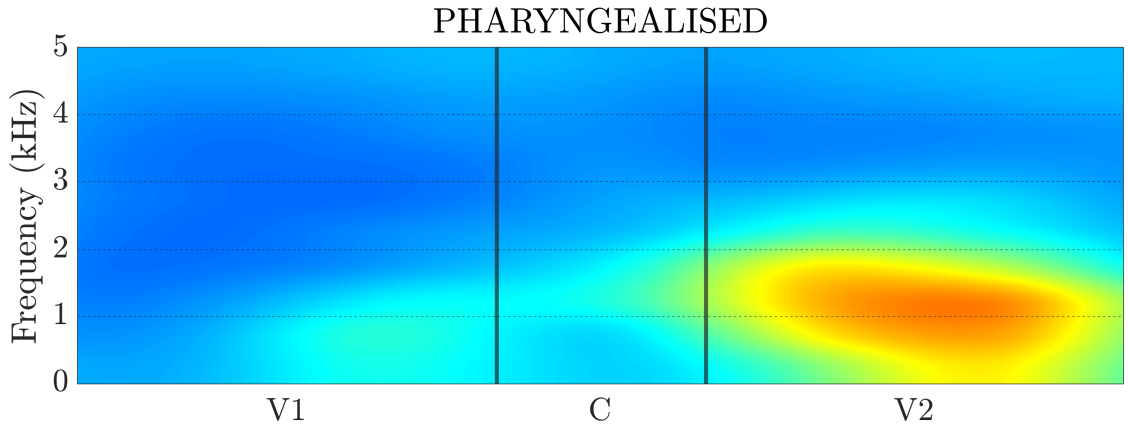



FIGURE 6.11 – Analyse de sensibilité à l'occlusion des consonnes pharyngalisées de l'arabe après classification par un réseau convolutif en différentes classes naturelles.

l'occasion à partir d'exemples recueillis sur Youtube, on voit à la Figure 6.12 que *record* semble suivre l'alternance de la règle simplifiée, alors que *concern* porte l'accent lexical sur sa dernière syllabe, quelle que soit sa catégorie grammaticale.

Or la perception de l'accent lexical est multidimensionnelle : de nombreux paramètres (f0, intensité, durée, timbre de la voyelle) parfois concourent à donner l'impression d'accentuation, et parfois s'opposent au point où il est difficile d'établir le schéma accentuel. Dans ces derniers cas, une solution intéressante a consisté pour nous à d'abord faire apprendre à partir des spectrogrammes de mots entiers la distinction entre les schémas accentuels /10/ et /01/ sur les items pour lesquels l'expert humain n'avait absolument aucun doute. Le modèle atteint 85 % de classification correcte. Dans un second temps, le modèle se substitue à l'expert humain pour décider du schéma accentuel des items ambigus.

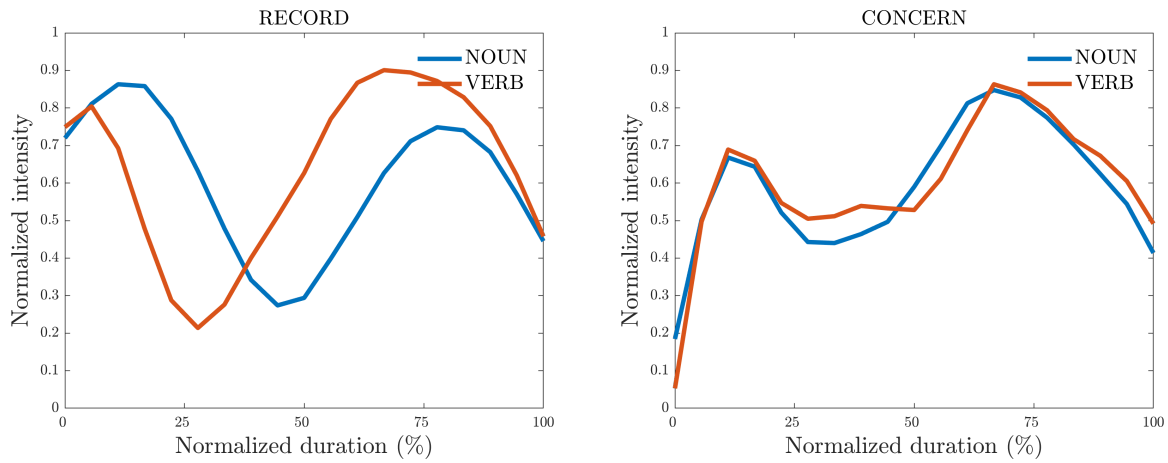



FIGURE 6.12 – Courbe d'intensité moyenne pour *record* (558 occurrences) et *concern* (242 occurrences), nom et verbe.

## 6.7 Vers un changement de paradigme

J'espère avoir démontré à travers ces quelques exemples que le *deep learning* appliqué à la phonétique constitue un nouveau paradigme particulièrement prometteur. Pour être plus précis, il s'agit en réalité d'une adaptation des techniques de la vision par ordinateur à l'analyse d'images issues de données phonétiques. Il y a bien eu quelques travaux précurseurs (Nagamine *et al.*, 2015 ; Pellegrini et Mouysset, 2016) s'intéressant aux applications du *deep learning* à la phonétique, mais je crois pouvoir dire que je suis parmi les tout premiers phonéticiens à développer cette approche.

Les avantages sont nombreux, et si ce nouveau programme de recherche est considéré avec le sérieux qu'il mérite, je ne doute pas un instant de l'importance du gain pour notre domaine. En premier lieu, une option viable consiste à se représenter les réseaux convolutifs profonds comme une méthodologie unique pourtant transférable en l'état à une grande variété de données différentes. On peut traiter sans modification du réseau des images d'objets, des spectrogrammes, ou encore des radiographies du poumon, etc. Les modèles que j'emploie sont capables d'extraire leurs propres descripteurs ; une segmentation manuelle précise n'est donc souvent pas requise puisque l'information pertinente, où qu'elle se situe dans l'image, est automatiquement extraite. Quand on sait à quel point la segmentation manuelle dans les corpus phonétiques est fastidieuse, l'intérêt d'économiser du temps s'impose de lui-même. De plus, la possibilité de rendre le processus entièrement automatique garantit une reproductibilité accrue, ce qui constitue un gage d'objectivité scientifique. Les méthodes de visualisation constituent elles-aussi un avantage indéniable : il s'agit là non seulement d'un atout pédagogique pour communiquer sa recherche au plus grand nombre, mais également, comme nous avons pu le constater fortuitement dans nos travaux (Section 6.5.1), d'indices précieux pour détecter un éventuel biais méthodologique.

En expérimentant, une analogie avec l'apprentissage chez l'humain m'est rapidement apparue. Comme le montre la revue de questions dans la thèse de Jennifer Krzonowski, on sait que l'acquisition de nouveaux contrastes phonémiques chez des apprenants est d'autant plus robuste que ces contrastes sont présentés avec des voix différentes ; c'est le principe du *high variability phonetic training*. En *machine learning*, il est parfois souhaitable d'ajouter artificiellement de la variation dans ses données pour accroître les capacités de généralisation d'un modèle ; c'est ce qu'on appelle « l'augmentation » de données, et j'ai pu constater de très nettes améliorations avec cette méthode. On entend parfois les collègues s'étonner d'une grande variation dans leurs données. Dans ma thèse, j'écrivais que c'est plutôt de l'absence de variation qu'il faudrait s'étonner. Aujourd'hui, avec mon

expérience de l'augmentation des données, j'ai à disposition un outil pédagogique qui démontre à quel point une absence de variation conduit à une impossibilité de catégoriser.

Je ne voudrais cependant pas laisser croire que l'intervention d'un expert humain dans le processus de recherche aura complètement disparu ; bien au contraire ! J'ai la conviction — et je l'expérimente moi-même au quotidien — que certes de plus en plus de tâches rébarbatives vont être confiées à la machine, mais avec une telle puissance, l'expert humain doit redoubler d'effort pour, par exemple, concevoir des protocoles de recueil de données avec un contrôle accru des éventuels biais. Ce point est particulièrement bien illustré dans la description de l'article de King et Ferragne (2019) à la Section 6.5. En effet, nous disposions d'un modèle présentant des taux de classification correcte très élevés, mais qui a eu l'intelligence, quoique artificielle, de tirer partie d'un biais inhérent aux données.

Les réactions des collègues face à ces idées nouvelles me semblent contrastées. La science, par nature, n'accepte pas facilement la nouveauté, ce qui constitue à la fois un défaut et une qualité : il m'incombe de faire la preuve de l'intérêt de ma démarche. La conférence invitée que j'ai donnée devant des collègues anglicistes (Ferragne, 2019) dans le cadre du colloque Phonologie de l'Anglais Contemporain à Aix-en-Provence en 2019 a été accueillie avec intérêt et bienveillance. Dans une communauté où la plupart des collègues n'utilisent que peu de méthodes d'analyse quantitative, la discussion, motivée par une véritable curiosité de la part de l'auditoire, a été sereine et enrichissante. En d'autres circonstances, et malgré les efforts que je consacre à trouver les mots pour bien expliquer ma démarche, il arrive ponctuellement que mes propositions soient rejetées. Les motivations sont extrêmement variées et ont toutes, à leur manière, une justification valable : la nouveauté peut effrayer, ma légitimité n'est pas encore totalement établie, il existe une défiance envers l'intelligence artificielle et « les algorithmes[61] », etc.

61. Il est amusant de noter à quel point un terme dénotant une méthode reproductible pour résoudre un problème en est venu, par connotation, à désigner une « cuisine » magique et incontrôlable n'agissant jamais pour le bien des humains.

# 7

# Bilan et perspectives

## Sommaire



## 7.1 *Things I have learned (so far)*

Ce dernier chapitre a pour objectif dans un premier temps d'établir le bilan de mon évolution professionnelle et de rappeler le fil conducteur qui lie le patchwork des nombreuses facettes de ma recherche. Je commenterai ensuite les changements survenus (ou en cours) dans mon domaine depuis ma prise de poste. Je terminerai enfin par quelques réflexions prospectives en m'appuyant sur mes projets actuels.

« Things I have learned (so far) », c'est comme cela que s'intitule l'article de Cohen (1990). Ce titre illustre bien ce que je pense devoir être mon attitude face à ma recherche. En effet, en dépit de notre expérience, il faut être disposé à tout moment à remettre en question nos positions, d'où l'ajout capital de « so far ». Nos savoirs sont périssables puisque, d'une part, le contexte dans lequel ils ont été construits est amené à changer (Oreskes, 2019), et d'autre part, d'après Popper (2002b), les données ne peuvent corroborer que provisoirement nos hypothèses dans l'attente de nouveaux tests qui viendront soit les corroborer à nouveau, soit les falsifier.

Une partie de la liste des « choses que j'ai apprises jusqu'ici » a été esquissée dans ce document ; je vais me contenter ici de résumer quelques points à la lumière du titre. La coloration « autobiographique », qui m'est venue assez naturellement, est en définitive conforme à la position de Oreskes (2019), qui pose qu'un résultat de recherche ne peut être apprécié pleinement qu'à la lumière des valeurs de l'individu qui le produit.

### 7.1.1 Du groupe à l'individu

L'expression « du groupe à l'individu » dans le titre de ce document reflète d'abord une réalité visible de mon parcours : ma thèse portait sur des groupes, en l'occurrence, des accents, et mes intérêts actuels font la part belle à l'étude des spécificités phonétiques du locuteur. Cette évolution est en fait trompeuse : je pense avoir toujours mis un point d'honneur à étudier en parallèle et le groupe, et l'individu. Pour s'en convaincre, il suffit de feuilleter ma thèse (Ferragne, 2008) et les articles qui en sont issus (par ex. Ferragne et Pellegrino, 2010b) : les mesures de tendances centrales et de dispersion (donc pour les groupes) côtoient des spectrogrammes reflétant des spécificités individuelles. Chaque locuteur dans la thèse appartient à son groupe, mais avec des degrés variables de conformité au prototype. Ces degrés avaient été estimés d'un côté à partir des données acoustiques, et de l'autre, par un phonéticien, en l'occurrence le Professeur Francis Nolan, qui avait bien voulu se prêter à une longue expérience de perception. Je me suis donc toujours efforcé de regarder simultanément ces deux niveaux de granularité — et cela ne s'est jamais démenti jusqu'à aujourd'hui — car l'un ne va pas sans l'autre.

D'ailleurs, c'est entre autres la nécessité d'avoir en permanence à disposition les métadonnées individuelles qui m'a poussé à lancer le développement du logiciel `ROCme!`. La nature du lien entre le groupe et l'individu dans un corpus ne va cependant pas de soi. J'y faisais déjà allusion dans ma thèse (Ferragne, 2008, p. 338 *sqq.*) en opposant ce que j'appelais à l'époque le « prototype probabiliste » au « prototype de l'expert ». Dans le

contexte de l'étude des accents de l'anglais, le premier est le locuteur affichant par exemple des valeurs acoustiques proches de la moyenne de son groupe. Le second, c'est celui qu'un expert humain a sélectionné comme étant typique de son accent. Dans le premier cas, le prototype est entièrement construit à partir des données, et donc aussi biaisé que la méthode qui a présidé à leur collecte. Puisque l'échantillonnage lors de l'enregistrement d'un corpus phonologique n'est généralement pas strictement aléatoire, le prototype obtenu risque de mal incarner l'accent dont il est censé représenter la tendance centrale. À l'inverse, si l'échantillonnage est dirigé par un expert qui « filtre » le recrutement, on injecte certes une touche de prescriptivisme, mais on garantit une certaine homogénéité.

Il faut bien sûr entendre ici « individu » non seulement comme un être humain, mais également dans le sens statistique, comme une occurrence du phénomène qu'on observe, qui sert, avec les autres occurrences du même type, à construire des généralisations. Cet individu statistique, cette occurrence unique de voyelle, il faut à tout moment pouvoir revenir à son niveau dans un scénario d'analyse. Car il faut en effet être en mesure de questionner son appartenance à son groupe présumé. Pour reprendre un de mes exemples fétiches, déjà utilisé dans ma thèse (Ferragne, 2008, p. 260) : si j'enregistre des locuteurs de Newcastle produisant la voyelle de FACE, j'aurai des occurrences de diphtongues fermantes et de diphtongues centripètes. Si je calcule une trajectoire formantique moyenne, j'obtiendrai probablement une monophtongue !

Un autre aspect lié à l'individu statistique — à la voyelle, à la consonne ou au contour intonatif individuels — qui me tient à cœur concerne tous les phénomènes liés à ce que j'ai appelé la « base de temps ». Ces concepts sont très brièvement mentionnés à la Section 4.4.3 : mesurer plusieurs voyelles à 20 % et 80 % de leur émission, comme c'est l'usage dans les études sur la dynamique des voyelles (Fox et Jacewicz, 2009 ; Elvin *et al.*, 2016), revient à effacer les différences de durée brute. Je crois qu'il est important de considérer que cela ne va pas de soi : par exemple, quand le débit de parole s'accélère, on a tendance à préserver la durée des transitions formantiques aux dépens des états stables (Janse, 2004), et ces mêmes transitions affichent, à débit supérieur, des valeurs initiales plus proches de la cible (Gay, 1978). Il paraît donc nécessaire d'envisager la possibilité de recourir à des techniques de recalage temporel (Gubian *et al.*, 2015), qui s'appuieraient par exemple sur la détection d'un élément phonétiquement pertinent (par ex. une accélération soudaine dans la courbe du formant) pour aligner toutes les voyelles. Une autre solution, que nous avons utilisée dans la comparaison de contours intonatifs (Guyot-Talbot *et al.*, 2016), consiste à appliquer des distorsions de phase locales en utilisant le *dynamic time warping*

(DTW).

L'intérêt que je porte aux questions d'alignement temporel ne se limite pas au signal audio. Le calcul de grandes moyennes dans l'analyse des potentiels évoqués en EEG soulève parfois la question de la variabilité de latence d'un essai à l'autre, et cette variabilité peut comporter des biais qui viennent fausser la représentation de l'onde qu'on obtient après le calcul de la moyenne (Luck, 2014). Nous avons employé l'une des techniques existantes pour corriger ces décalages de phase dans Heidlmayr *et al.* (2021) et Pélissier et Ferragne (2021).

Pour finir sur cette notion de rapport entre l'individu et le groupe, j'aimerais souligner un point élémentaire de l'inférence statistique, qui concerne le lien entre l'échantillon et la population qu'on souhaite représenter. Les premières pages des ouvrages d'introduction aux statistiques (par ex. Winer *et al.*, 1991 ; Baillargeon, 1982 ; Wonnacott et Wonnacott, 1991) sont souvent consacrées au principe-même de l'inférence : comment on généralise à une population à partir d'un échantillon. Pour que cette généralisation soit valide, il faut tirer aléatoirement des individus dans cette population. Le principe semble aller de soi et son souvenir est vite balayé par les calculs savants qui occupent les pages suivantes. Or ces calculs sont pris en charge par de nombreux logiciels alors que la question de l'échantillonnage ne l'est pas. Je pense pourtant, et j'espère l'avoir démontré au Chapitre 2, que cette question devrait être au centre de nos préoccupations.

### 7.1.2 Du corpus à l'expérimentation

Le Chapitre 2 développe mon positionnement épistémologique en affirmant ma très nette préférence pour les études expérimentales plutôt que la réutilisation de corpus déjà constitués. Puisque cet aspect à fait l'objet d'un chapitre complet, je ne vais pas y revenir dans le détail ici. Le raisonnement inductif, celui qu'on suit quand on souhaite « laisser parler » les données à partir d'un corpus, est très dangereux. Une illustration emblématique de réemploi d'un ensemble de données pour une question qui n'avait pas été posée au moment de sa constitution a été présentée dans la Figure 6.6 à la page 120. La notion de « facteur confondant » y est illustrée de manière éloquente.

En tant que consommateur assidu de *deep learning* et donc de *big data*, je ne peux cependant pas nier l'importance de disposer de données afin de laisser la machine en inférer des généralités. Mais peu importe la taille des données, puisque — pour reprendre l'exemple de la Figure 2.1 à la page 19 — la catégorie « Ohbot » n'étant pas représentée dans l'ensemble d'apprentissage du modèle Alexnet, il ne faut pas s'étonner que le Ohbot

soit reconnu comme étant un téléphone public. Alexnet n'est pas intrinsèquement ohbotophobe ; les millions d'images qui lui ont servi d'apprentissage ne comportaient pas de Ohbot. Le « coupable » n'est pas nouveau comme l'est l'intelligence artificielle ; il est au contraire aussi vieux que le problème de l'induction[62].

Au Chapitre 2, je me suis délibérément concentré sur l'utilisation de corpus à des fins de recherche fondamentale en linguistique et phonologie. Le statut des données pour le traitement automatique des langues et de la parole soulève quant à lui les mêmes problématiques, mais avec des conséquences différentes et donc des réponses différentes (voir Shah *et al.*, 2020, pour un aperçu). Par exemple, la sur- ou sous-représentation de certains accents dans les systèmes de reconnaissance de la parole était déjà une des motivations sous-jacentes de mon travail de thèse. Aujourd'hui, les systèmes sont bien plus performants, mais les disparités dans la prise en compte des accents subsistent. En effet, l'étude de Koenecke *et al.* (2020), qui teste, chez des locuteurs afro-américains et non afro-américains, les performances des systèmes de reconnaissances commerciaux les plus courants, montre des performances moins bonnes avec les Afro-américains. Cela plaide évidemment en faveur d'ensembles de données toujours plus inclusifs, et démontre que la responsabilité éthique ne pèse pas sur la technologie elle-même, mais bien sur les données. De mon point de vue, c'est la même responsabilité qui pèse sur les linguistes qui collectent ou réutilisent des corpus.

« Du corpus à l'expérimentation » exprime une caractéristique de mon parcours : des excursions de part et d'autre de la frontière qui sépare les corpus tout venant des données recueillies pour l'occasion, avec, depuis mon Master, une nette préférence pour les approches expérimentales. Mes incursions dans le « monde » des corpus généralistes m'ont permis de beaucoup réfléchir sur les travers du raisonnement inductif. Fort de cette réflexion, j'aborde mon retour à l'utilisation de grands corpus — dans certains de mes travaux récents impliquant le *deep learning*, comme Ferragne *et al.* (2019), la machine n'apprenant qu'à partir de beaucoup de données — avec un recul certain sur les biais inhérents aux données.

62. Formalisé par David Hume (Curd et Cover, 1998), donc environ 200 ans avant l'avènement de l'intelligence artificielle après la seconde guerre mondiale ; mais, bien sûr, les biais liés à l'inductivisme existent depuis toujours.

### 7.1.3 Du spectrogramme au *deep learning*

J'espère avoir démontré au Chapitre 6 que l'application de techniques de *deep learning* pourrait représenter un changement de paradigme pour notre discipline (Ferragne, 2019). Dans une publication récente (King et Ferragne, 2021), nous avons montré comment la classification automatique et la segmentation sémantique s'appuyant sur le *deep learning* pouvaient répondre à une question de phonétique articulatoire à partir d'images de la bouche des locuteurs. Ces mêmes techniques permettent de repérer des zones pertinentes dans des spectrogrammes (Ferragne *et al.*, 2019 ; Al-Tamimi et Ferragne, 2020), travail dont s'acquittait jusqu'ici l'œil humain. Les réseaux de neurones convolutifs, qu'il est virtuellement possible d'appliquer à tous les types d'image dont la phonétique s'occupe (échographies de la langue, imagerie à résonance magnétique, vidéos, spectrogrammes, etc.), présentent de nombreux avantages motivant ma prédiction d'un changement de paradigme. Ils se caractérisent par une modélisation « de bout en bout », ce qui permet une automaticité supérieure car il n'est même plus nécessaire de fournir explicitement à l'algorithme les portions de l'image contenant l'information pertinente. Cette automaticité garantit *de facto* une reproductibilité accrue puisqu'elle minimise la subjectivité humaine. Bien sûr, comme avec n'importe quelle technique, il convient de bien connaître les limites de la vision par ordinateur appliquée à la phonétique. Par exemple, on peut délibérément tromper ces algorithmes avec des manipulations d'images qui ne duperaient pas l'œil humain (Heaven, 2019). Mais je crois qu'en recherche fondamentale en phonétique et phonologie, où, par exemple, nous ne mettons pas au point de systèmes de conduite autonome, les risques sont limités.

Sur un plan moins méthodologique, j'ai suggéré la possibilité que les réseaux de neurones artificiels constituaient au moins une analogie utile avec la manière dont sont apprises et stockées les représentations phonologiques selon les modèles à exemplaires (Pierrehumbert, 2001, 2016). Mais cette affirmation mériterait de longs développements. Je retiens qu'avec le *deep learning* comme avec les modèles épisodiques, les représentations sont riches, la méthode d'apprentissage et de stockage n'est pas parcimonieuse ; le principe d'économie ne tient plus. La variation, naguère perçue comme un mal ou, au mieux, un mal nécessaire, est réhabilitée. D'ailleurs, expérimenter avec des techniques d'augmentation de données (ajouter du bruit dans les données d'apprentissage pour améliorer les capacités de généralisation d'un modèle) a fini de me convaincre à partir d'exemples visuels concrets que la variation est absolument indispensable à la catégorisation, ou comme le dit Taylor (1995, p. ix) : « the ability to categorize [is] to see similarity in diversity ».

## 7.2 Les grands chantiers

J'aborde à présent une esquisse de réflexion sur quelques éléments choisis de l'évolution du paysage scientifique ces dernières années. À défaut d'avoir des réponses toujours tranchées, il semble important au moins d'attirer l'attention sur les mutations passées et en cours. Le choix des quelques thématiques retenues et ce que je vais en dire est nécessairement très subjectif. J'espère cependant avoir convaincu les lecteurs que la subjectivité est un élément constitutif de l'objectivité scientifique, et qu'il vaut mieux l'afficher honnêtement pour l'ouvrir à la critique plutôt que tenter de la dissimuler.

### 7.2.1 La phonétique dans les Études Anglophones

Mon appréciation de l'évolution de la phonétique et de la linguistique dans la section disciplinaire des Études Anglophones ces dernières années comporte quelques limites car cette évolution s'est faite en parallèle de mes propres changements de statut et de ma connaissance toujours plus précise du métier.

Les départements d'anglais en France ont pour vocation principale de former des enseignants du secondaire notamment à travers la préparation de l'agrégation et du CAPES. Cette contrainte implique dans les faits que la phonétique (ou la phonologie) qu'on y pratique soit, au moins partiellement, en lien avec l'enseignement de l'anglais à des apprenants francophones. Il s'agit donc d'une phonétique « pédagogique », corrective, appliquée, qui est parfois en marge des grandes questions actuelles de recherche fondamentale.

Être phonéticien dans un département d'anglais impose donc de faire des choix concernant le juste degré de chevauchement entre mission pédagogique et thématiques de recherche. En effet, la solution qui consisterait à totalement dissocier sa recherche de son enseignement conduirait, dans mon cas, à vivre ma mission pédagogique comme un fardeau puisqu'elle s'effectuerait aux dépens de ma recherche. À l'inverse, m'astreindre à une recherche strictement en lien avec la pédagogie ou avec l'anglais m'empêcherait de suivre un grand nombre d'envies professionnelles. La diversité du Chapitre 5 raconte finalement l'histoire de l'équilibre auquel je suis parvenu.

Depuis ma prise de poste, j'espère avoir modestement participé à faire évoluer le rôle des phonéticiens dans les départements d'anglais, notamment en contribuant à dynamiser la recherche dans mon UFR. Cet élan a été facilité par l'obtention très tôt dans ma carrière de projets financés et par le maintien de mon lien avec les sciences cognitives d'une part, et les sciences de l'ingénieur de l'autre. Ma démarche a consisté à importer

des méthodes expérimentales et instrumentales qui, à ma connaissance, n'étaient que peu (ou pas) employées en Études Anglophones. J'ai très tôt introduit l'analyse quantitative et la programmation informatique dans mes cours. Je me suis toujours efforcé, en injectant des éléments de formation *à* et *par* la recherche dès la Licence, de préparer au mieux les étudiants à une recherche conforme à ce qui se fait au niveau international. L'image du phonéticien en Études Anglophones en France a évolué ; le stéréotype du professeur de diction laisse peu à peu place à une multiplicité de profils avec des compétences et des objets de recherche très variés. Bien loin d'être le seul artisan de cette mutation, je crois néanmoins y avoir contribué.

En Études Anglophones, tendre vers une recherche en linguistique et phonétique plus internationale, s'appuyant sur davantage de données et des dispositifs expérimentaux toujours plus sophistiqués engendre nécessairement des changements. Pour ne citer que deux aspects : il faut davantage de temps et d'argent que ce à quoi les départements d'anglais en France et leurs équipes d'accueil étaient habitués. Et le temps des enseignants-chercheurs n'étant pas extensible *ad libitum*, l'argent sert aussi à acheter du temps, sous forme de « personnes.mois » (post-docs, stagiaires, contract doctoraux, etc.). On dit souvent que le temps de la recherche est long (par rapport à celui de la politique ou des média) ; je crois pouvoir dire que celui de la recherche expérimentale et instrumentale visant une revue internationale l'est particulièrement. Des premières lectures en vue de la rédaction d'une demande de financement à la publication effective d'un article dans une revue très exigeante, il n'est pas rare que cinq années se soient écoulées.

Et ces cinq années sont jalonnées de nombreuses étapes qui apportent chacune une incertitude supplémentaire quant à la possibilité d'atteindre l'objectif : obtention non garantie du financement, délais dans la validation des aspects éthiques, dans la mise en place des fonds, problèmes de recrutement de participants, risque de ne pas obtenir de résultats probants, intégralité du processus conditionnée par la durée des contrats du personnel recruté pour l'occasion, par la disponibilité des membres du projet, procédure de relecture de l'article par les pairs qui peut être particulièrement longue, etc.

L'aperçu de toutes ces étapes démontre que lorsque plusieurs auteurs, mettons trois, figurent sur un article, ce qui historiquement n'était pas courant en Études Anglophones, il ne faut généralement pas considérer que chacun mérite un tiers de la rétribution (quelle qu'elle soit), mais bien plutôt que la qualité de l'article est trois fois supérieure à ce qu'elle aurait été avec un seul de ses auteurs. L'usage croissant de la *Contributor Roles*

*Taxonomy*[63] dans les revues internationales, qui permet de caractériser la nature de la contribution de chaque auteur à un article, assure d'ailleurs une plus juste reconnaissance des divers rôles (il y en actuellement 14) indispensables dans une recherche de pointe en phonétique, et peut potentiellement dissuader les inclusions abusives. Toutes ces pratiques sont relativement nouvelles en Études Anglophones, et elles génèrent de nombreuses questions comme la faisabilité d'une recherche ambitieuse en phonétique expérimentale et instrumentale hors UMR de sciences du langage.

Pour en revenir à la collaboration, pratique qui n'était pas aussi répandue qu'aujourd'hui en Études Anglophones (et qui ne l'est pas dans tous les sous-domaines) à l'époque de ma prise de poste, je crois qu'elle constitue, y compris quand elle est monodisciplinaire, une piste intéressante pour modérer les biais inhérents à la phonologie quantitative, qu'elle soit de corpus ou de laboratoire. En effet, c'est souvent un seul et même individu qui élabore une hypothèse, collecte les données pour la tester — et ce n'est souvent pas un test au sens de Popper (2002b,a), qui viserait à la réfuter, mais plutôt une recherche de confirmation — et segmente puis analyse des données. Cloisonner ces différentes étapes de la production scientifique en les confiant à des personnes différentes permettrait peut-être de simuler le principe du double aveugle, dont les vertus sont reconnues.

Comme je l'ai déjà noté dans cette section, la phonétique dans les Études Anglophones est très fortement liée à des besoins pédagogiques. Nous enseignons la prononciation de l'anglais, et notre connaissance du contexte sociolinguistique nous permet de :

1. Sensibiliser les étudiants à l'étendue de la variation phonétique et phonologique tout en explicitant les stéréotypes associés à telle ou telle variante ;
2. Encourager l'utilisation de deux accents non stigmatisés — *Received Pronunciation* et *General American* — et n'accepter dans nos évaluations en phonologie que ces deux variétés ;
3. Tenter de gommer l'accent étranger des étudiants afin que leur prononciation se rapproche de celles de locuteurs natifs.

L'équilibre entre ces trois éléments ne va pas de soi et s'annonce particulièrement difficile à maintenir dans un contexte où la notion d'inclusivité s'est imposée récemment parmi les valeurs fortes de notre société. Il ne fait aucun doute que s'exprimer avec un accent stigmatisé, étranger ou natif, peut constituer un obstacle dans la vie professionnelle et personnelle d'un individu (Baquiran et Nicoladis, 2019 ; Lev-Ari et Keysar, 2010 ; Hanzlíková et Skarnitzl, 2017 ; Frumkin, 2007 ; Rakić *et al.*, 2011 ; Dixon *et al.*, 2002).

63. https://casrai.org/credit/.

Il ne fait aucun doute non plus que si nous voulons continuer à évaluer la prononciation de nos étudiants, vu la multiplicité des accents et l'impossibilité pour les enseignants de tous les connaître, nous sommes contraints d'adopter un nombre restreint de variétés de référence. Doit-on seulement continuer à tenter de réduire le degré d'accent étranger et de faire correspondre la prononciation des étudiants à des normes dont le prestige est reconnu (voir par exemple Munro et Derwing, 1995 ; Diana, 2010 ; Isaacs, 2018) ? Peut-on faire dès maintenant le pari de l'inclusivité, dire à nos étudiants d'affronter le monde anglophone avec un accent non-natif assumé, et laisser à leurs interlocuteurs la charge de l'inclusion ? Plusieurs éléments, dont par exemple une recherche que nous avons menée récemment (voir Section 7.3.4), me portent à croire qu'il est trop tôt. Néanmoins, la question me paraît plus pressante qu'elle ne l'était quand j'ai commencé ma carrière. En effet, des cas-limites se présentent d'ores et déjà ; par exemple, faut-il recruter un lecteur qui s'exprime avec un accent stigmatisé et courir le risque que les étudiants influencés par son accent subissent la même discrimination que celle qui aurait consisté à ne pas recruter ce lecteur ?

### 7.2.2 *Open science, open data*

Il y a vingt ans, la question du partage dans la science était déjà d'actualité mais elle s'invite de plus en plus fréquemment dans les discussions ces dernières années. Garellek *et al.* (2020) proposent une synthèse exhaustive des enjeux du partage des données dans le domaine de la phonétique. Cette question est indissociable d'une part de la mise en exergue croissante des aspects éthiques (j'en reparle *infra* à la Section 7.2.3) dans nos pratiques, et d'autre part, des contraintes matérielles et temporelles liées au métier d'enseignant-chercheur.

Je pars du postulat que laisser ses données en libre accès représente un gain pour la communauté scientifique (j'émets cependant une réserve à la fin de cette section). Il s'agit d'une plus-value en termes de transparence et de reproductibilité, et cela garantit la pérennité des données là où, aujourd'hui, bien souvent, les données se perdent en même temps que l'investissement financier qui a permis leur collecte.

Cependant, partager ses données, c'est prêter le flanc à d'éventuelles critiques supplémentaires, se perdre dans la bureaucratie (toujours plus contraignante) qui accompagne les demandes d'autorisation liées aux aspects éthiques, passer du temps à rendre les données partageables (métadonnées, anonymisation, etc.) et, bien sûr, risquer que des laboratoires mieux dotés que le sien, en plus de jouir gratuitement de tout ce travail, décrochent l'appel

à projets auquel on a postulé… Au vu du travail supplémentaire nécessaire et des risques encourus, la route vers l'*open data* implique une véritable politique d'incitation. Garellek *et al.* (2020) mentionnent par exemple une possible obligation de joindre ses données à toute soumission d'article. Cette pratique représenterait un gain en termes de transparence, faciliterait probablement le processus de relecture, mais retarderait le moment de la soumission le temps que les données soient « mises en conformité ».

Depuis le milieu des années 2010, l'acceptation d'un article dans une revue de chez Elsevier s'accompagne de la proposition de publier les données liées à cet article dans une autre revue, *Data in Brief* ; ou encore, dans le cas d'une méthodologie ou d'une instrumentation innovante, de publier un article supplémentaire dans *MethodsX* ou *HardwareX*. Il existe donc déjà bel et bien une logique de l'incitation au partage, qui s'appuie sur la valorisation de l'effort consacré à la préparation des données et à l'innovation méthodologique. D'un côté, cela me réjouit puisque c'est la promesse d'un juste retour sur mon investissement dans les volets méthodologique et technologique de la recherche (voir Chapitre 3). D'un autre côté, si les données, la méthode et le matériel en viennent à être « rétribués » séparément comme des réalisations scientifiques en soi, je rejoins Garellek *et al.* (2020) pour poser la question de la pondération de ces réalisations par rapport à l'article scientifique auquel elles ont contribué.

La seule objection conceptuelle que je souhaiterais émettre vis-à-vis de l'*open data* réside dans le risque d'encourager la réutilisation des données pour répondre à une question qui n'a pas été posée au moment de leur collecte. Une telle éventualité favoriserait potentiellement l'inductivisme, que je dénonce tout au long de ce document, et en particulier au Chapitre 2. Et, en pratique, il me paraît légitime de faire valoir que rendre des données partageables prend du temps et que si les établissements souhaitent jouer le jeu de l'*open data*, ils doivent y consacrer des moyens concrets.

### 7.2.3 Standardisation, éthique et créativité

L'homogénéisation des pratiques semble être ce vers quoi tend la communauté scientifique. Cela passe en particulier par la réflexion omniprésente sur les aspects éthiques, concept qui recouvre un ensemble d'objets en réalité très variés. L'éthique englobe en effet une déontologie « générale » concernant la transparence et la reproductibilité des méthodes. Cet aspect-là comprend par exemple l'institutionnalisation de méthodes quantitatives, de bonnes pratiques graphiques pour représenter des données sans les distordre (Tufte, 2001), la mise en place de garde-fous méthodologiques comme le pré-enregistrement

(Kupferschmidt, 2018 ; Warren, 2018) et l'incitation à partager ses données dans une logique de science ouverte (Garellek *et al.*, 2020). L'éthique régit aussi la façon dont les scientifiques doivent se comporter avec les participants humains à leurs expériences (loi Jardé), et la façon dont les données personnelles doivent être traitées (Règlement Général sur la Protection des Données : RGPD). Je ne suis pas spécialiste de la question, et ne prétendrai pas avoir une connaissance étendue du sujet, mais puisque nous sommes confrontés à ce « dossier » au quotidien, nous avons tous notre mot à dire.

Tendre vers davantage d'éthique est louable, tant que le discernement prévaut sur l'application rigide d'un règlement ; tant que les recommandations ne se transforment pas en censure ; tant que les collègues sont « présumés éthiques » jusqu'à preuve du contraire. L'application en contexte de règles de bonne conduite est à mon avis un chantier bien plus délicat que celui de leur simple promulgation car tout cela a un coût pour les enseignants-chercheurs, non seulement en termes de contraintes pragmatiques, mais également sur le plan des libertés.

Un premier point très terre-à-terre (à défaut d'être original) que je souhaite mentionner, c'est la nécessité de disposer de moyens supplémentaires pour éviter une surcharge de travail dissuasive. Faire une demande d'agrément en vue d'une expérimentation sur la personne humaine, ou pré-enregistrer une étude pour laquelle une analyse de puissance statistique a révélé que le nombre de participants dont on a besoin est bien supérieur à celui sur lequel on avait initialement tablé, tous ces efforts dans le sens de la conformité à l'éthique génèrent un surcoût et pécuniaire, et en termes de temps de travail. Ce surcoût est vécu différemment selon qu'on est maître de conférences rattaché à ce qui s'appelait encore récemment une « équipe d'accueil » en Études Anglophones ou chercheur à plein temps dans une UMR, assisté de personnel de soutien à la recherche.

Le second point que je souhaite évoquer, c'est que l'homogénéisation grandissante des pratiques conduit par définition vers moins de diversité. Par principe, je me méfie des systèmes de pensée unique — comme celui qui fait l'apologie du test de l'hypothèse nulle en statistique et qui a conduit à tant de faux positifs dans l'histoire de la science (Cohen, 1994 ; Ziliak et McCloskey, 2008 ; Open Science Collaboration, 2015 ; Roettger, 2019, etc.) — et je préfère examiner les alternatives potentielles. Au vu de ce que j'ai écrit au Chapitre 2 et au début de cette section, la science semble aller vers davantage de raisonnement déductiviste, et moins d'inductivisme. Cette « poppérisation »— attention aux homophones ! — des pratiques devrait me réjouir. J'ai néanmoins la sensation que cette citation de Kuhn (1996, p. 166) n'a jamais été aussi vraie : « scientific training is

not well designed to produce the man who will easily discover a fresh approach ». En effet, comment favoriser la prise de risques, et l'audace des théories, que certains auteurs voient comme un gage de qualité (Serlin et Lapsley, 1990 ; Curd et Cover, 1998 ; Popper, 2002a,b), si c'est l'homogénéité qui est récompensée ? Est-il possible d'être audacieux, créatif (Heinze *et al.*, 2009), voire dissident, quand c'est le consensus qui est valorisé (Longino, 1990 ; Oreskes, 2019) ?

Une possibilité pour contrer cette tendance au conformisme consisterait peut-être à distinguer explicitement les études « observationnelles » des études expérimentales (van Belle, 2002), à revaloriser les premières, et à s'astreindre à n'employer l'inférence statistique que dans le second cas. Roettger (2019) et Roettger *et al.* (2019) suggèrent que les études soient a priori clairement identifiées comme soit exploratoires, soit confirmatoires. Je trouve cette dichotomie (qui recouvre en grande partie la précédente) tout à fait opportune— à ceci près que je préfèrerais « falsificatoire » à « confirmatoire » — car elle reprend ce que je pense être deux étapes différentes du processus de recherche, l'exploration précédant la tentative de falsification. La publication scientifique de haut niveau étant biaisée en défaveur des études exploratoires, cela contraint parfois les chercheurs à remanier leurs résultats exploratoires pour qu'ils aient l'air plus confirmatoires (Roettger *et al.*, 2019 ; Roettger, 2019). Une réhabilitation des études explicitement déclarées comme exploratoires — et qui, logiquement, n'auraient pas recours à l'inférence statistique mais à des techniques quantitatives appropriées (e.g. *clustering*, visualisation d'*embeddings*, etc.) — constitue une perspective intéressante. En effet, cela permettrait de retrouver un peu de spontanéité dans un contexte où prime l'anticipation, engendrée par la logique des appels à projets, les demandes d'autorisations qui se multiplient dans un système toujours plus bureaucratique, etc. En nous libérant de la contrainte d'obtenir rapidement des *p-values* significatives, des études exploratoires reconnues en tant que telles et publiées dans des revues prestigieuses limiteraient certainement le *p-hacking* et la crise de reproductibilité, et autoriseraient probablement davantage d'audace et de créativité.

J'aborde à présent un dernier point, où, comme pour les deux précédents, je me contenterai d'amorcer la réflexion. En 2002, j'ai donné mon premier cours à l'université ; il s'agissait d'un enseignement optionnel d'introduction à la linguistique. Claude Boisson m'avait donné carte blanche, et j'avais construit ce cours autour des dichotomies emblématiques de notre science : phonologie-phonétique, compétence-performance, langue-parole, *top-down-bottom-up*, etc. Au moment de présenter l'opposition « prescriptivisme-descriptivisme », un vent de liberté semblait traverser les préfabriqués du campus de Bron : oui, nous autres,

linguistes, nous avions pour particularité de nous affranchir du prescriptivisme des grammaires de notre enfance, et nous nous contentions de décrire la langue telle qu'elle était parlée. L'authenticité est d'ailleurs un argument phare justifiant la linguistique de corpus.

Le linguiste au sens large (incluant les spécialistes de traitement automatique) n'a-t-il donc pas de rôle normatif à jouer ? Pour reprendre un exemple entendu récemment dans une conférence (Hovy, 2020), en tapant « why are american » dans Google, l'algorithme de saisie semi-automatique proposait : « why are american [sic] so fat ». Cette suggestion est descriptivement correcte d'une part parce que parmi les pays de l'OCDE, c'est bien les États-Unis qui présentent le taux d'obésité le plus élevé[64], et d'autre part, parce que l'algorithme de suggestion de Google s'appuie notamment sur les recherches des autres internautes, et propose tout naturellement de nous faire gagner du temps en suggérant les recherches les plus populaires. La question se pose de savoir s'il est éthiquement souhaitable que cette particularité — l'obésité chez les Américains — soit mise en avant. Je constate incidemment que Google nous laisse la possibilité de « signaler des prédictions inappropriées » ; ne parvenant pas à reproduire moi-même cet exemple, je suppose qu'il a été signalé comme tel depuis. Je n'ai bien évidemment pas de réponse tranchée à apporter, mais, pour reprendre une question posée dans la salle lors d'un colloque consacré à l'intelligence artificielle : qui décide que c'est inapproprié ?

Dans le cas des suggestions de Google, comme dans celui des Afro-americains mal compris par les systèmes de reconnaissance de la parole d'Amazon ou d'IBM (Section 7.1.2), et comme pour la démarche qui consiste à réduire l'accent étranger de nos apprenants francophones pour le faire tendre vers l'une ou l'autre des deux variétés « autorisées » (fin de la Section 7.2.1), il me semble que le rôle des linguistes ne peut pas se limiter à rester strictement descriptifs. Avec une conscience sociolinguistique et du discernement, les linguistes et phonéticiens ont leur mot à dire dans les débats éthiques en lien avec leur discipline.

## 7.3 Projets actuels

À l'occasion de la rédaction de ce document de synthèse, c'est avec beaucoup d'enthousiasme que je réalise que mes projets actuels correspondent de très près à des envies scientifiques qui m'ont accompagné dès mes premiers pas dans la recherche. Si une légère coloration SHS se maintient depuis le début (la dimension sociophonétique), je dois cepen-

64. https://www.oecd.org/health/obesity-update.htm.

dant reconnaître que j'ai pris il y a vingt ans un tournant très quantitatif dans le but de fuir, pour ainsi dire, certaines facettes de ma formation initiale qui, quoique intéressantes, manquaient d'ancrage scientifique. Après toutes ces années à traiter automatiquement des données variées et complexes, à implémenter des méthodes de calculs parfois sophistiquées et, plus récemment, à défricher pour ma discipline les potentiels apports du *deep learning*, j'ai depuis peu inscrit à mon agenda quelques projets aux tonalités plus SHS, comme si ces vingt années m'avaient immunisé contre une possible suspicion de pratiquer une « science molle ».

### 7.3.1 Identité et voix chantée, le cas du Heavy Metal britannique

*Fear of the Dark*, *Heaven and Hell*, *Electric Eye* ou encore *Photograph* sont autant de titres de chansons de Metal britannique que je fredonne depuis plus de 30 ans. Et cela fait de nombreuses années que je caresse l'idée d'étudier les aspects phonétiques et phonologiques de ces chansons, inspiré par l'étude de Trudgill (1983) sur la prononciation des Beatles. Car en effet le Heavy Metal original est né du contexte socio-économique de l'Angleterre des années 1970. Et à ce titre, il présente une identité britannique forte ; est-elle cependant reflétée dans l'accent des chanteurs en dépit d'une tendance inhérente à la musique commerciale à américaniser la prononciation ?

Comme souvent dans un projet de recherche, c'est la rencontre d'une étudiante capable de porter une thématique aussi innovante qui m'a conduit à franchir le pas. Coline Caillol s'est depuis approprié le sujet, à mi-chemin entre sociologie, phonétique et voix chantée. Nos premiers travaux (Caillol et Ferragne, 2019), qui caractérisent notamment la variation du T Voicing, inaugurent une voie nouvelle dans les *metal studies* tout en perpétuant l'approche sociophonétique de l'anglais qui a servi de cadre à mes travaux de thèse.

### 7.3.2 Entraînements à la prononciation de l'anglais

Dans la Section 5.2.2, je fais référence aux travaux de thèse de Jennifer Krzonowski, pour lesquels des interfaces pour l'entraînement à la prononciation (voir par ex. celle présentée à la page 52) ont été développées. Comme je l'ai déjà mentionné, ces interfaces sont disponibles sur Github[65], et elles étaient probablement destinées à y rester telles quelles pour longtemps.

65. https://tinyurl.com/ys2urfu6.

Les circonstances en ont décidé autrement : c'est d'une part un appel à projets pédagogiques de l'IDEX Université de Paris et d'autre part la situation sanitaire nous contraignant au distanciel qui m'ont poussé à soumettre le projet *Solutions pour l'Enseignement de la Phonétique Appliquée aux Langues Étrangères* (SEPALE), qui a finalement été retenu. Après avoir procédé à des modifications mineures, Anne Guyot-Talbot, Sylvain Navarro et moi-même avons proposé ces interfaces dans le cadre du cours d'Oral à tous les étudiants de notre UFR.

Il s'agit là d'un projet avant tout pédagogique, mais qui permet néanmoins d'établir un lien solide entre enseignement et recherche. La suite du projet prévoit de solliciter les retours des étudiants dans le but d'améliorer l'ergonomie des interfaces, d'analyser les fichiers .log[66] pour mieux comprendre les habitudes de nos étudiants, et d'étendre nos exercices en proposant d'autres tâches et en incluant l'anglais américain.

La recherche sur l'acquisition de la phonologie de l'anglais par des apprenants francophones reste bien sûr presque naturellement à l'ordre du jour pour moi, qu'elle touche à des aspects pédagogiques très concrets, comme dans le projet SEPALE, ou à des questions de recherche plus fondamentale comme dans une de mes publications récentes (Heidlmayr *et al.*, 2021).

### 7.3.3 Représentation à l'écran des technologies de la parole

Mon intérêt pour l'influence des séries télévisées mettant en scène des experts en criminalistique sur notre perception de la science remonte au début des années 2000, à l'époque où j'avais un temps envisagé une thèse sur la comparaison de voix. Comme je l'explique dans l'Introduction, mon implication dans cette thématique est allée croissant parallèlement à mes dix années de mandats au conseil d'administration de l'AFCP. Après mon implication dans le projet ANR VoxCrim, j'ai trouvé opportun de proposer en quelque sorte le pendant SHS de ce dernier.

Le projet VoCSI-Telly (*Voice and Crime Scene Investigators on Telly*) vise donc à identifier la manière dont les séries télévisées embellissent le quotidien des scientifiques de la police au point de nous offrir la vision déformée d'une science totalement infaillible. Nous sommes en train de constituer une base de données d'extraits vidéos qui nous permettra d'élaborer des éléments de critique corrective en décortiquant les processus esthétiques et narratifs qui concourent à « glamouriser » la criminalistique.

66. Le détail de chaque session est enregistré automatiquement dans des fichiers « journaux » que les étudiants nous communiquent à la fin du semestre.

L'aspect ludique de ce projet ne doit cependant pas masquer le véritable enjeu sociétal qu'il comporte. Pour ne prendre qu'un seul exemple, l'étude de Call *et al.* (2013), qui s'appuie sur un questionnaire distribué à 60 jurés ayant siégé dans des affaires de coups et blessures volontaires, rapporte que 95 % d'entre eux regardaient CSI[67] et 73 % ont considéré que la série avait influencé leur verdict.

### 7.3.4 Un accent de culpabilité

Les stéréotypes liés aux accents suscitent depuis longtemps un grand intérêt chez moi. La culpabilité perçue d'un individu est modulée à la hausse si celui-ci s'exprime avec un accent (natif) stigmatisé (Dixon *et al.*, 2002). Et les opportunités de carrière de cet individu souffrent également de son accent (Rakić *et al.*, 2011). Par ailleurs, les locuteurs présentant un accent étranger sont globalement jugés comme moins fiables (Lev-Ari et Keysar, 2010), y compris s'ils sont témoins dans des procès (Frumkin, 2007) ou médecins (Baquiran et Nicoladis, 2019).

Quelques études utilisant les potentiels évoqués en électroencéphalographie ont mis en évidence des corrélats fiables de violations de stéréotypes liés au locuteur. Lattner et Friederici (2003) ont observé une onde P600 dans des violations de genre, par exemple lorsqu'un locuteur masculin déclare aimer porter du rouge à lèvres. Dans l'étude de van Berkum *et al.* (2008), c'est la violation de stéréotypes d'âge, de genre, et d'accent, qui a donné lieu à un effet N400 ; le même, quoique présentant une amplitude moindre, que celui qu'on observe en cas d'incongruence lexico-sémantique.

Dans le cadre du post-doctorat de Maud Pélissier au sein du projet ANR VoxCrim, nous avons répliqué cet effet N400 dans une étude sur les stéréotypes d'accent générés par la voix des locuteurs en français[68].

La question de la discrimination induite par l'accent a pris un tournant plus politique ces dernières années, avec l'apparition de termes comme « accentisme » ou « glottophobie » et, en France, l'adoption en novembre dernier de la « Proposition de loi visant à promouvoir la France des accents et à lutter contre les discriminations fondées sur l'accent ». Contribuer modestement à ce débat sociétal en employant des méthodes scientifiques éprouvées (comme nous l'avons fait dans Pélissier et Ferragne, 2021) constitue une forte motivation pour mes projets à venir.

67. En français : *Les Experts*.
68. Exemples de violations de stéréotypes *Tu préfères qu'on passe chez Ladurée pour les macarons ?* dit avec un accent de banlieue ; *Je passe des heures dans le hall de l'immeuble* dit avec un accent « bourgeois ».

Ce bref exposé de mes projets en cours offre un tour d'horizon de mes perspectives de recherche à moyen terme. Je ne crois pas déceler de changement de cap notoire si on omet une coloration SHS et des enjeux sociétaux peut-être plus visibles que par le passé. Pour la suite, j'espère continuer de développer la plupart des thèmes mentionnés dans ce document. Il faudra trouver le juste équilibre entre la tentation légitime de capitaliser sur un nombre restreint de « succès » récents — en termes de publications ou de projets — et continuer d'offrir aux étudiants la diversité, que je crois fertile, qui caractérise mon encadrement. Car en tant que chercheur *et* enseignant, la trajectoire professionnelle des étudiants est indissociable de mes perspectives de recherche. J'ai eu la chance d'encadrer des personnes aux profils très variés (comme le montre la Figure 7.1), et conserver cette variété dans la suite de ma carrière fait partie de mes priorités.

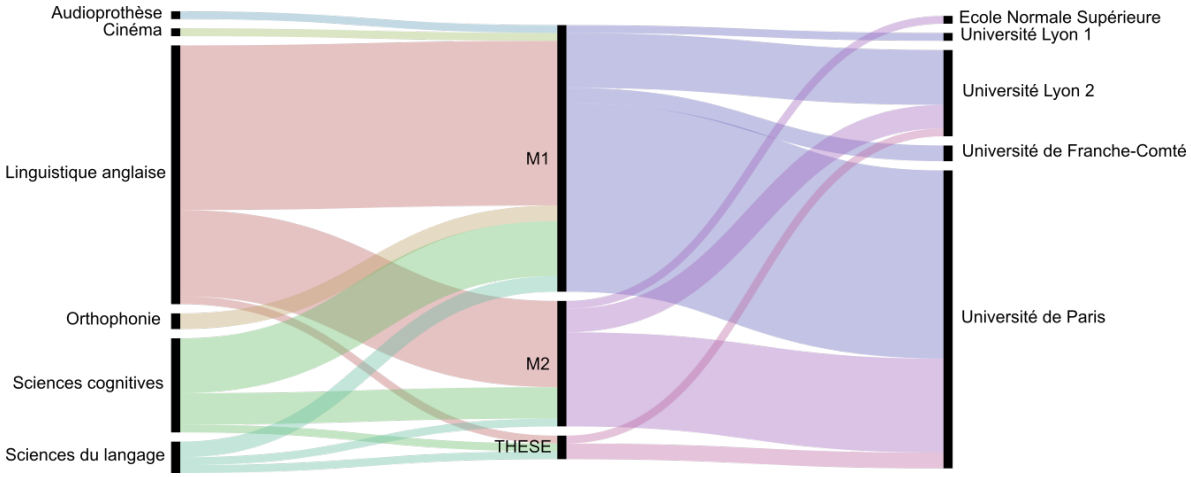



FIGURE 7.1 – Vue synoptique de mes encadrements par discipline, niveau d'études et établissement.

## 7.4 Conclusion

Dans ce document de synthèse, j'aurais pu me contenter de l'étude d'un objet précis. Les candidats qui auraient largement occupé 200 pages ne manquaient pourtant pas : contrastes dérivées dans les variétés d'anglais, acquisition de la phonologie de l'anglais par les francophones, caractéristiques phonétiques du locuteur, *deep learning* pour la recherche fondamentale en phonétique, etc. La curiosité étant plus forte que le conformisme dans la façon dont ma carrière se déroule, je crois qu'il fallait que ce document reflète cette réalité. Une grande diversité n'implique cependant pas une absence de cohérence. En effet, par exemple, la dimension sociale — ou sociétale comme on dit parfois — est omniprésente.

On la retrouve dans mes travaux sur les variétés d'anglais, dans le développement de techniques d'amélioration de la prononciation, ou encore dans la mise en évidence des stéréotypes liés à la voix par des approches expérimentales. Et puis, la diversité des approches, comme j'espère l'avoir démontré, c'est précisément ma singularité. J'ai certes des préférences méthodologiques, mais je suis néanmoins engagé dans un processus d'auto-formation constante avec pour but d'aborder les nouvelles questions de la manière la plus holistique et agnostique qui soit. Cette diversité, en particulier des méthodes que j'utilise, est un point fort indéniable dans la réalisation de la partie la plus gratifiante de mon métier : former des étudiants à la recherche et les voir développer leur propre identité scientifique.

# A

# Cartes SNR d'activation dans Ferragne *et al.* (2019)

Voici les cartes d'activation SNR pour les 45 locuteurs étudiés dans Ferragne *et al.* (2019). Les locuteurs sont regroupés par groupes de 15.

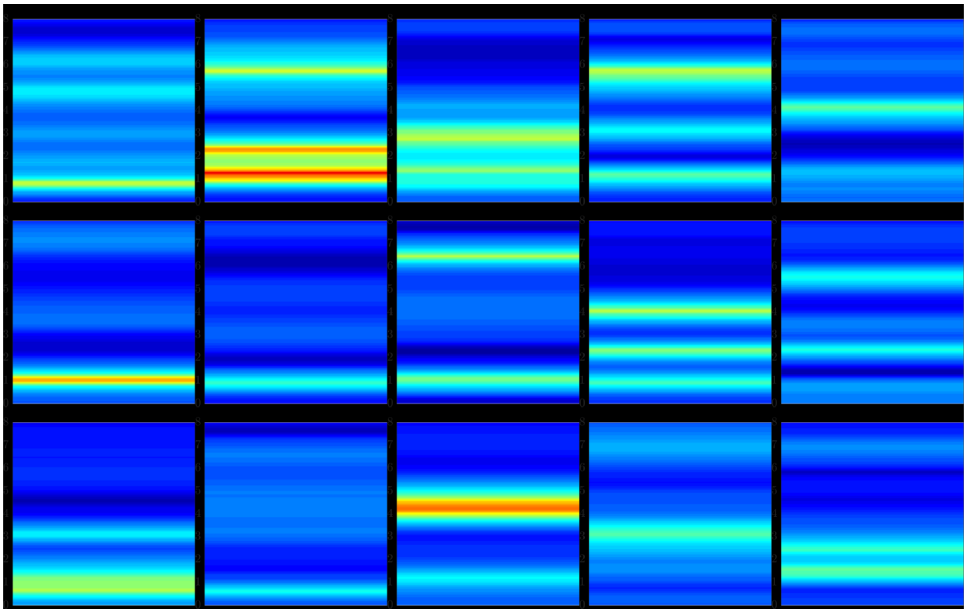


FIGURE A.1 – Cartes SNR pour les locuteurs 01 à 15.

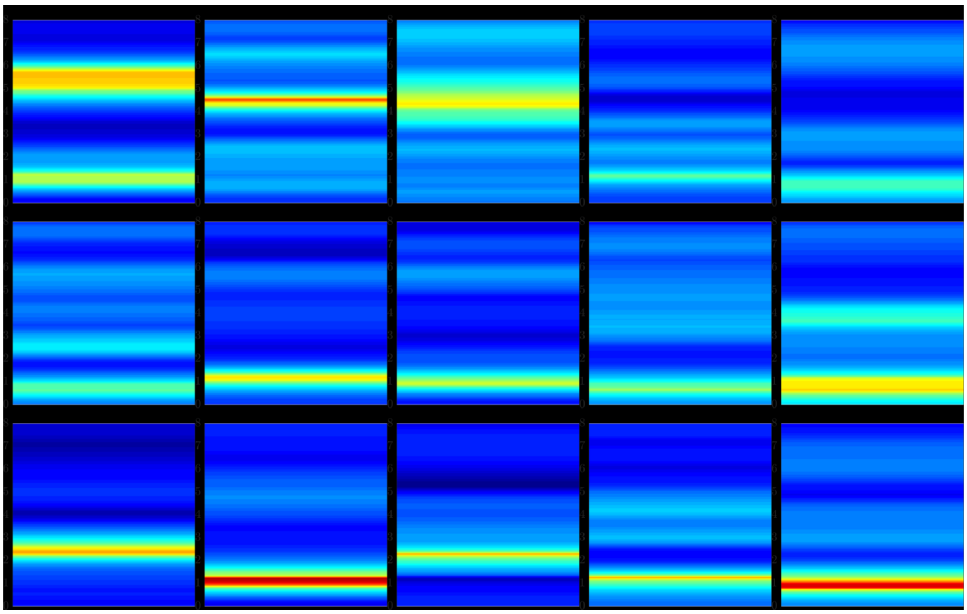


FIGURE A.2 – Cartes SNR pour les locuteurs 16 à 30.

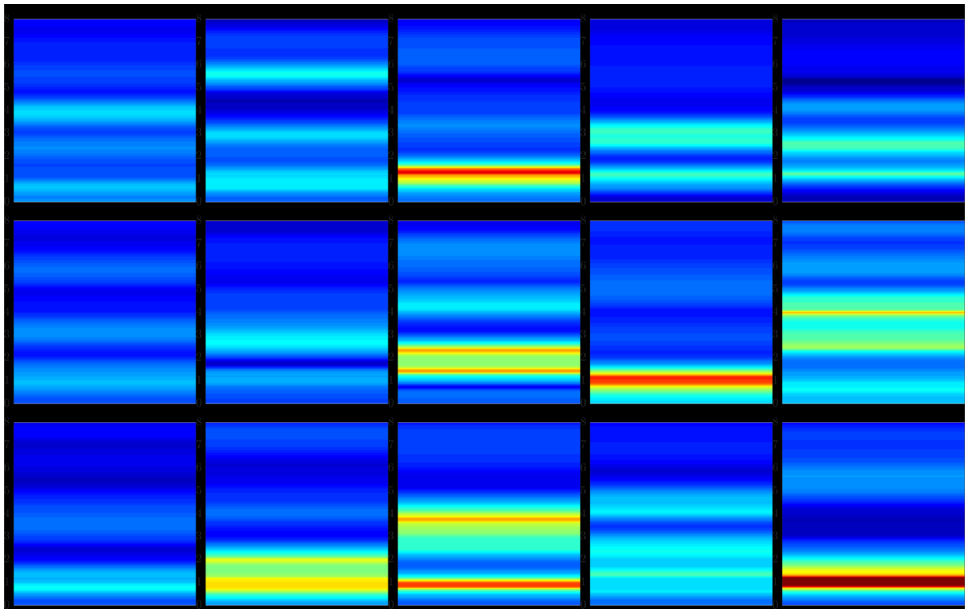


FIGURE A.3 – Cartes SNR pour les locuteurs 31 à 45.

# B

# Fonctionnement détaillé du Lab Monitor

Cette annexe décrit le prototype du Lab Monitor, un système de capteurs s'appuyant sur l'Internet des Objets (*Internet of Things*, IoT). J'ai conçu ce système pour optimiser l'utilisation d'une salle que je mets à disposition des étudiants pour la recherche en phonétique. L'intérêt de cette annexe n'est pas tant de comprendre le fonctionnement détaillé de ce système que de mesurer mon investissement personnel dans des aspects très techniques. J'ai choisi l'exemple du Lab Monitor car l'idée du système est assez singulière et aboutie.

Le système s'appuie sur un Raspberry Pi 3 B, un nano ordinateur low cost tournant sous Linux, auquel sont connectés un module Sense Hat et un détecteur de mouvement à infrarouge. Le prototype est présentée dans la Figure B.1.

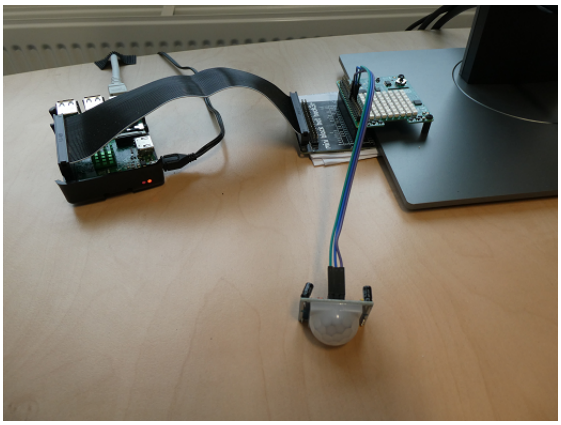


FIGURE B.1 – Prototype du Lab Monitor.

L'espace est composé de trois postes de travail. Le premier, polyvalent, est particulièrement adapté au calcul scientifique puisqu'il dispose d'une configuration puissante et d'une licence de `Matlab` très complète. Le second est dédié à l'électroglottographie et le troisième, à l'échographie. Mon premier objectif était donc de savoir quand ces postes étaient utilisés. C'est le logiciel `Nmap` qui, depuis le Raspberry Pi, effectue un balayage des adresses IP de ces trois machines afin de déterminer si elles sont connectées (et donc, démarrées). En fonction des options choisies, le logiciel peut renvoyer de faux positifs ; j'ai déterminé empiriquement qu'une recherche de l'adresse Mac de la machine concernée dans la sortie de la fonction `nmap` présentait des résultats relativement fiables. En résumé, un script shell tourne donc en arrière-plan sur le Raspberry Pi, lance la fonction `nmap` toutes les deux minutes, et écrit dans un fichier une valeur booléenne pour chaque machine, indiquant le statut de cette dernière : on ou off.

Puisque la fonction de cet espace ne se résume pas à l'utilisation des ordinateurs qui s'y trouvent, deux indicateurs supplémentaires complètent le dispositif. D'abord, un détecteur

de mouvements, qui, comme la plupart des capteurs de ce genre, analyse la température (et donc le rayonnement infrarouge) dans sa zone de détection pour conclure, en cas de variation abrupte, à une présence. Ensuite, l'utilisateur peut, par exemple dans le cas de passation d'une expérience pilote, signaler qu'il serait préférable de ne pas venir perturber l'expérience en appuyant sur le joystick du module Sense Hat.

Comme dans la plupart des projets IoT, le dispositif inclut également la collecte de données sur l'environnement ; en l'occurrence, la température et le taux d'humidité ambiants (capteurs embarqués sur le Sense Hat). La gestion du système est confiée à un programme compilé (sur le Raspberry Pi) que j'ai déployé à partir d'un modèle `Simulink`[69] élaboré pour l'occasion.

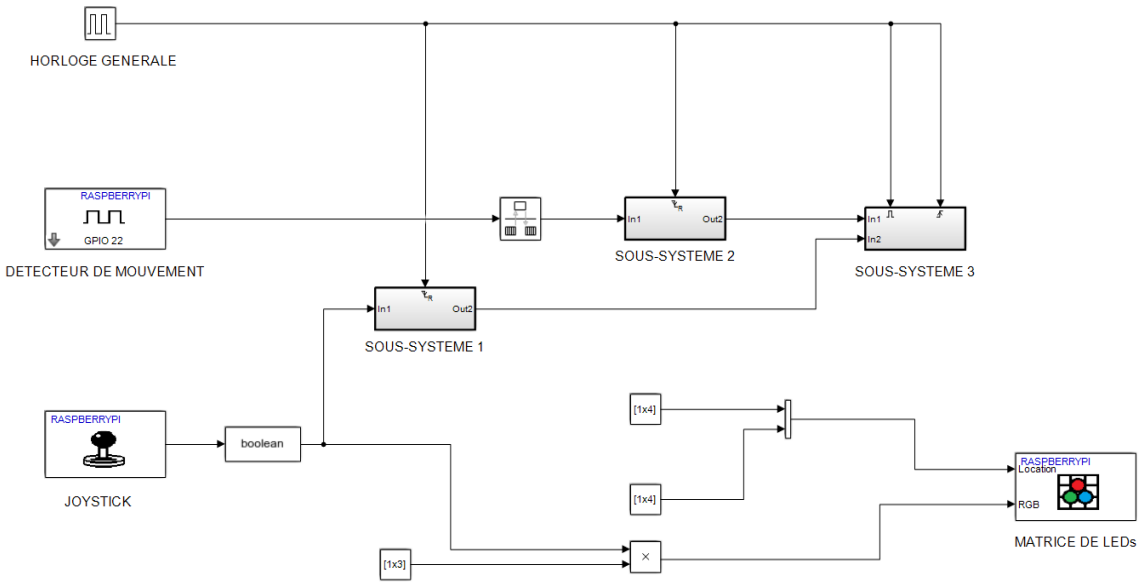



FIGURE B.2 – Vue synoptique du modèle Lab Monitor.

La Figure B.2 offre une vue synoptique du modèle. L'horloge générale en haut à gauche détermine la cadence de l'intégralité du modèle. Il s'agit d'un signal logique qui comporte un état haut toutes les 5 minutes. Cet état haut déclenche un événement dans les sous-systèmes 1, 2 et 3. L'événement le plus remarquable (dans le sous-système 3) consiste en l'envoi des données à la plateforme IoT Thingspeak. Au-dessous de l'horloge se trouve la sortie du détecteur de présence connectée à la broche GPIO 22 du Raspberry Pi. La sortie logique de cette broche est lue à une fréquence de 5 Hz et envoyée au sous-système 2.

69. Simulink est un logiciel édité par The Mathworks, qui permet de programmer par blocs.

Au-dessous, le joystick du Sense Hat est lui aussi lu toutes les 200 ms. Sa valeur, 0 s'il n'est pas actionné, ou 1, 2, 3, 4 ou 5 en fonction de l'action spécifique (à droite, à gauche, etc.) est convertie en valeur booléenne car c'est la simple détection d'une action qui nous intéresse ici. Cette valeur logique est envoyée simultanément au sous-système 1 et à la commande de la matrice de LEDs. La matrice de LEDs comporte deux entrées étiquetées Location et RGB. L'entrée Location prend comme argument constant les coordonnées des 4 LEDs qui forment les coins de la matrice de 64 ($8 \times 8$) diodes du Sense Hat. Ces coordonnées sont données par la concaténation des abscisses et des ordonnées inscrites dans les blocs où l'on voit figurer l'étiquette $1 \times 4$. L'entrée RGB reçoit l'information des 3 constantes de rouge, vert et bleu qui déterminent la couleur des LEDs (en l'occurrence, 0, 255, 255 pour cyan). La sortie convertie en booléen du joystick vient multiplier ces 3 constantes. En d'autres termes, sans action sur le joystick, les valeurs RGB sont multipliées par zero et les LEDs ne s'éclairent pas. Lorsque le joystick est actionné, les valeurs RGB sont multipliées par 1 et les LEDs des 4 coins de la matrice s'illuminent, signifiant ainsi à l'utilisateur que son action a été prise en compte.

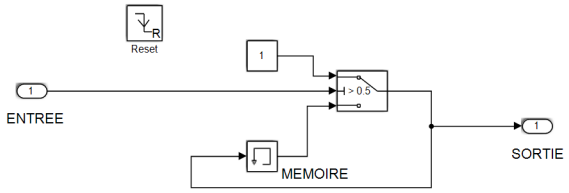



FIGURE B.3 – Détail du sous-système 1.

La Figure B.3 montre le fonctionnement du sous-système 1 (identique au 2). Ces deux sous-systèmes sont connectés aux sorties du détecteur de mouvement et du joystick. L'entrée, c'est-à-dire le signal logique en provenance de ces deux capteurs, contrôle un interrupteur. Tant que la valeur de contrôle reste à zéro, la sortie affiche zéro. Si le signal de contrôle passe à 1, l'interrupteur laisse alors passer la constante pré-définie dans l'entrée la plus haute dans le schéma (c'est-à-dire 1) et la boucle comportant un bloc mémoire renvoie la valeur 1 en sortie jusqu'à ce que le sous-système soit ré-initialisé (bloc Reset) par l'horloge générale. Autrement dit, si une présence est détectée ou si le joystick est actionné ponctuellement dans les 5 minutes qui précèdent l'envoi de données, l'événement reste en mémoire et il apparaîtra dans Thingspeak.

Pour finir, le sous-système 3 est représenté à la Figure B.4. Tout à gauche, on remarque les entrées 1 et 2 qui correspondent au signal du détecteur de présence et à celui du

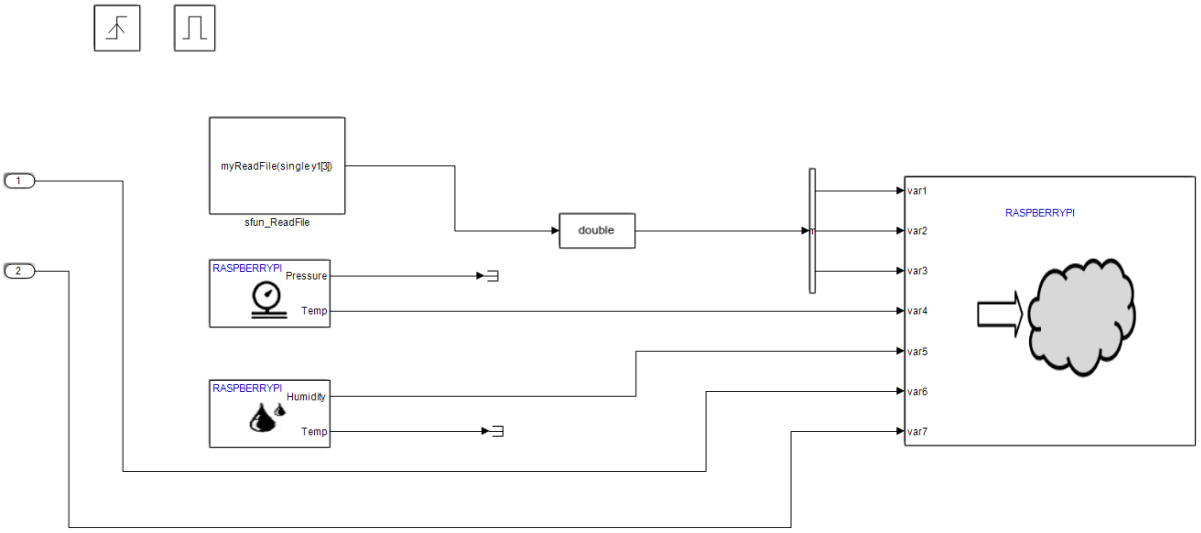



FIGURE B.4 – Détail du sous-système 3.

joystick (une fois passés par les sous-systèmes 1 et 2). Le bloc intitulé myReadFile permet de lire le fichier contenant l'état (on/off) des 3 ordinateurs de l'espace de travail. Les deux blocs au-dessous mesurent respectivement la température et l'humidité. Toutes ces variables sont connectées au bloc de droite, matérialisé par un nuage, qui permet d'écrire les résultats sur la chaîne Thingspeak créée pour l'occasion[70]. L'exécution de ce sous-système est déclenchée, comme mentionné plus haut, par le passage de l'horloge générale à un état haut toutes les 5 minutes.

J'ai par ailleurs développé une interface graphique qui permet d'analyser les données recueillies par le Lab Monitor ; j'en présente une capture d'écran à la Figure B.5.

Pour compléter le dispositif, j'ai réfléchi à une solution pour pouvoir échanger des fichiers entre les 3 postes de travail (voire avec d'autres machines sur le même réseau). J'ai donc créé mon propre système, qui est certes rudimentaire, mais qui permet de remplir exactement la fonction que j'avais à l'esprit. Le même Raspberry Pi a été mis à contribution pour cette partie du projet. J'ai créé un répertoire pour le stockage permanent de scripts et programmes à usage collectif et un pour l'échange ponctuel de fichiers (détruits automatiquement après un certain temps). J'ai écrit une interface en `C#`, baptisée PC2PI, qui s'appuie notamment sur la bibliothèque SSH.NET, et implémente la communication avec le protocole SFTP entre les machines équipées de PC2PI et le Raspberry Pi. Par choix, les informations de connexion sont codées en dur, ce qui rend les opérations de

70. La chaîne existe depuis le 19 février 2018 et est en accès public ici : https://thingspeak.com/channels/430548.

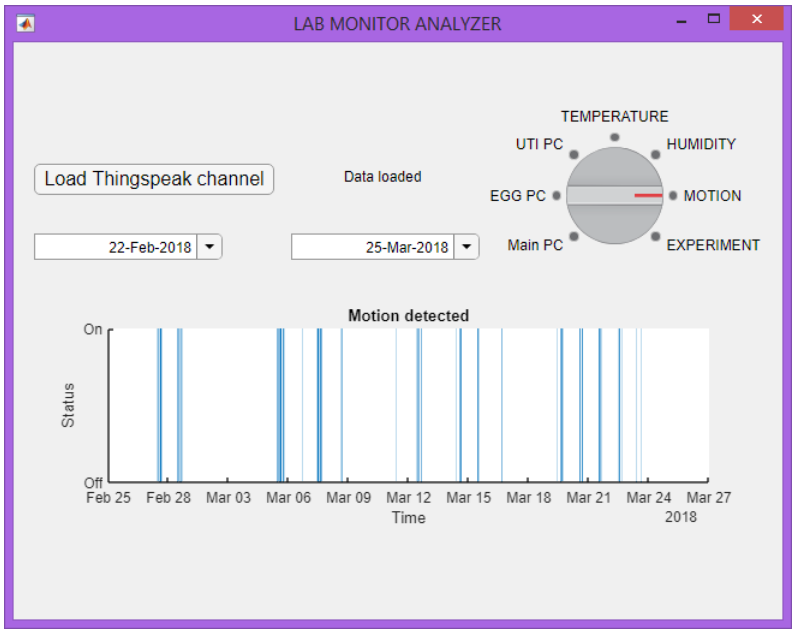



FIGURE B.5 – Interface pour l'analyse des données du Lab Monitor.

transfert très ergonomiques et instantanées, il suffit de choisir un fichier sur la machine source (serveur ou client) et de le transférer vers la machine cible (serveur ou client) en un clic.

En résumé, sur le plan technique, on peut véritablement parler d'un succès en termes de downsizing. Cette réalisation prouve en effet qu'avec un nano ordinateur à 35 euros et quelques capteurs, il est possible de mettre en place un laboratoire connecté et intelligent qui, outre la collecte et l'envoi de données dans l'Internet des Objects, se charge du stockage local de fichiers. Et il convient de noter qu'à ce stade nous sommes encore loin d'avoir exploité toutes les possibilités du dispositif. En effet, de nombreuses broches restent disponibles pour connecter non seulement d'autres capteurs, mais également pour programmer des actions. On pourrait par exemple imaginer des extensions domotiques, comme la commande d'un relai qui, si une présence est détectée et que la valeur de luminosité mesurée par un luxmètre est inférieure à un certain seuil, déclenche l'éclairage. Le potentiel est très vaste, mais il faut savoir rester raisonnable et se recentrer sur l'objectif initial : comprendre l'utilisation qui est faite de ce laboratoire pour pouvoir optimiser son fonctionnement.

En plus de la réalisation elle-même, ce projet m'a permis de travailler sous Linux pour la première fois, d'explorer le potentiel du Raspberry Pi et de l'électronique embarquée, de perfectionner mes connaissances en programmation procédurale (shell) et orientée-objet (`C#`), de découvrir la programmation par bloc et son potentiel pédagogique (`Simulink`),

et d'utiliser des outils liés à un domaine dans lequel j'étais particulièrement ignorant : les réseaux informatiques (communication SSH, SFTP, scanner de ports, etc.). Les retombées secondaires constituent peut-être l'intérêt majeur de cette démarche car, en effet, il est certain que ces nouvelles compétences vont rapidement être mises au service de la recherche et des étudiants.

# C

# Tableau de correspondance des URL réduites

| URL réduite | URL originale |
|---|---|
| https://tinyurl.com/ys2urfu6 | https://github.com/emmanuelferragne/Entrainements_Phono |
| https://tinyurl.com/nsupwav4 | https://www.youtube.com/watch?v=cJJtYiWifbY |
| https://tinyurl.com/ruzpjt69 | https://www.youtube.com/watch?v=LoCSa6Hx0VE |
| https://tinyurl.com/snvz387u | https://github.com/emmanuelferragne/CminR-Praatik |
| https://tinyurl.com/2jk8y3za | https://github.com/emmanuelferragne/ETVOYLA |
| https://tinyurl.com/yx4mf4ys | https://www.emmanuelferragne.com/post/covid-data/ |
| https://tinyurl.com/xbxxkxbk | https://www.emmanuelferragne.com/post/formant-dynamics/ |
| https://tinyurl.com/4nxzrepp | https://github.com/emmanuelferragne/dynamicFormantDataset |

# Index

# Bibliographie


Aaltonen, O., Niemi, P., Nyrke, T. et Tuhkanen, M. (1987). Event-related brain potentials and the perception of a phonetic continuum. *Biological Psychology*, 24(3):197–207.

Adda-Decker, M., Fougeron, C., Gendrot, C., Delais-Roussarie, E. et Lamel, L. (2012). La liaison dans la parole spontanée familière : une étude sur grand corpus. *Revue Française de Linguistique Appliquée*, XVII(1):113–128.

Al-Tamimi, J. et Ferragne, E. (2020). The phonetic basis of the guttural natural class in Levantine Arabic: Evidence from coarticulation and energy components using deep learning and random forests [communication orale]. *In LabPhon*, Vancouver.

Altman, Y. M. (2015). *Accelerating MATLAB® Performance: 1001 Tips to Speed up MATLAB Programs*. CRC Press, Boca Raton.

Amrhein, V., Greenland, S. et McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, 567(7748):305–307.

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge University Press, Cambridge.

Baillargeon, G. (1982). *Introduction à l'Inférence Statistique : Méthodes d'Échantillonnage, Estimation, Tests d'Hypothèses, Corrélation Linéaire, Droite de Régression et Test du Khi-Deux avec Applications Diverses*. Éditions SMG, Trois-Rivières, Québec.

Ballier, N. (2016). Du dictionnaire lexico-phonétisé aux corpus oraux, quelques problèmes épistémologiques pour l'école de Guierre. *Histoire Épistémologie Langage*, 38(2):23–40.

Baquiran, C. L. C. et Nicoladis, E. (2019). A doctor's foreign accent affects perceptions of competence. *Health Communication*, 35(6):726–730.

Bedoin, N., Abadie, R., Krzonowski, J., Ferragne, E. et Marcastel, A. (2019). A combined forced-attention dichotic listening – Go/Nogo task to assess response inhibition and interference suppression: An auditory event-related potential investigation. *Neuropsychology*, 33(8):1136–1150.

Bedoin, N., Ferragne, E., Lopez, C., Herbillon, V., de Bellescize, J. et des Portes, V. (2011). Atypical hemispheric asymmetries for the processing of phonological features in children with rolandic epilepsy. *Epilepsy & Behavior*, 21(1):42–51.

Bedoin, N., Ferragne, E. et Marsico, E. (2010). Hemispheric asymmetries depend on the phonetic feature: A dichotic study of place of articulation and voicing in French stops. *Brain and Language*, 115(2):133–140.

Bedoin, N., Krzonowski, J. et Ferragne, E. (2013). How voicing, place and manner of articulation differently modulate event-related potentials associated with response inhibition. *In Interspeech*, 906–910, Lyon.

Bloomfield, L. (1926). A set of postulates for the science of language. *Language*, 2(3):153–164.

Boë, L.-J. et Vilain, C.-E., dir. (2010). *Un Siècle de Phonétique Expérimentale, Fondation et Éléments de Développement : Hommage à Théodore Rosset et John Ohala*. ENS éditions, Lyon.

Bogliotti, C., Serniclaes, W., Messaoud-Galusi, S. et Sprenger-Charolles, L. (2008). Discrimination of speech sounds by children with dyslexia: Comparisons with chronological age and reading level controls. *Journal of Experimental Child Psychology*, 101(2):137–155.

Bonastre, J.-F. (2020). 1990-2020 : Retours sur 30 ans d'échanges autour de l'identification de voix en milieu judiciaire. *In* Adda, G., Amblard, M. et Fort, K., dir., *2e Atelier Éthique et TRaitemeNt Automatique des Langues (ETeRNAL)*, 38–47, Nancy.

Boulenger, V., Ferragne, E., Bedoin, N. et Pellegrino, F. (2011). Derived contrasts in Scottish English: An EEG study. *In ICPhS*, 352–355, Hong Kong.

Boulenger, V., Hoen, M., Ferragne, E., Pellegrino, F. et Meunier, F. (2010). Real-time lexical competitions during speech-in-speech comprehension. *Speech Communication*, 52(3):246–253.

Bybee, J. (2010). *Language, Usage and Cognition*. Cambridge University Press, Cambridge.

Caillol, C. et Ferragne, E. (2019). The sociophonetics of British Heavy Metal music: T Voicing and the FOOT-STRUT split. *In ICPhS*, 2650–2654, Melbourne.

Call, C., Cook, K., A., Reitzel, D., J. et McDougle, D., R. (2013). Seeing is believing: The CSI effect among jurors in malicious wounding cases. *Journal of Social, Behavioral, and Health Sciences*, 7(1):52–66.

Chattopadhay, A., Sarkar, A., Howlader, P. et Balasubramanian, V. N. (2018). Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. *In IEEE Winter Conference on Applications of Computer Vision (WACV)*, 839–847, Lake Tahoe.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. et Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. *In* Ferrari, V., Hebert, M., Sminchisescu, C. et Weiss, Y., dir., *Computer Vision – ECCV*, volume 11211, 833–851. Springer, Cham.

Chládková, K., Hamann, S., Williams, D. et Hellmuth, S. (2017). F2 slope as a perceptual cue for the front–back contrast in standard southern British English. *Language and Speech*, 60(3):377–398.

Chollet, F. (2018). *Deep Learning with Python*. Manning, Shelter Island.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Erlbaum Associates, Hillsdale, 2e édition.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, 45(12):1304–1312.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49(12):997–1003.

Cole, J. et Hasegawa-Johnson, M. (2012). Corpus phonology with speech resources. *In* Cohn, A. C., Fougeron, C. et Huffman, M. K., dir., *The Oxford Handbook of Laboratory Phonology*, 431–440. Oxford University Press, Oxford.

Collins, B. et Mees, I. M. (1999). *The Real Professor Higgins. The Life and Career of Daniel Jones*. De Gruyter, Berlin.

Curd, M. et Cover, J. A., dir. (1998). *Philosophy of Science: The Central Issues*. Norton, New York.

Dattalo, P. (2008). *Determining Sample Size: Balancing Power, Precision, and Practicality*. Oxford University Press, Oxford.

de Carvalho, J. B., Nguyen, N. et Wauquier-Gravelines, S. (2010). *Comprendre la Phonologie*. Presses Universitaires de France, Paris.

Delhoume, A. et Ferragne, E. (2018). Influence de la posture corporelle sur les paramètres acoustiques de la parole. *In Journées d'Études sur la Parole*, 568–575, Aix-en-Provence.

Dellwo, V., Ferragne, E. et Pellegrino, F. (2006). The perception of intended speech rate in English, French, and German by French speakers. *In Speech Prosody*, Dresde, Allemagne.

Dellwo, V., Steiner, I., Aschenberner, B., Dankovicova, J. et Wagner, P. S. (2004). BonnTempo-Corpus and BonnTempo-Tools: A database for the study of speech rhythm and rate. *In Interspeech-ICSLP*, 777–780, Jeju, Corée.

Detey, S., Durand, J., Laks, B. et Lyche, C., dir. (2016). *Varieties of Spoken French*. Oxford University Press, Oxford.

Diana, A. (2010). La phonétique dans l'enseignement de l'anglais aux spécialistes d'autres disciplines : Enjeux et priorités. *Cahiers de l'APLIUT*, 29(3):10–21.

Dixon, J. A., Mahoney, B. et Cocks, R. (2002). Accents of guilt?: Effects of regional accent, race, and crime type on attributions of guilt. *Journal of Language and Social Psychology*, 21(2):162–168.

Dmitrieva, O., Jongman, A. et Sereno, J. (2010). Phonological neutralization by native and non-native speakers: The case of Russian final devoicing. *Journal of Phonetics*, 38(3):483–492.

Dresher, B. E. (2009). *The Contrastive Hierarchy in Phonology*. Cambridge University Press, Cambridge.

Duda, R. O., Hart, P. E. et Stork, D. G. (2001). *Pattern Classification*. Wiley, New York, 2e édition.

Dumas, I. et Ferragne, E. (2001). La prosodie : un marqueur de politesse. *In Journées Prosodie*, s. p., Grenoble.

Dumas, I. et Ferragne, E. (2003). How does fundamental frequency correlate with perceived politeness in greetings? *In Interfaces Prosodiques*, s. p., Nantes.

Durand, J. (2009). On the scope of linguistics: Data, intuitions, corpora. *In* Kawaguchi, Y., Minegishi, M. et Durand, J., dir., *Corpus Analysis and Variation in Linguistics*, 25–52. Benjamins, Amsterdam.

Durand, J. et Lyche, C. (2016). Approaching variation in PFC: The liaison level. *In* Detey, S., Durand, J., Laks, B. et Lyche, C., dir., *Varieties of Spoken French*, 363–375. Oxford University Press, Oxford.

Elvin, J., Williams, D. et Escudero, P. (2016). Dynamic acoustic properties of monophthongs and diphthongs in Western Sydney Australian English. *The Journal of the Acoustical Society of America*, 140(1):576–581.

Epps, J., Smith, J. R. et Wolfe, J. (1997). A novel instrument to measure acoustic resonances of the vocal tract during phonation. *Measurement Science and Technology*, 8(10):1112–1121.

Everitt, B., dir. (2011). *Cluster Analysis*. Wiley, Chichester, 5e édition.

Feldman, N. H., Griffiths, T. L. et Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, 116(4): 752–782.

Ferragne, E. (2008). *Étude Phonétique des Dialectes Modernes de l'Anglais des Îles Britanniques : Vers l'Identification Automatique du Dialecte*. Thèse de Doctorat, Université Lumière Lyon 2, Lyon.

Ferragne, E. (2013). Automatic suprasegmental parameter extraction in learner corpora. *In* Díaz-Negrillo, A., Ballier, N. et Thompson, P., dir., *Automatic Treatment and Analysis of Learner Corpus Data*, 151–168. Benjamins, Amsterdam.

Ferragne, E. (2019). Phonetics and artificial intelligence: Ready for the paradigm shift? [conférence invitée]. *In Phonologie de l'Anglais Contemporain*, Aix-en-Provence.

Ferragne, E. (2020). The production and perception of derived phonological contrasts in selected varieties of English. *In* Przewozny, A., Viollain, C. et Navarro, S., dir., *The Corpus Phonology of English: Multifocal Analyses of Variation*, 30–49. Edinburgh University Press, Édimbourg.

Ferragne, E., Afonso-Santiago, J. et Pellegrino, F. (2010). Étude acoustique d'un contraste dérivé en anglais d'Écosse. *In Journées d'Études sur la Parole*, 381–384, Mons, Belgique.

Ferragne, E., Bedoin, N., Boulenger, V. et Pellegrino, F. (2011). The perception of a derived contrast in Scottish English. *In ICPhS*, 667–670, Hong Kong.

Ferragne, E., Flavier, S. et Fressard, C. (2013). ROCme! software for the recording and management of speech corpora. *In Interspeech*, 1864–1865, Lyon.

Ferragne, E., Gendrot, C. et Pellegrini, T. (2019). Towards phonetic interpretability in deep learning applied to voice comparison. *In ICPhS*, 790–794, Melbourne.

Ferragne, E. et Pellegrino, F. (2004). A comparative account of the suprasegmental and rhythmic features of British English dialects. *In Modélisations Pour l'Identification Des Langues*, s. p., Paris.

Ferragne, E. et Pellegrino, F. (2008). Le rythme dans les dialectes de l'anglais : Une affaire d'intensité ? *In Journées d'Études sur la Parole*, article 1678, Avignon.

Ferragne, E. et Pellegrino, F. (2010a). Vowel systems and accent similarity in the British Isles: Exploiting multidimensional acoustic distances in phonetics. *Journal of Phonetics*, 38(4):526–539.

Ferragne, E. et Pellegrino, F. (2010b). Formant frequencies of vowels in 13 accents of the British Isles. *Journal of the International Phonetic Association*, 40(01):1–34.

Ferragne, E. et Pellegrino, F. (2010c). Towards a knowledge-based system for accent classification in the British Isles [communication orale]. *In LabPhon*, Albuquerque.

Fox, R. A. et Jacewicz, E. (2009). Cross-dialectal variation in formant dynamics of American English vowels. *The Journal of the Acoustical Society of America*, 126(5):2603–2618.

Frumkin, L. (2007). Influences of accent and ethnic background on perceptions of eyewitness testimony. *Psychology, Crime & Law*, 13(3):317–331.

Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J. et Gravier, G. (2005). The ESTER phase II evaluation campaign for the rich transcription of French broadcast news. *In Interspeech*, 1149–1152, Lisbonne.

Garellek, M., Gordon, M., Kirby, J., Lee, W.-S., Michaud, A., Mooshammer, C., Niebuhr, O., Recasens, D., Roettger, T. B., Simpson, A. et Yu, K. M. (2020). Toward open data policies in phonetics: What we can gain and how we can avoid pitfalls. *Journal of Speech Science*, 9(1):3–16.

Gastner, M. T. et Newman, M. E. J. (2004). Diffusion-based method for producing density-equalizing maps. *Proceedings of the National Academy of Sciences*, 101(20):7499–7504.

Gay, T. (1978). Effect of speaking rate on vowel formant movements. *The Journal of the Acoustical Society of America*, 63(1):223–230.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, 33(5):587–606.

Gilquin, G. et Gries, S. T. (2009). Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory*, 5(1):1–26.

Goodfellow, I., Bengio, Y. et Courville, A. (2016). *Deep Learning*. MIT Press, Cambridge, MA.

Greenwald, A., Gonzalez, R., Harris, R. J. et Guthrie, D. (1996). Effect sizes and p values: What should be reported and what should be replicated? *Psychophysiology*, 33(2):175–183.

Gubian, M., Torreira, F. et Boves, L. (2015). Using functional data analysis for investigating multidimensional dynamic phonetic contrasts. *Journal of Phonetics*, 49:16–40.

Guiraud, H., Ferragne, E., Bedoin, N. et Boulenger, V. (2013). Adaptation to natural fast speech and time-compressed speech in children. *In Interspeech*, 1370–1374, Lyon.

Guiraud, H., Ferragne, E., Bedoin, N., Krifi-Papoz, S., Herbillon, V., Bascoul, A., Gonzalez-Monge, S. et Boulenger, V. (2014). Perception de la parole rapide chez les enfants présentant une dysphasie expressive. *In Journées d'Études sur la Parole*, 1–4, Le Mans.

Gut, U. et Voormann, H. (2014). Corpus design. *In* Durand, J., Gut, U. et Kristoffersen, G., dir., *The Oxford Handbook of Corpus Phonology*, 13–26. Oxford University Press, Oxford.

Guyot-Talbot, A., Heidelmayr, K. et Ferragne, E. (2016). Entraînements à la prosodie des questions ouvertes et fermées de l'anglais chez des apprenants francophones. *In Journées d'Études sur la Parole*, 265–273, Paris.

Hall, K. C. (2009). *A Probabilistic Model of Phonological Relationships from Contrast to Allophony.* Thèse de Doctorat, Ohio State University.

Hall, K. C. (2013). A typology of intermediate phonological relationships. *The Linguistic Review*, 30(2):215–275.

Hanzlíková, D. et Skarnitzl, R. (2017). Credibility of native and non-native speakers of English revisited: Do non-native listeners feel the same? *Research in Language*, 15(3):285–298.

Harrington, J. (2010). *Phonetic Analysis of Speech Corpora.* Wiley-Blackwell, Chichester.

Harris, J. (1990). Derived phonological contrasts. *In* Ramsaran, S., dir., *Studies in the Pronunciation of English : A Commemorative Volume in Honour of A.C. Gimson*, 87–105. Routledge, Londres.

He, K., Zhang, X., Ren, S. et Sun, J. (2016). Deep residual learning for image recognition. *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778, Las Vegas.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T. et Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLOS Biology*, 13(3):1–15.

Heaven, D. (2019). Why deep-learning AIs are so easy to fool. *Nature*, 574(7777):163–166.

Heidlmayr, K., Ferragne, E. et Isel, F. (2021). Neuroplasticity in the phonological system: The PMN and the N400 as markers for the perception of non-native phonemic contrasts by late second language learners. *Neuropsychologia*, 156:107831.

Heinze, T., Shapira, P., Rogers, J. D. et Senker, J. M. (2009). Organizational and institutional influences on creativity in scientific research. *Research Policy*, 38(4):610–623.

Hillenbrand, J. M. (2013). Static and dynamic approaches to vowel perception. *In* Morrison, G. S. et Assmann, P. F., dir., *Vowel Inherent Spectral Change*, 9–30. Springer, Berlin.

Hovy, D. (2020). Layers, biases, and responsability [conférence invitée]. *In 2e Atelier Éthique et TRaitemeNt Automatique des Langues (ETeRNAL)*, Nancy.

Isaacs, T. (2018). Shifting sands in second language pronunciation teaching and assessment research and practice. *Language Assessment Quarterly*, 15(3):273–293.

Îto, J. et Mester, A. (1995). Japanese phonology. *In* Goldsmith, J. A., dir., *Handbook of Phonological Theory*, 817–838. Blackwell, Cambridge, MA.

Janse, E. (2004). Word perception in fast speech: Artificially time-compressed vs. naturally produced fast speech. *Speech Communication*, 42(2):155–173.

Johnson, K. (2007). Decisions and mechanisms in exemplar-based phonology. *In* Sole, M.-J., Speeter Beddor, P. et Ohala, M., dir., *Experimental Approaches to Phonology*, 25–40. Oxford University Press, Oxford.

Kendall, T. et Vaughn, C. (2015). Measurement variability in vowel formant estimation : A simulation experiment. *In ICPhS*, article 797, Glasgow.

Kiefte, M., Neary, T. M. et Assmann, P. F. (2013). Vowel perception in normal speakers. *In* Ball, M. J. et Gibbon, F. E., dir., *Handbook of Vowels and Vowel Disorders*, 160–185. Psychology Press, New York.

Kim-Dufor, D.-H., Ferragne, E., Dufor, O., Astesano, C. et Nespoulous, J.-L. (2010). Perception and comprehension of linguistic and affective prosody in children with Landau-Kleffner syndrome. *In Speech Prosody*, article 885, Chicago.

Kim-Dufor, D.-H., Ferragne, E., Dufor, O., Astésano, C. et Nespoulous, J.-L. (2012). A novel prosody assessment test: Findings in three cases of Landau–Kleffner syndrome. *Journal of Neurolinguistics*, 25(3):194–211.

King, H. et Ferragne, E. (2018). La parole sans les lèvres : une étude acoustique et articulatoire. *In Journées d'Études sur la Parole*, 451–459, Aix-en-Provence.

King, H. et Ferragne, E. (2019). The contribution of lip protrusion to Anglo-English /r/: Evidence from hyper- and non-hyperarticulated speech. *In Interspeech*, 3322–3326, Graz, Autriche.

King, H. et Ferragne, E. (2020). Loose lips and tongue tips: The central role of the /r/-typical labial gesture in Anglo-English. *Journal of Phonetics*, 80:100978.

King, H. et Ferragne, E. (2021). Labiodentals /r/ here to stay: Deep learning shows us why. *Anglophonia*, 30:3424.

KLATT, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America*, 59(5):1208.

KOENECKE, A., NAM, A., LAKE, E., NUDELL, J., QUARTEY, M., MENGESHA, Z., TOUPS, C., RICKFORD, J. R., JURAFSKY, D. et GOEL, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

KRAEMER, H. C. et BLASEY, C. (2016). *How Many Subjects? Statistical Power Analysis in Research*. Sage, Los Angeles, 2e édition.

KRIZHEVSKY, A., SUTSKEVER, I. et HINTON, G. E. (2012). ImageNet classification with deep convolutional neural networks. *In* PEREIRA, F., BURGES, C. J. C., BOTTOU, L. et WEINBERGER, K. Q., dir., *Advances in Neural Information Processing Systems 25*, 1097–1105. Curran Associates, Red Hook.

KRONROD, Y., COPPESS, E. et FELDMAN, N. H. (2016). A unified account of categorical effects in phonetic perception. *Psychonomic Bulletin & Review*, 23(6):1681–1712.

KRZONOWSKI, J., FERRAGNE, E. et PELLEGRINO, F. (2016). Perception et production de voyelles de l'anglais par des apprenants francophones : effet d'entraînements en perception et en production. *In Journées d'Études sur la Parole*, 491–499, Paris.

KRZONOWSKI, J., PELLEGRINO, F. et FERRAGNE, E. (2018). Étude acoustique de la production de voyelles de l'anglais par des apprenants francophones. *In Journées d'Études sur la Parole*, 523–531, Aix-en-Provence.

KUHN, T. S. (1996). *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago, 3e édition.

KUPFERSCHMIDT, K. (2018). More and More Scientists Are Preregistering Their Studies. Should You? *Science*.

LABOV, W. (1972). *Sociolinguistic Patterns*. University of Pennsylvania Press, Philadelphia.

LADEFOGED, P. (2003). *Phonetic Data Analysis: An Introduction to Fieldwork and Instrumental Techniques*. Blackwell, Malden, MA.

LATTNER, S. et FRIEDERICI, A. D. (2003). Talker's voice and gender stereotype in human auditory sentence processing – evidence from event-related brain potentials. *Neuroscience Letters*, 339(3):191–194.

LE CERF, M. et FERRAGNE, E. (2020). Paramètres acoustiques et phonétiques dans la parole parkinsonienne avant et après traitement LSVT LOUD®. *In Journées d'Études sur la Parole*, 326–334, Nancy.

LE CUN, Y. (2019). *Quand la Machine Apprend : la Révolution des Neurones Artificiels et de l'Apprentissage Profond*. Odile Jacob, Paris.

Lev-Ari, S. et Keysar, B. (2010). Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of Experimental Social Psychology*, 46(6):1093–1096.

Liberman, M. Y. (2019). Corpus phonetics. *Annual Review of Linguistics*, 5(1):91–107.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A., van Ginneken, B. et Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88.

Lodge, K. (2009). *Fundamental Concepts in Phonology: Sameness and Difference*. Edinburgh University Press, Édimbourg.

Loftus, G. R. (1993). A picture is worth a thousand $p$ values: On the irrelevance of hypothesis testing in the microcomputer age. *Behavior Research Methods, Instruments, & Computers*, 25(2):250–256.

Longino, H. E. (1990). *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press, Princeton.

Luck, S. J. (2014). *An Introduction to the Event-Related Potential Technique*. MIT Press, Cambridge, MA, 2$^{e}$ édition.

McShane, B. B., Gal, D., Gelman, A., Robert, C. et Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, 73(sup1):235–245.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. et Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv e-prints*, arXiv :1908.09635.

Meyer, C. F. (2002). *English Corpus Linguistics: An Introduction*. Cambridge University Press, Cambridge.

Michardière, Q., Guyot-Talbot, A., Ferragne, E. et Pellegrino, F. (2016). Étude transversale du rythme de l'anglais chez des apprenants francophones. *In Journées d'Études sur la Parole*, 328–336, Paris.

Mielke, J. (2017). Visualizing phonetic segment frequencies with density-equalizing maps. *Journal of the International Phonetic Association*, 48(2):1–26.

Miller, J. L. et Volaitis, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception & Psychophysics*, 46(6):505–512.

Montreuil, J.-P. (2001). *La Phonologie de l'Anglais*. Presses Universitaires de Rennes, Rennes.

Morrison, G. S. (2013). Theories of vowel inherent spectral change. *In* Morrison, G. S. et Assmann, P. F., dir., *Vowel Inherent Spectral Change*, 31–47. Springer, Berlin.

Morrison, G. S. et Assmann, P. F., dir. (2013). *Vowel Inherent Spectral Change*. Springer, Berlin.

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

Munro, M. J. et Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, 45(1):73–97.

Näätänen, R., Paavilainen, P., Rinne, T. et Alho, K. (2007). The Mismatch Negativity (MMN) in basic research of central auditory processing: A review. *Clinical Neurophysiology*, 118(12):2544–2590.

Nabney, I. (2002). *NETLAB: Algorithms for Pattern Recognition*. Springer, Londres.

Nagamine, T., Seltzer, M. L. et Mesgarani, N. (2015). Exploring how deep neural networks form phonemic categories. *In Interspeech*, 1912–1916, Dresde, Allemagne.

Negnevitsky, M. (2002). *Artificial Intelligence: A Guide to Intelligent Systems*. Addison Wesley, New York.

Nguyen, N. (2016). Approaching variation in the Phonologie du Français Contemporain project: The segmental level. *In* Detey, S., Durand, J., Laks, B. et Lyche, C., dir., *Varieties of Spoken French*, 341–349. Oxford University Press, Oxford.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, 5(2):241–301.

Nosofsky, M., R. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(1):54–65.

Nusrat, S. et Kobourov, S. (2016). The state of the art in cartograms. *Computer Graphics Forum*, 35(3):619–642.

Nuzzo, R. (2014). Scientific method: Statistical errors. *Nature*, 506(7487):150–152.

O'Connor, J. D. (1991). *Phonetics: A Simple and Practical Introduction to the Nature and Use of Sound in Language*. Penguin Books, Londres.

Olson, J. M. (1976). Noncontiguous area cartograms. *The Professional Geographer*, 28(4):371–380.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.

Oreskes, N. (2019). *Why Trust Science?* Princeton University Press, Princeton.

Paradis, C. et Prunet, J.-F., dir. (1991). *The Special Status of Coronals: Internal and External Evidence*. Academic Press, San Diego.

Patel, A. D., Iversen, J. R. et Rosenberg, J. C. (2006). Comparing the rhythm and melody of speech and music: The case of British English and French. *The Journal of the Acoustical Society of America*, 119(5):3034–3047.

Pélissier, M. (2018). *Effets d'Entrainements Explicites et Implicites Sur l'Acquisition de la Syntaxe de l'Anglais par des Apprenants Francophones : Étude en Potentiels Évoqués*. Thèse de Doctorat, Université Paris Diderot, Paris.

Pélissier, M. et Ferragne, E. (2021). The N400 reveals implicit accent-induced prejudice [en révision]. *Speech Communication*.

Pellegrini, T. et Mouysset, S. (2016). Inferring phonemic classes from CNN activation maps using clustering techniques. *In Interspeech*, 1290–1294, San Francisco.

Pellegrino, F., Ferragne, E. et Meunier, F. (2010). 2010, a speech oddity: Phonetic transcription of reversed speech. *In Interspeech*, 1221–1224, Makuhari, Japon.

Peterson, G. E. et Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, 24(2):175–184.

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. *In* Bybee, J. L. et Hopper, P. J., dir., *Typological Studies in Language*, 137–157. Benjamins, Amsterdam.

Pierrehumbert, J. B. (2016). Phonological representation: Beyond abstract versus episodic. *Annual Review of Linguistics*, 2(1):33–52.

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: cerebral lateralization as 'Asymmetric Sampling in Time'. *Speech Communication*, 41(1):245–255.

Popiel, S. J. et McRae, K. (1988). The figurative and literal senses of idioms, or all idioms are not used equally. *Journal of Psycholinguistic Research*, 17(6):475–487.

Popper, K. R. (2002a). *Conjectures and Refutations: The Growth of Scientific Knowledge*. Routledge, Londres.

Popper, K. R. (2002b). *The Logic of Scientific Discovery*. Routledge, Londres.

Pota, S., Spinelli, E., Boulenger, V., Ferragne, E., Varnet, L., Hoen, M. et Meunier, F. (2012). La mie de pain n'est pas une amie : une étude EEG sur la perception de différences infraphonémiques en situation de variations. *In Journées d'Études sur la Parole*, 859–886, Grenoble.

Rakić, T., Steffens, M. C. et Mummendey, A. (2011). When it matters how you pronounce it: The influence of regional accents on job interview outcome: The role of accents in job interviews. *British Journal of Psychology*, 102(4):868–883.

Ramsay, J. O., Hooker, G. et Graves, S. (2009). *Functional Data Analysis with R and MATLAB*. Springer, Berlin.

Ramus, F., Nespor, M. et Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, 73:265–292.

Rastovic, A., Pélissier, M. et Ferragne, E. (2019). The perception of swear words by French learners of English: An experiment involving electrodermal activity. *Anglophonia*, 27:2254.

Rathcke, T. V. et Stuart-Smith, J. H. (2016). On the tail of the Scottish vowel length rule in Glasgow. *Language and Speech*, 59(3):404–430.

Rice, K. (2007). Markedness in phonology. *In* De Lacy, P. V., dir., *The Cambridge Handbook of Phonology*, 79–98. Cambridge University Press, Cambridge.

Roettger, T. B. (2019). Researcher degrees of freedom in phonetic research. *Laboratory Phonology : Journal of the Association for Laboratory Phonology*, 10(1):1.

Roettger, T. B., Winter, B. et Baayen, H. (2019). Emergent data analysis in phonetic sciences: Towards pluralism and reproducibility. *Journal of Phonetics*, 73:1–7.

Rossi, M. (1972). Le seuil différentiel de durée. *In* Valdman, A., dir., *Papers in Linguistics and Phonetics to the Memory of Pierre Delattre*. De Gruyter, Berlin.

Rousselot, J.-P. (1897). *Principes de Phonétique Expérimentale*. Welter, Paris.

Salsburg, D. S. (1985). The religion of statistics as practiced in medical journals. *The American Statistician*, 39(3):220–223.

Scobbie, J. M., Hewlett, N. et Turk, A. (1999). Standard English in Edinburgh and Glasgow: The Scottish vowel length rule revealed. *In* Foulkes, P. et Docherty, G., dir., *Urban Voices : Accent Studies in the British Isles*, 230–245. Arnold, Londres.

Scobbie, J. M. et Stuart-Smith, J. (2008). Quasi-phonemic contrast and the fuzzy inventory: Examples from Scottish English. *In* Avery, P., Dresher, B. E. et Rice, K., dir., *Contrast in Phonology*, 87–113. De Gruyter, Berlin.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. et Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *In IEEE International Conference on Computer Vision (ICCV)*, 618–626, Venise.

Serlin, R. C. et Lapsley, D. K. (1990). Meehl on theory appraisal. *Psychological Inquiry*, 1(2):169–172.

Shadle, C. H., Nam, H. et Whalen, D. H. (2016). Comparing measurement errors for formants in synthetic and natural vowels. *The Journal of the Acoustical Society of America*, 139(2):713–727.

Shah, D. S., Schwartz, H. A. et Hovy, D. (2020). Predictive biases in natural language processing models: A conceptual framework and overview. *In 58$^{th}$ Annual Meeting of the Association for Computational Linguistics*, 5248–5264, en ligne.

Shannon, C. E. et Weaver, W. (1975). *The Mathematical Theory of Communication*. University of Illinois Press, Urbana.

Simonsohn, U., Nelson, L. D. et Simmons, J. P. (2014). P-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, 9(6):666–681.

Simonyan, K. et Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *In International Conference on Learning Representations (ICLR)*, San Diego.

Squire, P. (1988). Why the 1936 literary digest poll failed. *The Public Opinion Quarterly*, 52(1):125–133.

Stephan, P. et Ferragne, E. (2012). Analyse acoustique de contrastes atypiques en anglais d'Irlande du Nord. *In Journées d'Études sur la Parole*, 121–127, Grenoble.

Tanner, J., Sonderegger, M., Stuart-Smith, J. et Fruehwald, J. (2020). Toward "English" phonetics: Variability in the pre-consonantal voicing effect across English dialects and speakers. *Frontiers in Artificial Intelligence*, 3:38.

Taylor, J. R. (1995). *Linguistic Categorization: Prototypes in Linguistic Theory*. Clarendon Press, Londres, 2e édition.

Taylor, J. R. (2012). *The Mental Corpus: How Language is Represented in the Mind*. Oxford University Press, Oxford.

Tilsen, S. et Arvaniti, A. (2013). Speech rhythm analysis with decomposition of the amplitude envelope: Characterizing rhythmic patterns within and across languages. *The Journal of the Acoustical Society of America*, 134(1):628–639.

Trudgill, P. (1983). *On Dialect: Social and Geographical Perspectives*. Blackwell, Oxford.

Trudgill, P. (2018). I'll git the milk time you bile the kittle do you oon't get no tea yit no coffee more oon't I: Phonetic erosion and grammaticalisation in East Anglian conjunction-formation. *In* Wright, L., dir., *Southern English Varieties Then and Now*, 132–147. De Gruyter, Berlin.

Tufte, E. R. (2001). *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, 2e édition.

van Belle, G. (2002). *Statistical Rules of Thumb*. Wiley, New York.

van Berkum, J. J. A., van den Brink, D., Tesink, C. M. J. Y., Kos, M. et Hagoort, P. (2008). The neural integration of speaker and message. *Journal of Cognitive Neuroscience*, 20(4):580–591.

Varshney, L. R. et Sun, J. Z. (2013). Why do we perceive logarithmically? *Significance*, 10(1):28–31.

Viollain, C. et Chatellier, H. (2018). De petits corpus pour une grande base de données sur l'anglais oral contemporain : quels enjeux à la lumière du programme PAC ? *Corpus*, 18:3222.

Warren, M. (2018). First analysis of 'pre-registered' studies shows sharp rise in null findings. *Nature*, d41586–018–07118–1.

Wasserstein, R. L. et Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, 70(2):129–133.

Watson, I., C. et Harrington, J. (1999). Acoustic evidence for dynamic formant trajectories in Australian English vowels. *Journal of the Acoustical Society of America*, 106(1):458–468.

Weenink, D. (2015). Improved formant frequency measurements of short segments. *In ICPhS*, article 445, Glasgow.

Wells, J. C. (1982). *Accents of English. The British Isles.* Cambridge University Press, Cambridge.

Wiener, N. (2007). *Cybernetics or Control and Communication in the Animal and the Machine.* MIT Press, Cambridge, MA, 2$^{e}$ édition.

Wiese, R. (2006). *The Phonology of German.* Oxford University Press, Oxford.

Williams, D. et Escudero, P. (2014). A cross-dialectal acoustic comparison of vowels in Northern and Southern British English. *The Journal of the Acoustical Society of America*, 136(5):2751–2761.

Winer, B. J., Brown, D. R. et Michels, K. M. (1991). *Statistical Principles in Experimental Design.* McGraw-Hill, New York, 3$^{e}$ édition.

Winter, B. (2019). *Statistics for Linguists: An Introduction Using R.* Routledge, New York.

Wonnacott, T. H. et Wonnacott, R. J. (1991). *Statistique.* Economica, Paris, 4$^{e}$ édition.

Zeiler, M. D. et Fergus, R. (2014). Visualizing and understanding convolutional networks. *In* Fleet, D., Pajdla, T., Schiele, B. et Tuytelaars, T., dir., *Computer Vision – ECCV*, volume 8689, 818–833. Springer, Cham.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. et Torralba, A. (2016). Learning deep features for discriminative localization. *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2921–2929, Las Vegas.

Ziliak, S. T. et McCloskey, D. N. (2008). *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives.* University of Michigan Press, Ann Arbor.

Zou, J. et Schiebinger, L. (2018). AI can be sexist and racist — it's time to make it fair. *Nature*, 559(7714):324–326.

# Résumé

Puisque les préférences épistémologiques individuelles des scientifiques ne manquent pas de façonner leurs résultats, ce document de synthèse présente dans un premier temps la position de l'auteur envers un certain nombre de questions méthodologiques en lien avec le domaine de la phonétique contemporaine. Les corpus, les techniques expérimentales, les méthodes quantitatives et le rôle de la technologie sont autant de concepts abordés dans le but de rendre les valeurs et les biais scientifiques de l'auteur plus explicites. Les chapitres suivants offrent une sélection de travaux menés par l'auteur depuis 2008, date de son doctorat. Ces travaux montrent une évolution de la phonétique acoustique s'appuyant sur un corpus à des protocoles plus expérimentaux impliquant une grande variété d'instruments et de types de données. De la classification automatique et la description acoustique-articulatoire des accents de l'anglais des Îles Britanniques au développement de l'hypothèse de la phonémicité gradiente ; de l'étude du rythme de la parole aux études psycholinguistiques avec des apprenants francophones de l'anglais, ce document couvre les résultats principaux et souligne comment cette vaste palette d'intérêts et de méthodes a été mise au service de deux buts cohérents : une approche agnostique face à de nouvelles énigmes et la possibilité d'aider efficacement les étudiants dans le développement de leur propre identité scientifique. Ce document aborde enfin, à partir d'exemples issus des recherches récentes de l'auteur, le changement de paradigme que va provoquer l'avènement du *deep learning* dans de nombreux domaines académiques.

**Mots-clés:** phonétique expérimentale, phonologie de l'anglais, deep learning.

# Abstract

Since scientists' individual epistemological preferences infallibly shape the output of their research, this thesis starts with a presentation of the author's position with respect to a number of methodological issues pertaining to the field of contemporary phonetics. Such concepts as corpora, experimental techniques, quantitative methods, and the role of technology are discussed with the aim of making the author's scientific values and biases more explicit. The following chapters offer a selection of research works the author has carried out since his PhD in 2008. They show an evolution from corpus-based acoustic phonetics to more experimental protocols involving a great diversity of instruments and data types. From the automatic classification and acoustic-articulatory description of British Isles accents to the development of the gradient phonemicity hypothesis; from the study of speech rhythm to psycholinguistic experiments with French learners of English, the thesis covers the main findings and highlights how this wide array of interests and methods has served two consistent goals: an agnostic approach to new puzzles, and the possibility to efficiently help students develop their own scientific identity. The final part of the thesis addresses the forthcoming paradigm shift that deep learning will bring about in many academic fields with illustrations from the author's recent work.

**Keywords:** experimental phonetics, English phonology, deep learning.