Modélisation bayésienne du développement conjoint de la perception, l'action et la phonologie

Marie-Lou Barnaud

UNIVERSITÉ GRENOBLE ALPES

THÈSE

pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTÉ UNIVERSITÉ GRENOBLE ALPES

Spécialité : **Ingénierie de la cognition, de l'interaction, de l'apprentissage et de la création**

Arrêté ministériel : 25 mai 2016

Présentée par
Marie-Lou BARNAUD

Thèse dirigée par **Jean-Luc SCHWARTZ** et
codirigée par **Julien DIARD** et **Pierre BESSIÈRE**

préparée au sein du **Laboratoire Grenoble Images Parole Signal & Automatique (GIPSA-Lab, UMR 5216)**
dans **l'Ecole Doctorale Ingénierie pour la Santé, la Cognition et l' Environnement (EDISCE)**

Modélisation bayésienne du développement conjoint de la perception, l'action et la phonologie

Thèse soutenue publiquement le **19 janvier 2018**,
devant le jury composé de :

Madame Sharon PEPERKAMP
Directrice de recherche, ENS-DEC, Rapporteuse
Monsieur Francis COLAS
Chargé de recherche, INRIA Nancy Grand Est, Rapporteur
Madame Janet PIERREHUMBERT
Professeure, Université d'Oxford, Examinatrice
Monsieur Laurent BESACIER
Professeur, Université Grenoble Alpes, Examinateur et Président du jury
Monsieur Jean-Luc SCHWARTZ
Directeur de recherche, Université Grenoble Alpes, Directeur de thèse
Monsieur Pierre BESSIÈRE
Directeur de recherche, Sorbonne Université, Codirecteur de thèse
Monsieur Julien DIARD
Chargé de recherche, Université Grenoble Alpes, Codirecteur de thèse

Table des matières

Table des figures

Liste des tableaux

CHAPITRE 1

Remerciements

Je laisserai de côté ma perplexité concernant le caractère public de cette formalité traditionnelle et me plierai aux règles. Par cette présente, je vais donc tenter de mettre en exergue les personnalités ayant marquées cette thèse et de notifier ma reconnaissance éternelle - cela va sans dire - à tous ceux qui ont rendus ce document possible. Cette section nécessitant une certaine concision, je ne m'étendrai pas sur la chance inouïe d'avoir été, et d'être toujours, aussi bien entourée. J'omettrai également, non sans regrets, les louanges que nécessiteraient certains brillants esprits. Cette prétérition achevée, c'est donc en toute simplicité que j'annonce le début de mes remerciements.

Cette thèse n'aurait pu être validée sans la présence d'un jury de qualité. C'est pourquoi je tiens à remercier l'ensemble du jury pour avoir accepté d'évaluer cette thèse. Je remercie tout particulièrement le président du jury, Monsieur Laurent Besacier, pour avoir accepté ce titre et avoir accordé sa confiance au travail présenté. Je remercie également les deux rapporteurs, Madame Sharon Peperkamp et Monsieur Francis Colas, pour avoir, d'une part, lu avec grande attention ce long, très long, manuscrit et, d'autre part, pour leurs remarques, conseils et corrections. Pour finir, je remercie Madame Janet Pierrehumbert pour avoir accepté de lire et d'écouter une thèse écrite et présentée en français.

Cette thèse n'aurait tout simplement jamais existé sans mes directeurs de thèse. Si quelqu'un me demande s'il est bien aisé de travailler avec ce trio, je répondrai sans hésiter par l'affirmative. Pour ne citer que quelques-unes de leurs particularités, je dirais que Jean-Luc Schwartz est un excellent pédagogue, toujours disponible, avec un remarquable esprit de synthèse ; Julien Diard est une personne de rigueur, efficace et doté d'un grand esprit critique ; Pierre Bessière est d'une patience infinie, réfléchi et toujours de bons conseils. Si chacun d'eux, à leur manière, représente un chercheur exemplaire, je pense néanmoins que leur équipe, soudée, performante et complémentaire est l'une de leur plus grande force. Merci de m'avoir guidée. Merci de m'avoir accompagnée. Merci de m'avoir soutenue. Merci pour tout ce que vous avez apporté.

Cette thèse n'aurait pu être menée à bien sans un laboratoire pour l'exécuter. J'ai eu la chance de pouvoir travailler dans de très bonnes conditions. J'ai eu toutes les ressources, l'aide et la compagnie nécessaire pour ce que ces trois années se déroulent à merveille. Merci donc à tous pour avoir rendu cette aventure plus simple et agréable.

Cette thèse ne serait pas une thèse si elle n'était pas reconnue comme telle. Merci aux structures et organisations offrant valeur, soutien, formation et crédit au travail mené. Leur apport est nécessaire pour que le doctorat soit apprécié à sa juste valeur.

Cette thèse n'aurait pas la même valeur si elle n'était pas partagée. Je remercie tous ceux m'ayant

permis de diffuser et présenter ces travaux. Merci à tous ceux ayant pris le temps d'écouter, de lire et d'échanger sur les différentes études de cette thèse. Merci à tous ceux ayant porté un intérêt à ce travail.

Enfin, je remercie tous ceux œuvrant chaque jour pour que la recherche continue d'avancer. Néanmoins, si tous ceux contribuant à la recherche ont été nécessaires pour l'avancée de cette thèse, ils n'auraient cependant pas été suffisants pour que je puisse l'achever.

Je remercie en premier lieu ma famille pour toujours croire en moi et pour m'apporter la force et le courage de poursuivre mes ambitions. Je ne tiens pas à décrire ici tout ce que vous avez fait pour moi, ni à préciser dans le détail l'apport et l'importance de chacun. Je vous remercie simplement pour être toujours, toujours là.

Je remercie en second lieu mes proches. Qu'ils soient de Grenoble ou d'ailleurs, ces quelques personnes sont très importantes pour moi et je les remercie pour être et avoir été présents pour moi. Les moments passés avec eux sont de précieux souvenirs qui ont contribués à me faire parcourir tout ce chemin.

Mais finalement si j'en suis arrivée ici, c'est aussi grâce à ces personnes d'un jour ou de tous les jours qui par leurs réflexions, leurs discussions, leurs aspirations et leurs actions ont influencé mes décisions, mes envies et mes projets. Si je peux écrire ces quelques lignes, c'est aussi grâce à eux. Je les remercie donc pour cela, même si, au fond, je ne suis pas certaine qu'ils l'aient fait exprès.

Je n'oublierai pas ceux qui ont contribué à égayer mon quotidien par leur art, efforts et créativité. Cela concerne en particulier tous ceux liés de loin ou de près à mes contributions sur MAL ou SC. Je vous dois une grande part de mes réflexions et de ma motivation. Même si je ne pourrai jamais tous les répertorier, je remercie également les manipulateurs d'air de tous les temps, notamment ceux qui composent et s'exercent avec virtuosité sur le noir et blanc. J'en profite pour remercier le créateur d'un jeu merveilleux qui a influencé mon parcours à plusieurs occasions. Il m'a appris à persévérer et à croire même en l'impossible. Je n'ai pas l'éternité devant moi, donc je ferai en sorte de le résoudre avant, deux fois.

Si je les ai tous évoqués, je n'ai pas raison de t'oublier. Toi sans qui rien de tout ça n'aurait de raison d'être. Même si tu n'en fais qu'à ta tête, je te le dis aussi : merci.

CHAPITRE 2

Introduction

Cette thèse s'inscrit dans le cadre des recherches en sciences cognitives, ce champ de recherche interdisciplinaire ayant pour objet d'étude la pensée et ses mécanismes. Parmi ses nombreuses disciplines, la **phonétique** est celle au cœur de l'ensemble de nos travaux. Du grec « $\Phi\omega\nu\eta\tau\iota\kappa o'\varsigma$ », phonétique signifie littéralement « qui concerne le son ou la parole ». Parmi les branches existantes, nous nous intéressons plus précisément à la phonétique cognitive, que nous définissons ici comme l'ensemble des représentations et processus mentaux permettant la perception et la production du son de parole.

Un des mystères les plus robustes en phonétique est la possibilité de découper les sons, a priori continus, en unités linguistiques discrètes, n'ayant aucune signification particulière, facilitant de façon essentielle la communication. Par la suite, nous nommons ces unités linguistiques indistinctement **catégories phonétiques** ou **unités distinctives**. Dans cette thèse, nous souhaitons en savoir davantage sur le fonctionnement de ce découpage et surtout sur les spécificités des unités distinctives obtenues. Leur compréhension générale étant beaucoup trop présomptueuse pour une simple thèse, nous nous focalisons sur trois de leur caractéristiques : leurs représentations, leur variabilité et leur contenu cognitif.

Premièrement, nous souhaitons mieux appréhender la manière dont les unités sont représentées dans le cerveau, c'est-à-dire en savoir davantage sur le rôle de leurs représentations. Il est couramment admis que les unités sont caractérisées, au minimum, par des **représentations auditives et motrices**. Si leurs activations conjointes durant la production et la perception paraissent maintenant admises dans la littérature, une autre question, tout aussi importante, semble à ce jour non résolue : quels sont les rôles fonctionnels respectifs des représentations auditives et motrices des unités distinctives ?

Deuxièmement, nous aimerions étudier la variabilité des unités distinctives. Nous savons que les unités distinctives sont très variables : elles varient bien sûr d'une langue à l'autre mais également d'un interlocuteur à l'autre. En nous concentrant sur les particularités auditives et motrices propres à chaque individu, nommées **idiosyncrasies**, nous désirons nous intéresser à une problématique majeure de la phonétique : le lien entre la perception et la production. L'existence d'un tel lien global paraît incontestable. Cependant, si les recherches actuelles tendent à montrer que cette relation concerne également les unités distinctives, sa caractérisation exacte n'est pas encore clairement déterminée. Pour cela, nous nous donnons pour objectif de répondre à la problématique suivante : en quoi les idiosyncrasies en perception sont-elles corrélées à celles en production ?

Troisièmement, nous voulons mieux définir le contenu cognitif interne des unités distinctives. Dans la littérature, les sons sont découpés en unités distinctives de différentes tailles. Nous en consi-

dérons deux : le **phonème**, unité minimale de contraste des éléments du lexique, et la **syllabe** [1], unité principale d'organisation des séquences de sons. Les débats pour savoir si le cerveau interprète le son en phonèmes ou en syllabes sont assez nombreux. Si les deux hypothèses semblent conjointement viables, leurs propriétés respectives sont beaucoup moins évidentes. En orientant cette recherche du point de vue des principes de la communication, nous nous demandons alors : comment ces différentes structures nous permettent de communiquer avec des interlocuteurs extérieurs ?

La ligne directrice adoptée pour répondre à ces trois questions est celle du **développement phonétique**. Nous considérons, en lien avec les études sur le développement de la phonétique, que les unités distinctives s'acquièrent et que ce développement est important pour mieux comprendre comment les unités distinctives sont caractérisées. Nous nous intéressons donc aussi bien à la manière dont la phonétique cognitive se développe, qu'aux conséquences de ce développement.

Les méthodologies utilisées pour étudier les unités distinctives sont nombreuses. L'approche considérée dans cette thèse est celle de la modélisation mathématique. Nous choisissons plus précisément la modélisation cognitive algorithmique, c'est-à-dire la modélisation des mécanismes internes du cerveau ayant pour but de mieux en comprendre le fonctionnement. Dans le vaste ensemble des techniques de modélisation, nous utilisons la modélisation bayésienne et plus exactement la programmation bayésienne qui propose les probabilités comme une extension de la logique pour formaliser le raisonnement rationnel. Nous réutilisons et adaptons une famille de modèles, nommée **COSMO** (signifiant « Communicating Objects using Sensori-Motor Operations »), qui modélise la communication parlée, et qui nous permet d'étudier dans un même cadre les représentations auditives et motrices.

2.1 Contributions

Les principales contributions apportées durant cette thèse visent à conduire à une meilleure compréhension de la caractérisation des unités distinctives. Elles permettent, d'une part, de rendre compte des spécificités de l'apprentissage et, d'autre part, de mettre en avant la complémentarité incontournable des représentations auditives et motrices.

En ce qui concerne l'apprentissage que nous modélisons, nos études expérimentales en simulations nous permettent de constater des différences d'apprentissage importantes entre les représentations auditives et les représentations motrices. Les premières sont apprises très rapidement et de façon assez directe tandis que les secondes nécessitent un processus beaucoup plus complexe, ce qui ralentit leur acquisition.

Par ailleurs, nous avons fait l'hypothèse que l'apprentissage des représentations motrices nécessite, d'un côté, d'apprendre la relation entre les représentations auditives et motrices, ce qu'on nomme apprentissage sensorimoteur, et, d'un autre côté, d'apprendre la relation entre les relations motrices et les catégories phonétiques, ce qu'on nomme apprentissage moteur. En analysant chacun de ces deux apprentissages, nous observons, dans un premier temps, que l'apprentissage sensorimoteur peut être facilité par différentes stratégies d'exploration. Nous en proposons trois basées respectivement

1. Par la suite, les phonèmes sont également nommés catégories phonémiques et les syllabes, catégories syllabiques.

sur l'interaction sociale, la focalisation sur les représentations motrices précédemment produites et la généralisation des données apprises. Dans un second temps, nous montrons que l'apprentissage moteur est plus performant lorsqu'il est guidé par un objectif communicatif plutôt que par un objectif purement imitatif.

Pour finir, nos résultats indiquent qu'il n'est pas nécessaire d'apprendre une représentation interne explicite des catégories phonétiques pour communiquer. Cela suggère que l'organisation cognitive des représentations auditives et motrices peut être apprise différemment pour deux interlocuteurs, sans que cela ne les empêche de se comprendre.

En ce qui concerne la complémentarité des représentations auditives et motrices, nos études expérimentales suggèrent trois observations. Premièrement, nous mettons en avant la propriété « bande étroite/bande large » stipulant que les représentations auditives sont plus précises et focalisées sur les données de l'environnement tandis que les représentations motrices sont plus diffuses et sont plus à même de traiter des données bruitées. Deuxièmement, nous remarquons que les représentations auditives semblent davantage guidées par des processus exogènes tandis que les représentations motrices paraissent, à l'inverse, se baser davantage sur des processus endogènes. Troisièmement, nous confirmons, suite à une précédente étude, que les voyelles sont mieux caractérisées par les représentations auditives et les consonnes mieux caractérisées par les représentations motrices.

Tous ces travaux sont effectués grâce à l'implémentation de plusieurs variantes du modèle générique COSMO. Bien que certaines d'entre elles ont déjà été antérieurement développées, tel COSMO 1D, une version unidimensionnelle du modèle, d'autres ont été développées principalement pour cette thèse. À ce titre, nous pouvons nommer principalement COSMO-V, permettant de manipuler des voyelles et COSMO SylPhon, utilisé pour étudier la structure syllabique et phonémique des unités distinctives.

2.2 Plan de thèse

Afin de guider le lecteur, nous proposons un descriptif rapide de chacun de nos chapitres. Cette thèse est organisée de la manière suivante.

Le deuxième chapitre est un état de l'art d'une partie de la phonétique. Nous commençons par discuter des représentations sensorielles et motrices des unités phonétiques selon trois points de vue : les représentations en perception, le lien entre les représentations en perception et en production et la structure cognitive des unités. Ensuite, nous nous intéressons plus spécifiquement à leur développement et nous étudions d'une part comment elles s'acquièrent aussi bien durant le développement de la perception que celui de la production et d'autre part comment se développent les structures cognitives qui permettent de les traiter.

Le troisième chapitre s'intéresse aux modèles phonétiques. La première section suit un plan similaire au deuxième chapitre. Dans une première sous-section, nous commençons par faire une revue de la littérature sur la manière dont les modèles computationnels centrés sur la perception phonétique intègrent les représentations sensorielles et motrices. Ensuite, nous nous intéressons à l'implémentation

du lien entre les représentations sensorielles et motrices en production. Enfin, nous nous intéressons à la structure cognitive. Dans une seconde sous-section, nous étudions l'implémentation et l'acquisition des représentations sensorielles et motrices dans les modèles de développement phonétique. Nous étudions d'une part comment les modèles réalisent l'apprentissage des représentations sensorielles et motrices et d'autre part comment ils prennent en compte la structure cognitive des unités phonétiques. Dans la deuxième section, nous présentons le modèle générique COSMO, que nous utilisons tout au long de cette thèse. Nous présentons pour cela aussi bien la manière dont est construite le modèle que son apprentissage.

Nos études, réalisées avec des variantes du modèle COSMO, ont été concentrées dans trois chapitres de ce manuscrit, abordant respectivement les trois aspects des unités distinctives que nous étudions : la caractérisation de leurs représentations, leur variabilité et leur structure cognitive.

Le quatrième chapitre décrit deux études relatives aux représentations auditives et motrices. Dans la première, nous comparons l'apprentissage auditif et moteur et mettons en évidence la propriété « bande étroite/bande large ». Ainsi, nous montrons que les deux voies auditives et motrices sont complémentaires : la première est apprise rapidement, de façon précise et se focalise sur les sons de l'environnement (bande étroite) ; la seconde est apprise plus lentement, de manière plus imparfaite, ce qui lui procure des capacités de généralisation (bande large). La seconde étude est consacrée à une analyse de trois principes pouvant faciliter l'apprentissage du lien entre les représentations motrices et sensorielles. Ils sont basés respectivement sur l'interaction sociale, la focalisation sur les représentations motrices précédemment produites et la généralisation des données apprises. Nous montrons que ces trois principes peuvent être avantageux pour faciliter l'apprentissage.

Le cinquième chapitre concerne deux études centrées sur les idiosyncrasies. Dans la première étude, nous abordons l'acquisition des idiosyncrasies de production. Nous comparons pour cela deux types d'apprentissage des représentations motrices : le premier basé sur la répétition des sons de l'environnement (apprentissage par imitation), le second basé sur la reproduction des phonèmes de l'environnement (apprentissage par communication). Nous montrons que seul l'apprentissage par communication fait apparaître des idiosyncrasies. En nous servant de ce résultat, nous étudions la corrélation entre les idiosyncrasies en perception et en production. À travers la comparaison entre nos résultats avec des données expérimentales, nous révélons l'importance des représentations motrices dans cette corrélation. Par ailleurs, ce résultat permet de souligner la complémentarité entre les représentations auditives, basées davantage sur des processus exogènes, et les représentations motrices, basées sur des processus endogènes.

Le sixième chapitre présente un modèle, nommé COSMO SylPhon, pour examiner la structure cognitive des unités distinctives, c'est-à-dire le découpage en catégories phonémiques et syllabiques. Cette étude permet notamment d'analyser de manière précise les invariants phonétiques, et particulièrement l'invariant consonantique et de voir s'il est possible d'apprendre les structures phonétiques sans connaissances préalables sur le nombre de catégories. Nous montrons, dans un premier temps, que notre modèle, bien que complexe, est capable d'apprendre des invariants phonétiques. Dans un deuxième temps, nos résultats révèlent que la communication ne nécessite pas forcément d'avoir des représentations phonémiques ou syllabiques identiques chez deux interlocuteurs. Dans un troisième temps, nous observons qu'un apprentissage syllabique, bien que plus long, permet d'apprendre de manière plus précise les phonèmes qu'un apprentissage phonémique et donc permet, à terme, de mieux

discriminer les unités distinctives.

Le septième et dernier chapitre fait une synthèse des résultats obtenus et présente une discussion générale sur différents aspects de la phonétique, de l'apprentissage et de la modélisation computationnelle. Il aborde également à titre de perspective, trois autres études complémentaires à cette thèse : la première propose une architecture neuronale du modèle COSMO, la seconde décrit un modèle prenant en compte les unités lexicales et la troisième expose un modèle intégratif muni de représentions visuelles et somatosensorielles en plus des représentations auditives et motrices.

CHAPITRE 3

Les unités distinctives, caractérisation et développement

Supposons un terme quelconque, par exemple « mato ». Pour traiter ce terme cognitivement, nous supposons qu'il est décomposé en unités distinctives, par exemple, en phonèmes [m, a, t, o] [1] ou en syllabes [ma, to]. Suite à cette décomposition, se pose une première question : comment se caractérisent les unités distinctives ? Nous sommes capables de les produire avec nos gestes moteurs. De même, nous sommes également capables de les percevoir à partir des signaux sonores correspondants. De par cette production et cette perception, il semblerait qu'elles puissent être aussi bien identifiées par des paramètres moteurs que par des paramètres acoustiques. Cependant, cela ne nous informe pas sur la manière dont elles sont caractérisées dans le cerveau afin d'être toujours produites et perçues comme une même unité distinctive. Dit autrement, qu'est-ce qui, dans le cerveau, correspond à une unité distinctive particulière ?

Par la suite, une deuxième question nous intéresse : comment se développent les unités distinctives ? En supposant que les unités distinctives se caractérisent aussi bien par des paramètres acoustiques et moteurs, comment ces paramètres sont-ils définis lors du développement de la parole ? À travers ces deux questions, nous recherchons les caractéristiques de « l'invariant » phonétique des unités distinctives. Dans cette thèse, cette notion est à prendre au sens large, correspondant à tout ensemble permettant de discriminer les unités distinctives les unes des autres.

Ces deux problématiques forment les deux sections principales de ce chapitre. Pour chacune d'elles, nous effectuons un état de l'art des recherches effectuées, principalement en psychologie et neurosciences. Nous nous intéressons également à une troisième question concernant la structure cognitives des unités distinctives. Dans cette thèse, nous considérons deux types d'unités distinctives : les phonèmes et les syllabes. Si nous reprenons notre exemple, de quelle manière se décompose le terme « mato » : [m, a, t, o], [ma, to] ou les deux simultanément ? Du fait que les deux premières questions sont majoritairement traitées du point de vue des phonèmes dans la littérature, nous nous focalisons, dans chaque section, d'abord sur les unités phonémiques. Cependant, pour bien prendre en compte cette troisième question, en fin de chaque section, nous détaillons les études s'intéressant à la structure cognitive des unités distinctives, afin de mieux comprendre si les unités distinctives du cerveau sont davantage des phonèmes ou des syllabes.

1. La notation entre crochets est discutable puisqu'il existe une différence conceptuelle entre une notation entre barres obliques (pour les unités phonologiques abstraites) et une notation entre crochets (pour les allophones dépendant du contexte). Comme nous ne souhaitons insister sur cette distinction conceptuelle dans le présent travail, nous utilisons par défaut les crochets tout au long de ce document.

3.1 Comment se caractérisent les unités distinctives ?

Cette première section est dédiée à la caractérisation des unités distinctives chez l'adulte. Nous abordons cette question du point de vue des études sur la perception qui débattent depuis longtemps de la nature des représentations des unités distinctives et, plus particulièrement, de la nature des représentations des phonèmes. Outre la présentation des différentes théories existant sur le sujet, cela nous permet de présenter le consensus actuel soutenant l'idée que les unités distinctives sont caractérisées aussi bien par des représentations auditives que motrices en perception. Par la suite, nous nous intéressons à la relation entre les représentations en perception et en production en nous basant sur des études expérimentales aux protocoles variés. Enfin, du fait que ces deux parties sont focalisées essentiellement sur des unités distinctives phonémiques, nous questionnons cette suprématie phonémique en nous intéressant aux études sur la structure des unités distinctives chez l'adulte. Nous en déduisons que les deux types d'unités distinctives, phonémiques et syllabiques, semblent coexister dans le cerveau.

3.1.1 La nature sensorielle et motrice des unités en perception

Considérant une unité distinctive phonémique, nous nous intéressons à la nature de l'invariant en perception. Il semble logique de supposer que cet invariant est essentiellement auditif puisque la perception semble se baser sur le son. Cependant, cette supposition pose quelques difficultés, ce qui a donné lieu à l'hypothèse que l'invariant pourrait être plutôt moteur voire perceptuo-moteur. Par la suite, la découverte des neurones miroirs et les nombreuses expérimentations en neurosciences qui en découlèrent confirment une activation des aires motrices du cerveau durant la perception de la parole, ce qui consolide l'hypothèse que les unités possèderaient, en partie, une représentation motrice et que l'invariant serait finalement perceptuo-moteur. Nous détaillons tout ceci dans cette première sous-section.

3.1.1.1 Les trois familles de théories en perception

Les phonèmes sont coarticulés. Cela signifie que les mouvements articulatoires utilisés pour un phonème se chevauchent avec ceux des phonèmes qui précèdent et qui suivent durant la production. En conséquence, non seulement le signal acoustique résultant est dépendant du contexte mais son découpage en unités phonémiques s'avère également non trivial. Ces conclusions sont notamment mentionnées par Liberman et ses collègues lors de leur essai (et échec) pour créer une machine de lecture basée sur un alphabet acoustique (voir, par exemple, Liberman, 1996, pour une synthèse de leurs travaux).

Pour être plus précis, illustrons le cas de la consonne [d]. Comme toute consonne plosive, celle-ci se caractérise, entre autres, par une transition formantique sur le second formant (Cooper et al., 1952; Delattre et al., 1955). De plus, cette transition formantique diffère selon la voyelle suivante. En prenant l'exemple de la voyelle [i] et de la voyelle [u], il est remarqué que la transition est haute et augmente vers le second formant de la voyelle [i] pour la syllabe [di], tandis qu'elle est basse et diminue vers le

second formant de la voyelle [u] pour la syllabe [du] (voir les cercles en pointillés de la Fig. 3.1).

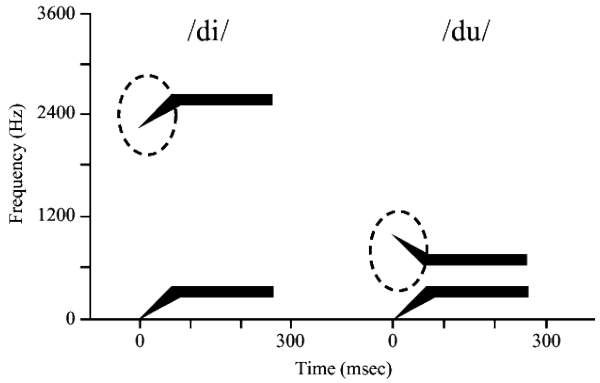


FIGURE 3.1 – Schéma du spectrogramme pour les syllabes synthétiques [di] et [du]. Représentation du premier et du second formants. Issu de Galantucci et al. (2006), adapté de Liberman et al. (1967)

Par ailleurs, Liberman et collègues notent que ces deux phonèmes [d] se produisent de la même manière. En effet, ils sont réalisés tous deux avec un geste de constriction de la pointe de la langue contre les alvéoles des dents (d'où son nom de consonne alvéolaire). Ils généralisent leurs observations et font alors l'hypothèse que, puisque la consonne [d] pour la syllabe [di] et celle pour la syllabe [du] se perçoivent comme un même phonème, mais possèdent des paramètres acoustiques différents, il n'y a pas d'invariant auditif pour ce phonème et, plus généralement, aucun invariant auditif du tout. En revanche, puisque la façon de produire un même phonème semble être similaire et indépendante du contexte, l'invariant phonémique serait en réalité moteur. Ils nomment cela : « la théorie motrice » (Liberman et al., 1967). Dans la dernière version de cette théorie, ils supposent que les invariants sont les « gestes intentionnels » (« intended gestures »), c'est-à-dire les commandes neuromotrices des articulateurs (Liberman et Mattingly, 1985).

Cette hypothèse ne fait pas l'unanimité puisque d'autres théories voient le jour. Certains se sont opposés non pas à l'hypothèse d'un invariant moteur mais à ce à quoi il correspond réellement. C'est notamment le cas de la « théorie réaliste directe », proposée entre autres par Fowler (1986) qui soutient que l'invariant ne correspond pas à des gestes moteurs intentionnels mais à des configurations articulatoires effectives, c'est-à-dire aux gestes physiques produits [2]. Cependant, les principaux opposants à la théorie motrice restent ceux remettant en question l'hypothèse d'un invariant moteur : les partisans des théories auditives et plus récemment, des théories perceptuo-motrices.

Les premiers affirment que l'invariant phonétique est auditif. Pour expliquer le fait que le phonème [d] est catégorisé comme un même phonème dans la syllabe [di] et la syllabe [du] malgré leurs différences acoustiques, les partisans de cette théorie supposent qu'un individu est capable de catégoriser des stimuli complexes à partir de multiples indices acoustiques incomplets (voir, par exemple, l'approche auditive générale de Diehl et al., 2004). Ainsi, il n'est pas nécessaire que la transition formantique du phonème [d] soit identique dans chaque contexte pour percevoir la consonne [d] puisque

2. Il s'agit de l'unique endroit où nous faisons cette distinction. Malgré leurs importantes différences théoriques qui ont suscité de nombreux débats, nous n'allons pas contraster ces deux notions dans cette thèse. Par la suite, nous confondons sous le même terme « moteur » aussi bien le niveau des configurations et gestes moteurs, que le niveau articulatoire, en supposant qu'ils réfèrent tous deux à la chaîne de production.

l'invariant auditif est basé sur de multiples autres indices acoustiques. De plus, ils affirment que l'argument d'un invariant moteur ne tient pas puisque certains oiseaux peuvent être entraînés à catégoriser correctement des syllabes commençant par [d], [b] ou [g] sans pour autant être capables de les produire (Kluender et al., 1987). Un argument similaire est également invoqué sur le développement, les bébés présentant des capacités de perception antérieures à leurs capacités de production, nous y reviendrons.

Les seconds supposent une utilisation conjointe des représentations auditives et motrices durant la perception. Par exemple, selon une première théorie, celle de Skipper et al. (2007), le stimulus perçu est décodé, d'une part, de manière sensorielle et, d'autre part, de manière motrice. Durant le décodage moteur, les commandes motrices correspondant au stimulus sont utilisées pour former une copie d'efférence afin de reproduire le stimulus sensoriel. Celle-ci est, par la suite, utilisée pour valider et affiner le décodage auditif afin de pouvoir interpréter correctement le phonème perçu. Ainsi, selon cette théorie, l'invariant phonétique serait intrinsèquement sensoriel puisque l'interprétation se fait grâce au stimulus auditif perçu et au stimulus auditif issu de la copie d'efférence mais la perception, elle, serait néanmoins sensorimotrice puisqu'elle s'effectue à partir d'un décodage double : l'un sensoriel et l'autre moteur.

Une seconde théorie est proposée par Schwartz et al. (2012), nommée « la théorie de la perception pour le contrôle de l'action » (PACT, « Perception-for-Action-Control Theory »). Elle repose sur l'hypothèse que les représentations motrices et sensorielles se développeraient de manière conjointe durant le développement, ce qui implique qu'elles s'influenceraient l'une et l'autre. Cela permettrait, par la suite, la création de cartes sensorimotrices. Ainsi, même si, comme dans la théorie de Skipper, le décodage et la catégorisation du son en unités phonétiques s'effectuent dans les aires auditives voire sensorielles, les représentations existantes dans ces aires seraient, par construction, sensorimotrices. Nous pouvons donc considérer que, selon cette théorie, les invariants seraient sensorimoteurs. Par ailleurs, il est également proposé que durant la perception, les représentations motrices pourraient également influencer le décodage sensoriel : la perception sensorielle serait donc, de tout point de vue, contrainte par les représentations motrices.

En résumé, il y a donc trois familles de théories : auditives, motrices et sensorimotrices pour lesquelles la perception s'effectue respectivement en utilisant uniquement les représentations auditives, uniquement celles motrices ou les deux conjointement. L'invariant, soit purement auditif, soit purement moteur, selon les théories auditives et motrices, s'avère plus complexe à définir dans les théories sensorimotrices du fait de l'utilisation combinée des deux représentations.

3.1.1.2 Les neurones miroirs et le rôle des représentations motrices en perception

Le débat entre les différentes familles de théories a fait un bond en avant suite aux avancées des neurosciences et particulièrement suite à la découverte des neurones miroirs et des expérimentations qui s'ensuivirent. Brièvement, les neurones miroirs sont des neurones s'activant aussi bien durant la production de l'action que durant la perception de cette même action. Observés chez les singes (Rizzolatti et al., 1996a), il a été fait l'hypothèse qu'il pourrait exister un système similaire chez l'humain, ce qu'on nomme le système miroir (Grafton et al., 1996; Iacoboni et al., 1999; Rizzolatti et al., 1996b).

En parole, cette découverte est venue accréditer l'hypothèse d'une théorie motrice ou sensorimotrice. De nombreuses expérimentations ont été réalisées en ce sens. Il a ainsi été découvert que durant la perception, certaines aires motrices du cerveau sont activées (Fadiga et al., 2002; Pulvermüller et al., 2006; Watkins et al., 2003; Wilson et al., 2004). Il a par ailleurs été remarqué que cette activation est renforcée quand le signal auditif est perturbé ou inhabituel : dans le bruit (Binder et al., 2004; Zekveld et al., 2006), avec des stimuli non natifs (Callan et al., 2014, 2004; Wilson et Iacoboni, 2006) ou avec des stimuli audiovisuels incongruents (Jones et Callan, 2003; Ojanen et al., 2005; Skipper et al., 2007).

Pour autant, une activation des aires motrices durant la perception ne signifie pas que les représentations motrices sont primordiales dans la perception. Plusieurs chercheurs se sont questionnés sur les conséquences réelles de cette activation sur la perception (Lotto et al., 2009; Toni et al., 2008).

Des réponses sont apportées par la neuroimagerie en étudiant non pas l'activation des aires motrices mais leur potentiel rôle durant la perception. À l'aide de la stimulation magnétique transcranienne (transcranial magnetic stimulation, TMS), il a par exemple été montré que des stimulations servant à perturber les aires motrices diminuent les performances de décodage (Meister et al., 2007; Möttönen et al., 2012; Möttönen et Watkins, 2009; Rogers et al., 2014; Sato et al., 2009). Toujours en utilisant la TMS, il a également été montré qu'une stimulation servant à exciter ces même aires peut favoriser la perception (D'Ausilio et al., 2012a, 2009; Grabski et al., 2013). À titre d'illustration, il a été observé lors de l'étude réalisée par D'Ausilio et al. (2009) qu'une stimulation sur la zone de la langue favorise la perception des phonèmes dentaux ([d] et [t]) tandis qu'une stimulation sur la zone des lèvres favorise la perception des phonèmes labiaux ([b] et [p]).

Quelques études comportementales permettent également de révéler le rôle du système moteur en perception. Dans une étude de Sato et al. (2011), des participants ont pour but de catégoriser un stimulus entre [pa] et [ta] après une séance lors de laquelle ils répètent soit 150 mouvements des lèvres, soit 150 mouvements de la langue. Il est remarqué que les participants ont une tendance à percevoir davantage de syllabes labiales ([pa]) que de syllabes dentales ([ta]) après avoir effectué les mouvements des lèvres par rapport à une situation contrôle sans mouvement. L'effet réciproque pour les dentales est également observé. Les auteurs considèrent que ce résultat s'explique par un biais des représentations motrices durant la perception.

Par ailleurs, plusieurs études ont été réalisées avec des patients aphasiques, c'est-à-dire ne parlant pas, possédant une lésion au niveau des aires motrices. Dans certaines d'entre elles, il est observé que ces patients ont de faibles performances de discrimination des unités phonétiques (Blumstein, 1995; Miceli et al., 1980), alors que d'autres chercheurs, plus récemment, ne remarquent aucun déficit de performance (Baker et al., 1981; Hickok et al., 2011; Rogalsky et al., 2011). Sans remettre en cause les précédentes recherches, ce résultat suggère que les représentations motrices ne sont, en réalité, pas forcément nécessaires à la perception.

En résumé, ces différentes études en neurosciences mettent en avant l'activation des aires motrices durant la perception, notamment sous certaines conditions non habituelles, et soulèvent l'importance de ces aires motrices en perception en montrant que leur altération peut modifier la discrimination ou la catégorisation.

3.1.1.3 Conclusion

Pour synthétiser, le débat autour de la nature sensorielle et motrice des unités distinctives s'est fait sur deux périodes assez distinctes. Il concerne initialement les trois familles de théories, auditives, motrices et perceptuo-motrices et la nature de l'invariant respectivement auditif, moteur ou sensori-moteur (voir, par exemple, Diehl et al., 2004; Galantucci et al., 2006; Perkell et Klatt, 1986; Samuel, 2011, pour des revues). Par la suite, grâce à la découverte des neurones miroirs et aux avancées récentes en neurosciences et neuroimagerie, le sujet du débat évolue et les chercheurs se questionnent davantage sur l'activation et l'influence des représentations motrices en perception (voir, par exemple, D'Ausilio et al., 2012b; Stasenko et al., 2013, pour des revues). Le lecteur intéressé peut également se reporter à Grabski (2012); Laurent (2014) pour plus de détails sur l'ensemble de ces deux périodes.

Comme en témoignent les revues récentes sur le sujet, le débat n'est actuellement pas terminé (Hickok, 2010; McGettigan et Tremblay, 2017; Skipper et al., 2017). Si la plupart des chercheurs s'accorde désormais sur le fait que les représentations motrices jouent un rôle durant la perception, l'importance de ce rôle reste encore à définir. Certains affirment que les représentations motrices sont primordiales pour la perception (Meister et al., 2007; Pulvermüller et Fadiga, 2010) tandis que d'autres soutiennent que ce rôle est minime (Scott et Johnsrude, 2003). Pour les derniers, les représentations motrices ne sont pas primordiales (« primary ») mais modulaires (« modulary ») et ne sont importantes que dans certaines conditions : soit lorsque la tâche à réaliser est cognitivement coûteuse ou lorsque les conditions sont bruitées (voir principalement Hickok, 2009).

Cependant, quelle que soit l'importance des représentations motrices, une question subsiste : si les représentations sensorielles et motrices sont utilisées les unes et les autres durant la perception, de façon équilibrées ou non, quelle est la nature de leur rôle ? Cette question est cruciale et a longtemps été abordée, notamment lors des débats sur les familles de théories de la perception. Avec l'arrivée des neurosciences, les recherches se sont davantage tournées vers une vision intégrative dans laquelle les chercheurs ont étudié si ces deux voies ont un rôle en perception plutôt qu'à chercher la nature de ce potentiel rôle. De notre côté, nous nous intéressons à la question d'origine et la traitons dans une de nos modélisations, au chapitre 4.

3.1.2 Perception et production : des invariants communs ?

La nature des invariants en perception est toujours sujette à débat notamment parce qu'il est difficile de définir le rôle réel des représentations motrices en perception. En revanche, la nature perceptuo-motrice des représentations en production semble, elle, beaucoup moins controversée. C'est pourquoi une manière d'étudier le problème en perception consiste à analyser le lien entre la perception et la production. Si ces deux processus ont des invariants communs, connaître la nature de l'un pourrait faciliter la compréhension de l'autre.

Il y a de nombreuses preuves de l'existence d'un lien global entre la perception et la production. Par exemple, l'effet Lombard, décrit par l'otolaryngologiste français Lombard au début du XXème siècle, stipule qu'un locuteur adapte le volume de sa voix au niveau de bruit ambiant. De la même manière, le locuteur adapte également sa production tout entière au niveau de bruit de l'environne-

ment (Lane et Tranel, 1971). Cela illustre le fait que la perception du locuteur influence d'une certaine manière ce qu'il produit. Au même titre, nous venons de voir, dans la section précédente, que les aires motrices utilisées durant la production sont non seulement actives durant la perception mais qu'elles peuvent l'influencer.

Pour autant, peut-on dire que les invariants utilisés en perception sont liés à ceux utilisés en production ? C'est ce que nous allons tenter d'analyser à travers trois types d'études. Les premières concernent l'adaptation perceptuo-motrice où nous observons, d'une part, les effets de la perception sur la production puis, d'autre part, les effets de la production sur la perception. Les secondes s'intéressent à la compatibilité perceptuo-motrice. Enfin, les troisièmes se rapportent aux idiosyncrasies. Toutes ces études suggèrent l'existence d'un lien entre les invariants en perception et ceux en production.

3.1.2.1 L'adaptation perceptuo-motrice : l'influence de la perception sur la production

Commençons par les expérimentations d'adaptation perceptuo-motrice dans lesquelles une tâche de perception provoque une adaptation de la production. Nous décrivons trois séries d'études : celles sur la fatigue auditive, celles sur les changements involontaires et celles sur la perturbation auditive.

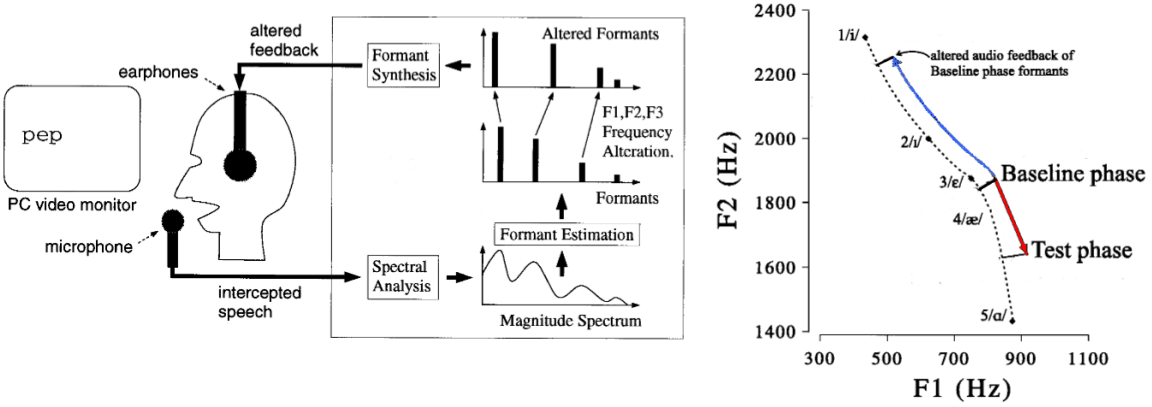
La première série d'études concerne les expérimentations de fatigue auditive, qui consistent à fatiguer le système perceptif afin d'observer les effets sur la production (Cooper, 1979; Cooper et al., 1976; Cooper et Lauritsen, 1974; Jamieson et Cheesman, 1987). Dans ces études, les participants ont pour tâche de prononcer une syllabe Consonne+Voyelle (CV) après avoir écouté en boucle une unité distinctive spécifique (une syllabe ou une voyelle). Les auteurs analysent les valeurs de VOT (Voice-Onset Time), caractérisant la production des consonnes plosives [3]. Pour la plupart des syllabes, ils montrent une adaptation des valeurs de VOT dans les consonnes produites, influencées par ce que les participants ont perçu. Par exemple, ils observent un déclin du VOT pour la consonne [p] de la syllabe [pi] après avoir écouté de façon répétitive la même syllabe [pi] alors qu'un tel déclin n'est pas observé avec l'écoute répétitive de la voyelle [i] (Cooper et Lauritsen, 1974). Ils interprètent ce phénomène comme la conséquence de la fatigue du mécanisme de VOT, partagé entre la perception et la production. Ce résultat peut également s'expliquer par ce qui est nommé l'effet d'adaptation sélective : l'écart observé de la frontière catégorielle vers un stimulus donné suite à une exposition prolongée à ce même stimulus (Eimas et Corbit, 1973).

Dans la seconde série d'études sur les changements involontaires, les expérimentations ont pour objectif d'analyser l'évolution de la production en fonction de ce qui est perçu (Sato et al., 2013, voir aussi Garnier et al., 2013). Ces expérimentations se basent sur un phénomène bien étudié de la littérature : la convergence phonétique, qui est la tendance à modifier sa production afin d'imiter involontairement ce que produit un interlocuteur (voir, par exemple, Aubanel, 2011; Babel, 2009; Lelong, 2012; Pardo, 2013, pour des revues). À cet effet, l'expérimentation de Sato et al. (2013) est composée de trois phases : une production de phonèmes présentés visuellement (baseline), une production de

3. Plus précisément, le VOT, aussi nommé Délai d'Établissement du Voisement, correspond au délai séparant le relâchement de la constriction de la plosive, caractérisée par une explosion acoustique (burst), et le début du voisement, marquant la mise en action des cordes vocales.

phonèmes présentés acoustiquement (test) et à nouveau une production de phonèmes présentés visuellement (after-effect). Les auteurs comparent d'abord les productions entre la phase de test et la baseline pour savoir si le fait d'entendre les phonèmes influence la perception. Ils montrent que les participants ont tendance à imiter le phonème perçu sans même qu'il soit explicitement demandé de le répéter. Ce phénomène subsiste lors de la phase d'after-effect, confirmant que l'adaptation est robuste. Selon les auteurs, ces résultats suggèrent que la production de la parole se module constamment à l'aide de la perception. Cette expérimentation montre également que la convergence phonétique n'est pas seulement un phénomène social mais provient également d'une adaptation sensorimotrice de plus bas niveau.

La troisième série d'études, relative à la perturbation auditive, analyse l'effet d'une perturbation auditive lors de la perception sur la production. Dans ce domaine, les pionniers sont Houde et Jordan (1998). Durant leur expérimentation, les participants doivent produire une syllabe Consonne + Voyelle + Consonne (CVC) affichée sur un écran. Équipés d'un casque, ils reçoivent un feedback auditif altéré en temps réel de leur propre production. L'altération consiste à modifier les valeurs des formants de la voyelle produite (voir Fig. 3.2 pour avoir une vue d'ensemble). Les résultats montrent que les participants ajustent leur production en fonction de ce qu'ils perçoivent. Par exemple, lorsque la voyelle [e] est modifiée avec des formants plus faibles en F1 et plus élevés en F2 de façon à ressembler davantage à un [i] (voir flèche bleue Fig. 3.2b), les participants compensent cette perturbation en prononçant le [e] avec des formants plus élevés en F1 et plus faibles en F2 par rapport à sa prononciation de base, c'est-à-dire une prononciation plus proche du [a] (voir flèche rouge Fig. 3.2b). Ces résultats suggèrent une adaptation sensorimotrice. Des études similaires concordantes ont été par la suite réalisées, confirmant la robustesse de cette expérimentation (Bourguignon et al., 2014; Max et al., 2003; Purcell et Munhall, 2006; Villacorta et al., 2007).



(a) Protocole expérimental (b) Résultat obtenu pour un participant

FIGURE 3.2 – Illustration de l'expérimentation de Houde et Jordan (1998). Figures adaptées de Houde et Jordan (2002)

En résumé, à travers différents protocoles d'expérimentations, ces trois séries d'études montrent une influence des processus de perception sur la production et semblent confirmer que les invariants phonétiques utilisés en perception et en production sont liés. Pour savoir si le lien entre perception et production est intégral, il devrait être possible d'observer des effets bilatéraux, c'est-à-dire des tâches

dans lesquelles la perception est altérée par la production.

3.1.2.2 L'adaptation perceptuo-motrice : l'influence de la production sur la perception

Les études principales d'adaptation perceptuo-motrice ayant étudié et observé l'influence de la production sur la perception sont des études de perturbation auditive. Une des premières études à observer un tel résultat est celle de Shiller et al. (2009). Le protocole expérimental est très similaire à celui de Houde et Jordan (1998), décrit précédemment. Les auteurs ajoutent, en plus, une tâche de perception avant et après la tâche de perturbation pour évaluer l'effet de la perturbation en production sur les performances de perception. Ils testent notamment trois groupes de participants : un groupe avec production et feedback altéré dans des conditions similaires à celles de Houde et Jordan (1998) (altered feedback, AF), un groupe avec production avec un feedback non altéré (unaltered feedback, UF) et un groupe écoutant simplement le feedback altéré sans production (passive feedback, PL). Les participants sont testés avec des séquences Consonne+Voyelle (CV) ou Consonne+Voyelle+Consonne (CVC) pour lesquelles la première consonne est un [s]. L'altération du feedback diminue la fréquence du centroïde spectral de la consonne [s] de manière à ce que le signal ressemble davantage à une consonne [ʃ]. En production, avec le groupe AF, ils observent une compensation contraire à la perturbation avec une augmentation de la fréquence du centroïde, similaire au résultat observé dans Houde et Jordan (1998). En perception, ils observent un déplacement de la frontière catégorielle entre le [s] et le [ʃ]. En fin d'expérimentation, les participants perçoivent davantage le signal comme une consonne [ʃ] plutôt qu'une consonne [s] par rapport à leur catégorisation avant l'expérimentation. Ces résultats sont concordants avec ceux obtenus en production. Cela montre que la perception est également altérée par le feedback auditif au même titre que la production.

Par ailleurs, le groupe PL ne montre aucune altération de sa perception (et bien entendu aucune altération de la production non plus) en fin d'expérimentation. Ceci semble suggérer que la modification pour le groupe AF serait sensorimotrice et qu'une adaptation de la production serait nécessaire pour modifier en conséquence la perception.

Par la suite, Lametti et al. (2014) ont étudié la cause de ce changement perceptif avec un protocole similaire. À la place des consonnes [s] et [ʃ], ils utilisent les voyelles [a,ɛ,i], respectivement à travers les mots [had, head, hid]. En partant du mot [head], ils étudient deux types d'altération : 1) une altération qui augmente le premier formant de la voyelle [ɛ] de façon à ressembler davantage à une voyelle [a] et 2) une altération qui diminue le premier formant de la voyelle [ɛ] de façon à ressembler davantage à une voyelle [i]. Pour ces deux types d'altération, ils étudient l'effet sur la catégorisation aussi bien entre [ɛ] et [a] qu'entre [ɛ] et [i]. Ils remarquent qu'il y a une différence de catégorisation entre [ɛ] et [a], uniquement pour les participants de la seconde altération (avec des formants qui diminuent pour ressembler à un [i]) : la voyelle [ɛ] ressemble davantage à un [a]. Inversement, les participants discriminent différemment [ɛ] et [a] uniquement lors de la première altération (avec des formants qui augmentent pour ressembler à un [a]). Ces résultats semblent donc montrer que la compensation en perception est due aux modifications en production plutôt qu'à l'altération perceptive elle-même.

Cependant, les résultats de Lametti et collègues n'ont pas été retrouvés dans d'autres expérimentations similaires, comme dans Schuerman et al. (2017). Néanmoins, en réalisant des enregistrements

EEG durant la tâche de catégorisation, les auteurs observent une corrélation entre l'altération du feedback auditif et un changement des potentiels évoqués, supportant malgré tout l'hypothèse d'une modification du percept sensoriel causée par l'altération du feedback auditif.

En résumé, ces quelques études basées sur une altération du feedback auditif ne montrent pas seulement un changement en production mais également un changement en perception. Ce résultat suggère que l'adaptation est bien sensorimotrice et qu'il existe véritablement un lien entre les invariants phonétiques en production et en perception.

3.1.2.3 Les effets de compatibilité perceptuo-motrices

Nous analysons maintenant un autre type d'études, celles sur la compatibilité perceptuo-motrice, qui mettent en évidence la relation entre la perception et la production en montrant que l'utilisation de l'une accélère l'autre.

Pour mettre en évidence ce phénomène, les études utilisent le « close shadowing » (répétition rapide). Cela consiste à identifier et répéter le plus rapidement possible un stimulus auditif. Une des premières études à avoir testé cette tâche est celle décrite par Porter Jr et Castellanos (1980). Durant une première expérimentation dite « à choix multiple », les participants doivent répéter le plus rapidement possible des séquences Voyelle + Consonne + Voyelle (VCV) ([aba], [apa], [ama], [aka], [aga]) dès qu'ils les entendent. Ils doivent, typiquement, prononcer la première voyelle dès son apparition et ensuite produire la consonne dès qu'ils l'entendent. Dans une seconde expérimentation dite « à choix unique », les participants réalisent une variante du close shadowing dans laquelle ils doivent prononcer aussi vite que possible la première voyelle et prononcer ensuite l'occurence [ba] dès qu'ils entendent la consonne. Les auteurs trouvent des temps de réaction plus rapide dans la seconde expérimentation « à choix unique » ($\approx$ 170 ms) par rapport à la première ($\approx$ 240 ms). Ils interprètent cette différence comme le temps nécessaire de prise de décision. Néanmoins, du fait de la rapidité d'imitation, ils supposent que le processus de production pourrait commencer avant même que la prise de décision en perception ne soit complètement terminée, suggérant ainsi l'existence de représentations partagées entre les processus de perception et de production.

Cette idée semble être corroborée par une étude reportée dans Luce (1986). Dans cette expérimentation, la tâche est similaire et également composée d'une expérimentation « à choix multiple » et « à choix unique ». En revanche, les auteurs utilisent non pas une répétition de la syllabe perçue mais une touche sur laquelle appuyer. Avec ce protocole, non seulement les temps de réaction sont plus longs que dans l'étude précédente (supérieurs à 300 ms, alors qu'ils ne dépassent pas 240 ms avec le précédent protocole) mais les écarts entre les deux expérimentations le sont également (environ 100 ms, alors qu'ils sont d'environ 70 ms avec le précédent protocole). Cette différence d'écart peut s'expliquer par un lien entre la perception et la production n'existant pas entre la perception et l'appui sur une touche. Galantucci et al. (2006) proposent que cette différence est un argument supplémentaire en faveur de l'existence d'un invariant moteur en perception et en production.

Par la suite, une étude similaire est réalisée par Fowler et al. (2003). Les auteurs trouvent les mêmes résultats que Porter Jr et Castellanos (1980) et observent, en plus, des temps de réaction plus rapides dans la tâche « à choix unique » quand les participants perçoivent la même syllabe que celle

qu'ils doivent produire. Cela semble confirmer le fait que les participants se servent des invariants moteurs perçus pour réaliser leur production.

En résumé, l'ensemble de ces études montrent que les temps de réaction diffèrent selon la congruence des tâches à réaliser en perception et en production. Cela semble suggérer l'utilisation d'invariants communs en perception et en production.

3.1.2.4 Études des idiosyncrasies

Les études sur l'adaptation et la compatibilité perceptuo-motrices montrent que les invariants en perception et en production semblent liés. Si cette hypothèse est correcte, il devrait être possible de réaliser des études dans lesquelles nous observons, pour chaque individu, des similitudes entre les invariants en perception et en production. Cela semble être le cas puisqu'une des expérimentations d'adaptation perceptuo-motrice souligne, entre autres, qu'une meilleure acuité auditive résulte en une plus forte compensation en production lors de l'altération du feedback (Villacorta et al., 2007).

Par ailleurs, il est supposé dans Perkell et al. (2004a) que plus un contraste entre deux unités phonétiques est précisément discriminé plus ce contraste est clairement et distinctement articulé. Pour tester cela, ils ont proposé une expérimentation composée d'une tâche de production et d'une tâche de perception. Durant la première, les participants répètent des phrases contenant entre autres les mots [cod, cud, who'd, hood] à différentes vitesse d'articulation. Durant la seconde, les participants entendent une suite de stimuli synthétiques sur un continuum dans lequel les voyelles évoluent soit entre [ɑ] and [ʌ], soit entre [u] and [ʊ]. Ils doivent respectivement discriminer soit les mots [cod] et [cud], soit les mots [who'd] et [hood]. Les auteurs observent que les participants avec une discrimination supérieure à la moyenne produisent de meilleurs contrastes acoustiques. De mêmes résultats sont obtenus avec une expérimentation testant les contrastes [s] et [ʃ] (Perkell et al., 2004b). Par la suite, des études complémentaires (Franken et al., 2015; Perkell et al., 2008) montrent que plus l'acuité auditive est forte plus les régions vocaliques sont restreintes et focalisées.

Les spécificités de perception et de production des unités phonétiques propres à chaque individu sont nommées les idiosyncrasies. Dans une étude de Bell-Berti et al. (1979), les participants doivent discriminer les voyelles [i] et [ɪ] dans un continuum vocalique. En étudiant leur manière de produire ces deux voyelles, les auteurs remarquent que des stratégies différentes de production mènent à des différences de perception. Similairement, dans une de ces études, Fox (1982) tente de corréler les différences individuelles des voyelles [i, ɪ, ɛ, æ, a, ʌ, o, ʊ, u] en perception avec les différences individuelles des voyelles en production. À l'aide de mesures de distance, il trouve plusieurs corrélations significatives entre les variations individuelles en perception et production. Newman (2003) entreprend le même genre d'études mais, cette fois-ci, avec des caractéristiques consonantiques : pour les plosives [b, p], le VOT et la fréquence centroide ; pour les fricatives [s, ʃ], la pente spectrale et les pics spectraux. L'auteur observe des corrélations individuelles significatives entre la perception et la production pour le VOT des plosives et pour les pics de fréquences des fricatives. Plus récemment, Ménard et Schwartz (2014) étudient la composition des répertoires des unités vocaliques du français en perception et en production. Après avoir effectué une tâche de production (voir Ménard et al. (2008) pour le détail) et une tâche de perception sur dix voyelles du français [a i u y e ɛ œ ø o ɔ],

ils analysent les valeurs de F1 normalisées de chaque sujet. Ils obtiennent deux résultats principaux. Premièrement, pour chaque participant étudié séparément, les voyelles perçues et produites de même aperture ont une valeur formantique de F1 identique, indépendante de l'arrondissement et du lieu d'articulation des voyelles. Deuxièmement, la distance en F1 entre deux voyelles perçues d'un participant est corrélée avec la distance en F1 de ces deux mêmes voyelles produites.

En résumé, à travers leurs résultats sur les idiosyncrasies en comparant les manières d'articuler avec les capacités de discrimination ou en mesurant de manière plus précise les spécificités des unités phonétiques en perception et en production pour chaque individu, les auteurs soutiennent l'existence d'un lien entre les invariants phonétiques en perception et production.

3.1.2.5 Conclusion

Pour synthétiser, les études sur l'adaptation perceptuo-motrice nous permettent de mettre en avant le fait que la perception et la production des unités phonétiques s'influencent mutuellement. Ensuite, à travers les études sur la compatibilité, nous avons vu que le processus de perception facilite l'utilisation du processus de production. Finalement, les études sur les spécificités phonétiques propres à chaque individu, c'est-à-dire les idiosyncrasies, sont également liées en perception et production. Ainsi, ces trois types d'études sont cohérents avec l'hypothèse d'un lien entre les invariants en perception et les invariants en production.

S'il devient de plus en plus convainquant que les invariants en perception et production sont liés, il reste à définir plus exactement la nature de cette connexion. Récemment, diverses théories ont été proposées afin d'éclaircir comment la production et la perception sont connectées (Pickering et Garrod, 2013; Remez, 2015). Nous proposons également un début de réponse à cette question dans une de nos modélisations au chapitre 5.

3.1.3 La structure sensorimotrice des unités

Précédemment, nous avons étudié la nature sensorielle et motrice de l'invariant phonétique. Après nous être d'abord focalisés sur la perception, nous avons relaté plusieurs études analysant le lien entre les invariants en perception et en production. Comme l'illustre l'ensemble de ces études ainsi qu'une revue récente sur le sujet (Schomers et Pulvermüller, 2016), les questionnements autour de la nature sensorielle et motrice de l'invariant phonétique se sont principalement centrés sur les phonèmes. Pour autant, la réalité cognitive phonémique de cet invariant n'est pas avérée et nécessite une analyse plus approfondie. Plusieurs études remettant en question l'existence cognitive du phonème ont proposé à la place d'autres structures alternatives, dont l'unité syllabique.

Cette section est consacrée au débat phonème/syllabe. Dans un premier temps, nous relatons d'abord les principales études comportementales autour de cette question en commençant par la remise en question du phonème et les arguments en faveur d'une unité syllabique puis en détaillant les propositions inverses. Dans un second temps, nous nous intéressons à ce débat du point de vue des neurosciences et mettons notamment en avant les principales avancées en faveur de l'existence en

parallèle d'une double structure cognitive, l'une syllabique et l'autre phonémique.

3.1.3.1 Les théories comportementales : phonèmes versus syllabes

Si les études comportementales des théories de la perception prennent comme objet d'étude l'unité phonémique, comme celles de Liberman et ses collègues, celle-ci ne fait pour autant pas l'unanimité (Liberman, 1957; Liberman et al., 1954). À la place, certains proposent que l'unité phonétique de base est la syllabe (Massaro, 1972; Mehler et Hayes, 1981; Mills, 1980; Stetson, 1951; Studdert-Kennedy, 1975). Par exemple, un des arguments cités dans Massaro (1972) concerne la discrimination entre [di] et [du] (voir section 3.1.1.1, Fig 3.1). Si l'invariant est syllabique, la distinction entre le phonème [d] de [di] et celui de [du] n'est plus une difficulté puisque les deux unités syllabiques sont distinctes. De même, l'existence cognitive d'une unité syllabique expliquerait pourquoi Liberman et ses collègues n'ont pas réussi à créer un alphabet acoustique phonémique (voir aussi Fowler, 1984).

Une des premières expérimentations remettant clairement en cause l'existence d'une unité phonémique est proposée par Savin et Bever (1970). Dans leur étude, ils réalisent une tâche de rapidité dans laquelle des participants doivent reconnaître soit des syllabes ([baeb] ou [saeb]) soit des phonèmes contenus dans cette syllabe (les consonnes [b] ou [s] pour un groupe, et la voyelle [ae] pour l'autre). Les auteurs remarquent que les participants reconnaissent plus rapidement les cibles syllabiques que celles phonémiques. Ils interprètent ce résultat comme une preuve que le phonème n'est pas l'unité de base et qu'il est issu de traitements des unités de plus haut niveau. Par la suite, des études similaires sur les temps de réaction sont réalisées supportant l'hypothèse que l'unité de base n'est pas phonémique mais syllabique (Foss et Swinney, 1973; Segui et al., 1981). À ce titre, Segui et al. (1981) terminent leur article par : « the syllable can be seen as the structural unit from which both higher and lower level analyses originate ».

Cependant, les études sur les temps de réaction ne convainquent pas tout le monde, même les chercheurs en faveur d'une unité syllabique (par exemple, Content et Frauenfelder, 2002; Cutler et al., 1986; McNeill et Lindig, 1973). Ces derniers affirment, entre autres, que ces expérimentations sont dépendantes du protocole expérimental utilisé. Il est, par exemple, montré que ces expérimentations sont dépendantes du lexique : les phonèmes des mots sont retrouvés plus rapidement que ceux des non-mots (Rubin et al., 1976).

Pour pallier cette difficulté, Mehler et al. (1981) proposent un nouveau type d'expérimentation. Bien que toujours basée sur les temps de réaction, leur tâche ne compare pas directement les temps de réaction des syllabes et des phonèmes. L'objectif est de détecter le plus rapidement possible soit une syllabe initiale CV (par exemple, [ba]), soit une syllabe initiale CVC (par exemple, [bal]) dans une séquence de mots bisyllabiques français. Pour chaque syllabe initiale visée, il existe dans le corpus une paire de mots commençant par les trois mêmes phonèmes (par exemple [bal]) mais possédant une structure syllabique initiale différente : soit CV (par exemple « balance »), soit CVC (par exemple « balcon »). Les paires ne sont, bien entendu, pas présentées dans la même séquence. Les auteurs comparent les temps de réaction des différentes syllabes recherchées par rapport à la structure syllabique des paires de mots. Ils montrent que les participants détectent plus rapidement les syllabes CV dans les mots CV que dans les mots CVC et inversement (voir Fig. 3.3). Pour reprendre l'exemple,

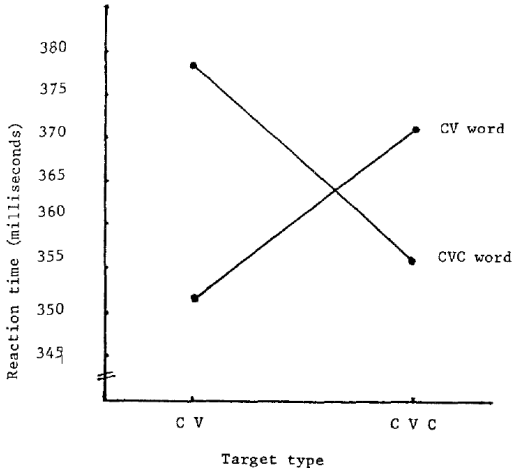


FIGURE 3.3 – Temps de réaction entre des mots CV et des mots CVC. Résultat repris de Mehler et al. (1981)

cela signifie que [ba] est plus facilement repéré dans « balance » et [bal] plus facilement repéré dans « balcon ». Ainsi, les participants détectent plus facilement une séquence de phonèmes quand elle correspond à la première syllabe d'un mot. Les auteurs concluent que ces résultats sont compatibles avec une unité syllabique de base.

Cette étude reste également dépendante du lexique puisqu'une reproduction dans des conditions similaires avec des pseudos-mots échoue à retrouver l'effet observé (Content et al., 2001). Par ailleurs, le même paradigme est appliqué sur des sujets anglophones, avec des mots anglais. Le phénomène n'est pas non plus observé (Cutler et al., 1983, 1986). Une explication avancée est que cet effet « syllabique » ne serait spécifique qu'à certaines langues, dont le français, et que les autres langues segmenteraient les mots différemment, comme l'anglais ou le japonais (Bruck et al., 1995; Finney et al., 1996; Mattys et Melhorn, 2005; Otake et al., 1993). Néanmoins, pour les langues avec une segmentation similaire au français, les mêmes résultats sont obtenus, comme le montre, par exemple, une expérimentation basée sur le même protocole, réalisée avec des illettrés portuguais (Morais et al., 1989).

D'autres expérimentations avec des illettrés contribuent également à faire valoir la syllabe comme unité de base (voir, par exemple, Bertelson, 1986; Morais, 1985, pour une revue). Dans une étude de Morais et al. (1979), les participants illettrés sont incapables d'ajouter ou d'enlever un phonème à un non-mot (ajouter un [b] à « abata » pour former « babata ») alors que les participants lettrés réussissent sans difficulté. Ainsi, même s'ils savent segmenter les mots en syllabes (Morais et al., 1989), les illettrés ne semblent pas capables de manipuler les phonèmes. Le phonème pourrait donc ne pas être une unité de base de la parole et n'émerger que lors de l'apprentissage de la lecture. Cependant, les auteurs rappellent que l'absence de capacité à manipuler consciemment les phonèmes ne correspond pas pour autant à une absence totale de traitement phonémique.

La syllabe est également vue comme une unité de base par certains chercheurs en production (voir Samlowski, 2016, pour une revue). Un des exemples les plus connus est l'existence d'un « mental

syllabary », c'est-à-dire la proposition que les syllabes sont stockées dans un inventaire accessible en production (voir, par exemple, Crompton, 1981; Levelt et al., 1999; Levelt et Wheeldon, 1994). Parmi les arguments en faveur d'une unité syllabique, ces chercheurs remarquent que certaines erreurs de production conservent la structure syllabique (par exemple, « guinea hig pair » pour« guinea pig hair »). D'autres arguments avancés se basent sur la statistique des syllabes. Par exemple, Levelt et Wheeldon (1994) testent la vitesse de production de mots bisyllabiques, fréquents ou non, contenant des syllabes, elles aussi, fréquentes ou non. Ils montrent que non seulement les participants prononcent plus lentement les mots moins fréquents que les mots plus fréquents, mais que, indépendamment de la fréquence des mots, ils prononcent plus lentement les syllabes peu fréquentes dans leur langue que les syllabes les plus fréquentes.

Mais l'hypothèse syllabique possède également son lot de critiques (Cutler et al., 2001) et certains chercheurs continuent à soutenir que l'unité phonétique primaire est phonémique (Fowler et al., 2016; Ohala et al., 1986). Reprenant l'expérimentation basée sur les temps de réponses de Savin et Bever (1970), Norris et Cutler (1988) effectuent une tâche dans laquelle les participants anglais doivent retrouver une syllabe ou un phonème parmi une liste de phrases. En contrôlant leur protocole expérimental en s'assurant que les participants n'établissent pas de stratégie particulière de segmentation et parcourent l'ensemble de la phrase pour chercher la syllabe ou le phonème souhaité, ils remarquent que les participants sont plus rapides pour détecter le phonème que la syllabe. Ainsi, ils concluent que la syllabe n'est pas l'unité de base, au moins pour l'anglais. Des résultats similaires sont par la suite obtenus (Pitt et Samuel, 1990), même sur des données en français (Pallier, 1997).

En suivant un autre cadre expérimental, Decoene (1993) obtient la même conclusion. Il réalise une série d'expérimentations basées sur une tâche de « primed matching » (voir, par exemple, Beller, 1971, pour le détail). Brièvement, dans cette tâche, les participants doivent décider le plus rapidement possible si une paire de mots néerlandais commence par la même unité ou non. Cette tâche peut être facilitée par deux effets : 1) lorsque l'unité initiale est montrée juste avant de présenter la paire de mots (le priming effect) et 2) lorsque les mots sont prototypiques, c'est-à-dire assez fréquents. Dans des expérimentations différentes, il utilise des unités soit phonémiques, soit syllabiques. Pour chacune d'elles, il teste les différentes combinaisons, c'est-à-dire avec ou sans priming effect et avec des mots prototypiques ou non. Parmi les résultats obtenus, Decoene observe que les participants répondent plus vite avec un priming effect et avec un mot prototypique lorsque l'unité est phonémique mais pas lorsqu'elle est syllabique. Il en déduit que le phonème est l'unité cognitive de base.

En production, certains chercheurs considèrent également le phonème comme l'unité cognitive de base (Dell, 1986; Fromkin, 1984). L'argument principal concerne ici aussi les erreurs de production. Bien que les erreurs de production conservent la forme syllabique de l'énoncé erroné, l'erreur elle-même concerne généralement un unique phonème. C'est le cas de la majorité des erreurs distinguées par Meringer et Mayer (1895) cités dans Levelt et al. (1999) : les échanges (« mell wade » pour « well made »), les anticipations (« taddle tennis » pour « paddle tennis ») et les persévérations (« been abay » for « been away »).

En résumé, ces différentes études montrent que la segmentation, le traitement et la manipulation des unités phonétiques sont très dépendantes de la tâche à effectuer, voire même de la langue considérée. Il reste alors difficile de départager ces deux unités, ce qui rappelle la conclusion faite par Lackner et Goldstein (1975) il y a plus de 30 ans : « It must be concluded that at present the

psychological representation of speech sounds remains an intriguing mystery. ». Comme l'écrivent Goldinger et Azuma (2003) ou McQueen et Cutler (2001), face à cet apparent obstacle, différentes solutions ont été énoncées : certains ont proposé une unité alternative, comme le trait (Marslen-Wilson et Warren, 1994, par exemple), d'autres une absence d'unités phonétiques (par exemple Mitterer et al., 2013), ou encore l'existence cognitive des deux unités (par exemple Healy et Cutting, 1976). C'est sur cette dernière proposition que nous allons nous baser pour présenter les expérimentations réalisées en neurosciences sur la structure cognitive des unités.

3.1.3.2 L'apport des neurosciences dans le débat

Nous avons observé la structure des unités à travers les études comportementales et avons relaté les différentes expérimentations en faveur d'une unité syllabique ou phonémique. Nous étudions maintenant ce sujet du point de vue des neurosciences avec trois types d'études : celles sur les potentiels évoqués, celles sur les oscillations neuronales et celles sur l'activation des aires du cerveau.

À l'aide de techniques d'enregistrement cérébral telles que l'électro-encéphalographie (EEG) ou la magnétoencéphalographie (MEG), il est possible d'observer les réponses neuronales sous forme de potentiels électriques. Lorsque le cerveau est stimulé par des données externes, les potentiels électriques varient ; on recueille alors les « potentiels évoqués ». Parmi les études sur les potentiels évoqués examinant les unités phonétiques, beaucoup d'entre elles analysent la négativité de discordance (mismatch negativity, MMN) qui est une composante des potentiels évoqués survenant lors d'un changement soudain (voir Näätänen et al., 2011, 2007, pour des revues). Une des études les plus marquantes sur la structure des unités phonétiques est celle de Näätänen et al. (1997) qui montrent que des traces phonémiques peuvent être retrouvées à travers les fluctuations de la MMN. L'expérimentation consiste à présenter de façon répétitive le phonème [e] à des participants finlandais (stimulus fréquent), qui est un phonème prototypique de leur langue et de le modifier certaines fois soit par un autre phonème prototypique du finnois, soit par le phonème [õ] qui n'est pas un phonème prototypique (mais qui l'est en estonien). Durant cette présentation, les auteurs mesurent la MMN et remarquent qu'elle est plus importante lorsque le changement s'effectue à l'aide d'un phonème prototypique qu'avec le phonème non prototypique. De plus, avec des participants estoniens, ils observent une plus grande activation de la MMN avec le phonème [õ] qu'avec les participants finlandais. De la même manière, Sharma et Dorman (1999) observent que la MMN est plus forte à la frontière entre deux catégories phonétiques qu'entre deux stimuli à l'intérieur d'une même catégorie. Ces résultats suggèrent qu'il existe une activation spécifique pour les phonèmes.

Avec des protocoles similaires, des traces d'unités syllabiques sont également mises en avant (Alho et al., 1998; Shtyrov et al., 1998, 2000). Par exemple, dans Alho et al. (1998), l'expérimentation consiste à présenter de façon répétitive aux participants le stimulus [da] et à utiliser un stimulus déviant [ba] ou [di]. Dans les deux conditions, les deux syllabes déviantes provoquent l'activation de la MMN malgré la consonne (le [d] de [da] et [di]) ou la voyelle (le [a] de [da] et [ba]) communes avec le phonème d'origine, supportant l'hypothèse d'une reconnaissance syllabique.

À l'aide de l'EEG et de la MEG, il est également possible d'étudier les potentiels électriques à travers d'autres types d'analyses, comme, par exemple les analyses temps-fréquence permettant

d'observer les oscillations neuronales. Les ondes sont généralement regroupées en cinq groupes : les ondes alpha variant entre 8,5 et 15 Hz, les ondes beta variant entre 15 et 30 Hz, les ondes gamma variant entre 40 et 80 Hz, les ondes delta inférieures à 3 Hz et les ondes theta variant entre 4 et 8 Hz. Parmi ces études se focalisant sur les unités de la parole, des traces syllabiques et phonémiques ont également été retrouvées. L'hypothèse principale de ces études consiste à considérer que la perception ne s'effectue pas seulement avec un décodage direct des caractéristiques acoustiques du son mais qu'elle nécessite la prise en compte de fenêtres temporelles de différentes tailles permettant de guider le décodage du flux de parole (voir par exemple Bastiaansen et Hagoort, 2006; Ghitza, 2011, pour des explications plus précises). Dit autrement, pour pouvoir décoder un mot, une syllabe ou un phonème, il faudrait d'abord savoir découper et traiter le signal acoustique en éléments de la taille d'un mot, d'une syllabe ou d'un phonème, ce que provoquerait les oscillations neuronales. Comme le rappelle Ghitza (2011), plusieurs données expérimentales semblent soutenir l'hypothèse qu'il existerait une correspondance entre les unités de parole et certaines oscillations neuronales (voir aussi Poeppel, 2003) :

> Phonetic features (duration of 20–50 ms) are associated with gamma (>50 Hz) and beta (15–30 Hz) oscillations, syllables, and words (mean duration of 250 ms) with theta (4–8 Hz) oscillations, and sequences of syllables and words embedded within a prosodic phrase (500–2000 ms) with delta oscillations (<3 Hz). (Ghitza, 2011, p.1)

En s'appuyant sur ces travaux, Giraud et Poeppel (2012) proposent par la suite une segmentation et un traitement phonétique du son en deux échelles de temps principales, gamma et theta, permettant de traiter les deux unités phonétiques de parole, respectivement le phonème et la syllabe (voir aussi Ghitza, 2013; Hyafil et al., 2015; Morillon et al., 2012).

Si certaines recherches se penchent sur l'étude temporelle du flux de parole, que ce soit en observant les potentiels évoqués ou les oscillations neuronales, d'autres se focalisent sur les aires du cerveau s'activant durant la perception ou la production de parole. Concernant les unités distinctives, les recherches regroupent phonétique et phonologie dans un même ensemble. De ce fait, il est remarqué que les traitements phonétiques et phonologiques s'effectuent principalement dans les aires suivantes : le sulcus/gyrus temporal supérieur (STS et STG), aux alentours de ce qui est parfois nommé l'aire de Wernicke, et le cortex frontal inférieur, aux alentours de ce qui est parfois nommé l'aire de Broca (voir Buchsbaum et al., 2001; Hickok et Poeppel, 2007; Siok et al., 2003, pour plus de détails).

Par la suite, les recherches se sont focalisées sur les aires spécifiques aux différentes unités. Bien que des aires différentes soient trouvées selon les études, celles-ci ne sont pas toujours concordantes. Par exemple, Jäncke et al. (2002) s'intéressent aux syllabes (CV) et les comparent, entres autres, aux voyelles. Ils trouvent une plus forte activation bilatérale au niveau du planum temporale et au niveau du mid-STS, toutes deux proches de l'aire de Wernicke. En revanche, aucune observation n'est faite aux alentours de l'aire de Broca. De leur côté, Peeva et al. (2010) s'intéressent aussi bien aux syllabes qu'aux phonèmes. Mais, contrairement à Jäncke et al. (2002), ils observent une plus forte activation aux alentours du cortex prémoteur ventral pour les processus syllabiques, c'est-à-dire aux alentours de l'aire de Broca, mais rien aux alentours de l'aire de Wernicke. Ils obtiennent des aires plus variées pour les processus phonémiques, pour lesquels ils observent une plus forte activation au niveau du pallidum, du gyrus temporal supérieur postérieur, de l'aire motrice supplémentaire (SMA) et même au niveau du cervelet latéral supérieur. Dans un autre cadre, Markiewicz et Bohland (2016) obtiennent également pour les syllabes une plus forte activation au niveau du cortex prémoteur ventral, mais ob-

servent une plus forte activation des phonèmes au niveau du STS et du sulcus frontal inférieur, ce qui ne correspond pas aux mêmes aires que Peeva et al. (2010) pour les phonèmes. Celles-ci ne sont citées qu'à titre d'illustrations et d'autres études proposent encore des aires différentes (Gelfand et Bookheimer, 2003; Hickok, 2012, voir Price, 2012 pour une revue). Ainsi, bien que des aires différentes soient obtenues, elles ne sont pas concordantes entre les études, ce qui n'est pas sans rappeler les différences obtenues par les études comportementales selon la tâche effectuée. Malgré ces différences, le fait que des aires différentes s'activent pour les phonèmes et les syllabes semble plutôt robuste. Il est, par exemple, montré que des aires différentes pour ces deux types d'unités phonétiques sont présentes, même dans des langues n'utilisant pas explicitement les phonèmes, comme le chinois (Siok et al., 2003; Yu et al., 2015).

En résumé, ces trois types d'études mettent en évidence la structure cognitive syllabique et phonémique des unités distinctives. Les études sur les potentiels évoqués montrent des réponses neuronales spécifiques face à un changement phonémique ou syllabique. Celles sur les oscillations neuronales mettent en parallèle certaines ondes neuronales respectivement gamma et theta pour traiter les unités phonémiques et syllabiques. Enfin, les études en neuroimagerie relèvent des aires en partie différentes pour traiter ces deux types d'unités distinctives.

3.1.3.3 Conclusion

Pour synthétiser, nous avons vu que le débat syllabe/phonème s'est d'abord établi à travers les études comportementales en faveur de l'une ou de l'autre de ces unités phonétiques. S'il y a manifestement des arguments en faveur de chacune des deux positions, plusieurs études ont par la suite pu mettre en évidence l'existence des deux structures cognitives, en parallèle, que ce soit dans les données comportementales ou neuroanatomiques/neurophysiologiques.

Essayons maintenant de replacer ces principaux résultats dans le contexte de la recherche sur les invariants phonétiques. S'il est vrai qu'il existe à la fois ces deux structures cognitives distinctes pour les unités phonétiques, comment sont représentés les invariants sensorimoteurs dans ce contexte ? Les dernières études sur la localisation des aires relatives aux deux types d'unités phonétiques laissent supposer qu'il existerait des aires auditives et motrices aussi bien pour les syllabes que pour les phonèmes. En suivant cette proposition, nous proposerons un modèle pour étudier comment les représentations sensorielles et motrices phonémiques et syllabiques s'articulent.

3.2 Comment se développent les invariants des unités distinctives ?

Dans la section précédente, nous avons observé comment les invariants des unités étaient caractérisés aussi bien selon leur nature sensorielle et motrice que selon leur structure cognitive phonémique et syllabique. Nous nous intéressons maintenant à leur développement et notamment à l'influence du développement sur leur caractérisation. Comme précédemment, nous commençons par décrire les études relatives à la nature des processus d'apprentissage perceptifs et moteurs avant de voir ce qu'ils nous disent sur la question du contenu, syllabique et phonémique, des unités.

3.2.1 Les étapes du développement : différences entre apprentissage des représentations sensorielles et des représentations motrices

Si les représentations sensorielles et motrices semblent être toutes deux utilisées pour caractériser les unités phonétiques, leur développement ne semble pas se faire de la même manière. Les représentations sensorielles, que nous analysons principalement à travers l'apprentissage perceptif, semblent être précoces et universelles pendant les premiers mois puis se focalisent sur les unités de la langue native du bébé. À l'opposé, les représentations motrices, que nous analysons principalement à travers l'apprentissage de la production, semble s'établir petit à petit en suivant différentes phases dont les deux plus connues sont les vocalisations et le babillage.

3.2.1.1 Apprentissage en perception

Dès les premiers mois, le bébé présente les signes de ce qui est nommé une perception catégorielle (voir, par exemple, Jusczyk, 1997; Vihman, 2013, pour des revues), c'est-à-dire qu'il discrimine très bien deux sons faisant partie de deux catégories phonétiques différentes mais difficilement deux sons appartenant à la même catégorie phonétique. Une expérimentation pionnière à ce sujet est celle d'Eimas et al. (1971) sur les bébés de 1 mois qui montre que les bébés sont capables de discriminer de manière catégorielle le [ba] du [pa]. Pour cela, il se sert de la méthode High-Amplitude Sucking (HAS). Utilisée surtout pour les nourrissons entre 0 et 4 mois, cette technique est basée sur le fait que la vitesse de succion d'une tétine fournie au nourrisson en début d'expérience s'accélère quand un changement est détecté par le bébé. Pour les expérimentations en phonétique, elle est particulièrement utilisée pour tester si les bébés sont capables de distinguer deux unités distinctives ou des groupes d'unités distinctives. Durant l'expérimentation, un ensemble de stimuli similaires (ici de type [ba]) est présenté au bébé jusqu'à ce que le taux de succion se stabilise ou descende en dessous d'un certain seuil (effet d'habituation). Ensuite, soit le stimulus continue à être diffusé, soit un nouveau stimulus apparaît (soit [pa], soit un [ba] acoustiquement différent, mais avec des écarts acoustiques entre ces deux sons identiques dans les deux cas). L'expérimentation consiste donc à comparer la vitesse de succion du bébé lorsque le second stimulus correspond à la même unité à celle lorsque le second stimulus correspond à une unité différente. L'expérimentation est concluante si la vitesse de succion augmente significativement face au nouveau stimulus. Dans leur étude, Eimas et al. (1971) observent une augmentation entre le [ba] et le [pa], mais pas entre les deux [ba], bien qu'ils soient acoustiquement différents, ce que les auteurs interprètent comme de la perception catégorielle. Cette expérimentation est par la suite reproduite et confirmée sur d'autres stimuli, témoignant de la robustesse du phénomène (Eimas, 1975; Eimas et Miller, 1980; Hillenbrand et al., 1979; Jusczyk et al., 1977; Trehub, 1973). Cependant, bien que cette perception catégorielle soit assez développée, certaines catégories, comme les fricatives (par exemple, [sa] vs. [za]) sont difficilement discriminées durant les premiers mois (Eilers et Minifie, 1975; Nittrouer, 2001) ou seulement dans certaines conditions (Jusczyk et al., 1979; Levitt et al., 1988).

En s'intéressant aux sons non-natifs, Trehub (1976) observe que les bébés anglais de 1 à 4 mois sont capables de discriminer plusieurs contrastes, respectivement présents en français et tchèque, mais pas en anglais ([pa] / [pã] et [za]-[ř a]), contrairement à des adultes anglais, qui ne peuvent discriminer

que des contrastes de leur langue, comme [li] / [ri]. De la même manière, Kuhl et al. (2006) remarquent que le contraste [r] / [l] est également discriminé par les natifs japonais de 6 à 8 mois, alors qu'il est mal discriminé par les japonais adultes, car il ne fait pas partie de leur système phonologique. Par ailleurs, dès les premiers mois, il apparaît des différences entre les bébés de langue native différente. Prenons l'exemple du VOT entre les phonèmes [b, p, p^h]. Avant 6 mois, les données font état de deux frontières catégorielles, une pour séparer chaque contraste (Lasky et al., 1975). Or, certaines langues ne présentent pas tous les contrastes et ne conservent qu'une seule frontière catégorielle : le français ou l'espagnol ont, par exemple, une unique frontière centrale située entre les deux frontières initiales tandis que l'anglais ne possède qu'une des frontières initiales parmi les deux (voir Fig .3.4). En conséquence, vers 6-8 mois, les bébés anglais ne discriminent pas le [ba] du [pa] alors que les bébés espagnols, au même âge, discriminent les trois contrastes (Eilers et al., 1979).

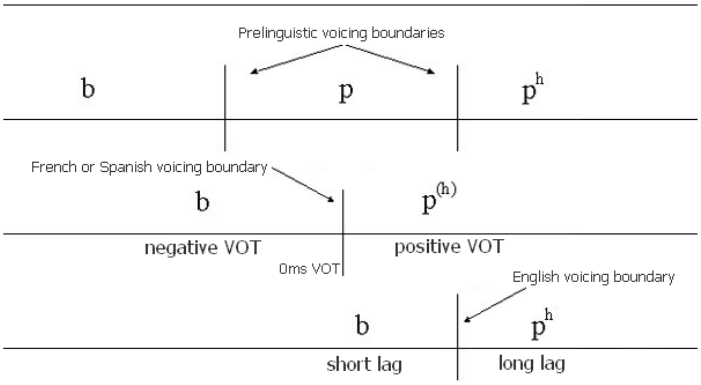


FIGURE 3.4 – Illustration des frontières entre les contrastes [b, p, p^h] selon le continuum VOT. Adapté de Serniclaes et Sprenger-Charolles (2003)

Il semble donc exister des différences d'apprentissage selon les langues. L'expérimentation de Werker et Tees (1984) va même plus loin. Remarquant qu'à 6-8 mois les bébés anglais sont encore capables de discriminer les contrastes non-natifs sur le lieu d'articulation des consonnes (Werker et al., 1981), ils étudient l'évolution de la perception de ces contrastes pour trois groupes de bébés anglais : 6-8 mois, 8-10 mois et 10-12 mois. Ils remarquent qu'à partir de 10-12 mois, les bébés semblent perdre totalement leur capacité à discriminer les lieux d'articulation des consonnes non-natives. C'est ce qu'on nomme le perceptual narrowing.

La technique pour tester la perception des bébés de cette tranche d'âge est nommée visually reinforced Head Turn technique (HT). Elle est basée sur le fait que l'attention du bébé se focalise plus longtemps vers les nouveaux stimuli. Le principe consiste à placer le bébé au centre d'une pièce dans laquelle des sons peuvent être présentés de deux côtés. Un premier son est joué d'un côté, puis, lorsque le bébé se désintéresse du son, un second son est présenté de l'autre, renforcé par l'animation d'un jouet. L'idée est de comparer l'attention du bébé entre un second son similaire au premier et un second son différent. Quand le bébé tourne sa tête significativement plus longtemps vers un nouveau stimulus que vers le stimulus actuel, on dit alors qu'il est capable de distinguer les deux sons (voir, par exemple, Nelson et al., 1995; Werker et al., 1997, pour plus de détails sur cette technique).

À l'aide de cette méthode, les auteurs testent la capacité des bébés à discriminer les contrastes an-

glais [ba]-[da], les contrastes Salish [k̀i]-[q̀i] et les contrastes Hindous [t̪a]-[ta]. Ils montrent d'abord qu'entre 6-8 mois les bébés anglais savent discriminer tous les contrastes. Ensuite, entre 8-10 mois, cette capacité disparait chez certains bébés pour les contrastes [k̀i]-[q̀i] et [t̪a]-[ta] et elle disparait totalement, pour ces deux contrastes, chez les bébés de 10-12 mois. En comparaison, ils montrent que les bébés Salish et Hindous de 10-12 mois conservent bien les contrastes de leur langue respective. Diverses expérimentations confirment ce résultat (Bosch et Sebastián-Gallés, 2003; Conboy et al., 2005; Kuhl et al., 2003, voir aussi Maurer et Werker, 2014 pour une revue). Néanmoins, cette observation ne concerne pas tous les contrastes non-natifs, puisque la perception de certains contrastes est conservée même après 1 an (par exemple les clics Zulu, voir Best et al., 1995). Cependant, il s'agit des contrastes non-natifs généralement aussi perçus par les adultes.

En plus de la perte de la perception des contrastes non natifs, les bébés améliorent leur perception des contrastes natifs. Par exemple, Kuhl et al. (1997) étudient la perception des contrastes [l] et [r] chez des bébés américains et japonais de 6-8 mois et de 10-12 mois en utilisant la technique HT. Ils observent qu'à 6-8 mois, les bébés discriminent les contrastes (environ 64 % de réponses correctes). À 10-12 mois, les bébés japonais, pour qui ce contraste n'est pas natif, discriminent moins bien ces contrastes (environ 60 %). À l'inverse, les bébés américains améliorent leur perception et discriminent mieux ces deux contrastes (environ 74 %). Cette expérimentation est reproduite avec les mêmes résultats (Kuhl et al., 2006) et un résultat similaire est obtenu avec des contrastes mandarins avec les bébés chinois natifs de cette langue (Tsao et al., 2000).

En résumé, l'apprentissage de la perception passe par deux étapes principales : d'abord le bébé semble posséder des capacités précoces universelles puisqu'il paraît capable, dès les premiers mois, de discriminer la plupart des contrastes, aussi bien ceux de sa langue maternelle que ceux d'autres langues. Par la suite, il focalise sa perception sur les contrastes de sa langue et perd la capacité de discriminer les contrastes non-natifs (voir de Boysson-Bardies et Hallé, 2004, pour une revue). Nous n'avons montré que les études comportementales mais les études de neuroimagerie réalisent des observations similaires (voir Conboy et al., 2008a; Kuhl, 2010, pour des revues).

3.2.1.2 Production : l'importance du babillage

Contrairement à la perception pour laquelle le bébé semble avoir des capacités précoces, la production n'est pas développée à la naissance. Le bébé est capable d'utiliser sa voix mais il ne maîtrise pas ses muscles lui permettant d'articuler des sons de parole. Ces capacités se développent au cours du temps. Nous retraçons, dans cette section, les principales étapes du développement phonétique en production. Il existe quelques différences parmi les théories s'intéressant à ce développement phonétique (voir Fig. 3.5, voir aussi Mowrer, 1980, pour une revue). Néanmoins, on retrouve globalement les deux phases principales sur lesquelles nous nous centrons : les vocalisations et le babillage.

Avant 2 mois, le bébé se sert principalement de sa voix pour crier, pleurer et de manière plus générale exprimer ses ressentis par des petits sons mais sa production n'est pas réellement contrôlée. Dans sa théorie, Oller (2000) les nomme des sons végétatifs et les distingue des sons de parole dans le développement qu'il nomme protophones. Pour lui, ces derniers sont spécifiques aux humains alors que les sons végétatifs peuvent être produits par d'autres espèces. Durant cette période, les premières

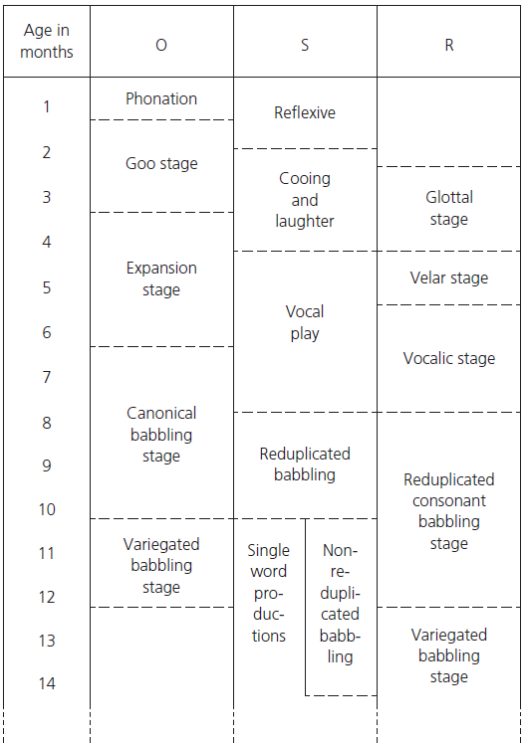


FIGURE 3.5 – Comparaison entre les théories de Oller (1980) (colonne marquée « O »), Stark (1980) (colonne marquée « S ») et Roug et al. (1989) (colonne marquée « R »). Repris de Vihman (2013), Fig 4.1, adapté de Roug et al. (1989)

vocalisations apparaissent, elles sont nommées « quasi-voyelles ». Elles possèdent les caractéristiques acoustiques des sons de paroles et spécialement celles des voyelles (Hollien, 1974) mais pas les caractéristiques motrices. En effet, elles sont produites avec le conduit vocal au repos, c'est-à-dire avec la bouche quasiment fermée, sans utilisation des articulateurs (les lèvres, la langue, etc.), ce qui diffère de la production des voyelles chez l'adulte (Oller et Eilers, 1988).

Entre 2 et 4 mois, d'autres vocalisations apparaissent, plus diversifiées, souvent en réponse à un interlocuteur. Elles sont parfois accompagnées d'une fermeture vélaire mal contrôlée, ressemblant à une fricative, ce qui donne lieu à ce qu'on nomme le « roucoulement » (« cooing » (Stark, 1980) ou « goo » (Oller, 1980) en anglais). Bien que très primaire, cette fermeture correspond aux origines de l'articulation. Ces productions sont d'abord réalisées de façon isolée puis, petit à petit, en série, séparées par un coup de glotte (« glottal stage » de Roug et al., 1989)).

Entre 4 et 7 mois, le bébé commence à mieux contrôler ses organes et muscles articulatoires. De la même manière, il contrôle également mieux certains paramètres acoustiques relatifs à la prosodie (hauteur, vitesse, intensité) et relatifs aux caractéristiques consonantiques comme les bruits de friction, les murmures nasaux ou les consonnes roulées. À ce stade, les voyelles sont produites comme celles des adultes, sans en avoir encore bien sûr la diversité et la maîtrise réelle.

À cette même période, commence la production d'éléments ressemblant à des syllabes. Elles correspondent à un mouvement partant d'une fermeture du conduit vocal (ressemblant à une consonne)

et finissant vers une voyelle bien formée. Cette séquence ne suit cependant pas le rythme des syllabes adultes : il manque la transition rapide entre la consonne et la voyelle (Oller, 2000). Ce sont les prémices d'une étape primordiale dans le développement de la production : le babillage. C'est pourquoi, cette étape est parfois nommée « babillage marginal ».

Le babillage en tant que tel apparait aux alentours de 7 mois. Il s'agit de la phase durant laquelle le bébé commence réellement à produire des gestes moteurs ressemblant à des syllabes ou qui en possèdent au moins les principales propriétés acoustiques en termes de durée et d'enchaînement. La proto-syllabe produite est une syllabe Consonne+Voyelle (CV), composée d'une consonne possédant une complète ou quasi-complète fermeture supraglottale et d'une transition rythmée comme celle d'un adulte vers une voyelle bien formée. L'étape de babillage est souvent composée de deux phases : le babillage canonique commençant au début du babillage et qui se caractérise par la répétition prolongée de mêmes proto-syllabes (par exemple, « babababa... ») et le babillage diversifié (« variegated ») ne commençant généralement pas avant 10-11 mois qui se caractérise par la production de proto-syllabes plus variées (par exemple « digadu... ») (Fagan, 2009; Oller, 1980; Stark, 1980).

Cette distinction entre babillage canonique et diversifié n'est néanmoins pas suivie par tous. Elbers (1982) voit davantage le babillage comme un processus continu d'exploration dans lequel le bébé construit peu à peu ses représentations phonétiques menant à une variation phonétique de plus en plus importante entre 6 et 12 mois. Ceci semble confirmé par diverses études comparant le babillage canonique et diversifié (Mitchell et Kent, 1990; Roug et al., 1989; Smith et al., 1989). À ce titre, Roug et al. (1989) montrent que les productions du babillage diversifié, bien que peu fréquentes au départ, apparaissent dès le début du babillage et augmentent de façon conséquente en fin de babillage.

S'il s'agit d'un processus d'exploration de plus en plus complexe, il peut être vu davantage comme un processus sensorimoteur général qu'un phénomène spécifique au développement de la parole et de la phonétique (Fagan, 2009; Kent, 1984). À ce titre, Thelen (1981) remarque qu'à un âge similaire les mouvements rythmiques et répétitifs se retrouvent pour d'autres parties du corps (les membres, les doigts, etc.). Des études complémentaires observent également un lien important entre le babillage et ces autres mouvements rythmiques (Ejiri, 1998; Ejiri et Masataka, 2001; Iverson et Fagan, 2004; Iverson et al., 2007; Iverson et Thelen, 1999). De la même manière, d'autres auteurs remarquent que les mouvements cycliques de la mâchoire apparaissent au début sans phonation (Meier et al., 1997; Roug et al., 1989).

D'un point de vue similaire, Davis et MacNeilage (1995) proposent l'hypothèse « Frame then Content » (F/C). Une des hypothèses de cette théorie est que le babillage est le résultat d'une oscillation plutôt que la combinaison d'une consonne et d'une voyelle indépendante. Ils supposent ainsi que les consonnes et voyelles effectuées lors du babillage sont articulatoirement liées, ce qui est effectivement vérifié. En effet, les consonnes labiales sont plus souvent combinées avec les voyelles centrales (par exemple, [ba]), dans un geste qui n'impliquerait, pour les auteurs, qu'un pur geste de mâchoire, (« pure frames »). De même, les consonnes alvéolaires sont plus souvent combinées avec les voyelles antérieures (par exemple, [di]) dans un geste d'ouverture qui se superposerait à une avancée globale de la langue (« fronted frames ») et les consonnes vélaires avec les voyelles postérieures (« backed frames », par exemple [gu]) (voir aussi MacNeilage, 1998; MacNeilage et al., 1997).

En résumé, la production semble suivre des étapes relativement précises. D'abord, le bébé com-

mence, à travers la vocalisation, à produire des voyelles de mieux en mieux formées. Puis, guidé par un processus rythmique, il apprend à travers le babillage à associer voyelles et consonnes afin de produire des syllabes de plus en plus évoluées.

3.2.1.3 Conclusion

Pour synthétiser, nous obtenons deux comportements développementaux bien différents lors de l'acquisition de la perception et de la production. Dès les premiers mois, alors que la perception permet de discriminer un grand nombre de contrastes phonétiques, la production ne permet que d'effectuer des vocalisations. Par la suite, cette dernière se développe à travers le babillage pour finalement être capable de produire des syllabes correctement formées. Entre temps, la perception se spécialise et perd peu à peu ses capacités universelles pour ne se focaliser que sur les contrastes natifs.

Dans toute cette section, nous avons fait l'hypothèse que l'acquisition des représentations auditives avait lieu lors du développement auditif, à l'aide de la perception, et que celle des représentations motrices s'effectuait pendant le développement moteur, à l'aide de la production. Cependant, cela ne signifie pas pour autant que la perception ou la production elle-même n'impliquent que des représentations respectivement auditives et motrices chez le bébé. Par exemple, il est montré que les aires auditives et motrices sont conjointement activées durant la perception à 7 et 12 mois (Kuhl et al., 2014). Nous avons simplement voulu mettre en avant les deux processus de développement pris indépendamment.

Par ailleurs, ces différences de développement, bien qu'intrigantes, semblent avoir été assez peu comparées. Pourtant, nous pourrions nous demander si elles ont une quelconque influence sur la caractérisation des unités phonétiques à l'âge adulte ? C'est ce que nous étudierons par la suite dans une de nos modélisations au chapitre 4.

3.2.2 Des représentations sensorielles et motrices si différentes ? Une focalisation sur les stimuli de l'environnement

Nous avons analysé précédemment l'acquisition des processus de perception et de production. Nous avons observé que dans les deux cas, l'apprentissage se fait assez différemment, le premier semblant être guidé par des connaissances précoces puis se focalisant sur les stimuli de la langue tandis que le second s'effectue de manière plus progressive. Nous allons maintenant nous pencher sur les différents mécanismes permettant ces différents apprentissages.

Cette section relate donc les mécanismes utilisés durant le développement de la perception et de la production et montre que, malgré les différences d'apprentissage, les deux processus sont guidés par l'interaction sociale et focalisés sur les stimuli de l'environnement.

3.2.2.1 Perception : un apprentissage statistique avant tout

Dans le contexte des débats multiples et souvent âpres entre approches behavioristes [4] et théories nativistes, notamment Chomskyennes [5], il est maintenant de plus en plus admis que les bébés semblent, en réalité, mettre en oeuvre des mécanismes d'apprentissage statistique leur permettant d'acquérir des connaissances sur la structure des stimuli de leur environnement. L'apprentissage statistique concerne un grand pan de la littérature linguistique incluant entre autres la segmentation des mots (Johnson et Tyler, 2010; Pelucchi et al., 2009; Saffran et al., 1996; Swingley, 2005), la syntaxe (Kidd, 2012; Thompson et Newport, 2007) et, bien sûr, la phonétique et la phonologie (Maye et al., 2008, 2002; Yoshida et al., 2010). Dans cette section, nous nous concentrons essentiellement sur l'apprentissage statistique phonétique.

Une des premières études à ce sujet est celle de Maye et al. (2002), analysant le rôle de la composition des données d'apprentissage dans les performances de catégorisation. Pour cela, ils testent des bébés de 6 à 8 mois sur des stimuli faisant partie d'un continuum [da]-[ta] composé de 8 intervalles de valeurs. Ils utilisent deux groupes : un groupe « unimodal » dans lequel les bébés perçoivent plus fréquemment les intervalles 4 et 5 (au centre du continuum) et un groupe « bimodal » dans lequel les bébés perçoivent plus fréquemment les intervalles 2 (plus proche d'un [da]) et 7 (plus proche d'un [ta]). Par la suite, ils examinent les bébés dans une tâche de perception sur ce même continuum. Les auteurs montrent que les bébés du groupe « unimodal » traitent l'ensemble des stimuli comme des stimuli similaires et ne semblent pas percevoir la différence entre les deux catégories phonétiques. Au contraire, les bébés du groupe « bimodal » font la distinction entre les deux catégories. Cela laisse supposer que les bébés apprennent à discriminer les sons en catégories en suivant la statistique de leur langue.

De plus, la statistique pourrait influer sur les distributions qui sont discriminées ou non en fin d'apprentissage. Ainsi, l'apprentissage focalisé sur les stimuli propres à la langue pourrait être une potentielle explication au perceptual narrowing (Werker, 1994). Pour appuyer ce propos, Anderson et al. (2003) montrent que la fréquence des stimuli de la langue natale influe sur la vitesse de la perte de discrimination des stimuli non-natifs. Ils testent des bébés anglais de 6 mois et demi et 8 mois et demi sur leur capacité à discriminer un contraste coronal et un contraste dorsal. Du fait que les consonnes coronales sont plus fréquentes que les dorsales en anglais, ils supposent que le premier contraste doit disparaître avant le second. Ils observent que les bébés de 6 mois et demi discriminent tous les contrastes et que ceux de 8 mois et demi discriminent moins bien le contraste coronal, ce qui est en accord avec leur hypothèse.

Par la suite, Maye et al. (2008) proposent deux autres expérimentations, dont l'objectif consiste à analyser si la familiarisation à des stimuli améliore la discrimination et si oui, si cette amélioration peut se généraliser à d'autres stimuli. Pour cela, ils étudient des bébés de 8 mois en les familiarisant à des stimuli appartenant à un continuum de VOT [da]-[ta] ou [ga]-[ka]. Les bébés sont séparés en deux groupes : un groupe « unimodal », dans lequel ils perçoivent davantage des stimuli au centre

4. Le langage est selon cette tradition un « comportement » qui s'apprend par une suite d'expériences et de renforcement (voir, par exemple, Skinner, 1957).

5. Les bébés possèderaient selon ce cadre théorique des connaissances linguistiques innées (voir, par exemple, Chomsky, 1959).

du continuum, un groupe « bimodal » de bébés qui perçoivent davantage des stimuli aux extrémités du continuum. Un groupe contrôle, testé sur des stimuli en dehors de ce continuum, sert de groupe de référence. Par la suite, avec la technique HT, ils testent les bébés sur leur discrimination des stimuli bimodaux. Ils montrent, d'une part, que ceux entrainés dans le groupe « bimodal » discriminent mieux que ceux des autres groupes mais qu'en plus, les bébés de ce même groupe, familiarisés avec les stimuli [da]-[ta], discriminent également mieux les stimuli [ga]-[ka] et réciproquement. Ils en déduisent que l'apprentissage statistique permet de mieux se focaliser sur les stimuli entendus et que cet apprentissage concerne des caractéristiques invariantes communes à diverses catégories phonétiques.

Si cet apprentissage statistique semble efficace, on peut se demander s'il permet d'acquérir et de distinguer les allophones des phonèmes. Les allophones correspondent aux différentes variantes d'un même phonème, ceci dans une langue donnée. Il y a deux sortes d'allophones, ceux dont les variations s'échangent librement et ceux qui sont utilisés de façon complémentaire dans certains contextes. Par exemple, [r] et [l] sont des phonèmes en français puisqu'on trouve des mots comme « bar » et « bal » où les deux contrastes permettent de différencier les mots. En revanche, [ʀ] et [ʁ] sont deux allophones du premier type du phonème [r] en français et peuvent être utilisés indifféremment l'un ou l'autre, tout comme [r] et [l] en japonais. De même, les consonnes [t] et [t^h] sont deux allophones anglais mais du second type car ils s'utilisent dans des contextes différents. La question est alors de savoir si les bébés sont capables d'apprendre à associer deux allophones d'un même phonème à une seule catégorie phonémique. Les distributions des occurrences des successions de sons dans la langue fournissent en effet des informations qui permettent, en théorie, d'apprendre la relation entre phonème et allophone Peperkamp et al. (2006). Si les allophones libres semblent pouvoir être fusionnés assez vite en une seule classe grâce à l'apprentissage statistique, comme le montre le déclin des contrastes non-natifs (Werker, 1994), l'apprentissage des allophones du second type est questionnable. En utilisant une grammaire artificielle, White et al. (2008) testent si les bébés de 8 et 12 mois peuvent faire la distinction entre des contrastes allophoniques contraints. Ils montrent que les bébés de 8 mois distinguent les deux contrastes allophoniques comme des phonèmes différents tandis que ceux de 12 mois, bien que percevant le contraste allophonique, apprennent à lui associer une unique catégorie. Ainsi, les bébés semblent avoir la capacité de faire la différence entre phonèmes et allophones durant leur apprentissage (voir aussi Seidl et Cristia, 2012, pour une revue).

L'acquisition de ce second type d'allophones fait le lien avec un autre type d'acquisition statistique, celui concernant les combinaisons de phonèmes propres à chaque langue, ce qu'on nomme généralement l'apprentissage phonotactique (Friederici et Wessels, 1993; Jusczyk et al., 1993; Jusczyk et Luce, 1994; Saffran, 2003). Par exemple, dans Jusczyk et Luce (1994), des bébés américains de 6 mois et 9 mois sont testés sur leur préférences phonotactiques. S'appuyant sur le fait que les bébés écoutent plus longtemps les stimuli avec lesquels ils sont familiers [6], ils utilisent la méthode HT sur deux listes d'items : l'une avec des combinaisons de phonèmes peu probables et l'autre très probables. Ils observent que les bébés de 9 mois, mais pas ceux de 6 mois, écoutent plus longtemps la liste avec les combinaisons hautement probables. Ils en déduisent que les bébés apprennent des éléments de la phonotactique de leur langue.

6. Nous avons relaté précédemment une étude qui s'appuyait sur l'hypothèse inverse selon laquelle les bébés s'orientent vers des stimuli nouveaux. Ces hypothèses sont souvent débattues dans la littérature du développement, mais nous ne les traitons pas dans cette thèse. Nous donnons juste les hypothèses de départ des études considérées.

Néanmoins, les prouesses de l'apprentissage statistique auditif, qu'elles soient phonémiques, allophoniques ou phonotactiques, nécessitent d'être nuancées ou remises en contexte. D'abord, l'apprentissage de certains contrastes phonétiques peut être influencé par d'autres modalités, par exemple visuelle, ou par des effets de contexte, qu'ils soient associés aux contraintes lexicales ou à d'autres facteurs cognitifs (Conboy et al., 2008b; Teinonen et al., 2008; Yeung et Werker, 2009). À titre d'illustration, Teinonen et al. (2008) testent l'apport de la modalité visuelle. Ils entraînent des bébés de 6 mois sur un continuum [ba-da] en suivant une distribution unimodale et analysent deux conditions visuelles. Dans la première, les bébés sont familiarisés sur des stimuli bimodaux dont l'articulation visuelle correspond à la syllabe entendue (le milieu du continuum étant pris comme point de séparation). Dans la seconde, les bébés sont familiarisés à des stimuli bimodaux dans lesquels l'articulation visuelle, quel que soit le stimulus acoustique, est celle de la syllabe [ba] pour un premier sous-groupe ou de la syllabe [da] dans un second sous-groupe. Dans la phase de test, ils observent que seuls les bébés du premier groupe perçoivent les contrastes [ba-da], montrant ainsi que la modalité visuelle peut conditionner l'efficacité et les résultats de l'apprentissage.

Par ailleurs, cet apprentissage ne semble avoir lieu qu'en cas d'interaction sociale (Kuhl et al., 2003). Dans leur expérimentation, Kuhl et al. (2003) testent l'apprentissage de contrastes en chinois mandarins chez des bébés américains de 9 mois dans deux conditions : soit les bébés sont directement exposés à des locuteurs chinois, soit ils écoutent et regardent une vidéo de locuteurs chinois, présentées sur un écran. Les auteurs montrent que seuls les bébés dans la première condition apprennent à discriminer les contrastes chinois, mettant ainsi en évidence l'importance de l'interaction sociale.

En résumé, le processus de perception semble être en grande partie dû à un apprentissage statistique qui permet non seulement de se focaliser sur les contrastes et spécificités propres à sa langue mais également de perdre la discrimination des contrastes qui ne sont pas utiles dans sa langue natale.

3.2.2.2 Mécanismes de l'apprentissage en production et points communs avec la perception

Comme nous venons de le voir, l'environnement acoustique dans lequel baigne le bébé dès le plus jeune âge influence le développement de sa perception. De la même façon, il joue un rôle également dans le développement de sa production. Cette influence de l'environnement se remarque dès les premiers jours de vie lors desquels il est observé que le bébé crie en suivant la mélodie de sa langue (Mampe et al., 2009). L'interprétation donnée est que l'influence de l'environnement sur la production commence in utero. À ce stade, cela semble concerner majoritairement les traits prosodiques. Lors du développement, la focalisation sur les caractéristiques prosodiques continue (par exemple, de Boysson-Bardies et al., 1984) mais celle sur les catégories phonétiques propres à la langue se remarque également.

Dans une première étude, de Boysson-Bardies et al. (1989) analysent les productions des voyelles de bébés de 10 mois ayant pour langue native le français, l'anglais, le cantonais ou l'arabe. Ils mettent en évidence des différences significatives entre les productions des bébés n'ayant pas la même langue native. Par la suite, une expérimentation similaire par de Boysson-Bardies et Vihman (1991) est effectuée sur des bébés ayant pour langue native le français, l'anglais, le japonais et le suédois. Des enregistrements sont effectués à partir de 9 mois pendant la phase de babillage et jusqu'à la produc-

tion des 25 premiers mots. Cette fois-ci, les auteurs montrent que les consonnes sont significativement différentes entre les bébés de différentes langues natives. Si l'apprentissage statistique explique la focalisation de l'apprentissage perceptif sur les sons de l'environnement, comment expliquer cette apparente focalisation en production ?

Dès la naissance, les bébés sont motivés par l'interaction sociale (Bloom, 1975; Kuhl, 2007; Stark, 1980). Nous avons vu l'importance de cette interaction en perception et il semblerait qu'elle ait également une influence sur la production. Par exemple, les bébés semblent produire davantage de sons de parole lorsque cette interaction respecte les « normes » d'une conversation. Dans cette optique, Bloom et al. (1987) testent les productions de bébés anglais de 0 à 3 mois. Dans une première phase, ils enregistrent les bébés avec et sans interaction avec un adulte. Durant l'interaction, ils mettent en place deux situations : soit les tours de parole sont respectés, c'est-à-dire que l'adulte répond après chaque production du bébé, soit les réponses de l'adulte s'effectuent selon un scénario préparé sans prendre en compte les productions du bébé. Dans une seconde phase, d'autres participants doivent ensuite définir si les enregistrements entendus correspondent selon eux à des sons de parole ou non. Les auteurs montrent que l'interaction engendre davantage de sons de parole. Ainsi, l'environnement dans lequel se trouve le bébé semble influencer directement sur sa production (voir aussi Bloom, 1988; Kuhl et Meltzoff, 1982; Masataka, 1993). Pour expliquer cet effet, Bloom (1988) suppose que l'interaction donne l'opportunité au bébé d'imiter les productions de l'adulte.

Cette hypothèse est renforcée par le fait que les bébés sont capables d'imitation dès les premiers jours de vie (Field et al., 1983; Meltzoff et Moore, 1977; Vinter, 1986). Plus radicalement, il est supposé que le bébé acquiert ses capacités de production en partie grâce à l'imitation, ce que Kuhl et Meltzoff (1996) résument sous le nom « d'apprentissage vocal » (vocal learning). Comme les résultats de Mampe et al. (2009), cette imitation semble concerner en grande partie les traits prosodiques (Kessen et al., 1979; Papoušek et Papoušek, 1981). Cependant, Kuhl et Meltzoff (1996) trouvent également une imitation pour les catégories phonétiques. Ils observent notamment que les bébés âgés entre 3 et 5 mois, écoutant une voyelle particulière parmi [i, a, u], produisent davantage de vocalisations ressemblant à cette voyelle.

Malgré tout, cette hypothèse imitative n'est pas retenue par tous. Avec des études similaires à celle de Bloom et al. (1987) sur des bébés de 7 à 10 mois capables de babiller, Goldstein et collègues observent que le babillage des bébés n'est pas une imitation de ce que produisent les adultes, même si les productions augmentent significativement avec l'interaction sociale (Goldstein et al., 2003; Goldstein et Schwade, 2008). Les auteurs supposent plutôt que les interactions permettent aux bébés de découvrir les régularités statistiques de la production des adultes, ce qui leur permet de guider le développement de leur propre production.

Si on suit cette seconde hypothèse, la statistique de l'environnement et l'interaction sociale permettraient de guider également la production tout comme elles guident la perception. Ainsi les deux processus, malgré leurs différences, se serviraient de mécanismes similaires pour leur apprentissage. Nous pouvons donc supposer en conséquence que le lien phonétique entre perception et production pourrait apparaître dès le plus jeune âge. C'est effectivement ce que semblent montrer au moins deux études sur le sujet (voir aussi Munson et al., 2011; Polka et al., 2007, pour des revues).

Dans la première, Vihman et Nakai (2003) testent la perception et la production de bébés anglais

et gallois. Dans une première phase, ils enregistrent deux fois par mois des bébés anglais et gallois âgés entre 10,5 et 12 mois durant des sessions d'une demi-heure où les bébés interagissent avec leur mère. À 12 mois et demi, ils les testent sur leur capacité à discriminer les contrastes [t] et [s] pour les bébés anglais et [b] et [g] pour les bébés gallois en utilisant la méthode HT. Les auteurs observent que les temps d'écoute de ces contrastes sont corrélés inversement à la capacité de production de ces mêmes contrastes. Dit autrement, les bébés écoutent les contrastes plus longtemps lorsqu'ils les produisent le moins souvent. Cette corrélation confirme l'existence d'un lien entre les représentations en production et en perception.

Dans une expérimentation similaire, DePaolis et al. (2011) testent la production et la perception de bébés anglais âgés entre 9 et 11 mois. Dans une première phase, ils enregistrent des bébés dans plusieurs périodes d'une demi-heure lors d'interaction avec leurs parents. De ces enregistrements, ils analysent les consonnes produites et, pour chaque bébé, ils les séparent en trois groupes : fréquemment produites (groupe « own »), peu produites par le bébé mais fréquemment produites à cet âge pour d'autres bébés (groupe « other »), et rarement produites (groupe « rare »). Dans une tâche de perception, en utilisant la méthode HT, ils comparent les enfants disposant d'une unique consonne « own » de ceux en disposant de plusieurs. Ils observent une tendance à ce que les enfants ayant une unique consonne « own » écoutent plus longtemps les contrastes « own ». Mais, de manière significative, les enfants disposant de plusieurs consonnes « own » préfèrent écouter les contrastes « other ». Aucune préférence n'est montrée pour les contrastes « rare ». Ils en déduisent que la production a une influence sur les préférences perceptives.

En résumé, bien que les mécanismes exacts fassent encore débat, la production paraît, elle aussi, influencée par l'environnement. Cela laisse supposer que les processus de perception et production phonétiques sont liés, ce que semblent confirmer les données expérimentales.

3.2.2.3 Conclusion

Pour synthétiser, les deux processus de perception et de production semblent tous deux se focaliser sur les stimuli de l'environnement. Le processus de perception paraît essentiellement basé sur un apprentissage statistique des stimuli de l'environnement. De son côté, le processus de production, bien que plus tardif, semble être influencé par l'environnement, à travers l'utilisation de l'interaction sociale.

Nous nous servons de ces observations dans nos futures modélisations, notamment pour réaliser le développement auditif et moteur de notre modèle.

3.2.3 La structure des unités distinctives

Nous avons étudié le développement des représentations motrices et sensorielles chez le bébé en nous appuyant sur des études sur le développement de la perception et de la production. Dans celles-ci, nous avons vu que le bébé semble apprendre progressivement les représentations sensorielles et motrices correspondant aux unités distinctives propres à sa langue. Néanmoins, il subsiste deux

questions : 1) les unités phonétiques elles-mêmes sont-elles innées ou acquises ? 2) Ces unités sont-elles syllabiques ou phonémiques ?

Ce sont les deux questions auxquelles nous nous intéressons dans cette section à travers trois parties. Nous étudions d'abord si les unités sont innées ou acquises. Ensuite, nous étudions les études présentant la syllabe comme unité de base. Nous terminons sur le développement phonémique.

3.2.3.1 Des unités phonétiques innées ou acquises ?

Du fait que le bébé perçoive dès les premiers mois la majorité des contrastes phonétiques, même ceux ne faisant pas partie de sa langue native, il semble intéressant de se demander s'il possède dès la naissance un bagage phonétique universel. Comme le rappelle Peperkamp (2003), il y a plusieurs avis sur le sujet. Une des théories est de supposer que les frontières entre les contrastes phonétiques correspondent à des changements acoustiques et auditifs généraux permettant au bébé de les discriminer dès la naissance (Kuhl, 2000). Ainsi, les catégories phonétiques seraient naturellement séparées les unes des autres. Cette hypothèse s'appuie notamment sur le fait que certains contrastes phonétiques sont discriminables par plusieurs animaux, tels que les singes ou les chinchillas (Kuhl et Miller, 1975; Kuhl et Padden, 1983). L'apprentissage se ferait par la suite par focalisation sur certaines de ces catégories grâce à l'apprentissage statistique. Cependant, comme le mentionne Pierrehumbert (2003), les catégories phonétiques sont spécifiques à chaque langue et l'hypothèse des frontières catégorielles universelles ne semble pas viable. La discrimination des contrastes phonétiques se ferait uniquement grâce à un apprentissage statistique sans frontières préalables. Cependant, il est possible de trouver un compromis entre les deux théories. Ainsi, Kuhl (2004) propose que les catégories sont bien au départ universelles mais qu'elles sont « primitives » et que le bébé se focaliserait par la suite, durant l'apprentissage, sur les frontières propres à sa langue. Ainsi, selon cette théorie, les frontières universelles serviraient d'amorce aux frontières définitives, spécifiques à chaque langue, acquises durant l'apprentissage.

Par ailleurs, on peut se demander si cette discrimination précoce est phonétique ou simplement auditive. Les expérimentations sur cette question étudient plus en détail le cerveau du bébé en utilisant des techniques d'enregistrement neuronal. Par exemple, l'expérimentation de Dehaene-Lambertz et Baillet (1998) montre que les bébés de 3 mois ont une représentation neuronale similaire à celle de l'adulte. Pour cela, ils enregistrent les potentiels évoqués lors d'un changement purement acoustique et d'un changement phonétique. Ils observent que, bien que les deux changements correspondent à un changement acoustique de même amplitude, les réponses électrophysiologiques sont plus élevées lors du changement phonétique. Ils en déduisent que les bébés discriminent les contrastes de façon phonétique et pas simplement de façon auditive. Une expérimentation similaire par Dehaene-Lambertz et Peña (2001) confirme ces résultats.

Cependant, cela ne signifie pas pour autant que les stimuli sont traités comme ceux de l'adulte. Par exemple, dans les données d'imagerie par magnétoencéphalographie de Kuhl et al. (2014), les auteurs observent les zones d'activation des bébés de 7 et 11-12 mois. Ils montrent qu'à 7 mois les aires auditives et motrices s'activent de façon équivalente aussi bien pour les contrastes natifs que non natifs. Cependant, à 11-12 mois, ils observent que les aires auditives s'activent davantage pour les

contrastes natifs et les aires motrices s'activent davantage pour les contrastes non-natifs. Ce dernier comportement correspond à ce qui est obtenu chez l'adulte contrairement au premier et laisse donc supposer que les bébés de moins de 7 mois ne réagissent pas aux stimuli de la même manière que les adultes. Par ailleurs, l'expérimentation de Cheour et al. (1998) montre qu'à 6 mois, les contrastes mesurés par l'amplitude de la réponse MMN respectent une hiérarchie auditive et non phonétique (plus d'amplitude pour des stimuli plus distincts, qu'ils fassent partie de la langue ou pas) alors qu'à 12 mois la hiérarchie devient phonétique et conforme à celle des adultes (plus d'amplitude pour des stimuli correspondant à deux catégories différentes, même s'ils sont acoustiquement proches).

En résumé, il reste encore difficile de savoir si les unités phonétiques sont présentes dès la naissance ou non. Il semblerait que le bébé soit guidé d'abord par des contrastes acoustiques qui deviennent très rapidement phonétiques, c'est-à-dire interprétés en lien avec la langue. Ainsi, même si les unités phonétiques ne correspondent pas au départ à celles de l'adulte, le bébé possède les mécanismes permettant de les acquérir.

3.2.3.2 Une structure cognitive principalement syllabique

La majorité des études sur le développement de la perception considère un codage phonémique en étudiant les contrastes consonantiques, comme l'illustrent par exemple les études sur la perception catégorielle. Néanmoins, du fait que les études sont réalisées plus généralement avec des syllabes, on pourrait également supposer que le bébé discrimine en réalité des syllabes. De plus, les recherches sur le développement et notamment celles sur le babillage semblent plutôt se centrer sur une structure syllabique.

Dans leur revue sur les représentations mentales des unités de parole durant l'acquisition, Hallé et Cristia (2012) considèrent que l'unité phonétique de base en perception est la syllabe, ce qu'ils affirment très clairement : « Young children code speech in terms of syllables ». Cette hypothèse est également soutenue par d'autres auteurs (Jusczyk, 1997; Mehler et Hayes, 1981), citant divers résultats comme arguments.

Parmi ceux les plus cités, les bébés posséderaient presque de façon innée la capacité de compter les syllabes plutôt que les phonèmes dans un énoncé. C'est ce que testent Bijeljac-Babic et al. (1993). En utilisant la méthode HAS, ils observent d'abord que les bébés de quatre jours savent discriminer un ensemble bisyllabique CVCV (par exemple, [rifu, kepa]) d'un ensemble trisyllabique CVCVCV (par exemple, [mazopu, rekiva...]) et inversement. Ils testent ensuite la discrimination d'ensembles bisyllabiques CVCV composés soit de 4 phonèmes (par exemple, [rifu, kepa]) , soit de 6 phonèmes (par exemple, [treklu, suldri]). Cependant, ils n'observent aucune déshabituation entre ces deux ensembles. Ils supposent donc que les bébés détectent les changements du nombre de syllabes mais pas ceux modifiant le nombre de phonèmes. Ils interprètent ceci en faveur d'un codage syllabique.

À l'aide d'une expérimentation voisine, Bertoncini et Mehler (1981) proposent une comparaison de stimuli syllabiques et non syllabiques. Dans leur étude, ils testent des bébés français de deux mois avec la méthode HAS sur leur capacité à distinguer des stimuli ayant une structure syllabique correcte (par exemple, [tap] vs. [pat]) ou des stimuli ne respectant pas cette structure syllabique (par exemple, [tʃp]-[pʃt]). Ils montrent que les bébés discriminent les premiers mais pas les seconds. Par ailleurs,

les bébés sont également capables de discriminer ces secondes structures lorsqu'elles sont comprises dans une structure syllabique (par exemple, [utʃpu]-[upʃtu]). Du fait que seuls les stimuli syllabiques sont discriminés, ils en déduisent que les bébés s'appuient sur un traitement syllabique.

Cependant, comme le précisent Bertoncini et al. (1988), ces deux résultats pourraient également s'expliquer par le fait que les bébés sont sensibles à la prosodie de leur langue et reconnaîtraient les ensembles rythmiques sans avoir pour autant un codage syllabique. Néanmoins, une comparaison plus directe de la discrimination phonémique et syllabique semble tout de même jouer en faveur de la syllabe. Par exemple, Jusczyk et Derrah (1987) proposent un protocole où des bébés de deux mois sont familiarisés à un ensemble de quatre syllabes CV partageant la même consonne C ([bi, bo, ba, bɔ̃]). Avec la méthode HAS, ils testent ensuite trois stimuli différents : 1) un stimulus possédant une voyelle V différente mais la même consonne C ([bu]), 2) un stimulus possédant une consonne C différente mais une voyelle V commune ([di]) et 3) un stimulus avec une nouvelle consonne C et une nouvelle voyelle V ([du]). Ils remarquent que les bébés se déshabituent dans les trois conditions. Cela va à l'encontre de l'hypothèse d'une unité phonémique. En effet, selon cette hypothèse, le bébé ne devrait pas réagir face au stimulus 1) (ou moins réagir) puisque la consonne [b] est commune aux stimuli avec lesquels il a été habitué. Or, le fait qu'il se déshabitue identiquement laisse penser qu'il ne remarque pas que la consonne [b] est commune aux stimuli avec lesquels il a été habitué et donc qu'il possède davantage un traitement syllabique où chaque syllabe est perçue indépendamment. Cette expérimentation est répliquée par Bertoncini et al. (1988). Par ailleurs, ces auteurs réalisent également le même type d'expérimentation en habituant les bébés de deux mois avec des syllabes CV partageant une même voyelle ([bi, si, li, mi]). En testant les stimuli avec 1) une nouvelle voyelle V ([ba]), 2) une nouvelle consonne C ([di]) ou 3) une nouvelle consonne C et voyelle V ([da]), ils obtiennent également une déshabituation pour les trois stimuli.

En production, les principales théories semblent également plutôt en faveur d'un codage initial syllabique. Du fait que les bébés ne savent produire des énoncés qu'à partir du babillage, les auteurs se focalisent surtout sur cette période. La question principale est de savoir si le babillage se fait en termes d'un codage phonémique associant une consonne C et une voyelle V ou en termes d'un codage syllabique global CV. La théorie Frame Then Content (MacNeilage et al., 1997, voir également section 3.2.1.2) privilégie la seconde option dans les stades initiaux du babillage (Hallé et Cristia, 2012, pour une revue)

En résumé, la majorité des études de perception, surtout comportementales, convergent vers l'hypothèse d'une primauté de l'unité syllabique dans les premiers stades du développement. Cette question ne semble pas vraiment avoir été étudiée actuellement par les études en neurosciences.

3.2.3.3 Le développement phonémique

Il y a assez peu d'études pour contre-argumenter et considérer plutôt un codage phonémique précoce chez le bébé. En effet, outre les travaux défendant le codage syllabique, un des arguments principaux contre un codage phonémique vient du fait que l'enfant apprend à manipuler consciemment les phonèmes assez tardivement par rapport à la syllabe (voir le concept de « phonological/phoneme awareness », Carroll et al., 2003; Fowler et al., 1991; Mann et Wimmer, 2002; Ziegler et Goswami,

2005).

Parmi les exceptions, on trouve les études sur l'apprentissage des voyelles (Kuhl et al., 1992; Polka et Werker, 1994). Par exemple, dans l'étude de Kuhl et al. (1992), des bébés américains et suédois de 6 mois sont testés sur leur discrimination du contraste [i], prototypique de l'anglais et [y], prototypique du suédois. Pour réaliser leur expérimentation, les auteurs se basent sur le fait que la discrimination des voyelles n'est pas catégorielle comme pour les consonnes mais que la discrimination est moins bonne entre une voyelle prototypique et une voyelle proche qu'entre deux voyelles non prototypiques. Cela est nommé « effet magnet ». En utilisant la technique HT, la discrimination des prototypes suédois et anglais est évaluée par rapport à celle d'autres voyelles proches. Les résultats montrent que les bébés américains discriminent mieux le prototype anglais et que les bébés suédois discriminent mieux le prototype suédois. Ils en déduisent que l'apprentissage perceptif améliore la perception des prototypes natifs, ce qui laisse supposer que les bébés apprennent bien les phonèmes de leur langue et ce, avec des effets de « perceptual narrowing » sur les voyelles dès 6 mois. A cette même période, les auteurs d'une étude de pupillométrie montrent que les bébés semblent pouvoir reconnaître une même consonne couplée avec différentes voyelles. En conséquence, ils supposent que les bébés de 6 mois seraient capables de retrouver l'invariant consonantique (Hochmann et Papeo, 2014).

Dans une expérimentation similaire en neurosciences, Cheour et al. (1998), comme nous l'avons vu précédemment (voir section 3.1.3.2), montrent la présence de traces phonémiques, spécifiques à chaque langue, dès 12 mois, ce qui reste assez tardif par rapport à la syllabe et par rapport au résultat précédent. Dans leur expérimentation, ils observent qu'à 12 mois, le bébé finlandais réagit davantage au contraste finnois qu'estonien et inversement, jouant en faveur d'un codage phonémique au moins vocalique.

Ainsi, il semble bien que s'opère au cours du développement, tant en production qu'en perception, un passage progressif de la syllabe, unité primaire et disponible dès les tous premiers temps de l'acquisition, vers le phonème, unité émergente acquise éventuellement plus tardivement, selon des processus qui restent encore largement à définir.

3.2.3.4 Conclusion

Pour synthétiser, nous avons d'abord observé que les processus de perception phonétique associés à la syllabe semblent être grossièrement similaires à ceux de l'adulte dès deux mois. De leur côté, les unités phonémiques ne semblent être réellement acquises que plus tardivement. Nous reviendrons, dans le chapitre 6, sur cette structure cognitive en présentant un modèle permettant d'étudier l'acquisition conjointe des syllabes et des phonèmes.

CHAPITRE 4

Modélisation des unités distinctives et présentation du modèle COSMO

Dans le chapitre précédent, nous nous sommes consacrés à l'analyse de la caractérisation et du développement des unités phonétiques à travers la revue de diverses études sur la parole. Celles-ci sont principalement basées sur des expérimentations sur l'humain, qu'elles soient comportementales ou neuroscientifiques. Dans toute la suite de cette thèse, nous nous focalisons toujours sur les unités phonétiques, mais en faisant appel à un tout autre domaine : la modélisation computationnelle.

Les modèles computationnels sont des outils adaptés pour tester des hypothèses sur les mécanismes internes d'un phénonème et semblent, en ce sens, appropriés pour tenter de mieux comprendre comment fonctionne le cerveau humain. La modélisation computationnelle s'applique aussi à la phonétique et de nombreux modèles ont été créés pour en comprendre les différents aspects.

Dans ce chapitre, en nous appuyant sur les faits et théories du chapitre précédent, nous décrivons, dans un premier temps, plusieurs modèles computationnels s'étant intéressés, directement ou non, aux unités phonétiques. Dans un second temps, nous présentons le modèle COSMO, que nous utilisons dans nos simulations.

4.1 Les modèles computationnels étudiant les unités phonétiques

L'ensemble de cette section suit le plan du chapitre précédent et se base ainsi sur les deux questions suivantes : 1) comment les unités phonétiques sont-elles caractérisées dans les modèles computationnels ? 2) Comment s'effectue le développement des unités phonétiques dans les modèles computationnels ? Nous tentons d'apporter des éléments de réponses dans les deux prochaines sous-sections.

Nous commençons par aborder les modèles computationnels phonétiques de perception et de production avant de nous focaliser spécifiquement sur les modèles s'intéressant au développement des unités phonétiques.

4.1.1 Comment les unités distinctives sont-elles caractérisées dans les modèles ?

Les théories auditives, motrices et perceptuo-motrices de la perception ont été longuement débattues afin de découvrir la nature des unités phonétiques. Si les récentes études en neuroimagerie

semblent montrer une activation commune des aires sensorielles et motrices durant la perception, le rôle exact de ces deux ensembles d'aires est toujours en discussion. La modélisation pourrait être un bon moyen d'analyser ce problème. C'est pourquoi, dans un premier temps, nous nous focalisons sur les modèles de perception. Deux aspects de cette thématique sont développés : d'une part, la nature des représentations utilisées dans les modèles de perception et, d'autre part, les résultats obtenus par les modèles s'étant particulièrement centrés sur cet enjeu perceptuo-moteur.

Une fois la perception étudiée, nous examinons le lien entre les représentations en perception et en production phonétiques. Celui-ci ayant été démontré dans diverses études, nous souhaitons analyser comment les modèles computationnels le prennent en compte. Partant du constat que les modèles de perception implémentent très rarement un mécanisme de production, nous nous intéressons, dans un second temps, aux quelques modèles de production ayant un lien perception/production. Sans pour autant les décrire dans leur globalité, nous étudions spécifiquement comment est implémenté ce lien sensorimoteur.

Pour finir, nous nous intéressons à la structure cognitive des unités phonétiques. Le phonème est généralement au cœur des études sur la nature des unités phonétiques. Néanmoins, les études relatives au contenu cognitif de ces unités ont également montré l'importance de l'unité syllabique. Bien que les avancées en neurosciences aient permis de proposer l'existence d'une double structure cognitive phonémique et syllabique, les relations entre les deux types d'unités restent toutefois à approfondir. Les modèles pourraient être un bon moyen d'étudier ces mécanismes. C'est pourquoi, dans un troisième temps, nous analysons comment le lien entre ces deux types d'unités est implémenté dans les modèles computationnels. Nous nous concentrons, d'une part, sur les unités phonétiques utilisées globalement dans les modèles et, d'autre part, sur les modèles s'étant particulièrement intéressés à ce lien.

4.1.1.1 La nature des unités phonétiques dans les modèles de perception

Comme le rappellent McClelland et Elman (1986), il y a généralement deux types de modèles de la perception phonétique. Les premiers concernent la reconnaissance de parole et ont pour but de construire une machine permettant de reconnaître le plus efficacement et le plus précisément possible les différentes unités (voir par exemple Benzeghiba et al., 2007; Sarma et Prasanna, 2017, pour des revues). Bien qu'en majorité focalisés sur les représentations auditives, certains d'entre eux montrent que les représentations motrices peuvent améliorer les performances de reconnaissance (Badino et al., 2014; Kirchhoff, 1998; Zolnay et al., 2005, voir aussi King et al., 2007 pour une revue). Mais ces modèles ayant principalement des préoccupations de performance et d'efficacité et non de réalisme, nous les laissons de côté pour nous focaliser sur le second type de modèles ayant un objectif davantage cognitif et psychologique puisqu'ils cherchent à mieux comprendre comment fonctionne le processus de perception.

Malgré les vifs débats ayant opposé différentes théories de la perception (voir section 3.1.1), une grande partie des modèles cognitifs laissent de côté la question des représentations phonétiques pour ne s'intéresser qu'au traitement des unités linguistiques en tant que tel (voir Scharenborg et Boves, 2010; Weber et Scharenborg, 2012, pour des revues). Ceci s'explique, en partie, par le fait qu'ils sont

avant tout des modèles de la perception générale prenant en compte les unités lexicales. De ce point de vue, le signal acoustique entrant n'est traité réellement ni de manière auditive, ni de manière motrice mais directement de manière linguistique. Dès le début de la perception, il est donc directement décomposé en un nombre fini d'unités discrètes abstraites prélinguistiques. Ces unités prélinguistiques sont généralement définies au niveau phonémique ou au niveau des traits phonétiques. Dans le premier cas, certains modèles se servent par exemple des phonèmes eux-mêmes (McQueen et al., 2000; Norris, 1994; Scharenborg et al., 2005), des allophones (Luce et al., 2000) ou encore des séquences probabilistes de phonèmes (Norris et McQueen, 2008). Dans le second cas, les traits phonétiques, bien qu'ils soient aussi des unités discrètes, prennent différentes formes : il peut s'agir d'unités binaires (Gaskell et Marslen-Wilson, 1997) ou d'un peu plus grande cardinalité (McClelland et Elman, 1986; Scharenborg, 2008), représentant aussi bien des caractéristiques auditives (le voisement) que motrices (le lieu d'articulation). En ce sens, les traits phonétiques se rapprochent des représentations auditives et motrices caractérisant les unités distinctives.

Afin d'illustrer plus clairement ce que nous nommons un modèle linguistique, la Fig. 4.1 montre le modèle MERGE (Norris et al., 2000). Dans ce modèle, les représentations linguistiques sont réparties en trois couches de réseaux de nœuds, chaque nœud correspondant à une unité linguistique. La première couche, l'input, est, par commodité, représentée par des unités phonémiques à reconnaître par le modèle. Les deux autres couches sont les niveaux de décision qui correspondent aux unités stockées dans le modèle. Il y a un niveau phonémique pour reconnaître les phonèmes et un niveau lexical pour reconnaître les mots.

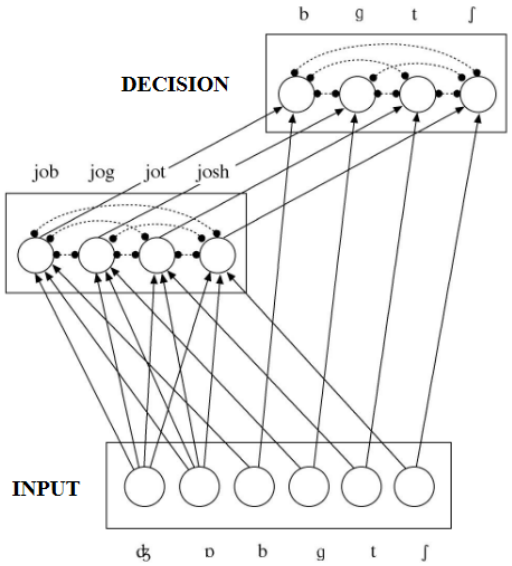


FIGURE 4.1 – Schéma du modèle MERGE. Issu de McQueen et al. (2000), similaire à la figure correspondante dans Norris et al. (2000)

Parmi les modèles s'intéressant à la nature des représentations des unités phonétiques, la plupart n'utilise que des représentations auditives (Clayards et al., 2008; Klatt, 1980; Kleinschmidt et Jaeger, 2011, 2015). Celles-ci peuvent aussi bien être des paramètres acoustiques prédéfinis (Voice Onset Time ou formants) que des représentations générales (séquences de spectres auditifs). Mais,

même parmi ces modèles, les recherches sur la nature exacte des unités et de l'invariant phonétique ne sont pas toujours entièrement développées. En effet, certains auteurs étant focalisés sur une problématique phonétique précise ne définissent pas les invariants phonétiques dans leur intégralité mais se concentrent uniquement sur certains contrastes leur permettant de tester et d'illustrer leurs hypothèses. C'est, par exemple, le cas de Kleinschmidt et Jaeger (2011) qui s'intéressent à l'adaptation phonétique et qui ne définissent que les invariants leur permettant de manipuler les contrastes consonantiques [b] et [d]. Pour ceux dont la problématique est plus générale, la question de l'invariant a plus d'importance. C'est par exemple le cas de Klatt (1980) qui se sert des caractéristiques auditives des diphones, c'est-à-dire de la succession des noyaux de deux phonèmes consécutifs, pour pouvoir catégoriser les phonèmes.

Si la majorité des modèles possèdent seulement des représentations auditives, il existe néanmoins quelques exceptions prenant en compte les représentations motrices. Citons deux modèles. Le premier, développé par l'équipe de Fadiga est, à notre connaissance, le seul modèle de perception s'intéressant spécifiquement au rôle des représentations motrices en perception (Badino et al., 2016; Canevari et al., 2013; Castellini et al., 2011). Le second, développé par l'équipe de Kröger, est, à notre connaissance, le seul à proposer un modèle cognitif global sensorimoteur (Eckers et al., 2013; Kröger et al., 2011; Kröger et Cao, 2015; Kröger et al., 2014, 2009). Néanmoins, il ne s'agit pas seulement d'un modèle de perception mais d'un modèle couplant perception et production.

Concernant le modèle de l'équipe de Fadiga, il s'agit d'un modèle de reconnaissance phonétique, assez proche des modèles de reconnaissance de parole aussi bien dans le déroulement des simulations effectuées que dans l'analyse des résultats mais qui, néanmoins, se préoccupe de la représentation interne du modèle. Nous l'intégrons donc également parmi les modèles cognitifs. Ce modèle a la particularité de contenir des représentations motrices qui sont utilisées dans le processus de perception. L'étude se concentrant le plus sur l'apport des représentations motrices en perception est celle de Castellini et al. (2011). Dans celle-ci, les auteurs vérifient, dans différentes conditions, si la prise en compte des représentations motrices améliore la catégorisation des consonnes [b-p] versus [d-t]. Le modèle est un réseau de neurones dans lequel les représentations auditives correspondent à des coefficients cepstraux [1] extraits du signal auditif et les représentations motrices correspondent soit à des trajectoires motrices (nommées « real motor »), soit à des positions articulatoires reconstruites à partir du signal auditif (nommées « reconstructed motor »). La discrimination des phonèmes s'effectue selon quatre conditions : une pour chacune des trois représentations, auditive et motrices, prises séparément et une mélangeant les représentations auditives et les représentations motrices « reconstructed motor ». La simulation se passe en deux phases : une phase d'entrainement dans laquelle le modèle est entrainé à reconnaître différentes unités phonétiques sur des signaux donnés et une phase de test dans laquelle sont testées les performances du modèle sur d'autres signaux. En réalisant différentes conditions d'entrainement et de test en termes de locuteurs et d'unités, ils montrent que l'utilisation des représentations motrices « real motor » donnent, dans chaque condition, les meilleures performances pour reconnaître les consonnes [b-p] versus [d-t] (< 8% d'erreur). Ils observent également que les représentations auditives donnent les moins bons taux de catégorisation (entre 6% et 37% d'erreurs pour les cas les plus difficiles) et que les deux autres sont sensiblement, mais significativement, meilleures (entre 5% et 35% d'erreurs pour les cas les plus difficiles). Ils en déduisent que les représentations motrices, au moins pour la discrimination des consonnes, améliorent la catégorisa-

1. Un cepstre est le résultat de la transformée de Fourier inverse du logarithme du spectre estimé d'un signal

tion, quoique très faiblement. Toutefois, les auteurs n'expliquent pas les raisons de ces performances. Il reste donc difficile à comprendre pourquoi le décodage moteur semble meilleur dans cette étude.

De son côté, l'équipe de Kröger propose, dans ses différentes études, un modèle neurocomputationnel global de la perception et de la production en accord avec les processus de perception et de production chez l'humain. Concernant les représentations des unités phonétiques, celles-ci sont codées par des ensembles de neurones regroupés sous le terme de cartes. Dans une de ses versions (Kröger et al., 2011), utilisée pour des simulations de perception, le modèle contient six cartes (voir Fig. 4.2 pour un schéma global du modèle) : deux cartes phonétiques (« phonetic map » et « phonemic map ») composées de phonèmes et syllabes, une carte auditive (« auditory map ») paramétrée par les trois premiers formants du signal acoustique F1, F2 et F3, une carte somatosensorielle (« somatosensory map ») informant sur l'ouverture du conduit vocal et deux cartes motrices (« motor plan » et « primary motor map ») caractérisées par deux paramètres sur le lieu d'articulation et un paramètre sur le mode d'articulation. Cependant, bien que le modèle contienne toutes ces cartes, les auteurs considèrent que la perception phonétique s'effectue uniquement à partir des représentations sensorielles (cartes auditives et somatosensorielles) jusqu'aux représentations linguistiques (cartes phonétiques et phonémiques) mais sans l'utilisation des cartes motrices. Ainsi, bien que le modèle contienne des connaissances sensorielles et motrices et un lien sensorimoteur, la réalisation de la tâche de perception implique, elle, uniquement des connaissances sensorielles.

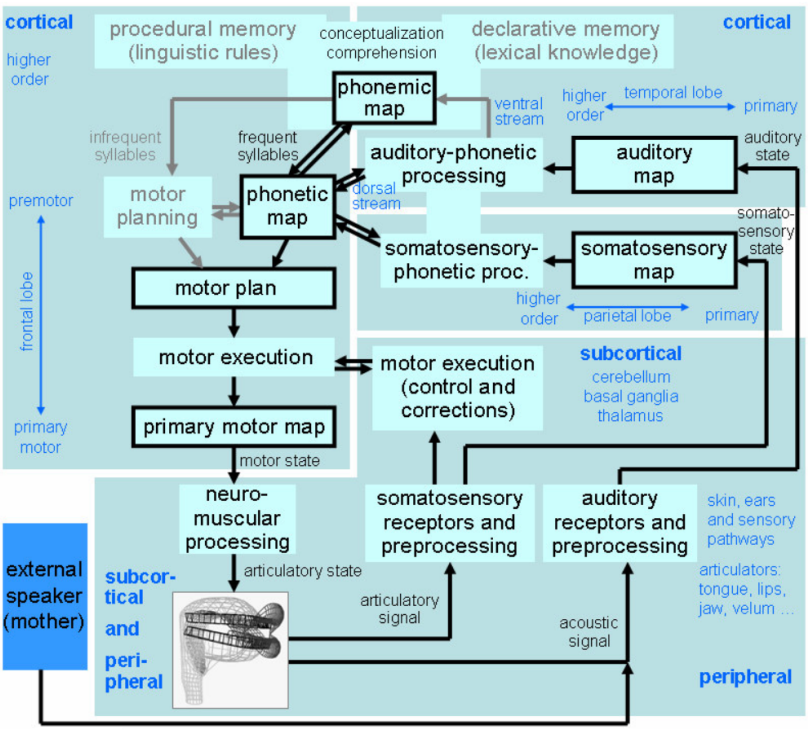


FIGURE 4.2 – Schéma du modèle de perception et de production de Kröger et collègues, issu de Kröger et al. (2011)

En résumé, il y a assez peu de modèles computationnels de perception s'intéressant à la nature des représentations phonétiques et aux invariants phonétiques. Quand les modèles ne sont pas uniquement

linguistiques, la plupart utilisent des représentations auditives sans prendre en compte les potentielles représentations motrices. Parmi ceux faisant figure d'exception, seul le modèle proposé par l'équipe de Fadiga semble réellement se questionner sur l'apport des représentations motrices. Néanmoins, à notre connaissance, aucun de ces modèles n'étudie finalement le rôle exact que jouent les invariants sensoriels et moteurs dans la perception.

4.1.1.2 La nature sensorimotrice des invariants dans les modèles de production

Dans le chapitre précédent, nous avons énuméré plusieurs études traduisant le lien existant entre les invariants phonétiques en perception et en production (voir section 3.1.2). Du côté des modèles de perception phonétique, l'étude des invariants phonétiques reste assez limitée et la plupart des modèles se focalisent sur les représentations auditives. Dans ce contexte, il semble peu aisé d'étudier le lien entre les représentations en perception et production. Du côté des modèles de production phonétique, il en existe davantage proposant des représentations sensorimotrices. Nous avons, par exemple, évoqué précédemment le modèle de l'équipe de Kröger qui, en plus d'être un modèle de perception, est également un modèle de production (Kröger et al., 2009). Cela vient principalement du fait que la perception, notamment la perception auditive de ses propres productions, est jugée importante pour la production. En effet, comme le rappellent Houde et Nagarajan (2011), même si, une fois les gestes moteurs appris et maîtrisés, la production de la parole peut se passer de ces retours auditifs, ils jouent néanmoins un rôle de feedback indispensable dans un certain nombre de cas, tant au niveau phonétique que prosodique. C'est pourquoi, plusieurs modèles de production se sont penchés sur leur implémentation et ont cherché à comprendre comment ils affectent la production. C'est sur ces modèles que nous nous focalisons dans cette partie, afin d'étudier comment les représentations sensorielles et motrices sont liées et comment ces liens influencent le processus de production.

Un des modèles de production les plus connus est le modèle DIVA de Guenther et ses collègues (Guenther, 1995, 2006; Guenther et Vladusich, 2012). Plusieurs modèles, dont celui de Kröger, que nous avons mentionné précédemment, sont basés sur lui. Ce modèle a connu plusieurs versions et améliorations au cours des années. Nous nous concentrons sur la version proposée par Tourville et Guenther (2011) qui décrit assez précisément l'interaction des représentations sensorimotrices en perception et production, schématisée Fig. 4.3. Conformément au fait que le modèle de Kröger est basé sur DIVA, nous y retrouvons la notion de cartes, celles-ci correspondant à des ensembles de neurones. Nous remarquons, pour commencer, que les unités phonétiques sont regroupées dans une carte nommée « Speech Sound Map ». Celle-ci est reliée à trois représentations, elles aussi organisées sous forme de cartes : une carte pour les représentations motrices « Articulatory Velocity and Position Maps », une carte pour les représentations auditives « Auditory Target Map » et une carte pour les représentations somatosensorielles « Somatosensory Target Map ».

Lors de la production, deux systèmes sont actifs : le système de contrôle feedforward, qui génère le geste de production, et le système de contrôle feedback, qui se charge du traitement du retour sensoriel de cette production. À l'aide du système feedforward, une unité phonétique est sélectionnée dans la carte « Speech Sound Map » et générée à l'aide des représentations motrices correspondantes dans la carte « Articulatory Velocity and Position Maps ». En parallèle, à l'aide du système feedback, l'unité phonétique choisie génère une prédiction auditive et somatosensorielle dans les cartes auditives

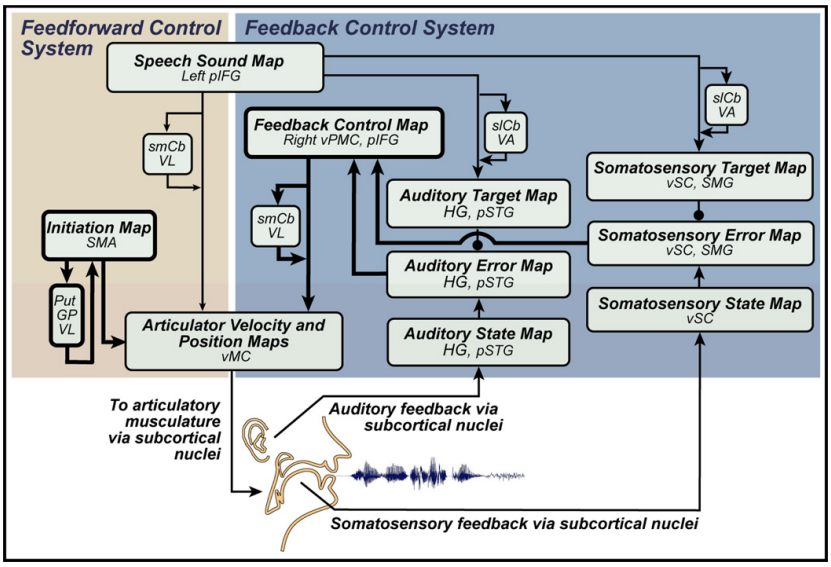


FIGURE 4.3 – Schéma du modèle DIVA. Issu de Tourville et Guenther (2011)

« Auditory Target Map » et « Somatosensory Target Map », prédiction qui est ensuite rétro-activée respectivement dans les cartes « Auditory Target Map » et « Somatosensory Target Map ». Les « error maps » correspondent donc à l'opposé des « target maps », c'est-à-dire que les neurones excités dans les target maps correspondent aux neurones inhibés dans les error maps et réciproquement.

Lors du retour auditif, le système reçoit et traite d'abord le retour sensoriel auditif et somatosensoriel respectivement dans une carte auditive nommée « Auditory State Map » et une carte somatosensorielle nommée « Somatosensory State Map ». Les neurones activés de ces deux « state maps » sont projetés dans les error maps préalablement activées. Ainsi, ces dernières correspondent, après le retour auditif, à la différence entre le stimulus attendu et le stimulus réel. Pour finir, le résultat des erreurs entre le stimulus attendu et le stimulus réel est projeté dans une carte nommée « Feedback Control Map », qui sert, par la suite, à corriger les gestes moteurs activés dans la carte des représentations motrices « Articulatory Velocity and Position Maps ». Ainsi, le lien entre les représentations motrices et les représentations sensorielles ne se fait que lors de la projection de la carte « Feedback Control Map » sur la carte « Articulatory Velocity and Position Maps ». Dans ce modèle, les deux représentations restent séparées l'une de l'autre et ont, chacune, un rôle différent dans le processus de production : les représentations motrices sont utilisées pour produire le geste moteur et les représentations sensorielles sont utilisées pour vérifier qu'il est correct et l'adapter, si besoin.

Supposant que le retour auditif seul n'est pas suffisant pour corriger la dynamique des gestes moteurs, Houde et collègues ont proposé une solution alternative nommée State Feedback Control (Houde et Nagarajan, 2011; Houde et al., 2007). La Fig. 4.4 reprend la version schématisée du modèle de 2007 qui est, à notre sens, la plus intuitive à comprendre. Ce modèle s'intéresse davantage au traitement du retour auditif lui-même qu'au modèle de production et donne, en ce sens, peu de détails sur les implémentations respectives des représentations. Il permet de calculer deux processus pour la production : un processus feedforward, dans lequel, comme dans le modèle de Guenther, une cible de parole est sélectionnée puis produite à l'aide des représentations motrices, et un processus de feedback, qui,

lui, diffère du modèle de Guenther. Celui-ci est défini de la manière suivante. En parallèle du processus feedforward, une copie des représentations motrices est conservée en interne. Elle est nommée copie d'efférence. Du fait qu'il n'est pas réaliste d'avoir directement un retour des gestes moteurs produits, ces représentations motrices correspondent à une estimation des gestes moteurs produits. Cette copie d'efférence est projetée dans un modèle du conduit vocal (voir « internal model of vocal apparatus » sur la figure). Cela permet d'avoir une approximation du geste moteur produit qui, d'une part, est utilisée pour adapter si besoin les représentations motrices futures et, d'autre part, est projetée dans un modèle interne fournissant les représentations sensorielles, et donc une estimation du signal, correspondant à ce geste moteur (voir « internal model of feedback delays » sur la figure). Lors du retour auditif, le signal obtenu est comparé avec le signal estimé et la différence obtenue est ensuite convertie (voir « Kalman gain ») et utilisée afin de modifier les représentations motrices estimées. Celles-ci sont enfin reproduites pour former un nouveau geste moteur, utilisé pour adapter les prochaines productions.

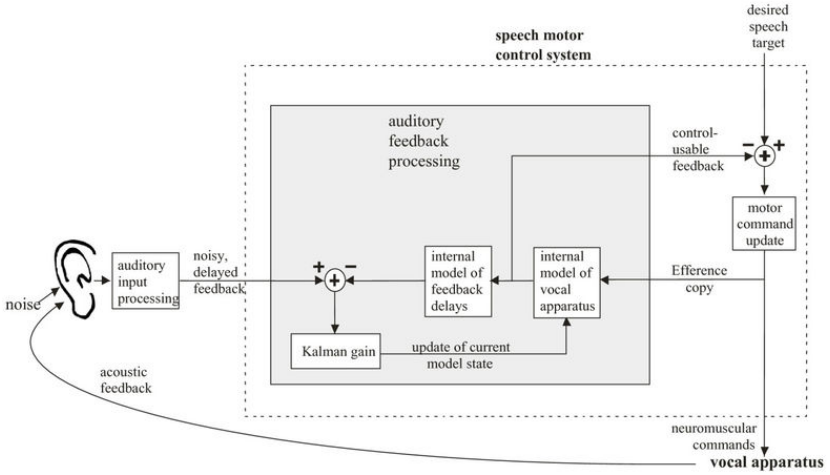


FIGURE 4.4 – Schéma du modèle State Feedback Control. Issu de Houde et al. (2007)

En résumé, ces deux modèles montrent que le retour auditif et le lien entre les représentations sensorielles et motrices peuvent s'effectuer de différentes manières : soit à travers la comparaison entre un stimulus directement estimé et le stimulus réel, comme dans le modèle de Tourville et Guenther (2011), soit à travers la comparaison entre le retour auditif d'une production estimée via un modèle interne et le retour auditif réel, comme dans le modèle de Houde et al. (2007). Le lecteur peut se reporter à Zheng (2012) pour plus de détails sur ces deux modèles. Par ailleurs, bien que le but de cette section est de présenter comment les modèles de production implémentent le lien sensorimoteur, il est important de préciser que tous les modèles de production ne considèrent pas l'existence d'un lien sensorimoteur (voir par exemple Gauvin et al., 2016, pour une revue). Dans certains d'entre eux, le contrôle interne de la validité de la production s'effectue par des processus soit purement moteurs (par exemple Nozari et al., 2011), soit purement sensoriels (par exemple Hartsuiker et Kolk, 2001).

4.1.1.3 La structure cognitive des unités dans les modèles

Comme nous l'avons vu dans le chapitre précédent, la structure cognitive des unités phonétiques a été longuement débattue. Nous nous intéressons, ici, à la façon dont les modèles computationnels l'implémentent. Comme le modèle que nous étudions dans les prochains chapitres est davantage centré sur des problématiques de perception que de production, nous nous concentrons principalement sur les modèles de perception comme point de comparaison. Nous renvoyons le lecteur, par exemple, à Bohland et al. (2010); Levelt (1999); Pierrehumbert (2003); Postma (2000); Schiller (2006); Walker (2016), pour plus de détails sur la structure cognitive des modèles phonétiques de production.

Parmi les modèles computationnels phonétiques de perception, un grand nombre d'entre eux possèdent une structure phonémique. Bien qu'il y en ait certainement davantage, nous nous focalisons sur deux types de modèles : les modèles spécialisés sur les mécanismes de certaines fonctions précises de perception et les modèles de perception globaux. Les premiers sont des modèles qui ne modélisent pas la perception dans son ensemble mais n'étudient que certaines fonctions inhérentes à la perception comme l'adaptation, la normalisation ou la généralisation (Clayards et al., 2008; Kleinschmidt et Jaeger, 2011, 2015; Richter et al., 2016). Généralement, ces modèles se concentrent sur quelques unités phonémiques précises, nécessaires pour illustrer la tâche qu'ils souhaitent effectuer. Il s'agit principalement d'unités catégorielles discrètes, indépendantes les unes des autres. À titre d'illustration, dans Kleinschmidt et Jaeger (2011), les auteurs se concentrent uniquement sur les consonnes [b-d]. Ces deux consonnes font partie d'un ensemble catégoriel discret, noté C, avec lequel ils peuvent manipuler un ensemble de probabilités, telles que la probabilité des unités via la distribution $P(C)$, la probabilité d'une catégorie sachant un stimulus x donné via la distribution $P(C|[X = x])$ ou encore la probabilité des stimuli sachant la catégorie c via la distribution $P(X|[C = c])$.

Les seconds modèles phonémiques, les modèles de perception globaux, implémentent le processus de perception dans son ensemble. Un grand nombre d'entre eux sont des modèles hiérarchiques possédant, d'une part, une structure phonémique et, d'autre part, une structure lexicale composée de mots. Globalement, ces modèles hiérarchiques se séparent en deux familles : ceux avec une structure feedforward et ceux avec une structure interactive (voir par exemple Guediche et al., 2014; Norris et al., 2000, pour le débat sur ces deux types de modèles). Les modèles feedforward sont les modèles implémentant une structure hiérarchique dans laquelle la perception s'effectue dans un sens unique, allant du son aux unités lexicales, en passant par les unités phonémiques, si nécessaires. Les unités phonémiques sont donc traitées à partir du son, et seulement du son. Plusieurs modèles de perception assez connus font partie de cette famille, notamment ceux développés par Norris et collègues tels que le modèle Race (Cutler et Norris, 1979), le modèle MERGE (Norris et al., 2000), cité précédemment, ou encore le modèle ShortList (Norris, 1994; Norris et McQueen, 2008). Le modèle le plus récent, ShortListB (Norris et McQueen, 2008) est un modèle probabiliste comme celui illustré précédemment. Comme dans le modèle de Kleinschmidt et Jaeger (2011), les phonèmes correspondent à des catégories discrètes, indépendantes les unes des autres, pour lesquelles il est possible de stocker diverses probabilités. Le modèle ShortListB possède en plus la possibilité de calculer la probabilité des mots à l'aide des connaissances phonémiques. Précisons que, bien que nous en ayons illustré deux dans cette partie, tous les modèles de perception ne sont pas probabilistes. Le modèle MERGE est, par exemple, un réseau connexionniste.

De leur côté, les modèles hiérarchiques interactifs sont des modèles impliquant un retour de la structure lexicale vers la structure phonétique, celui-ci étant dénommé feedback. Les unités phonémiques sont donc traitées à partir du son et des unités lexicales de plus haut niveau. Des modèles de perception, comme le célèbre modèle TRACE et ses variantes (McClelland et Elman, 1986; Mirman et al., 2006) ou le modèle proposé par Gaskell et Marslen-Wilson (1997), font partie de cette famille. Le modèle TRACE est, par exemple, composé d'une structure de traits phonétiques, en plus d'une structure phonémique et lexicale. Dans ce modèle, la structure phonémique fait le pont entre les deux autres structures puisqu'elle est, d'une part, connectée à la structure des traits et, d'autre part, connectée à la structure lexicale. Les informations entre ces niveaux s'échangent de manière bilatérale. Dans ce modèle, comme précédemment, la structure phonémique est composée d'unités phonémiques discrètes mais qui, contrairement aux autres modèles, ont la capacité de s'inhiber mutuellement.

En bref, malgré les différences d'implémentation et d'objectifs des modèles précédemment cités, les phonèmes sont représentés, à chaque fois, comme un ensemble d'unités discrètes. Globalement indépendants les uns des autres, à part dans le modèle TRACE où ils possèdent une fonction d'inhibition, les phonèmes n'ont aucune architecture, ni organisation. De plus, les modèles présentés étant principalement des modèles fonctionnels, soit computationnels, soit algorithmiques selon la taxonomie de Marr (1982), il n'est détaillé que leur relation avec les autres structures, qu'il s'agisse de sons, des traits phonétiques ou d'unités lexicales, mais aucun ne spécifie leur implémentation cognitive. Dernier point, la structure phonémique sert davantage d'outil que de sujet d'étude dans ces modèles. Ainsi, elle n'est ni remise en question, ni précisément étudiée. Néanmoins, la variété des modèles l'utilisant montre qu'elle semble être suffisante pour réaliser les principales fonctions de perception.

Les modèles de perception syllabiques sont moins courants. Citons cependant le travail de Sussman (1984) qui propose un modèle neuronal de représentation syllabique. L'hypothèse principale est que les syllabes sont des gabarits neuronaux (« frame ») pouvant être représentés sous la forme d'un réseau de neurones dans lequel chaque phonème de la syllabe souhaitée correspond à une cellule du réseau. Le contenu de la syllabe (« content »), c'est-à-dire aussi bien les phonèmes que les contraintes phonologiques propres à la langue, est stocké en mémoire et utilisé pour créer la syllabe en tant que telle. Ce serait donc le mélange frame/content qui permettrait une représentation syllabique.

Nous avons vu que les théories actuelles sont globalement en faveur d'une double structure cognitive, l'une phonémique et l'autre syllabique (se reporter à la section 3.1.3 pour le détail). Il existe également des modèles composés de cette double structure (Kiebel et al., 2009; Kröger et al., 2011, 2010, par exemple). La structure syllabique et phonémique du modèle de Kröger évolue selon les versions. Le modèle présenté par Kröger et al. (2011) possède deux cartes phonétiques (« phonetic map » et « phonemic map »), composées de phonèmes et syllabes. L'une des cartes (« phonetic map ») sert principalement pour la production, nous la mettons de côté. Dans la seconde (« phonemic map »), les unités phonétiques sont, chacune, constituées d'un ensemble de neurones sans qu'il n'y ait de différences notables entre les syllabes et les phonèmes.

À l'inverse, pour représenter les deux types d'unités, le modèle présenté par Kiebel et al. (2009) utilise une double structure hiérarchique, l'une étant phonémique, la seconde étant syllabique. Bien que fortement connectées, elles représentent deux espaces indépendants. Chaque structure est composée d'une séquence d'unités discrètes. Les séquences phonémiques permettent de reconnaître les syllabes et les séquences syllabiques permettent de reconnaître des séquences de plus haut niveau.

Dans ce modèle, les auteurs implémentent également les différences de temps de calcul. Ainsi, dans le modèle, le traitement phonémique est quatre fois plus rapide que le traitement syllabique.

Si ces auteurs proposent un modèle avec une double structure phonétique en accord avec les données expérimentales, elles ne sont ni spécifiquement étudiées, ni comparées. Pour cela, il n'y a, à notre connaissance, aucun modèle cognitif de perception. En revanche, il existe plusieurs modèles de reconnaissance de la parole ayant réalisé une comparaison entre une structure syllabique et une structure phonémique (e.g Bazzi et Glass, 2000; Ganapathiraju et al., 2001). Par exemple, le modèle proposé par Bazzi et Glass (2000) est, brièvement, un graphe pondéré dans lequel la reconnaissance consiste à trouver le meilleur chemin. Similairement aux modèles hiérarchiques présentés précédemment, il s'agit d'un modèle à deux niveaux, l'un phonétique, l'autre lexical. Bien que peu de détails soient donnés sur l'implémentation, les auteurs réalisent dans cette étude la comparaison entre un modèle avec, d'une part, un niveau phonétique phonémique et, d'autre part, un niveau phonétique syllabique. Pour cela, ils entrainent leur modèle sur un corpus puis testent ensuite la reconnaissance. En calculant le taux d'erreur phonétique moyen pour chaque modèle, ils observent que leur modèle syllabique est plus performant que leur modèle phonémique. En comparant un modèle syllabique avec un modèle basé sur des triphones, Ganapathiraju et al. (2001) montrent également la suprématie du modèle syllabique.

En résumé, la majorité des modèles de perception utilise une structure phonémique. Pour les quelques cas de modèles à double structure hiérarchique, aucun d'entre eux ne s'intéresse aux différences en termes d'organisation et de traitement. Seuls Kiebel et al. (2009) modélisent les différences de temps de traitement entre les deux voies. En terme de comparaison des deux voies, il n'y a, à notre connaissance, aucun modèle cognitif s'étant penché sur le sujet. En termes de performance, les modèles de reconnaissance montrent une amélioration en utilisant une structure syllabique. Cela nécessiterait quelques approfondissements pour mieux comprendre les implications cognitives d'un tel résultat.

4.1.1.4 Conclusion

Pour synthétiser, nous avons dédié cette section à la caractérisation des unités distinctives dans les modèles computationnels. Nous avons étudié, dans un premier temps, comment les représentations sensorielles et motrices des unités sont modélisées et, dans un second temps, nous nous sommes penchés sur leur structure cognitive.

En analysant les représentations des unités des modèles de perception, nous avons observé qu'il y a peu de modèles s'intéressant à l'étude conjointe des représentations sensorielles et motrices. De ce fait, le lien sensorimoteur entre les représentations sensorielles et motrices est absent dans la plupart de ces modèles. C'est pourquoi, nous avons étudié ce lien dans les modèles de production. Nous avons remarqué que ce lien peut être modélisé de différentes manières. Ensuite, lorsque nous nous sommes intéressés à la structure cognitive des unités, nous avons observé que les modèles de perception sont principalement focalisés sur les phonèmes et que peu d'entre eux analysent de façon conjointe les implications d'une structure phonémique et syllabique.

Cette revue des modèles existants nous sert à introduire le modèle que nous utilisons dans la suite

de ce document. Ainsi, dans la prochaine section, nous proposons un modèle de la communication composé de représentations sensorielles et motrices. Conformément aux modèles de production, il possède également un lien sensorimoteur. Dans les prochains chapitres, nous analysons conjointement, avec ce modèle, les représentations sensorielles et motrices durant une tâche de perception afin de discuter de leurs rôles respectifs dans différents contextes. Les études de modélisation des deux prochains chapitres sont réalisées avec des unités phonémiques. Dans le chapitre 6, une variante étendue de notre modèle est utilisée afin d'étudier la structure conjointe phonémique et syllabique.

4.1.2 Comment les unités distinctives se développent-elles dans les modèles ?

Dans la section précédente, nous avons étudié les représentations sensorielles et motrices ainsi que la structure cognitive des modèles computationnels, principalement de perception. Dans tous les exemples donnés, nous avons volontairement mis de côté une catégorie de modèles : ceux s'intéressant au développement des représentations phonétiques. Ils sont le sujet principal de cette section.

Comme c'est souvent le cas dans la littérature, la section consacrée au développement des représentations phonétiques est composée de deux parties comme dans le chapitre précédent. Dans un premier temps, nous avons observé le développement de la perception durant lequel s'effectue l'acquisition des représentations sensorielles. Dans un second temps, nous avons décrit le développement de la production, davantage focalisé sur l'acquisition des représentations motrices (voir section 3.2.1 pour le détail).

Nous reprenons cette distinction pour examiner comment est implémenté le développement phonétique dans les modèles computationnels. Ainsi, dans cette section, nous étudions d'abord, dans deux sous-sections différentes, les deux types de développements : l'acquisition des représentations sensorielles et l'apprentissage des représentations motrices. Puis, dans une troisième sous-section, nous nous focalisons sur la structure cognitive des modèles de développement.

4.1.2.1 Apprentissage sensoriel

Les modèles visant à simuler les mécanismes d'apprentissage sensoriel intègrent des processus de convergence vers les propriétés statistiques des données de l'environnement. Néanmoins, la nature de ces processus de convergence prend des formes variées. Dans un des premiers modèles emblématiques du domaine, le modèle WRAPSA (Word Recognition And Phonetic Structure Acquisition) de Jusczyk (1993), l'apprentissage implique un ensemble de détecteurs acoustiques généraux, dont les poids de connexion vers les décodeurs phonétiques, variant selon les langues, sont les paramètres des processus d'apprentissage. Les études sur le modèle TRACE, mentionné précédemment, impliquent, de leur côté, des processus d'apprentissage neuronal hebbien qui modifient la structure des réseaux connexionnistes à la base du modèle (voir, par exemple, Mirman et al., 2006).

Un cadre important dans ce domaine a été fourni ces vingt dernières années grâce au développement des théories exemplaristes, dans lesquelles il est postulé que les représentations cognitives pourraient être élaborées non pas à partir de représentations paramétriques explicites ou implicites

(distribuées dans des réseaux connexionnistes), mais directement par l'enregistrement et le stockage, sous une forme adéquate, des traces en mémoire des expériences vécues par le sujet. C'est ainsi qu'ont été développés des modèles de traitement perceptif (Johnson, 1997), de reconnaissance des mots (Goldinger, 1996, 1998) ou de production du langage (Pierrehumbert, 2001). Le cadre exemplariste a ainsi fourni la base de nombreux travaux en modélisation des processus d'acquisition du langage (Pierrehumbert, 2003).

Par la suite, nous nous focalisons sur les modèles opérant directement à base de distributions statistiques explicites, tel qu'il est supposé d'après les données expérimentales (voir section 3.2.2.1), sans évacuer l'hypothèse que des modèles exemplaristes pourraient également être considérés, ou couplés à des modèles probabilistes explicites (Pierrehumbert, 2016).

Un des premiers modèles à appliquer un apprentissage statistique performant est celui proposé par de Boer et Kuhl (2003). Dans celui-ci, les auteurs étudient et comparent l'apprentissage de la position des voyelles américaines [a, i, u] dans l'espace auditif à partir de deux corpus. Pour cela, l'espace auditif correspond à un espace à deux dimensions composé des deux premiers formants F1 et F2, suffisant pour distinguer les trois voyelles. Ces dernières sont modélisées par une mixture de gaussiennes, une gaussienne pour chaque voyelle. Les auteurs utilisent l'algorithme Expectation–Maximization (EM) pour réaliser l'apprentissage (Bilmes, 1998; Dempster et al., 1977) qui, brièvement, estime la position des trois gaussiennes à partir de la distribution globale des données du corpus choisi. En fin d'apprentissage, les auteurs observent que, pour au moins un des corpus, cet apprentissage permet de retrouver la position adéquate des trois voyelles.

Si cet apprentissage est efficace, il possède néanmoins quelques limites. L'une d'elles, notée également dans l'article, est de considérer que le nombre d'unités phonétiques, et donc de gaussiennes, est connu à l'avance. Une deuxième est que l'apprentissage se fait de manière « batch », c'est-à-dire que toutes les données du corpus sont fournies en même temps pour paramétrer le modèle. Les bébés semblent, à l'inverse, apprendre davantage de manière itérative. À ce titre, plusieurs études tentent d'effectuer un apprentissage itératif avec des mixtures de gaussiennes non initialisées avec le nombre correct d'unités (McMurray et al., 2009; Vallabha et al., 2007).

Par exemple, Vallabha et al. (2007) testent leur modèle sur un apprentissage, d'une part, des voyelles anglaises [I, i, ɛ, e] et, d'autre part, des voyelles japonaises [i, iː, e, eː]. Ils utilisent un espace auditif à trois dimensions : les deux premiers formants F1 et F2 et un paramètre de durée. La mixture utilisée contient, au départ, 1000 gaussiennes de même variance aléatoirement réparties dans l'espace auditif. En se servant d'un corpus de voyelles soit de locuteurs anglais, soit de locuteurs japonais, ils utilisent un algorithme basé sur une version itérative de l'algorithme EM (EM incrémental) pour paramétrer leurs gaussiennes. À la fin de l'apprentissage, ils remarquent que les gaussiennes ayant la plus forte probabilité correspondent aux voyelles du corpus appris. Ils montrent ainsi que leur modèle est adapté pour apprendre différentes distributions de voyelles et donc différentes langues, et ce, sans avoir une initialisation avec un nombre de catégories correct.

Cependant, McMurray et al. (2009), avec différentes conditions, obtiennent un résultat plus mitigé. Dans un espace auditif à une dimension correspondant au Voice Onset Time, ils testent l'apprentissage de deux catégories phonétiques anglaises. Au début de l'apprentissage, ils initialisent entre 10 et 20 gaussiennes en les répartissant également aléatoirement dans l'espace auditif et avec une même va-

riance. En se basant sur un corpus anglais, ils utilisent un algorithme itératif de descente de gradient, basé sur un critère de Maximum Likelihood Estimation (MLE) qui, globalement, estime la probabilité de chaque gaussienne, à chaque itération, pour le stimulus donné, et met à jour la mixture avec ces probabilités. En fin d'apprentissage, ils n'obtiennent pas de résultats concluants puisqu'il reste, en moyenne, 11 gaussiennes pour approximer les deux catégories. Cependant, les deux catégories sont mieux apprises lorsqu'ils ajoutent un mécanisme de compétition de type winner-take-all, dans lequel seule la gaussienne avec la plus haute probabilité est mise à jour à chaque itération, au lieu de l'ensemble des gaussiennes. Ils montrent ainsi que l'apprentissage statistique itératif fonctionne pour apprendre les catégories phonétiques, mais seulement avec un algorithme possédant les bons critères de mise à jour des paramètres comme un mécanisme de compétition.

Ce n'est, néanmoins, pas le seul moyen existant pour améliorer l'apprentissage de modèles multi-gaussiens. Par exemple, plusieurs études de Feldman et collègues montrent que l'apprentissage conjoint des catégories de plus haut niveau, comme les mots, améliore l'apprentissage statistique des catégories phonétiques, notamment celles se superposant dans l'espace auditif (Feldman et al., 2013a, 2009b, 2011). De la même manière, plusieurs études montrent que l'apport des règles phonologiques peut aider à apprendre les catégories phonétiques et notamment faire la distinction entre les contrastes phonémiques et les contrastes allophoniques (Dillon et al., 2013; Peperkamp et al., 2006). L'apport du lexique et des règles phonologiques pour réaliser l'apprentissage des catégories phonétiques peut d'ailleurs se faire dans un même modèle, comme le montre l'étude de Martin et al. (2013).

Bien que nous ayons uniquement décrit des modèles basés sur des mixtures de gaussiennes, il ne s'agit pas du seul moyen de faire de l'apprentissage statistique phonétique. Par exemple, certains modèles utilisent des cartes auto-organisatrices (Gauthier et al., 2007; Kröger et al., 2014; Salminen et al., 2009), des modèles de Markov cachés (Goldsmith et Xanthos, 2009; Taniguchi et al., 2016; Varadarajan et al., 2008) ou des méthodes de clustering (Coen, 2006). L'ensemble de ces modèles réussit avec succès à apprendre la statistique des unités phonétiques, ce qui montre que l'apprentissage statistique en tant que tel est relativement robuste.

En résumé, l'apprentissage statistique, seul ou accompagné d'autres mécanismes, semble globalement efficace pour apprendre les catégories phonétiques. Outre cet apprentissage, différents modèles se sont centrés sur l'apprentissage phonotactique (par exemple, Hayes et White, 2013; Hayes et Wilson, 2008; Magri, 2014). Néanmoins, ces modèles semblent plus centrés sur des préoccupations phonologiques ou syntaxiques que phonétiques.

4.1.2.2 Apprentissage moteur

Nous avons vu précédemment que lorsqu'un individu produit une unité phonétique, il reçoit un retour sensoriel de ce son. Cela lui permet, entre autres, d'assurer une bonne production et d'ajuster ses gestes moteurs, si nécessaire. Pour pouvoir réaliser ceci, l'individu en question doit non seulement savoir produire l'unité phonétique mais doit également posséder un lien entre ses représentations sensorielles et ses représentations motrices pour assurer le retour auditif. De la même manière, lorsque le bébé apprend à parler, il doit apprendre à produire les bonnes catégories phonétiques et doit posséder un retour auditif adéquat pour savoir si ce qu'il produit est correct. Le fait qu'il produise, dès le

babillage, des sons influencés par sa langue native laisse supposer qu'il existe effectivement un lien entre ce qu'il perçoit et ce qu'il produit. La plupart des modèles sur le développement de la production réalisent le développement moteur en deux phases. Dans la première phase, dite sensorimotrice, le modèle apprend à lier ses représentations motrices et ses représentations sensorielles. Dans la seconde, dite motrice, le modèle se focalise sur les signaux de son environnement afin de ne garder que les productions de sa langue native.

Lors de la phase sensorimotrice, la solution privilégiée est l'exploration. Brièvement, cela consiste à produire des gestes moteurs, percevoir les stimuli sensoriels correspondants, et associer les deux, ce qui semble se rapprocher du babillage effectué par le bébé. Du fait que les espaces à apprendre sont grands et que la relation sensorimotrice est surjective et non-linéaire, cette exploration peut très rapidement devenir complexe et coûteuse. C'est pourquoi les modélisateurs utilisent des techniques d'exploration adaptées pour faciliter l'apprentissage, centrées autour de deux choix.

Le premier choix à faire lors de l'exploration consiste à définir l'espace à explorer : soit l'espace des représentations motrices dans lequel sont choisis des gestes moteurs à produire, soit celui des représentations sensorielles dans lequel sont choisis des stimuli à reproduire. L'exploration correspondante est alors respectivement appelée « babillage moteur » (motor babbling) ou « babillage d'objectifs » (goal babbling). Si les théories et études comportementales débattent encore sur le sujet (voir par exemple Messum, 2008, pour une revue), les modèles semblent montrer de meilleures performances avec le babillage d'objectifs par rapport au babillage moteur (Philippsen et al., 2015, 2016; Rolf et Steil, 2012; Rolf et al., 2010, 2011).

Le second choix à faire lors de l'exploration consiste à définir une stratégie d'exploration. Le plus simple est de faire un apprentissage aléatoire. Néanmoins, il ne s'agit pas de la méthode la plus efficace. Une autre manière de faire consiste à définir une heuristique basée sur l'exploration des portions de l'espace où l'incertitude est très grande pour le modèle. Cette technique est appelée « apprentissage actif » (active learning). Une des méthodes d'apprentissage actif, nommée parfois « motivation intrinsèque » (intrinsic motivation) ou « apprentissage par curiosité », consiste à calculer, lors de l'exploration, les progrès du modèle et à choisir les points permettant un progrès maximal (Baranes et Oudeyer, 2010, 2013; Moulin-Frier et Oudeyer, 2012, 2013). Par exemple, Moulin-Frier et Oudeyer (2013) proposent une comparaison de modèles différant sur leur exploration de l'espace vocalique. En implémentant, pour chacun d'eux, une mixture de gaussiennes, ils comparent, d'une part, un apprentissage basé sur le babillage moteur et un apprentissage basé sur le babillage d'objectifs et, d'autre part, un apprentissage aléatoire avec un apprentissage actif. Ils montrent que, toutes techniques confondues, le babillage d'objectifs est plus précis que le babillage moteur et que lorsque le babillage d'objectifs est terminé, les gaussiennes sont mieux réparties dans l'espace auditif pour apprendre les voyelles. Par ailleurs, ils observent que, parmi les techniques d'exploration, le babillage d'objectifs actif donne de meilleures performances que le babillage d'objectifs basé sur une exploration aléatoire.

L'apprentissage actif par motivation intrinsèque n'est pas la seule stratégie d'exploration à être performante. Un des critères variables de cette stratégie concerne la mesure utilisée pour guider l'exploration (Howard et Messum, 2011; Murakami et al., 2015). Par exemple, au lieu de considérer le niveau de progression, Murakami et al. (2015) implémentent une mesure basée sur le niveau de confiance de la qualité d'apprentissage. Dans cette étude, le modèle continue d'apprendre un stimulus s tant que la précision d'apprentissage n'a pas atteint un certain seuil. Dans l'étude de Howard

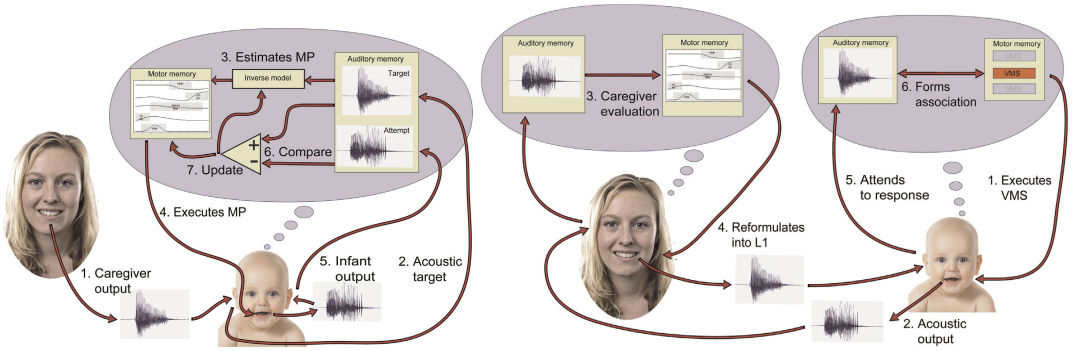
et Messum (2011), les critères privilégiés sont la saillance et la diversité du son. Mais, il est également possible d'envisager une stratégie d'exploration autre que l'apprentissage actif. Par exemple, au lieu d'avoir une stratégie auto-centrée, certains proposent que l'exploration soit facilitée par un renforcement social dans lequel un interlocuteur extérieur guide le choix des stimuli à explorer (Philippsen et al., 2014; Warlaumont, 2012; Warlaumont et al., 2013). Néanmoins, il a été montré que ce renforcement social pouvait quelquefois biaiser l'exploration (Warlaumont et al., 2011).

Si cet appel à un interlocuteur extérieur n'est peut-être pas le plus adapté pour faciliter la phase sensorimotrice, il est, en revanche, le plus souvent utilisé dans la phase motrice, lors de laquelle le modèle se focalise sur l'apprentissage de la production des unités de sa langue. Cet interlocuteur peut alors avoir plusieurs rôles. Dans un grand nombre de modèles, il fournit les stimuli à explorer (Lopes et al., 2009). On parle alors d'imitation. Plusieurs études montrent que cette phase d'imitation semble adaptée pour apprendre les unités phonétiques (Bailly, 1997; Guenther et Vladusich, 2012; Heintz et al., 2009; Hornstein et Santos-Victor, 2007; Kanda et al., 2008, 2009; Markey, 1994; Tourville et Guenther, 2011).

Contrairement à la phase d'exploration sensorimotrice, cet apprentissage social semble également plus adapté que l'apprentissage auto-centré. Par exemple, Westermann et Miranda (2004) proposent une comparaison entre un apprentissage auto-centré et un apprentissage social par imitation. Dans une première phase, le modèle commence par une exploration de son espace moteur grâce à laquelle il apprend la relation entre ce qu'il produit et ce qu'il perçoit. Dans la seconde phase, l'apprentissage auto-centré consiste à choisir un son proche de ceux déjà appris et à tenter d'inférer le geste moteur correspondant. L'apprentissage guidé consiste, lui, à choisir un des sons provenant des voyelles d'un corpus (soit français, soit allemand) et de tenter d'en inférer le geste moteur correspondant. En fin d'apprentissage, les résultats indiquent que l'apprentissage auto-centré favorise les portions de l'espace où il y a une correspondance quasi linéaire entre l'espace auditif et l'espace moteur. Dit autrement, les portions privilégiées sont celles pour lesquelles une faible variation dans l'espace moteur produit une faible variation dans l'espace auditif. Cependant, ces portions ne correspondent pas aux voyelles des corpus souhaités. L'apprentissage guidé, quant à lui, permet de se centrer sur les voyelles de l'environnement et se rapproche ainsi davantage des données expérimentales sur le bébé. Si l'apprentissage auto-centré semble suffisant pour catégoriser le signal en unités phonétiques quelconques (voir aussi Oudeyer, 2002), l'apprentissage social semble en revanche nécessaire pour se focaliser sur les unités de sa langue native. Néanmoins, même si l'apprentissage auto-centré seul ne suffit pas, Moulin-Frier et al. (2013) montrent qu'il peut accompagner l'apprentissage social et que les deux, conjointement, permettent au modèle d'acquérir des représentations motrices des unités phonétiques de l'environnement d'une manière similaire à celle du bébé.

Si l'apprentissage social à travers l'imitation, tel qu'il est présenté ici, permet effectivement de bien apprendre les catégories phonétiques, son réalisme est parfois remis en question (Howard et Messum, 2011). Ces auteurs montrent en effet que l'imitation est en réalité souvent plutôt inversée, conduisant non pas le bébé à imiter son parent, mais le parent, désireux de communiquer avec son bébé, à imiter en retour les productions spontanées de celui-ci. Ceci conduit à proposer, à la place, une autre variante, nommée apprentissage miroir, basée sur le fait que l'imitateur n'est pas le bébé mais l'interlocuteur. Brièvement, l'imitation dans le cadre d'un apprentissage phonétique consiste à supposer que le bébé reçoit un signal de son interlocuteur, tente de le reproduire, le compare avec ce qu'il a

entendu et met à jour ses représentations motrices en fonction de cette comparaison. L'apprentissage miroir consiste à supposer que le bébé produit un geste moteur, qui est, par la suite, imité par l'interlocuteur, puis qu'il apprend la correspondance entre son geste produit et le stimulus imité (Messum et Howard, 2015, voir aussi Fig. 4.5 pour une comparaison des deux apprentissages). Cette méthode semble plus réaliste que l'apprentissage par imitation classique. Plusieurs études montrent qu'elle est également efficace pour apprendre les unités phonétiques (Howard et Messum, 2011; Ishihara et al., 2008; Messum et Howard, 2015; Miura et al., 2007, 2008, 2012; Vaz et al., 2009; Yoshikawa et al., 2003).



(a) Apprentissage par imitation (b) Apprentissage miroir

FIGURE 4.5 – Différences entre l'apprentissage par imitation et l'apprentissage miroir. Schémas tirés de Messum et Howard (2015)

En résumé, nous avons considéré les modèles dans lesquels l'apprentissage moteur peut être découpé en deux phases : une phase sensorimotrice, similaire au babillage du bébé, dans laquelle le modèle apprend la relation entre ses représentations sensorielles et motrices, et une phase plutôt motrice, dans laquelle le modèle se focalise sur les unités de son environnement et dans laquelle il apprend donc plus spécifiquement la relation entre ses représentations motrices et les unités distinctives de sa langue. La première se fait généralement par exploration, la seconde est davantage basée sur les stimuli de l'environnement.

4.1.2.3 Le développement de la structure des unités

Les études sur le développement du bébé semblent montrer que le bébé acquiert, ou du moins utilise, une structure syllabique avant d'utiliser une structure phonémique. Le développement des phonèmes n'est pas non plus homogène puisque les voyelles semblent être apprises avant les consonnes. Dans cette section, nous étudions, d'une part, la structure des unités globalement utilisées par les modèles de développement et, d'autre part, nous portons un intérêt particulier aux modèles qui se sont intéressés aux mécanismes de développement de la structure cognitive des unités phonétiques.

Du côté des modèles de perception, les études principales sur le développement, précédemment citées, ont essentiellement été réalisées sur des phonèmes (de Boer et Kuhl, 2003; Dillon et al., 2013; Feldman et al., 2013a; Martin et al., 2013; McMurray et al., 2009; Vallabha et al., 2007). Les phonèmes utilisés sont assez diversifiés et varient considérablement d'une étude à l'autre. Pour commencer par

le plus abstrait, certaines études ne spécifient pas les catégories phonémiques qu'ils apprennent dans leur modèle. C'est par exemple le cas de McMurray et al. (2009) qui indiquent simplement apprendre deux phonèmes anglais de Voice Onset Time différents ou Feldman et al. (2009b) qui se servent de quatre catégories non spécifiées, notées A, B, C et D. D'autres études préfèrent avoir des unités plus concrètes mais ne se focalisent que sur un seul type de phonèmes : les voyelles. Comme vu précédemment, Vallabha et al. (2007) se basent, par exemple, sur les voyelles anglaises [I, i, ɛ, e] et sur les voyelles japonaises [i, iː, e, eː] tandis que de Boer et Kuhl (2003) utilisent les voyelles américaines [a, i, u]. Ces trois voyelles, communes à un grand nombre de langues, ont également été étudiées dans le modèle de Dillon et al. (2013) mais en utilisant un corpus de la langue inuktikut (un dialecte de l'inuit). Si toutes les études citées actuellement utilisent un nombre limité d'unités phonétiques à apprendre, inférieur à cinq, ce n'est pas toujours le cas. À titre d'illustration, Coen (2006) réalise l'apprentissage de dix voyelles américaines. Mais les études réalisant l'apprentissage le plus exhaustif de phonèmes sont celles privilégiant l'apprentissage conjoint des consonnes et voyelles. Par exemple, les différentes simulations de Martin et al. (2013) sont basées sur un corpus de plus d'une quarantaine de phonèmes du japonais et sur un corpus de plus d'une cinquantaine de phonèmes du néerlandais.

L'apprentissage décrit ici, bien qu'il soit toujours phonémique, diffère d'une étude à l'autre aussi bien de par la nature des catégories étudiées que par leur nombre. Notons que, même au sein de cette diversité, il n'y a pas, à notre connaissance, d'études réalisant uniquement l'apprentissage de consonnes. Cependant, il existe quelques modèles de perception réalisant un apprentissage à partir d'autres catégories phonétiques que les phonèmes. C'est notamment le cas des modèles n'étant pas basés sur l'apprentissage de mixtures de gaussiennes (Dupoux et al., 2011; Kröger et al., 2010) Par exemple, Dupoux et al. (2011) proposent un modèle d'apprentissage d'unités syllabiques. Brièvement, l'apprentissage de leur modèle consiste à stocker des fenêtres temporelles du signal acoustique puis de les comparer afin d'en extraire les similitudes et de retrouver les catégories phonémiques. L'apprentissage est réalisé sur différentes tailles de fenêtres : diphones, triphones ou des syllabes de tailles constantes.

Du côté des modèles de production, il en existe également un grand nombre qui se concentrent sur l'apprentissage des phonèmes, notamment les voyelles, que ce soit pour l'apprentissage du lien sensorimoteur (Moulin-Frier et Oudeyer, 2013; Murakami et al., 2015; Westermann et Miranda, 2004) ou durant la phase de focalisation sur les unités de la langues (Ishihara et al., 2008; Miura et al., 2012). Néanmoins, contrairement aux modèles de perception, il y en a également plusieurs qui s'intéressent à l'apprentissage des syllabes (Bailly, 1997; Messum et Howard, 2015; Philippsen et al., 2014; Warlaumont et al., 2013). Cela vient notamment du fait que la phase de babillage du bébé est reconnue pour être une production d'unités de type syllabique (« proto-syllabes »). Quelques modèles de production apprennent même une double structure phonémique et syllabique (Brandl et al., 2008; Vaz et al., 2009).

De plus, contrairement aux modèles de perception qui se focalisent plus sur le mécanisme d'apprentissage que sur la structure des unités apprises, certains modèles de production focalisent leur étude sur l'acquisition de cette structure. Par exemple, Warlaumont (2012) propose un modèle pour mieux comprendre l'apparition du babillage (canonique et varié) et donc mieux comprendre comment le bébé apprend à produire des syllabes. L'étude de Najnin et Banerjee (2016) s'intéresse à la même problématique.

En résumé, les modèles d'apprentissage sensoriel sont plus focalisés sur les phonèmes, similairement aux modèles de perception (voir section 4.1.1.3). Les modèles de production semblent plus diversifiés et certains d'entre eux focalisent même leur étude sur la compréhension de la structure cognitive des unités syllabiques. Néanmoins, il n'y a, à notre connaissance, aucun modèle s'intéressant conjointement à la structure phonémique et syllabique et qui compare leur apprentissage respectif.

4.1.2.4 Conclusion

Pour synthétiser, nous avons observé comment s'effectue le développement des unités cognitives. Nous avons étudié, d'une part, comment les modèles d'apprentissage réalisent les étapes du développement et, d'autre part, quelles structures cognitives étaient le plus souvent utilisées.

Concernant les étapes d'apprentissage, les modèles se basent sur le développement du bébé. Afin d'apprendre les représentations sensorielles correspondant aux unités distinctives, un certain nombre de modèles de développement se basent sur les mécanismes d'apprentissage statistique tels qu'ils ont été décrits pour le bébé. Du côté des représentations motrices, les modèles de développement s'effectuent, pour la plupart, en au moins deux phases : une phase d'exploration qui se rapproche du babillage du bébé, lors de laquelle se développe le lien entre les représentations sensorielles et motrices, et une phase de focalisation, qui se sert de l'imitation (du modèle apprenant ou de l'interlocuteur) pour apprendre la relation entre les représentations motrices et les unités distinctives de la langue donnée.

Dans la prochaine section, nous nous servons de ces étapes pour réaliser l'apprentissage du modèle que nous utilisons dans nos études. Par ailleurs, nous pensons que ces deux apprentissages, sensoriel et moteur, du fait de leurs différences, pourraient expliquer plusieurs phénomènes en perception, par exemple le fait que les aires motrices soient plus actives lors d'une perception bruitée. De plus, à notre connaissance, aucun modèle existant ne réalise une comparaison des comportements respectifs de ces deux apprentissages. C'est pourquoi, nous nous chargeons, dans plusieurs de nos études, de les comparer et surtout d'analyser leurs conséquences respectives sur la perception.

Par la suite, nous analysons la structure cognitive utilisée dans les modèles de perception. Nous avons vu que les modèles d'apprentissage sensoriel sont plus focalisés sur l'apprentissage phonémique tandis que les modèles d'apprentissage moteur sont plus diversifiés. Néanmoins, aucun d'entre eux ne semble avoir analysé conjointement les différences d'apprentissage des structures phonémiques et syllabiques. Nous tentons d'apporter quelques réponses à ce sujet dans le chapitre 6.

4.2 Le modèle COSMO

Cette section est dédiée au modèle utilisé durant toute cette thèse : le modèle COSMO. Elle recoupe, d'une part, l'analyse effectuée sur l'implémentation des représentations sensorielles et motrices, du lien sensorimoteur et de la structure cognitive des unités phonétiques et, d'autre part, celle effectuée sur l'implémentation du développement des unités phonétiques.

COSMO, signifiant « Communication about Objects using Sensory–Motor Operations », est un

modèle bayésien antérieurement conçu par l'équipe. Il a déjà été implémenté pour étudier divers aspects de la parole tels que l'émergence des systèmes phonologiques (Moulin-Frier et al., 2015) ou la perception et la production de la parole (Laurent, 2014; Laurent et al., 2017, 2013; Moulin-Frier et al., 2012).

Nous détaillons dans cette section le modèle générique. C'est le modèle pleinement utilisé dans les études du chapitre 4 et 5. Le chapitre 6, quant à lui, utilise une extension de COSMO. Nous laissons de côté cette extension, pour le moment, et nous nous concentrons sur le modèle de base, appelé aussi modèle générique.

Après une brève introduction sur ce qu'est un modèle bayésien, nous présentons le modèle COSMO dans son ensemble en précisant notamment comment sont modélisées les unités phonétiques et leurs représentations. Ensuite, nous décrivons comment est réalisé l'apprentissage phonétique en insistant particulièrement sur la manière dont sont développées les représentations sensorielles et motrices. Pour finir, nous terminons sur un exemple de questions, nous servant plusieurs fois durant cette thèse, concernant la modélisation des trois grandes théories de la perception.

4.2.1 Structure d'un programme bayésien

COSMO est un modèle bayésien élaboré à partir des principes de la programmation bayésienne (Bessière et al., 2013). Celle-ci est basée sur une approche subjectiviste des probabilités, dans laquelle la théorie des probabilités est vue comme une extension de la logique (Jaynes, 2003). En ce sens, les règles mathématiques des probabilités ne servent pas à calculer des fréquences ou des variables aléatoires mais sont davantage des outils utilisés pour réaliser des raisonnements rationnels (voir, par exemple, Colas et al., 2010; Lebeltel et al., 2004, pour plus de détails). Cette méthodologie semble donc adaptée pour traiter des problèmes liés à la compréhension du cerveau.

En pratique, un programme bayésien est une structure qui se construit en deux étapes principales nommées « description » et « question » (voir Fig. 4.6). La première étape commence par une phase dite de spécification durant laquelle sont définies les variables et les distributions du modèle. Plus précisément, dans celle-ci, on choisit d'abord les variables nécessaires pour le modèle. Elles forment les dimensions de la distribution conjointe qui est la distribution principale du modèle. La conjointe est ensuite décomposée en un produit de distributions plus simples. Pour finir, les formes paramétriques des distributions sont précisées ou une manière de les calculer est fournie si la forme en question est trop complexe à définir au préalable. Une fois le modèle spécifié, la phase d'identification commence, qui correspond à l'apprentissage du modèle. Quand le modèle est entièrement décrit et que l'étape de description est terminée, il est possible de l'utiliser à l'aide de ce qu'on appelle des questions. Une question correspond à une tâche qui nous sert à étudier le comportement de notre modèle et qui se calcule par inférence bayésienne à partir des distributions du modèle.

Ces étapes sont importantes à retenir car nous les retrouvons tout au long de cette thèse. Dans ce chapitre, nous nous focalisons sur la spécification et l'apprentissage générique du modèle COSMO. Dit autrement, nous présentons uniquement l'étape de « description » du modèle en omettant de détailler comment le modèle est implémenté. Les détails d'implémentation des différentes versions du modèle ainsi que l'étape de « question » sont le sujet des deux chapitres suivants dédiés aux

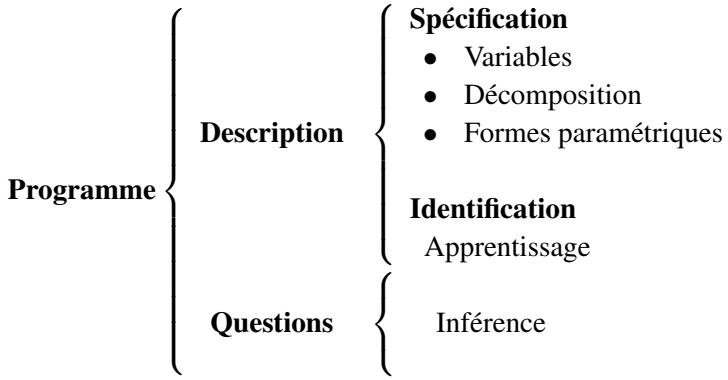


FIGURE 4.6 – Étapes d'un programme bayésien. Adapté de Bessière et al. (2013)

différentes études réalisées à l'aide du modèle COSMO.

4.2.2 Spécification du modèle COSMO

La spécification du modèle COSMO reprend les étapes du programme bayésien telles qu'elles sont détaillées dans la section précédente. Dans une première partie, nous décrivons les variables, ce qui nous permet d'écrire la conjointe du modèle. Dans une seconde partie, nous expliquons comment se décompose cette conjointe et détaillons les distributions qui en résultent.

4.2.2.1 Description des variables du modèle

Afin de bien comprendre comment est construit le modèle COSMO, nous commençons par décrire son origine et l'hypothèse sur laquelle il est basé. Cette explication a l'avantage de faciliter par la suite la description des variables du modèle.

Pour cela, imaginons une situation de communication orale la plus simple possible entre deux agents : un locuteur et un auditeur. Le locuteur souhaite transmettre un concept à l'auditeur. Pour cela, le locuteur utilise des représentations motrices correspondant au concept souhaité et les produit grâce à son conduit vocal. Cette production, le message, est ensuite transmise dans l'environnement et reçue sous la forme d'un signal sonore par l'auditeur. Ce dernier interprète alors le signal reçu pour retrouver le concept. On suppose que la communication est un succès si le concept compris par l'auditeur correspond à celui transmis par le locuteur. Cette situation très simple est schématisée Fig. 4.7 dans laquelle le concept correspond au phonème [a].

Dans le modèle COSMO, il est supposé que cette situation peut être internalisée dans le cerveau d'un unique agent, ce qui est nommée l'hypothèse d'internalisation. Selon cette hypothèse, les éléments de la communication internalisée correspondent aux variables du modèle, symbolisées, chacune, par un symbole précis. Ainsi, le concept de la communication devient un « objet » internalisé.

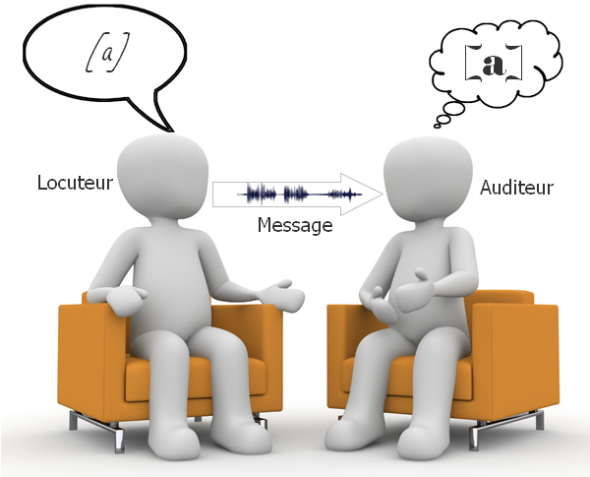


FIGURE 4.7 – Schéma d'une situation de communication simplifiée entre deux agents

Le terme « objet » se réfère, ici, à tout concept (objet, être, action, pensée) pouvant être communiqué. Il est noté O. Dans cette thèse, nous limitons le terme objet aux unités distinctives et particulièrement celles préalablement définies : les phonèmes et/ou les syllabes. Cela constitue ce qui est parfois nommé le niveau de « seconde articulation » du langage (Martinet, 1970). Il est donc important de mentionner, qu'à l'inverse, tous les éléments du niveau de « première articulation » (mots, morphèmes, structure syntaxique, unités sémantiques) sont écartés de cette thèse.

Dans la situation de communication, il y a deux concepts différents : celui pensé par le locuteur et celui interprété par l'auditeur. Dans le modèle, après internalisation, cela correspond à deux variables objets, notées respectivement O_S (S faisant référence ici au locuteur : « speaker ») et O_L (L faisant référence ici à l'auditeur : « listener »[2]). De plus, dans la vision du modèle choisie pour cette thèse, l'objet O_S correspond aux unités distinctives liées aux représentations motrices tandis que l'objet O_L correspond aux unités distinctives liées aux représentations sensorielles. Du fait de ce lien, et pour les distinguer, nous appelons dans la suite de cette thèse O_S, des « objets moteurs », et O_L, des « objets sensoriels ».

Focalisons-nous maintenant sur les gestes moteurs que le locuteur utilise pour produire le concept. Selon l'hypothèse d'internalisation, ces gestes correspondent aux représentations motrices présentes dans le cerveau. Il leur est attribué la lettre M faisant référence au terme « moteur » et qui, dans sa définition d'origine, correspond à tout élément capable de produire un mouvement. Cette lettre fait également écho au terme « moteur » du « cortex moteur ». La définition du mot « élément » et les représentations motrices considérées sont laissées volontairement floues ici puisqu'elles dépendent principalement du niveau d'analyse dans lequel on se place lors de l'implémentation du modèle. Il peut s'agir des organes du conduit vocal, des muscles, des articulations, etc., ceux-ci pouvant être considérés ensemble ou séparément. Ainsi, nous regroupons dans COSMO différents niveaux possibles d'analyse de la chaîne de production. Nous ne faisons notamment pas la distinction entre niveau

2. l'utilisation de terminologies anglaises se justifie par le fait que tous ces travaux ont été présentés dans plusieurs publications en anglais, et qu'il semble inadéquat de proposer deux terminologies en langues différentes, avec de forts risques de confusions.

articulatoire et niveau moteur ou entre le niveau des actions motrices, celui des commandes motrices et celui des programmes moteurs, bien que ces distinctions jouent un rôle important dans les débats sur les théories motrices (voir, par exemple, les différences sur ce point précis entre Fowler, 1986; Galantucci et al., 2006; Liberman et Mattingly, 1985).

De son côté, le signal sonore reçu et interprété par l'auditeur correspond, selon l'hypothèse d'internalisation, aux représentations sensorielles. Elles sont symbolisées par la lettre S. Dans la version actuelle du modèle, les représentations sensorielles correspondent uniquement aux représentations auditives. Tout comme les représentations motrices, elles peuvent être considérées à différents niveaux d'analyse, mais nous les considérons pour le moment dans leur ensemble.

Pour finir, bien qu'il ne fasse pas partie de la communication elle-même mais qu'il en soit plutôt une conséquence, le succès de la communication est également pris en compte dans le modèle. Selon l'hypothèse d'internalisation, la variable en question ne correspond plus au « succès » de la communication mais assure la cohérence entre les deux objets O_S et O_L. Elle est notée C. De manière imagée, cette variable de cohérence a le rôle d'un interrupteur : si la variable C est « allumée » alors il est considéré que les deux objets sont connectés et égaux. Si elle est « éteinte », les deux objets sont simplement considérés indépendants l'un de l'autre.

Précisons maintenant quelques détails techniques sur ces variables probabilistes. Ce sont toutes les cinq des ensembles finis et discrets. Néanmoins, malgré cette similitude, elles se distinguent de par leurs différences de cardinal. Il y a globalement trois types de variables : binaire, à faible cardinal et à fort cardinal. La seule variable binaire du modèle est la variable C. Il s'agit en réalité d'une variable booléenne vrai/faux. Les variables à faible cardinal sont les objets O_S et O_L qui représentent des ensembles catégoriels et, plus précisément dans notre cas, des catégories phonétiques. Ces variables catégorielles ont donc un nombre limité de valeurs. Cette modélisation des unités phonétiques en unités discrètes catégorielles est, comme nous l'avons vu en section 4.1.1.3, classiquement utilisée. À l'opposé, les représentations sensorielles S et motrices M sont les variables à fort cardinal. Du fait qu'elles traitent de phénomènes physiques de l'environnement, ces variables quantitatives, correspondant en fait à une discrétisation d'un espace continu, nécessitent un nombre important de valeurs afin de modéliser le plus précisément possible les phénomènes sensoriels et moteurs de l'environnement.

En résumé, le modèle est donc composé de cinq variables probabilistes qui sont le résultat de l'hypothèse d'internalisation d'une situation de communication. Outre le fait qu'elles soient symbolisées par les lettres O_S, O_L, S, M et C, qui permet d'écrire l'acronyme COSMO, ces cinq variables forment surtout la distribution conjointe $P(C\ O_L\ S\ M\ O_S)$, c'est-à-dire la distribution globale du modèle.

4.2.2.2 Description des distributions du modèle

Une fois l'espace décrit par la distribution conjointe caractérisé, il faut définir les distributions de probabilité du modèle. Ces distributions suivent les règles classiques du calcul probabiliste telles que la règle de normalisation, la règle du produit, la règle de marginalisation ou le théorème de Bayes (voir par exemple un rappel de ces règles dans Laurent, 2014, section 2.2).

Par exemple, en suivant la règle du produit, la distribution conjointe $P(C\ O_L\ S\ M\ O_S)$ peut se décomposer en une suite de distributions :

$$P(C\ O_L\ S\ M\ O_S) = P(C)\ P(O_L \mid C)\ P(S \mid C\ O_L)\ P(M \mid C\ O_L\ S)\ P(O_S \mid C\ O_L\ S\ M)\ . \quad (4.1)$$

Mais, comme le montre l'exemple ci-dessus, les distributions résultant de cette décomposition sont, parfois, aussi complexes à traiter que la distribution conjointe elle-même. Pour mieux manipuler cette conjointe, il est possible de faire des hypothèses simplificatrices, nommées hypothèses d'indépendances conditionnelles. Au lieu de considérer que toutes les variables sont dépendantes les unes des autres, nous supposons que certaines d'entre elles sont indépendantes et qu'il est possible de connaître leur probabilité conditionnellement à un nombre limité de variables. Ainsi, la conjointe se décompose en une suite de distributions conditionnelles préalablement choisies.

Outre le fait qu'elles dictent la composition de notre modèle, ces distributions ont un rôle bien particulier, notamment dans cette thèse. Le modèle COSMO étant ici envisagé comme un modèle computationnel cognitif, nous supposons que ce sont principalement ces distributions qui sont connues, stockées en mémoire et, pour certaines, apprises par un agent communicant. La décomposition choisie pour le modèle COSMO est la suivante :

$$P(C\ O_L\ S\ M\ O_S) = P(O_S)\ P(M \mid O_S)\ P(S \mid M)\ P(O_L \mid S)\ P(C \mid O_S\ O_L)\ . \quad (4.2)$$

L'ensemble du modèle et ses relations sont schématisés Fig. 4.8.

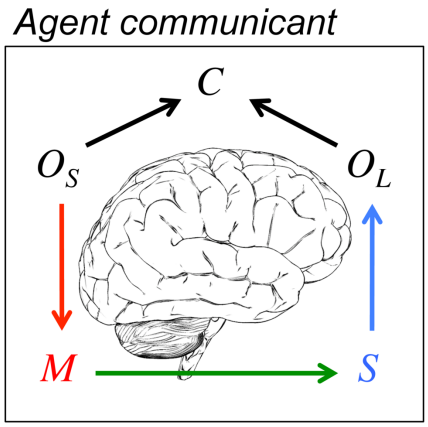


FIGURE 4.8 – Schéma du modèle COSMO

La première distribution $P(O_S)$ est le « prior sur les objets moteurs » ou « prior catégoriel moteur ». Cette distribution suppose que nous avons en mémoire une connaissance préalable sur les objets, indépendamment de toute autre variable. Cette distribution sert, entre autres, à exprimer la fréquence relative de chaque objet moteur du modèle et à indiquer que tel objet est plus probable qu'un autre.

La deuxième distribution $P(M \mid O_S)$ est le répertoire moteur. Elle correspond aux connaissances qu'a l'agent sur la relation entre les représentations motrices et les objets. Nous considérons que ces

connaissances sont contenues sous la forme d'un répertoire, formalisant le fait que l'agent possède en mémoire une distribution indiquant, pour chaque objet, la probabilité des représentations motrices.

La troisième distribution, $P(S \mid M)$, est le modèle interne. Elle correspond à une première application de l'indépendance conditionnelle puisque nous supposons ici qu'un signal acoustique est essentiellement causé par un geste moteur, et que connaître l'objet causant ce geste moteur n'apporte pas d'information supplémentaire. Concernant la probabilité elle-même, nous considérons que la relation entre les représentations sensorielles et les représentations motrices se fait à travers ce qui est généralement nommé un modèle interne direct. Ce type de modèle est souvent utilisé dans la littérature (Kawato, 1999; Wolpert et al., 1998) et également dans plusieurs modèles de production phonétique comme, par exemple, celui de Houde et Nagarajan (2011) vu en section 4.1.1.2. Dans le modèle COSMO, cela correspond à la distribution de probabilité des représentations sensorielles sachant les représentations motrices. Nous supposons ainsi que l'agent a en mémoire la probabilité des représentations sensorielles correspondant à chaque représentation motrice. Étant la seule distribution du modèle liant les représentations sensorielles aux représentations motrices, cela implique que le modèle ne stocke pas les représentations motrices sachant les représentations sensorielles, c'est-à-dire un modèle inverse, que l'on trouve aussi parfois dans la littérature. Celui-ci peut être calculé, mais il n'est pas stocké en mémoire.

La quatrième distribution $P(O_L \mid S)$ est le classifieur auditif, donnant la probabilité de l'objet O_L sachant les représentations sensorielles S. Cette distribution suppose que l'agent a en mémoire la probabilité des objets auditifs pour chaque représentation sensorielle. C'est donc une distribution permettant de catégoriser les représentations sensorielles. Comme le modèle interne, c'est une distribution également assez classique, qui se retrouve dans plusieurs modèles phonétiques, notamment les modèles de perception, comme ceux de Kleinschmidt et Jaeger (2011, 2015) ou Norris et McQueen (2008) comme cela a été illustré dans la section 4.1.1.3.

La cinquième distribution $P(C \mid O_S\ O_L)$ est le système de cohérence. Cette distribution permet de lier, si nécessaire, les objets catégoriels de la branche motrice O_S avec les objets catégoriels de la branche auditive O_L. Elle permet de déterminer l'état de l'interrupteur C. Si l'interrupteur est non activé, les deux objets ne sont pas liés et ils sont traités de façon totalement indépendante. Si l'interrupteur est activé, les deux objets sont liés. Dans ce cas, la probabilité de la variable de cohérence vaut 1 si, et seulement si, les deux objets sont identiques. Ils correspondent alors au même objet O. Ce système est intéressant puisqu'il nous permet de traiter, dans un même modèle, les objets indépendamment selon leur composante auditive ou motrice ou, au contraire, de les percevoir comme des objets perceptuo-moteur. D'un point de vue cognitif, il ne semble pas aberrant d'imaginer que la catégorisation se fait de manière séparée selon la modalité et d'imaginer un système de plus haut niveau capable d'intégrer ces différentes catégorisations aboutissant à un objet linguistique tel que nous le connaissons.

4.2.3 Description de l'apprentissage de COSMO

Dans cette partie, nous décrivons les étapes de l'apprentissage sans nous focaliser sur la nature des distributions. Comme nous l'avons fait dans la section précédente, nous préférons pour le moment

nous focaliser sur les mécanismes et les différentes phases de l'apprentissage. Comme précédemment, nous nous concentrons sur l'apprentissage de COSMO générique.

Dans le modèle COSMO, l'apprentissage se fait en trois temps. Dans un premier temps, nous réalisons un apprentissage sensoriel qui, comme nous l'avons supposé pour le développement du bébé et comme nous l'avons vu dans les modèles de développement, correspond à l'acquisition des représentations sensorielles. Ensuite, nous séparons l'apprentissage moteur en deux temps avec, d'une part, un apprentissage sensorimoteur dans lequel l'agent apprend son modèle interne et, d'autre part, un apprentissage moteur dans lequel l'agent apprend ses représentations motrices. Cela correspond aux deux phases que nous avons relevées précédemment dans les modèles de développement. Nous préférons les distinguer et les présenter à part dans notre modèle.

L'apprentissage de COSMO étant essentiellement un apprentissage social, nous commençons par présenter dans cette section comment les signaux de l'environnement sont fournis à l'agent apprenant. Ensuite, nous détaillons successivement les trois étapes d'apprentissage du modèle.

4.2.3.1 La présence d'un maître

Précédemment, nous avons vu que l'apprentissage du bébé est modulé par l'environnement. Plus exactement, les processus d'acquisition de perception et de production du bébé sont influencés par les signaux de son environnement (voir section 3.2.2), ce qui permet au bébé d'apprendre à percevoir et produire principalement les unités phonétiques spécifiques à sa langue. C'est pourquoi, il est important que notre agent puisse recevoir, durant son apprentissage, des signaux caractéristiques d'une langue donnée ou, au moins, les signaux des catégories phonétiques que nous souhaitons lui faire apprendre.

Pour cela, « l'environnement » correspond dans notre modèle à un agent COSMO particulier, nommé maître. Pour le distinguer de l'agent apprenant, ses distributions sont notées avec la mention $Maitre$ en exposant. Par exemple le prior catégoriel moteur $P(O_S)$ s'écrit pour le maître $P(O_S^{Maitre})$. Néanmoins, cela ne change en rien la nature des distributions et il reste un agent COSMO, comme l'agent apprenant. Cet agent maître peut s'apparenter à une personne sachant déjà communiquer : un tuteur, un parent, un membre de la famille, etc. Pour simplifier notre processus d'apprentissage, nous n'utilisons, dans nos études, qu'un unique agent maître que nous supposons commun à chaque simulation. L'agent maître possède des distributions de mêmes formes et de même nature qu'un agent apprenant. La seule différence est que les distributions du maître sont déjà configurées (le processus de développement est supposé déjà avoir eu lieu pour le maître à une étape préalable), contrairement à celles de l'agent.

Durant chaque étape de l'apprentissage, cet agent maître est chargé de fournir à l'agent apprenant des données de parole lui permettant d'apprendre ses distributions. Ce sont ces données qui permettront à l'agent apprenant de mettre à jour ses distributions de manière à apprendre les unités distinctives de son maître. Afin d'éviter des biais d'apprentissage associés à la caractérisation des signaux acoustiques, nous décidons que l'agent est parfaitement capable d'entendre son maître et que les productions des deux agents sont préalablement normalisées, c'est-à-dire ramenés à un triangle vocalique de référence. De plus, nous supposons que l'agent est capable de segmenter parfaitement les signaux et qu'il les traite les uns après les autres. Cela évite tous les problèmes inhérents à la

segmentation du signal, qui ne font pas partie du programme de cette thèse.

L'apprentissage est découpé en pas d'apprentissage pendant lesquels le maître fournit une donnée sensorielle à l'agent. Cela permet d'avoir un apprentissage itératif, ce que nous supposons primordial pour l'apprentissage phonétique. À chaque itération, la production du maître s'effectue de la manière suivante : le maître choisit un objet o à communiquer grâce à son prior catégoriel moteur $P(O_S^{Maitre})$. Ensuite, il utilise son répertoire moteur $P(M^{Maitre} \mid [O_S^{Maitre} = o])$ grâce auquel il tire une représentation motrice m correspondant à cet objet o, puis la produit. Cette production est transformée dans l'environnement et perçue par l'agent apprenant sous forme de signal auditif. Nous modélisons cette transformation également par une distribution de probabilité $P(S^{Env} \mid M^{Env})$, ce qui nous permet de faire varier la qualité du signal. Par exemple, cela nous permet de modéliser un signal bruité. En termes de correspondance, nous considérons actuellement dans le modèle que les représentations motrices M^{Maitre} sont équivalentes à la réalisation des gestes moteurs M^{Env} produits dans l'environnement. De la même manière, le signal auditif S^{Env} est dans notre modèle directement perçu comme une représentation sensorielle S sans traitement acoustique ou auditif préalable par l'agent apprenant. Ceci est illustré Fig. 4.9. Il s'agit bien entendu de simplifications. Plus techniquement, cela suppose qu'il existe deux variables et deux systèmes de cohérence pour assurer respectivement l'équivalence entre M^{Env} et M^{Maitre} et entre S et S^{Env}, comme il en existe un pour assurer l'équivalence entre O_S et O_L. Pour ne pas alourdir les calculs, nous ne les exprimons pas.

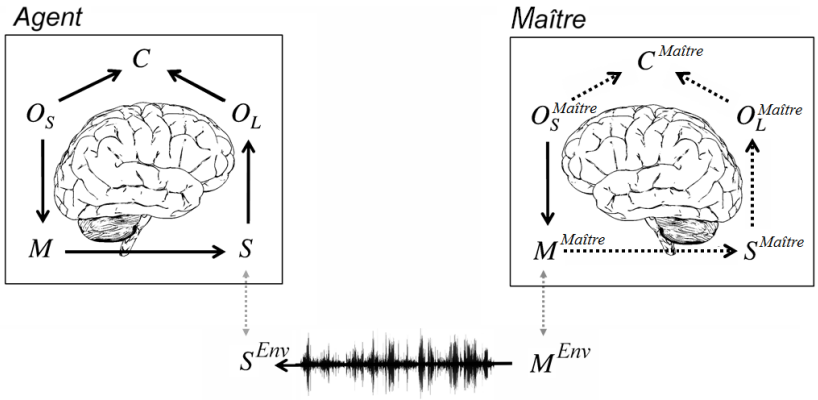


FIGURE 4.9 – Illustration de la production d'un son dans l'environnement par le maître. Les distributions non détaillées du maître sont notées en pointillés. Les équivalences entre les variables M et M^{Env} d'une part et S et S^{Env} sont marquées par une double flèche

Les formes paramétriques du répertoire moteur $P(M^{Maitre} \mid O_S^{Maitre})$ et de la distribution de l'environnement $P(S^{Env} \mid M^{Env})$ dépendant des simulations effectuées, sont précisées dans les chapitres suivants. Pour le moment, nous considérons simplement leurs formes génériques, qui sont suffisantes pour expliquer les étapes d'apprentissage de l'agent. Ces étapes, que nous allons maintenant décrire, sont présentées dans la Fig. 4.10. Pour plus de détails techniques, le lecteur pourra également se référer à l'annexe 9.1.

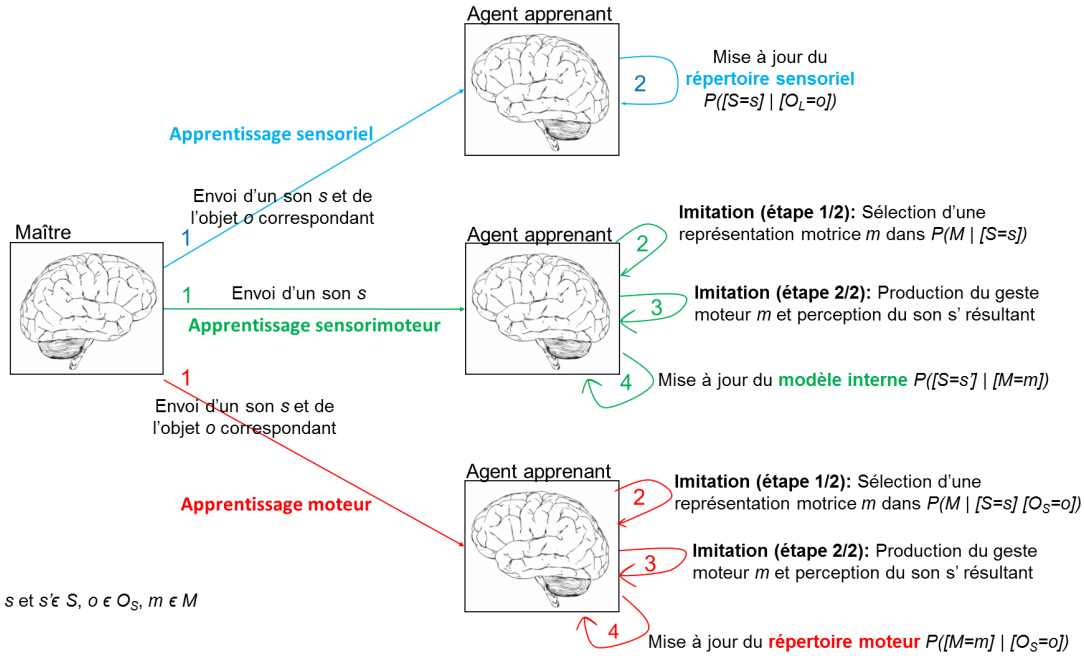


FIGURE 4.10 – Synthèse des phases d'apprentissage

4.2.3.2 L'apprentissage sensoriel

Durant l'apprentissage sensoriel, s'effectue la catégorisation sensorielle. Dans celle-ci, l'agent apprenant COSMO apprend à relier ses représentations auditives aux catégories phonétiques. C'est typiquement l'apprentissage effectué dans plusieurs modèles de développement précédemment présentés (de Boer et Kuhl, 2003; McMurray et al., 2009; Vallabha et al., 2007). Dans notre modèle, cet apprentissage revient à apprendre la relation entre les représentations auditives S et les objets sensoriels correspondants O_L, ce qui correspond à l'apprentissage du classifieur sensoriel $P(O_L \mid S)$. Apprendre cette distribution de façon directe ne nous semble pas chose aisée. Nous avons donc décidé de décomposer le problème en utilisant une astuce de la programmation bayésienne : les sous-programmes. Brièvement, un sous-programme consiste à calculer une distribution en la définissant elle-même comme la question d'un sous-modèle bayésien, composé de distributions plus simples à évaluer. Dans notre cas, nous considérons donc le sous-modèle composé de la conjointe $P(O_L\ S)$ qui se décompose par :

$$P(O_L\ S) = P(O_L)\ P(S \mid O_L)\ , \tag{4.3}$$

dans laquelle $P(O_L)$ représente un prior sur les objets sensoriels, autrement dit la probabilité des objets sensoriels et $P(S \mid O_L)$ représente le répertoire sensoriel. À l'aide de ces deux distributions, nous pouvons calculer le classifieur sensoriel de COSMO par l'inférence suivante :

$$P(O_L \mid S) = \frac{P(O_L)\ P(S \mid O_L)}{\sum_{O_L} P(O_L)\ P(S \mid O_L)}\ . \tag{4.4}$$

Nous supposons donc que le classifieur sensoriel est calculé et appris à partir de ces deux distribu-

tions qui sont des sous-connaissances de notre modèle. C'est donc sur elles que nous nous focalisons pour réaliser l'apprentissage sensoriel. Dans notre modélisation, la fréquence des unités distinctives, c'est-à-dire le prior des objets sensoriels, ne nous intéresse pas. Nous considérons donc que $P(O_L)$ est uniforme et ne s'apprend pas durant l'apprentissage. En revanche, l'apprentissage du répertoire sensoriel nous semble essentiel pour apprendre correctement la relation entre les représentations sensorielles et les objets. L'agent va donc se focaliser sur l'apprentissage de son répertoire auditif $P(S \mid O_L)$ préalablement défini.

Dans toutes nos simulations, la distribution $P(S \mid O_L)$, composée de nb_o objets O_L, correspond à un ensemble de nb_o gaussiennes de moyenne μ et de matrice de covariance σ, chaque gaussienne étant associée à un objet o. Ainsi, dès le début de l'apprentissage, l'agent apprenant et le maître ont le même nombre de catégories et donc de gaussiennes. Durant l'apprentissage, à chaque itération, l'agent apprenant reçoit le signal s produit par le maître ainsi que l'objet o sélectionné par le maître pour produire ce signal. L'apprentissage sensoriel est alors assez direct : l'agent met à jour les paramètres de la gaussienne correspondant à l'objet o avec le signal sonore s perçu. C'est donc un apprentissage itératif entièrement supervisé.

Le choix d'utiliser des gaussiennes n'est pas anodin. Comme nous l'avons vu précédemment en section 4.1.2.1, il s'agit d'une des distributions les plus utilisées pour réaliser l'apprentissage sensoriel catégoriel (de Boer et Kuhl, 2003; Feldman et al., 2013a; McMurray et al., 2009; Vallabha et al., 2007). Ces distributions permettent également de mettre en avant des phénomènes sensoriels bien connus dans la littérature tels que la perception catégorielle, le perceptual narrowing ou encore le perceptual magnet effect (Feldman et al., 2009a; Kleinschmidt et Jaeger, 2015).

Comme dans la majorité des modèles de développement de la perception, nous choisissons de réaliser un apprentissage sensoriel uniquement basé sur les signaux du maître. En effet, nous supposons qu'au départ l'apprentissage sensoriel est dépendant des signaux de l'environnement. Il est possible que, par la suite, l'apprentissage sensoriel se spécialise sur les signaux produits par l'agent lui-même mais nous ne le prenons pas en compte dans notre apprentissage. De même, nous ne prenons pas en compte la possible perception interne pendant laquelle l'agent imaginerait des sons dans sa tête et apprendrait à partir de là les catégories correspondantes.

Si l'utilisation du signal sensoriel est courante dans la littérature, le fait de donner l'objet o l'est beaucoup moins. Il peut néanmoins se justifier par la deixis, par exemple, en pointant du doigt un référent commun, grâce à laquelle le maître et l'agent se mettent d'accord sur l'objet désigné (Moulin-Frier et al., 2015). Ce mécanisme nécessite d'avoir un même nombre de catégories entre le maître et l'agent, comme dans le modèle de de Boer et Kuhl (2003). D'un point de vue technique, cela permet d'assurer que le nombre de gaussiennes en fin d'apprentissage correspond au nombre de catégories. Cela facilite la comparaison entre les catégories du maître et celles de l'agent et entre celles du répertoire sensoriel et du répertoire moteur. Cet apprentissage supervisé, aboutissant à un nombre d'objets communs entre l'agent et le maître, accélère donc l'apprentissage. Néanmoins, il peut sembler trop simplifié. C'est pourquoi, nous proposons, dans l'extension de COSMO, au chapitre 6, un apprentissage non supervisé dans lequel non seulement le maître ne fournit plus l'objet à l'agent mais dans lequel l'agent et le maître ne possèdent plus, non plus, le même nombre de catégories.

4.2.3.3 L'apprentissage sensorimoteur

L'apprentissage sensorimoteur consiste à faire apprendre à l'agent apprenant la relation entre ses représentations sensorielles et ses représentations motrices. Il s'agit de la phase sensorimotrice précédemment décrite dans les modèles de développement en section 4.1.2.2 et qui s'effectue, principalement, par exploration. Dans notre modèle, cela correspond à l'apprentissage du modèle interne $P(S \mid M)$ reliant les représentations motrices M aux représentations sensorielles S.

L'apprentissage sensorimoteur est plus complexe que l'apprentissage sensoriel puisque l'agent doit manipuler des représentations motrices M qui ne sont pas fournies par le maître. Pour pallier cette difficulté, nous utilisons un algorithme d'apprentissage nommé processus d'accommodation grâce auquel l'agent peut inférer les gestes moteurs à l'aide des données fournies par le maître. Comme lors de l'apprentissage sensoriel, à chaque itération, le maître produit un son s pour l'agent. Ensuite, l'agent infère un geste moteur grâce à sa distribution $P(M \mid [S = s])$:

$$P(M \mid [S = s]) \propto \sum_{O_S} P(O_S)P(M \mid O_S)P(S \mid M) \,, \tag{4.5}$$

inférant ainsi à l'aide de ses connaissances motrices une représentation motrice m. Tout comme $P(O_L)$, nous considérons que $P(O_S)$ est constamment uniforme.

À l'aide de cette inférence sur la distribution $P(M \mid [S = s])$, l'agent tire une représentation motrice m puis la produit dans l'environnement à l'aide de la distribution $P(S^{Env} \mid [M^{Env} = m])$, qui est la même que celle utilisée pour produire les signaux auditifs du maître (se référer à la Fig 4.9 en supposant que cette fois-ci, c'est l'agent qui produit). Le son résultant, perçu comme la représentation sensorielle s', et la représentation motrice inférée m, qui a servi à la produire, sont ensuite tous deux utilisés pour mettre à jour le modèle interne $P(S \mid M)$.

L'apprentissage du modèle interne est donc un apprentissage semi-supervisé mélangeant un processus d'exploration dans lequel l'agent apprend tout ce qu'il produit, avec un processus imitatif dans lequel les stimuli qu'il tente d'imiter sont ceux fournis par son maître. Ainsi, l'apprentissage par accommodation se classe, d'une part, parmi les apprentissages d'objectifs (Moulin-Frier et Oudeyer, 2013; Philippsen et al., 2016; Rolf et al., 2010) et, d'autre part, parmi les apprentissages d'exploration par guidage social (Philippsen et al., 2014; Warlaumont, 2012; Warlaumont et al., 2013).

4.2.3.4 L'apprentissage moteur

La dernière étape d'apprentissage correspond à ce que nous nommons l'apprentissage « moteur » et qui correspond à la deuxième phase d'apprentissage présenté dans la section 4.1.2.2. Elle correspond à la focalisation des représentations motrices sur les unités linguistiques de la langue donnée. Dans le modèle COSMO, cela correspond à l'acquisition du lien entre les représentations motrices et les objets moteurs c'est-à-dire à l'apprentissage du répertoire moteur $P(M \mid O_S)$.

Dans le modèle COSMO, cet apprentissage se fait, comme l'apprentissage sensorimoteur, par accommodation. Cependant, cette fois-ci, l'agent reçoit de son maître non seulement le stimulus s mais

également l'objet o que le maître communique. Ainsi, la première étape de l'apprentissage d'accommodation consiste à trouver une représentation motrice m correspondant à la représentation sensorielle s et à l'objet o, ce qui se fait par l'inférence suivante :

$$P(M \mid [S = s]\ [O_S = o]) \propto P(O_S)P(M \mid O_S)P(S \mid M)\ . \tag{4.6}$$

Du fait que $P(O_S)$ est constamment uniforme, nous pouvons l'enlever du calcul, ce qui donne

$$P(M \mid [S = s]\ [O_S = o]) \propto P(M \mid O_S)P(S \mid M)\ . \tag{4.7}$$

Ainsi, l'inférence se fait donc avec le répertoire moteur et le modèle interne de l'agent. A l'aide de cette inférence, l'agent tire une représentation motrice m. Avec celle-ci et l'objet fourni par le maître, il met ensuite son répertoire moteur $P(M \mid O_S)$ à jour.

Cet apprentissage est donc, comme l'apprentissage sensorimoteur, semi-supervisé. De plus, il est, comme cela est classique dans la littérature, guidé par un apprentissage social. Nous avons opté pour un processus d'imitation plutôt qu'un apprentissage miroir (voir la distinction pour rappel dans Messum et Howard, 2015) car nous considérons que l'apprentissage sensorimoteur réalisé précédemment est suffisant pour inférer une représentation motrice correspondant à l'objet o. Le fait que l'objet o soit fourni par le maître est moins courant dans la littérature et peut être considéré comme simplifié mais nous l'appliquons pour des raisons similaires à l'apprentissage sensoriel et pour rester consistant.

Comme pour l'apprentissage sensorimoteur, une étude détaillée dans les chapitres suivants compare plusieurs variantes de l'apprentissage moteur. Quand cela n'est pas précisé, c'est l'apprentissage décrit ici qui est appliqué. De plus, comme pour l'apprentissage sensoriel, l'apprentissage moteur utilisé dans le modèle COSMO du chapitre 6, bien que basé sur le même type d'apprentissage, s'effectue de manière non supervisée, sans que l'objet o soit fourni par le maître.

4.2.4 Modélisation des trois familles de théories de la perception

Pour étudier la complémentarité des voies auditive et motrice lors de la perception de la parole, nous modélisons avec le modèle COSMO les trois familles de théories de la perception : auditive, motrice et perceptuo-motrice, à l'aide des distributions du modèle. Cela nous permet, d'une part, de comparer les trois familles de théories dans un même cadre computationnel et, d'autre part, d'analyser le rôle de l'apprentissage dans la perception.

Afin de représenter les trois familles de théories, nous devons d'abord définir au préalable ce que représente une tâche de perception. En terme expérimental, cela consiste à demander à un sujet de catégoriser des sons. Dans le modèle COSMO, on suppose que cela correspond au calcul de l'objet o le plus probable d'un décodeur sachant des informations sensorielles S. En d'autres termes, il s'agit de calculer $o = \max(P(O \mid S))$. Dans notre modèle COSMO, nous avons une unique variable sensorielle S mais deux variables pour les objets O_L et O_S. Il existe donc plusieurs décodeurs pour réaliser notre tâche de perception :

— $P(O_L \mid S)$ représentant la probabilité des objets O_L sachant le signal sensoriel S,

— $P(O_S \mid S)$ représentant la probabilité des objets O_S sachant le signal sensoriel S,
— $P(O_L \mid S\,[C=1])$ représentant la probabilité de l'objet O_L sachant le signal sensoriel S, et sous la contrainte $O_S = O_L$. Ici, la variable C permet d'assurer le fait que O_L et O_S soient identiques. Le classifieur $P(O_L \mid S\,[C=1])$ peut donc aussi s'écrire $P(O_S \mid S\,[C=1])$.

Ces trois décodeurs correspondent à une question dans notre modèle (voir section 4.2.1). Nous calculons chacune d'elles par inférence bayésienne en nous servant des distributions apprises du modèle.

L'équation du décodeur $P(O_L \mid S)$ est :

$$P(O_L \mid S) = P(O_L \mid S)\ . \tag{4.8}$$

Cette équation peut sembler triviale. En effet, la partie gauche et la partie droite correspondent à la même distribution. Cependant, elles sont sémantiquement différentes. Ici, la partie gauche correspond au décodeur que nous souhaitons calculer qui peut se résumer par : « je souhaite trouver la probabilité des objets O_L sachant les signaux sensoriels S que j'ai perçu ». La partie droite, quant à elle, correspond à la distribution apprise du modèle qui sert à répondre à cette question, ici, le classifieur auditif de notre agent. Nous sommes dans le cas particulier où notre question correspond à une des représentations internes du modèle. Malgré sa trivialité, cette équation nous permet de voir que le modèle utilise uniquement son classifieur auditif pour décoder l'information, c'est-à-dire uniquement ses représentations auditives. Ainsi, le décodeur $P(O_L \mid S)$ peut être utilisé pour modéliser une tâche de perception selon les théories auditives.

L'équation du décodeur $P(O_S \mid S)$ est :

$$P(O_S \mid S) \propto \sum_M P(M \mid O_S)\ P(S \mid M)\ . \tag{4.9}$$

Cette équation montre que le calcul de la probabilité des objets O_S correspondant aux signaux sensoriels S fait appel uniquement au répertoire moteur $(M \mid O_S)$ et au modèle interne $P(S \mid M)$ de l'agent, révélant ainsi que le calcul du décodeur $P(O_S \mid S)$ ne nécessite que les représentations motrices de l'agent. Cette distribution est donc utilisée pour modéliser une tâche de perception selon les théories motrices. Formellement, les facteurs $P(S \mid M)$ et $P(M \mid O_S)$ permettent de traiter respectivement les deux enjeux des modèles de la théorie motrice : l'inversion articulatori-acoustique et le décodage articulatoire. L'équation 4.9 fournit donc une implémentation bayésienne de l'ensemble du processus, remplaçant les algorithmes traditionnels de choix d'un unique antécédent articulatoire avant décodage, par une sommation sur l'ensemble des configurations articulatoires possibles, pondérées par leur vraisemblance.

Enfin, l'équation du décodeur $P(O_L \mid S\,[C=1])$ est :

$$\begin{aligned} P(O_L \mid S\,[C=1]) &\propto P(O_L \mid S) \sum_M P(M \mid O_S)\ P(S \mid M)\ , \\ P(O_L \mid S\,[C=1]) &\propto P(O_L \mid S)\ P(O_S \mid S)\ . \end{aligned} \tag{4.10}$$

Cette équation est une fusion des deux décodeurs précédents. La distribution $P(O_L \mid S\,[C=1])$ utilise les représentations auditives du modèle à travers le décodeur $P(O_L \mid S)$ et les représentations motrices à travers le décodeur $P(O_S \mid S)$. Elle est donc utilisée pour modéliser une tâche de perception selon les théories perceptuo-motrices.

4.3 Conclusion

Dans la première section de ce chapitre, nous avons fait une revue de la littérature des modèles computationnels étudiant les représentations sensorielles et/ou motrices des unités phonétiques. Ceci nous a permis de faire le point sur les résultats actuels et de définir quelques questions qu'il semble intéressant d'approfondir, comme, par exemple, le rôle des représentations sensorielles et motrices en perception. Cette revue nous a également servi de contexte pour présenter le modèle computationnel COSMO, qui est le modèle utilisé durant cette thèse. Nous avons détaillé le modèle lui-même, son apprentissage et la modélisation des tâches de perception, adaptée à chacune des famille de théories de la perception.

Le cadre théorique étant posé et le modèle défini, nous pouvons, désormais, décrire l'ensemble des études que nous avons réalisées.

CHAPITRE 5

Études des étapes de l'apprentissage

Afin de mieux comprendre l'acquisition des représentations sensorielles et motrices des unités distinctives, ce chapitre est dédié à l'étude des différentes étapes de l'apprentissage : l'apprentissage sensoriel, l'apprentissage sensorimoteur et l'apprentissage moteur. Ces étapes ont fait l'objet de deux études spécifiques, réalisées à l'aide du modèle COSMO, présenté dans le chapitre précédent. Dans un premier temps, nous analysons l'apprentissage sensoriel et l'apprentissage moteur afin d'étudier leurs propriétés et leur rôle dans la structuration des représentations sensorielles et motrices lors de la perception de la parole. Dans un second temps, nous nous focalisons sur l'apprentissage sensorimoteur, indispensable pour faire le lien entre les représentations sensorielles et motrices, et comparons huit algorithmes basés sur différents principes d'apprentissage afin de tester leur efficacité en termes d'exploration et de qualité d'apprentissage.

5.1 Le rôle des représentations auditives et motrices en perception

Publications :

— Barnaud, M.-L., Laurent, R., Bessière, P., Diard, J., et Schwartz, J.-L. (2015c). Modeling concurrent development of speech perception and production in a Bayesian framework. In Workshop on Infant Language Development (WILD), Stockholm, Sweden

— Barnaud, M.-L., Diard, J., Bessière, P., et Schwartz, J.-L. (2015a). Modeling the concurrent development of speech perception and production in a Bayesian framework. In The 5th Joint IEEE International Conference on Developmental and Learning and on Epigenetic Robotics (ICDL-Epirob 2015), pages 248–249. Poster

— Barnaud, M.-L., Diard, J., Bessière, P., et Schwartz, J.-L. (2015b). Modeling the concurrent development of speech perception and production in a Bayesian framework. In Workshop on Probabilistic Inference and the Brain. Poster

— Laurent, R., Barnaud, M.-L., Schwartz, J.-L., Bessière, P., et Diard, J. (2017). The complementary roles of auditory and motor information evaluated in a Bayesian perceptuo-motor model of speech perception. Psychological Review, 124(5):572–602, section 2

Comment notre cerveau interprète-t-il le signal sonore en tant qu'unités phonétiques durant une tâche de perception ? Ce questionnement est le point de départ de cette première partie. Pour trouver une réponse, nous nous sommes intéressés aux théories de la perception qui, comme nous l'avons vu

dans le chapitre 2, étudient la caractérisation de l'invariant phonétique.

Pour rappel, ces théories peuvent globalement être regroupées en trois familles : les théories auditives, les théories motrices et les théories perceptuo-motrices. La nature de l'invariant phonétique a longtemps été débattue au travers des deux premières familles de théories. Grâce à l'avancée des techniques en neuroimagerie, un consensus est peu à peu apparu : les voies auditives et motrices seraient toutes deux utilisées durant la perception de la parole, donnant ainsi raison aux théories perceptuo-motrices. Néanmoins, même si la perception est réalisée à l'aide de ces deux voies, une question subsiste : quels sont leurs rôles respectifs ? À notre connaissance, les études comportementales, neuroscientifiques ou computationnelles n'ont pas tranché cette question. C'est donc la problématique que nous choisissons d'étudier.

À l'aide du modèle COSMO, nous implémentons ces trois familles de théories et les comparons. Cela nous permet de définir ce que nous nommons la propriété « bande étroite/bande large », propriété mettant en avant la différence de comportement des branches auditive et motrice face à un même stimulus auditif. Nous désignons sous le terme « bande étroite », la précision de la branche auditive et sous le terme « bande large » la capacité de généralisation de la branche motrice ; deux caractéristiques dont l'origine dépend des différences d'apprentissage.

Précisons, dès lors, que le modèle COSMO utilisé dans cette étude est un modèle minimaliste, nous permettant d'évaluer nos hypothèses dans un cadre très simplifié. Néanmoins, comme il est discuté par la suite, les résultats obtenus dans cette étude n'en sont pas moins génériques et devraient pouvoir être généralisés à d'autres modèles plus complexes.

5.1.1 Hypothèse de l'étude

Dans cette étude, nous supposons que l'invariant phonétique est perceptuo-moteur et que la perception fait appel à deux voies, l'une auditive et l'autre motrice. Nous souhaitons comprendre les rôles que peuvent jouer chacune de ces voies dans la perception. Une hypothèse possible serait de supposer qu'elles sont similaires voire redondantes : elles traiteraient de la même manière les informations reçues. D'un côté, la redondance permettrait d'assurer la perception même en cas de défaillance d'une des deux voies. D'un autre côté, elle pourrait aussi avoir un rôle de consolidation pour lequel les informations perçues par l'une des voies seraient confirmées par les informations perçues par la seconde.

Cette similarité entre les deux branches est le résultat obtenu par le modèle COSMO lorsque les apprentissages sensoriel et moteur sont réalisés dans des conditions dites « parfaites ». Il s'agit d'un théorème « d'indistinguabilité » puisque, sous ces conditions, les branches auditive et motrice du modèle contiennent exactement les mêmes informations et sont donc totalement indistinguables l'une de l'autre (Laurent, 2014). Les conditions « parfaites » d'apprentissage reposent sur trois hypothèses :

— un apprentissage parfait du classifieur auditif, c'est-à-dire un classifieur auditif ayant parfaitement appris la distribution des stimuli produits par le maître en fin d'apprentissage.
— un apprentissage parfait du modèle interne, c'est-à-dire un modèle interne identique à la transformation articulatori-acoustique de l'environnement.
— un apprentissage parfait du répertoire moteur, c'est-à-dire un répertoire moteur identique à celui du maître en fin d'apprentissage.

Bien entendu, ces trois hypothèses ne sont pas réalistes puisqu'il semble peu vraisemblable qu'un bébé puisse apprendre parfaitement son environnement et les représentations internes de ses tuteurs. Toutefois, du fait que ces conditions sont le seul moyen pour que les deux branches soient totalement identiques, un apprentissage réaliste suppose donc que les deux branches diffèrent et qu'elles ne contiennent pas exactement les mêmes informations. C'est pourquoi, même s'il existe indéniablement un certain niveau de redondance entre les deux voies, nous pouvons supposer qu'elles ne sont pas entièrement similaires, et la question que nous nous posons est celle de leurs spécificités respectives.

Du point de vue des neurosciences, un des faits les plus marquants allant à l'encontre d'une hypothèse de totale redondance est la plus grande activation des aires motrices dans des conditions adverses, par exemple dans du bruit ou lors d'une communication avec une personne ayant un accent différent du nôtre (voir section 3.1.1.2). Pourquoi les aires motrices seraient-elles plus actives dans des conditions adverses si les deux voies sont redondantes ? L'activation plus importante des aires motrices dans ces conditions laisse, au contraire, penser que la voie motrice a des caractéristiques que la voie auditive ne possède pas et réciproquement. C'est pourquoi, nous posons l'hypothèse d'une spécificité de chacune des deux voies.

5.1.2 Implémentation du modèle : COSMO 1D

Dans cette section, nous présentons les détails d'implémentation du modèle COSMO utilisé. Nous commençons par décrire l'espace de définition des variables et la forme paramétrique des distributions du modèle. Ensuite, nous décrivons comment sont initialisées les distributions du maître et de l'agent apprenant. Pour terminer, nous explicitons les détails d'apprentissage de l'agent apprenant.

5.1.2.1 Implémentation des variables du modèle

Afin de pouvoir clairement observer le comportement des distributions au cours de l'apprentissage, nous implémentons une version du modèle COSMO dans laquelle chaque variable est unidimensionnelle, c'est pourquoi elle est nommée COSMO 1D.

Nous nous plaçons dans le cadre donné par la théorie quantique (Stevens, 1989; Stevens et Keyser, 2010) et imaginons un paramètre acoustique quelconque du son provenant du conduit vocal et un paramètre articulatoire de ce conduit vocal. Selon cette théorie, les contrastes phonétiques exploiteraient des régions de l'espace dans lesquelles, lorsque le paramètre articulatoire varie, le paramètre acoustique est relativement stable dans une première portion de l'espace (notée I), puis varie brusquement dans une seconde portion de l'espace (notée II) et enfin redevient à peu près stable dans une troisième portion de l'espace (notée III). Tout ceci est schématisé sur la Fig. 5.1.

Dans ce contexte, nous supposons que les représentations sensorielles S du modèle COSMO correspondent au paramètre acoustique et les représentations motrices M correspondent au paramètre articulatoire. Ce sont toutes les deux des variables finies et discrétisées sur une dimension. On choisit pour ces deux variables l'intervalle $\{-140; +140\}$ avec un pas de discrétisation linéaire de 1 pour les représenter. Ce choix est arbitraire, la seule contrainte est d'avoir un espace suffisamment précis pour

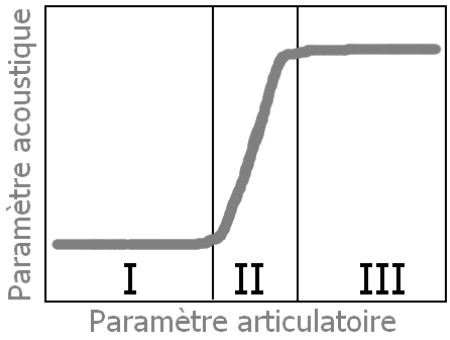


FIGURE 5.1 – Schéma de la relation entre un paramètre acoustique et un paramètre articulatoire selon la théorie quantique (Stevens, 1998, 2010)

pouvoir observer le comportement des distributions. De leur côté, les objets O_L et O_S correspondent à deux catégories phonétiques quelconques, ce qui est suffisant pour faire de la catégorisation. On note ainsi $O_L = \{o^-, o^+\}$ et $O_S = \{o^-, o^+\}$. Pour finir, la variable C est, comme précisé dans le modèle générique, une variable booléenne prenant comme valeur « vrai » (1) ou « faux » (0).

5.1.2.2 Implémentation des distributions du modèle

La distribution prior $P(O_S)$ indiquant la fréquence des objets ne nous intéresse pas dans cette étude. Pour cette raison, nous supposons que les deux objets considérés ont la même fréquence, ce qui nous permet d'implémenter le prior $P(O_S)$ comme une distribution uniforme. Les répertoires moteur $P(M \mid O_S)$ et sensoriel $P(S \mid O_L)$ [1] sont des ensembles de deux gaussiennes, une pour chaque objet. Chaque gaussienne possède une moyenne μ et un écart-type σ. Comme nos espaces sensoriels S et moteurs M sont des espaces discrets et finis, les gaussiennes utilisées dans nos études ne sont en réalité qu'une représentation discrétisée et tronquée des gaussiennes.

Illustrons-ceci avec un exemple : la distribution correspondant à l'objet o^+ du répertoire moteur. Elle se calcule ainsi :

$$Gauss(m) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{m-\mu}{\sigma})^2}, \tag{5.1}$$

$$P([M = m] \mid [O = o^+]) = \frac{Gauss(m)}{\sum_M Gauss}. \tag{5.2}$$

Ainsi, pour cette distribution, la probabilité de chaque point m de l'espace M discrétisé est calculée à l'aide de l'équation de la gaussienne (cf Eq. 5.1) puis l'ensemble de ces valeurs est, par la suite, normalisé pour sommer à 1 (cf Eq. 5.2). Bien entendu, nous illustrons ce cas avec une unique distribution du répertoire moteur mais cela concerne aussi bien les deux distributions du répertoire moteur que les deux distributions du répertoire sensoriel. Par la suite, pour simplifier la notation, nous nommons toute distribution calculée ainsi « distribution gaussienne ». Ainsi, nous pouvons décrire le modèle

1. Rappelons que le classifieur auditif $P(O_L \mid S)$ est calculé à partir d'un répertoire sensoriel $P(S \mid O_L)$. Nous ne nous focalisons donc, dans cette partie, que sur la description de ce répertoire sensoriel.

interne $P(S \mid M)$ comme, lui aussi, un ensemble de distributions gaussiennes. Il possède exactement 281 distributions gaussiennes, une pour chaque valeur de m dans l'intervalle $\{-140; +140\}$.

Terminons avec la distribution $P(C \mid O_S\, O_L)$. Quand la variable C vaut 1, les deux variables O_S et O_L sont connectées. C'est pourquoi, la probabilité de cette distribution vaut 1 si et seulement si les deux objets sont égaux. En revanche, quand C n'est pas spécifié, les deux variables O_S et O_L sont indépendantes (voir Gilet et al., 2011, pour plus de détails sur cette distribution).

5.1.2.3 Implémentation de l'environnement

Comme expliqué dans le chapitre précédent, l'apprentissage, tel qu'il est effectué, nécessite un maître. Pour rappel, le maître est un agent COSMO pour lequel nous ne nous préoccupons que des distributions $P(O_S^{Maitre})$ et $P(M^{Maitre} \mid O_S^{Maitre})$ lui servant à produire des stimuli pour l'agent apprenant. Afin que l'agent apprenant puisse se servir des productions du maître, nous avons également besoin de transformer les représentations motrices en représentations sensorielles perçues par l'agent. Ceci s'effectue par la transformation de la production du maître en stimulus perçu par l'agent, ce qui est représenté par la distribution $P(S^{Env} \mid M^{Env})$.

Plus précisément, les variables du maître sont implémentées de la même manière que celles de l'agent apprenant et représentent les mêmes informations. Ainsi, M^{Maitre} est, comme la variable M, l'espace articulatoire fini et discret dans l'intervalle $\{-140; +140\}$. De son côté, O_S^{Maitre} est, comme O_S, un espace catégoriel prenant les deux valeurs $\{o^-, o^+\}$. Concernant ses distributions, nous considérons, comme pour l'agent apprenant, que les deux objets ont la même fréquence d'apparition. C'est pourquoi $P(O_S^{Maitre})$ est, comme $P(O_S)$, une distribution uniforme. Le répertoire moteur $P(M^{Maitre} \mid O_S^{Maitre})$ du maître est également implémenté de manière similaire à celui de l'agent $P(M \mid O_S)$: il s'agit d'un ensemble de distributions gaussiennes telles qu'elles ont été définies précédemment (cf Eq. 5.1 et Eq. 5.2). Afin d'avoir des représentations motrices pour les deux objets bien séparables, nous choisissons arbitrairement que la distribution gaussienne $P(M^{Maitre} \mid [O_S^{Maitre} = o^-])$ a pour moyenne $\mu = -50$, que la distribution gaussienne $P(M^{Maitre} \mid [O_S^{Maitre} = o^+])$ a pour moyenne $\mu = -50$ et qu'elles ont toutes deux un écart-type $\sigma = 10$.

Concernant la transformation de la réalisation motrice en signal sonore, comme précisé précédemment, nous simplifions les représentations telles que le signal acoustique S^{Env} est équivalent aux représentations sensorielles S perçues par l'agent et la réalisation de la production M^{Env} est équivalente aux représentions motrices M^{Maitre} du maître (pour rappel, voir Fig. 5.2, équivalente à la Fig. 4.9 du chapitre précédent). De ce fait, S^{Env} et M^{Env} correspondent également tous deux à un espace fini et discret dans l'intervalle $\{-140; +140\}$. La transformation articulatori-acoustique $P(S^{Env} \mid M^{Env})$ est, comme le modèle interne de l'agent $P(S \mid M)$, un ensemble de 281 distributions gaussiennes, une pour chaque valeur de m dans l'intervalle $\{-140; +140\}$. Les écart-types de chaque distributions valent $\sigma = 1$ et symbolisent le bruit ambiant de l'environnement, supposé faible ici. Les valeurs des moyennes demandent un peu plus de calcul. En effet, comme nous nous plaçons dans le cadre donné par la théorie quantique (cf section 5.1.2.1), la transformation de la production d'un geste articulatoire m du maître en un signal sonore s perçu par l'agent doit posséder les caractéristiques évoquées dans cette théorie. Celle-ci ayant la forme d'une fonction sigmoïde (cf

Fig. 5.1), les moyennes de chaque distribution gaussienne de $P(S^{Env} \mid M^{Env})$ suivent donc une fonction sigmoïde, $\mu(m) = \frac{b \times tan^{-1}(a \times m)}{tan^{-1}(a \times b)}$. Le point d'origine de cette sigmoïde a été fixé à 0. Dans nos simulations, nous avons testé différentes valeurs de la pente a, allant du cas linéaire (a très petit, en l'occurrence $a = 0{,}01$) au cas non linéaire « à la Stevens » (a plus élevé, en l'occurrence $a = 0{,}1$). La valeur de la borne b est égale à 120 afin de ne pas être biaisé par les limites de notre intervalle, qui sont à 140. Cette implémentation nous permet ainsi de reproduire les trois phases supposées de la théorie.

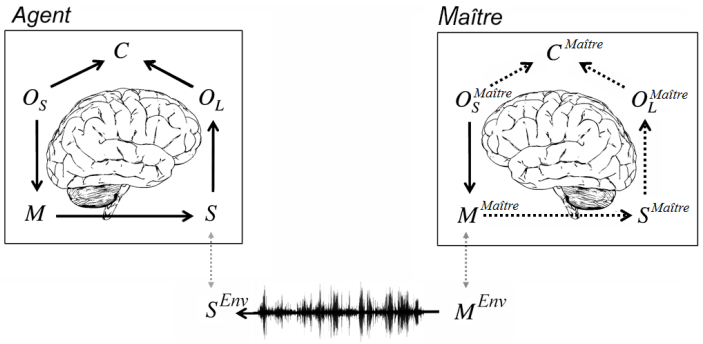


FIGURE 5.2 – Illustration de la production d'un son dans l'environnement par le maître. Les distributions non détaillées du maître sont notées en pointillés. Les équivalences entre les variables M et M^{Env} d'une part et S et S^{Env} sont marquées par une double flèche

Durant l'apprentissage, chaque objet o est sélectionné par le maître l'un après l'autre. À chaque itération, le maître produit un geste moteur m, relatif à l'objet sélectionné o, qui est, par la suite, transformé en signal sonore s dans l'environnement. Cela correspond à tirer un geste articulatoire m sur la distribution $P(M^{Maitre} \mid [O_S^{Maitre} = o])$ puis de tirer un son s sur la distribution $P(S^{Env} \mid [M^{Env} = m])$. Pour faciliter l'implémentation, nous réalisons en réalité un simple tirage sur la distribution $P(S^{Env} \mid O_S^{Maitre})$, calculée à l'avance :

$$P(S^{Env} \mid O_S^{Maitre}) = \sum_M P(S^{Env} \mid M^{Env})\ P(M^{Maitre} \mid O_S^{Maitre})\ . \tag{5.3}$$

Ainsi, lors de chaque itération, le maître choisit un objet o puis tire un signal sensoriel s à l'aide de la distribution $P(S^{Env} \mid O_S^{Maitre})$. Durant cette étude, nous effectuons douze simulations qui ont pour uniques différences les signaux sensoriels s tirés à chaque itération dans $P(S^{Env} \mid O_S^{Maitre})$. Cela nous permet de vérifier la stabilité des simulations. À titre d'illustration, l'ensemble des distributions composant cette équation sont représentées Fig. 5.3.

Bien que plusieurs valeurs de pente a soient testées pour définir $P(S^{Env} \mid M^{Env})$, nous illustrons, dans tout ce qui suit, uniquement les résultats obtenus avec la valeur de a égale à 0,1, dans le cas d'une transformation non-linéaire. En effet, les résultats s'avèrent être tout à fait semblables dans le cas linéaire (voir Laurent et al., 2017, pour plus de détails).

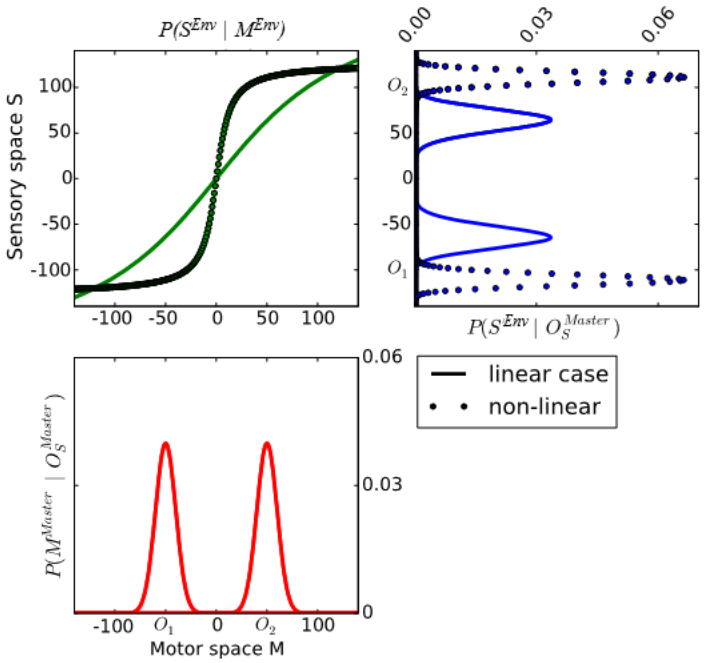


FIGURE 5.3 – Résumé des distributions du maître et de l'environnement. Le répertoire moteur du maître est représenté en bas à gauche (en rouge) et la transformation motrice-à-sensorielle est représentée en haut à gauche (en vert), pour les deux valeurs de a testées. Le résultat de ces deux processus est donné par les deux distributions en haut à droite (en bleu)

5.1.2.4 Implémentation de l'apprentissage du modèle

À l'initialisation, avant apprentissage, nous supposons que l'agent a un état de connaissance maximalement incertain dans ses distributions de probabilités. C'est pourquoi ses distributions $P(M \mid O_S)$, $P(S \mid O_L)$ et $P(S \mid M)$ approximent des distributions uniformes. Nous représentons cela par des moyennes situées au centre de l'espace et possédant un grand écart-type. Cela correspond, dans nos intervalles $[-140; +140]$, à une moyenne $\mu = 0$ et un écart-type $\sigma = 140$.

Ensuite, dans chacune de ces simulations, nous effectuons les trois apprentissages précédemment décrits dans le chapitre 3 : l'apprentissage sensoriel, durant lequel sont mis à jour les paramètres des distributions gaussiennes de $P(S \mid O_L)$, l'apprentissage sensorimoteur, durant lequel sont mis à jour les paramètres des distributions gaussiennes de $P(S \mid M)$ et l'apprentissage moteur, durant lequel sont mis à jour les paramètres des distributions gaussiennes de $P(M \mid O_S)$. Dans cette version, afin de faciliter leur comparaison, ces trois apprentissages sont appris en même temps et à partir des mêmes données. Elles durent chacune 20 000 itérations.

5.1.2.5 Implémentation des décodeurs

Comme nous l'avons vu dans le chapitre 3, les trois familles de théories peuvent être analysées, dans COSMO, à l'aide de trois décodeurs différents : le décodeur auditif $P(O_L \mid S)$ pour les théories

auditives, le décodeur moteur $P(O_S \mid S)$ pour les théories motrices et le décodeur perceptuo-moteur $P(O_S \mid S\,[C=1])$ pour les théories perceptuo-motrices.

Ce décodage nécessite quelques ajustements. En effet, nous souhaitons que, lors du décodage, certaines portions de l'espace, ayant une très faible probabilité, ne soient pas décodées comme un objet o^+ ou o^- mais soit perçues comme des zones équiprobables entre les deux objets. C'est pourquoi, afin de ne conserver que les portions de l'espace les plus représentatives de chaque catégorie, nous définissons un seuil de probabilité. Au dessus de ce seuil, les deux objets sont reconnus, en dessous de ce seuil, les deux objets sont équiprobables. Le seuil choisi pour cette étude vaut $se = \frac{1}{281}$. Cette valeur est la probabilité de la distribution uniforme de notre espace sensoriel discrétisé S.

En terme d'interprétation, il est possible d'imaginer ce seuil comme la présence d'une « catégorie poubelle », non définie. En dessous de ce seuil, l'agent décode en réalité le son non pas comme o^+ ou o^-, mais comme la catégorie poubelle. Cependant, comme il ne peut choisir qu'entre o^+ ou o^-, il sélectionne l'un ou l'autre, de façon équiprobable.

5.1.3 Outils d'évaluation

5.1.3.1 Analyse de l'apprentissage

Nous devons, au préalable, vérifier la qualité de l'apprentissage de notre modèle, c'est-à-dire vérifier que l'agent apprend convenablement ses distributions. Pour réaliser cette vérification, nous nous focalisons sur le système perceptif de notre modèle, indispensable pour étudier les trois familles de théories de la perception. De notre point de vue, ce système perceptif se note $P(S \mid O)$ et correspond à la probabilité des stimuli sensoriels S pour chaque objet o. Le système perceptif de notre modèle se décompose en deux simulations : d'une part, la branche prédictive auditive $P(S \mid O_L)$, que nous appelons juste « branche auditive » par la suite, qui correspond à la probabilité des stimuli pour chaque objet O_L, d'autre part, la branche prédictive motrice $P(S \mid O_S)$, que nous appelons juste « branche motrice » par la suite, qui correspond à la probabilité des stimuli pour chaque objet O_S.

Comme pour les décodeurs, ces distributions s'interprètent comme des questions dans le modèle COSMO, calculées par inférence à l'aide des distributions du modèle. La différence avec les décodeurs précédents vient du fait que nos branches perceptives ne correspondent pas à une tâche mais à l'analyse directe des distributions que l'agent a apprises. Pour métaphore, c'est comme si nous étions capables d'analyser le système de perception d'un individu en analysant directement le fonctionnement interne des branches prédictives de son cerveau. Cela n'est pas possible en neurosciences mais est permis par la modélisation. Nous allons donc étudier ce que nos agents ont appris.

Les équations correspondants à chaque branche perceptive sont :

$$P(S \mid O_L) = P(S \mid O_L)\,, \tag{5.4}$$

$$P(S \mid O_S) \propto \sum_M P(M \mid O_S)\, P(S \mid M)\,. \tag{5.5}$$

Ainsi, l'étude de la branche auditive nous permet d'étudier le répertoire auditif tandis que celle de la

branche motrice consiste à analyser conjointement le modèle interne et le répertoire moteur. À travers ces deux branches, nous pouvons donc observer l'évolution de nos trois distributions apprises.

Observer la qualité de l'apprentissage nécessite de connaître les valeurs des distributions du modèle (ou de leurs paramètres) au cours de l'apprentissage. Comme il est coûteux de conserver les valeurs de ces distributions pour chaque itération de l'apprentissage, nous n'avons conservé que les valeurs de certaines itérations. L'enregistrement des valeurs conservées s'est fait de manière logarithmique : un grand nombre d'enregistrements est réalisé lors des premières itérations puis la fréquence d'enregistrement diminue au fur et à mesure de l'apprentissage. En effet, au cours de simulations pilotes, nous avons pu observer que les données évoluent beaucoup au début de l'apprentissage et peu à la fin. Cela nécessite donc plus d'enregistrements en début qu'en fin d'apprentissage.

Nous choisissons d'étudier la qualité de l'apprentissage à travers la mesure d'entropie H, qui pour une distribution $P(X)$ donnée, se calcule par :

$$H(P(X)) = -\sum_{X} P(X) \times \log(P(X)) \,. \tag{5.6}$$

Pour chacune de nos douze simulations et pour chaque itération enregistrée, nous calculons d'abord l'entropie de $P(S \mid [O_L = o_-])$ et de $P(S \mid [O_L = o_+])$, puis nous calculons la moyenne de ces deux entropies. Nous obtenons donc douze mesures d'entropie de $P(S \mid O_L)$, une pour chaque simulation. Ensuite, nous calculons la moyenne et l'écart-type globaux toutes simulations confondues. Cela nous donne l'évolution moyenne de l'entropie pour $P(S \mid O_L)$ au cours de l'apprentissage. Nous effectuons la même procédure avec $P(S \mid O_S)$. Nous comparons ensuite ces deux mesures d'entropie avec celle de la distribution des signaux fournis par le maître, que nous calculons à partir de la distribution $P(S \mid O_S^{Maitre})$ (cf Eq. 5.3). Comme pour les précédentes, cette entropie correspond à la moyenne de l'entropie des deux objets. Comme elle est identique pour chaque simulation, aucune moyenne inter-simulation n'est nécessaire.

5.1.3.2 Comparaison des trois familles de théories

Nous souhaitons analyser les performances de notre modèle selon les différentes familles de théories de la perception, en testant les trois décodeurs précédemment définis. Pour ce faire, nous nous servons d'une tâche de catégorisation, nécessitant un décodeur $P(O \mid S)$, qui vaut respectivement $P(O_L \mid S)$, $P(O_S \mid S)$ ou $P(O_S \mid S\,[C = 1])$ selon que nous étudions les théories auditives, motrices ou perceptuo-motrices.

La tâche est la suivante : le maître sélectionne un des deux objets o et tire un geste moteur m correspondant à cet objet grâce à sa distribution $P(M^{Maitre} \mid [O_S^{Maitre} = o])$. Cette production est envoyée dans l'environnement et est perçue par l'agent comme un signal s, ce qui correspond à un tirage sur la distribution $P(S^{Env} \mid [M^{Env} = m])$ (rappelons que $S = S_{Env}$ et $M_{Maitre} = M_{Env}$, voir Fig. 5.2). Ensuite, nous utilisons le décodeur $P([O = o] \mid [S = s])$ afin de reconnaître l'objet o sachant le stimulus s.

Un de nos objectifs étant de comprendre pourquoi le système moteur est plus activé dans des conditions bruitées, nous réalisons la tâche de catégorisation dans différents niveaux de bruit. Cela

revient, dans la tâche de catégorisation, à augmenter l'écart-type des distributions gaussiennes de la distribution $P(S^{Env} \mid M^{Env})$. Durant l'apprentissage, ces distributions ont un écart-type de 1 pour simuler le bruit ambiant (niveau 0). Durant la tâche de catégorisation, nous faisons varier cet écart-type entre 1 et 11 pour simuler un niveau de bruit ambiant et dix niveaux de bruits supplémentaires.

Pour un niveau de bruit donné et pour une famille de théories donnée, afin de simuler un grand nombre de fois cette tâche, nous pouvons calculer plus simplement la matrice de confusion entre les objets du maître et ceux de l'agent, ce qui équivaut à la distribution $P(O \mid O_S^{Maitre})$:

$$P(O \mid O_S^{Maitre}) = \sum_{\substack{S=S^{Env} \\ M^{Env}=M^{Maitre}}} P(O \mid S)\, P(S^{Env} \mid M^{Env})\, P(M^{Maitre} \mid O_S^{Maitre})\,. \qquad (5.7)$$

Comme nous avons définis deux objets O et O_S^{Maitre}, la distribution $P(O \mid O_S^{Maitre})$ équivaut à une matrice 2×2. Dans cette matrice, les deux valeurs diagonales correspondent aux probabilités de reconnaissance de chaque objet. La probabilité de reconnaissance globale est donc la moyenne de ces deux valeurs. Nous réalisons les calculs de $P(O \mid O_S^{Maitre})$ et de la probabilité de reconnaissance pour chacune des douze simulations. Ensuite, comme pour l'entropie, nous calculons la probabilité moyenne de reconnaissance inter-simulation et l'écart-type entre les simulations. Nous obtenons ainsi un ensemble de probabilités de reconnaissance moyenne, pour chaque famille de théories, et pour chaque niveau de bruit.

Pour synthèse, la tâche de catégorisation et la matrice de confusion correspondante sont schématisées par la Fig. 5.4.

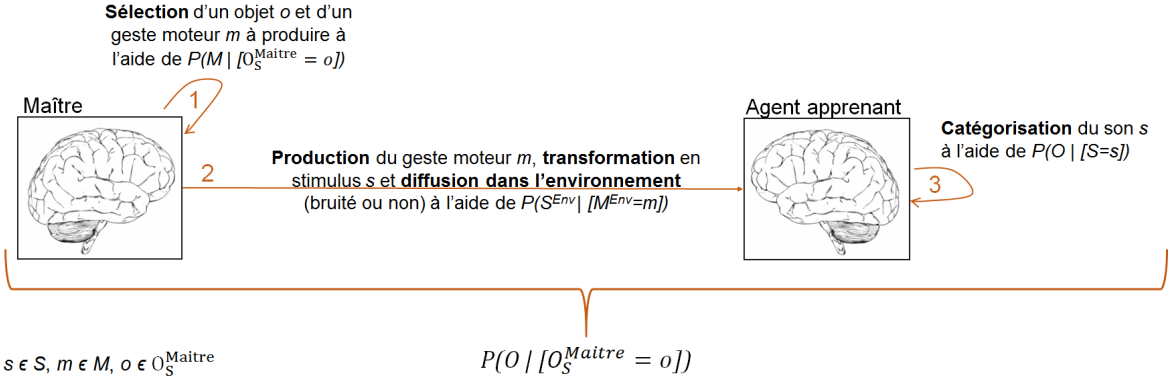


FIGURE 5.4 – Illustration de la tâche de catégorisation

5.1.4 Résultats

5.1.4.1 Évolution de l'apprentissage

Commençons par observer l'évolution de l'apprentissage des simulations de notre modèle en étudiant l'entropie des branches auditive et motrice au cours de l'apprentissage (voir Fig. 5.5). Du fait que les trois répertoires sont appris en même temps, nous avons superposé dans une même figure l'évolution de l'entropie de la branche auditive et de la branche motrice. L'analyse de cette évolution permet

de surligner trois différences entre les branches auditives et motrice : la vitesse d'apprentissage, la convergence et la variabilité.

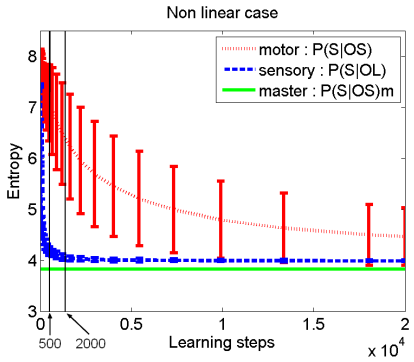


FIGURE 5.5 – Évolution des branches auditive et motrice au cours de l'apprentissage

Nous observons d'abord que les deux branches n'ont pas la même vitesse d'apprentissage. La branche sensorielle converge très rapidement. En effet, il faut moins de 1 000 itérations pour qu'elle atteigne son point de convergence. Au contraire, la branche motrice est beaucoup plus lente. Après 20 000 itérations, elle ne semble toujours pas avoir convergé puisque son entropie continue de diminuer. On peut alors supposer qu'il est beaucoup plus simple pour l'agent d'apprendre sa branche sensorielle que sa branche motrice. Ce comportement semble logique puisque le répertoire sensoriel est appris à partir d'un apprentissage supervisé tandis que le modèle interne et le répertoire moteur sont tous deux appris par accommodation. Il faut donc un certain temps avant que l'agent apprenne, d'un côté, à associer des représentations sensorielles perçues à des représentations motrices adéquates et, d'un autre côté, qu'il apprenne à associer ses représentations motrices à la catégorie correspondante.

Concernant leur point de convergence respectif, nous remarquons que l'entropie de la branche auditive converge vers celle du maître mais qu'elle conserve une erreur résiduelle. Le premier point suggère que la branche auditive réussit à apprendre correctement les données du maître. Concernant l'erreur résiduelle, elle s'explique par le fait que le répertoire sensoriel approxime les données d'une catégorie du maître comme une gaussienne alors que la distribution sensorielle de l'environnement $P(S^{Env} \mid O_S^{Maitre})$ qu'il apprend n'est pas gaussienne. Du côté de la branche motrice, son entropie est bien plus élevée que celle du maître et que celle de la branche auditive, ce qui suppose une approximation moins bonne que celle de la branche auditive. Le fait que l'entropie de la branche motrice soit plus élevée, mais qu'elle continue sa décroissance tout au long de l'apprentissage, laisse par ailleurs supposer qu'elle pourrait rejoindre l'entropie du maître si on prolongeait l'apprentissage. C'est en effet le comportement attendu selon le théorème d'indistinguabilité.

La troisième différence concerne la variabilité. L'entropie de la branche auditive est très stable d'une simulation à l'autre ce qui est cohérent avec le fait qu'elle approxime toujours très rapidement l'entropie du maître. En revanche, l'entropie de la branche motrice est très variable. Il apparaît ainsi que l'apprentissage moteur varie selon les simulations : certains agents arrivent à avoir des branches motrices proches de la distribution sensorielle de l'environnement très rapidement tandis que d'autres sont au contraire beaucoup moins précis et n'approximent que globalement la distribution du maître.

Nous déduisons de ces trois différences que l'apprentissage auditif, de par sa rapidité et sa préci-

sion, est une très bonne approximation des données du maître, et se focalise ainsi efficacement sur les régions sensorielles adéquates dans l'espace d'apprentissage. L'apprentissage de la voie motrice, au contraire, combinant les termes d'apprentissage sensorimoteur et moteur, fournit dans la majorité des cas une approximation plus lente et moins précise des données du maître, explorant des régions plus larges de l'espace sensoriel.

5.1.4.2 Comparaison des tâches de catégorisation

Nous comparons maintenant les trois décodeurs $P(O_L \mid S)$, $P(O_S \mid S)$ et $P(O_S \mid S\ [C = 1])$ à l'aide de la tâche de catégorisation définie en section 5.1.3.2. Les scores de reconnaissance des catégories pour les trois décodeurs et pour différents niveaux de bruit sont présentés Fig. 5.6. Les trois cases de cette figure correspondent à l'observation des résultats lorsque l'on arrête l'apprentissage à trois moment différents : 500 itérations (c'est-à-dire pour un volume d'apprentissage encore faible pour chacun des décodeurs), 2 000 itérations (à un moment où l'apprentissage sensoriel a convergé, mais pas l'apprentissage moteur) et 20 000 itérations (pour lequel l'apprentissage moteur a, à peu près, convergé). Nous étudions d'abord globalement les trois décodeurs avant de nous focaliser sur les spécificités relatives aux trois moments sélectionnés.

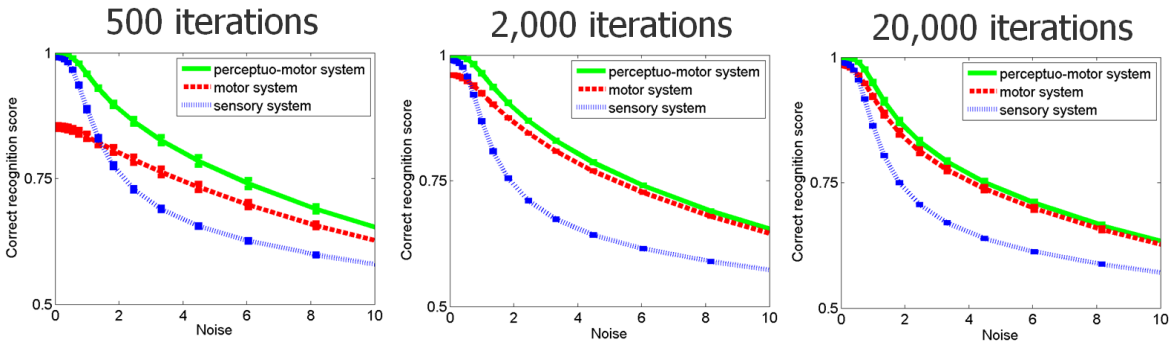


FIGURE 5.6 – Étude de la performance de catégorisation à différents niveaux d'apprentissage

Dans un premier temps, nous observons que, de manière globale, dans toutes les conditions, le décodeur perceptuo-moteur donne de meilleurs performances que les deux autres décodeurs. Il semble donc plus efficace de fusionner les décodeurs auditif et moteur que de les utiliser séparément. Nous remarquons que le décodeur perceptuo-moteur est très vite performant. En effet, dès 500 itérations, il catégorise parfaitement les deux objets (score de reconnaissance à 1). Cependant, la qualité de catégorisation diminue avec le niveau de bruit. Ce résultat est concordant avec les résultats obtenus dans la littérature (voir, par exemple, les études présentées section 3.1.1.2). Néanmoins, même avec un niveau de bruit très élevé (10 fois l'écart-type du niveau de bruit normal), le score de reconnaissance est au dessus du niveau du hasard (qui est à 0,5 puisqu'il y a deux objets).

Comparons maintenant les décodeurs auditif et moteur. Nous remarquons d'abord que le décodeur auditif est meilleur que le décodeur moteur dans des conditions non bruitées. Dans ces conditions, il est, comme le décodeur perceptuo-moteur, très rapidement performant, puisque son score de reconnaissance est quasiment parfait dès 500 itérations. En revanche, dès qu'un peu de bruit est ajouté,

le score diminue drastiquement et est inférieur à 75% pour un bruit à 2. Par ailleurs, le score de reconnaissance se stabilise par la suite quand le niveau de bruit augmente puisqu'il ne diminue que d'environ 10 à 15% entre un bruit à 2 et un bruit à 10. Du côté du décodeur moteur, les scores dans des conditions non bruitées sont moins bons que ceux du décodage sensoriel en début d'apprentissage (environ 80% de reconnaissance) mais s'en rapprochent avec l'apprentissage : le score est quasiment parfait à 20 000 itérations. Fait intéressant, quand du bruit est ajouté, les performances du décodeur moteur diminuent, bien sûr, mais deviennent meilleures que celles du décodeur auditif.

Focalisons-nous maintenant quelques instants sur l'apprentissage. Entre 500 et 20 000 itérations, nous observons finalement que le seul décodeur qui s'améliore significativement est le décodeur moteur. Les deux autres ne semblent pas beaucoup évoluer (moins de 5% d'amélioration) mais ce résultat est contrasté par le fait que dès 500 itérations, les décodeurs auditif et perceptuo-moteur semblent avoir déjà de très bons scores de catégorisation, surtout dans des conditions non bruitées. Ce constat est concordant avec le résultat que nous avons observé précédemment : la branche auditive converge très rapidement vers une distribution sensorielle similaire à celle de l'environnement. De ce fait, le décodeur auditif est, lui aussi, très rapidement très performant. En revanche, la branche motrice est apprise beaucoup plus lentement, ce qui explique que les performances augmentent avec l'apprentissage.

En fin d'apprentissage, nous observons d'abord que les scores des trois décodeurs sont tous les trois parfaits ou presque, en l'absence de bruit. Ce résultat nous rapproche du théorème d'indistingabilité dans lequel les représentations sensorielles et motrices sont identiques. Nous supposons que, si nous augmentions davantage l'apprentissage, les branches sensorielles et motrices deviendraient totalement indistinguables dans des conditions non bruitées et qu'elles fourniraient la même information. Par ailleurs, dans des conditions bruitées, nous observons en fin d'apprentissage que les performances du décodeur moteur sont, certes, toujours inférieures à celles du décodeur perceptuo-moteur mais en sont très proches. Ainsi, à fort bruit, le décodeur sensoriel ne peut plus fonctionner utilement, et seul le décodeur moteur est capable de pouvoir extraire des informations adéquates pour le décodage.

En résumé, la perception selon les théories auditives, que l'on modélise dans COSMO avec le décodeur auditif $P(O_L \mid S)$, obtient des scores quasi-parfaits dans des conditions non bruitées mais ses performances diminuent très rapidement et fortement (de plus de 25 % pour un bruit à 2) dès que le niveau de bruit augmente. En parallèle, la perception selon les théories motrices, que l'on modélise dans COSMO avec le décodeur moteur $P(O_S \mid S)$, bien qu'elle soit de plus en plus performante au cours de l'apprentissage, est moins efficace que le décodeur auditif dans des conditions non bruitées. La tendance s'inverse dès que du bruit est ajouté puisque nous observons que ses scores de reconnaissance dépassent ceux du décodeur auditif. Dans tous les cas, les scores sont les plus élevés quand les deux décodeurs sont fusionnés à travers le décodeur perceptuo-moteur.

5.1.4.3 Interprétation des résultats

Avant de discuter de la relation entre ces résultats et ceux de la littérature, il est important de mieux les comprendre. Notre question principale est : pourquoi le décodeur auditif est meilleur dans des conditions non bruitées et pourquoi le décodeur moteur est meilleur dans des conditions bruitées ?

Pour cela, nous sommes retournés étudier le système perceptif, notamment les deux branches auditive $P(S \mid O_L)$ et motrice $P(S \mid O_S)$, que nous avons précédemment définies en section 5.1.3.1 (voir Eq. 5.4 et Eq. 5.5). Pour chacune de ces branches, nous avons étudié leurs distributions respectives à 2 000 itérations, quand elles sont parfaitement distinguables l'une de l'autre. Nous avons ensuite analysé leur décodage quand elles reçoivent un stimulus bruité et non bruité. Le résultat est schématisé Fig. 5.7.

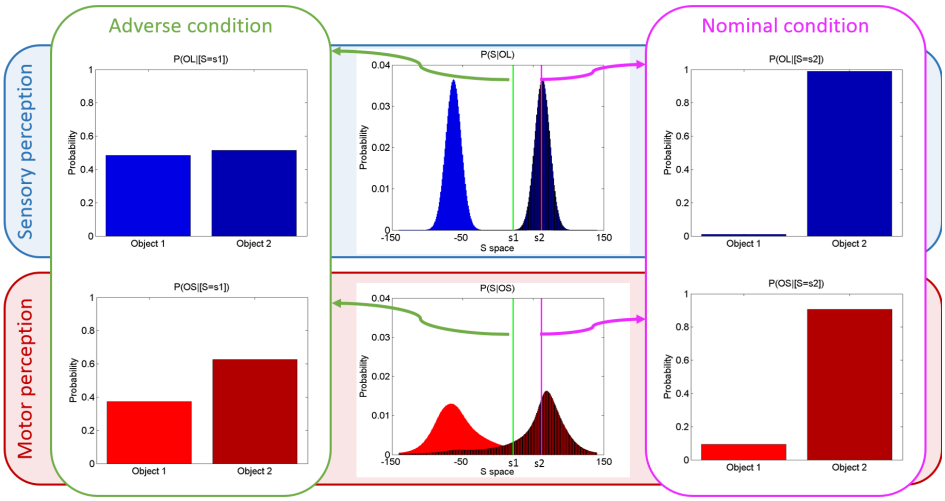


FIGURE 5.7 – Schéma illustrant le comportement des branches auditive (**Haut, en bleu**) et motrice (**Bas, en rouge**). Pour chaque branche, observation des résultats de perception pour un stimulus bruité (**Gauche, trait vert**) et non bruité (**Droite, trait rose**)

Commençons par la branche auditive $P(S \mid O_L)$. Du fait que cette branche est un reflet du répertoire auditif, nous observons que les distributions sont gaussiennes. De plus, comme nous l'avons déjà suggéré précédemment en étudiant l'entropie, elles sont une très bonne approximation de la distribution de l'environnement $P(S^{Env} \mid O_S^{Maitre})$. Ainsi, même si les distributions du maître ne sont pas gaussiennes, les prototypes sont centrés au même endroit et leurs variances respectives sont également très proches. Comme le confirme l'entropie, nous en déduisons que la branche auditive est une distribution approximant de manière très précise les données reçues de son environnement. De ce fait, nous observons que lorsqu'elle reçoit un stimulus non bruité, c'est-à-dire correspondant à un stimulus de l'environnement proche du prototype appris, le décodage est parfait (Fig. 5.7 en haut à droite). En revanche, quand elle reçoit un stimulus très bruité, c'est-à-dire un stimulus ne correspondant pas à ceux qu'elle a appris durant son apprentissage, le décodage est proche du hasard (Fig. 5.7 en haut à gauche). Cela vient du fait que le stimulus bruité sort de la couverture de la distribution gaussienne, et rentre dans une région pour laquelle la probabilité de chaque classe passe en dessous du seuil que nous avons défini pour le processus de catégorisation (voir section 5.1.2.5).

Passons à la branche motrice $P(S \mid O_S)$. Contrairement à la distribution $P(S \mid O_L)$, il s'agit d'une distribution non gaussienne, puisqu'elle est une somme de produits des distributions $P(M \mid O_S)$ et $P(S \mid M)$. On pourrait donc s'attendre à ce qu'elle approxime mieux les données du maître. Cependant, comme nous l'avons déjà suggéré en étudiant l'entropie, elle est une approximation plus grossière des stimuli du maître. En effet, même si elle est également centrée sur les prototypes du

maître, sa variance est plus grande, ce qui donne une distribution plus aplatie que son homologue auditif. Cela vient du fait que son apprentissage est plus lent et qu'elle converge moins efficacement vers la distribution des stimuli de l'apprentissage. Du fait de cette grande variance, nous observons que, lorsqu'elle reçoit un stimulus non bruité, le décodage est très bon mais pas totalement parfait (Fig. 5.7 en bas à droite). En revanche, quand elle reçoit un stimulus très bruité, le décodage est globalement performant ou, du moins, l'objet correspondant au stimulus bruité est reconnu dans la majorité des cas (Fig. 5.7 en bas à gauche). Nous en déduisons que la grande variance des distributions de la branche motrice permet de catégoriser les stimuli différant de ceux appris durant l'apprentissage.

Cette grande variance est liée au processus d'exploration sensorimotrice qui accompagne le processus d'apprentissage sensorimoteur et moteur, qui, rappelons-le, n'est pas supervisé par le maître, puisqu'il ne fournit aucune information motrice. De plus, dans cette étude, du fait que l'apprentissage sensorimoteur et moteur sont appris en même temps, l'exploration est davantage accentuée puisqu'au début de l'apprentissage, l'agent, n'ayant ni de connaissances sur son répertoire moteur, ni de connaissance sur son modèle interne, commence par associer de mauvaises représentations motrices aux catégories fournies par le maître. Bien entendu, cet effet diminue lorsque l'agent commence à apprendre son modèle interne. Malgré ces contraintes, le processus d'exploration a tout de même un avantage puisqu'il permet à l'agent de tester des régions sensorielles et motrices non prototypiques. Ceci pénalise la réponse à des stimuli typiques, mais facilite l'identification de stimuli atypiques.

Ainsi, nous pouvons conclure que ce sont les différences de variances qui sont à l'origine des différences de performance entre les deux branches perceptives. La branche auditive est plus performante dans des conditions non bruitées car elle possède des distributions de petite variance, piquées sur les signaux non bruités, tandis que la branche motrice est plus performante dans le bruit car elle possède des distributions de plus grande variance, plus étalées dans l'espace, capables de reconnaître des signaux éloignés des signaux prototypiques. Pour conclure, la branche auditive semble agir comme une « bande étroite » alors que la branche motrice semble agir comme une « bande large ». C'est ce que nous nommons, pour synthétiser, la propriété « bande étroite/bande large » (dans nos publications récentes en anglais, « auditory narrow motor wide »).

5.1.5 Discussion

5.1.5.1 Résumé et interprétation des résultats

Nous avons mis en évidence la propriété « bande étroite/bande large » de la perception de la parole dans laquelle les différences de variance des distributions des branches auditive et motrice seraient à l'origine des différences de catégorisation. La branche auditive de petite variance permet de catégoriser avec une grande précision les signaux prototypiques tandis que la branche motrice de plus grande variance permet de catégoriser des signaux bruités. Ceci met donc en avant la complémentarité de la branche auditive et de la branche motrice et explique également pourquoi une combinaison des deux branches, via le décodeur perceptuo-moteur, donne de meilleures performances de catégorisation.

Cette complémentarité entre les deux branches nous semble d'une grande importance et d'un certain intérêt théorique. Nous l'interprétons comme provenant de la différence d'apprentissage. En effet,

avant d'analyser les tâches de perception, nous avons étudié la vitesse d'apprentissage des deux distributions et avons observé que la branche auditive est rapide et converge rapidement vers l'entropie du maître tandis que la branche motrice est plus lente et n'a toujours pas convergé en fin d'apprentissage. L'apprentissage rapide de la branche auditive provoque une brusque diminution de la variance et produit des distributions de faibles variances, très piquées vers les signaux du maître. L'apprentissage de la branche motrice permet aussi de se centrer sur les signaux du maître mais, l'apprentissage étant plus long et plus complexe, les variances des distributions diminuent plus lentement.

Par ailleurs, ces résultats sont concordants avec la littérature puisque plusieurs études en neurosciences montrent une activation des aires motrices, surtout dans des tâches de perception en conditions bruitées. À titre d'illustration, Binder et ses collègues écrivent : « Lateral opercular areas (ventral BA 44 and 45) showed increases in activation as SNR decreased [...] » (which would show) « [...] enhanced activation of internal representations of the speech sounds as a template against which the sensory information could be matched » (Binder et al., 2004, (p. 298)). De la même manière, selon Zekveld : « only Broca's area (BA44) showed activation to unintelligible speech presented at low SNRs. » (Zekveld et al., 2006, (p. 1)). Nos simulations permettent, pour la première fois, à notre connaissance, de donner une base théorique et interprétative à ces observations récurrentes.

5.1.5.2 Limites et perspectives des résultats

Bien que les résultats soient concordants avec ceux de la littérature, cette étude possède plusieurs limites. La première concerne les paramètres arbitraires utilisés à savoir : le nombre de tirages et les paramètres de la sigmoïde. Le nombre de tirages a été choisi de façon arbitraire mais dure suffisamment pour que la branche auditive converge en fin d'apprentissage et que la branche motrice soit suffisamment apprise comme le montrent les tâches de perception. Les paramètres de la sigmoïde sont également arbitraires mais n'influent en fait pas sur le résultat. En effet, nous avons testé plusieurs paramètrages concernant cette sigmoïde et nous avons obtenu à chaque fois les mêmes résultats en termes d'apprentissage et de performances.

La seconde, plus importante, concerne la simplicité générale du modèle. En effet, cette étude est une version très simple de notre modèle. Il s'agit d'un avantage, puisque nous avons pu dans un cadre simplifié, étudier les comportements des branches auditives et motrices en perception. Il s'agit également d'une limite, puisque ce cadre simplifié manque de réalisme et qu'il convient de vérifier que la propriété définie soit toujours valide dans des simulations plus complexes.

Pour autant, nous affirmons que la propriété « bande étroite/bande large » présentée ici est tout de même générique. D'abord il est important de noter qu'elle n'apparaît pas dépendante de la linéarite versus non linéarité de la transformation $P(S^{Env} | M^{Env})$. D'ailleurs, nous pouvons déjà annoncer que ce résultat a été retrouvé avec succès dans un modèle COSMO plus complexe modélisant les syllabes (voir Laurent et al., 2017). De plus, les résultats observés dans cette étude dépendent essentiellement non pas de l'implémentation choisie mais de la structure de COSMO et des différences entre les processus d'apprentissage. En effet, l'apprentissage moteur, basé sur un processus d'inférence complexe, apparaît, dans tous les cas, computationnellement plus lourd que l'apprentissage auditif. Bien entendu, l'implémentation choisie possède un bon nombre de choix non-génériques tels

que l'utilisation d'une transformation articulatoire vers acoustique sigmoïdale ou encore l'utilisation d'un maître unique. Néanmoins, nous pensons que rendre notre modèle plus réaliste, à travers l'utilisation d'une transformation articulatoire vers acoustique plus complexe ou de multiples maîtres durant l'apprentissage, aurait pour conséquence d'augmenter la complexité générale du modèle et donc, en conséquence, la complexité du processus d'inférence utilisé durant l'apprentissage moteur. De ce fait, nous supposons que l'apprentissage moteur, dans des conditions réalistes, serait encore plus lent et diffus que celui présenté dans cette étude. Selon cette hypothèse, si nous obtenons la propriété « bande étroite/bande large » dans des conditions ultra-simplifiées, elle devrait également apparaître dans des situations plus complexes (voir Laurent et al., 2017, pour une discussion plus détaillée).

Cette propriété est notre interprétation des performances obtenues par les branches auditive et motrice. Il s'agit également d'une prédiction sur le fonctionnement du cerveau. Sans considérer que le cerveau opère sur des distributions aussi simples que celles présentées ici, nous supposons que les aires auditives sont plus efficaces dans des conditions non bruitées parce qu'elles sont précises et ont été conditionnées à répondre pour des signaux prototypiques. Au contraire, nous proposons que les aires motrices pourraient être plus efficaces dans des conditions bruitées parce qu'elles sont moins précises et ont une capacité de généralisation. Cette hypothèse est actuellement testée, dans notre équipe, dans une expérience de neuroimagerie fonctionnelle (fMRI) portant sur les phénomènes de répétition suppression, dans lesquels la répétition d'une stimulation comportant des propriétés stables d'un item au suivant produit une diminution progressive de la réponse neuronale et donc de la réponse indiquée par le débit sanguin (signal BOLD) en neuroimagerie (Dole et al., en préparation). L'expérience conduite consiste à présenter des stimuli vocaliques s'éloignant peu à peu d'une cible, d'un item à l'item suivant. Le raisonnement est que si le système auditif fonctionne à bande étroite, c'est-à-dire avec $P(S \mid O_L)$ de faible variance, l'éloignement progressif se traduirait très vite par la sortie du champ récepteur de l'ensemble de neurones correspondant, et donc que les phénomènes de diminution de la réponse disparaîtraient rapidement. Au contraire, dans le système moteur, même en éloignant le stimulus de la cible, on resterait dans le même champ récepteur, $P(S \mid O_S)$ de forte variance, et donc il y aurait effectivement une diminution de la réponse, même pour des stimuli assez éloignés de la cible. Les données de neuroimagerie semblent confirmer cette prédiction (voir aussi les prédictions présentées dans Laurent et al., 2017).

5.2 Analyse de l'apprentissage sensorimoteur

Publication :

— Barnaud, M.-L., Schwartz, J.-L., Diard, J., et Bessière, P. (2016b). Sensorimotor learning in a Bayesian computational model of speech communication. In The 6th Joint IEEE International Conference on Developmental and Learning and on Epigenetic Robotics (ICDL-Epirob 2016), Cergy-Pontoise, France

---

Dans l'étude précédente, nous nous sommes intéressés aux conséquences de l'apprentissage sensoriel et moteur à travers la comparaison des théories de la perception. Bien que nous considérons que

les résultats observés, notamment la propriété « bande étroite/bande large », soient génériques, nous souhaitons complexifier le modèle pour nos prochaines études. Néanmoins, cette complexité risque en conséquence de contraindre nos processus d'apprentissage, notamment l'apprentissage sensorimoteur. En effet, comme nous l'avons vu dans le chapitre précédent, l'implémentation de cet apprentissage a suscité beaucoup de recherches et d'intérêt en modélisation et nécessite la mise en place de plusieurs stratégies pour être mené efficacement.

C'est pourquoi, cette seconde étude a pour but de mieux comprendre ce qui favorise l'apprentissage sensorimoteur. Nous laissons donc de côté, pour le moment, les unités distinctives et nous nous focalisons uniquement sur le développement des représentations sensorielles et motrices. Pour cela, nous comparons huit algorithmes, basés sur trois composants : « le principe d'accommodation », qui est la base de notre algorithme par accommodation qui consiste en un mélange entre un apprentissage supervisé par un maître et le babillage d'objectifs (goal babbling), « le babillage idiosyncratique », qui favorise les gestes moteurs précédemment sélectionnés, et « l'extrapolation locale », qui explore le voisinage d'un geste moteur sélectionné pour faire de la généralisation. Nous montrons que l'utilisation de ces trois composants permet de réduire l'exploration de l'espace sensorimoteur tout en gardant une bonne qualité d'apprentissage.

5.2.1 Hypothèse de l'étude

Dans notre modèle COSMO, l'apprentissage sensorimoteur concerne l'apprentissage du modèle interne $P(S \mid M)$, correspondant à la probabilité des représentations sensorielles sachant les représentations motrices. L'apprentissage de cette distribution est complexe pour essentiellement deux raisons (citées par exemple dans Moulin-Frier et Oudeyer, 2013) : 1) les représentations sensorielles S et motrices M sont généralement de grandes dimensions, ce qui limite leur exploration, 2) la relation entre les représentations sensorielles et motrices est non linéaire, ce qui rend difficile l'exploration. Nous proposons, dans cette étude, trois composants dont nous faisons l'hypothèse qu'ils facilitent l'apprentissage sensorimoteur : le principe d'accommodation, le babillage idiosyncratique et l'extrapolation locale.

Le principe d'accommodation Comme nous l'avons vu précédemment, le principe d'accommodation consiste à inférer une représentation motrice m à partir d'un stimulus s envoyé par le maître puis à la produire pour apprendre la relation entre la représentation motrice m et le stimulus résultant s'.

Ce principe mélange ainsi deux éléments : d'une part, un apprentissage social, puisque le choix de la représentation motrice m est basé sur un processus d'imitation dans lequel l'agent apprenant tente de sélectionner une représentation motrice m lui permettant de reproduire le stimulus s qu'il a reçu du maître et, d'autre part, une phase d'exploration, puisque l'agent apprenant sélectionne des représentations motrices, les produit et apprend la relation entre les représentations motrices et les représentations sensorielles perçues et ce, même si le stimulus s' ne correspond pas au stimulus s du maître.

Il est ainsi supposé qu'au départ, comme l'agent n'a pas de connaissances préalables sur son

modèle interne, les représentations motrices sélectionnées ne correspondent pas ou très rarement au stimulus s fourni par le maître et que l'exploration est relativement aléatoire. Par la suite, en explorant peu à peu son espace, l'agent apprenant arrive petit à petit à imiter son maître. Nous considérons que, même pendant cet apprentissage, il est important que le modèle apprenne à favoriser le plus rapidement possible les portions des espaces sensoriel et moteur correspondant aux stimuli de son environnement. Cela lui évite de devoir apprendre de façon exhaustive à relier ses représentations motrices aux représentations sensorielles.

Le babillage idiosyncratique Ce mécanisme consiste à favoriser, parmi les configurations motrices m possibles correspondant à un stimulus s, celles qui ont été préalablement sélectionnées.

Cela s'apparente à une stratégie d'exploration auto-centrée basée sur le renforcement. Tout comme l'apprentissage actif favorisait les représentations motrices résultant en un progrès maximum, nous considérons que l'agent apprenant favorise les représentations motrices précédemment explorées. Nous supposons que cela permet, d'une part, de limiter l'exploration de l'espace moteur lorsqu'une représentation motrice précédemment explorée correspond à un stimulus s et, d'autre part, d'accélérer l'apprentissage, puisque les représentations motrices les plus adaptées pour un stimulus s sont plus souvent sélectionnées et de ce fait plus souvent mises à jour.

L'extrapolation locale Bien que la transformation sensorimotrice soit non-linéaire, nous supposons que le voisinage d'un geste moteur m est constitué de gestes moteurs m_v produisant des stimuli sensoriels s' similaires. Ainsi, lorsque l'agent apprenant infère une représentation motrice m et la produit afin d'apprendre la relation entre la représentation m et le stimulus s' résultant , il sélectionne également des représentations motrices proches m_v et apprend la relation entre ces représentations motrices proches et s'.

Derrière cette hypothèse, nous souhaitons prendre en compte la continuité de la relation articulatori-acoustique, c'est-à-dire le fait que des configurations motrices voisines fournissent des paramètres sensoriels proches. Ainsi, nous ne considérons plus l'apprentissage sensorimoteur comme la mise à jour d'une unique représentation motrice mais comme celle de l'ensemble d'une portion de l'espace moteur, ce qui permet de généraliser l'exploration d'une représentation motrice à l'ensemble des représentations motrices similaires. Par ailleurs, ce mécanisme réduit considérablement l'exploration puisque la sélection d'une représentation motrice permet d'apprendre également la relation sensorimotrice de son voisinage.

5.2.2 Description des variables du modèle

Nous nous plaçons dans un cadre plus réaliste que l'étude précédente dans lequel l'environnement considéré est composé de trois voyelles du français : [a i u]. Ce sont également trois voyelles que l'on retrouve régulièrement dans les langues du monde. Elles correspondent chacune, dans COSMO, à une valeur d'un objet O_S et O_L.

Afin de les produire, nous nous servons des paramètres articulatoires du modèle du conduit vo-

cal VLAM (« Variable Linear Articulatory Model », Maeda, 1990). Dans ce modèle, la valeur d'un paramètre articulatoire correspond à la position relative d'un articulateur dans le conduit vocal. Pour chaque configuration articulatoire donnée, le modèle est capable de fournir le stimulus acoustique correspondant, dans un espace formantique. Le modèle contient initialement 7 paramètres articulatoires : le corps de la langue (TB, tongue body), le dos de la langue (TD, tongue dorsum), le bout de la langue ($Apex$), la hauteur des lèvres (LH, lip height), l'arrondissement des lèvres (LP, lip protusion), le larynx ($Larynx$) et la mâchoire (Jaw) (voir Fig. 5.8, voir aussi Laurent, 2014, section 1.2, pour en savoir plus sur ce modèle). Dans COSMO, pour les voyelles considérées, nous utilisons uniquement trois paramètres articulatoires de VLAM, correspondant à nos représentations motrices M : TB, TD et LH. Ces paramètres évoluent dans l'intervalle $[-2; +4]$ pour TB, $[-4{,}5; +3{,}5]$ pour TD et $[-1; +5]$ pour LH, 0 étant la position neutre, de repos. Ces trois paramètres permettent de configurer les trois dimensions avant/arrière (essentiellement reliée à TB), haut/bas (essentiellement reliée à TD) et arrondi/non arrondi (LH) caractéristiques de l'espace des voyelles orales. Chaque dimension est discrétisée en 25 valeurs équitablement réparties, de manière à avoir un espace à 15 625 valeurs (25^3). Notons qu'à l'origine, les articulateurs sont définis dans l'intervalle $[-5; +5]$, mais nous avons enlevé les valeurs ne permettant pas de produire une voyelle, c'est-à-dire les valeurs conduisant à une fermeture du conduit vocal.

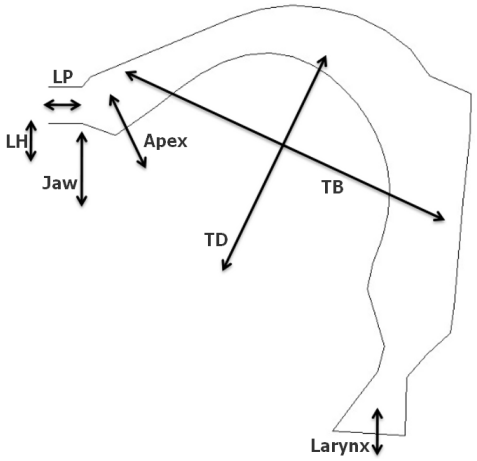


FIGURE 5.8 – Illustration des paramètres articulatoires du modèle VLAM

En ce qui concerne les représentations sensorielles, dans le modèle COSMO, elles correspondent aux valeurs des formants F1 et F2, ce qui est assez courant dans la littérature (voir, par exemple, les modèles vocaliques du chapitre précédent). L'espace sensoriel est, comme précédemment, un espace fini et discret où les valeurs sont exprimées en Barks. Les Barks sont une échelle perceptive de fréquence, de type quasi linéaire en basse fréquence (jusqu'à 700 Hz environ) et quasi logarithmique en haute fréquence (au-delà de 1 500 Hz environ). La formule de conversion des Hertz en Barks est tirée de Schroeder et al. (1979) :

$$z(Barks) = 7 \times Argsh\left(\frac{F(Hertz)}{650}\right) . \tag{5.8}$$

Dans le modèle, F1 est un espace de 59 valeurs, définies dans l'intervalle $[2{,}74; 6{,}95]$, avec un pas de discrétisation linéaire de 0,07, tandis que F2 est un espace de 73 valeurs, définies dans l'intervalle

$[5{,}61; 13{,}55]$ avec un pas de discrétisation linéaire de 0,11. Nous obtenons ainsi un espace sensoriel à 4 307 valeurs. Les intervalles ont été choisis de façon à contenir dans son ensemble le triangle vocalique. Les pas de discrétisation et les valeurs associées ont été obtenus initialement à partir d'un découpage en Hertz de 75 valeurs en F1 et d'un découpage deux fois plus grand de 150 valeurs en F2. Ce découpage initial, bien qu'arbitraire, a été jugé assez précis pour voir apparaître les spécificités d'apprentissage.

5.2.3 Description des distributions de l'agent

L'objectif principal étant d'étudier l'apprentissage sensorimoteur, nous nous concentrons essentiellement sur le modèle interne $P(S \mid M)$ de l'agent apprenant, qui est la distribution mise à jour lors de cet apprentissage. Les autres distributions du modèle, c'est-à-dire le prior des objets moteurs, le répertoire moteur, le classifieur auditif et le système de cohérence, ne sont pas considérées dans cette étude. Par ailleurs, afin de mettre en place le babillage idiosyncratique, une nouvelle distribution a été implémentée, un prior moteur $P(M)$. Ainsi, le modèle COSMO de l'agent apprenant se résume, dans cette étude, à un modèle sensorimoteur $P(S\ M)$ qui se décompose par :

$$P(S\ M) = P(M)\ P(S \mid M)\ . \tag{5.9}$$

Dans cette équation, le prior moteur $P(M)$ correspond à un ensemble de 15 625 valeurs, une pour chaque configuration de l'espace des représentations motrices. Cette distribution est uniforme à l'initialisation. De son côté, le modèle interne $P(S \mid M)$ correspond à un ensemble de gaussiennes. Du fait qu'il y a 15 625 valeurs dans l'espace des représentations motrices, il y a donc 15 625 gaussiennes dans cet ensemble. Leurs moyennes sont initialisées au centre de l'espace des représentations motrices et elles possèdent une grande variance initiale, de la taille de l'espace. Cela permet de simuler au départ une distribution uniforme.

Comme dans l'étude précédente, comme les représentations sensorielles S et motrices M sont discrètes et finies, les distributions gaussiennes correspondantes sont tronquées et discrétisées (voir Eq. 5.1 et Eq. 5.2). La différence est que, cette fois-ci, les gaussiennes sont multi-dimensionnelles. Elles sont donc paramétrées par une moyenne μ et une matrice de covariance Σ. En prenant l'exemple de la distribution $P([S = s] \mid [M = m])$ pour des représentations motrice m et sensorielle s quelconques, l'équation exacte d'une distribution gaussienne selon nos critères est :

$$Gauss_{multi}(s) = \frac{1}{|\Sigma|^{1/2}(2\pi)^{N/2}} e^{-\frac{1}{2}(s-\mu)^T \Sigma^{-1}(s-\mu)}\ , \tag{5.10}$$

$$P([S = s] \mid [M = m]) = \frac{Gauss_{multi}(s)}{\sum_S Gauss_{multi}}\ . \tag{5.11}$$

5.2.4 Description des distributions de l'environnement

L'environnement est représenté par un maître fournissant des stimuli sensoriels pour l'agent apprenant à travers la distribution $P(S^{Env} \mid O_S^{Maitre})$ (voir Eq. 5.3 de l'étude précédente), représentée

Fig. 5.9. Pour rappel, cette équation se calcule à partir du répertoire moteur du maître $P(M^{Maitre} \mid O_S^{Maitre})$, qui permet de produire un objet o donné et de la transformation articulatori-acoustique $P(S^{Env} \mid M^{Env})$ donnant signal sonore S^{Env} correspondant à une réalisation motrice M^{Env} (voir Fig.4.9).

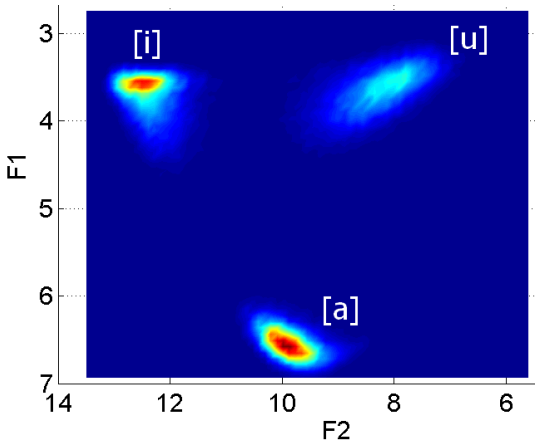


FIGURE 5.9 – Représentation de la distribution $P(S^{Env} \mid O_S^{Maitre})$ dans l'espace des représentations sensorielles, en Barks. L'axe des abscisses correspond au formant F2 inversé. L'axe des ordonnées correspond au formant F1 inversé. Les valeurs s'étendent du bleu foncé (faible probabilité) au rouge (forte probabilité)

Dans cette étude, le répertoire moteur du maître $P(M^{Maitre} \mid O_S^{Maitre})$ correspond à un ensemble de trois gaussiennes, une pour chaque objet, et la transformation des paramètres articulatoires du maître en signaux acoustiques $P(S^{Env} \mid M^{Env})$ est un ensemble de distributions de Dirac. En théorie, cette dernière est calculée à l'aide du modèle VLAM. En pratique, comme l'utilisation du modèle VLAM est coûteuse en temps, nous utilisons un dictionnaire de 542 085 configurations articulatoires, linéairement réparties dans l'espace M, pour lesquelles il a été préalablement calculé avec VLAM les deux premiers formants correspondants. Ainsi, la distribution $P(S^{Env} \mid M^{Env})$ est définie formellement comme un ensemble de 542 085 distributions de Dirac, valant 1 pour la valeur s^{Env} donnée par la ligne correspondante du dictionnaire et 0 pour les autres valeurs.

En termes d'initialisation, la moyenne μ des trois distributions gaussiennes de la distribution $P(M^{Maitre} \mid O_S^{Maitre})$ correspond à un prototype moteur, c'est-à-dire une configuration articulatoire de la voyelle correspondante. Pour les définir, nous nous servons des prototypes auditifs formantiques de chaque voyelle, obtenus par Meunier (2007). Plus précisément, pour chacun d'eux, nous calculons le point sensoriel du dictionnaire de VLAM le plus proche du prototype auditif et sélectionnons le prototype moteur correspondant.

Pour chaque gaussienne, la matrice de covariance vaut :

$$\Sigma = \begin{pmatrix} 0.1 & 0 & 0 \\ 0 & 0.1 & 0 \\ 0 & 0 & 0.1 \end{pmatrix} .$$

5.2.5 Description de l'apprentissage

Dans le chapitre précédent, nous avons présenté un algorithme d'apprentissage sensorimoteur, que nous utilisons dans la majorité de nos études. Dans cette étude, nous testons plusieurs variantes de cet algorithme pour analyser plus en détail l'apprentissage sensorimoteur. Dans tous nos algorithmes, nous considérons que l'apprentissage s'effectue de façon itérative, chaque itération se déroulant en trois phases : une phase de sélection, dans laquelle l'agent choisit ce qu'il souhaite apprendre, une phase de production, dans laquelle l'agent produit la représentation motrice m correspondant à ce qu'il souhaite apprendre, et une phase de mise à jour, dans laquelle l'agent met à jour, entre autres, la relation entre la représentation motrice m choisie et la représentation sensorielle s' résultant de sa production.

Commençons par détailler la phase de sélection. Comme nous l'avons observé dans le chapitre 3, dans les modèles computationnels, la sélection repose globalement sur deux choix : le premier choix concerne l'espace à explorer, le second concerne la stratégie d'exploration.

Concernant l'espace à explorer, comme précédemment rappelé, il y a deux possibilités. Soit l'espace à explorer est l'espace des représentations motrices, ce qui est nommé babillage moteur (« motor babbling »), soit il s'agit de l'espace des représentations sensorielles, ce qui est nommé babillage d'objectifs (« goal babbling »). Dans notre modèle, l'exploration via le babillage moteur consiste à sélectionner une représentation motrice m, puis à effectuer la phase de production et la phase de mise à jour. Dans le babillage d'objectifs, la production ne peut se faire directement puisque l'espace d'exploration est l'espace des représentations sensorielles. De ce fait, l'exploration via le babillage d'objectifs consiste à sélectionner une représentation sensorielle s, à inférer une représentation motrice m correspondant à cette valeur de s, puis à effectuer la phase de production et de mise à jour. Plus précisément, l'inférence consiste à réaliser un tirage via la distribution $P(M \mid S)$, qui se calcule dans le modèle actuel par :

$$P(M \mid S) \propto P(M)\ P(S \mid M)\ . \tag{5.12}$$

Le choix de l'espace d'exploration modifie donc également l'algorithme d'apprentissage.

Les stratégies d'exploration sont beaucoup plus nombreuses que les espaces à explorer. Nous n'avons pas la prétention d'être exhaustif et nous n'en avons testé que deux. La première est une stratégie d'exploration aléatoire dans laquelle les représentations sélectionnées, motrices dans le cas du babillage moteur ou sensorielles dans le cas du babillage d'objectifs, sont tirées aléatoirement dans l'espace. C'est cette stratégie d'exploration que nous utilisons pour comparer les deux espaces d'exploration, sensoriels et moteurs. Nous avons donc un algorithme d'exploration aléatoire avec babillage moteur, RMB (« random motor babbling ») et un algorithme d'exploration aléatoire avec babillage d'objectifs RGB (« random goal babbling »).

La seconde stratégie d'exploration n'est réalisée qu'avec le babillage d'objectifs. Il s'agit d'une stratégie d'exploration guidée socialement à l'aide de la distribution $P(S^{Env} \mid O_S^{Maitre})$ du maître, telle qu'elle a été décrite ci-dessus. Les trois objets O_S^{Maitre} ([a i u]) ayant une même fréquence d'apparition, ils sont tirés successivement, les uns après les autres, durant les itérations de l'apprentissage. Du fait que le maître ne peut fournir que des représentations sensorielles, seul le babillage d'objectifs a été testé avec un apprentissage guidé. Cela correspond à ce que nous nommons le principe d'accom-

modation. L'algorithme basé sur ce principe est nommé AGB (« accommodation goal babbling »).

Passons à la phase de production. Celle-ci consiste à produire la représentation motrice m et à percevoir la représentation sensorielle s' correspondante. La transformation de l'articulation en signal acoustique s'effectue, comme pour le maître, avec le modèle VLAM, via la distribution $P(S^{Env} \mid M^{Env})$. Pour des raisons de simplifications, et comme pour le maître, il est considéré que les représentations motrices M de l'agent sont équivalentes à leur réalisation M^{Env} et que les représentations sensorielles S de l'agent sont équivalentes au stimuli S^{Env} résultant de cette production (voir Fig.4.9).

Pour la phase de mise à jour, trois stratégies ont été mises en place. La première est une stratégie consistant à mettre simplement à jour le modèle interne $P(S \mid M)$ avec la représentation motrice m et la représentation sensorielle s'. Le prior moteur $P(M)$, lui, reste uniforme et n'influe pas sur l'apprentissage. C'est la stratégie qui est appliquée pour les algorithmes cités précédemment : RMB, RGB et AGB.

La deuxième stratégie est une stratégie d'exploration auto-centrée dans laquelle sont favorisées les représentations motrices précédemment sélectionnées. Elle consiste à mettre à jour le prior moteur $P(M)$ avec la représentation motrice m sélectionnée en plus du modèle interne $P(S \mid M)$. C'est ce que nous nommons le babillage idiosyncratique. L'algorithme correspondant à cette stratégie est nommé IGB (« idiosyncratic goal babbling). Il a été implémenté via le babillage d'objectifs avec une stratégie d'exploration guidée. Ainsi, il ne diffère d'AGB que dans la phase de mise à jour. Le fait d'avoir ces deux algorithmes permet de comparer l'influence du prior moteur lors de l'inférence d'une représentation motrice via la distribution $P(M \mid S)$ (voir Eq. 5.12).

Précisons la phase de mise à jour du prior. Dans les autres algorithmes, le prior $P(M)$ est constamment uniforme. Comme il évolue dans l'algorithme IGB, nous avons défini une manière de le mettre à jour. Nous considérons donc que le prior moteur est calculé dans cet algorithme à partir d'un paramètre sum_m qui correspond au nombre de fois qu'une valeur m a été sélectionnée. Ainsi, il possède 15 625 valeurs, c'est-à-dire le nombre de représentations motrices dans l'espace moteur. À chaque itération, ce paramètre est incrémenté pour la représentation motrice m sélectionnée et le prior $P(M)$ est ensuite mis à jour, ce qui s'effectue comme suit :

$$sum_m(m) = sum_m(m) + 1 \,, \tag{5.13}$$

$$P(M) = \frac{sum_m}{\sum_M sum_m} \tag{5.14}$$

L'importance du prior sélectionné peut varier selon l'initialisation de sum_m au début de l'apprentissage. En effet, si sum_m est élevé, l'incrémentation et la mise à jour ont peu d'effet sur le prior. Si au contraire, elle est faible, l'incrémentation et la mise à jour modifient grandement le prior. Nous avons testé plusieurs valeurs du paramètre d'initialisation, notées fi (« fréquence initiale ») : 1, ce qui correspond également à la valeur du paramètre d'incrémentation, $\frac{1}{15\,625}$ qui correspond à la probabilité initiale du prior $P(M)$ et une valeur intermédiaire 0,1.

La troisième stratégie correspond à une généralisation de l'exploration. Plus précisément, cela consiste à mettre à jour le voisinage de la représentation motrice m choisie. Cette troisième stratégie a été implémentée sur tous les algorithmes précédemment cités. Nous avons donc quatre nouveaux

algorithmes que nous nommons RMB-LE, RGB-LE, AGB-LE et IGB-LE correspondant respectivement aux algorithme RMB, RGB, AGB et IGB pour lesquels a été ajoutée une extrapolation locale, « LE » faisant référence ici à « local extrapolation ».

Plus précisément, une fois la représentation motrice m sélectionnée et la représentation sensorielle s' résultant de la production perçue, il est défini un ensemble de voisins m_v. Il s'agit des points de l'espace M les plus proches de la représentation motrice m. Pour choisir les voisins, nous définissons une gaussienne très piquée sur la représentation motrice m sélectionnée : sa moyenne vaut m et sa variance vaut 0,05. Les voisins correspondent aux points de l'espace des représentations motrices M dont la probabilité dépasse 0,1. Ensuite, pour chacun de ces voisins, la phase de mise à jour est appliquée pour les distributions concernées. Pour tous les algorithmes avec extrapolation locale, le modèle interne $P(S \mid M)$ est mis à jour avec, d'une part, la représentation motrice m et la valeur sensorielle s' et, d'autre part, les représentations motrices m_v et la valeur sensorielle s'. De plus, pour l'algorithme IGB-LE, le prior moteur $P(M)$ est mis à jour pour l'ensemble des valeurs m et m_v.

En résumé, nous avons donc défini huit algorithmes, variant soit sur la phase de sélection, soit sur la phase de mise à jour. Les caractéristiques de chacun sont récapitulées Tableau 5.1.

Algorithmes	Phase de sélection			Phase de mise à jour	
	Aléatoire moteur	Aléatoire sensorielle	Accommodation (guidée sensorielle)	Babillage idiosyncratique	Extrapolation locale
RMB	X				
RGB		X			
AGB			X		
IGB			X	X	
RMB-LE	X				X
RGB-LE		X			X
AGB-LE			X		X
IGB-LE			X	X	X

TABLE 5.1 – Résumé des spécificités de chaque algorithme

5.2.6 Outils d'évaluation

Pour chaque algorithme, l'apprentissage dure 100 000 itérations. À la fin de l'apprentissage, les algorithmes sont évalués sur deux critères : la qualité d'apprentissage et la quantité d'exploration.

La qualité de l'apprentissage est estimée par la capacité de l'agent à pouvoir reproduire les stimuli du maître. Pour réaliser cette analyse, nous mettons en place une tâche qui consiste à faire répéter à l'agent des données fournies par un maître. Plus précisément, cette tâche consiste dans le modèle à calculer la matrice de confusion $P(S^{Prod} \mid S^{Attendu})$ correspondant à la probabilité des signaux sensoriels S^{Prod} produits par l'agent, sachant les représentations sensorielles $S^{Attendu}$ fournies par un maître.

Le maître utilisé est celui défini précédemment, employé durant l'apprentissage par accommo-

dation. Les représentations sensorielles perçues $S^{Attendu}$ dans l'environnement sont tirées selon la distribution $P(S^{Env})$ qui se calcule par :

$$P(S^{Env}) = \sum_{O^{Maitre}} P(S^{Env} \mid O^{Maitre}) \,, \tag{5.15}$$

avec la distribution $P(S^{Env} \mid O^{Maitre})$ telle qu'elle a été définie à l'Eq. 5.3. Cela revient donc à calculer $P(S^{Env} \mid O^{Maitre})$, tous objets O^{Maitre} confondus.

L'inférence correspondant à la matrice de confusion est la suivante :

$$\begin{aligned} P(S^{Prod} \mid S^{Attendu}) &= \sum_{M} P(S^{Prod} \mid M)\, P(M \mid S^{Attendu}) \\ &\propto \sum_{M} P(S^{Prod} \mid M)\, P(M)\, P(S^{Attendu} \mid M) \,. \end{aligned}$$

Dans cette équation, le calcul s'effectue à l'aide de deux distributions : $P(M \mid S^{Attendu})$ et $P(S^{Prod} \mid M)$. La distribution $P(M \mid S^{Attendu})$ se calcule à l'aide du prior moteur $P(M)$ et du modèle interne $P(S \mid M)$ de l'agent (voir Eq. 5.12). De son côté, la distribution $P(S^{Prod} \mid M)$ correspond à la transformation des gestes moteurs en sons. Elle est équivalente à $P(S^{Env} \mid M^{Env})$.

Par la suite, nous ne conservons que la diagonale de cette matrice, notée $\mathrm{Diag}_{\mathrm{ConfMat}}$ qui donne, pour chaque représentation sensorielle $S^{Attendu}$ fournie par le maître, la probabilité que l'agent apprenant la reproduise à l'identique via le signal S^{Prod}. À l'aide de cette diagonale, nous calculons une mesure d'erreur, définie comme l'opposé (1 - $\mathrm{Diag}_{\mathrm{ConfMat}}$) de la moyenne des probabilités.

La méthode pour quantifier l'exploration est plus simple. Il s'agit de compter le nombre de représentations motrices différentes qui ont été sélectionnées durant l'apprentissage afin d'évaluer le volume de l'espace moteur exploré. Dans ce calcul, nous ne comptons que les représentations motrices choisies durant la phase de sélection. Ainsi les valeurs de voisinages utilisées pour les algorithmes avec extrapolation locale (RMB-LE, RGB-LE, AGB-LE et IGB-LE) ne sont pas considérées dans ce calcul. Bien entendu, il ne s'agit pas d'une tâche réaliste puisque cette mesure n'est pas quantifiable chez l'humain. C'est un des avantages d'un modèle computationnel.

5.2.7 Résultats

Nous analysons, un par un, les trois composants que nous avons mis en place lors de l'apprentissage sensorimoteur : le principe d'accommodation, le babillage idiosyncratique et l'extrapolation locale.

5.2.7.1 Le principe d'accommodation

Pour analyser les effets de l'accommodation, nous comparons les algorithmes RMB, RGB et AGB. Dans un premier temps, nous analysons en fin d'apprentissage $\mathrm{Diag}_{\mathrm{ConfMat}}$ pour ces trois

algorithmes. Ce résultat est présenté dans l'espace des représentations sensorielles dans la Fig. 5.10 de la même manière que la distribution $P(S^{Env} \mid O_S^{Maitre})$ du maître dans la Fig. 5.9.

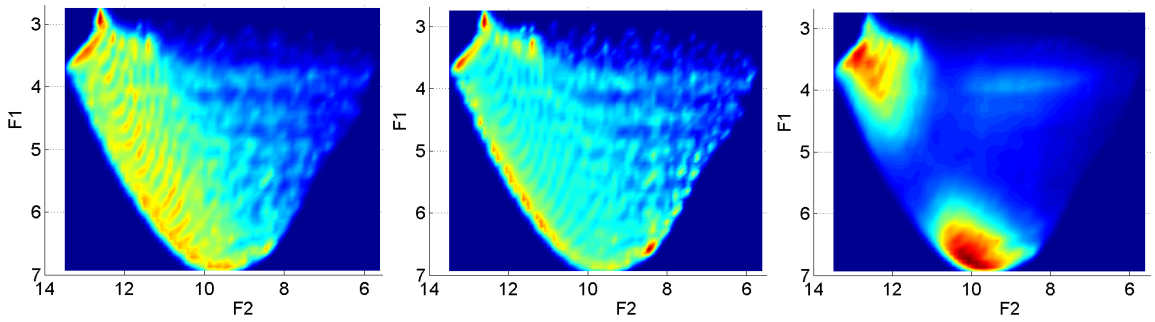


FIGURE 5.10 – Observation de $\mathrm{Diag}_{\mathrm{ConfMat}}$ pour les trois algorithmes RMB (**Gauche**), RGB (**Milieu**) et AGB (**Droite**). L'axe des abscisse correspond au formant F2 inversé et l'axe des ordonnées correspond au formant F1 inversé, en Barks. Les valeurs s'étendent du bleu foncé (faible probabilité) au rouge (forte probabilité)

Nous observons, d'une part, que pour les trois algorithmes, les valeurs de la partie gauche du triangle vocalique allant du [i] au [a], semblent mieux apprises que les autres valeurs du triangle vocalique. Plus précisément, dans l'algorithme RMB, cela semble concerner de façon globalement homogène toutes les valeurs de la diagonale gauche. Dans l'algorithme RGB, il semble y avoir trois niveau de précision, les valeurs les mieux apprises sont celles sur le bord gauche du triangle vocalique, ensuite ce sont les valeurs centrales de la partie gauche et enfin les autres valeurs du triangle vocalique. L'algorithme AGB fait un peu figure d'exception, nous observons que l'algorithme s'est essentiellement centré sur l'apprentissage des valeurs autour du [i] et du [a] du maître. Néanmoins, nous observons, qu'en plus de ces deux valeurs, la portion de l'espace sensoriel correspondant à la voyelle [u] du maître a une plus forte probabilité que la majorité des autres valeurs du triangle vocalique, même si elle reste plus faible que celle autour du [i] et du [a].

Le fait que les valeurs de la partie gauche du triangle vocalique soient mieux apprises que les autres n'est pas surprenant. En effet, rappelons que certaines représentations motrices correspondent à la même représentation sensorielle, propriété classiquement appelée « many-to-one ». Cela signifie donc qu'il y a plus d'antécédents moteurs correspondant à la partie gauche du triangle vocalique. Cette supposition est en accord avec les données expérimentales (voir par exemple Ménard et al., 2004, Fig. 3).

Pour interpréter plus précisément les résultats obtenus, rappelons que l'algorithme RMB sélectionne ses données en choisissant aléatoirement des représentations motrices à apprendre. Le résultat obtenu est donc représentatif de la proportion des représentations motrices correspondant à chaque représentation sensorielle. Cela implique un apprentissage plus précis de la partie gauche du triangle vocalique.

La phase de sélection dans l'algorithme RGB se fait en choisissant de façon aléatoire des représentations sensorielles. Les résultats obtenus sont révélateurs d'un apprentissage en deux phases. D'une part, l'apprentissage commence par une phase d'exploration aléatoire des représentations motrices comme dans l'algorithme RMB car l'agent ne connaît pas encore son modèle interne et ne peut

donc pas inférer les représentations motrices correctes correspondant aux représentations sensorielles tirées aléatoirement. De cela, en résulte un apprentissage de la partie gauche du triangle vocalique, comme pour l'algorithme RMB. Par la suite, le modèle interne étant peu à peu appris, l'agent arrive à inférer les représentations motrices correctes correspondant aux représentations sensorielles sélectionnées. Cependant, du fait que l'espace sensoriel est rectangulaire, certaines représentations sensorielles choisies aléatoirement sont en dehors du triangle vocalique. Ainsi, afin de les reproduire le mieux possible, l'agent choisit des représentations motrices donnant des représentations sensorielles au bord de l'espace vocalique, ce qui résulte en un fort apprentissage des valeurs sensorielles au bord du triangle vocalique. Ce même type de comportement a été observé précédemment (Moulin-Frier et Oudeyer, 2013).

Enfin, la phase de sélection dans AGB correspond au principe d'accommodation et s'effectue par guidage social à l'aide d'un maître. Comme pour l'apprentissage RGB, il est probable qu'au début de l'apprentissage, l'inférence des données motrices correspondant aux données du maître soient aléatoires et que, par la suite, l'agent, par apprentissage, arrive à se focaliser sur les données fournies par le maître. Comme il y a plus de représentations motrices pour la partie gauche du triangle vocalique, l'agent arrive d'abord à se focaliser sur les sons [a] et [i] fournis par le maître. Cela expliquerait pourquoi les sons [a] et [i] sont beaucoup mieux appris que les sons [u].

Nous comparons, dans un second temps, la mesure d'erreur et la proportion de l'espace exploré pour ces trois algorithmes au cours de l'apprentissage. Ceci est illustré Fig. 5.11 sur les courbes en pointillés. Globalement, nous observons que l'évolution de la mesure d'erreur est assez lente pour les trois algorithmes puisqu'il faut entre 25 000 et 30 000 itérations pour atteindre 0,5 et davantage pour converger vers une erreur minimum : 45 000 pour RGB, 80 000 pour AGB et plus de 100 000 pour RMB. Ainsi, RGB semble plus rapide que les deux autres algorithmes. En terme d'exploration, tous les algorithmes ont une exploration qui se fait de manière quasi exponentielle : très rapide au début et diminuant petit à petit. Plus de la moitié des points ont été observés dès 10 000 itérations. L'algorithme RGB est celui qui explore le plus rapidement l'ensemble de l'espace : 45 000 itération tandis que les deux autres terminent leur exploration aux alentours de 80 000 itérations. Cela correspond au moment où l'erreur de mesure converge pour RGB et AGB.

Interprétons ces résultats. L'algorithme RMB sert ici d'algorithme de référence pour savoir comment l'exploration s'effectue simplement en choisissant des représentations motrices aléatoires. Même après 100 000 itérations, l'erreur de mesure avec cet algorithme n'a pas convergé et n'a pas atteint le minimum malgré le fait que l'ensemble des points aient été observés. Cela vient du fait que certains points importants de l'espace moteur n'ont été explorés que très rarement, de façon insuffisante pour bien reproduire les sons du maître.

Le fait que les deux autres algorithmes, RGB et AGB, se comportent quasiment comme RMB au début d'apprentissage, jusqu'à environ 18 000 itérations, semble confirmer l'hypothèse qu'ils commencent tous les trois par une phase d'exploration dans laquelle les représentations motrices sont choisies de manière aléatoire car l'agent n'a pas de connaissance sur son modèle interne. Quand environ les deux tiers des points de l'espace ont été explorés, la connaissance sur le modèle interne devient suffisante pour que le modèle RGB prenne une trajectoire d'apprentissage différente. Le fait d'explorer aléatoirement l'espace des représentations sensorielles et d'avoir des représentations sensorielles non réalisables permet à l'agent de rapidement explorer l'ensemble des représentations motrices. Par

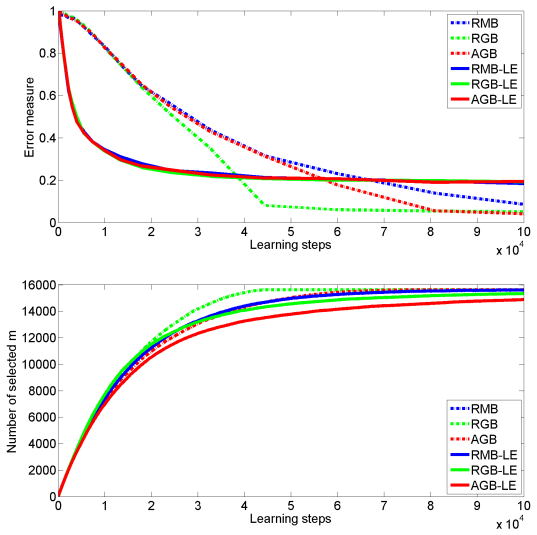


FIGURE 5.11 – Mesure d'erreur (**Haut**) et quantité d'exploration (**Bas**) à chaque itération, au cours de l'apprentissage, pour les algorithmes RMB/RMB-LE, RGB/RGB-LE et AGB/AGB-LE

ailleurs, du fait que ce sont les représentations sensorielles qui sont choisies aléatoirement, l'ensemble de l'espace sensoriel est aussi rapidement appris, ce qui permet à l'agent d'avoir une erreur de mesure minimale lorsque l'ensemble des points des représentations motrices a été exploré.

En ce qui concerne l'algorithme AGB, il faut plus de temps pour qu'il diffère de RMB. Cela s'explique certainement par le fait qu'il tente de n'explorer que les sons des voyelles [a, i, u]. Comme les représentations motrices de la voyelle [u] sont difficiles à trouver car moins nombreuses, l'agent doit explorer aléatoirement plus longtemps l'espace moteur. Nous pouvons supposer que les représentations motrices de la voyelle [u] sont trouvées lorsqu'il commence à différer de l'algorithme RMB. À ce moment là, il a néanmoins déjà exploré plus de 14 000 points sur 15 625. Cela lui permet d'avoir une erreur de mesure qui décroit plus vite que l'algorithme RMB, même si son exploration reste similaire.

5.2.7.2 Le babillage idiosyncratique

Passons maintenant au babillage idiosyncratique. Pour évaluer son effet, nous comparons les algorithmes AGB et IGB qui ne diffèrent que sur ce critère. L'analyse s'effectue comme précédemment par comparaison de la qualité d'apprentissage et de la quantité d'exploration.

En analysant la mesure d'erreur calculée à partir de $\mathrm{Diag}_{\mathrm{ConfMat}}$, nous remarquons dans un premier temps que la mesure d'erreur reste très élevée (entre 0,5 et 0,9) pour l'algorithme IGB, quelle que soit la valeur d'initialisation fi choisie. Nous nous intéressons donc peu à la comparaison des algorithmes dans l'espace sensoriel en fin d'apprentissage mais nous concentrons sur l'évolution de l'apprentissage au cours du temps, représentée sur la Fig 5.12, en pointillés.

Globalement, nous remarquons que l'algorithme IGB explore moins rapidement l'espace des re-

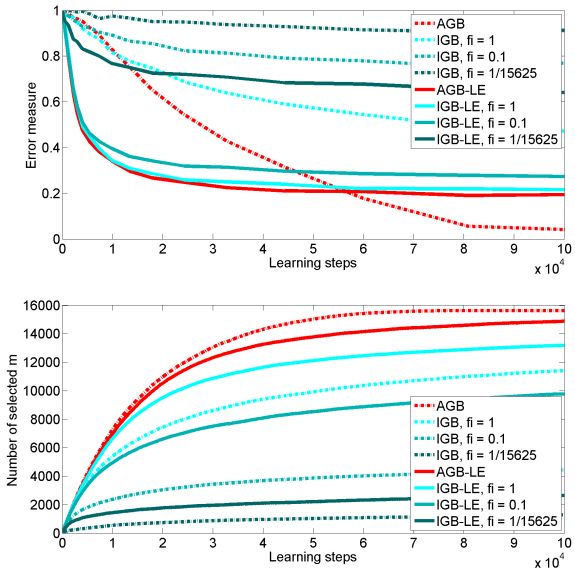


FIGURE 5.12 – Mesure d'erreur (**Haut**) et quantité d'exploration (**Bas**) à chaque itération, au cours de l'apprentissage pour les algorithmes AGB/AGB-LE et IGB/IGB-LE, avec différentes valeurs d'initialisation

présentations motrices que l'algorithme AGB au cours de l'apprentissage, toutes valeur fi confondues, mais qu'en conséquence l'erreur de mesure diminue également beaucoup moins rapidement. De plus, nous observons que plus la mesure d'initialisation fi est faible, moins il y a d'exploration.

Le fait que l'algorithme IGB explore moins que l'algorithme AGB est causé par la mise à jour du prior $P(M)$. Comme attendu, la mise à jour du prior favorise l'exploration des représentations motrices précédemment sélectionnées même si celles-ci ne correspondent pas forcément à celles données par le maître. Cela est notamment visible lorsque le prior évolue très vite, c'est-à-dire quand la valeur d'initialisation fi est faible. En conséquence, le nombre de points explorés est plus petit que pour l'algorithme AGB mais la mesure d'erreur reste plus élevée.

5.2.7.3 L'extrapolation locale

Nous terminons par l'analyse du dernier composant : l'extrapolation locale. Celui-ci ayant été testé sur tous les algorithmes précédemment définis, nous le comparons avec les algorithmes analogues. Nous commençons par comparer les algorithmes différant sur la phase de sélection, c'est-à-dire RMB/RMB-LE, RGB/RGB-LE et AGB/AGB-LE. Ensuite, nous analysons plus précisément la combinaison entre l'extrapolation locale et le babillage idiosyncratique via la comparaison entre IGB et IGB-LE.

Pour la comparaison des algorithmes différant sur la phase de sélection, nous nous basons sur la Fig. 5.11, traits pleins. De manière générale, nous remarquons que la mesure d'erreur des algorithmes de sélection avec extrapolation locale est quasiment identique au cours de l'apprentissage. Par ailleurs,

nous observons que pour une quantité d'exploration proche des algorithmes sans extrapolation locale, la mesure d'erreur diminue et converge beaucoup plus rapidement. En revanche, elle conserve, à convergence, une mesure d'erreur plus élevée que les algorithmes sans extrapolation locale.

En ce qui concerne la baisse rapide de l'exploration, cela vient du fait qu'à chaque itération, un grand nombre de représentations motrices sont mises à jour. Néanmoins, comme la mise à jour ne se fait pas avec la véritable valeur mais avec une valeur potentiellement proche, le modèle interne conserve une erreur de reproduction par rapport aux autres algorithmes à convergence. En ce qui concerne l'exploration, du fait que nous n'ayons compté que les représentations motrices sélectionnées et non pas l'ensemble du voisinage, l'exploration des algorithmes avec extrapolation locale est similaire aux autres. Néanmoins, l'apprentissage guidé, couplé avec une extrapolation locale, permet d'explorer moins qu'avec une sélection aléatoire. Cela signifie que l'extrapolation locale permet d'explorer plus efficacement l'ensemble et le guidage permet de se focaliser plus rapidement et évite une exploration complète de l'espace.

Si nous comparons maintenant avec l'algorithme avec babillage idiosyncratique (voir Fig. 5.12, traits pleins), nous remarquons globalement que la mesure d'erreur et la quantité d'exploration dépendent du paramètre initial fi. Néanmoins, l'évolution de la mesure d'erreur est similaire à celle des algorithmes précédents.

Comme précédemment, la valeur d'initialisation a beaucoup d'influence sur l'apprentissage et donne plus ou moins d'importance au prior moteur $P(M)$. L'extrapolation locale couplée à un babillage idiosyncratique permet une moindre exploration, ce qui vient du fait qu'un nombre plus importants de points du prior sont mis à jour lors de l'extrapolation locale. Néanmoins, cette exploration avec extrapolation locale reste plus faible avec babillage idiosyncratique pour une erreur quasi-similaire lorsque la valeur d'initialisation est suffisament élevée. Cela signifie que dans ces conditions, la mise à jour du prior moteur nécessite moins d'exploration pour reproduire correctement l'environnement.

5.2.7.4 Synthèse des résultats

Pour synthétiser l'ensemble de nos résultats, nous reprenons l'ensemble de nos résultats obtenus en fin d'apprentissage. Dans un premier temps, nous calculons la quantité d'exploration de l'espace moteur, c'est-à-dire le pourcentage de cases explorées parmi les 15 625 cases de l'espace moteur. Le résultat est ramené entre 0 et 1 (0 pour une exploration nulle et 1 pour une exploration totale). Dans un second temps, nous calculons la qualité de l'apprentissage, qui est l'opposé de l'erreur d'apprentissage (0 pour un apprentissage médiocre, 1 pour un apprentissage parfait). Nous reportons ces deux mesures sur un même graphique, présenté Fig. 5.13.

Ce graphique permet de mettre en avant trois principales observations. Premièrement, à exploration complète, les algorithmes AGB, RGB et RMB ont tous trois une grande qualité d'apprentissage. AGB reste tout de même l'algorithme avec les meilleures performances. Deuxièmement, IGB-LE, avec une initialisation à 1, possède une qualité d'apprentissage quasiment identiques aux autres algorithmes avec extrapolation locale (RMB-LE, RGB-LE et AGB-LE) mais en explorant moins l'espace. Les trois principes, regroupés dans un même algorithme, semblent donc être un compromis intéressant entre quantité d'exploration et qualité d'apprentissage. Troisièmement, lorsque l'exploration est très

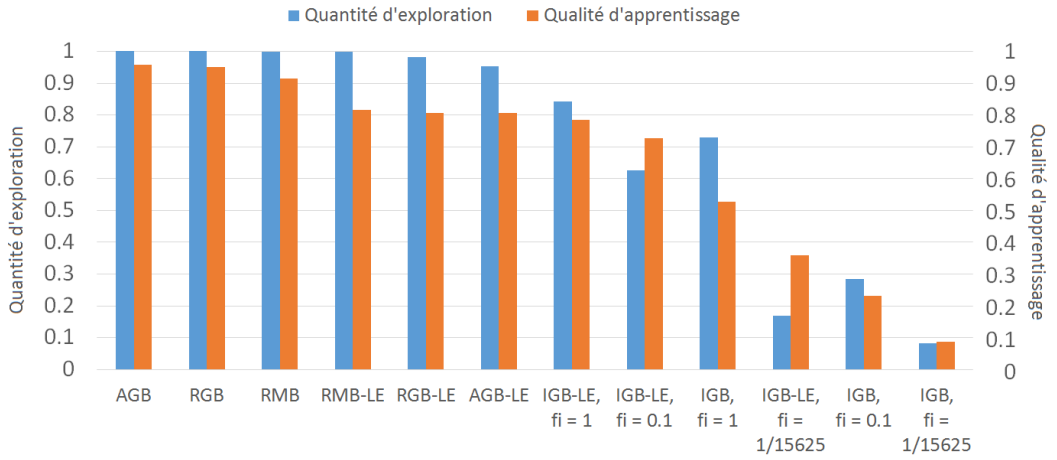


FIGURE 5.13 – Synthèse des résultats en fin d'apprentissage : quantité d'exploration (en bleu) et qualité d'apprentissage (en orange), le tout ordonné selon la qualité d'apprentissage

faible, la qualité d'apprentissage l'est également, ce qu'illustre les algorithmes IGB et IGB-LE avec une faible valeur d'initialisation.

5.2.8 Discussion

5.2.8.1 Résumé des résultats

En résumé, nous avons exploré le comportement de huit algorithmes sensorimoteurs. Nous nous sommes particulièrement intéressés aux effets de trois critères : le principe d'accommodation, le babillage idiosyncratique et l'extrapolation locale. Nous avons observé qu'un algorithme muni de ces trois critères est le meilleur compromis pour obtenir, non seulement une focalisation sur les données de l'environnement, mais également une faible erreur de reproduction en explorant le moins possible l'espace moteur. En effet, nous avons observé, dans un premier temps, que les algorithmes sans extrapolation locale ni babillage idiosyncratique permettent d'obtenir une faible mesure d'erreur mais nécessitent une exploration totale de l'espace moteur. Parmi ces algorithmes, celui avec un principe d'accommodation permet de se focaliser sur les données de l'environnement. L'ajout du babillage idiosyncratique, sans extrapolation locale, diminue considérablement l'exploration mais ne permet pas d'avoir une faible mesure d'erreur. Lorsque l'extrapolation locale est ajoutée, sans babillage idiosyncratique, la mesure d'erreur diminue très rapidement mais l'exploration bien qu'un peu plus faible que les algorithmes de base reste élevée en fin d'apprentissage. Lorsque le babillage idiosyncratique et l'extrapolation locale sont tous deux combinés, la mesure d'erreur obtenue est faible et l'algorithme nécessite moins d'exploration. Il s'agit donc du meilleur compromis.

En termes d'interprétation, le principe d'accommodation est un apprentissage guidé dans lequel l'agent tente de reproduire les données fournies par un maître. De ce fait, il permet d'apprendre en priorité les données de l'environnement. Le babillage idiosyncratique assure la focalisation sur les représentations motrices précédemment apprises à travers la mise à jour des données de l'environne-

ment. De par cette focalisation, l'exploration de l'espace des représentations motrices est diminuée. Enfin, l'extrapolation locale explore le voisinage d'une représentation motrice sélectionnée et permet d'approximer l'espace des représentations motrices. Cette approximation a pour conséquence de tester des points de l'espace assez diversifiés ce qui permet de trouver plus rapidement les points donnant une faible mesure d'erreur.

5.2.8.2 Limites et perspectives

Les résultats obtenus, bien que satisfaisants, pourraient être améliorés. Nous citons quelques exemples ici. Une première amélioration concerne la mesure d'erreur finale des algorithmes avec extrapolation locale. En effet, comme nous l'avons vu, ces algorithmes conservent en fin d'apprentissage une erreur résiduelle plus élevée que ceux sans extrapolation locale tels que RMB, RGB et AGB. Cela vient du fait que l'algorithme réalise une approximation des représentations motrices au voisinage de celle sélectionnée en considérant que le voisinage possède la même valeur que la représentation motrice sélectionnée. Une possibilité pour résoudre ce problème serait de cesser l'extrapolation locale lorsque la mesure d'erreur commence à converger. Cela permettrait d'ajuster plus finement les représentations motrices aux bonnes représentations sensorielles.

Une seconde amélioration pourrait également être réalisée avec l'algorithme IGB. En effet, nous avons vu que, pour ces algorithmes, l'exploration est limitée mais ils conservent une forte mesure d'erreur. Un moyen pour éviter ce biais serait de complexifier le processus d'apprentissage et de ne mettre à jour le prior moteur que lorsqu'il correspond à une donnée de l'environnement. Cela nécessiterait une vérification que la donnée produite s' s'approche bien de la donnée fournie par le maître s. Une troisième amélioration pourrait également être proposée sur le principe d'accommodation. Comme nous l'avons vu en comparant les algorithmes RMB, RGB et AGB, au départ l'exploration est aléatoire car l'agent n'a pas de connaissance sur son modèle interne et une exploration aléatoire de l'espace sensoriel permet de réduire plus rapidement la mesure d'erreur que l'algorithme AGB. Ainsi, il peut être supposé qu'une exploration aléatoire au début d'apprentissage, couplée à l'extrapolation locale, permettrait d'obtenir encore plus rapidement une faible mesure d'erreur. Par la suite, le principe d'accommodation, couplé au babillage idiosyncratique, permettrait de se focaliser sur les données de l'environnement en se concentrant précisément sur certaines représentations motrices.

Passons maintenant aux limites de cette étude. La plupart de ces limites concernent le réalisme de l'apprentissage sensorimoteur effectué. Nous en citons quelques unes ici. Premièrement, le choix du modèle à apprendre est discutable. Comme pour le modèle COSMO, nous avons décidé que notre agent possède un modèle interne direct $P(S \mid M)$. Dans cet étude, celui-ci est couplé à un prior moteur $P(M)$. Si le modèle interne a été beaucoup étudié pour explorer l'apprentissage sensorimoteur (voir chapitre précédent), ce n'est pas la seule alternative possible. La distribution $P(S\ M)$ pourrait par exemple être considérée dans son ensemble (Moulin-Frier et Oudeyer, 2013) ou décomposée selon l'équation :

$$P(S\ M) = P(S)P(M \mid S)\ , \tag{5.16}$$

avec apprentissage d'un prior sensoriel $P(S)$ et d'un modèle inverse $P(M \mid S)$ (voir par exemple Philippsen et al., 2014, qui considèrent l'existence d'un modèle direct et inverse). Nous avons choisi le modèle direct car il est plus simple à implémenter.

Deuxièmement, le réalisme des critères implémentés est lui aussi contestable. Nous avons vu précédemment que le bébé se focalise, dès la phase de babillage, sur les données de son environnement. C'est ce que permet le principe d'accommodation. Néanmoins, la manière de faire pourrait sans doute être améliorée. Par exemple, le fait de n'avoir qu'un maître n'est pas réaliste. Néanmoins, nous considérons que dans cette étude, ajouter d'autres maîtres aurait peu d'influence sur nos observations. En ce qui concerne le babillage idiosyncratique, plusieurs études montrent que chaque individu réalise des gestes moteurs qui lui sont propres, même ceux ayant grandi dans le même environnement (Rapin et al., 2017), ce qui semble valider l'hypothèse qu'il existe bien une focalisation sur des représentations motrices précises chez chaque individu. Cependant, la mise à jour du prior moteur $P(M)$ nécessiterait des analyses plus approfondies. Il semble, par exemple, peu probable que le prior soit uniforme au début de l'apprentissage.

5.3 Conclusion

Ces deux études nous mènent à deux conclusions. Dans un premier temps, à travers la première étude, nous estimons que, même dans des simulations très simplifiées, nous sommes capables d'observer des propriétés génériques du modèle. C'est ainsi que nous avons mis en avant la propriété «bande étroite/bande large ». Dans un second temps, la seconde étude montre, à travers l'exemple de l'apprentissage sensorimoteur, que l'implémentation des apprentissages reste un problème complexe, nécessitant des réflexions aussi bien sur les stratégies d'exploration que sur les méthodes de mise à jour. L'étude que nous avons menée à cet effet, davantage penchée sur des préoccupations relevant de la robotique cognitive que du développement phonétique, a mis en avant trois principes pouvant faciliter l'apprentissage du lien entre les représentations sensorielles et motrices : le principe d'accommodation, le babillage idiosyncratique et l'extrapolation locale. Si nous souhaitons complexifier davantage nos implémentations, il semble donc nécessaire de mieux comprendre comment l'apprentissage se complexifie lui aussi. Ceci nous conduit à étudier la variabilité de l'apprentissage, et, plus précisément, les idiosyncrasies, qui sont les thématiques que nous abordons dans le chapitre suivant.

CHAPITRE 6

Variabilité des unités distinctives

Dans le chapitre précédent, nous avons observé, dans un premier temps, que les différences d'apprentissage entre les représentations sensorielles et motrices des catégories phonétiques induisent des différences d'exploration et des différences de variance entre les distributions, ce qui nous a amené à proposer l'existence d'une propriété « bande étroite/bande large ». Dans un second temps, en nous penchant sur le développement sensorimoteur, nous avons abordé la complexité de l'exploration sensorimotrice à travers plusieurs algorithmes et montré que le choix des représentations motrices choisies influence la qualité de l'apprentissage sensorimoteur. Nous pouvons donc affirmer à travers ces deux études que les différences d'apprentissage influent directement sur la précision et la variabilité des distributions.

Dans ce chapitre, nous nous intéressons plus précisément à cette variabilité en axant nos recherches sur les différences d'apprentissage inter-locuteurs. Les idiosyncrasies, évoquées dans le chapitre 2, sont un bon moyen de les étudier. Pour rappel, il s'agit des différences de perception et production propres à chaque individu, qui résistent même après normalisation des différences physiques. Nous présentons deux études basées sur les idiosyncrasies, toutes deux réalisées avec le modèle COSMO. Dans la première, nous nous intéressons à leur apparition au cours de l'apprentissage moteur et nous observons leur conséquence sur la production de la parole. Dans la seconde, nous étudions le comportement des idiosyncrasies durant la perception en analysant les représentations sensorielles et motrices des unités phonétiques à travers les trois familles de théories de la perception.

6.1 Etude de l'apparition des idiosyncrasies

Publication :

— Barnaud, M.-L., Diard, J., Bessière, P., et Schwartz, J.-L. (2016a). Assessing Idiosyncrasies in a Bayesian Model of Speech Communication. In Proceedings of Interspeech 2016, pages 2080–2084

Comme évoqué dans le chapitre 2, les idiosyncrasies ont globalement été assez peu explorées et leurs propriétés restent assez mystérieuses. Dans cette étude, nous souhaitons analyser leur apparition durant le développement, aspect des idiosyncrasies qui, à notre connaissance, n'a pas encore été examiné. Nous faisons l'hypothèse que ces idiosyncrasies apparaissent lors de l'apprentissage moteur

et, plus précisément, lorsque l'agent cherche à inférer des gestes moteurs correspondant à ce qu'il perçoit. À l'aide du modèle COSMO, nous comparons, durant une tâche de production, deux modèles différant par leur apprentissage moteur : dans le premier, l'agent a pour but de reproduire les signaux du maître qu'il perçoit, ce que nous nommons « imitation », dans le second, il a pour but de reproduire l'objet sélectionné par le maître, ce que nous nommons « communication ». Les résultats montrent que seul le second modèle peut générer des idiosyncrasies.

6.1.1 Hypothèse de l'étude

Commençons par une clarification terminologique. Une idiosyncrasie en production de parole peut se traduire, pour une même unité phonétique, soit par des variations articulatoires à résultat acoustique constant, soit par des variations acoustiques, impliquant, bien sûr, également des variations articulatoires. C'est sur ces secondes que porte le présent chapitre.

Définissons maintenant les hypothèses possibles concernant leur apparition. Pour cela, reprenons les résultats obtenus précédemment. Nous avons montré, dans la première étude du chapitre précédent, avec COSMO 1D, que l'apprentissage auditif conduit à une perception auditive très piquée sur les signaux du maître. Nous en déduisons ainsi que, durant son apprentissage auditif, l'agent apprend à reproduire les signaux du maître et, de façon plus générale, les signaux de son environnement. À travers nos simulations, nous avons d'ailleurs observé que les représentations auditives sont très similaires d'une simulation à l'autre. Nous supposons que ce type de comportement n'est pas adapté pour faire apparaître les idiosyncrasies acoustiques qui, par définition, impliquent des différences entre les individus.

Par ailleurs, nous avons également observé, dans cette même étude sur COSMO 1D, que la perception motrice semble centrée sur les prototypes du maître, mais en introduisant, en plus, une capacité de généralisation. Cela ne semble pas provenir de l'apprentissage sensorimoteur, puisque, comme le montre la seconde étude du chapitre précédent, l'apprentissage sensorimoteur par accommodation, même s'il commence par explorer diverses portions de l'espace sensoriel, finit par se focaliser sur l'apprentissage des signaux de l'environnement.

Il reste donc l'apprentissage moteur lors duquel l'agent apprend son répertoire moteur. Nous supposons que l'agent, à travers son apprentissage moteur, pourrait ne pas reproduire exactement les signaux du maître mais apprendre des gestes moteurs adéquats capables de reproduire globalement ce que produit le maître. Dans cette étude, nous interprétons l'expression « ce que produit le maître » de deux façons. Dans un premier temps, nous considérons « ce que produit le maître » comme un signal sonore. Selon cette interprétation, le but de l'agent apprenant, durant son apprentissage moteur, est donc de sélectionner des gestes moteurs capables d'imiter les signaux du maître. Nous nommons ce processus « l'apprentissage par imitation ». Dans un second temps, nous interprétons « ce que produit le maître » comme la réalisation des objets phonétiques. Dans ce cas, l'agent apprenant a pour but de sélectionner des gestes moteurs capables de reproduire l'objet produit par le maître. Nous nommons ceci « l'apprentissage par communication » puisque, dans cette situation, l'agent apprenant a pour but de produire les mêmes objets que son maître, qui sont les éléments de la communication.

Notre étude repose sur la différence entre ces deux types d'apprentissage, que nous jugeons pri-

mordiale pour mieux comprendre comment apparaissent les idiosyncrasies. Notre hypothèse est que les idiosyncrasies acoustiques n'apparaissent pas entre les agents dans l'apprentissage par imitation puisqu'ils cherchent à reproduire exactement les signaux du maître. En revanche, dans l'apprentissage par communication, nous supposons que chaque agent choisit ses propres gestes moteurs pour reproduire l'objet produit par le maître et que les signaux acoustiques correspondant à ces gestes produits peuvent être différents. Nous anticipons le fait que ce processus devrait faire apparaître des idiosyncrasies acoustiques.

6.1.2 Implémentation du modèle

Les deux études de ce chapitre sont dédiées à l'analyse des idiosyncrasies. C'est pourquoi, nous utilisons la même version du modèle COSMO. Celle-ci est proche de la version utilisée dans la deuxième étude du chapitre précédent, notamment en ce qui concerne l'implémentation des espaces sensoriel et moteur. Néanmoins, quelques adaptations sont nécessaires pour étudier les idiosyncrasies, notamment en ce qui concerne les catégories distinctives. Pour cela, nous nous inspirons des résultats de Ménard et Schwartz (2014), celle-ci servant d'étude de référence dans la prochaine partie.

6.1.2.1 Implémentation des variables de l'agent

Comme dans l'étude de Ménard et Schwartz (2014), les catégories phonétiques sur lesquelles nous nous basons sont des voyelles. C'est pourquoi cette nouvelle version de notre modèle se nomme COSMO-Voyelle, abrégée parfois par COSMO-V. Dans cette version, nous considérons sept voyelles [a i u e ɛ o ɔ] qui forment un système très utilisé dans les langues du monde (Schwartz et al., 1997). Elles ont l'intérêt de pouvoir être catégorisées selon deux paramètres formantiques composant le triangle vocalique. Ainsi, bien que Ménard et Schwartz (2014) en utilisent davantage dans leur étude (qui comprend également la série des voyelles antérieures arrondies [y œ ø]), nous nous focalisons sur ces sept voyelles pour faciliter l'implémentation du modèle. Ces voyelles correspondent aux sept valeurs des objets O_L et O_S.

Nos voyelles sont caractérisées dans l'espace auditif et moteur. Nous utilisons les mêmes paramètres, définis dans les mêmes intervalles de valeurs, que dans la seconde étude du chapitre précédent. Pour rappel, pour les caractériser dans l'espace sensoriel S, nous utilisons les formants F1 et F2. Pour les caractériser dans l'espace moteur, nous utilisons les trois paramètres articulatoires, issus du modèle articulatoire du conduit vocal VLAM (voir section 5.2.2) : l'un réglant la hauteur des lèvres (LH : Lip Height), le deuxième réglant la position horizontale de la langue (TB : Tongue Body), le troisième réglant la position verticale de la langue (TD : Tongue Dorsum). Les intervalles de valeurs des formants et des articulateurs sont également similaires à ceux de l'étude précédente. La seule différence est que, dans cette étude, les dimensions motrices sont discrétisées en 15 valeurs équitablement réparties, de manière à avoir un espace à 3 375 valeurs (15^3). Cette réduction du nombre de valeurs de l'espace moteur permet d'accélérer les calculs tout en restant suffisamment élevé pour ne pas biaiser les calculs et observations que nous réalisons.

6.1.2.2 Implémentation des distributions

Nous définissons les distributions $P(M \mid O_S)$, $P(S \mid O_L)$ et $P(S \mid M)$ par des ensembles de distributions gaussiennes multi-dimensionnelles discrétisées et tronquées comme dans le chapitre précédent (voir Eq. 5.10 et Eq. 5.11). Elles sont paramétrées par une moyenne μ et une matrice de covariance Σ. Le prior $P(O_S)$ est une distribution uniforme puisque les fréquences des objets ne nous intéressent pas.

Dans cette étude, le répertoire moteur $P(M \mid O_S)$ et le répertoire auditif $P(S \mid O_L)$ sont des ensembles de sept distributions gaussiennes, une pour chaque objet O (O_S et O_L). Le répertoire interne $P(S \mid M)$ est, quant à lui, un ensemble de 3 375 distributions gaussiennes, une pour chaque valeur de M.

6.1.2.3 Implémentation de l'environnement

Comme précédemment, les stimuli de l'environnement sont fournis par un maître à partir de la distribution $P(S \mid O_S^{Maitre})$. Celle-ci est schématisée Fig. 6.1.

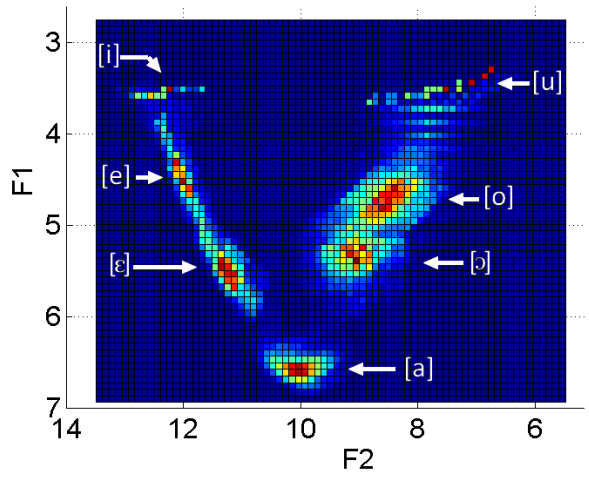


FIGURE 6.1 – Distribution des stimuli du maître que perçoit l'agent durant son apprentissage. L'axe des abscisses correspond au formant F2 inversé et l'axe des ordonnées correspond au formant F1 inversé, en Barks. Les valeurs s'étendent du bleu foncé (faible probabilité) au rouge (forte probabilité). Les cases représentent la discrétisation

Pour rappel, cette distribution $P(S \mid O_S^{Maitre})$ est calculée principalement avec le répertoire moteur $P(M^{Maitre} \mid O_S^{Maitre})$ et la transformation des productions motrices en signaux $P(S^{Env} \mid M^{Env})$. Le maître étant un agent COSMO, les variables sont similaires à celles de l'agent apprenant. Ainsi, l'espace des objets O_S^{Maitre} du maître contient sept valeurs, caractérisant les sept voyelles choisies, et les représentations motrices M^{Maitre} correspondent aux trois paramètres articulatoires de VLAM : TB, TD et LH. De la même manière, le geste moteur M^{Env} correspond également aux trois paramètres TB, TD et LH tandis que S^{Env} correspond aux deux premiers formants F1 et F2.

Les deux distributions $P(M^{Maitre} \mid O_S^{Maitre})$ et $P(S^{Env} \mid M^{Env})$ correspondent toutes deux à des ensembles de distributions gaussiennes, paramétrées par une moyenne μ et une matrice de covariance Σ (cf Eq. 5.10 et Eq. 5.11).

En termes d'initialisation, les distributions sont définies globalement de la même manière que dans la seconde étude du chapitre précédent (voir section 5.2.4) : la moyenne μ de chacune des sept distributions gaussiennes de la distribution $P(M^{Maitre} \mid O_S^{Maitre})$ est un prototype articulatoire de la voyelle correspondante et la variance de la matrice de covariance vaut 0,1. Cette variance a été conservée de manière à simuler la variabilité intra-locuteur du maître tout en assurant la séparabilité des voyelles dans l'espace moteur. En ce qui concerne $P(S^{Env} \mid M^{Env})$, la moyenne μ de chaque gaussienne correspond à la valeur des formants (en Barks) fournie pour chacune des configurations du dictionnaire de VLAM et la matrice de covariance, exprimée (en Barks), vaut :

$$\Sigma = \begin{pmatrix} 0.01 & 0 \\ 0 & 0.01 \end{pmatrix} .$$

La variabilité de production du maître, ainsi que celle issue des artéfacts de l'environnement, sont donc considérées dans cette étude comme très faibles. Cette quasi-absence de variabilité permet de faciliter l'analyse des mécanismes permettant l'apparition des idiosyncrasies, dépendant uniquement des différences entre individus.

6.1.3 Apprentissage du modèle

6.1.3.1 Généralités sur l'apprentissage

L'apprentissage concerne les trois phases du développement présentées dans le chapitre 3 : l'apprentissage sensoriel, l'apprentissage sensorimoteur et l'apprentissage moteur. Dans cette étude, elles sont réalisées successivement et de manière indépendante. Ceci est important pour l'apprentissage moteur que nous souhaitons étudier indépendamment des deux autres. Nous supposons donc que l'apprentissage sensoriel et sensorimoteur sont déjà terminés (ou du moins bien avancés) quand l'apprentissage moteur commence.

Les trois apprentissages ont chacun été réalisés sur 25 000 itérations. Nous implémentons 12 simulations qui diffèrent uniquement par le tirage des stimuli envoyés par le maître via la distribution $P(S \mid O_S^{Maitre})$. Ces 12 simulations peuvent être vues comme 12 agents différents, opérant dans le même environnement d'apprentissage (même maître et mêmes caractéristiques d'implémentation). Au début de l'apprentissage, les distributions gaussiennes $P(M \mid O_S)$, $P(S \mid M)$ et $P(S \mid O_L)$ de l'agent apprenant sont initialisées avec une moyenne μ située au centre de l'espace correspondant et avec des termes diagonaux de la matrice de covariance Σ de la taille de cet espace. Cela permet de simuler des distributions quasi-uniformes en début d'apprentissage.

6.1.3.2 Description des deux méthodes d'apprentissage

Décrivons maintenant plus en détail comment s'effectue l'apprentissage moteur (pour rappel, voir section 4.2.3.4). Comme énoncé précédemment (voir section 6.1.1), nous considérons deux types d'apprentissage moteur : un apprentissage par imitation et un apprentissage par communication.

Dans COSMO, la différence entre ces deux apprentissages moteurs s'effectue au niveau de l'inférence d'un geste moteur lorsque le maître a envoyé à l'agent un signal acoustique s et l'objet correspondant o. Lors de l'apprentissage par imitation, cette inférence s'effectue grâce à un tirage sur la distribution $P(M \mid [S = s]\ [O_S = o])$ (voir Eq. 4.7 du chapitre 3). Lors de l'apprentissage par communication, elle s'effectue à l'aide d'un tirage sur la distribution $P(M \mid [O_S = o]\ [C = 1])$. Mise à part cette différence, le reste de l'apprentissage moteur est identique entre les deux méthodes.

Etudions plus précisément ces deux distributions. Comme il a déjà été précisé dans le chapitre 3 (voir Eq. 4.7), la distribution $P(M \mid [S = s]\ [O_S = o])$ utilisée dans l'apprentissage par imitation se décompose comme suit :

$$P(M \mid [S = s]\ [O_S = o]) \propto P(M \mid [O_S = o])P([S = s] \mid M)\ . \tag{6.1}$$

Dans cette équation, le répertoire moteur $P(M \mid O_S)$ permet de choisir une représentation motrice correspondant à l'objet o du maître tandis que le modèle interne $P(S \mid M)$ permet de choisir une représentation motrice correspondant au stimulus s envoyé par le maître. Comme les stimuli s et l'objet o du maître sont toujours envoyés simultanément, si la représentation motrice sélectionnée permet de reproduire la représentation sensorielle s perçue, alors elle permet également de reproduire l'objet o envoyé. C'est pourquoi, nous supposons que, si le modèle interne a bien convergé en fin d'apprentissage sensorimoteur, alors les représentations motrices sélectionnées à partir du modèle interne $P([S = s] \mid M)$ permettent non seulement de reproduire les bonnes représentations sensorielles s mais également les bons objets o. Le répertoire moteur $P(M \mid O_S)$ permettrait de valider ce dernier choix. Dans cette situation, tous les agents devraient donc se focaliser sur la reproduction des représentations sensorielles du maître et aucune idiosyncrasie ne devrait apparaître entre eux.

De son côté, la distribution $P(M \mid [O_S = o][C = 1])$ se décompose par :

$$\begin{aligned} P(M \mid [O_S = o][C = 1]) &\propto P(M \mid [O_S = o]) \sum_S P(S \mid M)P([O_L = o] \mid S)\ , &(6.2)\\ &\propto P(M \mid [O_S = o])P(M \mid [O_L = o])\ . &(6.3)\end{aligned}$$

Cette équation est composée du répertoire moteur $P(M \mid O_S)$, qui permet de choisir une représentation motrice correspondant à l'objet o du maître, et d'une combinaison entre le modèle interne $P(S \mid M)$ et le décodeur auditif $P(S \mid O_L)$, équivalente à $P(M \mid O_L)$, qui permet, elle aussi, de choisir une représentation motrice correspondant à l'objet o du maître. Tandis que la première de ces deux distributions a comme effet de privilégier les représentations motrices déjà apprises, la seconde oriente le tirage vers des stimuli susceptibles d'être bien décodés par le maître, sans être nécessairement ceux que le maître lui-même utilise. Du fait que les deux distributions utilisées, $P(M \mid O_S)$ et $P(M \mid O_L)$, infèrent des représentations motrices non pas à partir des stimuli de l'environnement mais des objets phonétiques, nous nous attendons à ce que le répertoire appris ne soit pas une copie de l'environnement et qu'il apparaisse des idiosyncrasies sensorielles acoustiques entre les agents.

6.1.4 Outils d'évaluation

6.1.4.1 Etude de l'évolution de l'apprentissage

Nous étudions d'abord l'évolution de l'apprentissage au cours du temps. Comme nous nous intéressons aux idiosyncrasies acoustiques, nous étudions l'évolution du système perceptif durant l'apprentissage et, plus particulièrement, celle des branches auditive $P(S \mid O_L)$ et motrice $P(S \mid O_S)$. Bien que l'implémentation ait changé, elles se définissent comme dans le chapitre précédent par (voir Eq. 5.4 et Eq. 5.5) :

$$P(S \mid O_L) = P(S \mid O_L) \,, \tag{6.4}$$

$$P(S \mid O_S) \propto \sum_M P(M \mid O_S)\, P(S \mid M) \,. \tag{6.5}$$

Pour les étudier, nous nous basons sur l'étude de l'entropie (voir Eq. 5.6) au cours de l'apprentissage. La méthode est exactement la même que dans le chapitre précédent : nous calculons, pour chaque agent, l'entropie de la distribution $P(S \mid [O = o])$ pour chaque objet o. Ensuite, nous moyennons ces différentes entropies pour avoir l'entropie moyenne pour chaque distribution $P(S \mid O)$. Enfin, nous moyennons les entropies obtenues par chaque agent. Nous obtenons ainsi l'évolution moyenne de l'entropie pour la branche auditive $P(S \mid O_L)$, d'une part, et pour la branche motrice $P(S \mid O_S)$, d'autre part.

6.1.4.2 Etude des idiosyncrasies

Nous étudions par la suite l'apparition des idiosyncrasies acoustiques lors des deux apprentissages moteurs que nous avons définis. Pour cela, nous analysons le résultat acoustique d'une tâche de production dans laquelle les agents produisent les sept voyelles demandées.

Concrètement, cela revient à analyser la distribution $P(S^{Env} \mid O_S)$, qui se définit par :

$$P(S^{Env} \mid O_S) \propto \sum_M P(M \mid O_S)\, P(S^{Env} \mid M^{Env}) \,. \tag{6.6}$$

Dans cette équation, le répertoire moteur $P(M \mid O_S)$ est utilisé pour produire un geste moteur m correspondant à un objet o. Ensuite, cette production est envoyée dans l'environnement et transformée en signal acoustique s grâce à la distribution $P(S^{Env} \mid M^{Env})$, précédemment définie.

Pour voir s'il apparaît des idiosyncrasies, nous étudions, pour chaque voyelle, les valeurs individuelles acoustiques moyennes de chaque agent, notées mi (moyennes individuelles). Elles correspondent, dans le modèle COSMO, à la position du stimulus moyen produit, pour chaque voyelle, dans F1/F2 de l'espace auditif S. Autrement dit, cela revient à calculer la moyenne de la distribution $P(S^{Env} \mid O_S)$ pour chaque voyelle O_S.

6.1.5 Résultats

6.1.5.1 Evolution de l'apprentissage

Nous nous intéressons, dans un premier temps, à l'évolution de l'apprentissage de nos deux versions du modèle, au moyen des outils définis dans la section 6.1.4.1. Les résultats sont présentés dans la Fig. 6.2. Pour ne pas alourdir la figure, nous nous concentrons sur l'apprentissage moyen et les différences d'apprentissage entre agents ne sont pas affichées.

Dans cette étude, comme les apprentissages sont réalisés de manière séparée, nous découpons l'évolution de l'entropie en deux étapes, durant chacune 25 000 itérations. La première, (itérations 1 à 25 000 sur la figure), concerne l'évolution conjointe de l'apprentissage sensoriel et sensorimoteur, qui est commune aux deux versions. La seconde (itérations 25 001 à 50 000 sur la figure) concerne l'évolution de l'apprentissage moteur, qui est réalisé soit par un apprentissage par imitation, soit par un apprentissage par communication.

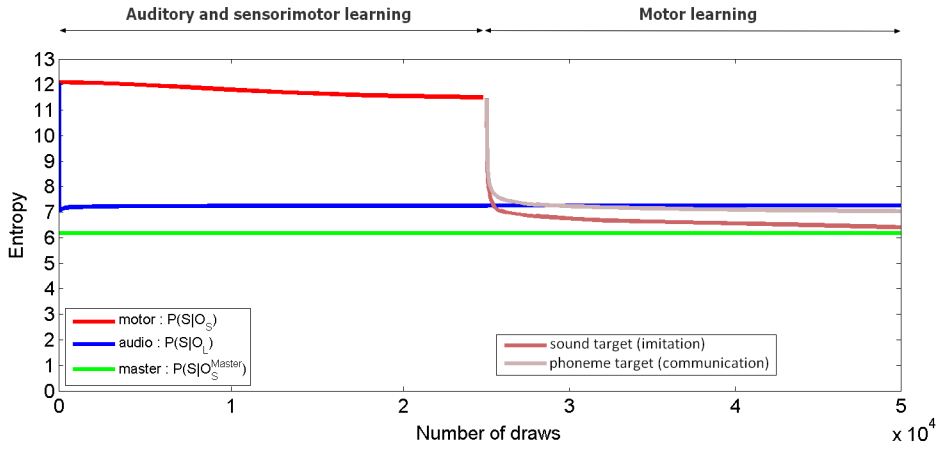


FIGURE 6.2 – Evolution de l'entropie des branches auditive et motrice au cours de l'apprentissage. Entropie du maître (en vert) et entropie de la branche auditive (en bleu). Entropie de la branche motrice : durant l'apprentissage sensorimoteur (en rouge), durant l'apprentissage par imitation (en marron), durant l'apprentissage par communication (en beige)

Commençons par analyser la partie commune aux deux versions : l'apprentissage auditif et l'apprentissage sensorimoteur. Au début de l'apprentissage, la branche auditive et la branche motrice ont toutes les deux une entropie beaucoup plus grande que le maître (environ 12 versus environ 6). Cela correspond, dans notre étude, à l'entropie d'une distribution uniforme. Par la suite, les deux branches perceptives n'évoluent pas de la même manière.

Nous observons que la branche auditive est très rapidement apprise. En moins de 500 itérations, elle a déjà atteint son point de convergence [1]. De plus, son point de convergence est un peu plus élevé que l'entropie du maître, ce qui suppose une erreur résiduelle. Cela s'explique par le fait que le réper-

1. Après la descente, nous observons une légère remontée, due à l'apprentissage de données sensorielles de faibles probabilités.

toire auditif $P(S \mid O_L)$ tente d'approximer avec des gaussiennes les distributions non gaussiennes du maître. Ces résultats sont en accord avec ce que nous avons observé dans le chapitre précédent.

La branche motrice, quant à elle, possède une entropie qui diminue très lentement (elle passe de 12 à 11,5 en 25 000 itérations). Cela vient du fait que durant cette étape, seul le modèle interne $P(S \mid M)$ est appris et que le répertoire moteur $P(M \mid O_S)$ est uniforme. Or, la branche motrice $P(S \mid O_S)$ nécessite l'apport des deux distributions pour pouvoir associer correctement des stimuli S à des objets O_S. Ici, l'apprentissage moteur ne permet que d'associer des stimuli S à des gestes moteurs M. Le fait que l'entropie diminue est seulement dû au fait que l'agent se spécialise sur les stimuli S du maître.

Passons maintenant à l'étape concernant l'apprentissage moteur. Dans cette étape, les apprentissages sensoriel et sensorimoteur sont considérés terminés. La branche auditive reste donc constante durant cette étape. Cela ne pose pas de problème puisqu'elle a atteint son point de convergence. Nous ne l'étudions pas davantage et nous nous concentrons sur la branche motrice. Celle-ci ne prend en compte finalement que l'évolution de l'entropie du répertoire moteur puisque celle du modèle interne est considérée terminée et la distribution reste constante durant cette étape. Nous avons schématisé les entropies des branches motrices de nos deux versions sur la même Fig. 6.2 afin de comparer leurs évolutions respectives.

Dans les deux versions, nous observons que l'apprentissage moteur se fait en deux phases. Dans une première phase, très rapide, ne durant que quelques centaines d'itérations, l'entropie de la branche motrice chute brusquement. Dans les deux cas, elle passe d'environ 11 à environ 7 ou 8. Ensuite, dans la seconde phase, l'entropie continue de diminuer mais ne semble pas avoir atteint totalement son point de convergence après 25 000 itérations. La brusque chute de l'entropie s'explique par le fait que l'agent commence à associer ses gestes moteurs à des objets. Ainsi, le répertoire moteur cesse d'être uniforme et sa variance diminue drastiquement pour se centrer globalement sur les objets du maître. La seconde phase s'explique par le fait que petit à petit, le répertoire moteur se spécialise sur les données du maître. Il diminue donc de plus en plus sa variance pour approximer le mieux possible celle du maître. Le fait qu'elle n'ait pas encore totalement convergé au bout de 25 000 itérations est en accord avec ce que nous avions observé de la branche motrice dans l'étude précédente.

Par ailleurs, nous remarquons que les deux apprentissages deviennent plus précis que l'apprentissage auditif. Ceci peut sembler contradictoire avec l'étude concernant la propriété « bande étroite/bande large » mise évidence au chapitre précédent mais s'explique par deux raisons. Premièrement, l'apprentissage sensorimoteur dure suffisamment longtemps pour que le modèle interne approxime de façon quasi parfaite la distribution $P(S^{Env} \mid M^{Env})$, qui transforme les gestes moteurs produits en son. Cela permet de diminuer considérablement les erreurs résiduelles d'apprentissage de la branche motrice. Même si cela semble peu réaliste, ceci est préférable dans cette étude, pour observer correctement les différences entre l'apprentissage moteur par imitation et l'apprentissage moteur par communication, sans être biaisé par des difficultés d'apprentissage du modèle interne. Deuxièmement, la branche auditive conserve, en revanche, une erreur résiduelle du fait de l'utilisation de distributions gaussiennes dans le répertoire auditif. De ce fait, elle est beaucoup moins précise que la branche motrice. Pour retrouver la propriété « bande étroite/bande large », il faudrait, d'une part, complexifier l'apprentissage de la distribution $P(S^{Env} \mid M^{Env})$ et, d'autre part, améliorer l'apprentissage de la distribution $P(S \mid O_L)$. Nous étudions en partie ceci dans le prochain chapitre.

Si les deux versions de l'apprentissage moteur montrent un comportement global commun, il subsiste quelques différences. La différence principale est le point de convergence. Dans l'apprentissage par imitation, après la première phase de l'apprentissage moteur, l'entropie est d'environ 7 tandis qu'elle est d'environ 8 pour l'apprentissage par communication. À ce stade, l'apprentissage par imitation semble donc plus précis que celui par communication. Par la suite, l'entropie de la branche motrice lors de l'apprentissage par imitation reste toujours plus faible et s'approche de l'entropie du maître. Un apprentissage plus poussé de 50 000 itérations montre, qu'en réalité, son entropie converge vers une valeur légèrement inférieure à celle du maître, ce qui laisse supposer que si l'apprentissage dure trop longtemps, il y a sur-apprentissage. De son côté, l'entropie de la branche motrice lors de l'apprentissage par communication se stabilise légèrement en dessous de l'entropie de la branche auditive. Une analyse plus poussée montre qu'elle conserve une erreur résiduelle lorsqu'elle atteint la convergence. Cela suggère donc que la branche motrice est moins semblable à celle du maître par apprentissage par communication.

En résumé, les deux apprentissages moteurs semblent globalement donner une branche motrice qui approxime très bien le maître. Nous pouvons donc désormais étudier les idiosyncrasies.

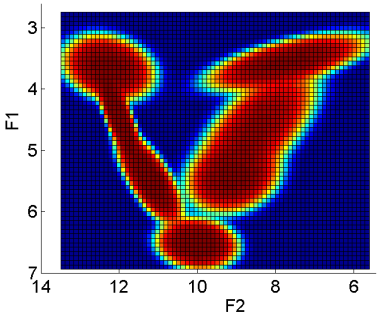
6.1.5.2 Etude des idiosyncrasies

Les idiosyncrasies acoustiques s'observent en analysant les valeurs des mi de nos agents apprenants. Elles sont calculées à partir de la méthode donnée en section 6.1.4.2 et représentées sur la Fig. 6.3 dans l'espace auditif S, soit F1/F2, disposé de manière à faire apparaître le triangle vocalique. Les trois extrémités de ce triangle [a i u], sont donc classiquement, selon cette configuration, disposées respectivement en bas au centre, en haut à gauche et en haut à droite de cette espace. Sur une même figure, les valeurs des mi de chaque agent sont représentées sous forme de points colorés (une couleur par voyelle).

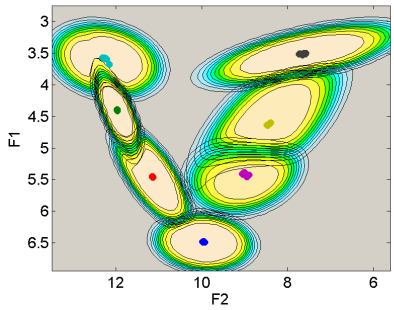
Afin de savoir si les valeurs des mi sont bien positionnées par rapport aux voyelles qu'elles représentent, nous avons, en plus, affiché sur ces même figures les courbes d'isoprobabilités de l'espace de chaque voyelle calculées à l'aide du classifieur auditif $P(O_L|\ S)$. Celui-ci est calculé par inversion du répertoire auditif $P(S\ |\ O_L)$, en ne conservant que les portions de l'espace les plus représentatives de chaque voyelle, c'est-à-dire celles dépassant une certaine probabilité. Le seuil de probabilité minimal choisi vaut $se = \frac{1}{4307}$, cette valeur étant la probabilité de la distribution uniforme de notre espace sensoriel discrétisé S (voir section 5.1.2.5 pour plus de détails sur le seuil). Le répertoire auditif $P(O_L|\ S)$ est jugé suffisamment bien appris pour servir de distribution de référence et nous permet ainsi d'avoir une information sur l'espace sensoriel dédié à chaque voyelle (voir Fig 6.3a).

Par ailleurs, afin de quantifier plus précisément la dispersion des valeurs des mi et donc, les idiosyncrasies, nous calculons, pour chaque apprentissage et pour chaque voyelle, les écarts-types en F1 et F2 des mi des douze simulations effectuées. Ils sont reportés Table 6.1.

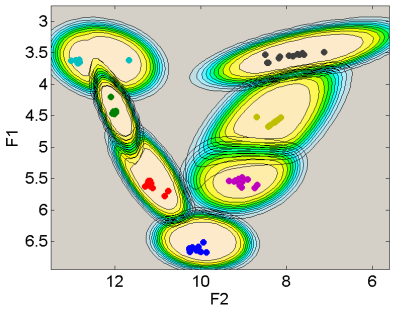
À partir de la Fig. 6.3 et de la Table 6.1, nous observons, premièrement, que les valeurs des mi lors de l'apprentissage par imitation sont toutes très proches voire identiques (dispersions quasi nulles). De plus, elles se rapprochent du centre de l'espace de chaque voyelle, ce qui semble montrer que les agents apprennent une distribution similaire à leur branche auditive et, de fait, similaire à celle du



(a) Représentation de $P(O_L \mid S)$



(b) App. imitation



(c) App. communication

FIGURE 6.3 – Illustration des représentations auditives $P(S)$. L'axe des abscisses correspond au formant F2 inversé et l'axe des ordonnées correspond au formant F1 inversé, en Barks. (a) Représentation du classifieur auditif $P(O_L \mid S)$. Les valeurs s'étendent du bleu foncé (faible probabilité) au rouge (forte probabilité). (b,c) Observation des mi dans l'espace auditif S : (b) en fin d'apprentissage par communication, (c) en fin d'apprentissage par imitation. Dans les panneaux (b) et (c) nous avons superposé les courbes d'isoprobabilité des distributions $P(O_L \mid S)$ pour chacune des 7 voyelles

maître. À terme, si l'apprentissage se prolonge, nous observons que les variations des valeurs des mi disparaissent totalement. Nous en déduisons que l'apprentissage par imitation n'est pas une méthode adaptée pour faire apparaître et conserver des idiosyncrasies acoustiques.

Deuxièmement, nous observons, qu'à l'inverse, les valeurs des mi lors de l'apprentissage par communication sont différentes entre les agents. Cependant, elles semblent suivre quelques règles particulières. Tout d'abord, elles sont toutes situées dans l'espace dédié à la voyelle à laquelle elles correspondent. Cela confirme qu'il y a un bon apprentissage même si les valeurs des mi sont différentes. Pour expliquer ceci, reprenons l'équation 6.3 :

$$P(M \mid [O_S = o]\,[C = 1]) \propto P(M \mid [O_S = o]) \sum_S P(S \mid M) P([O_L = o] \mid S)\,, \tag{6.7}$$

$$\propto P(M \mid [O_S = o]) P(M \mid [O_L = o])\,. \tag{6.8}$$

Lorsque l'apprentissage moteur par communication commence, les apprentissages sensoriel et sensorimoteur sont terminés, ce qui signifie que le modèle interne $P(S \mid M)$ et le classifieur auditif $P(O_L \mid S)$ sont déjà appris. Cela signifie donc non seulement que ces distributions sont constantes durant l'apprentissage moteur mais qu'elles sont adaptées pour trouver le bon objet o du maître. Comme

TABLE 6.1 – Ecart-types des valeurs de F1 et F2 pour chaque voyelle cible, pour l'apprentissage par imitation et par communication

Voyelles	App. imitation	App. communication
[a]	(0,0230 ; 0,0516)	(0,1546 ; 0,4443)
[i]	(0,1306 ; 0,1571)	(0,0762 ; 1,1695)
[u]	(0,0265 ; 0,1985)	(0,1849 ; 1,3839)
[e]	(0,0197 ; 0,0125)	(0,2390 ; 0,1158)
[ɛ]	(0,0198 ; 0,0283)	(0,2455 ; 0.5454)
[o]	(0,0497 ; 0,1025)	(0,1728 ; 0,4891)
[ɔ]	(0,0747 ; 0,1833)	(0,1822 ; 0,6316)

il est montré Fig. 6.3a, il existe plusieurs représentations sensorielles correspondant à un objet o selon $P(O_L \mid S)$: chaque voyelle correspond à une portion de l'espace sensoriel S et n'importe quelle représentation sensorielle de cette portion de l'espace peut donc être choisie pour reproduire la voyelle. Par suite, le modèle interne $P(S \mid M)$ permet de retrouver l'ensemble des représentations motrices M correspondant à ces représentations sensorielles S. Ainsi, la distribution $P(M \mid [O_L = o])$ fournit les représentations motrices pouvant reproduire l'ensemble des représentations sensorielles correspondant à la voyelle O_L. Un tirage sur cette distribution assure donc de choisir une représentation motrice correspondant à la bonne voyelle même si elle ne correspond pas forcément à la représentation sensorielle du maître. Cela explique pourquoi les valeurs des mi sont bien positionnées. De son côté, le répertoire moteur $P(M \mid O_S)$ permet de se focaliser sur des représentations motrices spécifiques. Au début de l'apprentissage, il est proche d'une uniforme, l'apprentissage s'effectue donc principalement avec la distribution $P(M \mid O_L)$ et permet de sélectionner n'importe quelle représentation motrice. Ensuite, au cours de l'apprentissage, il est peu à peu mis à jour ce qui permet à l'agent de se focaliser sur les représentations motrices préalablement choisies et donc de renforcer les idiosyncrasies.

Concernant la dispersion des valeurs des mi dans l'espace acoustique, nous observons qu'elle varie entre les voyelles et entre les agents. Si certaines, comme pour la voyelle [u], sont très dispersées d'un agent à l'autre et occupent une grande partie de l'espace de la voyelle, d'autres sont plus regroupées sur une portion de l'espace. Par exemple, la plupart des valeurs pour la voyelle [e] sont globalement situées au centre de l'espace, sauf pour un agent. De même, la plupart des valeurs pour la voyelle [a] est regroupée à l'extrémité basse de l'espace de la voyelle, sauf celle d'un agent. Après vérification, les valeurs des mi « isolées » par rapport aux autres ne correspondent pas toujours à celles du même agent. Un fait intéressant est que les valeurs des mi obtenues semblent avoir tendance à s'éloigner des lieux de confusion. Par exemple, une partie de l'espace de la voyelle [i] est confondue avec celle de la voyelle [e]. Dans ce cas, les valeurs des mi sont situées au deux extrémités de ce lieu de confusion. Cela a pour conséquence de les éloigner du centre de l'espace de la voyelle [i]. D'autres analyses seraient nécessaires pour vérifier que cette tendance se retrouve à chaque fois ou s'il s'agit simplement d'une coïncidence.

Ainsi, ces observations montrent que nous n'obtenons des variations entre les valeurs des mi et donc des idiosyncrasies que par l'apprentissage par communication. Cependant, une absence d'idiosyncrasies acoustiques ne signifie pas qu'il n'y a pas d'idiosyncrasies du tout. En effet, une même représentation acoustique peut être produite par plusieurs représentations motrices différentes, étant donnée la relation « *many-to-one* » existant entre gestes moteurs et résultats acoustiques. Des agents

produisant le même son peuvent donc le produire de différentes manières. De plus, comme le maître ne donne aucune information sur le geste moteur qu'il produit, chaque agent peut donc inférer le geste moteur qu'il préfère lors de son apprentissage moteur.

Afin de vérifier ces suppositions, nous observons les répertoires moteurs $P(M \mid O_S)$ en fin d'apprentissage par imitation, pour chaque agent. La tâche correspondante consiste à faire produire des voyelles à l'agent et à observer, dans l'espace moteur M, les configurations articulatoires utilisées pour produire cette voyelle. Nous observons que les répertoires moteurs sont tous différents, pour nos douze agents, c'est-à-dire que les configurations motrices d'une voyelle se situent, pour chaque agent, dans des portions de l'espace moteur différentes. À titre d'illustration, nous représentons Fig. 6.4 deux répertoires moteurs dans l'espace moteur M. Les points colorés correspondent aux points de plus hautes probabilités ($> 0{,}02$) pour chaque voyelle. Chaque couleur correspond à une voyelle différente. Bien entendu, même si nous ne l'illustrons pas, les répertoires moteurs de chaque agent sont également différents lors de l'apprentissage par communication. Ainsi, quel que soit l'apprentissage, nous obtenons des idiosyncrasies motrices et ce, même en l'absence d'idiosyncrasies acoustiques.

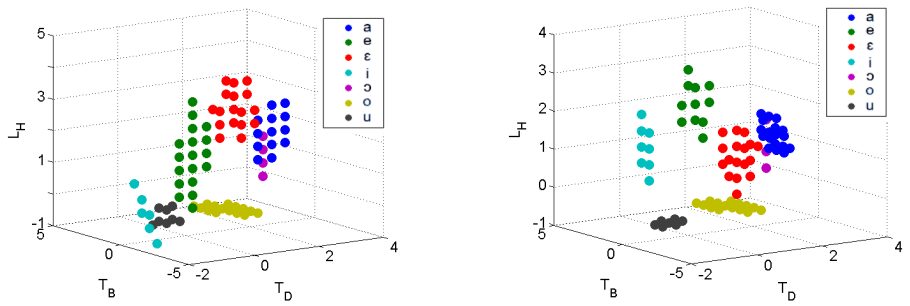


FIGURE 6.4 – Illustration des répertoires moteurs de deux agents lors de l'apprentissage par imitation

6.1.6 Discussion

6.1.6.1 Synthèse

Nous avons montré, dans un premier temps, qu'un apprentissage en deux phases, dans une version de COSMO vocalique, permet d'obtenir des scores d'entropies proches de celles du maître et, dans un second temps, que l'apprentissage par communication, contrairement à l'apprentissage par imitation, permet d'obtenir des idiosyncrasies acoustiques.

Comme l'agent apprenant est capable d'acquérir des distributions proches de celles de son environnement, nous en déduisons que le modèle COSMO-Voyelle et son apprentissage sont adaptés pour apprendre les distributions, et ce pour les deux types d'apprentissage moteur. Néanmoins, leurs comportements sont différents en ce qui concerne les idiosyncrasies.

Durant l'apprentissage par imitation, aucune idiosyncrasie acoustique n'apparaît en fin d'apprentissage. Cela s'explique par le fait que les représentations motrices sont essentiellement choisies de façon à ce qu'elle correspondent aux représentations sensorielles s perçues du fait de l'utilisation du

modèle interne $P(S \mid M)$ dans l'inférence $P(M \mid [O_S = o]\ [S = s])$ (voir Eq. 6.1). Ainsi, en fin d'apprentissage, la production de l'agent via la distribution $P(S^{Env} \mid O_S)$ est similaire à la distribution de production $P(S \mid O_S^{Maitre})$ du maître. Le répertoire moteur $P(M \mid O_S)$, quant à lui, a davantage un rôle d'ancrage et permet de se focaliser sur certaines représentations motrices particulières. Il n'a pas d'influence sur les idiosyncrasies acoustiques mais en présente en revanche sur les idiosyncrasies motrices. En effet, il permet de focaliser l'apprentissage sur une portion de l'espace moteur, différente entre les agents, ce qui fait apparaître des idiosyncrasies motrices.

Durant l'apprentissage par communication, des idiosyncrasies acoustiques apparaissent en fin d'apprentissage. Cela s'explique par le fait que l'agent ne cherche pas à inférer exactement les représentations sensorielles s perçues de son environnement mais cherche à inférer des représentations motrices correspondant au bon objet O. Ceci est notamment possible grâce à l'utilisation de la distribution $P(M \mid O_L)$ dans l'inférence $P(M \mid [O_S = o]\ [C = 1])$ (voir Eq. 6.3). Cette distribution, calculée à partir du modèle interne et du classifieur auditif, permet de trouver des représentations motrices correspondant aux objets o. Or, comme il existe plusieurs représentations sensorielles permettant de reproduire les objets, chaque agent choisit par tirage des représentations motrices permettant de reproduire une des représentations sensorielles possibles. Comme chaque agent réalise des tirages différents, les distributions de production de chaque agent $P(S^{Env} \mid O_S)$ sont différentes entre elles et différentes de celles du maître en fin d'apprentissage, mettant en évidence les idiosyncrasies acoustiques. Le répertoire moteur, quant à lui, lorsqu'il est mis à jour, permet d'ancrer les choix de chaque agent, ce qui renforce les idiosyncrasies acoustiques et les idiosyncrasies motrices.

6.1.6.2 Limites

Bien entendu, si ces résultats sont concluants, ils méritent d'être généralisés. Cette généralisation concerne notamment les catégories phonétiques telles que les consonnes et l'utilisation de représentations motrices et sensorielles permettant de traiter l'ensemble de ces catégories. La généralisation concerne également la complexification des simulations. Bien que la version COSMO-Voyelle soit plus complexe que COSMO-1D, plusieurs aspects nécessiteraient une complexification pour être plus réalistes.

Un point central est celui de l'existence de plusieurs maîtres durant l'apprentissage. Bien évidemment, la présence de maîtres différents peut produire tout naturellement des idiosyncrasies quel que soit le mode d'apprentissage mis en jeu. Une version stricte du mécanisme d'apprentissage par imitation proposé précédemment consisterait à apprendre la distribution moyenne des stimuli de l'environnement, c'est-à-dire fusionner l'ensemble des distributions sensorielles des différents maître et en apprendre la moyenne. Cependant, il n'y aurait toujours pas d'idiosyncrasies car le moyennage les ferait disparaître.

À la place du moyennage et de la simple imitation, on pourrait imaginer que l'apprentissage par imitation inclue un mécanisme de sélection permettant à l'agent apprenant de choisir entre ses différents maîtres un maître spécifique, ce qui permettrait alors de conserver des propriétés d'idiosyncrasie d'une génération à la suivante. Le fait d'observer subjectivement des ressemblances entre la voix d'un enfant et celle de l'un de ses parents semble confirmer la vraisemblance de cette possibilité. Néan-

moins, une étude récente de Rapin et al. (2017) suggère que les productions vocaliques entre frères présentent certes des corrélations significatives, mais globalement assez réduites. Ces corrélations, mesurées sur la production de voyelles orales du français, sur 10 paires de frères, locuteurs adultes du canadien français, ne sont significatives que pour 1/3 des voyelles étudiées, et ne permettaient d'expliquer que 25% de la variance des productions d'un sujet donné.

Par ailleurs, supposer l'imitation des idiosyncrasies des maîtres ne résout pas le problème de la génération d'idiosyncrasies. Celui-ci est simplement reporté puisqu'il reste à comprendre pourquoi les maîtres eux-mêmes présentent des idiosyncrasies. Ainsi, il nous semble que l'existence du mécanisme d'apprentissage par communication, seul capable de générer des idiosyncrasies et pas simplement d'en reproduire, est une hypothèse plausible. Cette forme d'apprentissage est également compatible avec un mécanisme de sélection du maître, ce qui peut permettre d'expliquer à la fois les influences des parents sur les enfants, et l'existence d'idiosyncrasies « libres », non produites par l'environnement.

L'utilisation du multi-maîtres n'est qu'une possible complexification de l'apprentissage. Néanmoins, comme l'illustre cet exemple, nous affirmons que notre étude peut être généralisée et que, même dans des situations plus complexes, l'apprentissage par communication fournit des idiosyncrasies plus adaptées et plus réalistes que l'apprentissage par imitation.

6.2 Corrélation des idiosyncrasies en perception et en production

---

Publication :

— Barnaud, M.-L., Diard, J., Bessière, P., et Schwartz, J.-L. (en révision). Computational simulations of perceptuo-motor idiosyncrasies support the involvement of motor knowledge in speech perception

---

Depuis le début de cette thèse, nous avons discuté le fait que les catégories phonétiques sont, entre autres, caractérisées par des représentations sensorielles et motrices. Dans le chapitre 2, nous avons notamment relaté que les représentations sensorielles et motrices utilisées durant des tâches de perception semblent, en plus, connectées à celles utilisées dans des tâches de production. Cependant, nous ne savons pas réellement quelle est la nature de cette connexion. Un des moyens utilisés pour examiner ce lien se fait à travers les idiosyncrasies. Dans cette étude, nous réutilisons les idiosyncrasies en production, observées dans la section précédente, pour analyser d'une part, le lien entre les représentations en perception et en production et, d'autre part, la nature des représentations en perception.

Dans un premier temps, nous faisons l'hypothèse avec le modèle COSMO que les représentations sensorielles et motrices sont utilisées de la même manière en perception et en production. Pour vérifier cette supposition, nous tentons de reproduire avec COSMO les résultats de Ménard et Schwartz (2014), étudiant les corrélations entre les idiosyncrasies en perception et en production. Dans un second temps, nous comparons une production motrice avec une perception auditive, motrice

et perceptuo-motrice, reflétant les trois familles de théories de la perception. Les résultats montrent que la corrélation entre une production motrice et une perception motrice ou perceptuo-motrice est similaire à celle observée expérimentalement par Ménard et Schwartz (2014). Nous en déduisons que les représentations motrices sont non seulement importantes en perception mais qu'elles pourraient être de même nature que celles utilisées en production.

6.2.1 Hypothèse de l'étude

Comme l'a montré le débat entre les théories de la perception (voir section 3.1.1), la nature des représentations des unités phonétiques reste mal élucidée. Les partisans des théories auditives affirment que les catégories phonétiques sont principalement caractérisées par des représentations auditives tandis que les partisans des théories motrices supposent que les catégories phonétiques sont principalement caractérisées par des représentations motrices. Grâce aux avancées des neurosciences, il est désormais supposé que les catégories phonétiques semblent être en réalité des unités perceptuo-motrices, caractérisées par des représentations à la fois auditives et motrices. Néanmoins, le rôle et l'importance respectifs des représentations durant la perception reste encore à préciser.

Par ailleurs, plusieurs études ont montré qu'il existe un lien entre la perception et la production et que la perturbation de l'une pouvait influencer l'autre (voir section 3.1.2). Néanmoins, la nature exacte de ce lien reste encore mystérieuse. En particulier, il reste encore beaucoup à découvrir sur la manière dont sont liées les catégories phonétiques en perception et en production.

Une des premières questions consiste à se demander si les représentations sensorielles et motrices utilisées en perception et en production sont connectées. Pour notre modèle, nous faisons cette hypothèse. En effet, de par sa construction, nous supposons qu'avec le modèle COSMO tout traitement du langage s'effectue avec les mêmes représentations sensorielles S et motrices M. Dans ce cadre, chaque tâche effectuée avec le modèle COSMO correspond à une « question », calculée par inférence en utilisant les distributions du modèle. Comme il n'y a qu'un unique répertoire moteur, qu'un unique répertoire auditif et qu'un unique modèle interne dans le modèle COSMO, cela suppose qu'ils sont tous trois utilisés aussi bien dans des tâches de perception que dans des tâches de production.

Ces travaux reposent donc sur deux hypothèses. Premièrement, nous supposons que les mêmes représentations peuvent être utilisées pour les deux tâches. En nous intéressant plus spécifiquement aux représentations motrices, nous supposons notamment que les mêmes représentations motrices sont utilisées pour réaliser des tâches de perception et de production. Deuxièmement, en accord avec les théories perceptuo-motrices, nous supposons que les représentations motrices sont nécessaires dans le processus de perception.

Pour tester ces hypothèses, nous avons besoin, d'une part, d'étudier dans un même contexte des tâches de production et de perception pour évaluer notre première hypothèse et, d'autre part, de comparer les théories de la perception pour évaluer notre seconde hypothèse. Pour cela, nous nous servons, une nouvelle fois, des idiosyncrasies.

Plus précisément, nous nous basons sur la reproduction, avec COSMO-V, de l'étude de Ménard et Schwartz (2014), étudiant la corrélation des idiosyncrasies voyelles perçues et produites. Après avoir

décrit, d'une part, l'étude de Ménard et Schwartz (2014) et, d'autre part, les spécificités du modèle utilisé, l'analyse se fait en trois temps. Dans un premier temps, nous réutilisons les résultats obtenus précédemment, pour effectuer une tâche de production et calculer les idiosyncrasies acoustiques de production. Dans un deuxième temps, nous développons les trois tâches de perception relatives aux trois familles de théories de la perception et calculons les idiosyncrasies acoustiques de perception. Dans un troisième temps, nous calculons, dans notre modèle, la corrélation entre les idiosyncrasies de production et de perception et la comparons avec celle de l'étude de Ménard et Schwartz (2014).

6.2.2 Description de l'étude de Ménard et Schwartz (2014)

Cette section décrit, de manière synthétique, les objectifs et la méthode de l'étude de Ménard et Schwartz (2014). Pour le reste de ce chapitre, elle est désormais notée M&S.

En résumé, cette étude a pour premier objectif d'étudier l'organisation des voyelles perçues du français dans l'espace acoustique. Le second objectif est de comparer l'organisation des voyelles perçues et produites, en s'appuyant sur une étude précédente dans laquelle Ménard et al. (2008) ont analysé, avec les mêmes participants, l'organisation des voyelles produites du français. Ils reportent deux résultats. Premièrement, pour chaque sujet étudié séparément, les voyelles perçues et produites de même aperture ont une valeur formantique en F1 similaire, indépendante de l'arrondissement et du lieu d'articulation des voyelles. Deuxièmement, la distance en F1 entre deux voyelles perçues d'un sujet est corrélée avec la distance en F1 de ces deux mêmes voyelles produites. Les auteurs interprètent cela comme la trace de l'existence d'un lien entre les catégories phonémiques perçues et produites dans le cerveau.

Plus précisément, pour chaque tâche de production et de perception, ils étudient dix voyelles orales du français : [i y u e ø o ɛ œ ɔ a]. Elles sont respectivement produites et perçues par douze participants âgés entre 4 et 39 ans répartis dans trois groupes d'âge (environ 4 ans, environ 8 ans et adultes). Tous les participants sont natifs du français et n'ont aucun problème d'audition ou d'articulation. Brièvement, la tâche de production, décrite plus précisément par Ménard et al. (2008), consiste à répéter dix fois les dix voyelles isolément et de manière prolongée. La tâche de perception, décrite plus précisément par M&S, consiste à écouter une centaine de stimuli synthétiques correspondant chacun à une des dix voyelles et à identifier sur un écran la voyelle perçue. Les stimuli synthétiques ont été générés avec VLAM : ils correspondent à différents conduits vocaux et les stimuli de chaque voyelle possèdent une large variabilité sur l'espace formantique F1/F2. Définis ainsi, ces stimuli contiennent donc de la variabilité inter-locuteur et intra-locuteur.

Durant la phase d'analyse, M&S ont récupéré les valeurs formantiques F1 et F2, en Hz, des stimuli auditifs produits et perçus et les ont convertis en Barks (voir Eq. 5.8). Afin de réduire la variabilité intra-locuteur, ils ont calculé pour chaque voyelle, produite d'une part et perçue d'autre part, la valeur moyenne de F1 et de F2 des stimuli auditifs enregistrés. Ils se sont ensuite focalisés sur l'organisation des voyelles sur la dimension F1 pour observer les idiosyncrasies. Dans un premier temps, les voyelles ont été regroupées selon quatre degrés d'aperture : haut ([i u y]), mi-haut ([e o ø]), mi-bas ([ɛ œ ɔ]) et bas ([a]). Dans un second temps, pour chaque participant, ils ont calculé la valeur moyenne m_1 des voyelles hautes [i u y] et nommé m_4 la valeur de la voyelle basse [a] (voir Fig. 6.5). Dans un troisième

temps, afin de réduire la variabilité inter-locuteur et pour pouvoir comparer les participants entre eux, ils ont normalisé la dimension F1 de chaque participant par une transformation affine dans l'échelle des Barks, de telle manière à ce que la valeur normalisée de F1 pour la moyenne m_1 soit égale à 0 et celle pour la moyenne m_4 soit égale à 1. Tout ce processus a été réalisé séparément en perception et en production. Par la suite, ils n'analysent plus que les voyelles mi-hautes et mi-basses puisque les autres servent pour la normalisation. Pour chacune d'elles, ils peuvent calculer la valeur normalisée en F1. En notant $m_{voyelle}$ la moyenne de la voyelle recherchée, l'équation correspondante est :

$$F1_{Normalise} = \frac{m_{voyelle} - m_1}{m_4 - m_1} \, . \tag{6.9}$$

Ceci est également schématisé Fig. 6.5.

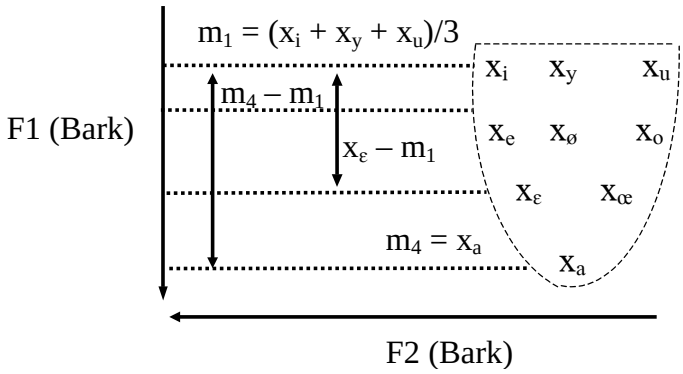


FIGURE 6.5 – Schéma illustrant le calcul de la distance relative dans F1. Repris de Ménard et Schwartz (2014)

L'analyse de ces valeurs relatives en F1 montre l'apparition d'idiosyncrasies aussi bien en perception qu'en production : chaque participant a des valeurs normalisées de F1 en production et en perception qui lui sont propres. En effectuant, pour chaque voyelle, une régression linéaire sur les couples de valeurs obtenues en perception et en production pour chaque sujet, les auteurs obtiennent une corrélation variant entre 0.6 et 0.8. Ce résultat met fortement en évidence le lien entre les idiosyncrasies de production et de perception. Par ailleurs, M&S n'obtiennent aucun résultat significatif sur leurs groupes d'âge, ce qui laisse supposer que la corrélation entre idiosyncrasies de production et idiosyncrasies de perception apparaît dès le plus jeune âge testé.

6.2.3 Implémentation du modèle et de l'apprentissage

Dans cette étude, nous réutilisons le modèle COSMO-V et les douze mêmes agents de la version d'apprentissage par communication tels qu'il ont été définis dans l'étude précédente. Pour rappel, les variables catégorielles des objets O_L et O_S sont composées de sept voyelles [a i u e ɛ o ɔ], les représentations motrices M sont définies par trois paramètres articulatoires TB, TD, LH, issus du modèle articulatoire VLAM, et les représentations auditives S sont définies par les deux formants F1 et F2, en Barks. En ce qui concerne les distributions, les répertoires auditif et moteur ainsi que le modèle interne sont des ensembles de gaussiennes multidimensionnelles, le prior des objets moteurs

est uniforme et le système de cohérence est basé sur une distribution de Dirac (voir section 6.1.2 pour plus de détails).

Concernant l'apprentissage, celui-ci s'effectue, une nouvelle fois, en trois étapes successives : un apprentissage auditif, un apprentissage sensorimoteur et un apprentissage moteur. Les deux premières étapes s'effectuent de la manière décrite dans le chapitre 3. En ce qui concerne l'apprentissage moteur, nous utilisons l'apprentissage par communication, qui a été défini précédemment comme la méthode la plus appropriée pour que les idiosyncrasies acoustiques apparaissent (voir section 6.1.3 pour plus de détail sur l'apprentissage).

6.2.4 Analyse du modèle

Contrairement aux études précédentes, nous n'analysons ni la qualité ni l'évolution de l'apprentissage puisque cela a déjà été fait dans l'étude précédente. Nous nous focalisons sur ce dont nous avons besoin pour comparer notre modèle avec l'étude de M&S : une tâche de production et une tâche de perception. Ces deux tâches sont évaluées une fois l'apprentissage des douze agents terminés.

6.2.4.1 Tâche de production

Lors de la tâche de production, nous souhaitons connaître, pour agent apprenant, le signal auditif produit en moyenne pour chaque voyelle. D'un point de vue expérimental, la tâche de production consisterait à faire produire des voyelles aux agents apprenants et à enregistrer le signal auditif résultant. Ensuite, la recherche du signal auditif moyen équivaudrait à calculer, pour chaque voyelle, la moyenne des signaux produits.

Dans COSMO, nous effectuons cette tâche grâce à la distribution $P(S^{Env} \mid O_S)$. Déjà décrite et utilisée précédemment pour comparer l'apprentissage par imitation et l'apprentissage par communication, nous la reprenons telle quelle dans cette étude. De la même manière, pour obtenir le signal auditif moyen, nous calculons, comme précédemment, les valeurs des moyennes individuelles (mi) de nos agents (voir section 6.1.4.2 pour plus de détails).

6.2.4.2 Tâche de perception

Lors de la tâche de perception, nous souhaitons définir, pour chaque agent apprenant, la moyenne des signaux auditifs perçus correspondant à chaque voyelle. Dit autrement, nous souhaitons déterminer l'ensemble des signaux qui sont perçus comme une même voyelle et calculer ensuite leur moyenne. D'un point de vue expérimental, la tâche de perception consisterait à faire catégoriser un ensemble de sons aux agents apprenants. Par la suite, la recherche du signal moyen consisterait à organiser les signaux testés par voyelle et à calculer le signal moyen pour chacune d'elles. Dans COSMO, le plus simple pour obtenir le signal perçu moyen pour chaque voyelle est de calculer la moyenne de la distribution $P(S \mid O)$ donnant la probabilité des signaux S pour chaque objet O.

Nous souhaitons, en plus, étudier la tâche de perception selon les trois familles des théories de la perception. En nous basant sur une tâche de catégorisation réalisée à l'aide d'un décodeur $P(O \mid S)$, nous avons vu, dans la section 4.2.4 du chapitre 3, que la nature de la perception dépend essentiellement de l'objet O considéré. En effet, choisir l'objet O_L permet d'implémenter une tâche de catégorisation selon les théories auditives via le décodeur auditif $P(O_L \mid S)$, choisir l'objet O_S permet d'implémenter une tâche de catégorisation selon les théories motrices via le décodeur moteur $P(O_S \mid S)$ et choisir un objet commun O_L et O_S, par l'activation de la variable de cohérence C, permet d'implémenter une tâche de catégorisation selon les théories auditives via le décodeur perceptuo-moteur $P(O_S \mid S\,[C=1])$.

Pour caractériser ces trois familles de théories, nous utilisons la distribution $P(S \mid O)$ qui est proportionnelle à la distribution $P(O \mid S)$, puisque les priors sur les objets O sont uniformes. Ainsi, les trois formules de $P(S \mid O)$ se définissent par :

$$\begin{aligned}
P(S \mid O_L) &\propto P(O_L \mid S) \text{ (voir Eq. 4.8)}\,, && (6.10)\\
P(S \mid O_S) &\propto P(O_S \mid S)\\
&\propto \sum_M P(M \mid O_S) P(S \mid M) \text{ ,(voir Eq. 4.9)}\,, && (6.11)\\
P(S \mid O_L\,[C=1]) &\propto P(O_L \mid S) P(O_S \mid S)\\
&\propto P(O_L \mid S) \sum_M P(M \mid O_S)\; P(S \mid M) \text{ (voir Eq. 4.10)}\\
&\propto P(S \mid O_L)\; P(S \mid O_S)\,. && (6.12)
\end{aligned}$$

Par la suite, $P(S \mid O_L)$ est donc la distribution utilisée pour les théories auditives, $P(S \mid O_S)$ est la distribution utilisée pour les théories motrices et $P(S \mid O_L\,[C=1])$ est la distribution utilisée pour les théories perceptuo-motrices. Pour chacune de ces distributions, nous calculons le stimulus moyen, pour chaque voyelle o. Nous obtenons ainsi les valeurs des moyennes individuelles mi en perception. Pour mieux les distinguer, nous appelons mi_A les valeurs des mi issues de la distribution auditive $P(S \mid O_L)$, mi_M celles issues de la distribution motrice $P(S \mid O_S)$ et mi_{PM} celles issues de la distribution perceptuo-motrice $P(S \mid O_L\,[C=1])$.

6.2.4.3 Corrélation production/perception

Maintenant que nos deux tâches et les valeurs de leurs mi respectives sont définies, nous calculons les corrélations entre ces valeurs. Pour cela, nous nous servons d'une procédure similaire à celle de M&S :

1. Nous récupérons la valeur en F1 des valeurs des mi.
2. Nous regroupons les voyelles selon les quatre degrés d'aperture précédemment définis : les voyelles hautes ([i u]), les mi-hautes ([e o]), mi-basses ([ɛ ɔ]) et basse ([a]).
3. Nous calculons la moyenne m_1 pour les voyelles hautes et nous utilisons la valeur de la mi de la voyelle [a] comme valeur m_4.
4. Nous utilisons la formule de la valeur normalisée de F1 (voir Eq. 6.9) pour calculer la valeur

normalisée des voyelles mi-hautes et mi-basses [2].

Nous obtenons ainsi des valeurs normalisées de F1 pour les quatre voyelles mi-hautes et mi-basses [e o ɛ ɔ] pour tous nos agents et ce, pour les valeurs des mi en perception d'une part (pour chacune des trois familles de théories perceptives) et en production d'autre part. Nous avons ainsi pu calculer la corrélation entre les idiosyncrasies de perception et les idiosyncrasies de production, pour chacune des trois théories perceptives.

6.2.5 Résultats

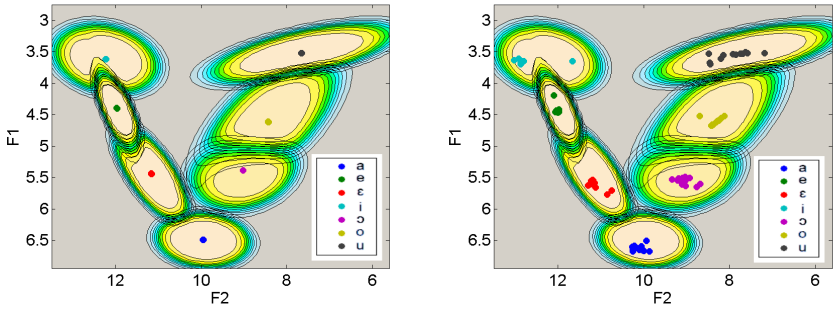
6.2.5.1 Idiosyncrasies en perception et en production

Nous commençons par étudier les valeurs des mi issues des tâches de perception (voir Fig. 6.6). Celle-ci sont construites exactement sur le même principe que dans l'étude précédente : elles sont affichées dans l'espace auditif F1/F2 organisé de manière à voir apparaître le triangle vocalique. En fond, la distribution $P(O_L \mid S)$ est représentée sous forme de courbes d'isoprobabilités afin d'afficher les contours associés à chaque voyelle et vérifier que les valeurs des mi obtenues sont correctement placées dans la portion de l'espace de la voyelle à laquelle elles correspondent. Pour rappel et pour comparaison, les valeurs des mi pour la tâche de production sont également affichées. La Fig 6.6d est donc identique à la Fig. 6.3c.

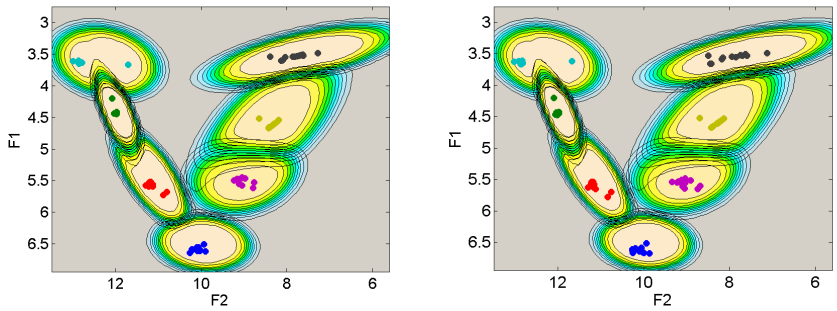
De manière générale, nous observons que les agents ont des valeurs de mi différentes dans les tâches de perception motrice et perceptuo-motrice mais pas dans la tâche de perception auditive. En effet, dans ce dernier cas, les valeurs des mi_A sont les mêmes pour tous les agents et sont situées au centre de chaque voyelle. Ce résultat n'est pas surprenant : le répertoire auditif $P(S \mid O_L)$ est, pour chaque agent, une approximation gaussienne de la distribution de production $P(S \mid O_S^{Maitre})$ du maître. Comme chaque agent réalise son apprentissage sensoriel à l'aide du même maître, les agents approximent tous la même distribution $P(S \mid O_S^{Maitre})$. Par conséquent, les valeurs de leurs mi sont toutes identiques et sont, par ailleurs, toutes situées au centre de l'espace de chaque voyelle. Il n'apparait donc aucune idiosyncrasie acoustique dans cette situation. Pour faire le lien avec l'étude précédente, l'apprentissage sensoriel agit donc comme l'apprentissage par imitation.

Au contraire, les valeurs des mi_M sont différentes entre les agents. La dispersion moyenne de ces valeurs est de 0,1837 Barks en F1 et de 0,6800 Barks en F2. Par ailleurs, nous observons qu'elles sont très similaires aux valeurs des mi en production dont la dispersion moyenne est de 0,1793 Barks en F1 et de 0,6828 Barks en F2 (voir Table 6.1 pour plus de détails sur la dispersion). Si nous reprenons les équations correspondantes (voir Eq. 6.6 et Eq.6.11), nous analysons que les valeurs des mi_M sont calculées à l'aide du répertoire moteur $P(M \mid O_S)$ et du modèle interne $P(S \mid M)$ tandis que les valeurs des mi en production sont calculées à partir du répertoire moteur $P(M \mid O_S)$ et de la transformation d'un geste en son $P(S^{Env} \mid M^{Env})$. Comme il s'agit des mêmes agents, les répertoires moteurs, dans ces deux équations, sont les mêmes. Si les valeurs des mi_M et celles des mi en production sont similaires, cela signifie donc que le modèle interne $P(S \mid M)$ est similaire à la distribution

2. Pour rappel, du fait de la formule utilisée, la valeur normalisée des voyelles hautes vaut 0 et celle de la voyelle basse vaut 1.



(a) mi_A (b) mi_M



(c) mi_{PM} (d) mi en production

FIGURE 6.6 – Illustration des valeurs des mi en fin d'apprentissage pour les tâches de perception auditive (a), motrice (b) et perceptuo-motrice (c) et pour la tâche de production (d)

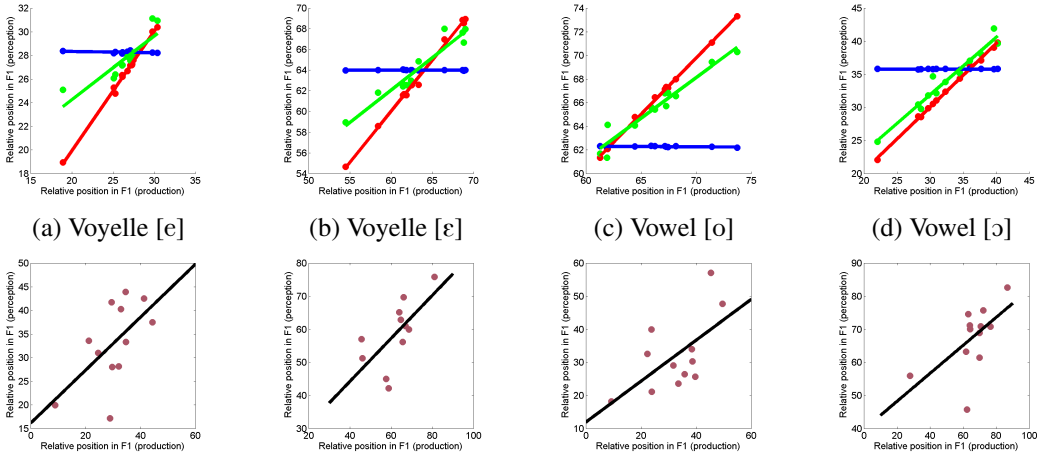
$P(S^{Env} \mid M^{Env})$. Ainsi, nous en déduisons que lors de l'apprentissage sensorimoteur, le modèle interne $P(S \mid M)$ apprend à approximer de façon quasi parfaite la distribution $P(S^{Env} \mid M^{Env})$. C'est la conclusion à laquelle nous sommes déjà parvenus lors de l'étude précédente (voir section 6.1.5.1). Il y a donc des idiosyncrasies en perception selon les théories motrices et celles-ci sont quasiment identiques à celles en production.

Pour finir, les valeurs des mi_{PM} sont également différentes entre agents. Leur localisation est similaire à celles des mi_M (et donc, par suite, à celles des mi de production). Elles sont cependant moins dispersées puisqu'elles ont une dispersion moyenne de 0,1522 Barks en F1 et de 0,5609 Barks en F2. Puisque la distribution utilisée pour réaliser la perception perceptuo-motrice est une fusion entre la distribution de perception auditive et la distribution de perception motrice, les valeurs des mi_{PM} sont un mélange entre les valeurs des mi_A pour lesquelles la dispersion est quasiment nulle et celles des mi_M. Il y a donc des idiosyncrasies également en perception selon les théories perceptuo-motrices, mais celles-ci diffèrent des idiosyncrasies en perception motrice et des idiosyncrasies en production.

6.2.5.2 Comparaison des deux études

Maintenant que nous avons observé qu'il apparaît des idiosyncrasies, nous pouvons étudier la corrélation entre les idiosyncrasies obtenues en perception et en production et surtout les comparer

avec celles obtenues dans l'étude de M&S. Nous affichons, sur la Fig. 6.7, les valeurs des mi pour les trois tâches de perception, comparées à celles de la tâche de production, ainsi que les droites de régression correspondantes. Nous reportons pour référence, les points et droites obtenus par M&S. Nous reportons également les pentes de régression dans la Table 6.2 et les coefficients de corrélation dans la Table 6.3.



(a) Voyelle [e] (b) Voyelle [ɛ] (c) Vowel [o] (d) Vowel [ɔ]

(e) Voyelle [e] M&S (f) Voyelle [ɛ] M&S (g) Vowel [o] M&S (h) Vowel [ɔ] M&S

FIGURE 6.7 – Idiosyncrasies couplées des tâches de perception/production. Haut : Résultats des simulations. Points et courbes de régression linaire pour les idiosyncrasies couplées des prédictions auditive (en bleu), motrice (en rouge) et perceptuo-motrice (en vert). Bas : Points expérimentaux et régressions linéaires issues des données de M&S. Chaque colonne correspond à l'une des quatre voyelles étudiées, dans l'ordre : [e ɛ o ɔ]

D'une manière générale, nous observons que les valeurs des mi couplées obtenues avec le modèle COSMO ont un comportement global similaire entre chaque voyelle. En étudiant plus spécifiquement la table des pentes de régression, nous observons que la pente est quasiment nulle pour la perception auditive, aux alentours de 1 pour la perception motrice et variant entre 0,54 et 0,87 pour la perception perceptuo-motrice, conformément à nos analyses précédentes.

Comparons maintenant nos résultats avec ceux de M&S. Du fait que nos simulations n'ont pas pour objectif de reproduire exactement les données obtenues par M&S, mais plutôt leur comportement global, notre comparaison se veut davantage qualitative que quantitative. Nous réalisons trois observations importantes.

Premièrement, les valeurs des mi couplées dans nos simulations sont moins dispersées que celles de M&S. En effet, les valeurs normalisées en F1 des valeurs des mi de production varient entre 10 et 20% tandis que celles de M&S varient de plus de 50%. De même, en perception, les valeurs normalisées en F1 des valeurs des mi de nos simulations varient au maximum de 20% tandis que celles de M&S varient de plus de 40%. Nous reviendrons sur ce point en discussion.

Deuxièmement, les pentes de régression obtenues par M&S varient entre 0,4 et 0,6, ce qui est significativement supérieur à 0. Ce résultat est incompatible avec les pentes des idiosyncrasies couplées de perception auditive dont les pentes sont nulles. Ce résultat semble, en revanche, compatible avec les

TABLE 6.2 – Pentes de régression entre les données en production et en perception pour chaque voyelle [e o ɛ ɔ] de l'étude de M&S comparées aux pentes de régression entre les tâches de perception et de production dans COSMO pour les systèmes de perception auditif, moteur et perceptuo-moteur (PM). Nous affichons les niveaux de significativité pour les pentes des données de M&S selon le format suivant : (supérieur à 0, inférieur à 1). Les niveaux de significativité sont notés $*$ pour $p < 0.05$, o pour $p \geq 0.05$

Voyelles	M&S	Auditif	Moteur	PM
[e]	0.5612 $(*,o)$	-0.0103	1.0097	0.5395
[ɛ]	0.6505 $(*,o)$	-0.0009	0.9915	0.6290
[o]	0.6195 $(*,o)$	-0.0044	0.9527	0.6898
[ɔ]	0.4236 $(*,*)$	0.0000	0.9619	0.8685

TABLE 6.3 – Coefficients de corrélation entre les données en production et en perception pour chaque voyelle [e o ɛ ɔ] de l'étude de M&S, associées aux coefficients de corrélation entre les données en perception et en production des simulations dans COSMO pour les systèmes de perception moteur et perceptuo-moteur (PM)

Vowels	M&S	Motor	PM
[e]	0.5965	0.9976	0.8728
[ɛ]	0.6554	0.9974	0.9573
[o]	0.6153	0.9989	0.9651
[ɔ]	0.6024	0.9987	0.9757

pentes des idiosyncrasies couplées de perception motrice, dont les valeurs sont proches de 1, puisque les pentes obtenues par M&S ne sont pas significativement inférieure à 1 (sauf pour la voyelle [ɔ]). Néanmoins, les pentes des idiosyncrasies couplées de perception perceptuo-motrice semblent plus similaires à celles de M&S.

Troisièmement, les coefficients de corrélation obtenus par M&S varient autour de 0,60 tandis que les coefficients obtenus pour nos simulations dépassent globalement 0,90 pour les prédictions motrices et perceptuo-motrices. Cela vient du fait que nos données de simulations sont moins bruitées que les données expérimentales.

6.2.6 Discussion

6.2.6.1 Synthèse

Nous avons, dans un premier temps, observé des idiosyncrasies auditives dans chacune de nos tâches sauf pour les valeurs des mi_A. Dans un second temps, en comparant nos données avec celles de M&S, nous avons remarqué que la corrélation obtenue était non nulle et en accord avec celle trouvée par M&S lorsque la tâche de perception s'effectuait avec un système de prédiction moteur ou perceptuo-moteur.

Ces résultats suggèrent que la composante motrice semble nécessaire en perception. Pour aller

plus loin, nous avons observé que les corrélations entre les données en perception et en production obtenues avec le système de prédiction perceptuo-moteur sont les plus proches de celles obtenues par M&S, ce qui suggère que la perception perceptuo-motrice est plus pertinente que les deux autres.

6.2.6.2 Interprétation

L'absence de variabilité dans les valeurs des mi_A peut surprendre. Comme dit précédemment, cela s'explique par le fait que le répertoire auditif $P(S \mid O_L)$ sur lesquelles elles s'appuient est similaire entre tous les agents puisqu'il approximent pour chaque agent le même environnement. En cela, les valeurs des mi_A se rapprochent de celles des mi obtenues par l'apprentissage par imitation dans la section précédente. En supposant que l'environnement diffère entre les agents, nous pouvons nous attendre à ce que les valeurs des mi_A varient elles aussi et donc que des idiosyncrasies apparaissent. À l'inverse, suite à l'apprentissage par communication les valeurs des mi_M diffèrent et ce, même avec un environnement identique.

À ce stade, l'interprétation à laquelle nous arrivons est que les représentations fournies par la voie de décodage auditive, ici correspondant aux valeurs des mi_A, sont issues des données de l'environnement tandis que les représentations issues de la voie de décodage moteur, correspondant aux valeurs des mi_M, sont basées sur des composantes endogènes, des tirages propres à chaque agent. Au travers des idiosyncrasies, ces deux comportements rendent compte de deux nouvelles différences entre l'apprentissage sensoriel et l'apprentissage moteur. Ainsi, nous proposons que l'apprentissage sensoriel est avant tout un apprentissage exogène centré sur les données de l'environnement tandis que l'apprentissage moteur est avant tout un apprentissage endogène relatif à chaque agent. Cela pourrait être une nouvelle explication de la complémentarité entre les représentations sensorielles et motrices que nous avons évoquée au chapitre précédent.

6.3 Conclusion

Ce chapitre était dédié à l'analyse des idiosyncrasies. Dans une première étude, nous avons analysé leur apparition et avons remarqué qu'elles dépendent principalement de l'apprentissage moteur, dans sa version dite « d'apprentissage communicatif ». Dans une seconde étude, nous avons comparé la corrélation entre les idiosyncrasies en perception et en production avec des données expérimentales et avons obtenu une similarité lorsque les représentations motrices sont utilisées en perception. Ces deux résultats accentuent les différences entre les représentations sensorielles et motrices et confirment l'importance des représentations motrices en perception. Ces études restent néanmoins centrées sur les voyelles. Une nouvelle étape de complexification est de s'intéresser à des unités distinctives plus variées commes des consonnes ou des syllabes. C'est le sujet du prochain chapitre.

CHAPITRE 7

La composition interne des unités phonétiques et COSMO SylPhon

Publication :

— Barnaud, M.-L., Diard, J., Bessière, P., et Schwartz, J.-L. (2017a). Assessing phonological learning in COSMO, a Bayesian model of speech communication. In The 7th Joint IEEE International Conference on Developmental and Learning and on Epigenetic Robotics (ICDL-Epirob 2017)

Dans les deux précédents chapitres, les unités distinctives utilisées ont évolué d'unités discrètes abstraites à des phonèmes vocaliques. Cependant, l'utilisation et l'apparition des phonèmes, surtout durant l'apprentissage, sont souvent remises en question : certains chercheurs affirment notamment que l'unité de base apparaissant lors du développement n'est pas le phonème mais la syllabe (voir section 3.2.3).

Dans ce chapitre, nous nous focalisons davantage sur l'étude de la structure cognitive des unités et étudions en particulier l'acquisition des syllabes et des phonèmes. Dans ce contexte, nous avons deux objectifs : nous souhaitons vérifier, d'une part, qu'un agent apprenant COSMO est capable d'apprendre des unités distinctives et, d'autre part, qu'il est capable de communiquer avec son maître. Cet apprentissage s'effectue sous deux conditions : 1) l'agent n'a pas de connaissances préalables sur le nombre et le contenu de ses catégories et 2) l'invariant consonantique n'est pas explicitement implémenté.

Ces deux conditions sont les challenges de ce chapitre et diffèrent des précédentes études effectuées. Après les avoir précisées, nous nous intéressons à l'implémentation du modèle. En plus de la prise en compte des deux conditions, l'utilisation des unités distinctives de différents types, syllabiques et phonémiques, nécessite une extension du modèle. Nous présentons, à cette occasion, une nouvelle version du modèle COSMO, nommée COSMO SylPhon (Syl pour syllabe et Phon pour phonème). Par la suite, nous détaillons les étapes d'apprentissage et leurs résultats respectifs. Pour finir, nous analysons la capacité de communication de ce nouveau modèle.

7.1 Hypothèses de l'étude

Implémenter une version syllabique du modèle COSMO n'est pas novateur. Cela a déjà été réalisé par Laurent et al. (2017) dans une extension du modèle nommée COSMO-S. Ce modèle montre notamment que la complémentarité entre les représentations auditives et motrices s'étend au contenu cognitif des unités distinctives puisque les voyelles semblent mieux discriminées par les représentations auditives tandis que les consonnes semblent mieux discriminées par les représentations motrices.

Le modèle COSMO SylPhon, que nous proposons dans cette étude, s'inspire de ce modèle pour sa partie syllabique et l'étend, en plus, à l'apprentissage des catégories phonémiques. Par ailleurs, son implémentation et son apprentissage diffèrent du modèle COSMO-S sur deux points cruciaux. Premièrement, l'agent COSMO apprenant ne possède plus de connaissances préalables sur le nombre exact d'unités distinctives dans sa langue et n'a plus d'accès direct aux catégories phonétiques du maître. Deuxièmement, dans COSMO SylPhon, les invariants phonémiques, notamment l'invariant consonantique, ne sont plus implémentés de façon aussi explicite que dans COSMO-S. Ceci nécessite, en conséquence, de contraindre l'apprentissage.

7.1.1 Condition 1 : Des connaissances limitées sur les unités distinctives

7.1.1.1 Un nombre d'unités discrètes non spécifié

Dans les études et versions précédentes de COSMO, que ce soit celles réalisées avant cette thèse, comme COSMO-S, ou celles présentées aux chapitres 4 et 5, comme COSMO-Voyelle, l'agent apprenant possède toujours des connaissances sur le nombre de catégories à apprendre et ce, dès le début de l'apprentissage. C'est ainsi que, dans COSMO-Voyelle du chapitre précédent, l'agent apprenant possède sept objets O_S et O_L.

Cette considération permet de simplifier le modèle et son apprentissage mais reste peu réaliste, surtout au début de l'apprentissage. En effet, il y a peu de raisons valables pour qu'un bébé ait une connaissance préalable du nombre de phonèmes de sa langue, d'autant plus que celui-ci diffère d'une langue à l'autre. C'est pourquoi, nous souhaitons, dans COSMO SylPhon, que l'agent apprenant découvre de lui-même le nombre de catégories de son environnement.

7.1.1.2 Un apprentissage sans deixis

En plus de n'avoir aucune connaissance du nombre exact de catégories phonétiques de sa langue, nous supposons que l'agent est incapable de reconnaître les catégories phonétiques fournies par le maître durant son apprentissage. Ce point diffère également des précédentes études réalisées avec COSMO. En effet, dans celles-ci, l'agent reconnaît la catégorie de l'objet correspondant au stimulus envoyé par le maître, par exemple, par deixis. Dans COSMO SylPhon, l'agent apprenant ne reçoit que les stimuli auditifs du maître, sans information fournie sur les catégories correspondantes

Finalement, du fait de ces deux contraintes, l'agent apprenant perd la capacité à associer facile-

ment ses objets à des unités distinctives. Pour souligner cette différence, ses objets seront désormais nommés noyaux. À travers cette première condition, notre objectif est donc de savoir si l'agent réussit à apprendre des catégories phonétiques qui peuvent être comprises par le maître.

7.1.2 Condition 2 : Un invariant consonantique non explicite

Afin de comprendre la deuxième condition fixée, nous expliquons nos hypothèses concernant la nature des unités distinctives considérées dans cette étude. Elles sont essentielles puisque la composition du modèle COSMO SylPhon dépend en grande partie de cette réflexion.

7.1.2.1 Représentation motrice et acoustique des unités distinctives

Dans cette étude, nous nous focalisons sur les syllabes CV, c'est-à-dire composées d'une consonne et d'une voyelle. Cette vision simplifiée des syllabes a le mérite d'éviter l'importante combinatoire syllabique, les questionnements autour de sa composition, et facilite également la décomposition du modèle. Elle nous permet de nous centrer sur le type de syllabes le plus simple, le plus répandu dans les langues du monde, et qui est aussi celui qui survient dans les premiers temps du développement de la parole, au sein du babillage. La syllabe CV, choisie pour la présente étude, permet également de prendre en compte les deux catégories principales du phonème, c'est-à-dire la consonne et la voyelle, ce qui est un bon point d'entrée pour étudier les différentes unités phonétiques.

Passons aux phonèmes. Nous avons déjà eu l'occasion de modéliser des voyelles dans les simulations des chapitres précédents. Pour cela, nous avons utilisé le modèle articulatoire du conduit vocal VLAM que nous avons présenté à cette occasion. Comme décrit précédemment, VLAM est un modèle du conduit vocal qui, pour une configuration donnée de paramètres articulatoires, renvoie les formants du signal sensoriel correspondants. Ce fonctionnement est approprié pour les voyelles, comme l'ont montré, par exemple, les études réalisées dans le chapitre 5.

La modélisation des consonnes avec VLAM s'avère en revanche plus complexe. Contrairement aux voyelles qui peuvent être assimilées à une ouverture du conduit vocal, les consonnes correspondent globalement à une fermeture du conduit vocal. Dans cette étude, nous nous focalisons sur certaines consonnes particulières : les consonnes plosives. Nous faisons ce choix parce que le modèle VLAM ne permet pas de synthétiser aisément d'autres types de consonnes. En effet, VLAM étant un modèle articulatoire 2D, il ne peut pas synthétiser les latérales (comme le [l]), qui nécessitent une prise en compte d'un espace articulatoire en 3D [1]. De même, VLAM ne permet pas non plus de synthétiser les fricatives car elles nécessitent un modèle de génération de bruit, absent dans le modèle acoustique de VLAM. Enfin, les consonnes nasales nécessitent la modélisation d'un tuyau nasal, qui est absent dans VLAM car il s'agit d'un modèle du conduit vocal oral.

Néanmoins, l'utilisation des consonnes plosives pose un problème de définition du son correspon-

1. Le phonème [l], comme le phonème [t], est produit par une fermeture du conduit vocal au niveau des alvéoles. Cependant, pour le [l], la fermeture est seulement effective dans la partie médiane du conduit vocal et l'air peut s'écouler sur les parties latérales, d'où son nom.

dant à la consonne. Les plosives peuvent être caractérisées par de multiples indices acoustiques (voir une revue dans la thèse de Laurent, 2014, Chapitre 5, section 5.1.3). Un corrélat majeur est la valeur des formants caractérisant la configuration articulatoire juste avant la fermeture ou après l'ouverture. C'est l'option choisie dans cette étude, en prenant, pour caractériser les plosives sur VLAM, la configuration des paramètres articulatoires juste avant la fermeture, ce qui nous fournit, d'une part, une représentation des plosives et, d'autre part, le signal sensoriel correspondant.

Les unités distinctives considérées dans le modèle COSMO SylPhon sont les mêmes que celles de COSMO-S (Laurent et al., 2017). L'unique différence est qu'elles ne servent, dans COSMO-S, qu'à caractériser, d'une part, les syllabes CV et, d'autre part, les portions C et V de ces syllabes, alors que dans COSMO SylPhon, ces portions renvoie explicitement à la caractérisation des phonèmes, ce qui fournit la base de l'implémentation purement phonémique du modèle.

7.1.2.2 Les invariants des unités distinctives

Suite aux résultats obtenus avec COSMO-S, nous faisons l'hypothèse que les unités vocaliques sont mieux représentées dans la branche auditive du modèle tandis que les unités consonantiques sont mieux représentées dans la branche motrice du modèle.

Afin d'obtenir un résultat similaire, cela suppose que l'agent apprenant ait la possibilité d'apprendre correctement ces unités phonémiques. L'implémentation des voyelles, déjà apprises avec succès dans le chapitre précédent, ne pose pas de difficulté particulière. Elles sont clairement définies aussi bien dans l'espace moteur que dans l'espace auditif, quoique mieux contraintes et plus focalisées dans le second espace.

Le cas des consonnes est plus complexe. En effet, un problème concernant les consonnes et notamment les plosives est la coarticulation (voir section 3.1.1.1 du chapitre 2) : leur production est très dépendante des productions temporellement environnantes. Les plosives sont des consonnes qui ne peuvent être conçues de manière isolée du fait que l'échappement d'air à l'ouverture est conditionné par les productions vocaliques environnantes et particulièrement celle de la voyelle qui suit. Du fait de l'utilisation des syllabes CV, nous supposons, d'une part, que les productions environnantes ne concernent que la voyelle qui suit et, d'autre part, que l'invariant consonantique se trouve dans le geste moteur permettant de passer de la configuration motrice consonantique à la configuration motrice vocalique.

Nous émettons une troisième hypothèse : ce geste est effectué de la manière la plus simple possible. Plus précisément, nous suggérons que le geste moteur permettant de passer de la consonne à la voyelle ne nécessite qu'un unique articulateur, différent pour chaque plosive, constituant l'invariant consonantique. Cette hypothèse peut se justifier par l'utilisation d'une parcimonie articulatoire limitant les efforts à faire lors de la production. De plus, en suivant les principes de la théorie « Frame-then-Content » (MacNeilage et Davis, 1990) émanant des études sur le développement du bébé (voir section 3.2.1.2, chapitre 2), nous estimons qu'un agent apprenant commence à contrôler son conduit vocal en effectuant également des gestes simples, en limitant le nombre d'articulateurs à déplacer. De ce fait, il devrait en priorité apprendre un invariant consonantique composé de gestes simples. Dans COSMO-S, cette hypothèse est modélisée chez l'agent apprenant en contraignant l'espace moteur

consonantique à trois mouvements spécifiques permettant de réaliser les trois consonnes du modèle. Dans COSMO SylPhon, nous enlevons cette contrainte mais nous implémentons une amorce au début de l'apprentissage dans laquelle l'agent apprenant possède une plus forte probabilité pour des gestes moteurs simples, nécessitant peu d'articulateurs. Cette contrainte ne garantit en rien la convergence de l'agent vers des gestes moteurs simplifiés. Notre objectif est donc de savoir si cette contrainte est suffisante pour apprendre l'invariant consonantique choisi et les consonnes d'une manière générale.

En résumé, à travers ces deux conditions, nous souhaitons répondre à deux questions. La première concerne la convergence du modèle : est-ce que l'agent est capable d'apprendre des unités distinctives sous ces conditions ? Si oui, quels sont les invariants des unités ? La seconde concerne la communication du modèle : est-ce que l'agent, quelles que soient ses unités distinctives, est capable de communiquer avec son maître ? Bien entendu, nous nous servons également de ces deux questions pour comparer l'apprentissage phonémique et l'apprentissage syllabique.

7.2 Description du modèle

7.2.1 Trois modèles génériques COSMO

Si COSMO SylPhon est une version plus aboutie que le modèle original, elle possède avec lui de nombreux points communs et sa description ne peut se faire sans parler du modèle de base. Pour faciliter la compréhension et pour passer facilement de l'un à l'autre, dans ce chapitre, l'extension est toujours appelée COSMO SylPhon et la version de base est toujours nommée COSMO.

Nous souhaitons utiliser des phonèmes (voyelles et consonnes) et des syllabes. Pour construire COSMO SylPhon, nous partons de trois modèles COSMO théoriques : un pour les syllabes, un pour les consonnes et un pour les voyelles. Ils sont représentés Fig. 7.1.

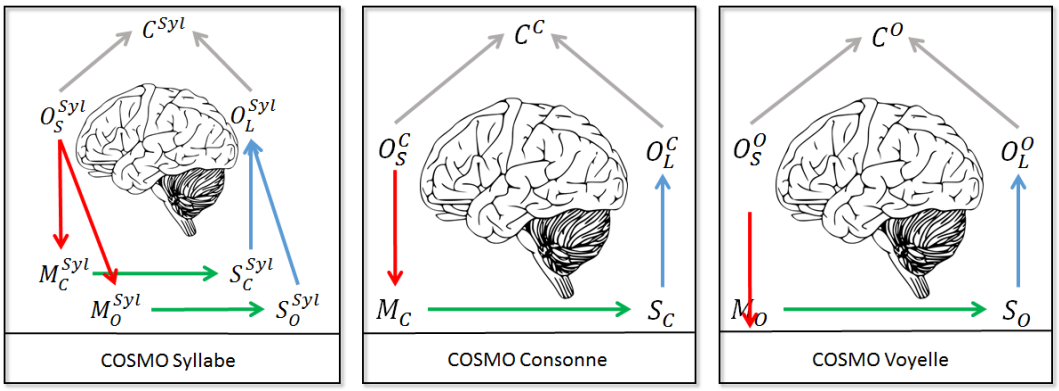


FIGURE 7.1 – Illustration des trois modèles COSMO, associés à chaque type d'unités phonétiques

Dans ces figures et, par la suite, pour différencier les variables communes, nous utilisons l'exposant Syl pour les syllabes, O pour les voyelles (O comme « Open » car elles correspondent à une ouverture du conduit vocal) et C pour les consonnes (C comme « Closed », car elles correspondent à une fermeture du conduit vocal). Les modèles « Consonne » et « Voyelle » sont semblables au

COSMO de base. Concernant le modèle « Syllabe », les syllabes considérées étant des variables CV, nous supposons qu'il possède deux représentations motrices et deux représentations sensorielles : l'une pour les consonnes (respectivement M_C^{Syl} et S_C^{Syl}) et l'autre pour les voyelles (respectivement M_O^{Syl} et S_O^{Syl}).

Énumérons maintenant les principales adaptations à faire dans COSMO pour réaliser l'extension souhaitée : i) le répertoire auditif devient explicite, ii) les objets sont renommés noyaux et iii) la variable motrice consonantique nécessite d'être décomposée.

7.2.1.1 Du classifieur auditif au répertoire auditif

Pour la première modification, nous remplaçons le classifieur auditif $P(O_L \mid S)$ de COSMO par le répertoire auditif $P(S \mid O_L)$. À ce stade, nous conservons ici, par simplicité, des variables génériques O, S, M et C sans rappeler les indices Syl, C ou O, ce qui permet d'appliquer ce qui suit, indifféremment, aux modèles Syllabe, Consonne ou Voyelle. Ainsi, le répertoire auditif devient une distribution apparaissant dans la décomposition de la conjointe du modèle tandis que $P(O_L \mid S)$ devient une distribution calculée par inférence.

Concrètement que cela change-t-il ? Une des règles de la programmation bayésienne est qu'il n'est pas possible d'avoir dans la décomposition deux fois la même variable en partie gauche d'une distribution. Or, c'est ce que nous obtenons ici puisque nous cherchons à faire apparaître à la fois le modèle interne $P(S \mid M)$ et le répertoire auditif $P(S \mid O_L)$, qui contiennent tous deux la variable sensorielle S en partie gauche. Pour y remédier, une des solutions est de dupliquer la variable concernée. Nous conservons la notation S pour les représentations sensorielles liées aux objets et notons S^M les représentations sensorielles liées aux représentations motrices. Les deux variables sensorielles S et S^M sont ensuite liées par une variable de cohérence λ_{SM} (de même nature mathématique que la variable de cohérence C), assurant que les deux variables sensorielles sont identiques. La nouvelle décomposition de COSMO est donc la suivante :

$$\begin{aligned} & P(C\ O_S\ S\ M\ O_L\ S^M\ \lambda_{SM}) \\ & \quad = \quad P(O_S)\ P(M \mid O_S)\ P(S^M \mid M)\ P(S \mid O_L)\ P(C \mid O_S\ O_L)\ P(\lambda_{SM} \mid S^M\ S) \end{aligned} \tag{7.1}$$

7.2.1.2 Des objets au noyaux

Comme nous l'avons précédemment énoncé, nous souhaitons souligner le fait que les agents n'ont pas de connaissances préalables sur le nombre de catégories phonétiques et qu'ils sont incapable d'associer précisément un de leur objet à une catégorie phonétique lors de l'apprentissage. C'est pourquoi nous remplaçons la terminologie des variables « objets » O par des variables « noyaux » N. Nous avons donc désormais, dans COSMO, les variables N_S et N_L à la place, respectivement, de O_S et O_L.

7.2.1.3 La décomposition de la variable motrice consonantique

La troisième modification concerne la représentation de l'invariant consonantique dans l'espace moteur. Comme nous l'avons précédemment énoncé, l'invariant consonantique, tel que nous le modélisons, correspond à la nature de l'articulateur (unique) permettant de passer de la configuration consonantique à la configuration vocalique. Afin de pouvoir l'observer dans le modèle, nous implémentons désormais deux variables motrices consonantiques : d'une part, la variable M_C, qui représente la configuration motrice consonantique de base et, d'autre part, la variable ΔM qui représente le geste permettant le passage de la configuration consonantique à la configuration vocalique. Cette seconde variable, selon nos hypothèse, porterait donc l'invariant consonantique.

Computationnellement, les deux variables sont très liées et peuvent être considérée comme émanant d'un changement de repère. Ainsi, M_C se calcule simplement à l'aide de la variable vocalique ΔM et M_O par :

$$M_C = \Delta M + M_O \ . \tag{7.2}$$

Dans COSMO SylPhon, la variable M_C est, d'une part, considérée comme prior $P(M_C)$ possédant des contraintes biologiques associées à sa nature de configuration fermée, et, d'autre part, reliée à M_O et ΔM, du fait du changement de repère, grâce à la distribution $P(M_C \mid \Delta M \ M_O)$. Comme pour le cas du répertoire auditif, cela pose problème dans COSMO car il y a deux distributions avec M_C en partie gauche. C'est pourquoi nous dédoublons M_C en deux variables : M_C et M_{NC} liées par une variable de cohérence λ_{MC}. De la même manière, comme il existe un prior moteur vocalique $P(M_O)$ associé au statut de configuration ouverte, nous dédoublons également M_O en deux variables : M_O et M_{NO} liées par une variable de cohérence λ_{MO}. Nous obtenons ainsi les distributions $P(M_C)$, $P(M_O)$ et $P(M_{NC} \mid \Delta M \ M_{NO})$.

Nous reportons ces trois changements dans les modèles théoriques, représentés Fig. 7.2. Nous constatons que le modèle théorique consonantique est désormais dépendant du modèle vocalique.

7.2.2 Le modèle COSMO SylPhon

La dernière étape pour construire COSMO SylPhon consiste à regrouper les trois modèles théoriques pour n'en former qu'un seul. En effet, un critère majeur du modèle est qu'il soit possible d'étudier, dans un même modèle, un modèle phonémique (C et V) et un modèle syllabique (CV) dans une même architecture, de la même manière que COSMO permet d'étudier conjointement, au sein d'un même modèle, les représentations motrices et sensorielles.

Les modèles vocaliques et consonantiques théoriques sont déjà reliés suite aux modifications apportées au modèle. Il reste donc à relier le modèle syllabique aux modèles phonémiques. Pour cela, nous fusionnons les différentes variables motrices M_C, d'une part, et M_O, d'autre part, ainsi que les variables sensorielles S_C, d'une part, et S_O, d'autre part, que nous considérons communes aux modèles syllabiques et phonémiques. Nous ajoutons ensuite des variables de cohérence entre les variables de même nature : S_C, S_O, ΔM, M_{NC} et M_{NO} . Le modèle final est illustré Fig. 7.3.

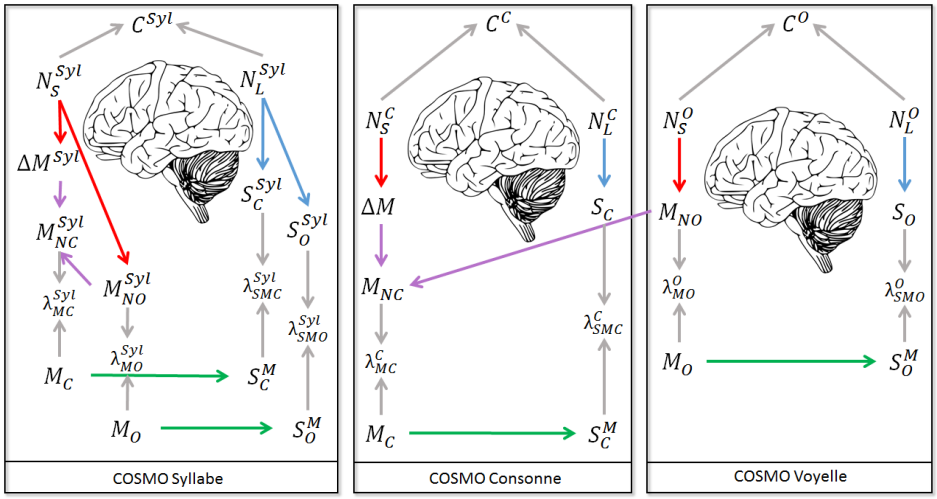


FIGURE 7.2 – Illustration des trois modèles COSMO, associés à chaque type d'unités phonétiques, après les trois modifications

Tout ceci pris en compte, l'ensemble des variables du modèle forme la conjointe que nous décomposons de la façon suivante :

$$
\begin{aligned}
&P(N_S^C\ N_L^C\ N_S^O\ N_L^O\ N_S^{Syl}\ N_L^{Syl}\ \Delta M\ M_{NC}\ M_{NO}\ \Delta M^{Syl}\ M_{NC}^{Syl}\ M_{NO}^{Syl}\ M_C\ M_O\\
&S_C\ S_O\ S_C^{Syl}\ S_O^{Syl}\ S_C^M\ S_O^M\ C_C\ C_O\ C_{Syl}\ \lambda_{\Delta M}\ \lambda_{NC}\ \lambda_{NO}\ \lambda_{MC}^C\ \lambda_{MO}^O\ \lambda_{MC}^{Syl}\ \lambda_{MO}^{Syl}\\
&\lambda_{SC}\ \lambda_{SO}\ \lambda_{SMC}^C\ \lambda_{SMO}^O\ \lambda_{SMC}^{Syl}\ \lambda_{SMO}^{Syl})\\
&= P(N_S^C)\ P(N_L^C)\ P(N_S^O)\ P(N_L^O)\ P(N_S^{Syl})\ P(N_L^{Syl})P(M_O)P(M_C)\\
&\quad P(\Delta M \mid N_S^C)\ P(M_{NO} \mid N_S^O)\ P(\Delta M^{Syl}\ M_{NO}^{Syl} \mid N_S^{Syl})\\
&\quad P(S_C \mid N_L^C)\ P(S_O \mid N_L^O)\ P(S_C^{Syl}\ S_O^{Syl} \mid N_L^{Syl})\\
&\quad P(S_C^M \mid M_C)\ P(S_O^M \mid M_O)\\
&\quad P(M_{NC} \mid \Delta M\ M_{NO})\ P(M_{NC}^{Syl} \mid \Delta M^{Syl}\ M_{NO}^{Syl})\\
&\quad P(\lambda_{NC} \mid M_{NC}\ M_{NC}^{Syl})\ P(\lambda_{\Delta M} \mid \Delta M\ \Delta M^{Syl})\ P(\lambda_{NO} \mid M_{NO}\ M_{NO}^{Syl})\\
&\quad P(\lambda_{MC}^C \mid M_C\ M_{NC})\ P(\lambda_{MO}^O \mid M_O\ M_{NO})\ P(\lambda_{MC}^{Syl} \mid M_C\ M_{NC}^{Syl})\ P(\lambda_{MO}^{Syl} \mid M_O\ M_{NO}^{Syl})\\
&\quad P(\lambda_{SMC} \mid S_C\ S_C^M)\ P(\lambda_{SMO} \mid S_O\ S_O^M)\ P(\lambda_{SMC}^{Syl} \mid S_C^{Syl}\ S_C^M)\ P(\lambda_{SMO}^{Syl} \mid S_O^{Syl}\ S_O^M)\\
&\quad P(\lambda_{SC} \mid S_C\ S_C^{Syl})\ P(\lambda_{SO} \mid S_O\ S_O^{Syl})\ P(C_C \mid O_S^C\ O_L^C)\ P(C_O \mid O_S^O\ O_L^O)\ P(C_{Syl} \mid O_S^{Syl}\ O_L^{Syl})
\end{aligned}
\tag{7.3}
$$

L'ensemble de cette composition peut sembler imposant. Néanmoins, ces distributions peuvent être ordonnées et regroupées en six catégories : les priors, les répertoires moteurs, les répertoires auditifs, les modèles internes, les dépendances consonantiques et les systèmes de cohérence . Les quatre premières catégories correspondent globalement aux types de distributions que nous avions précédemment dans COSMO. La cinquième catégorie est, quant à elle, une conséquence directe de la représentation consonantique que nous avons déjà longuement détaillée dans la section précédente. La sixième catégorie, les systèmes de cohérence, est déjà présente au préalable, et assure de fait la cohérence de l'ensemble du modèle. Précisons maintenant un peu plus chacune de ces catégories.

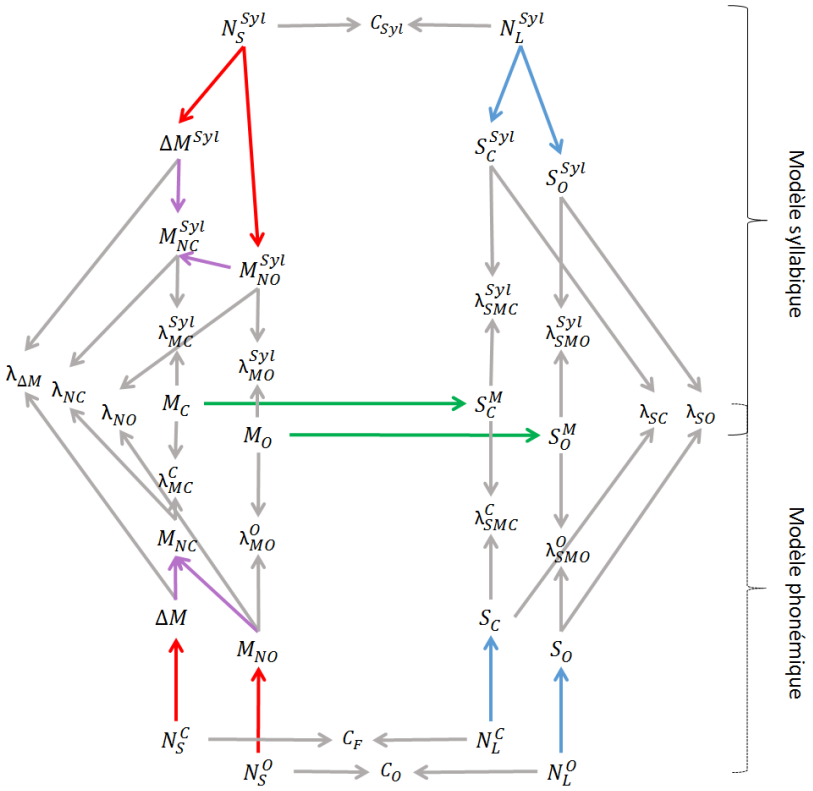


FIGURE 7.3 – Illustration du modèle COSMO SylPhon

7.2.3 Description des distributions

Les priors Dans COSMO SylPhon, la majorité des priors concerne les priors sur les noyaux, c'est-à-dire, d'une part, $P(N_S^C)$, $P(N_S^O)$ et $P(N_S^{Syl})$, que nous regroupons sous l'expression générique priors des noyaux moteurs $P(N_S)$ et, d'autre part, $P(N_L^C)$, $P(N_L^O)$ et $P(N_L^{Syl})$ que nous regroupons sous l'expression générique prior des noyaux auditifs $P(N_L)$.

Concernant leur forme paramétrique, les distributions $P(N_S)$ et $P(N_L)$ correspondent à la probabilité des noyaux discrets servant à décrire les unités distinctives. Cette probabilité dépend entièrement de l'apprentissage et évolue au cours de celui-ci. Nous décidons que les priors suivent une loi de succession de Laplace et qu'elle est similaire initialement à une distribution uniforme.

Les autres types de prior sont les priors moteurs vocalique $P(M_O)$ et consonantique $P(M_C)$. Ils sont l'un et l'autre uniformes sur l'ensemble des configurations correspondantes, respectivement ouvertes pour $P(M_O)$ et quasi fermées pour $P(M_C)$. Nous reviendrons sur le contenu de ces espaces.

Les répertoires moteurs Il y a trois types de répertoires moteurs : le répertoire moteur pour les consonnes $P(\Delta M \mid N_S^C)$, le répertoire moteur pour les voyelles $P(M_{NO} \mid N_S^O)$ et le répertoire moteur pour les syllabes $P(\Delta M^{Syl}\ M_{NO}^{Syl} \mid N_S^{Syl})$. Ces trois répertoires moteurs sont, comme le répertoire moteur du modèle COSMO, représentés par un ensemble de gaussiennes. Cependant, contrairement au modèle COSMO dans lequel chaque distribution gaussienne est reliée à un objet et donc à une unité

distinctive, dans COSMO SylPhon, chaque distribution gaussienne est reliée à un noyau N_S. Ainsi, une unité distinctive du maître est potentiellement représentée articulatoirement par un ensemble de plusieurs noyaux gaussiens pour l'agent apprenant.

Les répertoires auditifs Similairement aux répertoires moteurs, il y a trois types de répertoires auditifs : le répertoire auditif pour les consonnes $P(S_C \mid N_L^C)$, le répertoire auditif pour les voyelles $P(S_O \mid N_L^O)$ et le répertoire auditif pour les syllabes $P(S_C^{Syl} S_O^{Syl} \mid N_L^{Syl})$. Comme pour les répertoires moteurs, ils sont également représentés par un ensemble de distributions gaussiennes, une pour chaque noyau N_L.

Les modèles internes Il y a deux modèles internes, un pour les consonnes $P(S_C^M \mid M_C)$ et un pour les voyelles $P(S_O^M \mid M_O)$. Les modèles internes sont similaires à celui de COSMO et sont représentés par un ensemble de gaussiennes, un pour chaque configuration motrice considérée.

Les dépendances consonantiques Ces distributions servent à lier les deux types de variables motrices consonantiques. Pour cela, la probabilité d'une dépendance consonantique vaut 1 si et seulement si l'équation $M_{NC} = \Delta M + M_{NO}$ (resp. $M_{NC}^{Syl} = \Delta M^{Syl} + M_{NO}$ pour les syllabes) est respectée (voir Eq. 7.2).

Les systèmes de cohérence Ce sont les distributions servant à lier des variables similaires du modèle. Il y en a seize dans COSMO SylPhon. Ce sont les distributions contenant « C » ou « λ » en partie gauche. Elles sont construites exactement sur le même principe que le système de cohérence de COSMO dont la description est donnée section 4.2.2.2. Il faut simplement les considérer comme des « interrupteurs » permettant de lier les variables entre elles. Si la variable de cohérence vaut 1, les variables en partie droite de la distribution sont connectées et sont égales l'une à l'autre. En revanche, si la variable de cohérence n'est pas spécifiée, les variables en partie droite de la distribution ne sont pas connectées et sont indépendantes l'une de l'autre.

7.2.4 Implémentation

Le modèle étant défini, nous décrivons maintenant plus en détail comment les variables et distributions sont implémentées dans nos simulations.

7.2.4.1 Les variables

Nous décrivons l'implémentation des variables en les regroupant par variables de même famille. Il y en a quatre dans le modèle : les noyaux N, les représentations motrices M et ΔM, les représentations sensorielles S et les variables de cohérence C et λ.

Les noyaux Nous nous intéressons à trois sortes d'unités distinctives : les voyelles, les consonnes et les syllabes. Les voyelles considérées sont les mêmes que celles du chapitre 5 : [a i u e ɛ o ɔ]. Les consonnes considérées sont les plosives [b d g]. Nous supposons, arbitrairement, que ces plosives sont voisées mais, comme le modèle articulatoire VLAM ne permet pas de manipuler le paramètre de

voisement, nous aurions pu sans distinction considérer qu'il s'agit de leurs homologues non voisés [p t k] qui ont essentiellement les mêmes propriétés que [b d g] en termes de trajectoires formantiques. Nous considérons donc les 21 syllabes CV [ba da ga bi di gi bu du gu be de ge bɛ dɛ gɛ bo do go bɔ dɔ gɔ].

Comme nous n'avons pas de réel critère pour définir le nombre de noyaux, la seule contrainte que nous nous sommes donnés est que le modèle contient plus de noyaux que d'unités distinctives à apprendre. Nous notons respectivement : nb_{SC} et nb_{LC} le nombre de noyaux pour les consonnes dans N_S^C et N_L^C, supérieur à 3, nb_{SO} et nb_{LO}, le nombre de noyaux pour les voyelles dans N_S^O et N_L^O, supérieur à 7, et nb_{SSyl} et nb_{LSyl} le nombre de noyaux pour les syllabes dans N_S^{Syl} et N_L^{Syl}, supérieur à 21.

Les représentations motrices Celles-ci sont, comme dans COSMO, des configurations articulatoires des articulateurs du modèle articulatoire VLAM. Comme dans COSMO-Voyelle, nous gardons trois paramètres articulatoires pour représenter les gestes moteurs vocaliques M_O et M_O^{Syl} : la hauteur des lèvres (LH), le corps de la langue (TB) et le dos de la langue (TD). Ces trois paramètres ne sont pas suffisants pour représenter les consonnes plosives considérées. C'est pourquoi nous considérons deux autres paramètres de VLAM qui sont : la pointe de la langue (Apex), qui est notamment nécessaire pour représenter les plosives [d] et la mâchoire (Jaw) qui est le support de toutes les articulations consonantiques. Nous obtenons donc des espaces moteurs vocaliques à trois dimensions (ou cinq dimensions en considérant Jaw et Apex à 0, leur valeur de repos) et des espaces moteurs consonantiques à cinq dimensions.

Comme précédemment, les représentations motrices sont des ensembles finis et discrétisés. Les espaces ΔM (resp. ΔM^{Syl}) se calculent directement à partir des valeurs des espaces consonantiques M_{NC} et vocaliques M_{NO}. Ainsi, les valeurs de l'espace ΔM s'obtiennent en calculant $\Delta M = M_{NC} - M_{NO}$ et celles de l'espace ΔM^{Syl} s'obtiennent en calculant $\Delta M^{Syl} = M_{NC}^{Syl} - M_{NO}^{Syl}$. Pour les autres dimensions motrices, en s'inspirant du modèle COSMO-V, les valeurs des paramètres sont contenues dans l'intervalle $[-5, +5]$ et chaque dimension est discrétisée en 15 cases. Les variables des espaces moteurs vocaliques M_O, M_{NO} et M_{NO}^{Syl} contiennent donc 3 375 valeurs, considérées équiprobables selon les priors uniformes $P(M_O)$, comme dans le chapitre précédent, et celles des espaces moteurs consonantiques M_C, M_{NC} et M_{NC}^{Syl} contiennent 759 375 cases.

Le nombre de cases pour les espaces moteurs consonantiques nous semble trop conséquent, surtout lorsque nous implémentons le modèle interne. Or, tous les points de l'espace ne représentent pas une plosive (voir section 7.1.2.1). En effet, certaines configurations conduisent à un conduit vocal ouvert qui représente une voyelle et non une consonne plosive. D'autres, au contraire, conduisent à une fermeture totale du conduit vocal qui ne permet pas de calculer la résultante acoustique. C'est pourquoi nous avons réalisé un travail préalable sur cet espace pour ne conserver que les configurations consonantiques telles que nous les définissons. Plus spécifiquement, nous n'avons conservé que les cases, parmi les 759 375 cases de l'espace, qui possèdent au moins une configuration correspondant à une ouverture relative, c'est-à-dire une aire de la constriction, entre 0,05 et 0,07 cm^2 (calculée avec VLAM). En effectuant ce test pour chaque case, par un tirage aléatoire de dix configurations, nous n'avons conservé que 187 547 cases, considérées équiprobables selon les priors uniformes $P(M_C)$.

Les représentations auditives Elles sont, comme dans COSMO-Voyelle, caractérisées par des paramètres formantiques, dont l'unité de mesure est le Bark. Comme précédemment, les variables sensorielles vocaliques S_O et S_O^{Syl} sont décrites par les formants F1 et F2. Comme l'agent doit apprendre les mêmes sept voyelles que dans le chapitre précédent, nous savons que ces deux formants sont suffisants pour les caractériser. Pour les variables sensorielles consonantiques S_C et S_C^{Syl}, nous choisissons de les caractériser via les formants F2 et F3. Ces deux formants semblent suffisants pour caractériser les plosives du français (voir par exemple Laurent et al., 2017). Ainsi, nous obtenons des espaces sensoriels vocaliques et consonantiques bidimensionnels.

Les variables sensorielles sont finies et discrétisées. Nous discrétisons chaque dimension en 25 cases. Le formant F1 est défini dans l'intervalle $[2{,}3\ ;\ 7{,}1]$ Barks, le formant F2 (pour les variables sensorielles vocaliques et consonantiques) est défini dans l'intervalle $[4{,}7\ ;\ 13{,}8]$ Barks et le formant F3 est défini dans l'intervalle $[12{,}8\ ;\ 16{,}5]$ Barks. La discrétisation dans chacun de ces intervalles est faite de manière linéaire.

Les variables de cohérence Toutes les variables de cohérence, aussi bien les variables C que les variables λ, sont définies de la même manière que dans COSMO. Il s'agit donc de variables booléennes prenant les valeurs « vrai » (1) ou « faux » (0).

7.2.4.2 Les distributions de probabilité

Pour décrire comment sont implémentées les distributions, nous reprenons les six catégories que nous avons utilisées précédemment en section 7.2.3. Comme l'implémentation des systèmes de cohérence et des dépendances consonantiques ne diffère pas de leur définition générale, décrite ci-dessus, nous nous concentrons sur les quatre autres.

Les priors sur les noyaux Les priors sur les noyaux sont des distributions contenant le même nombre de valeurs que les noyaux qu'elles caractérisent : nb_{SC} pour $P(N_S^C)$, nb_{LC} pour $P(N_L^C)$, nb_{SO} pour $P(N_S^O)$, nb_{LO} pour $P(O_L^O)$, nb_{SSyl} pour $P(N_S^{Syl})$ et nb_{LSyl} pour $P(N_L^{Syl})$. Ces distributions suivent une loi de succession de Laplace. Plus précisément, elles se définissent à partir d'une forme initiale uniforme, qui évoluent ensuite à la manière d'un histogramme, en incrémentant les cases observées (chaque fois qu'un noyau est sélectionné dans l'apprentissage, son nombre d'observations est augmenté de 1). Ainsi, en notant obs_n le nombre d'observations du noyau n d'une des distribution prior de la forme $P(N)$ qui porte sur K noyaux, nous avons :

$$P([N = n]) = \frac{1 + obs_n}{K + \sum_N obs_n} \,. \tag{7.4}$$

Les répertoires moteurs Comme dans nos études précédentes, les répertoires moteurs sont des ensembles de distributions gaussiennes tronquées et discrétisées, paramétrées par une moyenne μ et une matrice de covariance Σ (cf Eq.5.11). Pour chaque répertoire moteur, il y a autant de distributions gaussiennes que de noyaux. Il y a donc nb_{SC} distributions gaussiennes dans le répertoire

$P(\Delta M \mid N_S^C)$, nb_{SO} distributions gaussiennes dans le répertoire $P(M_{NO} \mid N_S^O)$ et nb_{SSyl} distributions gaussiennes dans le répertoire $P(\Delta M^{Syl}\ M_{NO}^{Syl} \mid N_S^{Syl})$.

Les répertoires auditifs Comme les répertoires moteurs, les répertoires auditifs sont, une nouvelle fois, des ensembles de distributions gaussiennes tronquées et discrétisées. Il y a également autant de distributions gaussiennes que de noyaux pour chaque répertoire auditif, c'est-à-dire nb_{LC} distributions gaussiennes dans le répertoire $P(S_C \mid N_L^C)$, nb_{LO} distributions gaussiennes dans le répertoire $P(S_O \mid N_L^O)$ et nb_{LSyl} distributions gaussiennes dans le répertoire $P(S_C^{Syl}\ S_O^{Syl} \mid N_L^{Syl})$.

Les modèles internes Ce sont eux aussi des ensembles de distributions gaussiennes tronquées et discrétisées. Ils possèdent une distribution gaussienne pour chaque configuration motrice considérée. Il y a donc 3 375 distributions gaussiennes pour le modèle interne vocalique $P(S_O^M \mid M_O)$ et 187 547 distributions gaussiennes pour le modèle interne consonantique $P(S_C^M \mid M_C)$.

7.2.4.3 L'initialisation

Au début de l'apprentissage, comme dans COSMO, nous supposons que les ensembles de gaussiennes (répertoires auditifs, répertoires moteurs et modèles internes) approximent des distributions uniformes. Pour cela, nous considérons que leurs moyennes μ sont regroupées au centre de leurs espaces respectifs et que les valeurs diagonales des matrices de covariances Σ sont élevées (environ la taille de l'espace).

Les deux exceptions à cette initialisation des répertoires sont le répertoire moteur consonantique et le répertoire moteur syllabique. En effet, rappelons que nous souhaitons que le passage d'une configuration articulatoire consonantique à une configuration articulatoire vocalique s'effectue à l'aide d'un geste simple, impliquant le mouvement d'un articulateur principal, caractéristique de la consonne. Nous ne souhaitons pas, comme dans COSMO-S, implémenter cette contrainte directement. Nous souhaitons évaluer si l'agent est capable de l'apprendre sans produire de configurations consonantiques aberrantes cognitivement. Pour ce faire, nous l'implémentons comme une « amorce » : nous supposons que le répertoire consonantique (resp. la partie consonantique du répertoire syllabique) possède initialement des gaussiennes positionnées à 0 (valeur de repos) avec une grande variance sur un des articulateurs consonantiques TD, LH ou Apex et une petite variance sur les trois articulateurs restants. Elles ont également une grande variance sur Jaw, que nous considérons comme l'articulateur commun entre toutes les consonnes. Cette initialisation est une forme d'implémentation du calendrier développemental de la théorie Frame then Content : au départ les gestes du babillage sont stéréotypés autour de l'articulateur porteur, la mâchoire, et d'un articulateur spécifique. La question posée est de savoir si, au cours de l'apprentissage, l'agent parviendra à maintenir ces coordinations et à sélectionner les coordinations adéquates pour chaque consonne.

Contrairement à COSMO, les priors des objets nécessitent également une initialisation. Au début d'apprentissage, il n'y a aucune observation. Le paramètre obs_n vaut donc 0 (voir Eq. 7.4). Ainsi, Les priors sont, comme les ensembles de gaussiennes, uniformes en début d'apprentissage.

7.3 Apprentissage

Nous conservons, dans cette étude, les trois apprentissages, sensoriel, sensorimoteur et moteur, ainsi que l'interaction avec un maître. Néanmoins, contrairement aux apprentissages définis précédemment, le maître ne fournit plus à l'agent les catégories phonétiques correspondant aux stimuli sensoriels qu'il envoie. Ainsi, lors de l'apprentissage, en plus d'apprendre ses représentations sensorielles et motrices, l'apprentissage des répertoires sensoriels et moteurs se fait de façon totalement non supervisée.

L'ensemble des apprentissages effectués est représenté Fig. 7.4. Certains détails techniques de cet apprentissage sont précisés en annexe 9.2. Contrairement aux chapitres précédents, nous ne séparons pas ici la méthode des résultats. Nous présentons successivement les principes d'apprentissage de chaque bloc du modèle, ainsi que les principaux résultats observés.

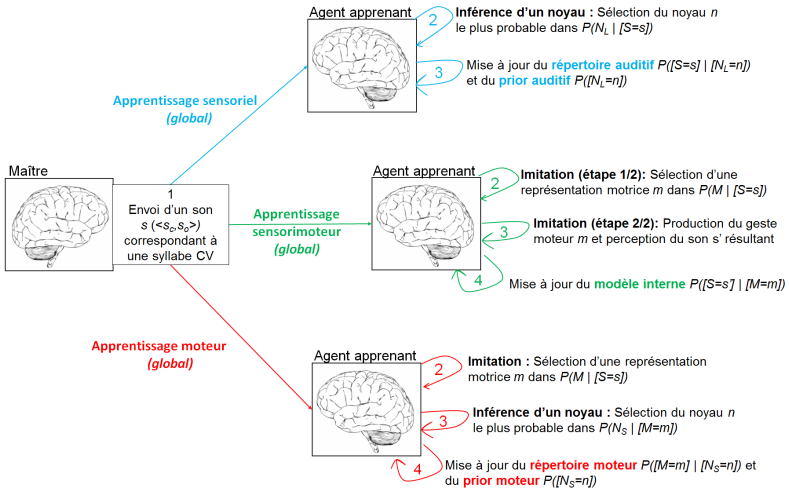


FIGURE 7.4 – Synthèse des apprentissages effectués avec le modèle COSMO Sylphon, ce scénario est décliné pour chacun des modèles syllabique, consonantique et vocalique

7.3.1 Élaboration des données d'apprentissage

7.3.1.1 Élaboration du répertoire syllabique du maître

Si l'agent apprenant apprend au fur et à mesure ses distributions phonémiques et syllabiques, ce n'est pas le cas du maître qui possède, au préalable, des connaissances motrices de ses unités phonétiques pour pouvoir les communiquer convenablement à l'agent lors de son apprentissage. Nous supposons que le maître ne communique que des syllabes à l'agent apprenant.

Plutôt que de construire une distribution complète du répertoire moteur du maître dont la forme est inconnue, nous élaborons un dictionnaire de configurations articulatoires. Les syllabes du maître, comme celle de l'agent, sont des syllabes CV. Le répertoire syllabique moteur du maître peut donc

se décomposer d'une part en représentations motrices consonantiques M_C^{Maitre} et d'autre part en représentations motrices vocaliques M_O^{Maitre}.

Commençons par les représentations vocaliques. Celles-ci sont organisées sous la forme d'un dictionnaire vocalique contenant 5 000 configurations articulatoires pour chaque voyelle du modèle. Sachant qu'il y a 7 voyelles, [a i u e ɛ o ɔ], il y a donc 35 000 configurations articulatoires au total dans ce dictionnaire vocalique. La construction de ce dictionnaire est réalisée à l'aide d'un tirage de 5 000 points dans des distributions gaussiennes vocaliques. Ces dernières sont similaires aux distributions gaussiennes du répertoire moteur vocalique du maître définies dans les précédentes études. Pour rappel, elles correspondent à un ensemble de distributions gaussiennes, une pour chaque voyelle, ayant chacune pour paramètres une moyenne μ et une matrice de covariance Σ. La moyenne correspond à un prototype moteur, calculé avec VLAM à partir des prototypes auditifs des voyelles définis par (Meunier, 2007). Les coefficients diagonaux de la matrice de covariance valent 0,1.

Passons aux représentations consonantiques. Comme nous l'avons présenté précédemment, nous considérons que la consonne est composée d'une configuration articulatoire proche de la fermeture de la plosive choisie et dépendante de la voyelle qui suit. C'est pourquoi, le dictionnaire consonantique est totalement dépendant du dictionnaire vocalique et s'élabore en fonction de lui. Outre leurs dépendances aux voyelles, les configurations articulatoires consonantiques respectent une seconde règle, celle de ne bouger que deux articulateurs par rapport à la voyelle : celui lié à la mâchoire (Jaw dans VLAM) et un articulateur spécifique à la consonne. Cela nous permet d'avoir un invariant consonantique composé d'un articulateur unique, comme nous l'avons précédemment supposé. Nous conservons en plus la mâchoire en tant qu'articulateur commun entre toutes les productions consonantiques. Il est donc important de noter que la présence d'un invariant consonantique correspondant à la sélection d'un articulateur spécifique à la consonne, en plus de la mâchoire, est considérée comme acquise par le maître.

Pour rappel, nous nous intéressons à trois consonnes dans le modèle : [b d g]. Le phonème [b] est une labiale, cela signifie que son lieu d'articulation, c'est-à-dire son lieu de fermeture, est les lèvres. Nous supposons que son élaboration consiste à manipuler l'articulateur LH (hauteur des lèvres) pour passer de la voyelle à la consonne. Le phonème [d] est une alvéolaire. Pour toucher l'alvéole, nous utilisons la point de la langue (apex). Nous allons supposons donc que l'élaboration du phonème [d] se fait grâce à la manipulation de l'articulateur $Apex$. Enfin, le phonème [g] est une vélaire, c'est-à-dire que son lieu de fermeture est l'arrière du palais, et l'articulateur permettant de réaliser cette fermeture est le dos de la langue, c'est-à-dire, dans VLAM, l'articulateur TD (Tongue Dorsum).

Dans ces conditions, pour chaque configuration motrice du dictionnaire vocalique, nous commençons par choisir la plosive à produire et son articulateur prédéfini. Ensuite, nous tirons aléatoirement cinq valeurs pour le paramètre choisi dans l'intervalle $[-5\ ;5]$ et, pour chacune d'elles, nous effectuons un parcours de l'articulateur Jaw pour trouver une configuration consonantique viable. Parmi les configurations résultantes, nous ne conservons que celles pour lesquelles la fermeture du conduit vocal est comprise dans un certain intervalle. Cet intervalle assure que la configuration est proche de la fermeture mais qu'elle n'est pas complète. Idéalement, nous aurions dû obtenir 25 000 (5 * 5 000) configurations articulatoires consonantiques pour chaque voyelle, mais comme un certain nombre de configurations ne correspondent pas à des consonnes, la moyenne se situe plutôt aux alentours de 17 200. Cela correspond tout de même à plus de 120 000 configurations articulatoires pour chaque

consonne (au total pour les 7 voyelles).

Pour faciliter la sélection des syllabes, le dictionnaire consonantique est organisé de façon syllabique : chaque configuration articulatoire consonantique est associée à sa consonne mais également à la voyelle avec laquelle elle a été construite. En résumé, sélectionner une syllabe pour le maître consiste à sélectionner une configuration articulatoire consonantique dans le dictionnaire consonantique correspondant à cette syllabe et de récupérer dans un même temps la configuration articulatoire vocalique ayant servi à construire cette configuration articulatoire consonantique.

Pour finir, soulignons à nouveau un point essentiel de nos simulations : nous considérons que le maître effectue ses consonnes par un geste en quelque sorte « invariant » puisqu'utilisant un seul articulateur spécifique en plus de la mâchoire. La question que nous posons in fine est de savoir si l'agent est capable d'identifier cet invariant en apprenant des configurations articulatoires de même nature que celles du maître.

7.3.1.2 Élaboration de la transformation articulatori-acoustique

La production syllabique est, par la suite, envoyée dans l'environnement et reçue par l'agent sous la forme d'un signal acoustique. Pour réaliser cette transformation, nous utilisons le modèle articulatoire VLAM comme nous l'avions précédemment utilisé dans le chapitre 5 dans les simulations de COSMO. Comme les gestes moteurs du maître correspondent dans COSMO SylPhon à un ensemble de configurations articulatoires regroupées dans un dictionnaire consonantique et vocalique, nous calculons, avec VLAM, la résultante acoustique de chacune de ces configurations. Nous avons ainsi un dictionnaire de signaux sensoriels correspondant aux configurations articulatoires du maître. C'est ce dictionnaire qui est utilisé pour représenter les signaux sensoriels reçus par l'agent lors de son apprentissage. Durant chaque échange entre le maître et l'agent, une syllabe est sélectionnée et un signal sensoriel correspondant à cette syllabe est tiré dans le dictionnaire des signaux sensoriels. Une illustration de ce dictionnaire dans l'espace F1/F2 pour les voyelles et F2/F3 pour les consonnes est représentée Fig. 7.5.

7.3.2 L'apprentissage sensoriel

7.3.2.1 Description de l'apprentissage

L'apprentissage sensoriel concerne l'apprentissage de la branche auditive. Comme les autres apprentissages, il s'effectue à l'aide du maître. Celui-ci produit des syllabes qui sont reçues par l'agent sous forme de signaux acoustiques. Contrairement à COSMO, le maître ne fournit pas à l'agent les unités correspondant aux signaux acoustiques. L'agent apprenant doit les inférer, ce qui constitue la principale difficulté de cet apprentissage sensoriel.

Dans COSMO SylPhon, cet apprentissage est en réalité composé de trois sous-apprentissages sensoriels différents : un apprentissage sensoriel consonantique, un apprentissage sensoriel vocalique et un apprentissage sensoriel syllabique. Dans chaque cas, il s'agit d'apprendre les répertoires auditifs

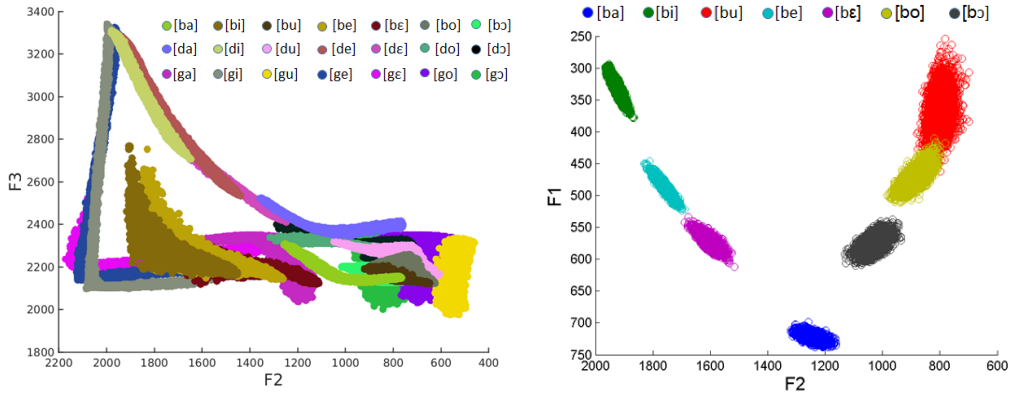


FIGURE 7.5 – Représentation des consonnes des syllabes (gauche) et des voyelles (droite) dans l'espace sensoriel, en Hz. Les syllabes sont affichées dans l'espace : F2 en abscisse et F3 en ordonnée. Les voyelles sont affichées dans l'espace : F2 en abscisse et F1 en ordonnée, les deux inversés de manière à faire apparaître le triangle vocalique

respectifs et les priors des noyaux auditifs associés. Pour simplifier l'explication, nous commençons par présenter l'apprentissage sensoriel de façon générale sans préciser s'il est consonantique, vocalique ou syllabique. Nous utilisons, pour cela, les variables génériques S pour les représentations auditives et N_L pour les noyaux.

D'un point de vue computationnel, l'apprentissage des répertoires auditifs $P(S \mid N_L)$ et des priors des objets auditifs $P(N_L)$ est équivalent à un apprentissage itératif d'une mixture de gaussiennes. En effet, en considérant les priors $P(N_L)$ comme les poids associés à chaque noyau gaussien, la mixture de gaussiennes correspondante s'écrit :

$$P(S) = \sum_{N_L} P(N_L)\ P(S \mid N_L)\ . \tag{7.5}$$

Les étapes de l'apprentissage sont les suivantes :

1. L'agent apprenant reçoit le signal sensoriel s correspondant à la syllabe choisie par le maître. Nous supposons que l'agent apprenant est capable de recevoir deux signaux sensoriels séparément : un correspondant à la partie consonantique de la syllabe et un correspondant à la partie vocalique de la syllabe $s =< s_c, s_o >$.
2. Il utilise le signal sensoriel pour inférer un noyau n pouvant correspondre à ce signal s en sélectionnant le noyau le plus probable de la distribution $P(N_L \mid [S = s])$.
3. Il met à jour les paramètres μ et Σ de la gaussienne du répertoire auditif $P(S \mid [N_L = n])$ avec le stimulus s perçu et le prior concerné $P(N_L)$ pour la valeur n.

À l'aide de ces étapes, détaillons plus précisément chaque apprentissage. Durant l'apprentissage vocalique, l'agent n'utilise que le signal s_o pour inférer un noyau n^o à partir de la distribution $P(N_L^O \mid [S_O = s_o])$. Il met ensuite à jour les paramètres de sa distribution $P(S_O \mid N_L^O)$ et de son prior $P(N_L^O)$. Durant l'apprentissage consonantique, l'agent n'utilise que le signal s_c pour inférer

un noyau n^c à partir de la distribution $P(N_L^C \mid [S_C = s_c])$. Il met ensuite à jour les paramètres de sa distribution $P(S_C \mid N_L^C)$ et de son prior $P(N_L^C)$. Enfin, durant l'apprentissage syllabique, l'agent utilise le signal sensoriel s de la syllabe pour inférer un noyau n^{syl} à partir de la distribution $P(N_L^{Syl} \mid [S_C^{Syl} = s_c]\ [S_O^{Syl} = s_o])$. Il met ensuite à jour les paramètres de sa distribution $P(S_C\ S_O \mid N_L^{Syl})$ et de son prior $P(N_L^{Syl})$.

Les inférences respectives calculées pour chaque apprentissage auditif sont :

$$P(N_L^O \mid [S_O = s_o]) \propto P(N_L^O)\ P([S_O = s_o] \mid N_L^O)\ , \tag{7.6}$$

$$P(N_L^C \mid [S_C = s_c]) \propto P(N_L^C)\ P([S_C = s_c] \mid N_L^C)\ , \tag{7.7}$$

$$P(N_L^{Syl} \mid [S_C^{Syl} = s_c]\ [S_O^{Syl} = s_o]) \propto P(N_L^{Syl})\ P([S_C^{Syl} = s_c]\ [S_O^{Syl} = s_o] \mid N_L^{Sy})\ . \tag{7.8}$$

7.3.2.2 Détails sur l'apprentissage

L'apprentissage sensoriel est réalisé avec cinq agents apprenants différents, pour assurer la robustesse des observations. Néanmoins, nous ne présentons par la suite que les résultats d'un unique agent. Les résultats sont similaires pour les autres.

Chaque sous-apprentissage sensoriel dure 100 000 itérations. Bien que différents nombres de noyaux ont été testés, nous montrons, par la suite, les résultats obtenus pour nb_{LO} (du prior $P(N_S^O)$) et nb_{LC} (du prior $P(N_S^C)$) à 20 et pour nb_{LSyl} (du prior $P(N_S^{Syl})$) à 60.

Nous effectuons un enregistrement des paramètres à certaines valeurs au cours de l'apprentissage, dix-neuf au total : beaucoup au début de l'apprentissage, puisque l'on suppose que l'agent varie beaucoup à ce moment là, et de moins en moins par la suite, puisqu'il est supposé que l'agent se stabilise et converge.

7.3.2.3 Analyse de l'apprentissage

Afin d'analyser la qualité de l'apprentissage, nous évaluons la distribution sensorielle de l'agent au cours des trois sous-apprentissages. Dans le modèle, cette distribution correspond à la mixture de gaussiennes $P(S)$ (respectivement vocalique $P(S_O)$, consonantique $P(S_C)$ et syllabique $P(S_C^{Syl}\ S_O^{Syl})$).

Dans un premier temps, nous comparons cette distribution $P(S)$ à celle de l'environnement. Cette dernière est calculée à partir du dictionnaire sensoriel de l'environnement (correspondant aux productions du maître) : l'ensemble des valeurs du dictionnaire est réparti dans un espace sensoriel similaire à celui de l'agent afin d'obtenir deux distributions comparables l'une à l'autre. La comparaison s'effectue par la suite à travers le calcul de la divergence de Kullblack-Leibler (KL divergence) entre ces deux distributions. Cette mesure, non symétrique, est définie par :

$$D_{KL}(P(A)||P(B)) = -\sum_i P(A(i))\ ln\left(\frac{P(B(i))}{P(A(i))}\right)\ . \tag{7.9}$$

où A et B correspondent aux deux espaces différents à comparer. Dans notre cas, nous calculons les deux divergences possibles : la KL divergence du maître par rapport à l'agent et celle de l'agent

par rapport au maître. Nous effectuons ensuite la moyenne des deux mesures afin d'obtenir la KL divergence symétrique moyenne :

$$Dmean_{KL}(P(A)||P(B)) = \frac{1}{2}D_{KL}(P(A)||P(B)) + \frac{1}{2}D_{KL}(P(B)||P(A)) \; . \tag{7.10}$$

Elle est représentée, pour nos trois apprentissages sur la Fig. 7.6.

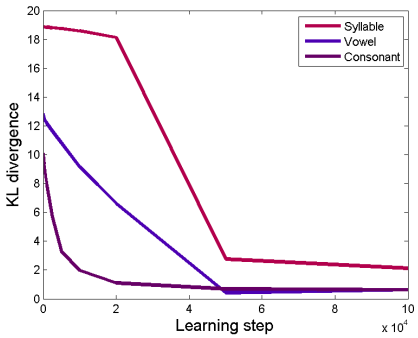


FIGURE 7.6 – KL divergence moyenne pour les trois sous-apprentissages sensoriels à différents moments de l'apprentissage entre 0 et 100 000 itérations

Afin d'interpréter cette mesure, rappelons que lorsque la KL divergence atteint 0 cela signifie que les deux distributions comparées sont identiques. Plus cette valeur est élevée, plus les distributions sont différentes. Nous observons donc que lors de l'apprentissage, la KL divergence des trois distributions diminue, ce qui signifie qu'elles ressemblent de plus en plus à celle du maître.

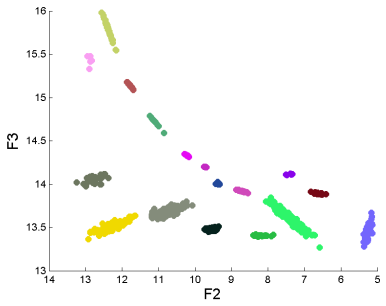
En termes de convergence, nous remarquons que les distributions phonémiques (vocalique et consonantique) convergent vers une valeur proche de 0. Ainsi, ces deux distributions semblent très similaires à celle du maître en fin d'apprentissage. La distribution syllabique, quant à elle, converge aux alentours de 2. Elle semble donc moins similaire à celle du maître que les distributions phonémiques. Néanmoins, nous observons que cette valeur semble encore légèrement diminuer au cours du temps, suggérant ainsi qu'elle peut correspondre davantage à celle de l'environnement si nous continuons l'apprentissage.

Cette observation nous amène à comparer les vitesses d'apprentissage. Pour commencer, nous notons que les trois distributions comparées n'ont pas la même valeur de KL divergence à l'initialisation, témoin des ressemblances initiales avec l'environnement. De plus, la distribution syllabique étant dans un espace à quatre dimensions (au lieu de deux pour les distributions phonémiques), il semble logique qu'elle ait la valeur la plus élevée en début d'apprentissage.

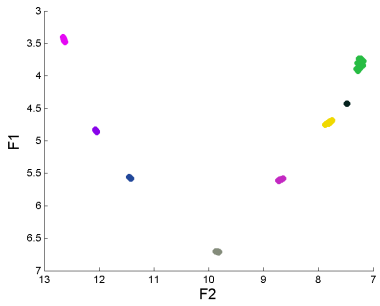
La distribution consonantique est celle qui converge le plus rapidement. Cela s'explique par le fait que la distribution consonantique du maître est plus étalée dans l'espace sensoriel et donc plus facile à apprendre par l'ensemble des noyaux. Au contraire, la distribution vocalique se trouve davantage concentrée dans des portions précises de l'espace, ce qui ralentit un peu la convergence. La distribution syllabique étant à 4 dimensions, elle est logiquement la plus longue des trois à converger.

Après avoir comparé l'évolution globale des apprentissages, nous nous intéressons à leur distributions en fin d'apprentissage et à la répartition des noyaux dans l'espace sensoriel. Pour cela, nous

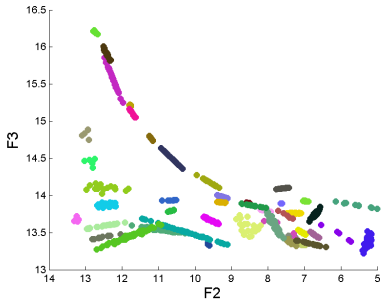
étudions les noyaux les plus représentatifs. La méthode est la suivante : nous commençons par tirer 50 points pour chaque catégorie phonétique du dictionnaire sensoriel syllabique. Ensuite, pour l'ensemble de ces points (en prenant soit le signal vocalique, soit le signal consonantique, soit le signal syllabique), nous calculons le noyau gaussien le plus probable dans chaque distribution sensorielle de l'agent (respectivement vocalique, consonantique et syllabique), c'est-à-dire le noyau le plus probable dans la distribution $P(N_L \mid [S = s])$, pour chacune des 50 données sensorielles s. Pour finir, nous effectuons un tirage de cinq points sur la distribution gaussienne associée au noyau sélectionné. Cela nous permet d'une part, d'avoir un aperçu des noyaux les plus représentatifs de l'environnement et d'autre part, d'avoir un aperçu de la variance des noyaux sélectionnés. Tout ceci est illustré, pour un agent, Fig. 7.7.



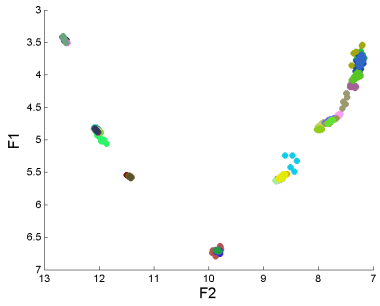
(a) Noyaux consonantiques



(b) Noyaux vocaliques



(c) Noyaux syllabiques, consonnes



(d) Noyaux syllabiques, voyelles

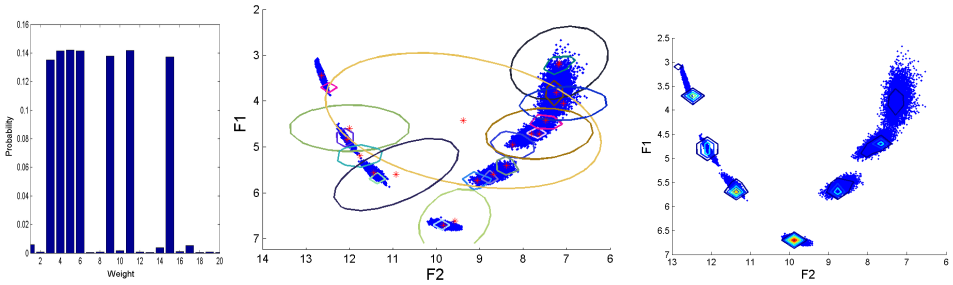
FIGURE 7.7 – Illustration des noyaux tirés pour un agent en fin d'apprentissage pour les données de l'environnement. Dans chaque figure, les points d'une même couleur correspondent à la même distribution gaussienne. Ceux des deux distributions syllabiques correspondent aussi à la même distribution gaussienne

Si nous comparons ces distributions à celles de l'environnement (voir Fig. 7.5), nous retrouvons une forme globale similaire, ce qui est cohérent avec les faibles valeurs de la KL divergence observées précédemment en fin d'apprentissage. En ce qui concerne les distributions phonémiques (consonantique, Fig. 7.7a, et vocalique, Fig. 7.7b), nous constatons que les noyaux sont répartis dans des portions bien spécifiques de l'espace et qu'il n'y a aucun chevauchement apparent. La variance des distributions gaussiennes vocaliques semble petite, ce qui augmente l'écart entre les noyaux gaussiens. Par ailleurs, nous remarquons que les noyaux gaussiens sont réparties globalement en suivant la répartition des sept voyelles de l'environnement, ce qui montre que l'agent a réparti ses noyaux gaussiens entre les voyelles. À l'inverse, les noyaux gaussiens consonantiques ne correspondent pas

chacun à une consonne. Plus frappant encore, les noyaux ne semblent pas correspondre à une portion des consonnes, ni même aux syllabes. En effet, il n'y a pas assez de noyaux pour représenter l'ensemble des syllabes et, de plus, un même noyau semble pouvoir correspondre à des consonnes différentes. Cela vient du fait que les consonnes sont très difficiles à apprendre dans l'espace sensoriel car elles correspondent à des domaines acoustiques complexes qui présentent même un certain niveau de recouvrement (voir Fig. 7.5, gauche).

De son côté, la distribution syllabique possède non seulement plus de noyaux représentatifs que les distributions phonémiques mais ceux-ci se superposent aussi bien dans l'espace consonantique que dans l'espace vocalique. L'analyse directe est donc plus difficile. Il semble y avoir plusieurs noyaux dans chaque portion de l'espace des voyelles et chacun d'eux correspond à une portion différente dans l'espace consonantique. La distribution vocalique pourrait donc faciliter le découpage de l'espace consonantique et permettre plus facilement de retrouver les syllabes. Dans une prochaine section, nous observerons si cela est suffisant pour retrouver entièrement les syllabes du maître.

Pour finir, pour chaque distribution sensorielle, nous avons analysé la répartition des poids des noyaux $P(N_L)$ et la répartition globale des noyaux gaussiens dans les espaces sensoriels à travers la distribution $P(S \mid N_L)$ en fin d'apprentissage. À titre d'illustration, nous représentons cela Fig. 7.8, pour les voyelles.



(a) $P(N_L^O)$ (b) $P(S_O \mid N_L^O)$ (c) $P(S_O)$

FIGURE 7.8 – Illustration des distributions sensorielles. (a) Répartition des noyaux gaussiens sous la forme d'un histogramme. (b) Répartition des distributions gaussiennes sous forme d'ellipses colorées dans l'espace sensoriel F1/F2, en Barks. (c) Mixture de gaussiennes correspondant à la distribution sensorielle de l'agent sous forme de courbes d'isoprobabilités. Pour ces deux dernières figures, les moyennes des distributions gaussiennes sont affichée sous la forme d'une étoile rouge et la distribution du maître est représentée, à titre de comparaison, sous la forme de points bleus.

De cette analyse, nous remarquons, dans un premier temps, que les noyaux ne sont pas tous appris. En effet, en fin d'apprentissage, comme l'illustre la Fig. 7.8a, seule une partie des noyaux a été mise à jour durant l'apprentissage et les autres ont une probabilité proche de zéro. Dans un second temps, nous observons que les noyaux les plus appris sont disposés dans les portions adéquates de l'environnement et possèdent une petite variance (voir, par exemple, Fig. 7.8c, les distributions centrées sur les données du maître). À l'inverse, les distributions non apprises possèdent une grande variance et sont, pour la plupart, en dehors des portions de l'espace de l'environnement.

7.3.3 Apprentissage sensorimoteur

7.3.3.1 Description de l'apprentissage

L'apprentissage sensorimoteur concerne l'apprentissage des modèles internes vocalique et consonantique. Du fait que le modèle interne consonantique dépend du modèle interne vocalique, nous supposons que l'apprentissage du modèle interne vocalique se fait avant celui du modèle interne consonantique. Dans les deux cas, ils s'apprennent, comme dans COSMO, par l'algorithme d'accommodation. Nous décrivons d'abord les étapes globale de cet algorithme avant d'expliquer les particularités de chacun d'eux. Nous utilisons les variables génériques S^M pour les représentations auditives et M pour les représentation motrices.

Les étapes sont les suivantes :

1. L'agent apprenant reçoit le signal sensoriel s correspondant à la syllabe choisie par le maître, décomposé en une partie consonantique s_c et une partie vocalique s_o.
2. Il sélectionne une représentation motrice m par tirage sur la distribution $P(M \mid [S^M = s])$.
3. Il produit la représentation motrice m sélectionnée et perçoit le stimuli s'. Celui-ci s'effectue avec les dictionnaires calculés à l'aide de VLAM, également utilisés pour définir les distributions du maître.
4. Il met à jour la moyenne et la matrice de covariance de sa distribution $P(S^M \mid [M = m])$ avec le paramètre s'.

Durant l'apprentissage sensorimoteur vocalique, l'agent n'utilise que le signal acoustique vocalique s_o fourni par le maître. Il infère ensuite une représentation motrice m_o à partir de la distribution $P(M_O \mid [S_O^M = s_o])$, la produit, puis met à jour les paramètres de sa distribution $P(S_O^M \mid M_O)$. De même, durant l'apprentissage sensorimoteur consonantique, l'agent n'utilise que le signal acoustique vocalique s_c fourni par le maître. Il infère ensuite une représentation motrice mc à partir de la distribution $P(M_C \mid [S_C^M = s_c])$, la produit, puis met à jour les paramètres de sa distribution $P(S_C^M \mid M_C)$.

Les inférences correspondantes aux deux apprentissages sont les suivantes :

$$P(M_O \mid [S_O^M = s_o]) \propto P(M_O)\, P([S_O^M = s_o] \mid M_O)\,, \tag{7.11}$$

$$P(M_C \mid [S_C^M = s_c]) \propto P(M_C)\, P([S_C^M = s_c] \mid M_C)\,. \tag{7.12}$$

7.3.3.2 Détails sur l'apprentissage

Cet apprentissage nous intéressant moins dans cette étude, il n'a été appris qu'une seule fois. Il est donc exactement le même pour chaque agent. Pour assurer un début de convergence, les deux sous-apprentissages sensorimoteurs durent 500 000 itérations.

Contrairement aux autres apprentissages, l'analyse se fait ici au cours du temps, après chaque itération. Ainsi, toutes les données ayant été générées lors de l'apprentissage sensorimoteur sont considérées dans l'analyse.

7.3.3.3 Analyse de l'apprentissage

Nous souhaitons vérifier que l'apprentissage sensorimoteur a bien convergé. Pour cela, nous l'évaluons sur sa capacité à bien reproduire les données de l'environnement. Ainsi, au cours de l'apprentissage, à chaque itération, nous comparons la donnée s envoyée par le maître à la donnée s' reproduite par l'agent. La comparaison s'effectue par le calcul de la distance euclidienne entre les deux points, en Barks. Pour avoir un aperçu globale de l'évolution temporelle des distributions, nous calculons, tous les 1 000 points, la moyenne des 1 000 dernières distance euclidiennes obtenues et nous la sauvegardons. L'ensemble de ces points, pour chaque apprentissage sensorimoteur, est représenté Fig. 7.9.

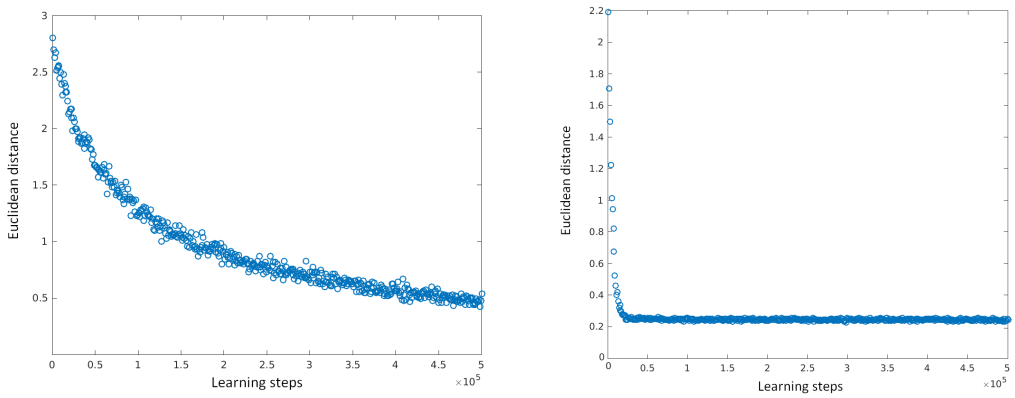


FIGURE 7.9 – Illustration de l'évolution de l'apprentissage sensorimoteur relatif aux consonnes (gauche) et aux voyelles (droite), en Barks

Le modèle interne vocalique converge aux alentours de 20 000 itérations avec un écart entre la distribution de l'agent et celle de l'environnement aux alentours de 0,2 Barks, ce qui est très faible. L'apprentissage vocalique est donc beaucoup plus rapide et plus précis que l'apprentissage sensorimoteur consonantique, dont la distance euclidienne calculée diminue lentement et ne semble pas avoir absolument convergé après 500 000 itérations. La distance euclidienne calculée en fin d'apprentissage est d'environ 0,5 Barks, ce qui est un peu élevé mais reste raisonnable (rappelons que l'espace F2 couvre 9 Barks et l'espace F3 en couvre environ 4). Outre les artéfacts liés à l'apprentissage, la distance euclidienne obtenue s'explique par la discrétisation réalisée. L'agent ne peut reproduire exactement une valeur de l'environnement du fait que nous découpons l'espace sensoriel en un ensemble relativement grossier de cases.

7.3.4 Apprentissage moteur

7.3.4.1 Description de l'apprentissage

L'apprentissage moteur concerne l'apprentissage de la branche motrice. Dans COSMO SylPhon, cela concerne l'apprentissage des répertoires moteurs et des priors sur les noyaux moteurs. Il est en ce sens très proche de l'apprentissage sensoriel. Il est également composé de trois sous-apprentissages :

un apprentissage moteur consonantique, un apprentissage moteur vocalique et un apprentissage moteur syllabique. Nous le présentons, comme pour l'apprentissage sensoriel, de façon globale en nous servant des variables génériques N_S pour les noyaux moteurs, S^M pour les représentations sensorielles, M pour les représentations motrices liées au représentations sensorielles, M_N pour les représentations motrices liées aux noyaux et λ_M pour la variable de cohérence liant les deux représentations motrices.

Comme l'agent apprend les répertoires moteurs $P(M_N \mid N_S)$ et leur prior associé $P(N_S)$, cet apprentissage peut être assimilé à l'apprentissage d'une mixture de gaussiennes tel que :

$$P(M_N) = \sum_{N_S} P(N_S)\ P(M_N \mid N_S)\ . \tag{7.13}$$

L'apprentissage moteur se fait par accommodation. Les étapes sont les suivantes :

1. L'agent apprenant reçoit le signal sensoriel s correspondant à la syllabe choisie par le maître, constitué du couple $< s_c, s_o >$.
2. Il sélectionne une représentation motrice m grâce à un tirage sur la distribution $P(M \mid [S^M = s])$ calculée par inférence.
3. L'agent infère le noyau n correspondant à la représentation motrice m en sélectionnant le noyau le plus probable dans $P(N_S \mid [M = m]\ [\lambda_M = 1])$.
4. L'agent met à jour la moyenne et la variance de sa distribution $P(M_N \mid [N_S = n])$ avec le paramètre m et le prior concerné $P(N_S)$ pour la valeur n.

Précisons maintenant les trois sous-apprentissages. Durant l'apprentissage vocalique, l'agent n'utilise que le signal s_o pour inférer une représentation motrice m_o à partir de la distribution $P(M_O \mid [S_O^M = s_o])$. Ensuite, il infère un noyau n^o, par inférence sur $P(N_S^O \mid [M_O = m_o]\ [\lambda_{MO}^O = 1])$. Pour finir, il met à jour les paramètres de sa distribution $P(M_{NO} \mid N_S^O)$ et de son prior $P(N_S^O)$.

Durant l'apprentissage consonantique, l'agent utilise les deux signaux s_o et s_c. Pour faciliter les calculs, il infère d'abord une représentation motrice m_o à partir de la distribution $P(M_O \mid [S_O^M = s_o])$ puis un geste m_c à partir de la distribution $P(M_C \mid [S_C^M = s_c][M_O = m_o])$. Cela évite de devoir calculer $P(M_C \mid [S_C^M = s_c]\ [S_O^M = s_o])$, ce qui est computationnellement plus coûteux. Ensuite, il infère un noyau n^c à partir de la distribution $P(N_S^C \mid [M_C = m_c]\ [M_O = m_o]\ [\lambda_{MC}^C = 1]\ [\lambda_{MO}^O = 1])$. Pour finir, il met à jour les paramètres de sa distribution $P(\Delta M \mid N_S^C)$ et de son prior $P(N_S^C)$.

Durant l'apprentissage syllabique, l'agent utilise le signal sensoriel s de la syllabe pour inférer une représentation motrice $< m_c, m_o >$ à partir de la distribution $P(M_C\ M_O \mid [S_C^M = s_c]\ [S_O^M = s_o])$. Ensuite, il infère un noyau n^{syl} à partir de la distribution $P(N_S^{Syl} \mid [M_C = m_c]\ [M_O = m_o][\lambda_{MC}^{Syl} = 1]\ [\lambda_{MO}^{Syl} = 1])$. Pour finir, il met à jour les paramètres de sa distribution $P(\Delta M^{Syl}\ M_{NO}^{Syl} \mid N_S^{Syl})$ et de son prior $P(N_S^{Syl})$.

Les inférences correspondantes, non déjà explicitées, sont les suivantes :

$$P(M_C\ M_O \mid [S_C^M = s_c]\ [S_O^M = s_o]) \tag{7.14}$$
$$\propto\ P(M_O)\ P([S_O^M = s_o] \mid M_O)\ P(M_C \mid M_O])\ P([S_C^M = s_c] \mid M_C)\ ,$$
$$P(N_S^O \mid [M_O = m_o]\ [\lambda_{MO}^O = 1]) \tag{7.15}$$
$$\propto\ P(N_S^O)\ P([M_{NO} = m_o] \mid N_S^O)\ ,$$
$$P(N_S^C \mid [M_C = m_c]\ [M_O = m_o]\ [\lambda_{MC}^C = 1]\ [\lambda_{MO}^O = 1]) \tag{7.16}$$
$$\propto\ P(N_S^C)\ P([\Delta M = m_c - m_o] \mid N_S^C)\ ,$$
$$P(N_S^{Syl} \mid [M_C = m_c]\ [M_O = m_o]\ [\lambda_{MC}^{Syl} = 1]\ [\lambda_{MO}^{Syl} = 1]) \tag{7.17}$$
$$\propto\ (N_S^{Syl})\ P([M_{NO} = m_o]^{Syl}\ [\Delta M^{Syl} = m_c - m_o] \mid N_S^{Syl})\ .$$

7.3.4.2 Détails sur l'apprentissage

Dans ses détails d'implémentation, l'apprentissage moteur est assez proche de l'apprentissage sensoriel. Il est donc réalisé lui-aussi avec cinq agents apprenants différents mais nous ne présentons que les résultats d'un unique agent.

L'apprentissage moteur est réalisé après l'apprentissage sensorimoteur. Chaque étape de l'apprentissage moteur étant assez longue computationnellement, chaque sous-apprentissage ne dure que 50 000 itérations. Nous illustrons les résultats obtenus pour nb_{SO} (du prior $P(N_S^O)$) et nb_{SC} (du prior $P(N_S^C)$) à 50 et pour nb_{SSyl} (du prior $P(N_S^{Syl})$) à 60.

Nous effectuons un enregistrement des paramètres à certaines valeurs au cours de l'apprentissage, dix-sept au total : beaucoup au début de l'apprentissage et de moins en moins par la suite.

7.3.4.3 Analyse de l'apprentissage

Afin d'analyser la qualité d'apprentissage des trois sous-apprentissages moteurs, nous souhaitons comparer les distributions $P(M)$ de l'agent avec celles de l'environnement. Cependant, l'apprentissage moteur permet d'apprendre des distributions motrices et non pas des distributions sensorielles. Nous étudions donc, non pas les distributions motrices directement, mais leur transformation auditive après production. Pour réaliser cette transformation, nous utilisons les dictionnaires calculés avec VLAM. Nous obtenons ainsi trois distributions $P(S)$ que nous comparons avec la distribution de l'environnement.

Pour commencer, nous calculons, comme lors de l'apprentissage sensoriel, la KL divergence moyenne (voir Eq. 7.10) entre la distribution $P(S)$ de l'agent et celle de l'environnement. Celle-ci est représentée Fig. 7.10.

Nous observons que les trois sous-apprentissages convergent tous très rapidement, en moins de 1 000 itérations. Cela montre que grâce à l'apprentissage sensorimoteur, déjà appris, les distributions motrices peuvent se stabiliser très rapidement. Cependant, elles conservent une erreur non né-

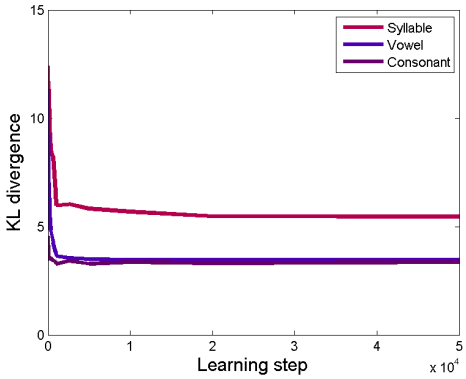


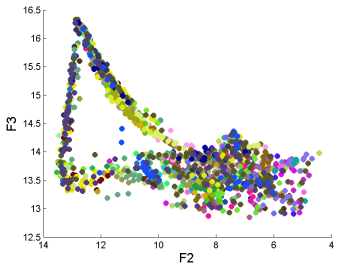
FIGURE 7.10 – Évolution de la KL divergence moyenne au cours du temps entre la distribution motrice de l'agent après production et la distribution sensorielle de l'environnement

gligeable : environ 4 pour les phonèmes et environ 8 pour les syllabes, cette différence pouvant s'expliquer par la différence de dimensions entre les espaces phonémiques et syllabiques. Ainsi, même si l'apprentissage se fait très rapidement, les distributions motrices convergent vers une distribution qui ne semble pas totalement similaire à celle de l'environnement.

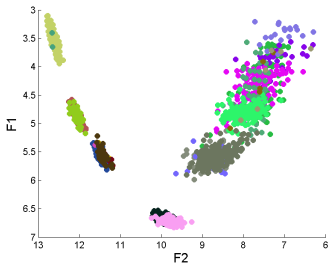
Pour mieux appréhender si ces distributions permettent de reproduire les données de l'environnement, nous les examinons dans l'espace sensoriel en fin d'apprentissage. Afin de ne pas observer uniquement la distribution globale mais aussi les distributions gaussiennes composant cette distribution, nous calculons, comme lors de l'analyse de l'apprentissage sensoriel, les noyaux gaussiens les plus représentatifs des données de l'environnement. Pour ce faire, nous inférons d'abord, à l'aide des modèles internes, les représentations motrices les plus probables correspondant aux données sensorielles de l'environnement. Ensuite, nous déterminons les noyaux moteurs les plus probables pour chaque représentation motrice. Enfin, nous produisons cette représentation motrice, ce qui nous donne un point dans l'espace sensoriel. L'ensemble des points obtenus pour un agent pour chacun des sous-apprentissage est illustré Fig. 7.11.

Globalement, ces figures illustrent le fait que les distributions de l'environnement sont, dans leur forme, bien reproduites (voir Fig. 7.5 pour comparaison) et que cela nécessite l'utilisation d'un nombre de noyaux assez conséquent. Par ailleurs, la distribution phonémique semble une nouvelle fois regrouper ces noyaux selon les voyelles de l'environnement. Néanmoins, nous remarquons que plusiseurs noyaux semblent utilisés pour la même catégorie. Il est plus difficile de juger si c'est le cas pour les branches consonantique et syllabique car les distributions gaussiennes se chevauchent dans la plupart des portions de l'espace sensoriel.

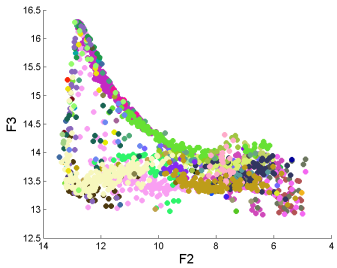
Nous souhaitons voir si ce chevauchement observé dans l'espace sensoriel se retrouve également dans l'espace moteur, notamment pour les consonnes. Pour cela, nous analysons les distributions gaussiennes directement dans l'espace moteur ΔM. Elles sont difficile à illustrer du fait que l'espace consonantique est à cinq dimensions. C'est pourquoi nous projetons cet espace dans des plans, dans lesquels nous affichons les distributions gaussiennes dont les noyaux sont les plus probables (de probabilité supérieure à 0,01). À titre d'exemple, nous montrons Fig. 7.12 le résultat obtenu pour un agent pour les dimensions TD, LH et $Apex$.



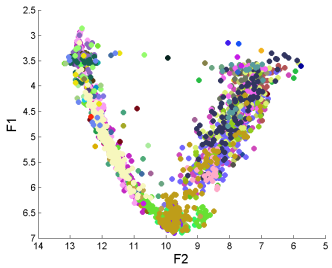
(a) Noyaux consonantiques



(b) Noyaux vocaliques



(c) Noyaux syllabiques, consonnes



(d) Noyaux syllabiques, voyelles

FIGURE 7.11 – Illustration des noyaux obtenus pour un agent en fin d'apprentissage correspondant aux données de l'environnement. Dans chaque figure, les points d'une même couleur correspondent à la même distribution gaussienne. Ceux des deux distributions syllabiques correspondent aussi à la même distribution gaussienne

Le choix des trois dimensions TD, LH et $Apex$ n'est pas anodin puisque nous attendons que ce soit dans ces dimensions qu'apparaissent les invariants consonantiques. En effet, pour rappel, l'hypothèse de départ est que la plosive [b] est associée à un mouvement des lèvres (LH), la plosive [d] à un mouvement de la pointe de la langue ($Apex$) et la plosive [g] à un mouvement du dos de la langue (TD). Nous avons initialisé les distributions gaussiennes en position de repos (0), de façon à ce qu'elles aient une grande variance sur la dimension Jaw et parmi une des dimensions TB, TD, LH ou $Apex$ et une petite variance sur les autres dimensions. Nous souhaitons en ce sens influencer l'agent apprenant, en début d'apprentissage, à ne bouger qu'un unique articulateur, en plus de la mâchoire, et par la suite, évaluer sa capacité à maintenir ce fonctionnement pour reproduire les données du maître

La Fig. 7.12 nous permet de faire deux observations. Premièrement, les noyaux ont conservé le bootstrap initial : ils ne présentent une grande variance que sur une des trois dimensions observées. Les distributions ont une petite variance et sont centrées sur 0, c'est-à-dire en position de repos, sur les autres dimensions. Ainsi, l'agent semble avoir réussi à apprendre les invariants consonantiques comme nous le souhaitions. Deuxièmement, nous observons que chaque dimension possède plusieurs noyaux, ce qui signifie que pour un agent, une même consonne est représentée par plusieurs noyaux. Le point remarquable est que, bien que les distributions gaussiennes consonantiques de l'agent se chevauchent dans l'espace sensoriel, les distributions sont séparables dans l'espace moteur.

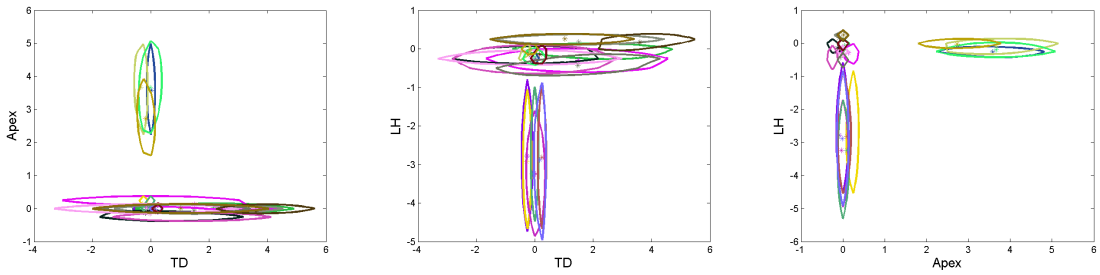


FIGURE 7.12 – Illustration des distributions gaussiennes consonantiques dans les dimensions TD, LH et $Apex$ de l'espace moteur ΔM. Chaque ellipse d'une même couleur représente la même distribution gaussienne

7.4 Communication avec le maître

Nous avons décrit et analysé l'apprentissage de l'agent. En fin d'apprentissage, l'agent semble conserver un certain nombre de noyaux lui permettant de représenter la distribution des stimuli de l'environnement. Si certains de ces noyaux semblent correspondre aux catégories phonétiques du maître, comme les voyelles, ceci reste moins clair pour les consonnes et les syllabes. Puisque rien, dans l'apprentissage, ne permet à l'agent apprenant de converger vers un code identique à celui du maître, nous ne pouvons espérer « voir » apparaitre explicitement les phonèmes ou les syllabes dans les distributions apprises. En revanche, même si l'agent converge vers un codage différent de celui du maitre, cela ne l'empêche pas forcément de résoudre des tâches de communication. Nous vérifions donc dans cette section que le maître et l'agent peuvent communiquer.

7.4.1 Communication via la branche sensorielle

Sachant que l'agent n'a pas les mêmes objets que le maître, il est difficile de vérifier qu'ils se comprennent de façon directe. Dans la branche sensorielle, un des moyens d'analyser cette communication est d'inférer le noyau correspondant à une catégorie phonétique prononcée par le maître, de générer un signal avec le noyau choisi et de vérifier que le signal résultant est bien compris par le maître comme appartenant à la même catégorie que celle qu'il avait sélectionnée initialement. Dans la suite, nous appellons « retranscription » l'ensemble de cette boucle maître, agent, maître.

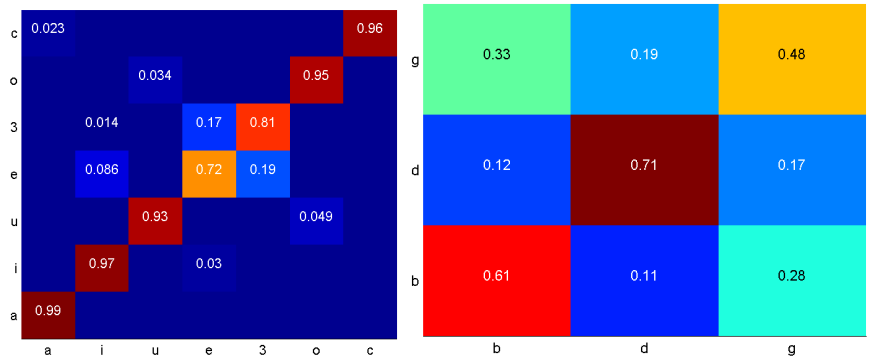
En termes computationnels, cela consiste à calculer la matrice de confusion $P(O^R \mid O^E)$ où O^E (E pour « Envoyé ») correspond à la catégorie phonétique envoyée par le maître[2] et O^R (R pour « Reçu ») à l'objet interprété, en fin de boucle, par le maître. En supposant S, la notation générique pour les représentations sensorielles et N_L celle pour les noyaux de l'agent, cette distribution se cal-

2. Pour rappel, contrairement à l'agent, le maître a des objets O correspondant chacun à des catégories phonétiques.

cule par :

$$P(O^R \mid O^E) \propto \sum_{S^R,S^E,N_L} P(O^R \mid S^R)\, P(S^R \mid N_L)\, P(N_L \mid S^E)\, P(S^E \mid O^E)\,. \tag{7.18}$$

Dans cette équation, $P(S^R \mid N_L)$ correspond à un des répertoires auditifs de l'agent apprenant (respectivement vocalique, consonantique et syllabique) et $P(N_L \mid S^E)$ se calcule de la même manière qu'aux Eq. 7.6 (vocalique), 7.7 (consonantique) et 7.8 (syllabique). Les matrices observées sont représentées Fig. 7.13 pour les phonèmes et Fig. 7.14 pour les syllabes. Nous calculons pour chacune d'elle une mesure de performance, traduisant les situations de communication correcte, c'est-à-dire où le maitre reconnaît le même objet qu'il avait fourni initialement. Cela correspond à la moyenne des valeurs de la diagonale de la matrice.



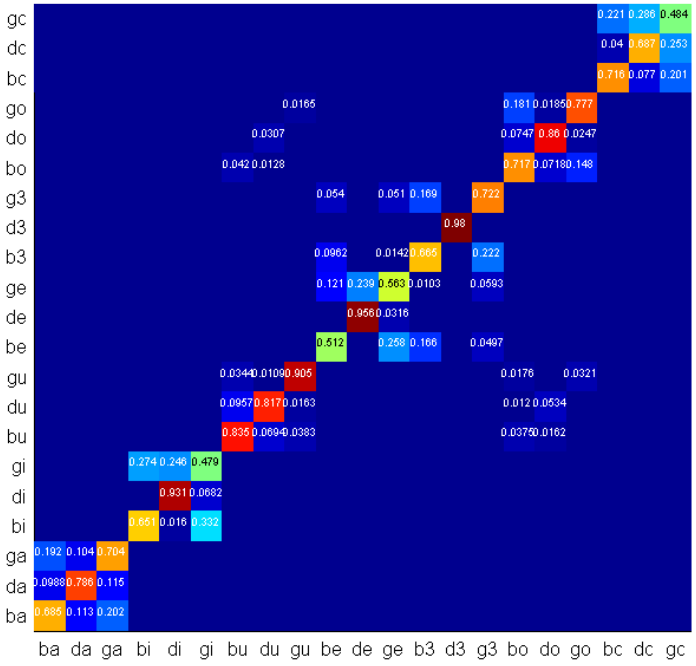
(a) Voyelles, performance = 0.90 (b) Consonnes, performance = 0.60

FIGURE 7.13 – Illustration des matrices de confusion phonémiques. Seules les valeurs au dessus de 0,01 sont notées

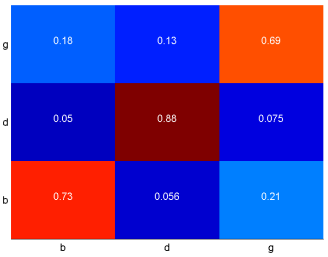
En ce qui concerne les phonèmes (Fig. 7.13), nous remarquons que le maître, après « rebond sur l'agent », reconnaît globalement très bien les voyelles (score de performance supérieur à 0,90). Néanmoins, les voyelles [e] et [ɛ] se confondent parfois. Cela signifie que globalement les noyaux de l'agent correspondent à une unique voyelle, ce qui est conforme à l'observation réalisée sur les noyaux vocaliques (voir Fig. 7.7d et Fig. 7.5 pour comparaison). Outre le fait que les noyaux vocaliques les plus représentatifs sont très séparables dans l'espace sensoriel, nous avons vu que les distributions gaussiennes associées semblent avoir une petite variance. Il est donc possible que l'agent choisisse un mauvais noyau lorsque le maître fournit à l'agent des données sensorielles éloignées des prototypes des catégories, ce qui donne une mauvaise retranscription.

De leur côté, les consonnes sont moins bien retranscrites que les voyelles (performance globale de 0,60). Même si les bonnes consonnes sont le plus souvent reconnues (la probabilité du hasard est de 0,33), il y a de nombreuses confusions. Cela s'explique par le fait que les noyaux consonantiques de l'agent, bien qu'ils décrivent correctement la distribution globale de l'environnement, ne sont pas toujours associés à une unique catégorie consonantique (voir Fig. 7.7a et Fig. 7.5 pour comparaison).

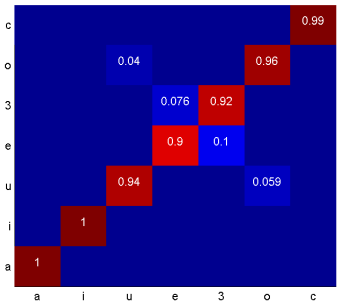
En ce qui concerne les syllabes (Fig. 7.14), nous remarquons qu'elles sont globalement moins



(a) Syllabes CV, performance = 0.73



(b) C, performance = 0.76



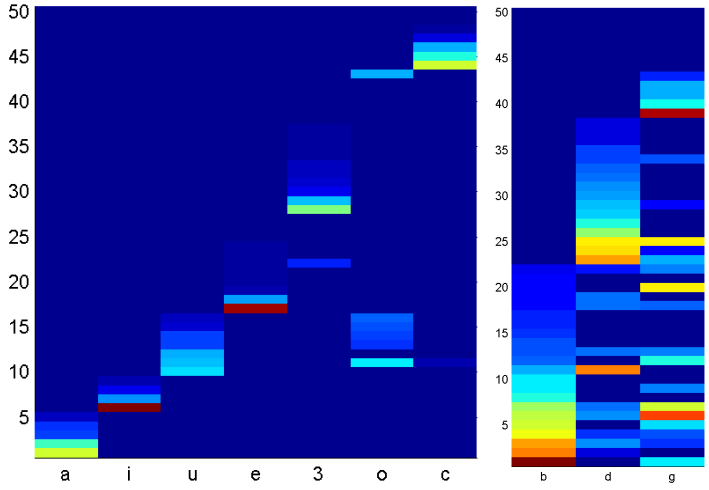
(c) V, performance = 0.96

FIGURE 7.14 – Illustration des matrices de confusion syllabiques globales (gauche) et regroupées par consonnes (droite, haut) ou par voyelles (droite, bas). Seules les valeurs au dessus de 0,01 sont notées

bien retranscrites que les voyelles mais mieux retranscrites que les consonnes. Néanmoins, après regroupement par phonèmes (consonnes d'une part et voyelles d'autre part), nous remarquons que la performance syllabique est plus élevée que la performance phonémique (score de 0,76 au lieu de 0,60 pour les consonnes et score de 0,96 au lieu de 0,90 pour les voyelles). Ainsi, l'apprentissage syllabique, apprenant les consonnes et les voyelles de manière conjointe, dans un même espace, semble conduire à de meilleures performances.

7.4.2 Communication via la branche motrice

Nous souhaitons vérifier la communication de l'agent avec le maître également dans la branche motrice. Une retranscription telle que celle effectuée pour la branche sensorielle demande d'inférer les noyaux moteurs correspondant aux sons du maître et de les reproduire. La production des noyaux pose quelques problèmes, notamment dans le cas consonantique. En effet, même si nous inférons un invariant consonantique, celui-ci ne peut être produit dans l'espace sensoriel sans voyelle associée. Or, les essais réalisés montrent que la voyelle est primordiale lors de la retranscription et que l'utilisation d'une voyelle aléatoire (par exemple, celle envoyée par le maître) produit de mauvaises transcriptions consonantiques. De ce fait, nous ne passons pas, cette fois-ci, par un système de retranscription et nous comparons directement les catégories du maître avec les noyaux de l'agent. Les matrices correspondantes sont illustrées Fig. 7.15 pour les phonèmes et Fig. 7.16 pour les syllabes.



(a) Voyelles (b) Consonnes

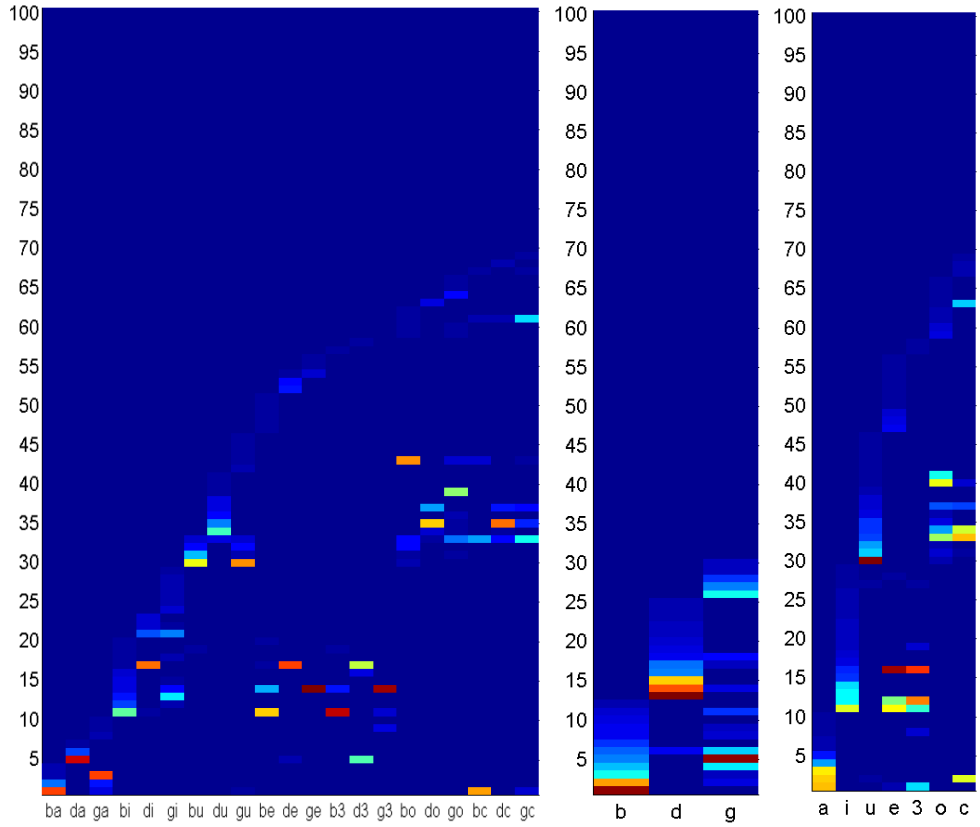
FIGURE 7.15 – Illustration des matrices maîtres/agents phonémiques : voyelles (gauche) et consonnes (droite). Les couleurs vont du marron (forte probabilité) au bleu foncé (faible probabilité, inférieure à 0,01). Les noyaux sont triés selon leur probabilité, pour chaque catégorie phonétique, de gauche à droite

De manière générale, nous remarquons que les catégories phonétiques sont toutes représentées par plusieurs noyaux. Cependant, nous notons également que tous les noyaux ne sont pas associés à des catégories phonétiques spécifiques : certains ont des probabilités faibles pour l'ensembles des catégories.

Focalisons-nous sur les phonèmes (Fig.7.15). Du côté des voyelles, nous observons que chaque catégorie phonétique est associée à un groupe de noyaux spécifique, sauf les catégories [u] et [o] qui semblent utiliser des noyaux communs. Cela suggère que malgré quelques confusions, l'agent arrive à la capacité d'apprendre les catégories vocaliques et de communiquer avec lui. Du côté des consonnes, le comportement des noyaux est assez diversifié. Si certains noyaux semblent être utilisé pour caractériser l'ensemble des plosives, d'autres sont plus spécifiques à une catégorie consonantique particulière. Ainsi, même si certains noyaux sont communs à plusieurs catégories, chaque consonne semble avoir des noyaux qui lui sont spécifiques. Du fait de ces différents types de noyaux, savoir si l'agent peut communiquer avec son maître n'est pas clairement décidable. De plus, les consonnes n'étant pas prononçables de façon isolée, il faudrait vérifier si les noyaux permettent de produire différentes voyelles.

Passons maintenant aux syllabes (Fig. 7.16). Nous présentons d'abord les syllabes dans leur ensemble avant de les regrouper par voyelles et par consonnes. Nous remarquons, comme pour les phonèmes, que si certaines syllabes ont des catégories qui leurs sont propres, d'autres sont partagées entre plusieurs catégories (voir les portions entourées en jaune). Par exemple, le noyau 30 est partagé entre [bu] et [gu], ce qui est cohérent avec la proximité acoustique de ces deux catégories de syllabes.

En regroupant par phonèmes les unités syllabiques, nous observons que les noyaux semblent davantage correspondre à une unique catégorie pour les consonnes que dans le modèle phonémique. En



(a) Syllabes CV (b) Consonnes C (c) Voyelles V

FIGURE 7.16 – Illustration des matrices maîtres/agents syllabiques (gauche), regroupées par consonnes (milieu) et regroupées par voyelles (droite). Les couleurs vont du marron (forte probabilité) au bleu foncé (faible probabilité, inférieure à 0.01). Les noyaux ont été triés selon leur probabilité pour chaque catégorie phonétique, de gauche à droite

revanche, il apparaît davantage de confusion pour les voyelles : les mêmes noyaux sont utilisés pour les voyelles [o] et [ɔ] ainsi que pour les voyelles [e] et [ɛ]. Ces mélanges risque de se ressentir dans la communication.

7.5 Discussion générale

7.5.1 Synthèse

Nous nous sommes fixés comme objectif d'analyser la convergence du modèle et la communication avec le maître et ce, sans que l'agent n'ait de connaissance sur le nombre de ses catégories, sans apprentissage supervisé des catégories et sans invariant consonantique explicite.

En termes de convergence, les études réalisées en calculant la KL divergence moyenne entre le maître et l'agent montrent que la branche auditive, calculée à l'aide des priors et des répertoires audi-

tifs, converge globalement vers des distributions sensorielles proches de celles des stimuli de l'environnement. L'agent apprenant conserve néanmoins des erreurs résiduelles en fin d'apprentissage. La branche motrice, calculée à l'aide des priors et des répertoires moteurs après transformation des représentations motrices en signal sensoriel, converge vers des distributions moins ressemblantes à celles des stimuli de l'environnement. Ainsi, la branche auditive est plus précise pour reproduire les données environnementales que la branche motrice, ce qui est cohérent avec nos études précédentes. De la même manière, les branches phonémiques apparaissent également plus ressemblantes aux données de l'environnement que la branche syllabique.

En termes de communication, nous observons que, globalement, l'agent arrive à communiquer avec son maître et ce, sans avoir de catégories phonétiques similaires à celles du maître. Il reste quelques confusions entre certaines catégories phonétiques mais la communication générale paraît effective. Nos deux objectifs semblent donc remplis.

7.5.2 Le cas de l'invariant consonantique

Comme nous l'avons précédemment décrit, l'invariant consonantique a nécessité quelques réflexions préalables. Le principe des simulations effectuées consistait à supposer, d'une part, l'existence de gestes prototypiques caractérisés par un invariant consonantique et, d'autre part, à tester si l'existence d'une initialisation utilisant le même type de propriétés motrices dans le cadre de la théorie Frame-then-Content pouvait conduire l'agent à récupérer des gestes aux invariants consonantiques similaires.

Les résultats montrent que cette initialisation est suffisante pour faire apprendre à l'agent des gestes moteurs contenant les mêmes formes d'invariance prototypique que celles du maître. Cela suggère que les propriétés d'invariance n'ont pas besoin d'être explicitement implémentées chez un agent mais qu'elles peuvent résulter de contraintes d'apprentissage.

7.5.3 Comparaison entre les noyaux et les unités distinctives

Un fait assez remarquable de cette étude est que l'agent semble globalement capable de communiquer avec son maître alors qu'il ne possède pas des noyaux similaires aux catégories phonétiques du maître. Dit autrement, deux agents semblent pouvoir communiquer sans posséder la même structure cognitive interne.

Bien entendu, cela ne s'effectue pas sous n'importe quelle condition. Nous remarquons, globalement, qu'une catégorie du maître correspond à plusieurs noyaux chez l'agent et qu'un noyau chez celui-ci ne correspond qu'à une catégorie phonétique du maître. Il existe donc bien un lien entre les noyaux et les unités distinctives mais celui-ci n'est pas bijectif. Cela s'explique par le fait que les noyaux correspondent à des gaussiennes dans l'espace auditif et moteur alors que les catégories phonétiques n'ont globalement pas une forme gaussienne. Il faut donc plusieurs noyaux, placés chacun sur une portion de l'espace de l'unité distinctive, pour la représenter dans son ensemble. Dans nos expériences précédentes, le nombre de gaussiennes étant posé a priori égal au nombre d'unités, il ne

pouvait y avoir qu'une gaussienne par unité distinctive, et donc, chacune de ces gaussiennes devait approximer la distribution non-gaussienne des signaux sensoriels ou moteurs. Ici, avec possiblement plus d'un noyau gaussien par unité distinctive, nous améliorons l'approximation mais perdons la bijection entre noyaux et unités. Néanmoins, nous remarquons que tous les noyaux ne sont pas utilisés pour réaliser l'apprentissage. L'agent ne se focalise que sur l'apprentissage d'un petit groupe de noyaux et laisse les autres dégénérer vers des probabilités très faibles. Nous pourrions supposer qu'à terme, les noyaux non utilisés ne soient pas conservés.

Une question se pose : si les noyaux ne correspondent pas exactement aux catégories phonétiques, comment un agent peut-il les reconnaître ? Il y a plusieurs hypothèses possibles. Nous en proposons deux : soit ses noyaux correspondent à des catégories phonétiques à part entière, soit les noyaux répondant à la même unité distinctive peuvent se regrouper pour former une unique catégorie.

Si ces noyaux correspondent à des catégories à part entière, cela signifie que l'agent a un niveau de précision sur ces catégories beaucoup plus élevé que celui du maître. Cependant, il n'y a pas vraiment de raison de penser qu'un agent apprendrait des catégories plus précises que celles de son maître. En revanche, sans pour autant être des catégories, ses noyaux pourraient être des allophones : ils correspondraient alors à différents éléments d'une même catégorie. Cela nous amène à la seconde hypothèse suggérant que les noyaux se regroupent autour d'une même catégorie phonétique. Nous pourrions alors imaginer l'existence d'un processus de plus haut niveau cognitif permettant de regrouper les noyaux associés à une catégorie.

Cognitivement, la pré-catégorisation des phonèmes apparaît comme un résultat acceptable pour la production de parole. La « régularisation » des phonèmes, c'est-à-dire le regroupement des noyaux proches en une même catégorie phonémique explicite ne serait nécessaire et utile que plus tard dans le développement, par exemple, lors de l'apprentissage de la lecture. Cela pourrait expliquer les différences de compétences phonémiques entre les lettrés et les illettrés (Morais et al., 1979).

7.5.4 Comparaison entre les branches auditives et les branches motrices

Nous avons pour le moment effectué des apprentissages sensoriel et moteur séparés. De ce fait, nous observons que les noyaux utilisés pour l'apprentissage moteur ne correspondent pas forcément à ceux utilisés pour l'apprentissage sensoriel. En nous basant sur les études sur les corrélations entre perception et production (voir section 3.1.2 du chapitre 2), il est peu probable que l'agent conserve une telle disparité entre ces deux branches.

Il serait intéressant, à terme, de combiner les noyaux des deux apprentissages ou de les apprendre de manière simultanée, par exemple grâce à la variable de cohérence C. Pour faire le lien avec la section précédente, il est possible que cette combinaison fasse apparaître plus clairement les catégories phonétiques et facilite la communication avec le maître. Si cette supposition s'avère vrai, la pré-catégorisation observée dans les deux branches actuellement, donnerait lieu, après fusion, à des unités distinctives perceptuo-motrices.

Par ailleurs, en comparant les apprentissages auditifs et moteurs, nous remarquons que les phonèmes et les syllabes sont globalement bien appris après les deux apprentissages. Néanmoins, en

examinant de manière plus précise ces résultats, nous observons des tendances similaires à celles obtenues par Laurent et al. (2017) : les voyelles sont mieux caractérisées dans l'espace auditif que les consonnes (voir Fig. 7.16) et, inversement, les consonnes sont mieux caractérisées dans l'espace moteur que les voyelles.

7.5.5 Comparaison entre l'apprentissage phonémique et syllabique

L'apprentissage phonémique semble plus complexe que l'apprentissage syllabique, notamment l'apprentissage consonantique. En effet, du fait de sa dépendance à la voyelle, l'apprentissage consonantique nécessite constamment l'utilisation des représentations vocaliques lors de l'apprentissage. En termes computationnel, il semble donc plus simple d'apprendre les syllabes.

Par ailleurs, les performances en fin d'apprentissage semblent montrer que l'apprentissage syllabique permet de mieux communiquer avec le maître. En effet, les résultats obtenus dans la branche sensorielle montrent que les consonnes et voyelles issues de l'apprentissage syllabique sont mieux apprises que lors de l'apprentissage phonémiques.

Sans résoudre le débat phonème/syllabes, ces deux premiers constats viennent affermir la théorie supposant que les syllabes sont les unités principales du développement de la parole. Bien entendu, des analyses complémentaires seraient nécessaires pour confirmer cette tendance. Une perspective possible pour cette étude serait de tenter de reproduire les résultats obtenus dans certaines études de la littérature. Par exemple, celle de Jusczyk et Derrah (1987) propose une familiarisation des bébés à des syllabes CV partageant la même consonne C. Une phase de test avec des syllabes différant soit sur la consonne C, soit sur la voyelle V, soit sur les deux phonèmes CV, montre que les bébés se déshabituent dans les trois conditions, suggérant ainsi qu'ils ne semblent pas caractériser la consonne initiale des CV. En comparant les noyaux obtenus avec COSMO SylPhon pour les différentes phases de l'expérimentation, on pourrait considérer qu'un agent se déshabitue s'il choisit un noyau différent dans la phase de test par rapport à la phase de familiarisation. Selon ces hypothèses, une prédiction serait que l'apprentissage syllabique reproduit mieux les résultats de cette étude que l'apprentissage vocalique. Plus globalement, l'étude des dynamiques d'apprentissage des différentes branches, sensorielle vs. motrice et phonémique vs. syllabique, de COSMO SylPhon devrait fournir un très riche cadre d'analyse des données sur le développement des compétences perceptives chez le bébé et le jeune enfant.

7.5.6 Conclusion

Cette étude n'est pour le moment pas complètement aboutie et nécessite des approfondissements. Néanmoins, d'ores et déjà, COSMO SylPhon offre un cadre riche permettant d'associer représentations sensorielles et motrices et structures phonémiques et syllabiques. Cette étude nous semble ouvrir de nombreuses pistes de réflexion et de travail visant à consolider et développer les différentes analyses que nous en avons faites : sur la nature des invariants phonémiques ; sur les rôles respectifs des syllabes et des phonèmes dans l'apprentissage et la communication ; sur la nature des processus de communication, et l'émergence des catégories au cours du développement.

CHAPITRE 8

Discussion

Cette thèse et sa rédaction ont été menées selon un axe bien particulier : la modélisation computationnelle de l'apprentissage de la phonétique cognitive. Cet axe est composé de trois domaines : la phonétique cognitive, le sujet principal de cette thèse, l'apprentissage, notre point de vue privilégié et la modélisation computationnelle, notre méthodologie. La première section de ce chapitre est dédiée à une synthèse et une discussion autour de cet axe.

Même si l'axe de cette thèse est resté globalement inchangé dans ce manuscrit, nous nous sommes permis d'élargir le modèle durant le déroulement de cette thèse et ce, en plusieurs occasions. D'abord, nous nous sommes intéressés à la viabilité cognitive de notre modèle computationnel COSMO du point de vue des neurosciences. Cela nous a permis de tenter de lier les composantes du modèle à ses potentiels corrélats neuroanatomiques dans le cerveau et surtout de décrire certains mécanismes du cerveau à l'aide de notre modèle. L'ensemble de cette étude est regroupé sous l'expression « COSMO Neuro ». Ensuite, nous nous sommes intéressés à l'influence de l'apprentissage lexical sur l'apprentissage phonétique. Cela nous a permis de développer une nouvelle version du modèle COSMO, nommée « COSMO WordPhon ». Enfin, nous avons décrit théoriquement le modèle COSMO afin qu'il soit capable de traiter un éventail plus large de spécificités phonétiques, notamment celles concernant la production phonétique ou la perception multimodale. Le modèle théorique résultant est appelé « COSMO multisensoriel ». Ces trois perspectives sont l'objet de la seconde section dédiée aux perspectives du modèle.

8.1 Synthèse

8.1.1 Synthèse globale

Après avoir introduit nos objectifs au chapitre 1, les chapitres 2 et 3 ont permis de rappeler l'état des connaissances sur les unités distinctives et sur les modèles computationnels associés, et de présenter le cadre computationnel du modèle COSMO. Nos études, réalisées avec ce modèle, ont été concentrées dans trois chapitres de ce manuscrit, les chapitres 4, 5 et 6, abordant respectivement trois aspects des unités distinctives de la parole : leurs caractérisations sensorielle et motrice, leur variabilité et leur contenu cognitif.

Le chapitre 4 contient deux études sur la caractérisation sensorielle et motrice réalisée avec le modèle COSMO, tel qu'il est décrit dans le chapitre 3. La première étude est consacrée à l'apprentissage des représentations sensorielles et motrices et à leur effet sur la perception. Inspirés par les théories de

la perception de la parole, nous analysons le rôle des branches auditive et motrice durant une tâche de perception. Nous montrons notamment que des différences non négligeables d'apprentissage résultent en l'apparition d'une complémentarité des deux branches en perception : la caractérisation auditive permet une perception précise et centrée sur les données apprises tandis que la caractérisation motrice possède un plus grand pouvoir de généralisation et semble capable de traiter plus efficacement des données bruitées. C'est ce que nous résumons sous la propriété « bande étroite/bande large ».

La deuxième étude est consacrée à l'apprentissage sensorimoteur. Nous comparons pour cela huit algorithmes d'apprentissage basés sur trois propriétés : le principe d'accommodation, qui s'appuie sur une exploration guidée socialement, le babillage idiosyncratique, qui favorise l'exploration des représentations motrices préalablement sélectionnées, et l'extrapolation locale, qui généralise l'apprentissage d'une représentation motrice aux représentations motrices similaires. Nous montrons que l'exploration sensorimotrice, longue et coûteuse sans stratégie préalable, est améliorée par ces trois principes.

Le chapitre 5 concerne également deux études réalisées avec le modèle COSMO mais, cette fois-ci, centrées sur la variabilité des représentations sensorielles et motrices à travers l'analyse des idiosyncrasies. La première étude est consacrée à l'apparition des idiosyncrasies. En supposant qu'un apprentissage moteur nécessite une phase d'imitation dans laquelle l'agent apprenant tente de trouver les représentations motrices correspondant à ce qu'il perçoit, nous observons qu'une imitation centrée sur les phonèmes reconnus permet l'apparition d'idiosyncrasies contrairement à une imitation centrée sur le signal sensoriel perçu. Nous en déduisons que l'apprentissage moteur semble à l'origine de ces idiosyncrasies et que cet apprentissage nécessite un objectif communicatif.

La deuxième étude nous permet d'analyser, à l'aide des idiosyncrasies, le lien entre les représentations sensorielles et motrices en perception et en production. En prenant comme objectif la reproduction des résultats de Ménard et Schwartz (2014), nous montrons que les corrélations entre les idiosyncrasies en perception et en production ne peuvent apparaître que si le modèle utilise sa composante motrice durant la tâche de perception, confirmant ainsi le rôle de la branche motrice en perception de la parole et soulignant l'apparente nature perceptuo-motrice des unités distinctives.

Le chapitre 6 nous permet d'aborder le contenu cognitif des unités phonétiques, c'est-à-dire la nature de leurs représentations cognitives sous-jacentes. Nous nous focalisons particulièrement sur les particularités des composantes vocalique, consonantique et syllabique de notre modèle et sur la manière dont s'apprennent les représentations sensorielles et motrices dans chacune de ces composantes. Ce chapitre nécessite une extension du modèle COSMO, que nous nommons COSMO SylPhon, associant notamment une structure syllabique et une structure phonémique. Nous étudions dans un même cadre l'apprentissage sensoriel et moteur de ces deux structures et comparons les spécificités de chacune. Nous observons principalement qu'un apprentissage sensoriel permet une bonne catégorisation vocalique et syllabique mais une mauvaise catégorisation consonantique tandis qu'un apprentissage moteur facilite davantage la catégorisation consonantique que vocalique. COSMO SylPhon étant un modèle plus complexe que son homologue générique, il nous permet également de nous confronter à plusieurs questions aussi bien d'ordre phonétique et théorique (par exemple, sur la représentation des consonnes) que computationnelles (par exemple, sur l'apprentissage itératif de mixtures de gaussiennes).

8.1.2 Synthèse des enjeux phonétiques

Comme le résume la section précédente, tout au long de ce manuscrit, nous nous focalisons sur trois aspects de la phonétique, à savoir : la caractérisation, la variabilité et le contenu cognitif des unités distinctives. Pour chacun de ces points, nous rappelons nos hypothèses de modélisation et nous discutons des limites des simulations effectuées et de quelques perspectives futures.

8.1.2.1 Caractérisation des unités distinctives

Nous discutons ici de notre implémentation des représentations internes des unités de la parole. La discussion s'effectue autour de deux points : les représentations utilisées dans notre modèle et leur composition.

Les représentations considérées Dans toutes les versions de COSMO, nous n'avons considéré que deux types de représentations des catégories phonétiques : des représentations sensorielles, principalement auditives et des représentations motrices. Comme nous l'avons vu dans le chapitre 2, ce sont les deux principales représentations permettant de caractériser les catégories phonétiques.

Les représentations sensorielles dans COSMO générique correspondent dans nos implémentations à des représentations auditives. Si elles sont assez peu détaillées dans COSMO 1D, elles deviennent plus précises dans COSMO-V et COSMO SylPhon dans lesquels elles correspondent à l'espace formantique. L'espace formantique est souvent utilisé en phonétique pour caractériser les propriétés acoustiques des catégories phonétiques. Cependant, les formants ne sont qu'une composante des représentations auditives. Ces dernières sont vraisemblablement plus riches et comprennent, par exemple, des propriétés acoustiques variées : des bruits d'explosion et de friction, des pentes spectrales, des bandes passantes, des équilibres entre régions spectrales, du voisement, des propriétés prosodiques, etc.

Par ailleurs, dans le modèle COSMO générique, la variable S est sensorielle. Même si nous l'interprétons comme auditive puisque les catégories phonétiques sont souvent définies comme telle, nous pourrions envisager que cette variable S soit un espace multisensoriel incluant d'autres modalités comme, par exemple, les représentations somatosensorielles (Patri et al., 2016), ou la vision. Une version plus élaborée des représentations sensorielles est, à ce titre, présentée dans la prochaine section.

De leur côté, les représentations motrices sont également peu détaillées dans COSMO 1D, puis sont implémentées comme un espace de configurations articulatoires dans COSMO-V et COSMO SylPhon, notamment du fait de l'utilisation du modèle VLAM. La notion de « configuration » fait ici référence à la forme générale du conduit vocal à un moment donné tandis que la notion « d'articulatoire » fait référence aux articulateurs utilisés pour réaliser cette configuration. Nous en considérons trois : les lèvres, la langue et la mâchoire. Ainsi ce que nous nommons « configuration articulatoire » correspond plus exactement à la forme et à la position de certains articulateurs à un moment donné (voir également la distinction entre moteur et articulatoire, dans le chapitre 3). Cette caractérisation est, entre autres, incomplète puisqu'elle ne permet pas de modéliser toutes les catégories phonétiques. Le problème apparaît notamment dans l'implémentation des consonnes du modèle COSMO SylPhon.

Bien que nous ayons pu contrer ces inconvénients et développer une version de COSMO SylPhon suffisamment satisfaisante pour nos études, une version améliorée de cette espace moteur est souhaitable.

Il se peut que ces articulateurs ne soient pas assez précis. Une des améliorations possibles serait d'avoir un espace articulatoire de plus grande dimension, pour prendre en compte d'autres paramètres articulatoires. Nous pourrions également remplacer les articulateurs par les muscles du conduit vocal ou par l'ensemble des cavités de résonance formant le conduit vocal (voir par exemple Schroeter et Sondhi, 1994, pour une revue de modèles existants). Néanmoins, connaître la position des composants du conduit vocal n'est en réalité par suffisant pour produire du son. Il faut également un modèle des cordes vocales. Ainsi, VLAM nécessite d'être couplé à un modèle implémentant le contrôle de la source vocale pour pouvoir produire toutes les catégories phonétiques. Autrement dit, le modèle trachée/cordes vocales implémente l'air à transformer en son, tandis que VLAM implémente la cavité permettant de réaliser cette transformation. D'un point de vue computationnel, intégrer un modèle de source vocale ajoute certainement de nouveaux paramètres et donc de nouvelles dimensions pour caractériser les représentations motrices.

L'absence de structure des composantes sensorielles et motrices Outre les problèmes relatifs au choix et au nombre de dimensions des espaces sensoriels et moteurs, il reste un problème majeur non évoqué : la hiérarchie des représentations. Prenons le cas de l'espace sensoriel dans COSMO générique. Dans nos implémentations, nous ne modélisons que l'espace sensoriel correspondant aux catégories phonétiques choisies. Or, durant une tâche quelconque, par exemple, une tâche de perception, le cerveau ne reçoit pas directement le signal acoustique sous la forme d'une représentation auditive prétraitée, reliée directement aux catégories phonétiques correspondantes. Le signal acoustique reçu doit être traité au préalable avant de pouvoir être perçu comme une catégorie phonétique (voir par exemple Poeppel et al., 2012). Le fait que nous utilisions un signal sensoriel synthétique préalablement découpé facilite le problème. Cela évite, d'une part, tout le traitement sensoriel lié à la segmentation du son et, d'autre part, facilite le prétraitement pour ne garder que les paramètres utiles à la catégorisation. Mais, si nous envisageons d'utiliser une représentation sensorielle plus réaliste, cela nécessitera l'ajout d'une ou plusieurs variables sensorielles, convenablement structurées, indépendantes des catégories phonétiques, et permettant de passer du signal sensoriel acoustique aux représentations auditives adéquates pour la catégorisation phonétique.

Par ailleurs, nous pouvons considérer que cette décomposition est commencée dans COSMO SylPhon puisque le modèle comprend des représentations auditives liées aux syllabes, des représentations auditives liées aux phonèmes et des représentations auditives liées aux représentations motrices. Si, à l'origine, ces trois sortes de variables sont un besoin computationnel, cette décomposition peut également être envisagée d'un point de vue théorique. Dans le modèle, ces trois types de représentations sont modélisés sous la même forme formantique (F1/F2 pour les représentations vocaliques et F2/F3 pour les représentations consonantiques) et connectés par une variable de cohérence pour assurer leur égalité. Nous pourrions les envisager comme trois types de variables sensorielles indépendants, représentant chacun une partie du signal sensoriel de base. Par la suite, une variable de plus haut niveau pourrait assurer la liaison entre chacune de ces représentations ou les fusionner pour avoir une vue d'ensemble du signal perçu. Néanmoins, ces pistes de réflexion nécessiteraient une évaluation rigoureuse en lien avec la littérature en neurosciences. Quelle que soit la nature de ces réflexions, le

développement de multiples représentations sensorielles reste nécessaire si nous souhaitons mettre en œuvre un modèle cognitif plus réaliste.

Les mêmes questions se posent sur les représentations de l'espace moteur. Nos représentations motrices sont pour le moment très limitées. Cette limite a déjà été un obstacle notamment dans la variante COSMO SylPhon puisque nous avons dû explicitement créer une représentation motrice ΔM, associée aux catégories consonantiques, distincte de la représentation motrice M^F, associée aux représentations sensorielles. Cependant, cette séparation entre les deux représentations motrices n'est qu'une infime partie de la complexité de l'espace moteur.

8.1.2.2 Variabilité des unités distinctives

Divers aspects de la variabilité des unités phonétiques ont été étudiés dans ce manuscrit. Nous commençons par discuter de la variabilité intra-locuteur avant de nous intéresser à la variabilité inter-locuteur. Qu'elle soit inter ou intra-locuteur, nous montrons que la variabilité des unités phonétiques apparaît comme cruciale dans nos analyses.

Variabilité intra-locuteur Un premier type de variabilité jouant un rôle important dans nos modèles est la variabilité intra-locuteur c'est-à-dire la variabilité des unités phonétiques se trouvant chez un même locuteur. À travers l'utilisation de représentations gaussiennes, nous avons implicitement supposé que les catégories phonétiques sont, dans chacun de leur répertoire, représentées par un prototype (la moyenne des gaussiennes) mais qu'un écart autour de ce prototype ne perturbe pas la catégorisation (la variance des gaussiennes).

Cette variabilité est primordiale car elle est au cœur de la propriété « bande étroite/bande large » que nous avons définie. En effet, la variabilité intra-locuteur est ce qui diffère majoritairement entre la branche auditive et la branche motrice de notre modèle : une petite variabilité, symbolisée par une faible variance, comme celle de la branche auditive, implique une catégorisation efficace des signaux prototypiques, et une grande variabilité, symbolisée par une forte variance, comme celle de la branche motrice, implique des capacités de généralisation et de meilleures performances dans des conditions adverses.

Ce modèle gaussien, associant position d'un prototype et variance autour, est en accord avec les données de la littérature et est également similaire aux différentes implémentations que l'on trouve dans des modèles computationnels phonétiques (de Boer et Kuhl, 2003; Feldman et al., 2009a; Kleinschmidt et Jaeger, 2011; McMurray et al., 2009). Néanmoins, l'utilisation de gaussiennes n'est pas le seul moyen pour la représenter. Il serait intéressant de vérifier si les propriétés observées dans nos études se retrouvent en utilisant d'autres formes paramétriques plus adaptées pour représenter la complexité du spectre de parole.

Variabilité inter-locuteur Un second aspect de la variabilité des unités phonétiques concerne la variabilité inter-locuteur, c'est-à-dire les différences existant entre les différents agents. Celle-ci est analysée notamment à travers les idiosyncrasies auditives comme le montrent les études du chapitre 5.

Dans ces études, nous cherchons d'abord à définir comment les idiosyncrasies apparaissent durant l'apprentissage et nous nous focalisons ensuite sur les corrélations entre les idiosyncrasies en perception et celles en production. Les idiosyncrasies étudiées restent néanmoins limitées puisque nous nous focalisons dans nos études sur les différences formantiques. Il est probable que la variabilité idiosyncratique concerne davantage de paramètres.

Même si la variabilité inter-locuteur des unités phonétiques est centrale dans deux de nos études, il est cependant nécessaire qu'elle soit liée à d'autres types de variabilités. En effet, nous observons dans nos simulations qu'une diversité des développements, c'est-à-dire différents agents apprenants, dans un même contexte d'apprentissage, autrement dit avec un maître unique, génère des représentations différentes d'un agent à l'autre. L'utilisation de multiples agents avec un maître unique nous permet de repérer les phénomènes dépendant des données d'apprentissage de ceux plus robustes et intrinsèques à notre modélisation. Ainsi, dans COSMO SylPhon, nous pouvons obtenir d'un agent à l'autre, d'une part, des différences de positionnement des noyaux gaussiens, voire de nombre de noyaux sélectionnés et, d'autre part, des tendances à la convergence sur un nombre proche de noyaux (autour de 7 pour les voyelles). De plus, les différences observées n'empêchent pas de converger vers un système viable, permettant de communiquer efficacement avec le maître. Cela montre que la structure des représentations des catégories phonétiques peut différer d'un agent à l'autre sans pour autant perturber la communication.

Un autre facteur important de variabilité inter-locuteur, que nous n'avons pas considéré dans ce travail, est celui de la variabilité morphologique liée à l'âge, à la taille et au sexe. Les formants associés à un phonème donné dépendent en effet de la taille et de la forme du conduit vocal du locuteur (voir une revue dans Ménard et al., 2004). Ces différences conduisent au problème, bien connu, de la normalisation ayant déclenché de nombreuses études. Celles-ci peuvent s'organiser en deux axes principaux : la normalisation extrinsèque, impliquant un apprentissage de propriétés d'un locuteur donné (voir par exemple Johnson, 1995), et la normalisation intrinsèque, dans laquelle des paramètres extraits du signal incident conduisent au calcul de paramètres normalisés supposés éliminer la majeure partie des influences de ces variations entre locuteurs (voir par exemple Ménard et al., 2002).

Cependant, dans COSMO, tous les agents sont identiques. En effet, tous nos agents se servent d'un même conduit vocal issu du modèle VLAM pour réaliser la transformation d'une représentation motrice en signal sensoriel. Ce conduit vocal a les mêmes propriétés quel que soit l'agent et ne prend notamment pas en compte les différences entre un bébé et un adulte. Ainsi, notre choix de simplicité, impliquant des agents et un maître morphologiquement identiques, peut être vu comme un cadre computationnel dans lequel les données perçues et produites sont préalablement normalisées, comme le proposent les études sur la normalisation intrinsèque. La question de la normalisation reste néanmoins un enjeu important pour la suite de ces travaux. La mise en œuvre de paradigmes d'interaction entre agents COSMO morphologiquement différents pourrait produire des résultats intéressants dans le développement d'outils de normalisation adaptés à la situation de communication.

8.1.2.3 Structure cognitive des unités distinctives

Comment sont structurées les unités distinctives dans le cerveau ? Nous discutons de cette question autour de trois notions : leur caractère implicite ou explicite, leur composition phonétique et syllabique et la multiplicité de leurs représentations.

L'existence d'unités distinctives discrètes explicites remise en question Utiliser le modèle COSMO générique nous amène à considérer que les catégories phonétiques existent explicitement et correspondent à des unités discrètes. Cette hypothèse est soutenue dans toutes les implémentations du modèle COSMO de base que ce soit COSMO-1D ou COSMO-V. Dans ce modèle, une unité phonétique est représentée par un objet o et celui-ci correspond par la suite à une unique distribution gaussienne moteur dans le répertoire moteur et une unique distribution gaussienne sensoriel dans le répertoire auditif. Ainsi, chaque agent COSMO possède dès le départ et avant tout apprentissage, une représentation discrète complète de ses catégories phonétiques.

Par la suite, cette considération a été remise en question dans les choix de modélisation de sa variante COSMO SylPhon. Dans ce dernier modèle, les catégories phonétiques, qu'elles soient phonémiques ou syllabiques, ne sont plus explicitement représentées dans le modèle. Pour rappel, les catégories phonétiques correspondent à des noyaux gaussiens présents en nombre bien supérieur aux catégories phonétiques qu'ils sont censés représenter. Cela signifie que les catégories phonétiques ne sont plus des unités discrètes à part entière mais qu'elles correspondent à des unités dépendantes de l'espace dans lequel elles sont représentées (dans notre cas soit sensoriel, soit moteur). Par ailleurs, à la fin de l'apprentissage, même si nous considérons que seuls les noyaux gaussiens les plus utilisés, c'est-à-dire ayant un poids non négligeable, sont représentatifs des catégories phonétiques, leur nombre n'est pas toujours équivalent au nombre de catégories puisqu'il existe des catégories phonétiques représentées par plusieurs noyaux gaussiens.

Par la suite, nous pourrions imaginer un mécanisme cognitif de plus haut niveau permettant de passer de cette représentation implicite sous forme de noyaux gaussiens à des catégories phonétiques discrètes explicites. Comme nos noyaux gaussiens sont reliés soit à l'espace sensoriel, soit à l'espace moteur, nous pourrions imaginer des catégories phonétiques discrètes sensorimotrices reliées aux noyaux gaussiens sensoriels et moteurs des catégories correspondantes. Par exemple, en réalisant une table de correspondance entre les noyaux sensoriels et les noyaux moteurs, nous pourrions former l'ensemble des unités phonétiques sensorimotrices. Cette fusion des noyaux gaussiens sensoriels et moteurs en catégories phonétiques sensorimotrices serait en accord avec les théories perceptuo-motrices dont avons discuté préalablement. Néanmoins, le mécanisme permettant de passer des noyaux gaussiens aux catégories phonétiques sensorimotrices n'a rien d'évident et mériterait certainement de plus amples réflexions notamment sur les méthodes de classification.

Actuellement, aucun de nos modèles ne permet réellement de trancher en faveur de l'une ou l'autre des solutions. Néanmoins, sans pour autant affirmer que les catégories phonétiques se limitent à une représentation implicite ou explicite, notre modélisation permet de soulever une question importante sur l'implémentation des catégories phonétiques.

Syllabes vs. phonèmes La nature des catégories phonétiques dans la version de base de COSMO n'est pas une question centrale. En effet, COSMO, tel qu'il est présenté dans le chapitre 3, est composé

d'objets O_S et O_L qui ne sont pas spécifiés et qui peuvent donc représenter n'importe quelle catégorie phonétique. Dans COSMO-1D, utilisé dans la première étude du chapitre 4, l'implémentation reste d'ailleurs assez floue sur la nature des objets puisqu'il s'agit simplement de contrastes phonétiques. Les deux objets O_S et O_L peuvent donc aussi bien représenter des phonèmes que des syllabes. Dans cette étude, nous avons laissé volontairement cette question floue car non seulement elle n'est pas le sujet central de l'étude, mais elle permet de montrer que les résultats obtenus ne sont pas spécifiques à la nature des catégories phonétiques choisies. Un apprentissage supervisé permettrait donc de faire émerger n'importe quelle unité phonétique.

C'est à travers COSMO SylPhon que nous abordons davantage la question de la nature des unités phonétiques. Ce modèle nous permet notamment d'étudier l'émergence des catégories phonémiques à partir de l'unité syllabique. Cela nous permet de montrer que, d'une part, les syllabes et les phonèmes peuvent tous deux émerger à partir d'un apprentissage non supervisé basé sur la perception d'unités syllabiques et que, d'autre part, les phonèmes émergent de deux façon différentes : les voyelles sont mieux apprises dans la branche sensorielle du modèle tandis que les consonnes sont mieux apprises dans la branche motrice. Ce résultat suggère que les phonèmes pourraient être acquis sur la base de composantes perceptuo-motrices.

Bien entendu, notre modélisation possède un certain nombre de limites concernant les catégories étudiées. Nous en soulevons deux principales. La première est de ne considérer que des syllabes CV. Nous pourrions envisager des syllabes plus complexes pouvant traiter tout type de syllabe comme, par exemple, la syllabe [strakt]. La seconde est de ne modéliser que des consonnes plosives. Pour ces deux limites, il s'agit davantage de choix computationnels que de choix théoriques. Une perspective serait donc d'avoir une version de COSMO SylPhon capable de gérer des unités phonétiques plus complexes. L'idéal serait d'avoir un modèle pouvant traiter l'ensemble des unités d'une langue donnée.

Une double représentation Une particularité du modèle COSMO est d'avoir deux variables pour les catégories phonétiques. Il y a d'une part des catégories phonétiques associées aux représentations sensorielles, O_L, et d'autre part des catégories phonétiques associées aux représentations motrices, O_S. Cette particularité a ses avantages puisqu'elle nous permet de représenter et comparer dans un même modèle les différentes théories de la perception. Néanmoins, la question de savoir s'il s'agit d'un choix théorique ou purement computationnel peut se poser.

Premièrement, il s'agit d'un choix en partie computationnel puisque la programmation bayésienne et la décomposition de la conjointe interdisent d'avoir deux fois la même variable en partie gauche. Dans le modèle COSMO, il est donc impossible d'avoir, en même temps, un prior $P(O)$ et un classifieur sensoriel $P(O|S)$. C'est pour cela que la variable O est décomposée en deux objets O_S et O_L et est reliée ensuite par une variable de cohérence C. Comme déjà expliqué précédemment, le rôle de la variable de cohérence est similaire à celui d'un interrupteur. S'il est « non activé », les deux objets sont indépendants et agissent de manière séparée. S'il est « activé », les deux objets sont reliés et considérés comme identiques. Cela peut être considéré aussi bien comme l'existence d'une fusion entre les deux objets que comme l'utilisation d'un unique objet O. Ainsi, bien qu'il y ait deux objets dans le modèle, ceci peut s'interpréter comme l'utilisation d'un unique objet sous certaines conditions. Le modèle en lui-même ne permet donc pas de trancher.

Deuxièmement, d'un point de vue théorique, cette vision dichotomique semble compatible avec certaines données neuropsychologiques, comme celles obtenues par Jacquemot et al. (2007), les ayant conduit à proposer dans leur modèle de mémoire à court terme phonologique l'existence d'un « phonological input » et d'un « phonological output » indépendants. Par ailleurs, cette vision est également compatible avec une ré-interprétation du modèle COSMO dans le cadre de la théorie de l'esprit. Dans ce cadre, nous pourrions supposer qu'un des objets fait référence à l'objet pensé par l'agent tandis que l'autre fait référence à l'objet supposé être pensé par l'interlocuteur. Ce second objet permettrait donc à un agent d'interpréter l'intention du locuteur avec qui il communique (voir Laurent, 2014, pour une discussion plus détaillée sur cette interprétation).

En résumé, dans cette thèse, les deux objets permettent de construire une théorie perceptuo-motrice computationnelle avec un invariant sensori-moteur construit comme une fusion d'invariants sensoriels et moteurs. Comme nous l'avons vu, ceci semble également nécessaire pour différencier la nature des catégories phonétiques : les voyelles, davantage sensorielles, et les consonnes, davantage motrices. Des analyses complémentaires sur cette dichotomie pourraient ouvrir des perspectives importantes dans les relations phonétique-phonologie.

8.1.3 Synthèse des connaissances sur l'apprentissage

Nous avons passé en revue notre modèle sous l'angle de la phonétique à travers nos trois aspects privilégiés. Une des facettes de cette thèse est notre intérêt pour le développement et l'apprentissage, dont les choix et les conséquences se retrouvent dans l'ensemble de nos études. Pour cette raison, il nous semble important de discuter de notre manière de l'implémenter ainsi que des limites et perspectives possibles pour l'améliorer. La discussion est menée autour de trois aspects : les étapes choisies, les scénarios établis et l'utilisation du maître.

8.1.3.1 La dynamique d'apprentissage

Dans toutes nos études, l'apprentissage du modèle COSMO nécessite trois phases : un apprentissage sensoriel pendant lequel l'agent apprend à relier ses représentations sensorielles à ses catégories phonétiques, un apprentissage sensorimoteur durant lequel l'agent à relier ses représentations sensorielles à ses représentations motrices et un apprentissage moteur durant lequel l'agent apprend à relier ses représentations motrices aux catégories phonétiques. En revanche, la séquence diffère selon les études. Dans la première étude du chapitre 4, avec COSMO 1D, nous supposons que les trois phases s'effectuent en parallèle. Dans les deux études du chapitre 5, ainsi que dans COSMO SylPhon, nous considérons que l'apprentissage sensorimoteur et l'apprentissage moteur s'effectuent l'un après l'autre et tous deux en parallèle avec l'apprentissage sensoriel.

Néanmoins, du fait que l'apprentissage sensoriel et l'apprentissage sensorimoteur soient indépendants l'un de l'autre, ils pourraient, dans chaque étude, être considérés en séquence plutôt qu'en parallèle sans que cela n'ait d'effet sur l'apprentissage de l'agent. Ce comportement semble plus en accord avec les données de la littérature. En effet, comme nous l'avions remarqué dans le chapitre 2 et dans les calendriers développementaux (voir par exemple Kuhl, 2004), le bébé semble capable de

percevoir du son de parole avant de pouvoir en produire. Cela suggèrerait donc que l'apprentissage sensoriel commence avant l'apprentissage sensorimoteur et moteur. Cependant, une autre hypothèse pourrait également expliquer ces observations. Dans nos études, nous remarquons que l'apprentissage sensoriel converge rapidement tandis que les apprentissages sensorimoteur et moteur sont beaucoup plus lents. Ainsi, même en supposant un apprentissage parallèle, nous prédisons que l'agent apprend à percevoir, même si ce n'est qu'une perception auditive, avant de savoir produire du son. Il serait donc possible d'imaginer que chez le bébé, les deux apprentissages commencent également en même temps mais que leurs dynamiques soient différentes.

D'autres facteurs contribuent au décalage entre perception et action : le démarrage de l'apprentissage sensoriel dès les dernières étapes de la vie prénatale ou la difficulté spécifique de la coordination source/conduit vocal qui retarde le démarrage du babillage canonique. Par ailleurs, la notion de calendrier développemental génétiquement programmé (Werker et Hensch, 2015) fournit un support naturel à l'hypothèse d'un décalage entre apprentissage sensoriel et sensorimoteur. Néanmoins, notre étude développementale dans COSMO 1D ouvre une piste de réflexion intéressante sur la possibilité, qu'en partie, les décalages de calendrier soient la simple conséquence de différences de complexité dans les processus d'apprentissage.

La relation entre l'apprentissage sensorimoteur et l'apprentissage moteur reste beaucoup plus floue notamment parce qu'il est difficile de savoir à quel moment le bébé apprend réellement à relier ses représentations motrices aux catégories phonétiques. D'un côté, comme l'apprentissage moteur est, en partie, basé sur un apprentissage sensorimoteur, il est plus simple, dans notre modèle, de considérer que ces deux phases s'effectuent séquentiellement. Cela signifierait que le bébé ne commence à relier ses représentations motrices à ses catégories phonétiques qu'une fois qu'il a appris à relier ses représentations motrices à ses représentations sensorielles.

Néanmoins, dans COSMO-1D, nous montrons également que les deux apprentissages peuvent s'effectuer en parallèle, sans que cela n'ait de réelle conséquence sur l'apprentissage final. Là encore, de par leur différence de dynamique, même si les deux apprentissages commencent en même temps, cela n'empêche pas le bébé d'acquérir d'abord la relation entre ses représentations motrices et ses représentations sensorielles et de ne réussir que plus tard à relier ses représentations motrices à ses catégories phonétiques. Bien que les schémas développementaux proposent généralement un décalage entre apprentissage sensorimoteur plus précoce et apprentissage moteur plus tardif (Kuhl, 2004), ces suppositions demandent de plus amples investigations.

8.1.3.2 Les scénarios d'apprentissage

Outre les trois phases d'apprentissage, trois scénarios d'apprentissage se retrouvent dans nos études : un apprentissage supervisé, un apprentissage non supervisé et un apprentissage par accommodation. Nous allons détailler chacun d'eux.

L'apprentissage supervisé est utilisé dans les implémentations de COSMO 1D et COSMO-V lors de l'apprentissage sensoriel. Durant cet apprentissage, le maître fournit le son et la catégorie phonétique à l'agent et celui-ci met directement à jour les paramètres de son répertoire auditif avec ces deux informations. Nous avons fait ce choix en nous basant principalement sur les précédent travaux

effectués sur le modèle COSMO (voir par exemple Laurent, 2014; Moulin-Frier et al., 2012). Cet apprentissage a le mérite d'être simple à mettre en place, ce qui, du point de vue computationnel, n'est pas négligeable, mais qui, du point de vue cognitif, n'est pas réellement un argument. Nous savons que le bébé est influencé par les signaux sensoriels de sa langue, nous pouvons donc supposer qu'il perçoit effectivement le signal sensoriel provenant de l'extérieur et s'en serve pour mettre à jour ses paramètres. Nous avons modélisé cela à l'aide d'un maître mais tout signal de l'environnement ayant les propriétés de sa langue aurait pu faire l'affaire. Néanmoins, le fait qu'il soit capable d'interpréter directement la catégorie phonétique correspondant à ce signal pose davantage de questions. Nous avons deux hypothèses : soit le maître est capable de fournir à l'agent une information sur la catégorie en question, soit l'agent possède déjà des connaissances innées sur ses catégories phonétiques et est directement capable de les associer à des sons.

Étudions le premier cas en supposant que le maître réussisse à informer l'agent de la catégorie phonétique correspondant au signal sensoriel. Une des hypothèses retenues pour expliquer ce comportement est la deixis (voir Laurent, 2014; Moulin-Frier, 2011, pour plus de détails). Cela suppose que le bébé perçoit chaque son en relation avec un élément extérieur et que le bébé se sert de cet élément extérieur pour labeliser le son entrant. Par exemple, imaginons une situation dans laquelle le bébé est capable de reconnaître un biberon et que dans ce mot, il ne retienne que la syllabe [bi]. Ainsi, à chaque fois que le maître prononce le mot biberon en présence du biberon, le bébé peut associer le son « bi » à la syllabe [bi]. Cet exemple, et cette vision en général, présentent néanmoins de nombreuses limites : si deux éléments différents sont identifiés par une même syllabe, disons [bi], le bébé risque d'apprendre deux catégories phonétiques pour la même syllabe [bi]. De même, il est difficilement justifiable de supposer que le bébé identifie chaque élément de son environnement par une catégorie phonétique et non par un mot. En revanche, du fait que cette situation soit en accord avec la littérature en ce qui concerne les mots, il peut être supposé que la catégorie phonétique puisse être déduite du mot à partir d'un traitement lexical mais que ce traitement serait réalisé implicitement dans le modèle COSMO.

Passons au second cas pour lequel le bébé possèderait déjà des connaissances sur ses catégories phonétiques. Cette hypothèse est souvent exploitée par les phonologues mais assez peu par les phonéticiens. Par ailleurs, même s'il est vrai que les bébés possèdent une perception catégorielle de façon précoce, celle-ci semble davantage provenir d'un mécanisme général auditif que de mécanismes linguistiques (Vihman, 2013, voir par exemple). Il n'y a donc aucune preuve actuelle que le bébé possède déjà une connaissance de ses catégories phonétiques à la naissance et encore moins qu'il sache les reconnaître dans un signal sensoriel. En effet, la perception catégorielle montre que le bébé sait mieux discriminer les sons dans certaines régions de l'espace acoustique mais cela n'implique pas forcément qu'il sache les catégoriser en conséquence.

Ces deux hypothèses mises à part, les chercheurs semblent supposer, comme nous l'avons vu dans le chapitre 2, que les catégories phonétiques s'obtiennent, en partie, par un apprentissage statistique non supervisé. C'est, entre autres, pour cette raison que nous avons testé dans COSMO SylPhon un apprentissage non supervisé dans lequel l'agent n'a accès qu'au signal sensoriel du maître. Dans celui-ci, le bébé apprend à relier ses représentations sensorielles aux catégories phonétiques seulement à partir du son qu'il perçoit. Cela correspond à l'apprentisssage réalisé dans plusieurs modèles d'apprentissage sensoriel évoqués au chapitre 3. Une perspective pour cet apprentissage phonétique

serait la prise en compte des mots. En effet, à partir d'un certain âge, quand le bébé commence à apprendre les mots, ceux-ci viennent influencer la relation entre les représentations auditives et les catégories phonétiques (Feldman et al., 2013b). Dans cette thèse, nous avons explicitement choisi de ne pas prendre en compte les mots. Une perspective sur cette thématique est cependant proposée dans la prochaine section.

Le dernier apprentissage est l'apprentissage par accommodation que nous utilisons pour l'apprentissage sensorimoteur et l'apprentissage moteur. L'accommodation consiste en un apprentissage statistique basé sur l'imitation des signaux reçus par l'agent. Cet apprentissage semble crédible puisque nous savons d'une part que le bébé est capable d'imiter des sons et que d'autre part ce qu'il produit est influencé par les sons de sa langue. Ceci laisse supposer, qu'en effet, il semble capable d'apprendre à produire en se basant sur ce qu'il perçoit. Néanmoins, comme le développement sensorimoteur est difficile à analyser, une évaluation précise de notre hypothèse d'accommodation à partir d'une « vérité de terrain » semble, actuellement, hors de portée.

Une des phases d'apprentissage manquantes est l'apprentissage du contrôle du conduit vocal. Avant que le bébé ne produise des sons correspondant à du son perçu ou à des catégories phonétiques, le développement semble supposer qu'il apprend au préalable à contrôler son conduit vocal. C'est d'abord ce qu'il semble se passer lors de la phase de « cooing », lors de laquelle il produit des proto-voyelles. Après cette première phase, le bébé apprend à produire réellement des voyelles. De la même manière, avant la phase de babillage, le bébé apprend à contrôler son conduit vocal et produit un babillage marginal, dans lequel les séquences d'ouverture et fermeture ne respectent pas encore les caractéristiques temporelles des syllabes comme dans le babillage canonique. Cependant, le fait que cette phase d'apprentissage manque dans COSMO est principalement dû au fait que nous ne modélisons pas les effecteurs du bébé et que nous ne faisons pas la différence entre geste prédit et geste produit. Ainsi, les prédictions de notre agent COSMO apprenant correspondent directement à ce qu'il produit, comme s'il maîtrisait déjà parfaitement son conduit vocal. Cela mériterait sans aucun doute d'être amélioré.

Par ailleurs, l'apprentissage par accommodation pourrait être aussi amélioré sur sa composante sociale. En effet, même si le bébé semble effectivement baser son apprentissage sur ce qu'il perçoit, le retour de son environnement semble également important. Il a, par exemple, été montré par Goldsmith et Xanthos (2009), qu'un bébé vocalisait mieux et davantage s'il avait un retour direct de sa mère après ses vocalisations. Dans nos implémentations, cette composante a été assez peu développée. Il sera donc essentiel, dans une phase future de cette recherche, de travailler sur des scénarios d'interaction plus réalistes, en s'inspirant des nombreuses données et observations sur les conditions de développement du langage et le rôle des interactions sociales dans les processus d'apprentissage (voir, par exemple, Kuhl, 2007).

8.1.3.3 Le maître

Un ou plusieurs maîtres Dans chacune des implémentations du modèle COSMO, nous avons fait le choix de n'implémenter qu'un seul maître, qui est, pour rappel, l'agent communiquant des stimuli pour l'agent apprenant lors des différentes phases d'apprentissage. Le choix d'un maître unique facilite

non seulement l'implémentation mais aussi la comparaison des différentes simulations et donc des différents agents apprenants.

Cependant, à part dans de rares cas, il semble inenvisageable que le bébé apprenne ses catégories phonétique à partir des signaux d'une seule personne. Il serait donc judicieux de proposer dans des prochaines versions de COSMO, un apprentissage multi-locuteur. Cela nécessite néanmoins des réflexions préalables sur les différences d'implémentation concernant chaque maître, sur la proportion de stimuli fournis par chacun, et sur les différences possibles du scénario d'apprentissage.

Une question serait de savoir si en introduisant plusieurs maîtres, la variabilité des catégories phonétiques augmente de la même manière que s'il existait une sorte de « méta maître » possédant une grande variabilité dans ses signaux, ou si la présence de plusieurs maîtres nécessite plutôt d'apprendre des modèles différents ainsi que des stratégies d'adaptation propres à chaque maître comme dans les modèles de Johnson (1997) ou Kleinschmidt et Jaeger (2015).

Par ailleurs, la présence de plusieurs maîtres pourrait faciliter la comparaison et l'influence des différents scénarios d'apprentissage. Nous pourrions, par exemple, imaginer deux maîtres : l'un passif, ne faisant que communiquer de façon indirecte à l'agent apprenant en lui fournissant des stimuli sans retour particulier, et un maître actif communiquant avec le bébé de façon directe et réalisant des retours sur ses productions.

La nécessité du maître Nous venons d'évoquer le maître en supposant pour de futures simulations qu'il ne soit plus unique. Nous pouvons néanmoins remettre en question sa nécessité durant l'apprentissage. En effet, dans toutes les implémentations du modèle COSMO, nous supposons que le maître intervient dans chaque phase d'apprentissage et à chaque itération de l'apprentissage mais cette hypothèse est discutable.

Du fait que le bébé semble apprendre par apprentissage statistique à relier ses catégories phonétiques à ses représentations auditives, il semble effectivement important que le bébé ait connaissance des signaux de l'environnement, et donc ait besoin d'un tuteur, pour effectuer son développement auditif. L'implémentation d'un maître semble donc importante pour cette phase d'apprentissage. Durant le développement sensorimoteur et moteur le bébé semble vocaliser et babiller dans sa langue maternelle. Cependant, il faut se rappeler que l'étape de vocalisation, tout comme l'étape de babillage, sont précédées d'une phase de pré-vocalisation et pré-babillage. Si les étapes finales semblent être influencées par l'environnement, les étapes préalables, davantage exploratoires, peuvent être réalisées de façon indépendante de l'environnement.

Dès lors, nous pourrions imaginer dans COSMO un apprentissage sensorimoteur et un apprentissage moteur qui se réalise de façon autonome dans laquelle l'agent apprend à contrôler son conduit vocal. Puis, une fois que l'agent a exploré de façon satisfaisante les possibilités de son conduit vocal, il tenterait de reproduire ce qu'il entend. Cette étape d'exploration sensorimotrice n'est pas novatrice et a déjà été implémentée dans plusieurs modèles comme nous l'avons vu dans le chapitre 3, depuis les premiers modèles dans les années 90 (par exemple Bailly, 1997; Guenther, 1995), jusqu'aux simulations récentes dans le cadre de la robotique cognitive (par exemple Oudeyer et al., 2010). Ces derniers développent notamment un mécanisme d'exploration des gestes moteurs en se basant sur le

principe de curiosité.

Par ailleurs, les deux types d'apprentissage imaginés, l'un d'exploration autonome et l'autre d'imitation des stimuli ne sont pas forcément contradictoires. Il est possible d'envisager que ces deux mécanismes se réalisent en parallèle. Si cette hypothèse s'avère crédible, il faudrait réfléchir à la manière d'implémenter ce parallélisme dans COSMO. Le principe d'accommodation s'en rapproche mais nécessiterait en plus un mécanisme de sélection auto-centré. Dans tous les cas, ces perspectives s'apparentent à la réflexion plus générale sur les rôles respectifs de l'environnement et de l'apprentissage personnel dans l'acquisition des catégories phonétiques.

8.1.4 Synthèse des aspects computationnels

8.1.4.1 La temporalité

La parole étant dynamique et évolutive, elle nécessite une prise en compte de la temporalité. Nous avons déjà abordé, dans la section précédente, la dynamique de l'apprentissage. Nous nous nous intéressons ici à une autre dynamique, portant sur des échelles bien différentes, beaucoup plus petites, de la parole et de son traitement.

Dans le modèle COSMO, nous ne traitons pas le signal en continu, mais nous découpons le signal sonore en instants de temps aussi bien pour l'apprentissage que lors d'une tâche à effectuer. Chaque instant de temps correspond à un stimulus de durée égale à la durée de la catégorie phonétique à traiter. Par exemple, dans les études du chapitre 5 de COSMO, chaque instant de temps permet d'obtenir un stimulus sensoriel dont les caractéristiques correspondent au phonème. Cette façon de faire se retrouve dans un grand nombre de modèles computationnels, comme ceux évoqués au chapitre 3.

D'un point de vue théorique, cette hypothèse suppose que l'humain est capable de segmenter le signal sonore au niveau du phonème. Cependant, la limite d'un tel découpage est qu'il ne prend pas en compte la dépendance entre les différentes unités de temps. Comme nous ne traitons que les voyelles dans les études du chapitre 5, qui ont un signal sensoriel relativement stable et indépendant des autres catégories phonémiques, nous pouvons nous permettre ce découpage. Néanmoins, ceci est plus compliqué avec les consonnes. Comme nous le voyons dans l'implémentation de COSMO SylPhon dans lequel les stimuli sont découpés non pas en phonèmes mais en syllabes CV, le stimulus de la consonne est dépendant de celui de la voyelle suivante. Ceci nous oblige à définir une représentation des consonnes prenant en compte cette dépendance. Ainsi, le découpage en instant de temps facilite le traitement des stimuli mais permet difficilement de gérer la dépendance existant entre les différentes unités. Par ailleurs, le problème s'accentuerait si nous utilisions des stimuli de parole naturels au lieu de stimuli synthétiques comme c'est le cas actuellement.

Outre la temporalité du signal de parole, d'autres temporalités sont à prendre en compte comme celle du traitement lui-même. En effet, le traitement de parole ne se fait pas de manière instantanée mais de manière progressive dans le cerveau. Cet aspect est rarement considéré dans les modèles computationnels. Une des exceptions est celui de Kiebel et al. (2009) dans lequel le traitement des phonèmes est quatre fois plus rapide que celui des syllabes. Il n'est absolument pas pris en compte

actuellement dans le modèle COSMO. En effet, nous nous préoccupons de la manière dont est traité le signal mais nous ne nous préoccupons pas de l'ordre ni de la séquence de ce traitement. Cela n'est pas formulable explicitement dans le modèle mais pourrait être interprété en s'aidant des inférences réalisées pour chaque tâche étudiée (Barnaud et al., 2017).

8.1.4.2 La non prise en compte du continu

Depuis le début de cette thèse, nous nous plaçons dans un cadre computationnel discret et tronqué. Ainsi, chaque variable a un nombre déterminé de valeurs et les distributions associées correspondent, en réalité, à des points sur cet espace de valeurs.

Bien que ceci accélère nos calculs et nos simulations, cela pose plusieurs questions notamment sur la manière de discrétiser et sur le nombre de valeurs à considérer. Sans prendre en compte les espaces catégoriels qui sont des espaces discrets par définition, nous avons choisi de discrétiser les représentations sensorielles et motrices de façon linéaire. Il s'agit du choix le plus simple mais pas forcément du plus représentatif. En effet, les valeurs sont souvent concentrées sur des portions bien précises de l'espace et non étalées sur son ensemble. Il pourrait être envisagé de discrétiser l'espace non pas linéairement mais en fonction de l'information reçue.

Par ailleurs, nous avons tenté d'avoir un compromis satisfaisant entre temps de calcul et précision. Des études quantitatives plus poussées seraient certainement à mener sur la pertinence de la discrétisation choisie. Nous pourrions également envisager une version de COSMO prenant en compte des variables continues et non plus discrètes. Cependant, il n'est pas absurde de considérer que les espaces physiquement continus comme l'espace moteur et l'espace sensoriel soient traités de manière discrète dans le cerveau. Bien que très nombreux, nous avons un nombre discret de neurones codant l'information perçue ou produite. Il est donc probable que le signal sensoriel ou les gestes moteurs soient traités également de façon discrète par les neurones du cerveau. C'est ce type de considération que l'on retrouve dans les modèles connexionnistes.

8.1.5 Synthèse des aspects de modélisation

Dans cette partie, nous discutons de l'utilisation des modèles computationnels en général avant de discuter plus en détail des modèles bayésiens et de COSMO.

8.1.5.1 Les différents rôles du modèle COSMO et des études réalisées

Au final, nous pouvons nous demander ce qu'apporte la modélisation et plus particulièrement le modèle COSMO. Dans cette thèse, le modèle COSMO est utilisé en tant que modèle cognitif avec pour objectif général de mieux comprendre comment le cerveau traite les sons de la parole pour accéder aux unités distinctives. Pour cela, il est utilisé de différentes manières.

Un des premiers rôles de COSMO est d'interpréter des comportements observés dans la littéra-

ture et actuellement non expliqués. Il a été utilisé en ce sens dans la première étude du chapitre 4 pour tenter d'interpréter les différences entre les représentations sensorielles et motrices et de comprendre pourquoi les représentations motrices semblent plus utilisées durant une tâche de perception dans des conditions adverses. Nous avons interprété ce phénomène comme une conséquence des différences d'apprentissage sensoriel et moteur, ce qui nous a conduit à définir la propriété « bande étroite/bande large ». Cette interprétation nous permet de faire des prédictions qui pourront, par la suite, être vérifiées par des études comportementales ou en neurosciences.

Un des autres rôles du modèle COSMO est de tenter de réfuter des hypothèses à partir de la comparaison de modèles. Plus précisément, cela consiste à implémenter deux modèles computationnels similaires se distinguant par une hypothèse de modélisation et de les tester sur un fait expérimental rapporté dans la littérature. Si cette hypothèse de modélisation est pertinente pour le résultat testé, alors un des deux modèles reproduira le résultat trouvé dans la littérature et pas l'autre (en tout cas, l'un des deux modèles reproduira l'observation mieux que l'autre). À partir de cette comparaison, nous considérons comme erroné le modèle n'ayant pas réussi à reproduire le résultat et prédisons que l'hypothèse de modélisation du second modèle est plus plausible. Cette approche, et notamment le processus de réfutation, est assez proche des expérimentations comparatives réalisées en psychologie. Nous avons, par exemple, utilisé le modèle COSMO en ce sens dans les deux études du chapitre 5. Dans la première étude, la comparaison d'un apprentissage par imitation et par communication nous conduit à montrer qu'un apprentissage par imitation ne permet pas de faire apparaître des idiosyncrasies, contrairement au second. Dans la seconde étude, une comparaison des trois théories de perception a été réalisée dans le but de reproduire les résultats obtenus dans Ménard et Schwartz (2014). Nous observons que les représentations motrices sont nécessaires pour reproduire les résultats.

Ces études computationnelles permettent d'éclairer l'interprétation des résultats expérimentaux et d'en tirer des conséquences théoriques. Elles génèrent en retour des prédictions qui conduisent à des pistes d'expérimentation dans de nouvelles directions.

8.1.5.2 Les avantages du modèle COSMO

Nous avons choisi le modèle COSMO dans cette thèse, d'abord, par continuité avec les travaux précédents de l'équipe, mais aussi parce qu'il nous semble adapté pour étudier des questions importantes sur le traitement et le fonctionnement des unités phonétiques. Dès lors, une question se pose tout naturellement : un autre modèle aurait-il pu être pertinent, voire plus pertinent ?

Pour commencer, sans considérer que COSMO est le modèle computationnel le plus pertinent existant, nous montrons dans ce manuscrit que le modèle COSMO est un modèle nous permettant de répondre à l'ensemble de nos études, aussi diversifiées soient-elles. La majorité des modèles analysés dans le chapitre 3 sont, à l'inverse, spécialisés sur une tâche précise et n'auraient pas permis de réaliser l'ensemble de nos études. Cela s'explique principalement pour deux raisons.

La première raison est le caractère générique du modèle COSMO, ce qui est, sans doute, un de ses plus grands avantages par rapport aux autres modèles computationnels existants. Celui-ci nous offre la possibilité d'étudier plusieurs aspects de la phonétique dans un même modèle et de nous placer à différents niveaux d'analyse. C'est ainsi que dans la première étude du chapitre 4, nous étudions

les conséquences de l'apprentissage auditif et moteur sur le fonctionnement global et général des branches auditive et motrice avec un modèle très simple et générique que nous nommons COSMO 1D. Puis, dans le chapitre 5, nous précisons assez finement nos représentations et implémentons une version du modèle nommée COSMO-V pour pouvoir la comparer aux résultats de l'étude de Ménard et Schwartz (2014). Ainsi, si chaque étude et chaque implémentation a son propre niveau d'analyse, le modèle en lui-même peut s'adapter pour répondre à des problématiques de différents niveaux. Néanmoins, comme l'a montré le modèle COSMO SylPhon, il arrive que le modèle seul ne suffise pas et qu'il soit étendu pour prendre en compte des représentations plus évoluées. Pour ces études plus complexes, il est possible qu'un modèle pré-existant comme celui de Kröger et al. (2011) aurait été plus judicieux mais nous avons préféré conserver la base de COSMO pour rester en accord avec nos précédentes études.

La seconde raison est que le modèle COSMO est basé sur le principe de « questions », chaque question correspondant à une tâche que peut effectuer le modèle. Ce principe, en plus d'offrir la possibilité de traiter des tâches diverses, nous permet de bien séparer la construction du modèle et son apprentissage de la phase d'évaluation consistant à réaliser différentes tâches aussi bien liées à la perception qu'à la production.

Un autre avantage de COSMO est qu'il nous permet de considérer, dans un même cadre computationnel, l'apport des représentations motrices et des représentations sensorielles des unités distinctives. Grâce à cela, nous pouvons définir, dans un même modèle, les trois familles de théories de la perception à savoir les théories auditives, motrices et perceptuo-motrices, et observer leur effets respectifs durant la perception. Le modèle COSMO est un des seuls à faire cette distinction entre ces trois caractérisations et à centrer le modèle sur les catégories phonétiques au lieu de le centrer sur les espaces de représentations.

Parmi les modèles de la littérature considérant les catégories phonétiques, nous avons observé, dans le chapitre 3, que seuls quelques rares modèles de perception possèdent également des représentations auditives et motrices, les autres utilisant principalement des représentations auditives. Nous avons relevé notamment deux modèles (chacun avec leurs déclinaisons respectives) : le modèle proposé par l'équipe du « MirrorNeurons and Interaction Lab », à Ferrare en Italie (Canevari et al., 2013; Castellini et al., 2011) et celui de Kröger et al., à Aachen en Allemagne (Kröger et al., 2011, 2009). Nous aurions pu donc potentiellement utiliser ces deux modèles pour certaines de nos études.

Pour finir, si le modèle COSMO est effectivement adapté aux études réalisées, cela n'empêche pas les autres modèles d'être plus performants sur d'autres aspects. Par exemple, les deux modèles ci-dessus ont l'avantage de pouvoir traiter un signal acoustique réel. Ce n'est pas impossible dans le modèle COSMO, comme déjà expliqué précédemment, mais cela n'a actuellement pas encore été réalisé et demanderait, sans aucun doute, quelques adaptations.

8.1.5.3 Interprétation du modèle COSMO

Interprétation générale du modèle L'avantage, mais aussi l'inconvénient du modèle COSMO générique est qu'il est assez général pour pouvoir être interprété de différentes manières. Nous avons déjà évoqué le fait qu'il pouvait être envisagé, par exemple, comme un modèle de la théorie de l'esprit,

même si ce n'est pas l'orientation prise dans cette thèse.

Dans cette thèse, le modèle COSMO est toujours présenté comme un modèle cognitif du traitement phonétique, centré sur les catégories phonétiques. De ce fait, les représentations motrices et sensorielles considérées ne correspondent qu'aux représentations utiles à la catégorisation. Cela explique en partie pourquoi la variable S est une variable auditive prétraitée normalisée caractérisant directement les catégories phonétiques auditives et pourquoi la variable M mélange aussi bien geste moteur prédit, geste moteur produit et retour somatosensoriel pour être reliée de manière directe aux catégories phonétiques motrices.

Si le modèle COSMO SylPhon semble un peu plus complexe, c'est essentiellement parce qu'il prend en compte des catégories phonétiques de natures différentes. Les représentations auditives et motrices sont, elles, toujours aussi simplifiées et implémentées de manière à être directement reliées aux catégories phonétiques. La seule exception notable concerne la variable motrice initiale M de COSMO qui, dans le modèle, est décomposée pour être, d'une part, une variable motrice M_F, prenant en compte la coarticulation, directement liée aux représentations sensorielles et, d'autre part, une variable motrice ΔM qui caractérise l'espace catégoriel des consonnes.

Le niveau d'analyse selon la hiérarchie de Marr La hiérarchie de Marr propose trois niveaux de représentations pour les modèles computationnels, qui sont, du plus conceptuel au plus proche du substrat physique : un niveau computationnel, un niveau algorithmique et un niveau implémentation. Le premier niveau concerne les fonctions d'ensemble du modèle en termes d'entrées, de sorties et d'objectif global. Le second niveau concerne les algorithmes et représentations permettant de calculer chaque brique de fonctions et d'en donner une explication du fonctionnement. Le troisième niveau concerne la manière dont sont implémentés (physiquement, biologiquement, chimiquement) ces algorithmes dans un agent.

Prenons l'exemple de la modélisation de la classification phonémique. Si un modèle ne se préoccupe que de la classification dans sa globalité en expliquant qu'il prend en entrée un stimulus sensoriel et qu'il donne en sortie le phonème correspondant, alors il se place au niveau computationnel. Dans ce cas, la classification phonémique est comme une boîte noire dont nous ne savons rien. Si un modèle se préoccupe de mettre en place un algorithme expliquant comment passer du stimulus au phonème, il se place au niveau algorithmique. Si enfin un modèle se préoccupe de savoir comment chaque neurone code le stimulus pour trouver le phonème correspondant, il se place au niveau implémentation.

Ainsi, le modèle COSMO est un modèle bayésien algorithmique (Diard, 2015), et, plus exactement, un modèle algorithmique de la phonétique cognitive. Chaque brique de fonctions correspond à une question traitée par inférence dans le modèle, qui explique l'algorithme à suivre pour la réaliser. De ce fait, le modèle COSMO se place clairement dans une démarche explicative cherchant, non pas à décrire, mais à expliquer, ou du moins, donner des mécanismes possibles afin de mieux comprendre comment les unités phonétiques sont traitées dans le cerveau. Cela confirme ce que nous avons expliqué précédemment sur les rôles de COSMO.

Les connaissances et les processus Dans nos modèles, nous souhaitons faire la différence entre deux types de distributions : les distributions apparaissant dans la définition de la distribution conjointe

et celles inférées en réponse à une question posée au modèle. Cette distinction n'est pas spécifique à COSMO. Il s'agit plutôt d'une caractéristique de la programmation bayésienne. Selon cette vision, les premières sont considérées comme les représentations dont les valeurs sont stockées dans le cerveau. Nous ne rentrons pas dans les détails de ce stockage qui concernent les études sur la mémoire et qui dépassent largement les préoccupations de cette thèse. Nous considérons simplement qu'elles sont stockées d'une façon ou d'une autre. Nous considérons également que ce sont les représentations apprises durant le développement du bébé. C'est pourquoi, nous mettons à jour ces distributions lors de l'apprentissage de notre agent COSMO. C'est ce que nous regroupons sous le terme général de « connaissances du modèle ».

À l'opposé, les tâches cognitives correspondent à des questions qui sont des distributions ne correspondant généralement pas directement à ce qui est stocké mais à un traitement, calculé à l'aide des représentations stockées en mémoire. Par inférence, l'agent utilise les distributions du modèle pour calculer la distribution recherchée. Nous supposons un processus similaire dans le cerveau. La question et la distribution associée correspondent à une tâche particulière pour laquelle le bébé fait appel aux représentations stockées dans son cerveau pour y répondre. C'est ce que nous regroupons sous le terme général de « processus du modèle ».

Il s'agit pour le moment d'une hypothèse mais cette différenciation entre connaissance et processus questionne l'interprétation des résultats en neurosciences. En effet, si cette hypothèse s'avère vraie, nous pouvons nous demander ce qui est observé en neurosciences quand une zone s'active. Cela correspond-il à ce qui est stocké en mémoire ou est-ce le résultat d'un calcul ?

Prenons un exemple issu de COSMO. Dans COSMO, nous calculons par exemple une tâche de perception d'un son selon les théories perceptuo-motrices de cette façon :

$$P(O \mid S\,[C=1]) \propto P(O_L \mid S) \sum_M P(M \mid O_S)\; P(S \mid M) \tag{8.1}$$

La distribution en partie gauche est la distribution calculée et celles en partie droite sont les distributions du modèle qui pour nous sont stockées. Lorsqu'on observe une activation dans une aire particulière, observe-t-on l'activation des distributions stockées (donc les distributions en partie droite) ou observe-t-on le résultat du calcul se diffusant dans les aires correspondantes (donc la distribution en partie gauche) ? Selon la réponse à cette question, les résultats peuvent différer. Cette question ne concerne pas les expérimentations étudiant les aires du cerveau dans leur globalité (par exemple en utilisant des technique IRM ou TMS) mais devient cruciale lorsque nous nous intéressons à l'organisation et aux réponses neuronales. Cette interrogation peut donc devenir véritablement importante si les études en neurosciences utilisent des technologies de neuroimagerie de plus en plus précises (voir Barnaud et al., 2017, pour plus de détail sur cette discussion).

8.2 Perspectives : trois extensions du modèle COSMO

Comme le montre l'ensemble de nos développements sur COSMO de cette thèse, notre cadre computationnel est riche. Afin de conclure ce manuscrit, nous proposons trois perspectives, qui sont

trois autres manières d'explorer le modèle. Ces développements sont plus « exploratoires » que nos contributions des chapitres précédents. Le premier axe de développement de COSMO concerne les neurosciences. Dans une étude spécifique, nous avons proposé l'architecture d'une implémentation neuronale possible du modèle et avons tenté d'interpréter plusieurs données expérimentales de neurosciences avec cette architecture, que nous nommons « COSMO Neuro ». Le second axe concerne la prise en compte du lexique. Pour cela, nous avons étudié l'influence du lexique sur l'apprentissage des catégories phonétiques grâce à une variante du modèle que nous nommons « COSMO WordPhon ». Le troisième axe concerne la multisensorialité pour laquelle nous définissons un modèle, que nous nommons « COSMO multi-sensoriel », actuellement purement théorique, où COSMO ne contient plus uniquement une branche sensorielle auditive mais également des branches somatosensorielle et visuelle.

8.2.1 COSMO Neuro

Publications :

— Barnaud, M.-L., Bessière, P., Diard, J., et Schwartz, J.-L. (2017, en révision). Reanalyzing neurocognitive data on the role of the motor system in speech perception within COSMO, a Bayesian perceptuo-motor model of speech communication

— Barnaud, M.-L., Diard, J., Bessière, P., et Schwartz, J.-L. (2017b). Perceptuo-motor speech units in the brain with COSMO, a Bayesian model of communication. In Proceedings of the 11th International Seminar on Speech Production

---

Le premier axe abordé concerne le lien entre le modèle COSMO et les données neurocognitives. Afin d'étudier une nouvelle fois le rôle et l'importance des représentations motrices durant la perception, nous tentons d'expliquer et d'interpréter les résultats obtenus dans deux études en neurosciences cognitives. Du fait que les résultats considérés sont issus d'analyses de réponses corticales, cette étude nécessite la proposition d'une architecture neuronanatomique pour le modèle COSMO visant à définir dans quelles aires du cerveau sont stockées les connaissances du modèle et où s'effectuent les différentes inférences.

Pour cette étude, la version de COSMO considérée est la version générique, décrite au chapitre 3. Le cœur de cette étude est la compréhension du processus de perception, que nous modélisons ici par le décodeur perceptuo-moteur $P(O_S \mid S\,[C = 1])$ composé d'un décodeur auditif $P(O_L \mid S)$ et d'un décodeur moteur $P(O_S \mid S)$.

8.2.1.1 Les trois propriétés

En nous basant sur les résultats obtenus dans cette thèse ou dans de précédentes études, nous supposons que les représentations motrices intervenant dans le processus de décodage des unités phonétiques présentent trois propriétés principales : redondance, complémentarité et spécificité.

Pour présenter ces trois propriétés, rappelons que le décodage perceptuo-moteur s'effectue, dans COSMO, par fusion des décodeurs auditif $P(O_L \mid S)$ et moteur $P(O_S \mid S)$. La propriété de redondance vient de l'hypothèse que la fusion de ces deux décodeurs fournit une information en partie redondante, ce qui renforce la robustesse du décodage perceptuo-moteur. Néanmoins, bien qu'ils soient en partie redondants, nous supposons également qu'ils sont complémentaires comme cela a déjà été évoqué dans le chapitre 4 : le décodage auditif serait précis et s'adapterait au mieux à la distribution de l'environnement, tandis que le décodage moteur possèderait une plus grande variance ce qui lui confèrerait des capacités de généralisation. Nous avons nommé cette complémentarité « bande étroite/bande large ». Enfin, en plus d'être complémentaires, les représentations motrices et sensorielles seraient chacune spécifique, ou, en tout cas, mieux adaptées, à certaines unités phonétiques. En effet, comme nous l'avons vu avec COSMO SylPhon, les voyelles sont mieux caractérisées par les représentations auditives et les consonnes mieux caractérisées par les représentations motrices. Nous estimons que ces trois rôles permettent d'expliquer en grande partie les résultats des études en neurosciences que nous choisissons d'analyser.

8.2.1.2 L'architecture neuronale du modèle COSMO

Nous commençons par proposer une architecture neuroanatomique explicitant dans quelles aires du cerveau les distributions du modèle COSMO sont stockées et calculées. Comme nous l'avons précédemment observé dans le chapitre 2, les aires respectives associées à chaque représentation diffèrent selon les expérimentations. Pour notre architecture, nous nous sommes principalement inspirés de l'organisation neuroanatomique de Hickok et Poeppel (2007) et Rauschecker et Scott (2009) avec l'hypothèse d'une voie dorsale qui connecterait le réseau de traitement auditif (cortex auditif primaire et secondaire, planum temporale PT, sulcus temporal postérieur supérieur pSTS) à un réseau articulatoire dans le lobe frontal (gyrus frontal inférieur IFG, cortex prémoteur PMC, insula antérieure, cortex moteur primaire M1) en passant par une interface sensorimotrice dans le lobe pariétal inférieur (et jusqu'à la frontière pariéto-temporale du sillon latéral à l'intérieur du planum temporale Spt).

Dans ce contexte, l'architecture neuronale proposée pour le modèle est représentée Fig. 8.1. Les distributions représentées sur cette figure sont données à titre illustratif et correspondent au modèle COSMO 1D utilisé dans le chapitre 4, les aires correspondantes sont indiquées par une flèche bleue en pointillés. Ainsi, nous supposons que le répertoire auditif $P(S \mid O_L)$ serait stocké dans les aires temporales postérieures supérieures. Le modèle interne $P(S \mid M)$ serait stocké dans l'interface sensorimotrice de l'aire Spt. Enfin, le répertoire moteur $P(M \mid O_S)$ serait stocké globalement dans les aires frontales, probablement à différents niveaux, allant de simples variables articulatoires situées plutôt dans M1 à des programmes moteurs associés à des unités phonétiques situées plutôt dans PMC.

Dans cette même figure, nous avons également représenté les processus de décodage. Ceux-ci sont spécifiés algorithmiquement dans les cadres de différentes couleurs et localisés dans les aires du cerveau par les flèches de couleurs correspondantes. Ainsi, nous supposons que le décodage auditif $P(O_L \mid S)$ s'effectuerait probablement dans l'aire pSTS ou au niveau du gyrus supramarginal SMG. Le décodage moteur $P(O_S \mid S)$ s'achèverait dans les aires frontales, probablement dans le cortex prémoteur, après avoir notamment traversé différentes aires pariétales. Enfin, le résultat du décodage moteur serait renvoyé dans les aires temporales, possiblement au niveau de l'aire Spt, afin que s'ef-

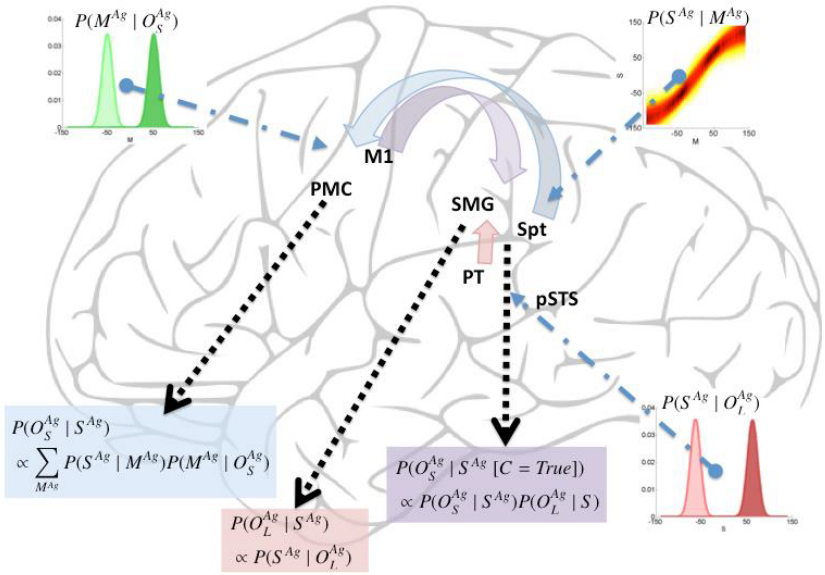


FIGURE 8.1 – Une possible architecture corticale pour le modèle COSMO

fectue le décodage perceptuo-moteur $P(O_S \mid S\ [C = 1])$.

Cette proposition présente bien évidemment de nombreuses limites et simplifications. Par exemple, comme nous l'avons dit préalablement, le fait que les représentations auditives et motrices soient limitées à une unique variable, respectivement S et M, qui réfèrent en réalité chacune à une hiérarchie complexe, rend difficile le positionnement des distributions dans le cerveau. Cela explique pourquoi certaines relations ne sont que partiellement spécifiées ou associées à des localisations possiblement multiples.

8.2.1.3 Les deux études réanalysées à la lumière de COSMO

La première étude choisie est celle de Möttönen et Watkins (2009) évoquée dans le chapitre 2. Dans cette étude, les auteurs montrent qu'une perturbation des aires frontales du cerveau modifie la discrimination des catégories phonétiques dans des tâches de perception. L'expérimentation consiste à perturber le cortex moteur primaire gauche, grâce à l'application d'une stimulation magnétique transcranienne répétée (rTMS) , durant une tâche de perception. Cette dernière est, plus précisément, une tâche de catégorisation et de discrimination d'un continuum de stimuli acoustiques entre les syllabes [ba] et [da]. Deux zones du cortex moteur primaire ont été perturbées : la « région des lèvres » et la « région de la main », cette dernière servant de contrôle. Les résultats montrent que la catégorisation et la discrimination sont toutes deux perturbées durant la stimulation de la région des lèvres, mais pas de la région de la main, ce qui montrerait bien un rôle spécifique du cortex primaire orofacial dans la perception de la parole.

Pour expliquer ce phénomène, nous supposons, conformément à l'architecture corticale proposée, qu'une perturbation des aires frontales du cerveau correspond dans notre modèle à une perturbation du

répertoire moteur. Du fait de cette supposition, comme le décodage moteur $P(O_S \mid M)$ se fait à l'aide du répertoire moteur, la perturbation des aires frontales impliquerait une modification de ce décodage moteur. Or, selon la propriété de redondance, le décodage se fait par fusion d'un décodage auditif $P(O_L \mid M)$ et d'un décodage moteur $P(O_S \mid M)$. Quand les deux informations sont concordantes et donc redondantes, la discrimination est correcte. Or, la perturbation du répertoire moteur modifie en conséquence le décodage moteur, ce qui perturbe la fusion et donc le décodage perceptuo-moteur. En reproduisant, avec le modèle COSMO, une perturbation du répertoire moteur durant le processus de décodage en augmentant la variance d'une des distributions gaussiennes du répertoire, nous parvenons à reproduire, au moins qualitativement, la dégradation des performances de résultats de discrimination observée par Möttönen et Watkins (2009), renforçant ainsi notre interprétation.

La seconde étude choisie est celle de Cheung et al. (2016). Dans cette étude, les auteurs montrent que les régions motrices activées durant la perception dépendent davantage de propriétés auditives que motrices, remettant ainsi implicitement en question les théories motrices de la perception. Pour réaliser leur expérimentation, les auteurs comparent les réponses neurophysiologiques (obtenues par enregistrements corticaux de surface sur des patients épileptiques implantés) durant des tâches de perception et de production. La tâche de production consiste à produire à haute voix les unités phonétiques tandis que la tâche de perception consiste à les écouter passivement. Les unités phonétiques considérées sont des syllabes Consonne+Voyelle (CV) pour lesquelles V correspond à la voyelle [a] et C correspond à une consonne parmi huit consonnes de l'anglais américain [b d g p t k s ʃ]. Les zones étudiées sont le gyrus temporal supérieur (STG) pour les régions supposées « auditives » et le cortex sensorimoteur (SMC) pour les régions supposées « motrices », toutes deux traitées par une grille d'électrodes permettant l'analyse de données de haute résolution spatiale et temporelle. Les auteurs observent que durant la tâche de la production, l'activité neuronale de la région SMC est somatotopiquement organisée en trois groupes, correspondant chacun à des propriétés articulatoires : les labiales [p b], les vélaires [k g] et les alvéolaires [t d s ʃ]. À l'inverse, durant la tâche de perception, l'activité neuronale de la région STG est somatotopiquement organisée en trois groupes correspondant plutôt à des propriétés acoustiques : les voisées [b d g], les non voisées [p t k] et les fricatives [s ʃ]. De façon intriguante, les analyses montrent que l'organisation neuronale de la région SMC durant la tâche de perception ressemble davantage à celle de la région STG durant cette même tâche, qu'à la région SMC durant la tâche de production. Les auteurs en déduisent que « le cortex moteur ne contient pas de représentations articulatoires des actions perçues en parole, mais réprésente plutôt des informations vocales auditives » [1].

Nous proposons une interprétation différente de ces observations. En effet, pour expliquer ce phénomène, nous supposons que l'observation de la tâche de production dans SMC consiste à observer la branche de production motrice $P(M \mid O_S)$. Comme précisé précédemment, nous supposons que l'observation de la tâche de perception dans STG consiste à observer le décodeur auditif $P(O_L \mid S)$ et que l'observation de la tâche de perception dans SMC consiste à observer le décodeur moteur $P(O_S \mid S)$. Ensuite, selon la propriété de spécificité et en accord avec plusieurs données de la littérature, nous supposons que les catégories phonétiques sont organisées selon des propriétés articulatoires dans le répertoire moteur $P(M \mid O_S)$ et que les catégories phonétiques sont organisées selon des propriétés auditives dans le répertoire auditif $P(S \mid O_L)$. En tant que modèle interne $P(S \mid M)$, nous utilisons,

1. « motor cortex does not contain articulatory representations of perceived actions in speech, but rather, represents auditory vocal information »

par simplification, un modèle sigmoïde, comme celui utilisé dans la première étude du chapitre 4. Ensuite, nous réalisons nos tâches de perception et production. Nous observons que la branche de production motrice $P(M \mid O_S)$, qui se calcule de façon directe à l'aide du répertoire moteur $P(M \mid O_S)$, est organisée selon des propriétés articulatoires, ce qui est concordant avec l'observation faite dans SMC durant la tâche de production. De même, la distribution du décodage auditif $P(O_L \mid S)$, calculée de façon directe à l'aide du répertoire auditif $P(S \mid O_L)$, est organisée selon des propriétés auditives, ce qui est concordant avec l'observation faite dans STG durant la tâche de perception. Enfin, le décodage moteur $P(O_S \mid S)$, calculé à partir du répertoire moteur $P(M \mid O_S)$ et du modèle interne $P(S \mid M)$, est davantage organisé selon des propriétés auditives qu'articulatoire, montrant ainsi que l'inférence impliquant le répertoire moteur et le modèle interne résulte en une organisation finale différente de celle du répertoire moteur. Ce résultat est aussi concordant avec l'observation faite dans SMC durant la tâche de perception.

Ainsi, en comparant nos distributions avec les aires du cerveau analysées par Cheung et al. (2016), nous sommes arrivés aux mêmes observations mais notre interprétation est différente. Selon nous, le décodage moteur $P(O_S \mid S)$ ne correspond ni à des représentations des actions perçues (comme dans le répertoire moteur $P(M \mid O_S)$) ni à des informations vocales auditives (comme dans le répertoire auditif $P(S \mid O_L)$), mais à un troisième type d'information issu du calcul entre le répertoire moteur et le modèle interne. Ainsi, l'organisation de SMC durant la tâche de perception proche de l'organisation auditive de STG durant cette même tâche ne remet pas en cause selon nous l'apport des représentations motrices durant la perception de la parole. Mieux encore, cela nous conduit à faire l'hypothèse que, conformément à l'hypothèse « bande étroite/bande large », le pattern de similarité des activités neuronales changerait en cas de perception dans le bruit : dans ce cas, nous prédisons une plus grande proximité des réponses dans les aires SMC en perception et en production (Barnaud et al., 2017b).

8.2.2 COSMO WordPhon

Le deuxième axe étudié concerne l'étude de l'influence des connaissances lexicales sur les catégories phonétiques. Il s'agit donc d'une extension de COSMO vers des unités de parole de plus haut niveau qui sort du cadre de la phonétique et des unités de seconde articulation que nous avons constamment maintenu durant cette thèse. Cette étude a été menée principalement par Ali Saghiran, étudiant en double diplôme du Master 2 Sciences Cognitives et de l'Institut National Polytechnique de Grenoble, durant son stage de fin d'année que nous avons encadré durant cette thèse. Pour plus de détail, nous invitons le lecteur à se reporter à Saghiran (2017).

L'objectif principal de cette étude est d'analyser si l'apprentissage lexical a une influence sur l'apprentissage phonétique. Du fait que le modèle COSMO de base, présenté au chapitre 3, ne possède qu'un type d'unités phonétique, nous avons donc développé une nouvelle version du modèle, celle-ci prenant en compte les mots, nommée COSMO WordPhon, inspirée de COSMO SylPhon.

8.2.2.1 Le modèle

Le modèle COSMO WordPhon ne concerne et n'étend que la branche auditive de COSMO. Il ne contient donc que des objets et des représentations sensorielles. Il y a trois types de variables : des mots, notés W, des unités phonétiques notées G et des représentations sensorielles S, toutes trois organisées de façon hiérarchique comme on le trouve dans la littérature (McClelland et Elman, 1986, voir aussi chapitre 3).

Les variables G du modèle correspondent, comme dans COSMO SylPhon, à N_g noyaux gaussiens symbolisant les unités phonétiques et caractérisés par leur moyenne μ, leur matrice de covariance Σ et leur poids de pondération N. Pour simplifier les simulations, les mots W sont des combinaisons de deux unités phonétiques G_1 et G_2.

Le modèle se décompose de deux manières différentes. Dans une version π_C, nous considérons que les noyaux gaussiens G_1 et G_2 représentent le même espace. Dans une version π_D, nous considérons qu'il s'agit de deux espaces indépendants. Les deux versions sont schématisées Fig. 8.2 et se décomposent par :

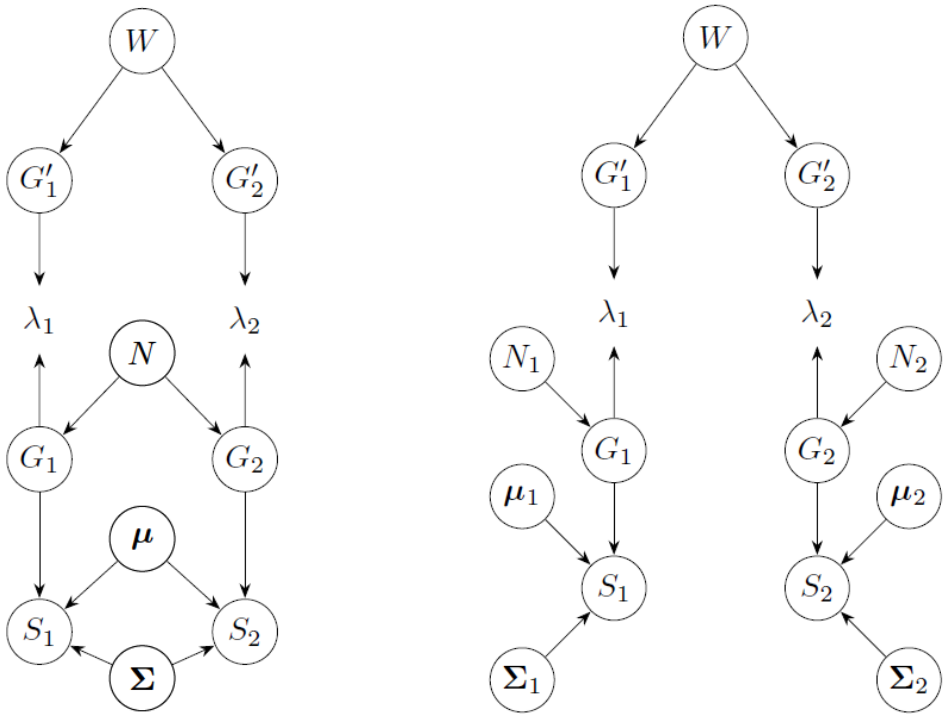
$$\begin{aligned} P(W\ G_1\ G_2\ G'_1\ G'_2\ \lambda_1\ \lambda_2\ S_1\ S_2\ N\ \mu\ \Sigma \mid \pi_C) = \\ P(W)\ P(G'1\ G'2 \mid W)\ P(\lambda_1 \mid G1\ G'1)\ P(\lambda_2 \mid G_2\ G'_2) \\ P(G_1 \mid N)\ P(G_2 \mid N)\ P(S_1 \mid G_1\ \mu\ \Sigma)\ P(S_2 \mid G_2\mu\ \Sigma)\ P(N)\ P(\mu)\ P(\Sigma)\ . \end{aligned} \quad (8.2)$$

$$\begin{aligned} P(W\ G_1\ G_2\ G'_1\ G'_2\ \lambda_1\ \lambda_2\ S_1\ S_2\ N_1\ \mu_1\ \Sigma_1\ N_2\ \mu_2\ \Sigma_2 \mid \pi_D) = \\ P(W)\ P(G'1\ G'2 \mid W)\ P(\lambda_1 \mid G1\ G'1)\ P(\lambda_2 \mid G_2\ G'_2)\ P(G_1 \mid N_1)\ P(G_2 \mid N_2) \\ P(S_1 \mid G_1\ \mu_1\ \Sigma_1)\ P(S_2 \mid G_2\ \mu_2\ \Sigma_2)\ P(N_1)\ P(\mu_1)\ P(\Sigma_1)\ P(N_2)\ P(\mu_2)\ P(\Sigma_2)\ . \end{aligned} \quad (8.3)$$

Les éléments π_C et π_D ont été omis de leur décomposition respective pour ne pas alourdir les équations. La distribution $P(W)$ est le prior sur les mots. Le répertoire lexical $P(G'1\ G'2 \mid W)$ associe chaque mot W aux unités phonétiques correspondantes. De même, chaque unité phonétique est associée à une représentation sensorielle via les répertoires auditifs respectivement $P(S_1 \mid G_1\ \mu\ \Sigma)$ et $P(S_2 \mid G2\ \mu\ \Sigma)$ quand les espaces G_1 et G_2 sont identiques (et partagent donc une même moyenne μ et une même variance Σ) et $P(S_1 \mid G_1\ \mu_1\ \Sigma_1)$ et $P(S_2 \mid G_2\ \mu_2\ \Sigma_2)$ lorsque les espaces G_1 et G_2 sont différents. Les distributions $P(G_1 \mid N)$, $P(G_2 \mid N)$ ainsi que les distributions $P(G_1 \mid N_1)$ et $P(G_2 \mid N_2)$ correspondent au poids associé à chaque gaussienne. Les valeurs λ sont des variables de cohérence, tout comme l'est C dans COSMO. Ainsi, les systèmes de cohérence $P(\lambda_1 \mid G_1\ G'_1)$ et $P(\lambda_2 \mid G_2\ G'_2)$ permettent d'assurer la cohérence entre les unités phonétiques du répertoire lexical et les unités phonétiques du répertoire auditif.

8.2.2.2 L'apprentissage

Dans COSMO WordPhon, l'apprentissage se fait à deux niveaux : un apprentissage lexical durant lequel l'agent apprend à associer les mots aux unités phonétiques et un apprentissage sensoriel durant lequel l'agent apprend à associer les unités phonétiques aux représentations sensorielles. Deux



(a) Version π_C (b) Version π_D

FIGURE 8.2 – Représentation de COSMO WordPhon. Repris de Saghiran (2017)

variantes d'apprentissage sont proposées : un apprentissage bottom-up (BtUp) dans lequel les deux niveaux d'apprentissage s'effectuent de façon indépendante et séquentielle et un apprentissage feedback (FdBk) dans lequel l'apprentissage lexical guide l'apprentissage des phonèmes. Comme dans les versions précédentes de COSMO, l'apprentissage se fait de manière itérative grâce à un maître. Celui-ci envoie à l'agent deux informations : un mot et le signal sensoriel correspondant. Dans les deux apprentissages, afin de mettre à jour ses distributions, l'agent doit inférer des unités phonétiques, ainsi que les paramètres associés, en fonction de ce que le maître lui envoie.

Dans l'apprentissage BtUp, l'inférence des unités phonétiques se fait uniquement grâce au signal sensoriel envoyé par le maître. Ainsi, pour chaque noyau G_i :

$$P(G_i \mid [S_i = s_i]\ N_i\ \mu_i\ \Sigma_i) \propto P(G_i \mid N_i)\ P([S_i = s_i] \mid G_i\ \mu_i\ \Sigma_i)\ . \tag{8.4}$$

L'agent utilise donc uniquement son répertoire auditif et la distribution de pondération pour inférer les unités phonétiques. Grâce à ces informations, il peut donc mettre à jour d'une part son répertoire auditif et sa distribution de pondération à l'aide du signal sensoriel du maître et des noyaux inférés et d'autre part son répertoire lexical à l'aide du mot envoyé par le maître et des noyaux inférés.

Dans l'apprentissage FdBk, l'inférence des unités phonétiques se fait à l'aide des deux informa-

tions envoyées par le maître. Ainsi :

$$P(G_1\ G_2 \mid [S1 = s_1]\ [S_2 = s_2\ ]\ [\lambda_1 = 1]\ [\lambda_2 = 1]\ [W = w]\ N_1\ \mu_1\ \Sigma_1 N_2\ \mu_2\ \Sigma_2) = P(G_1\ G_2 \mid [W = w])\ P(G_1 \mid N_1)\ P([S_1 = s_1] \mid G_1\ N_1\ \mu_1\ \Sigma_1) P(G_2 \mid N_2)\ P([S_2 = s_2] \mid G_2\ N_2\ \mu_2\ \Sigma_2)\ . \quad (8.5)$$

Dans cette variante, l'agent utilise en plus son répertoire lexical pour réaliser l'inférence des noyaux gaussiens. Ensuite, la mise à jour s'effectue de la même manière que précédemment.

8.2.2.3 Les résultats

Dans les simulations réalisés avec COSMO WordPhon, les mots W sont composés de deux voyelles V1 et V2, V1 étant une des sept voyelles [a i u e ɛ o ɔ] et V2 une des cinq voyelles [a i u e ɛ]. Les noyaux gaussiens G représentent donc des voyelles. Le signal sensoriel correspond aux deux formants F1 et F2, que nous avons précédemment utilisés dans plusieurs de nos simulations. Les prototypes sensoriels des voyelles du maître sont les mêmes que ceux du chapitre 5 et sont donc basés sur les données prototypiques définies par Meunier (2007). L'agent possède, de son côté, cinquante noyaux gaussiens.

Un des premiers résultats obtenus est que l'entropie des répertoires lexicaux et auditifs du maître ont convergé en fin d'apprentissage. Cela signifie donc que l'apprentissage réalisé est suffisant pour que les connaissances de l'agent deviennent stables. Un second résultat est que l'agent réussit à apprendre son répertoire auditif dans les deux modèles π_C et π_D et dans les deux apprentissages BtUp et FdBk. Cela signifie que dans toutes ces méthodes, chaque noyau gaussien est positionné dans une des régions sensorielles correspondant à une des voyelles du maître en fin d'apprentissage. En revanche, il y a plusieurs différences en ce qui concerne le répertoire lexical. Nous observons notamment que l'apprentissage du modèle π_C semble plus réaliste que celui du modèle π_D. Pour être plus précis, nous remarquons qu'une séparation des espaces G_1 et G_2 provoquait une divergence des noyaux. Ainsi, le noyau correspondant à une voyelle dans l'espace G_1 ne correspond plus à la même voyelle dans l'espace G_2 et réciproquement. Bien que des effets de séquences puissent exister, ce cas n'est pas généralisable à l'ensemble des unités phonétiques d'un mot.

La comparaison entre l'apprentissage BtUp et FdBk s'effectue par la suite uniquement dans le modèle π_C du fait de son plus grand réalisme. Quelques différences sont d'abord observées entre l'apprentissage BtUp et FdBk durant l'apprentissage auditif. Il semble que l'apprentissage BtUp permet d'inférer le nombre exact de catégories puisqu'en fin d'apprentissage il y a sept noyaux gaussiens avec un poids de pondération non négligeable, chacun correspondant à une des voyelles du maître. À l'inverse, en fin d'apprentissage FdBk, nous observons que même si chaque noyau gaussien correspond à une unique voyelle, il y a plusieurs noyaux gaussien pour une même voyelle. Il y a donc plus de noyaux gaussiens restant en fin d'apprentissage que de voyelles initiales. Ainsi, l'apprentissage lexical, guidant l'inférence des noyaux gaussiens dans l'apprentissage FdBk, semble perturber l'apparition du nombre exact de catégories phonétiques.

En étudiant le répertoire lexical en fin d'apprentissage, il semblerait que cette spécificité de l'apprentissage FdBk puisse s'interpréter comme la distinction entre des allophones. Plus précisément,

en étudiant la distribution $P(G_1\ G_2)$, c'est-à-dire une marginalisation du repertoire lexical sur les mots, nous observons qu'il apparaît 35 combinaisons de noyaux. Cette valeur correspond au nombre de mots du maître, cela signifie donc que l'agent possède bien, dans son répertoire lexical, un nombre de combinaisons et donc de mots similaire à celui du maître. Cependant, parmi ces combinaisons, les noyaux pour réaliser chaque voyelle ne sont pas forcément les mêmes. Ainsi, un même mot se fait toujours avec la même combinaison mais le noyau d'une voyelle ne sert pas forcément pour toutes les combinaisons contenant cette voyelle. Cela peut donc être assimilé à un apprentissage des allophones.

En ce qui concerne l'apprentissage BtUp, en étudiant la même distribution $P(G_1\ G_2)$, nous remarquons que l'agent possède également 35 combinaisons de haute probabilité, correspondant aux 35 mots du maître. Mais, cette fois-ci, le noyau d'une voyelle est utilisé pour chaque combinaison contenant cette voyelle. Par contre, il est observé que le répertoire lexical contient des combinaisons additionnelles de probabilités faibles mais non négligeables, c'est-à-dire que certains noyaux sont, dans de rares cas, utilisés pour réaliser les mots en plus des combinaisons de base. Il est supposé que ce cas survient lorsque le signal sensoriel d'une voyelle est éloigné de son prototype.

En résumé, l'apport de l'information lexicale a à la fois une conséquence « négative » et une conséquence « positive » pour l'émergence des unités phonémiques sous-jacentes. D'un côté, elle peut provoquer des différences de représentations d'une voyelle selon le mot dans lequel elle apparait, ce qui peut permettre l'apprentissage d'allophones qui devront être regroupés autour d'une même catégorie phonémique à un niveau ultérieur. D'un autre côté, elle a pour effet d'améliorer la convergence phonémique et de conduire à un nombre final de combinaisons plus stables (voir aussi Jacobs et al., 1991; McMurray et al., 2009).

8.2.3 COSMO multi-sensoriel

Le troisième axe concerne l'extension de COSMO à des représentations sensorielles autres qu'auditives. Cette dernière partie s'est faite en collaboration avec Jean-François Patri et Pascal Perrier, deux autres membres de notre équipe travaillant également sur des modèles bayésiens davantage focalisés sur le contrôle moteur et la production de la parole.

Bien que COSMO soit un modèle de la communication, nous avons discuté précédemment du fait qu'il reste globalement incomplet dans sa forme de base. Il est notamment difficile de l'utiliser en production du fait qu'il n'existe aucune branche somatosensorielle. De la même manière, la perception reste uniquement focalisée sur l'audition. Or, il a été montré dans de nombreuses publications que la composante visuelle de l'information fournie par le locuteur est également utilisée pour percevoir la parole (voir Schwartz et al., 2010, 1998; Summerfield, 1987, pour des revues). Nous proposons donc un modèle théorique prenant en compte ces deux branches. À ce stade, aucune implémentation n'a encore été réalisée mais nous discutons dans cette partie des possibilités d'études pouvant servir pour ce modèle.

8.2.3.1 Description du modèle

Le modèle, toujours basé sur COSMO, contient le même type de variables et de répertoires que dans le modèle d'origine. Ainsi, nous retrouvons des représentations motrices M, des représentations sensorielles S, des objets O et même la variable de cohérence C. Néanmoins, il y a désormais trois types de représentations sensorielles : S_A sont les représentations auditives, S_S sont les représentations somatosensorielles et S_V sont les représentations visuelles. En ce qui concerne les distributions, le modèle contient, entre autres, un répertoire moteur $P(M \mid O)$, trois répertoires sensoriels $P(S \mid O)$ et trois modèles internes $P(S \mid M)$, un pour chaque type de représentations sensorielles .

Plus précisément, pour bien faire la distinction entre les variables et pour permettre d'écrire correctement le modèle bayésien correspondant, certaines variables sont dupliquées et renommées en conséquence. Ainsi, le répertoire moteur s'écrit dans ce modèle $P(M^O \mid O_M)$ où M^O fait référence aux représentations motrices associées aux phonèmes et O_M aux objets associés aux représentations motrices. Les représentations sensorielles associées aux objets sont nommées S_A^O, S_S^O et S_V^O et sont respectivement associées aux objets O_A, O_S et O_V. Pour faire le lien avec le modèle COSMO, nous pouvons ainsi dire que l'objet O_S correspond désormais à l'objet O_M et l'objet O_L correspond, dans ce modèle, à l'objet O_A. L'ensemble des objets sont, comme dans COSMO, associés à une variable de cohérence C dans le système de cohérence $P(C \mid O_M\ O_A\ O_S\ O_V)$. En ce qui concerne les trois modèles internes, nous définissons la variable M et trois variables sensorielles S_A, S_S et S_V qui forment respectivement les trois répertoires internes $P(S_A \mid M)$, $P(S_S \mid M)$ et $P(S_V \mid M)$. Nous choisissons de bien faire la distinction, d'une part, entre les représentations motrices du répertoire moteur M^O et les représentations motrices des modèles internes M et, d'autre part, entre les représentations sensorielles des répertoires sensoriels S^O et les variables sensorielles des modèles internes S. Néanmoins, ces variables sont respectivement associées à leur homologue par une variable de cohérence, qui a la même fonction que C, et qui se définit dans un système de cohérence.

Par ailleurs, outre ces similitudes avec COSMO, nous souhaitons marquer plus clairement la distinction entre les stimuli de l'environnement (variables physiques) et les valeurs sensorielles induites (variables mentales), et entre les gestes moteurs produits (physiques) et les commandes motrices programmées (variables mentales). Ainsi, la variable $M^{\mathtt{P}}$ correspond à une commande motrice, $S^{\mathtt{P}}$, ou plus précisément $S_A^{\mathtt{P}}$, $S_S^{\mathtt{P}}$ et $S_V^{\mathtt{P}}$, représentent les valeurs sensorielles, $\mathtt{P}_M$ représente le geste moteur réellement produit et $\mathtt{P}_A$, $\mathtt{P}_S$, $\mathtt{P}_V$ représentent les stimuli respectivement auditifs, somatosensoriels et visuels. Les variables $\mathtt{P}$, cette lettre faisant référence à « physic », correspondent donc aux données « physiques », les variables mentales, sensorielles et motrices, étant respectivement les variables S et M. Les distributions résultantes, $P(\mathtt{P}_M \mid M^{\mathtt{P}})$ et $P(S^{\mathtt{P}} \mid \mathtt{P})$, sont nommées systèmes de transformation.

Pour finir, tous les objets, les représentations cinématiques, ainsi que les représentations motrices M sont définies dans des priors. La décomposition globale du modèle résultant, schématisée Fig. 8.3,

est la suivante :

$$
\begin{aligned}
&P(\, O_M\, O_A\, O_S\, O_V\, M^O\, M\, M^{\mathtt{P}}\, S_A^O\, S_S^O\, S_V^O\, S_A\, S_S\, S_V\, S_A^{\mathtt{P}}\, S_S^{\mathtt{P}}\, S_V^{\mathtt{P}} \\
&\quad \mathtt{P}_M\, \mathtt{P}_A\, \mathtt{P}_S\, \mathtt{P}_V\, C\, \lambda_{MO}\, \lambda_{M\mathtt{P}}\, \lambda_{SAO}\, \lambda_{SSO}\, \lambda_{SVO}\, \lambda_{SA\mathtt{P}}\, \lambda_{SS\mathtt{P}}\, \lambda_{SV\mathtt{P}}) \\
&\quad = P(O_M)\, P(O_A)\, P(O_S)\, P(O_V)\, P(\mathtt{P}_A)\, P(\mathtt{P}_S)\, P(\mathtt{P}_V)\, P(M)\, P(M^{\mathtt{P}}) P(M^O \mid O_M) \\
&\qquad P(S_A^O \mid O_A)\, P(S_S^O \mid O_S)\, P(S_V^O \mid O_V)\, P(S_A \mid M)\, P(S_S \mid M)\, P(S_V \mid M) \\
&\qquad P(\mathtt{P}_M \mid M^{\mathtt{P}})\, P(S_A^{\mathtt{P}} \mid \mathtt{P}_P)\, P(S_S^{\mathtt{P}} \mid \mathtt{P}_S)\, P(S_V^{\mathtt{P}} \mid \mathtt{P}_V)\, P(C \mid O_M\, O_A\, O_S\, O_V) \\
&\qquad P(\lambda_{MO} \mid M\, M^O)\, P(\lambda_{M\mathtt{P}} \mid M\, M^{\mathtt{P}})\, P(\lambda_{SAO} \mid S_A\, S_A^O)\, P(\lambda_{SSO} \mid S_S\, S_S^O) \\
&\qquad P(\lambda_{SVO} \mid S_V\, S_V^O)\, P(\lambda_{SA\mathtt{P}} \mid S_A\, S_A^{\mathtt{P}})\, P(\lambda_{SS\mathtt{P}} \mid S_S\, S_S^{\mathtt{P}})\, P(\lambda_{SV\mathtt{P}} \mid S_V\, S_V^{\mathtt{P}})\, .
\end{aligned}
$$

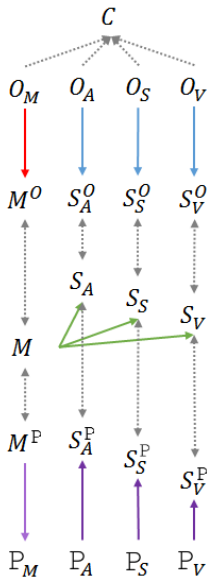


FIGURE 8.3 – Représentation du modèle COSMO « multi sensoriel ». Les distributions sont schématisées par des flèches de différentes couleurs : (rouge) le répertoire moteur, (bleu) les répertoires sensoriels, (vert) les modèles internes, (violet) les systèmes de transformations, (gris) les systèmes de cohérence. Pour ne pas alourdir la figure les λ liant d'une part les représentations motrices et d'autre part les représentations sensorielles sont schématisés par des doubles flèches

Construit comme tel, ce modèle possède donc deux améliorations principales par rapport au modèle COSMO d'origine : d'une part, la prise en compte d'une branche sensorielle non plus uniquement auditive mais également somatosensorielle et visuelle et d'autre part, la différenciation entre les éléments physiques du modèle et les représentations internes.

8.2.3.2 Discussion autour du modèle

Bien qu'il ne soit pas encore implémenté, le premier objectif de ce modèle est d'être intégrateur. Il nous permet ainsi de relier les modélisations réalisées avec COSMO, présentées dans cette thèse ou dans de précédentes études (Barnaud et al., en révision; Laurent et al., 2017; Moulin-Frier et al.,

2012), et les modélisations réalisées par Jean-François Patri utilisant un modèle permettant d'étudier efficacement le contrôle moteur (voir par exemple Patri et al., 2016, soumis).

Ce modèle fournit un cadre naturel pour étudier les processus d'intégration sensorimotrice ainsi que la relation entre les représentations motrices et les représentations sensorielles. De façon plus concrète, il permettrait, par exemple, d'étudier des phénomènes d'interaction audiovisuelle et le rôle que semble y jouer le lien sensorimoteur (voir Treille, 2017, pour plus de détails).

Dans un deuxième temps, ce modèle nous permettrait d'étudier plus efficacement que dans COSMO le lien entre la perception et la production. Dans cette optique, il pourrait, par exemple, servir à mieux comprendre la nature des interactions entre les sons et les perturbations somatosensorielles telles qu'elles sont étudiées par Ito et al. (2009) ou mieux comprendre la nature du lien sensorimoteur observé par Shiller et al. (2009) ou Lametti et al. (2014). Un exemple de modélisation est d'ailleurs proposé par Patri et al. (soumis) sur ce sujet.

8.3 Conclusion

Cette thèse avait pour objectif d'approfondir les connaissances déjà existantes sur la nature et la structure cognitive des unités distinctives, en utilisant le modèle bayésien COSMO. En intégrant, au sein d'un même modèle, les représentations auditives et motrices des unités distinctives, nous avons pu traiter différents aspects des unités distinctives : leur apprentissage, la complémentarité des représentations, la corrélation des idiosyncrasies en perception et en production, et l'acquisition des différentes structures cognitives.

Nos travaux nous ont permis de mettre en évidence l'existence d'une complémentarité entre les représentations auditives et motrices. En nous basant sur les théories de la perception, nous avons d'abord énoncé la propriété « bande étroite/bande large ». Par la suite, à travers l'étude de l'acquisition de la structure phonémique des unités, nous avons observé que les consonnes semblent mieux apprises à l'aide des représentations motrices et les voyelles semblent mieux apprises à l'aide des représentations auditives. Par ailleurs, toutes ces études, notamment celles sur les idiosyncrasies, nous ont permis de souligner le rôle et l'importance des représentations motrices en perception.

Pris ensemble, nos résultats indiquent également que le modèle COSMO est adapté pour étudier les unités distinctives, mais, comme le montrent les trois extensions présentées dans ce chapitre, il peut également s'étendre à d'autres domaines. COSMO offre un cadre au riche potentiel et cette thèse n'est qu'une étape vers le développement de modèles complets de la communication parlée, incluant l'apprentissage, le fonctionnement en ligne, et le lien entre la phonétique et la phonologie.

CHAPITRE 9

Annexe : Précisions sur la mise à jour des paramètres dans COSMO

Cette partie donne quelques détails supplémentaires sur la phase de mise à jour des paramètres réalisée dans la majorité des différents apprentissages de COSMO. Nous commençons par décrire la phase de mise à jour des gaussiennes générique, servant dans la majorité des versions de COSMO. Ensuite, nous présentons le cas particulier de la phase de mise à jour utilisée dans COSMO SylPhon.

9.1 Mise à jour des paramètres dans COSMO générique

Pour rappel, les distributions apprises dans la majorité des versions de COSMO, c'est-à-dire le répertoire auditif $P(S \mid O_L)$, le répertoire moteur $P(M \mid O_S)$ et le modèle interne $P(S \mid M)$, sont des ensembles de gaussiennes.

Chaque distribution gaussienne possède lors de son apprentissage trois paramètres : la somme des valeurs observées s, la somme des carrés des valeurs observées s^2 et le nombre de valeurs observées n. Ces trois paramètres permettent de calculer les deux paramètres classiques d'une distribution gaussienne : la moyenne μ et la matrice de covariance Σ (ou l'écart-type σ dans le cas unidimensionnel).

À l'initialisation, les distributions gaussiennes approximent une distribution uniforme. Nous décrivons cela en définissant une gaussienne dont la moyenne est localisée au centre de l'espace, et dont la variance est grande. Cela revient à considérer un ensemble de points répartis aléatoirement dans l'espace puis à calculer leur moyenne et leur dispersion. Ceci nous permet d'initialiser facilement aussi bien la moyenne et la matrice de covariance que les trois paramètres s, s^2 et n.

Lors de la mise à jour d'une distribution gaussienne g avec la valeur v, les paramètres μ et Σ sont calculés de la façon suivante :

$$\begin{array}{llllll} n & = & n + nb\,, & s & = & s + nb * v\,, \\ \mu & = & \dfrac{s}{n}\,, & \Sigma & = & \dfrac{s^2}{n} - \dfrac{s \times s^T}{n^2}\,. \end{array} \qquad \begin{array}{lll} s^2 & = & s^2 + nb * v \times v^T\,, \end{array} \tag{9.1}$$

La valeur nb symbolise le poids que nous souhaitons accorder aux nouvelles valeurs apprises. Elle diffère selon les simulations et selon les apprentissages. Il est notamment possible de la comparer à la valeur de n initiale : une grande valeur de nb par rapport au n initial permet de faire évoluer rapidement la moyenne et la matrice de covariance des gaussiennes. À l'inverse, une faible valeur de

nb par rapport à n ralentit l'apprentissage. Par ailleurs, dans nos simulations, la valeur de nb, une fois fixée, reste constante durant l'apprentissage. Cela signifie que les nouvelles valeurs ont de moins en moins d'importance au fur et à mesure de l'apprentissage. La distribution gaussienne se stabilise donc peu à peu et il devient plus difficile en fin d'apprentissage de la faire évoluer.

9.2 Mise à jour des paramètres dans COSMO SylPhon

Les répertoires auditifs, les répertoires moteurs et les modèles internes sont toujours des ensembles de gaussiennes dans COSMO SylPhon. De ce fait, ces distributions possèdent les mêmes trois paramètres : la somme des valeurs observées s, la somme des carrés des valeurs observées s^2 et le nombre de valeurs observées n, permettant de calculer les paramètres classiques μ et Σ.

Les étapes de l'apprentissage sensorimoteur de COSMO SylPhon sont très proches de celles des autres versions de COSMO. De ce fait, la phase de mise à jour des gaussiennes est similaire à celle présentée ci-dessus. En revanche, l'apprentissage auditif et l'apprentissage moteur diffèrent des autres versions de COSMO sur deux points importants. Premièrement, ces apprentissages concernent non seulement les répertoires respectifs auditifs et moteurs mais également les priors sur les noyaux, ce qui les rapproche d'un apprentissage d'une mixture de gaussienne plutôt que d'un ensemble de gaussiennes. Ainsi, chaque distribution gaussienne est associée à un noyau gaussien dont la pondération est donnée par le prior. Deuxièmement, dans COSMO Sylphon, l'apprentissage est proche d'un algorithme EM itératif dans lequel un noyau est inféré à chaque nouvelle itération alors que l'identité du noyau est directement donné dans les précédentes versions de COSMO. En conséquence, le choix de la distribution gaussienne à mettre à jour dépend non seulement de la localisation dans l'espace (μ) et de la variance (Σ) de chaque gaussienne mais également de la pondération du noyau, ce qui est fourni par le prior. Lors de la phase de mise à jour des apprentissages auditifs et moteurs, il est donc nécessaire d'effectuer simultanément la mise à jour des distributions gaussiennes et la mise à jour des poids.

Pour rappel, les priors des noyaux possèdent un nombre d'observation obs_n qui est incrémenté, pour le noyau sélectionné, à chaque mise à jour (voir Eq. 7.4). L'initialisation et l'incrémentation de ce paramètre doit être réalisé de telle sorte à ce qu'il ne nuise pas à la mise à jour de la distribution gaussienne. En effet, si la valeur de obs_n augmente très rapidement, alors certains noyaux auront une très forte pondération et pourront être choisis dans les itérations suivantes au détriment de leur localisation et de leur variance dans l'espace. Inversement, si la valeur de obs_n n'augmente pas assez rapidement, la pondération des gaussiennes devient inutile. Ainsi, il faut assurer un équilibre entre les dynamiques des apprentissages des paramètres des noyaux et de leur prior.

Durant l'apprentissage moteur, l'équilibre entre ces dynamiques a été choisi empiriquement.. En revanche, une méthode plus spécifique a été mise en place durant l'apprentissage auditif. Nous avons pris en compte le fait qu'à l'initialisation, les gaussiennes ont une grande variance et sont situées au centre de l'espace. Lors de l'apprentissage, leurs variances diminuent et elles se déplacent vers une portion spécifique de l'espace. Cependant, si deux unités distinctives à découvrir se trouvent dans deux portions de l'espace très proches l'une de l'autre, il est probable qu'une même gaussienne se localise entre les deux et conserve une variance permettant d'englober ces deux portions. Afin d'éviter ce cas

de figure, une solution consiste à mettre à jour dès le début de l'apprentissage le plus grand nombre de gaussiennes afin qu'elles se spécifient chacune dans une petite portion de l'espace. Pour ce faire, la stratégie utilisée consiste à favoriser les gaussiennes ayant une grande variance.

Choisir les gaussiennes de plus grande variance peut se faire à l'aide du prior sur les noyaux. Nous définissons un nouveau paramètre det qui contient la racine carré du déterminant de la matrice de covariance de la gaussienne associée. La mise à jour du prior pour le noyau n sélectionné se fait donc ainsi :

$$\begin{aligned} obs_n &= obs_n + 1 \,, \\ det(n) &= \sqrt{|\Sigma(n)|} \,, \\ P([N = n] \mid \pi_N) &= \frac{obs_n \times det(n)}{\sum_{G=i} obs_i \times det(i)} \,. \end{aligned}$$

Cette stratégie est employée pendant 25 000 itérations. Par la suite et jusqu'à la fin de l'apprentissage, la méthode initiale sans prise en compte de la variance est utilisée.

Bibliographie

Alho, K., Connolly, J. F., Cheour, M., Lehtokoski, A., Huotilainen, M., Virtanen, J., Aulanko, R., et Ilmoniemi, R. J. (1998). Hemispheric lateralization in preattentive processing of speech sounds. Neuroscience Letters, 258(1) :9–12. (Cité en page 24.)

Anderson, J. L., Morgan, J. L., et White, K. S. (2003). A statistical basis for speech sound discrimination. Language and Speech, 46(2-3) :155–182. (Cité en page 33.)

Aubanel, V. (2011). Variation phonologique régionale en interaction conversationnelle. PhD thesis, Aix Marseille 1. (Cité en page 15.)

Babel, M. E. (2009). Phonetic and social selectivity in speech accommodation. PhD thesis, University of California, Berkeley. (Cité en page 15.)

Badino, L., Canevari, C., Fadiga, L., et Metta, G. (2014). An auto-encoder based approach to unsupervised learning of subword units. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014), pages 7634–7638. (Cité en page 44.)

Badino, L., Canevari, C., Fadiga, L., et Metta, G. (2016). Integrating articulatory data in deep neural network-based acoustic modeling. Computer Speech & Language, 36 :173–195. (Cité en page 46.)

Bailly, G. (1997). Learning to speak. Sensori-motor control of speech movements. Speech Communication, 22(2-3) :251–267. (Cité en pages 58, 60 et 185.)

Baker, E., Blumstein, S. E., et Goodglass, H. (1981). Interaction between phonological and semantic factors in auditory comprehension. Neuropsychologia, 19(1) :1–15. (Cité en page 13.)

Baranes, A. et Oudeyer, P.-Y. (2010). Intrinsically motivated goal exploration for active motor learning in robots : A case study. In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2010), pages 1766–1773. (Cité en page 57.)

Baranes, A. et Oudeyer, P.-Y. (2013). Active learning of inverse models with intrinsically motivated goal exploration in robots. Robotics and Autonomous Systems, 61(1) :49–73. (Cité en page 57.)

Barnaud, M.-L., Bessière, P., Diard, J., et Schwartz, J.-L. (2017, en révision). Reanalyzing neurocognitive data on the role of the motor system in speech perception within COSMO, a Bayesian perceptuo-motor model of speech communication. (Cité en pages 187 et 191.)

Barnaud, M.-L., Diard, J., Bessière, P., et Schwartz, J.-L. (2015a). Modeling the concurrent development of speech perception and production in a Bayesian framework. In The 5th Joint IEEE International Conference on Developmental and Learning and on Epigenetic Robotics (ICDL-Epirob 2015), pages 248–249. Poster.

Barnaud, M.-L., Diard, J., Bessière, P., et Schwartz, J.-L. (2015b). Modeling the concurrent development of speech perception and production in a Bayesian framework. In Workshop on Probabilistic Inference and the Brain. Poster.

Barnaud, M.-L., Diard, J., Bessière, P., et Schwartz, J.-L. (2016a). Assessing Idiosyncrasies in a Bayesian Model of Speech Communication. In Proceedings of Interspeech 2016, pages 2080–2084.

Barnaud, M.-L., Diard, J., Bessière, P., et Schwartz, J.-L. (2017a). Assessing phonological learning in COSMO, a Bayesian model of speech communication. In The 7th Joint IEEE International Conference on Developmental and Learning and on Epigenetic Robotics (ICDL-Epirob 2017).

Barnaud, M.-L., Diard, J., Bessière, P., et Schwartz, J.-L. (2017b). Perceptuo-motor speech units in the brain with COSMO, a Bayesian model of communication. In Proceedings of the 11th International Seminar on Speech Production. (Cité en page 196.)

Barnaud, M.-L., Diard, J., Bessière, P., et Schwartz, J.-L. (en révision). Computational simulations of perceptuo-motor idiosyncrasies support the involvement of motor knowledge in speech perception. (Cité en page 202.)

Barnaud, M.-L., Laurent, R., Bessière, P., Diard, J., et Schwartz, J.-L. (2015c). Modeling concurrent development of speech perception and production in a Bayesian framework. In Workshop on Infant Language Development (WILD), Stockholm, Sweden.

Barnaud, M.-L., Schwartz, J.-L., Diard, J., et Bessière, P. (2016b). Sensorimotor learning in a Bayesian computational model of speech communication. In The 6th Joint IEEE International Conference on Developmental and Learning and on Epigenetic Robotics (ICDL-Epirob 2016), Cergy-Pontoise, France.

Bastiaansen, M. et Hagoort, P. (2006). Oscillatory neuronal dynamics during language comprehension. Progress in Brain Research, 159 :179–196. (Cité en page 25.)

Bazzi, I. et Glass, J. (2000). Heterogeneous lexical units for automatic speech recognition : preliminary investigations. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2000), volume 3, pages 1257–1260. (Cité en page 53.)

Bell-Berti, F., Raphael, L. J., Pisoni, D. B., et Sawusch, J. R. (1979). Some relationships between speech production and perception. Phonetica, 36(6) :373–383. (Cité en page 19.)

Beller, H. K. (1971). Priming : effects of advance information on matching. Journal of Experimental Psychology : General, 87(2) :176–182. (Cité en page 23.)

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., et Wellekens, C. (2007). Automatic speech recognition and speech variability : A review. Speech Communication, 49(10) :763–786. (Cité en page 44.)

Bertelson, P. (1986). The onset of literacy : Cognitive processes in reading acquisition. Cambridge, MA : MIT Press. (Cité en page 22.)

Bertoncini, J., Bijeljac-Babic, R., Jusczyk, P. W., Kennedy, L. J., et Mehler, J. (1988). An investigation of young infants' perceptual representations of speech sounds. Journal of Experimental Psychology : General, 117(1) :21–33. (Cité en page 40.)

Bertoncini, J. et Mehler, J. (1981). Syllables as units in infant speech perception. Infant Behavior and Development, 4 :247–260. (Cité en page 39.)

Bessière, P., Mazer, E., Ahuactzin, J. M., et Mekhnacha, K. (2013). Bayesian Programming. CRC Press, Boca Raton, Florida. (Cité en pages vi, 62 et 63.)

Best, C. T., McRoberts, G. W., LaFleur, R., et Silver-Isenstadt, J. (1995). Divergent developmental patterns for infants' perception of two nonnative consonant contrasts. Infant Behavior and Development, 18(3) :339–350. (Cité en page 29.)

Bijeljac-Babic, R., Bertoncini, J., et Mehler, J. (1993). How do 4-day-old infants categorize multisyllabic utterances ? Developmental Psychology, 29(4) :711–721. (Cité en page 39.)

Bilmes, J. A. (1998). A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. Technical Report TR 97-021, International Computer Science Institute. (Cité en page 55.)

Binder, J. R., Liebenthal, E., Possing, E. T., Medler, D. A., et Ward, B. D. (2004). Neural correlates of sensory and decision processes in auditory object identification. Nature Neuroscience, 7(3) :295–301. (Cité en pages 13 et 92.)

Bloom, K. (1975). Social elicitation of infant vocal behavior. Journal of Experimental Child Psychology, 20(1) :51–58. (Cité en page 36.)

Bloom, K. (1988). Quality of adult vocalizations affects the quality of infant vocalizations. Journal of Child Language, 15(3) :469–480. (Cité en page 36.)

Bloom, K., Russell, A., et Wassenberg, K. (1987). Turn taking affects the quality of infant vocalizations. Journal of Child Language, 14(2) :211–227. (Cité en page 36.)

Blumstein, S. E. (1995). The neurobiology of the sound structure of language. In Gazzaniga, M. S., editor, The Cognitive Neurosciences, pages 915–929. Cambridge, MA : MIT Press. (Cité en page 13.)

Bohland, J. W., Bullock, D., et Guenther, F. H. (2010). Neural representations and mechanisms for the performance of simple speech sequences. Journal of Cognitive Neuroscience, 22(7) :1504–1529. (Cité en page 51.)

Bosch, L. et Sebastián-Gallés, N. (2003). Language experience and the perception of a voicing contrast in fricatives : Infant and adult data. In Proceedings of the 15th International Conference of Phonetic Sciences, pages 1987–1990. (Cité en page 29.)

Bourguignon, N. J., Baum, S. R., et Shiller, D. M. (2014). Lexical-perceptual integration influences sensorimotor adaptation in speech. Frontiers in Human Neuroscience, 8(208) :1–9. (Cité en page 16.)

Brandl, H., Wrede, B., Joublin, F., et Goerick, C. (2008). A self-referential childlike model to acquire phones, syllables and words from acoustic speech. In IEEE International Conference on Developmental and Learning (ICDL 2008), pages 31–36. (Cité en page 60.)

Bruck, M., Treiman, R., et Caravolas, M. (1995). Role of the syllable in the processing of spoken English : Evidence from a nonword comparison task. Journal of Experimental Psychology : Human Perception and Performance, 21(3) :469–479. (Cité en page 22.)

Buchsbaum, B. R., Hickok, G., et Humphries, C. (2001). Role of left posterior superior temporal gyrus in phonological processing for speech perception and production. Cognitive Science, 25(5) :663–678. (Cité en page 25.)

Callan, D. E., Callan, A., et Jones, J. A. (2014). Speech motor brain regions are differentially recruited during perception of native and foreign-accented phonemes for first and second language listeners. Frontiers in Neuroscience, 8(275) :1–15. (Cité en page 13.)

Callan, D. E., Jones, J. A., Callan, A. M., et Akahane-Yamada, R. (2004). Phonetic perceptual identification by native-and second-language speakers differentially activates brain regions involved with acoustic phonetic processing and those involved with articulatory–auditory/orosensory internal models. NeuroImage, 22(3) :1182–1194. (Cité en page 13.)

Canevari, C., Badino, L., D'Ausilio, A., Fadiga, L., et Metta, G. (2013). Modeling speech imitation and ecological learning of auditory-motor maps. Frontiers in Psychology, 4(364) :1–12. (Cité en pages 46 et 189.)

Carroll, J. M., Snowling, M. J., Stevenson, J., et Hulme, C. (2003). The development of phonological awareness in preschool children. Developmental Psychology, 39(5) :913–923. (Cité en page 40.)

Castellini, C., Badino, L., Metta, G., Sandini, G., Tavella, M., Grimaldi, M., et Fadiga, L. (2011). The use of phonetic motor invariants can improve automatic phoneme discrimination. PLoS One, 6(9) :e24055. (Cité en pages 46 et 189.)

Cheour, M., Ceponiene, R., Lehtokoski, A., Luuk, A., Allik, J., Alho, K., et Näätänen, R. (1998). Development of language-specific phoneme representations in the infant brain. Nature Neuroscience, 1(5) :351–353. (Cité en pages 39 et 41.)

Cheung, C., Hamilton, L. S., Johnson, K., et Chang, E. F. (2016). The auditory representation of speech sounds in human motor cortex. ELife, 5 :e12577. (Cité en pages 195 et 196.)

Chomsky, N. (1959). A review of BF Skinner's verbal behavior. Language, 35(1) :26–58. (Cité en page 33.)

Clayards, M., Tanenhaus, M. K., Aslin, R. N., et Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. Cognition, 108(3) :804–809. (Cité en pages 45 et 51.)

Coen, M. H. (2006). Self-supervised acquisition of vowels in American English. In Proceedings of the National Conference on Artificial Intelligence, volume 21, pages 1451–1456. (Cité en pages 56 et 60.)

Colas, F., Diard, J., et Bessière, P. (2010). Common Bayesian models for common cognitive issues. Acta Biotheoretica, 58(2-3) :191–216. (Cité en page 62.)

Conboy, B. T., Rivera-Gaxiola, M., Klarman, L., Aksoylu, E., et Kuhl, P. K. (2005). Associations between native and nonnative speech sound discrimination and language development at the end of the first year. In Supplement to the Proceedings of the 29th Boston University Conference on Language Development. (Cité en page 29.)

Conboy, B. T., Rivera-Gaxiola, M., Silva-Pereyra, J., et Kuhl, P. K. (2008a). Event-related potential studies of early language processing at the phoneme, word, and sentence levels. In Friederici, A. D. et Guillaume, T., editors, Early Language Development : Bridging Brain and Behaviour, volume 5, pages 23–64. Amsterdam : John Benjamins Publishing Company. (Cité en page 29.)

Conboy, B. T., Sommerville, J. A., et Kuhl, P. K. (2008b). Cognitive control factors in speech perception at 11 months. Developmental Psychology, 44(5) :1505–1512. (Cité en page 35.)

Content, A. et Frauenfelder, U. H. (2002). La syllabe comme unité de perception de la parole : un état de la question. 24ème Journées d'Études sur la Parole. (Cité en page 21.)

Content, A., Meunier, C., Kearns, R. K., et Frauenfelder, U. H. (2001). Sequence detection in pseudowords in French : Where is the syllable effect ? Language and Cognitive Processes, 16(5-6) :609–636. (Cité en page 22.)

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., et Gerstman, L. J. (1952). Some experiments on the perception of synthetic speech sounds. The Journal of the Acoustical Society of America, 24(6) :597–606. (Cité en page 10.)

Cooper, W. E. (1979). Speech perception and production : Studies in selective adaptation. Ablex Publishing Corporation, Norwood, NJ. (Cité en page 15.)

Cooper, W. E., Ebert, R. R., et Cole, R. A. (1976). Speech perception and production of the consonant cluster [st]. Journal of Experimental Psychology : Human Perception and Performance, 2(1) :105–114. (Cité en page 15.)

Cooper, W. E. et Lauritsen, M. R. (1974). Feature processing in the perception and production of speech. Nature, 252(5479) :121–123. (Cité en page 15.)

Crompton, A. (1981). Syllables and segments in speech production. Linguistics, 19(7-8) :663–716. (Cité en page 23.)

Cutler, A., McQueen, J. M., Norris, D., et Somejuan, A. (2001). The roll of the silly ball. In Dupoux, E., editor, Language, Brain and Cognitive Development : Essays in honor of Jacques Mehler, pages 181–194. Cambridge, MA : MIT Press. (Cité en page 23.)

Cutler, A., Mehler, J., Norris, D., et Segui, J. (1983). A language-specific comprehension strategy. Nature, 304(5922) :159–160. (Cité en page 22.)

Cutler, A., Mehler, J., Norris, D., et Segui, J. (1986). The syllable's differing role in the segmentation of French and English. Journal of Memory and Language, 25(4) :385–400. (Cité en pages 21 et 22.)

Cutler, A. et Norris, D. (1979). Monitoring sentence comprehension. In Cooper, W. E. et Walker, E. C. T., editors, Sentence processing : Psycholinguistic studies presented to Merrill Garrett. Hillsdale, NJ : Lawrence Erlbaum Associates. (Cité en page 51.)

D'Ausilio, A., Bufalari, I., Salmas, P., et Fadiga, L. (2012a). The role of the motor system in discriminating normal and degraded speech sounds. Cortex, 48(7) :882–887. (Cité en page 13.)

D'Ausilio, A., Craighero, L., et Fadiga, L. (2012b). The contribution of the frontal lobe to the perception of speech. Journal of Neurolinguistics, 25(5) :328–335. (Cité en page 14.)

D'Ausilio, A., Pulvermüller, F., Salmas, P., Bufalari, I., Begliomini, C., et Fadiga, L. (2009). The motor somatotopy of speech perception. Current Biology, 19(5) :381–385. (Cité en page 13.)

Davis, B. L. et MacNeilage, P. F. (1995). The articulatory basis of babbling. Journal of Speech, Language, and Hearing Research, 38(6) :1199–1211. (Cité en page 31.)

de Boer, B. et Kuhl, P. K. (2003). Investigating the role of infant-directed speech with a computer model. Acoustics Research Letters Online, 4(4) :129–134. (Cité en pages 55, 59, 60, 70, 71 et 177.)

de Boysson-Bardies, B. et Hallé, P. (2004). Des « capacités précoces » à l'élaboration du premier lexique. In Ferrand, L. et Grainger, J., editors, Psycholinguistique cognitve, chapter 15, pages 291–305. De Boeck Supérieur, Louvain-la-Neuve. (Cité en page 29.)

de Boysson-Bardies, B., Hallé, P., Sagart, L., et Durand, C. (1989). A crosslinguistic investigation of vowel formants in babbling. Journal of Child Language, 16(1) :1–17. (Cité en page 35.)

de Boysson-Bardies, B., Sagart, L., et Durand, C. (1984). Discernible differences in the babbling of infants according to target language. Journal of Child Language, 11(1) :1–15. (Cité en page 35.)

de Boysson-Bardies, B. et Vihman, M. M. (1991). Adaptation to language : Evidence from babbling and first words in four languages. Language, 67(2) :297–319. (Cité en page 35.)

Decoene, S. (1993). Testing the speech unit hypothesis with the primed matching task : phoneme categories are perceptually basic. Attention, Perception, & Psychophysics, 53(6) :601–616. (Cité en page 23.)

Dehaene-Lambertz, G. et Baillet, S. (1998). A phonological representation in the infant brain. NeuroReport, 9(8) :1885–1888. (Cité en page 38.)

Dehaene-Lambertz, G. et Peña, M. (2001). Electrophysiological evidence for automatic phonetic processing in neonates. NeuroReport, 12(14) :3155–3158. (Cité en page 38.)

Delattre, P. C., Liberman, A. M., et Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. The Journal of the Acoustical Society of America, 27(4) :769–773. (Cité en page 10.)

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. Psychological Review, 93(3) :283. (Cité en page 23.)

Dempster, A. P., Laird, N. M., et Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society : Series B (Statistical Methodology), 39(1) :1–38. (Cité en page 55.)

DePaolis, R. A., Vihman, M. M., et Keren-Portnoy, T. (2011). Do production patterns influence the processing of speech in prelinguistic infants ? Infant Behavior and Development, 34(4) :590–601. (Cité en page 37.)

Diard, J. (2015). Bayesian Algorithmic Modeling in Cognitive Science. Habilitation à diriger des recherches (HDR), Université Grenoble-Alpes. (Cité en page 190.)

Diehl, R. L., Lotto, A. J., et Holt, L. L. (2004). Speech perception. Annual Review of Psychology, 55 :149–179. (Cité en pages 11 et 14.)

Dillon, B., Dunbar, E., et Idsardi, W. J. (2013). A single-stage approach to learning phonological categories : Insights from Inuktitut. Cognitive Science, 37(2) :344–377. (Cité en pages 56, 59 et 60.)

Dole, M., Vilain, C., Vilain, A., et Loevenbruck, H. & Schwartz, J.-L. (en préparation). (Cité en page 93.)

Dupoux, E., Beraud-Sudreau, G., et Sagayama, S. (2011). Templatic features for modeling phoneme acquisition. In Proceedings of the 31st Annual Conference of the Cognitive Science Society, pages 219–224. (Cité en page 60.)

Eckers, C., Kröger, B. J., Sass, K., et Heim, S. (2013). Neural representation of the sensorimotor speech–action-repository. Frontiers in Human Neuroscience, 7(121) :1–10. (Cité en page 46.)

Eilers, R. E., Gavin, W., et Wilson, W. R. (1979). Linguistic experience and phonemic perception in infancy : A crosslinguistic study. Child Development, 50(1) :14–18. (Cité en page 28.)

Eilers, R. E. et Minifie, F. D. (1975). Fricative discrimination in early infancy. Journal of Speech, Language, and Hearing Research, 18(1) :158–167. (Cité en page 27.)

Eimas, P. D. (1975). Auditory and phonetic coding of the cues for speech : Discrimination of the [rl] distinction by young infants. Attention, Perception, & Psychophysics, 18(5) :341–347. (Cité en page 27.)

Eimas, P. D. et Corbit, J. D. (1973). Selective adaptation of linguistic feature detectors. Cognitive Psychology, 4(1) :99–109. (Cité en page 15.)

Eimas, P. D. et Miller, J. L. (1980). Discrimination of information for manner of articulation. Infant Behavior and Development, 3 :367–375. (Cité en page 27.)

Eimas, P. D., Siqueland, E. R., Juscyk, P., et Vigorito, J. (1971). Speech perception in infants. Science, 171(3968) :303–306. (Cité en page 27.)

Ejiri, K. (1998). Relationship between rhythmic behavior and canonical babbling in infant vocal development. Phonetica, 55(4) :226–237. (Cité en page 31.)

Ejiri, K. et Masataka, N. (2001). Co-occurences of preverbal vocal behavior and motor action in early infancy. Developmental Science, 4(1) :40–48. (Cité en page 31.)

Elbers, L. (1982). Operating principles in repetitive babbling : A cognitive continuity approach. Cognition, 12(1) :45–63. (Cité en page 31.)

Fadiga, L., Craighero, L., Buccino, G., et Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles : a TMS study. European Journal of Neuroscience, 15(2) :399–402. (Cité en page 13.)

Fagan, M. K. (2009). Mean length of utterance before words and grammar : Longitudinal trends and developmental implications of infant vocalizations. Journal of Child Language, 36(3) :495–527. (Cité en page 31.)

Feldman, N. H., Griffiths, T. L., Goldwater, S., et Morgan, J. L. (2013a). A role for the developing lexicon in phonetic category acquisition. Psychological Review, 120(4) :751–778. (Cité en pages 56, 59 et 71.)

Feldman, N. H., Griffiths, T. L., et Morgan, J. L. (2009a). The influence of categories on perception : Explaining the perceptual magnet effect as optimal statistical inference. Psychological Review, 116(4) :752–782. (Cité en pages 71 et 177.)

Feldman, N. H., Griffiths, T. L., et Morgan, J. L. (2009b). Learning phonetic categories by learning a lexicon. In Proceedings of the 31st Annual Conference of the Cognitive Science Society, pages 2208–2213. (Cité en pages 56 et 60.)

Feldman, N. H., Myers, E. B., et White, K. S. (2011). Learners use word-level statistics in phonetic category acquisition. In Proceedings of the 35th Boston University Conference on Language Development, pages 197–209. (Cité en page 56.)

Feldman, N. H., Myers, E. B., White, K. S., Griffiths, T. L., et Morgan, J. L. (2013b). Word-level information influences phonetic learning in adults and infants. Cognition, 127(3) :427–438. (Cité en page 184.)

Field, T. M., Woodson, R., Greenberg, R., et Cohen, D. (1983). Facial expression by neonates. Annual Progress in Child Psychiatry and Child Development, 16 :119–125. (Cité en page 36.)

Finney, S. A., Protopapas, A., et Eimas, P. D. (1996). Attentional allocation to syllables in American English. Journal of Memory and Language, 35(6) :893–909. (Cité en page 22.)

Foss, D. J. et Swinney, D. A. (1973). On the psychological reality of the phoneme : Perception, identification, and consciousness. Journal of Verbal Learning and Verbal Behavior, 12(3) :246–257. (Cité en page 21.)

Fowler, A. E., Brady, S., et Shankweiler, D. P. (1991). How early phonological development might set the stage for phoneme awareness. Phonological processes in literacy : A tribute to Isabelle Y. Liberman, 106 :97–117. (Cité en page 40.)

Fowler, C. A. (1984). Segmentation of coarticulated speech in perception. Attention, Perception, & Psychophysics, 36(4) :359–368. (Cité en page 21.)

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. Journal of Phonetics, 14(1) :3–28. (Cité en pages 11 et 65.)

Fowler, C. A., Brown, J. M., Sabadini, L., et Weihing, J. (2003). Rapid access to speech gestures in perception : Evidence from choice and simple response time tasks. Journal of Memory and Language, 49(3) :396–413. (Cité en page 18.)

Fowler, C. A., Shankweiler, D. P., et Studdert-Kennedy, M. (2016). Perception of the speech code revisited : Speech is alphabetic after all. Psychological Review, 123(2) :125–150. (Cité en page 23.)

Fox, R. A. (1982). Individual variation in the perception of vowels : Implications for a perception-production link. Phonetica, 39(1) :1–22. (Cité en page 19.)

Franken, M. K., McQueen, J. M., Hagoort, P., et Acheson, D. J. (2015). Assessing the link between speech perception and production through individual differences. In Proceedings of the 18th International Congress of Phonetic Sciences. (Cité en page 19.)

Friederici, A. D. et Wessels, J. M. (1993). Phonotactic knowledge of word boundaries and its use in infant speech perception. Attention, Perception, & Psychophysics, 54(3) :287–295. (Cité en page 34.)

Fromkin, V. A., editor (1984). Speech errors as linguistic evidence. Mouton de Gruyter, The Hague. (Cité en page 23.)

Galantucci, B., Fowler, C. A., et Turvey, M. T. (2006). The motor theory of speech perception reviewed. Psychonomic Bulletin & Review, 13(3) :361–377. (Cité en pages vi, 11, 14, 18 et 65.)

Ganapathiraju, A., Hamaker, J., Picone, J., Ordowski, M., et Doddington, G. R. (2001). Syllable-based large vocabulary continuous speech recognition. IEEE Transactions on Speech and Audio Processing, 9(4) :358–366. (Cité en page 53.)

Garnier, M., Lamalle, L., et Sato, M. (2013). Neural correlates of phonetic convergence and speech imitation. Frontiers in Psychology, 4(600) :1–15. (Cité en page 15.)

Gaskell, M. G. et Marslen-Wilson, W. D. (1997). Integrating form and meaning : A distributed model of speech perception. Language and Cognitive Processes, 12(5-6) :613–656. (Cité en pages 45 et 52.)

Gauthier, B., Shi, R., et Xu, Y. (2007). Learning phonetic categories by tracking movements. Cognition, 103(1) :80–106. (Cité en page 56.)

Gauvin, H. S., De Baene, W., Brass, M., et Hartsuiker, R. J. (2016). Conflict monitoring in speech processing : an fMRI study of error detection in speech production and perception. NeuroImage, 126 :96–105. (Cité en page 50.)

Gelfand, J. R. et Bookheimer, S. Y. (2003). Dissociating neural mechanisms of temporal sequencing and processing phonemes. Neuron, 38(5) :831–842. (Cité en page 26.)

Ghitza, O. (2011). Linking speech perception and neurophysiology : speech decoding guided by cascaded oscillators locked to the input rhythm. Frontiers in Psychology, 2(130) :1–13. (Cité en page 25.)

Ghitza, O. (2013). The theta-syllable : a unit of speech information defined by cortical function. Frontiers in Psychology, 4(138) :1–5. (Cité en page 25.)

Gilet, E., Diard, J., et Bessière, P. (2011). Bayesian action–perception computational model : interaction of production and recognition of cursive letters. PLoS One, 6(6) :e20387. (Cité en page 81.)

Giraud, A.-L. et Poeppel, D. (2012). Cortical oscillations and speech processing : emerging computational principles and operations. Nature Neuroscience, 15(4) :511–517. (Cité en page 25.)

Goldinger, S. D. (1996). Words and voices : episodic traces in spoken word identification and recognition memory. Journal of Experimental Psychology : Learning, Memory, and Cognition, 22(5) :1166–1183. (Cité en page 55.)

Goldinger, S. D. (1998). Echoes of echoes ? An episodic theory of lexical access. Psychological Review, 105(2) :251–279. (Cité en page 55.)

Goldinger, S. D. et Azuma, T. (2003). Puzzle-solving science : The quixotic quest for units in speech perception. Journal of Phonetics, 31(3) :305–320. (Cité en page 24.)

Goldsmith, J. A. et Xanthos, A. (2009). Learning phonological categories. Language, 85(1) :4–38. (Cité en pages 56 et 184.)

Goldstein, M. H., King, A. P., et West, M. J. (2003). Social interaction shapes babbling : Testing parallels between birdsong and speech. Proceedings of the National Academy of Sciences, 100(13) :8030–8035. (Cité en page 36.)

Goldstein, M. H. et Schwade, J. A. (2008). Social feedback to infants' babbling facilitates rapid phonological learning. Psychological Science, 19(5) :515–523. (Cité en page 36.)

Grabski, K. (2012). Les cartes sensorimotrices de la parole : Corrélats neurocognitifs et couplage fonctionnel des systèmes de perception et de production des voyelles du Français. PhD thesis, Université de Grenoble. (Cité en page 14.)

Grabski, K., Tremblay, P., Gracco, V. L., Girin, L., et Sato, M. (2013). A mediating role of the auditory dorsal pathway in selective adaptation to speech : A state-dependent transcranial magnetic stimulation study. Brain Research, 1515 :55–65. (Cité en page 13.)

Grafton, S. T., Arbib, M. A., Fadiga, L., et Rizzolatti, G. (1996). Localization of grasp representations in humans by positron emission tomography. Experimental Brain Research, 112(1) :103–111. (Cité en page 12.)

Guediche, S., Blumstein, S. E., Fiez, J. A., et Holt, L. L. (2014). Speech perception under adverse conditions : insights from behavioral, computational, and neuroscience research. Frontiers in Systems Neuroscience, 7 :126. (Cité en page 51.)

Guenther, F. H. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. Psychological Review, 102(3) :594–621. (Cité en pages 48 et 185.)

Guenther, F. H. (2006). Cortical interactions underlying the production of speech sounds. Journal of Communication Disorders, 39(5) :350–365. (Cité en page 48.)

Guenther, F. H. et Vladusich, T. (2012). A neural theory of speech acquisition and production. Journal of Neurolinguistics, 25(5) :408–422. (Cité en pages 48 et 58.)

Hallé, P. et Cristia, A. (2012). Global and detailed speech representations in early language acquisition. In Fuchs, S., Weirich, M., Pape, D., et Perrier, P., editors, Speech production and perception : Planning and dynamics, pages 11–38. Frankfurt am Main : Peter Lang. (Cité en pages 39 et 40.)

Hartsuiker, R. J. et Kolk, H. H. (2001). Error monitoring in speech production : A computational test of the perceptual loop theory. Cognitive Psychology, 42(2) :113–157. (Cité en page 50.)

Hayes, B. et White, J. (2013). Phonological naturalness and phonotactic learning. Linguistic Inquiry, 44(1) :45–75. (Cité en page 56.)

Hayes, B. et Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. Linguistic Inquiry, 39(3) :379–440. (Cité en page 56.)

Healy, A. F. et Cutting, J. E. (1976). Units of speech perception : Phoneme and syllable. Journal of Verbal Learning and Verbal Behavior, 15(1) :73–83. (Cité en page 24.)

Heintz, I., Beckman, M. E., Fosler-Lussier, E., et Ménard, L. (2009). Evaluating parameters for mapping adult vowels to imitative babbling. In Proceedings of Interspeech 2009, pages 688–691. (Cité en page 58.)

Hickok, G. (2009). The functional neuroanatomy of language. Physics of Life Reviews, 6(3) :121–143. (Cité en page 14.)

Hickok, G. (2010). The role of mirror neurons in speech perception and action word semantics. Language and Cognitive Processes, 25(6) :749–776. (Cité en page 14.)

Hickok, G. (2012). Computational neuroanatomy of speech production. Nature Reviews Neuroscience, 13(2) :135. (Cité en page 26.)

Hickok, G., Costanzo, M., Capasso, R., et Miceli, G. (2011). The role of Broca's area in speech perception : Evidence from aphasia revisited. Brain and Language, 119(3) :214–220. (Cité en page 13.)

Hickok, G. et Poeppel, D. (2007). The cortical organization of speech processing. Nature Reviews Neuroscience, 8(5) :393–402. (Cité en pages 25 et 193.)

Hillenbrand, J., Minifie, F. D., et Edwards, T. J. (1979). Tempo of spectrum change as a cue in speech-sound discrimination by infants. Journal of Speech, Language, and Hearing Research, 22(1) :147–165. (Cité en page 27.)

Hochmann, J.-R. et Papeo, L. (2014). The invariance problem in infancy : A pupillometry study. Psychological science, 25(11) :2038–2046. (Cité en page 41.)

Hollien, H. (1974). On vocal registers. Journal of Phonetics, 2 :125–143. (Cité en page 30.)

Hornstein, J. et Santos-Victor, J. (2007). A unified approach to speech production and recognition based on articulatory motor representations. In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2007), pages 3442–3447. (Cité en page 58.)

Houde, J. F. et Jordan, M. I. (1998). Sensorimotor adaptation in speech production. Science, 279(5354) :1213–1216. (Cité en pages vi, 16 et 17.)

Houde, J. F. et Jordan, M. I. (2002). Sensorimotor adaptation of speech I : Compensation and adaptation. Journal of Speech, Language, and Hearing Research, 45(2) :295–310. (Cité en pages vi et 16.)

Houde, J. F. et Nagarajan, S. S. (2011). Speech production as state feedback control. Frontiers in Human Neuroscience, 5(82) :1–14. (Cité en pages 48, 49 et 67.)

Houde, J. F., Nagarajan, S. S., et Heinks-Maldonado, T. (2007). Dynamic cortical imaging of speech compensation for auditory feedback perturbations. The Journal of the Acoustical Society of America, 121(5) :3045–3045. (Cité en pages vi, 49 et 50.)

Howard, I. S. et Messum, P. (2011). Modeling the development of pronunciation in infant speech acquisition. Motor Control, 15(1) :85–117. (Cité en pages 57, 58 et 59.)

Hyafil, A., Fontolan, L., Kabdebon, C., Gutkin, B., et Giraud, A.-L. (2015). Speech encoding by coupled cortical theta and gamma oscillations. ELife, 4 :e06213. (Cité en page 25.)

Iacoboni, M., Woods, R. P., Brass, M., Bekkering, H., Mazziotta, J. C., et Rizzolatti, G. (1999). Cortical mechanisms of human imitation. Science, 286(5449) :2526–2528. (Cité en page 12.)

Ishihara, H., Yoshikawa, Y., Miura, K., et Asada, M. (2008). Caregiver's sensorimotor magnets lead infant's vowel acquisition through auto mirroring. In IEEE International Conference on Developmental and Learning (ICDL 2008), pages 49–54. (Cité en pages 59 et 60.)

Ito, T., Tiede, M., et Ostry, D. J. (2009). Somatosensory function in speech perception. Proceedings of the National Academy of Sciences, 106(4) :1245–1248. (Cité en page 203.)

Iverson, J. M. et Fagan, M. K. (2004). Infant vocal–motor coordination : precursor to the gesture–speech system ? Child Development, 75(4) :1053–1066. (Cité en page 31.)

Iverson, J. M., Hall, A. J., Nickel, L., et Wozniak, R. H. (2007). The relationship between reduplicated babble onset and laterality biases in infant rhythmic arm movements. Brain and Language, 101(3) :198–207. (Cité en page 31.)

Iverson, J. M. et Thelen, E. (1999). Hand, mouth and brain. the dynamic emergence of speech and gesture. Journal of Consciousness Studies, 6(11-12) :19–40. (Cité en page 31.)

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., et Hinton, G. E. (1991). Adaptive mixtures of local experts. Neural Computation, 3(1) :79–87. (Cité en page 200.)

Jacquemot, C., Dupoux, E., et Bachoud-Lévi, A.-C. (2007). Breaking the mirror : Asymmetrical disconnection between the phonological input and output codes. Cognitive Neuropsychology, 24(1) :3–22. (Cité en page 181.)

Jamieson, D. G. et Cheesman, M. F. (1987). The adaptation of produced voice-onset time. Journal of Phonetics, 15(1) :15–27. (Cité en page 15.)

Jäncke, L., Wüstenberg, T., Scheich, H., et Heinze, H.-J. (2002). Phonetic perception and the temporal cortex. NeuroImage, 15(4) :733–746. (Cité en page 25.)

Jaynes, E. T. (2003). Probability theory : The logic of science. Cambridge University Press, Cambridge, UK. (Cité en page 62.)

Johnson, E. K. et Tyler, M. D. (2010). Testing the limits of statistical learning for word segmentation. Developmental Science, 13(2) :339–345. (Cité en page 33.)

Johnson, K. (1995). Talker variability in vowel perception. The Journal of the Acoustical Society of America, 98(5) :2949–2950. (Cité en page 178.)

Johnson, K. (1997). Speech perception without speaker normalization : An exemplar model. In Johnson, K. et Mullennix, J. W., editors, Talker variability in speech processing, pages 145–165. Morgan Kaufmann Publishers Inc., San Francisco, CA. (Cité en pages 55 et 185.)

Jones, J. A. et Callan, D. E. (2003). Brain activity during audiovisual speech perception : an fMRI study of the McGurk effect. NeuroReport, 14(8) :1129–1133. (Cité en page 13.)

Jusczyk, P. W. (1993). From general to language-specific capacities : The WRAPSA model of how speech perception develops. Journal of Phonetics, 21 :3–28. (Cité en page 54.)

Jusczyk, P. W. (1997). The Discovery of Spoken Language. Cambridge, MA : MIT Press. (Cité en pages 27 et 39.)

Jusczyk, P. W. et Derrah, C. (1987). Representation of speech sounds by young infants. Developmental Psychology, 23(5) :648. (Cité en pages 40 et 171.)

Jusczyk, P. W., Friederici, A. D., Wessels, J. M., Svenkerud, V. Y., et Jusczyk, A. M. (1993). Infants' sensitivity to the sound patterns of native language words. Journal of Memory and Language, 32(3) :402–420. (Cité en page 34.)

Jusczyk, P. W. et Luce, P. A. (1994). Infants' sensitivity to phonotactic patterns in the native language. Journal of Memory and Language, 33(5) :630–645. (Cité en page 34.)

Jusczyk, P. W., Murray, J., et Bayly, J. (1979). Perception of place of articulation in fricatives and stops by infants. In Proceedings of the Biennial Meeting of the Society for Research in Child Development, San Francisco. (Cité en page 27.)

Jusczyk, P. W., Rosner, B. S., Cutting, J. E., Foard, C. F., et Smith, L. B. (1977). Categorical perception of nonspeech sounds by 2-month-old infants. Attention, Perception, & Psychophysics, 21(1) :50–54. (Cité en page 27.)

Kanda, H., Ogata, T., Komatani, K., et Okuno, H. G. (2008). Segmenting acoustic signal with articulatory movement using recurrent neural network for phoneme acquisition. In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2008), pages 1712–1717. (Cité en page 58.)

Kanda, H., Ogata, T., Takahashi, T., Komatani, K., et Okuno, H. G. (2009). Continuous vocal imitation with self-organized vowel spaces in recurrent neural network. In IEEE International Conference on Robotics and Automation (ICRA 2009), pages 4438–4443. (Cité en page 58.)

Kawato, M. (1999). Internal models for motor control and trajectory planning. Current Opinion in Neurobiology, 9(6) :718–727. (Cité en page 67.)

Kent, R. D. (1984). Psychobiology of speech development : Coemergence of language and a movement system. American Journal of Physiology : Regulatory, Integrative and Comparative Physiology, 246(6) :R888–R894. (Cité en page 31.)

Kessen, W., Levine, J., et Wendrich, K. A. (1979). The imitation of pitch in infants. Infant Behavior and Development, 2 :93–99. (Cité en page 36.)

Kidd, E. (2012). Implicit statistical learning is directly associated with the acquisition of syntax. Developmental Psychology, 48(1) :171–184. (Cité en page 33.)

Kiebel, S. J., Von Kriegstein, K., Daunizeau, J., et Friston, K. J. (2009). Recognizing sequences of sequences. PLoS Computational Biology, 5(8) :e1000464. (Cité en pages 52, 53 et 186.)

King, S., Frankel, J., Livescu, K., McDermott, E., Richmond, K., et Wester, M. (2007). Speech production knowledge in automatic speech recognition. The Journal of the Acoustical Society of America, 121(2) :723–742. (Cité en page 44.)

Kirchhoff, K. (1998). Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments. In Proceedings of the 5th International Conference on Spoken Language Processing, pages 891–894. (Cité en page 44.)

Klatt, D. H. (1980). Speech perception : A model of acoustic-phonetic analysis and lexical access. In Cole, R. A., editor, Perception and production of fluent speech, pages 243–288. Hillsdale, NJ : Erlbaum. (Cité en pages 45 et 46.)

Kleinschmidt, D. F. et Jaeger, T. F. (2011). A bayesian belief updating model of phonetic recalibration and selective adaptation. In Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics, pages 10–19. Association for Computational Linguistics. (Cité en pages 45, 46, 51, 67 et 177.)

Kleinschmidt, D. F. et Jaeger, T. F. (2015). Robust speech perception : Recognize the familiar, generalize to the similar, and adapt to the novel. Psychological Review, 122(2) :148–203. (Cité en pages 45, 51, 67, 71 et 185.)

Kluender, K. R., Diehl, R. L., et Killeen, P. R. (1987). Japanese quail can learn phonetic categories. Science, 237(4819) :1195–1197. (Cité en page 12.)

Kröger, B. J., Birkholz, P., Kannampuzha, J., et Neuschaefer-Rube, C. (2011). Categorical perception of consonants and vowels : evidence from a neurophonetic model of speech production and perception. In Esposito, A., Esposito, A. M., Martone, R., Müller, V. C., et Scarpetta, G., editors, Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues, pages 354–361. Springer, Berlin. (Cité en pages vi, 46, 47, 52 et 189.)

Kröger, B. J., Birkholz, P., Lowit, A., et Neuschaefer-Rube, C. (2010). Phonemic, sensory, and motor representations in an action-based neurocomputational model of speech production. In Maassen, B. et van Lieshout, P., editors, Speech motor control : New developments in basic and applied research, pages 23–36. Oxford University Press, Oxford. (Cité en pages 52 et 60.)

Kröger, B. J. et Cao, M. (2015). The emergence of phonetic-phonological features in a biologically inspired model of speech processing. Journal of Phonetics, 53 :88–100. (Cité en page 46.)

Kröger, B. J., Kannampuzha, J., et Kaufmann, E. (2014). Associative learning and self-organization as basic principles for simulating speech acquisition, speech production, and speech perception. EPJ Nonlinear Biomedical Physics, 2(1) :1–28. (Cité en pages 46 et 56.)

Kröger, B. J., Kannampuzha, J., et Neuschaefer-Rube, C. (2009). Towards a neurocomputational model of speech production and perception. Speech Communication, 51(9) :793–809. (Cité en pages 46, 48 et 189.)

Kuhl, P. K. (2000). Language, mind, and brain : Experience alters perception. In Gazzaniga, M., editor, The New Cognitive Neurosciences, volume 2, pages 99–115. Cambridge, MA : MIT Press. (Cité en page 38.)

Kuhl, P. K. (2004). Early language acquisition : cracking the speech code. Nature Reviews Neuroscience, 5(11) :831–843. (Cité en pages 38, 181 et 182.)

Kuhl, P. K. (2007). Is speech learning "gated" by the social brain ? Developmental Science, 10(1) :110–120. (Cité en pages 36 et 184.)

Kuhl, P. K. (2010). Brain mechanisms in early language acquisition. Neuron, 67(5) :713–727. (Cité en page 29.)

Kuhl, P. K., Kiritani, S., Deguchi, T., Hayashi, A., Stevens, E. B., Dugger, C. D., et Iverson, P. (1997). Effects of language experience on speech perception : American and Japanese infants' perception of /ra/ and /la/. The Journal of the Acoustical Society of America, 102(5) :3135–3136. (Cité en page 29.)

Kuhl, P. K. et Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. Science, 218(4577) :1138–1141. (Cité en page 36.)

Kuhl, P. K. et Meltzoff, A. N. (1996). Infant vocalizations in response to speech : Vocal imitation and developmental change. The Journal of the Acoustical Society of America, 100(4) :2425–2438. (Cité en page 36.)

Kuhl, P. K. et Miller, J. D. (1975). Speech perception by the chinchilla : Voiced-voiceless distinction in alveolar plosive consonants. Science, 190(4209) :69–72. (Cité en page 38.)

Kuhl, P. K. et Padden, D. M. (1983). Enhanced discriminability at the phonetic boundaries for the place feature in macaques. The Journal of the Acoustical Society of America, 73(3) :1003–1010. (Cité en page 38.)

Kuhl, P. K., Ramirez, R. R., Bosseler, A., Lin, J.-F. L., et Imada, T. (2014). Infants' brain responses to speech suggest analysis by synthesis. 111(31) :11238–11245. (Cité en pages 32 et 38.)

Kuhl, P. K., Stevens, E. B., Hayashi, A., Deguchi, T., Kiritani, S., et Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. Developmental Science, 9(2) :F13–F21. (Cité en pages 28 et 29.)

Kuhl, P. K., Tsao, F.-M., et Liu, H.-M. (2003). Foreign-language experience in infancy : Effects of short-term exposure and social interaction on phonetic learning. Proceedings of the National Academy of Sciences, 100(15) :9096–9101. (Cité en pages 29 et 35.)

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., et Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. Science, 255(5044) :606–608. (Cité en page 41.)

Lackner, J. R. et Goldstein, L. M. (1975). The psychological representation of speech sounds. The Quarterly Journal of Experimental Psychology, 27(2) :173–185. (Cité en page 23.)

Lametti, D. R., Rochet-Capellan, A., Neufeld, E., Shiller, D. M., et Ostry, D. J. (2014). Plasticity in the human speech motor system drives changes in speech perception. Journal of Neuroscience, 34(31) :10339–10346. (Cité en pages 17 et 203.)

Lane, H. et Tranel, B. (1971). The lombard sign and the role of hearing in speech. Journal of Speech, Language, and Hearing Research, 14(4) :677–709. (Cité en page 15.)

Lasky, R. E., Syrdal-Lasky, A., et Klein, R. E. (1975). VOT discrimination by four to six and a half month old infants from Spanish environments. Journal of Experimental Child Psychology, 20(2) :215–225. (Cité en page 28.)

Laurent, R. (2014). COSMO : un modèle bayésien des interactions sensori-motrices dans la perception de la parole. PhD thesis, Université de Grenoble. (Cité en pages 14, 62, 65, 78, 96, 140, 181 et 183.)

Laurent, R., Barnaud, M.-L., Schwartz, J.-L., Bessière, P., et Diard, J. (2017). The complementary roles of auditory and motor information evaluated in a Bayesian perceptuo-motor model of speech perception. Psychological Review, 124(5) :572–602. (Cité en pages 62, 82, 92, 93, 138, 140, 148, 171 et 202.)

Laurent, R., Schwartz, J.-L., Bessière, P., et Diard, J. (2013). A computational model of perceptuo-motor processing in speech perception : learning to imitate and categorize synthetic CV syllables. In Bimbot, F., editor, Proceedings of Interspeech 2013, pages 2796–2800. International Speech Communication Association (ISCA). (Cité en page 62.)

Lebeltel, O., Bessière, P., Diard, J., et Mazer, E. (2004). Bayesian robot programming. Autonomous Robots, 16(1) :49–79. (Cité en page 62.)

Lelong, A. (2012). Convergence phonétique en interaction. PhD thesis, Université Grenoble-Alpes. (Cité en page 15.)

Levelt, W. J. (1999). Models of word production. Trends in Cognitive Sciences, 3(6) :223–232. (Cité en page 51.)

Levelt, W. J., Roelofs, A., et Meyer, A. S. (1999). A theory of lexical access in speech production. Behavioral and Brain Sciences, 22(1) :1–38. (Cité en page 23.)

Levelt, W. J. et Wheeldon, L. (1994). Do speakers have access to a mental syllabary ? Cognition, 50(1) :239–269. (Cité en page 23.)

Levitt, A., Jusczyk, P. W., Murray, J., et Carden, G. (1988). The perception of place of articulation contrasts in voiced and voiceless fricatives by two-month-old infants. Journal of Experimental Psychology : Human Perception and Performance, 14 :361–368. (Cité en page 27.)

Liberman, A. M. (1957). Some results of research on speech perception. The Journal of the Acoustical Society of America, 29(1) :117–123. (Cité en page 21.)

Liberman, A. M. (1996). Speech : A special code. Cambridge, MA : MIT press. (Cité en page 10.)

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., et Studdert-Kennedy, M. (1967). Perception of the speech code. Psychological Review, 74(6) :431–461. (Cité en pages vi et 11.)

Liberman, A. M., Delattre, P. C., Cooper, F. S., et Gerstman, L. J. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. Psychological Monographs : General and Applied, 68(8) :1–13. (Cité en page 21.)

Liberman, A. M. et Mattingly, I. G. (1985). The motor theory of speech perception revised. Cognition, 21(1) :1–36. (Cité en pages 11 et 65.)

Lopes, M., Melo, F. S., Kenward, B., et Santos-Victor, J. (2009). A computational model of social-learning mechanisms. Adaptive Behavior, 17(6) :467–483. (Cité en page 58.)

Lotto, A. J., Hickok, G., et Holt, L. L. (2009). Reflections on mirror neurons and speech perception. Trends in Cognitive Sciences, 13(3) :110–114. (Cité en page 13.)

Luce, P. A., Goldinger, S. D., Auer, E. T., et Vitevitch, M. S. (2000). Phonetic priming, neighborhood activation, and PARSYN. Attention, Perception, & Psychophysics, 62(3) :615–625. (Cité en page 45.)

Luce, R. D. (1986). Response times : Their role in inferring elementary mental organization. Oxford University Press, New-York, Oxford. (Cité en page 18.)

MacNeilage, P. F. (1998). The frame/content theory of evolution of speech production. Behavioral and Brain Sciences, 21(4) :499–511. (Cité en page 31.)

MacNeilage, P. F. et Davis, B. L. (1990). Acquisition of speech production : Frames, then content. In Jeannerod, M., editor, Attention and Performance 13 : Motor representations and control, pages 453–476. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ. (Cité en page 140.)

MacNeilage, P. F., Davis, B. L., et Matyear, C. L. (1997). Babbling and first words : Phonetic similarities and differences. Speech Communication, 22(2-3) :269–277. (Cité en pages 31 et 40.)

Maeda, S. (1990). Compensatory articulation during speech : Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In Hardcastle, W. J. et Marchal, A., editors, Speech Production and Speech Modelling, pages 131–149. Kluwer Academic Publishers, Springer, Berlin. (Cité en page 96.)

Magri, G. (2014). Error-driven versus batch models of the acquisition of phonotactics : David defeats Goliath. In Proceedings of the Annual Meetings on Phonology, pages 1–12. (Cité en page 56.)

Mampe, B., Friederici, A. D., Christophe, A., et Wermke, K. (2009). Newborns' cry melody is shaped by their native language. Current Biology, 19(23) :1994–1997. (Cité en pages 35 et 36.)

Mann, V. et Wimmer, H. (2002). Phoneme awareness and pathways into literacy : A comparison of German and American children. Reading and Writing, 15(7) :653–682. (Cité en page 40.)

Markey, K. L. (1994). The sensorimotor foundations of phonology : a computational model of early childhood articulatory and phonetic development. PhD thesis, University of Colorado. (Cité en page 58.)

Markiewicz, C. J. et Bohland, J. W. (2016). Mapping the cortical representation of speech sounds in a syllable repetition task. NeuroImage, 141 :174–190. (Cité en page 25.)

Marr, D. (1982). Vision : A computational investigation into the human representation and processing of visual information. Cambridge, MA : MIT Press. (Cité en page 52.)

Marslen-Wilson, W. D. et Warren, P. (1994). Levels of perceptual representation and process in lexical access : Words, phonemes, and features. Psychological Review, 101(4) :653–674. (Cité en page 24.)

Martin, A., Peperkamp, S., et Dupoux, E. (2013). Learning phonemes with a proto-lexicon. Cognitive Science, 37(1) :103–124. (Cité en pages 56, 59 et 60.)

Martinet, A. (1970). Éléments de linguistique générale. A. Colin., Paris. (Cité en page 64.)

Masataka, N. (1993). Effects of contingent and noncontingent maternal stimulation on the vocal behaviour of three-to four-month-old Japanese infants. Journal of Child Language, 20(2) :303–312. (Cité en page 36.)

Massaro, D. W. (1972). Preperceptual images, processing time, and perceptual units in auditory perception. Psychological Review, 79(2) :124. (Cité en page 21.)

Mattys, S. L. et Melhorn, J. F. (2005). How do syllables contribute to the perception of spoken English ? Insight from the migration paradigm. Language and Speech, 48(2) :223–252. (Cité en page 22.)

Maurer, D. et Werker, J. F. (2014). Perceptual narrowing during infancy : A comparison of language and faces. Developmental Psychobiology, 56(2) :154–178. (Cité en page 29.)

Max, L., Wallace, M. E., et Vincent, I. (2003). Sensorimotor adaptation to auditory perturbations during speech : Acoustic and kinematic experiments. In Proceedings of the 15th International Congress of Phonetic Sciences, pages 1053–1056. (Cité en page 16.)

Maye, J., Weiss, D. J., et Aslin, R. N. (2008). Statistical phonetic learning in infants : Facilitation and feature generalization. Developmental Science, 11(1) :122–134. (Cité en page 33.)

Maye, J., Werker, J. F., et Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. Cognition, 82(3) :B101–B111. (Cité en page 33.)

McClelland, J. L. et Elman, J. L. (1986). The TRACE model of speech perception. Cognitive Psychology, 18(1) :1–86. (Cité en pages 44, 45, 52 et 197.)

McGettigan, C. et Tremblay, P. (2017). Links between perception and production : examining the roles of motor and premotor cortices in understanding speech. In Gaskell, M. et Rueschemeyer, S.-A., editors, Oxford Handbook of Psycholinguistics. Oxford University Press, Oxford. (Cité en page 14.)

McMurray, B., Aslin, R. N., et Toscano, J. C. (2009). Statistical learning of phonetic categories : insights from a computational approach. Developmental Science, 12(3) :369–378. (Cité en pages 55, 59, 60, 70, 71, 177 et 200.)

McNeill, D. et Lindig, K. (1973). The perceptual reality of phonemes, syllables, words, and sentences. Journal of Verbal Learning and Verbal Behavior, 12(4) :419–430. (Cité en page 21.)

McQueen, J. M. et Cutler, A. (2001). Spoken word access processes : An introduction. Language and Cognitive Processes, 16(5-6) :469–490. (Cité en page 24.)

McQueen, J. M., Cutler, A., et Norris, D. (2000). Why Merge really is autonomous and parsimonious. In 2000 ISCA Tutorial and Research Workshop (ITRW) on Spoken Word Access Processes, pages 47–50. (Cité en pages vi et 45.)

Mehler, J., Dommergues, J. Y., Frauenfelder, U. H., et Segui, J. (1981). The syllable's role in speech segmentation. Journal of Verbal Learning and Verbal Behavior, 20(3) :298–305. (Cité en pages vi, 21 et 22.)

Mehler, J. et Hayes, R. (1981). The role of syllables in speech processing : Infant and adult data. Philosophical Transactions of the Royal Society of London B : Biological Sciences, 295(1077) :333–352. (Cité en pages 21 et 39.)

Meier, R. P., McGarvin, L., Zakia, R. A., et Willerman, R. (1997). Silent mandibular oscillations in vocal babbling. Phonetica, 54(3-4) :153–171. (Cité en page 31.)

Meister, I. G., Wilson, S. M., Deblieck, C., Wu, A. D., et Iacoboni, M. (2007). The essential role of premotor cortex in speech perception. Current Biology, 17(19) :1692–1696. (Cité en pages 13 et 14.)

Meltzoff, A. N. et Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. Science, 198(4312) :75–78. (Cité en page 36.)

Ménard, L. et Schwartz, J.-L. (2014). Perceptuo-motor biases in the perceptual organization of the height feature in french vowels. Acta Acustica united with Acustica, 100(4) :676–689. (Cité en pages iv, viii, 19, 113, 125, 126, 127, 128, 174, 188 et 189.)

Ménard, L., Schwartz, J.-L., et Aubin, J. (2008). Invariance and variability in the production of the height feature in French vowels. Speech Communication, 50(1) :14–28. (Cité en pages 19 et 127.)

Ménard, L., Schwartz, J.-L., et Boë, L.-J. (2004). Role of vocal tract morphology in speech development : Perceptual targets and sensorimotor maps for synthesized French vowels from birth to adulthood. Journal of Speech, Language, and Hearing Research, 47(5) :1059–1080. (Cité en pages 103 et 178.)

Ménard, L., Schwartz, J.-L., Boë, L.-J., Kandel, S., et Vallée, N. (2002). Auditory normalization of French vowels synthesized by an articulatory model simulating growth from birth to adulthood. The Journal of the Acoustical Society of America, 111(4) :1892–1905. (Cité en page 178.)

Meringer, R. et Mayer, K. (1895). Versprechen und Verlesen : eine Psychologish-linguistische Studie. Stuttgart : Göschensche Verlagsbuschhandlung. (Cité en page 23.)

Messum, P. (2008). The role of imitation in learning to pronounce. PhD thesis, University of London. (Cité en page 57.)

Messum, P. et Howard, I. S. (2015). Creating the cognitive form of phonological units : The speech sound correspondence problem in infancy could be solved by mirrored vocal interactions rather than by imitation. Journal of Phonetics, 53 :125–140. (Cité en pages vi, 59, 60 et 73.)

Meunier, C. (2007). Phonétique acoustique. In Auzou, P., Rolland-Monnoury, V., Pinto, S., et Ozsancak, C., editors, Les dysarthries, pages 164–173. Solal, Marseille. (Cité en pages 98, 151 et 199.)

Miceli, G., Gainotti, G., Caltagirone, C., et Masullo, C. (1980). Some aspects of phonological impairment in aphasia. Brain and Language, 11(1) :159–169. (Cité en page 13.)

Mills, C. B. (1980). Effects of context on reaction time to phonemes. Journal of Verbal Learning and Verbal Behavior, 19(1) :75–83. (Cité en page 21.)

Mirman, D., McClelland, J. L., et Holt, L. L. (2006). An interactive hebbian account of lexically guided tuning of speech perception. Psychonomic Bulletin & Review, 13(6) :958–965. (Cité en pages 52 et 54.)

Mitchell, P. R. et Kent, R. D. (1990). Phonetic variation in multisyllable babbling. Journal of Child Language, 17(2) :247–265. (Cité en page 31.)

Mitterer, H., Scharenborg, O., et McQueen, J. M. (2013). Phonological abstraction without phonemes in speech perception. Cognition, 129(2) :356–361. (Cité en page 24.)

Miura, K., Yoshikawa, Y., et Asada, M. (2007). Unconscious anchoring in maternal imitation that helps find the correspondence of a caregiver's vowel categories. Advanced Robotics, 21(13) :1583–1600. (Cité en page 59.)

Miura, K., Yoshikawa, Y., et Asada, M. (2008). Realizing being imitated : Vowel mapping with clearer articulation. In IEEE International Conference on Developmental and Learning (ICDL 2008), pages 262–267. (Cité en page 59.)

Miura, K., Yoshikawa, Y., et Asada, M. (2012). Vowel acquisition based on an auto-mirroring bias with a less imitative caregiver. Advanced Robotics, 26(1-2) :23–44. (Cité en pages 59 et 60.)

Morais, J. (1985). Literacy and awareness of the units of speech : Implications for research on the units of perception. Linguistics, 23(5) :707–722. (Cité en page 22.)

Morais, J., Cary, L., Alegria, J., et Bertelson, P. (1979). Does awareness of speech as a sequence of phones arise spontaneously ? Cognition, 7(4) :323–331. (Cité en pages 22 et 170.)

Morais, J., Content, A., Cary, L., Mehler, J., et Segui, J. (1989). Syllabic segmentation and literacy. Language and Cognitive Processes, 4(1) :57–67. (Cité en page 22.)

Morillon, B., Liégeois-Chauvel, C., Arnal, L. H., Bénar, C.-G., et Giraud, A.-L. (2012). Asymmetric function of theta and gamma activity in syllable processing : an intra-cortical study. Frontiers in Psychology, 3 :248. (Cité en page 25.)

Möttönen, R., Dutton, R., et Watkins, K. E. (2012). Auditory-motor processing of speech sounds. Cerebral Cortex, 23(5) :1190–1197. (Cité en page 13.)

Möttönen, R. et Watkins, K. E. (2009). Motor representations of articulators contribute to categorical perception of speech sounds. Journal of Neuroscience, 29(31) :9819–9825. (Cité en pages 13, 194 et 195.)

Moulin-Frier, C. (2011). Rôle des relations perception-action dans la communication parlée et l'émergence des systèmes phonologiques : étude, modélisation computationnelle et simulations. PhD thesis, Université de Grenoble. (Cité en page 183.)

Moulin-Frier, C., Diard, J., Schwartz, J.-L., et Bessière, P. (2015). COSMO ("Communicating about Objects using Sensory-Motor Operations") : A Bayesian modeling framework for studying speech communication and the emergence of phonological systems. Journal of Phonetics, 53 :5–41. (Cité en pages 62 et 71.)

Moulin-Frier, C., Laurent, R., Bessière, P., Schwartz, J.-L., et Diard, J. (2012). Adverse conditions improve distinguishability of auditory, motor, and perceptuo-motor theories of speech perception : An exploratory Bayesian modelling study. Language and Cognitive Processes, 27(7-8) :1240–1263. (Cité en pages 62, 183 et 202.)

Moulin-Frier, C., Nguyen, S. M., et Oudeyer, P.-Y. (2013). Self-organization of early vocal development in infants and machines : the role of intrinsic motivation. Frontiers in Psychology, 4(1006) :1–20. (Cité en page 58.)

Moulin-Frier, C. et Oudeyer, P.-Y. (2012). Curiosity-driven phonetic learning. In The 2nd Joint IEEE International Conference on Developmental and Learning and on Epigenetic Robotics (ICDL-Epirob 2012), pages 1–8. (Cité en page 57.)

Moulin-Frier, C. et Oudeyer, P.-Y. (2013). Exploration strategies in developmental robotics : a unified probabilistic framework. In The 3rd Joint IEEE International Conference on Developmental and Learning and on Epigenetic Robotics (ICDL-Epirob 2013), pages 1–6. (Cité en pages 57, 60, 72, 94, 104 et 109.)

Mowrer, D. E. (1980). Phonological development during the first year of life. Speech and Language : Advances in Basic Research and Practice, 4 :99–142. (Cité en page 29.)

Munson, B., Edwards, J., et Beckman, M. E. (2011). Phonological representations in language acquisition : Climbing the ladder of abstraction. In Cohn, A. C. et Fougeron, C. & Huffman, M. K., editors, The Oxford Handbook of Laboratory Phonology, pages 288–309. Oxford : Oxford University Press. (Cité en page 36.)

Murakami, M., Kröger, B. J., Birkholz, P., et Triesch, J. (2015). Seeing [u] aids vocal learning : Babbling and imitation of vowels using a 3D vocal tract model, reinforcement learning, and reservoir computing. In The 5th Joint IEEE International Conference Developmental Learning and Epigenetic Robotics (ICDL-Epirob 2015), pages 208–213. (Cité en pages 57 et 60.)

Näätänen, R., Kujala, T., et Winkler, I. (2011). Auditory processing that leads to conscious perception : a unique window to central auditory processing opened by the mismatch negativity and related responses. Psychophysiology, 48(1) :4–22. (Cité en page 24.)

Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., Vainio, M., Alku, P., Ilmoniemi, R. J., Luuk, A., Allik, J., Sinkkonen, J., et Alho, K. (1997). Language-specific phoneme representations revealed by electric and magnetic brain responses. Nature, 385(6615) :432–434. (Cité en page 24.)

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. (Cité en page 24.)

Najnin, S. et Banerjee, B. (2016). Emergence of vocal developmental sequences in a predictive coding model of speech acquisition. In Proceedings of Interspeech 2016, pages 1113–1117. (Cité en page 60.)

Nelson, D. G. K., Jusczyk, P. W., Mandel, D. R., Myers, J., Turk, A., et Gerken, L. (1995). The head-turn preference procedure for testing auditory perception. Infant Behavior and Development, 18(1) :111–116. (Cité en page 28.)

Newman, R. S. (2003). Using links between speech perception and speech production to evaluate different acoustic metrics : A preliminary report. The Journal of the Acoustical Society of America, 113(5) :2850–2860. (Cité en page 19.)

Nittrouer, S. (2001). Challenging the notion of innate phonetic boundaries. The Journal of the Acoustical Society of America, 110(3) :1598–1605. (Cité en page 27.)

Norris, D. (1994). Shortlist : A connectionist model of continuous speech recognition. Cognition, 52(3) :189–234. (Cité en pages 45 et 51.)

Norris, D. et Cutler, A. (1988). The relative accessibility of phonemes and syllables. Perception & Psychophysics, 43(6) :541–550. (Cité en page 23.)

Norris, D. et McQueen, J. M. (2008). Shortlist B : a Bayesian model of continuous speech recognition. Psychological Review, 115(2) :357. (Cité en pages 45, 51 et 67.)

Norris, D., McQueen, J. M., et Cutler, A. (2000). Merging information in speech recognition : Feedback is never necessary. Behavioral and Brain Sciences, 23(3) :299–325. (Cité en pages vi, 45 et 51.)

Nozari, N., Dell, G. S., et Schwartz, M. F. (2011). Is comprehension necessary for error detection ? A conflict-based account of monitoring in speech production. Cognitive Psychology, 63(1) :1–33. (Cité en page 50.)

Ohala, J. J., Derwing, B. L., Nearey, T. M., et Dow, M. L. (1986). On the phoneme as the unit of the "second articulation". Phonology, 3(1) :45–69. (Cité en page 23.)

Ojanen, V., Möttönen, R., Pekkola, J., Jääskeläinen, I. P., Joensuu, R., Autti, T., et Sams, M. (2005). Processing of audiovisual speech in Broca's area. NeuroImage, 25(2) :333–338. (Cité en page 13.)

Oller, D. K. (1980). The emergence of the sounds of speech in infancy. In Yeni-Komshian, G., Kavanagh, J. F., et Ferguson, C. A., editors, Child Phonology, Vol. 1,, pages 93–112. New York, NY : Academic Press. (Cité en pages vi, 30 et 31.)

Oller, D. K. (2000). The emergence of the capacity for speech. Lawrence Erlbaum Associates, Hillsdale, NJ. (Cité en pages 29 et 31.)

Oller, D. K. et Eilers, R. E. (1988). The role of audition in infant babbling. Child Development, 59(2) :441–449. (Cité en page 30.)

Otake, T., Hatano, G., Cutler, A., et Mehler, J. (1993). Mora or syllable ? Speech segmentation in Japanese. Journal of Memory and Language, 32(2) :258–278. (Cité en page 22.)

Oudeyer, P.-Y. (2002). Phonemic coding might result from sensory-motor coupling dynamics. In Proceedings of the 7th International Conference on Simulation of Adaptive Behavior : From Animals to Animats, pages 407–416. MIT Press. (Cité en page 58.)

Oudeyer, P.-Y., Baranes, A., et Kaplan, F. (2010). Intrinsically motivated exploration for developmental and active sensorimotor learning. In Sigaud, O. et Peters, J., editors, From motor learning to interaction learning in robots, pages 107–146. Springer, Berlin. (Cité en page 185.)

Pallier, C. (1997). Phonemes and syllables in speech perception : size of the attentional focus in French. In Proceedings of Eurospeech-97. University of Patras, Rion, Greece. (Cité en page 23.)

Papoušek, M. et Papoušek, H. (1981). Musical elements in the infant's vocalization : Their significance for communication, cognition, and creativity. In Lipsitt, L. P. et Rovee-Collier, C. K., editors, Advances in Infancy Research. Ablex Publishing Corporation, Norwood, NJ. (Cité en page 36.)

Pardo, J. S. (2013). Measuring phonetic convergence in speech production. Frontiers in Psychology, 4(559) :1–5. (Cité en page 15.)

Patri, J.-F., Diard, J., et Perrier, P. (2016). Modélisation bayésienne de la planification motrice des gestes de parole : Evaluation du rôle des différentes modalités sensorielles. In 31ème Journées d'Études sur la Parole, pages 419–427. (Cité en pages 175 et 203.)

Patri, J.-F., Diard, J., Schwartz, J.-L., et Pascal, P. (soumis). What drives the perceptual change resulting from speech motor adaptation ? Evaluation of hypotheses in a Bayesian modeling framework. PLoS Computational Biology. (Cité en page 203.)

Peeva, M. G., Guenther, F. H., Tourville, J. A., Nieto-Castanon, A., Anton, J.-L., Nazarian, B., et Alario, F.-X. (2010). Distinct representations of phonemes, syllables, and supra-syllabic sequences in the speech production network. NeuroImage, 50(2) :626–638. (Cité en pages 25 et 26.)

Pelucchi, B., Hay, J. F., et Saffran, J. R. (2009). Statistical learning in a natural language by 8-month-old infants. Child Development, 80(3) :674–685. (Cité en page 33.)

Peperkamp, S. (2003). Phonological acquisition : Recent attainments and new challenges. Language and Speech, 46(2-3) :87–113. (Cité en page 38.)

Peperkamp, S., Le Calvez, R., Nadal, J.-P., et Dupoux, E. (2006). The acquisition of allophonic rules : Statistical learning with linguistic constraints. Cognition, 101(3) :B31–B41. (Cité en pages 34 et 56.)

Perkell, J. S., Guenther, F. H., Lane, H., Matthies, M. L., Stockmann, E., Tiede, M., et Zandipour, M. (2004a). The distinctness of speakers' productions of vowel contrasts is related to their discrimination of the contrasts. The Journal of the Acoustical Society of America, 116(4) :2338–2344. (Cité en page 19.)

Perkell, J. S. et Klatt, D. H., editors (1986). Invariance and variability in speech processes. Lawrence Erlbaum Associates, Hillsdale, NJ. (Cité en page 14.)

Perkell, J. S., Lane, H., Ghosh, S., Matthies, M. L., Tiede, M., Guenther, F. H., et Ménard, L. (2008). Mechanisms of vowel production : auditory goals and speaker acuity. In Proceedings of the 8th International Seminar on Speech Production, pages 29–32. (Cité en page 19.)

Perkell, J. S., Matthies, M. L., Tiede, M., Lane, H., Zandipour, M., Marrone, N., Stockmann, E., et Guenther, F. H. (2004b). The distinctness of speakers' /s/—/ʃ/ contrast is related to their auditory discrimination and use of an articulatory saturation effect. Journal of Speech, Language, and Hearing Research, 47(6) :1259–1269. (Cité en page 19.)

Philippsen, A. K., Reinhart, F., et Wrede, B. (2015). Efficient bootstrapping of vocalization skills using active goal babbling. In International Workshop on Speech Robotics at Interspeech. (Cité en page 57.)

Philippsen, A. K., Reinhart, R. F., et Wrede, B. (2014). Learning how to speak : Imitation-based refinement of syllable production in an articulatory-acoustic model. In The 4th Joint IEEE International Conference on Developmental and Learning and on Epigenetic Robotics (ICDL-Epirob 2017), pages 195–200. (Cité en pages 58, 60, 72 et 109.)

Philippsen, A. K., Reinhart, R. F., et Wrede, B. (2016). Goal babbling of acoustic-articulatory models with adaptive exploration noise. In The 6th Joint IEEE International Conference on Developmental and Learning and on Epigenetic Robotics (ICDL-Epirob 2017), pages 72–78. (Cité en pages 57 et 72.)

Pickering, M. J. et Garrod, S. (2013). An integrated theory of language production and comprehension. Behavioral and Brain Sciences, 36(4) :329–347. (Cité en page 20.)

Pierrehumbert, J. B. (2001). Exemplar dynamics : word frequency, lenition, and contrast. In Bybee, J. et Hopper, P., editors, Frequency effects and the emergence of linguistic structure, pages 137–157. Amsterdam : John Benjamins Publishing Company. (Cité en page 55.)

Pierrehumbert, J. B. (2003). Phonetic diversity, statistical learning, and acquisition of phonology. Language and Speech, 46(2-3) :115–154. (Cité en pages 38, 51 et 55.)

Pierrehumbert, J. B. (2016). Phonological representation : beyond abstract versus episodic. Annual Review of Linguistics, 2 :33–52. (Cité en page 55.)

Pitt, M. A. et Samuel, A. G. (1990). Attentional allocation during speech perception : How fine is the focus ? Journal of Memory and Language, 29(5) :611–632. (Cité en page 23.)

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. (Cité en page 25.)

Poeppel, D., Overath, T., Popper, A. N., et Fay, R. R. (2012). The Human Auditory Cortex. Springer New York. (Cité en page 176.)

Polka, L., Rvachew, S., et Mattock, K. (2007). Experiential influences on speech perception and speech production in infancy. In Hoff, E. et Shatz, M., editors, Blackwell Handbook of Language Development, pages 153–172. Wiley Online Library. (Cité en page 36.)

Polka, L. et Werker, J. F. (1994). Developmental changes in perception of nonnative vowel contrasts. Journal of Experimental Psychology : Human Perception and Performance, 20(2) :421. (Cité en page 41.)

Porter Jr, R. J. et Castellanos, F. X. (1980). Speech-production measures of speech perception : Rapid shadowing of VCV syllables. The Journal of the Acoustical Society of America, 67(4) :1349–1356. (Cité en page 18.)

Postma, A. (2000). Detection of errors during speech production : A review of speech monitoring models. Cognition, 77(2) :97–132. (Cité en page 51.)

Price, C. J. (2012). A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. NeuroImage, 62(2) :816–847. (Cité en page 26.)

Pulvermüller, F. et Fadiga, L. (2010). Active perception : sensorimotor circuits as a cortical basis for language. Nature Reviews Neuroscience, 11(5) :351–360. (Cité en page 14.)

Pulvermüller, F., Huss, M., Kherif, F., del Prado Martin, F. M., Hauk, O., et Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. Proceedings of the National Academy of Sciences, 103(20) :7865–7870. (Cité en page 13.)

Purcell, D. W. et Munhall, K. G. (2006). Adaptive control of vowel formant frequency : Evidence from real-time formant manipulation. The Journal of the Acoustical Society of America, 120(2) :966–977. (Cité en page 16.)

Rapin, L., Schwartz, J.-L., et Ménard, L. (2017). Are idiosyncrasies in vowel production free or learned ? A study of variants of the French vowel system in biological brothers. The Journal of the Acoustical Society of America, 141(5) :3582–3582. (Cité en pages 110 et 125.)

Rauschecker, J. P. et Scott, S. K. (2009). Maps and streams in the auditory cortex : nonhuman primates illuminate human speech processing. Nature Neuroscience, 12(6) :718–724. (Cité en page 193.)

Remez, R. E. (2015). Analogy and disanalogy in production and perception of speech. Language, Cognition and Neuroscience, 30(3) :273–286. (Cité en page 20.)

Richter, C., Feldman, N. H., Salgado, H., et Jansen, A. (2016). A framework for evaluating speech representations. In Proceedings of the 38th Annual Conference of the Cognitive Science Society, pages 1919–1924. (Cité en page 51.)

Rizzolatti, G., Fadiga, L., Gallese, V., et Fogassi, L. (1996a). Premotor cortex and the recognition of motor actions. Cognitive Brain Research, 3(2) :131–141. (Cité en page 12.)

Rizzolatti, G., Fadiga, L., Matelli, M., Bettinardi, V., Paulesu, E., Perani, D., et Fazio, F. (1996b). Localization of grasp representations in humans by PET : 1. Observation versus execution. Experimental Brain Research, 111(2) :246–252. (Cité en page 12.)

Rogalsky, C., Love, T., Driscoll, D., Anderson, S. W., et Hickok, G. (2011). Are mirror neurons the basis of speech perception ? Evidence from five cases with damage to the purported human mirror system. Neurocase, 17(2) :178–187. (Cité en page 13.)

Rogers, J. C., Möttönen, R., Boyles, R., et Watkins, K. E. (2014). Discrimination of speech and non-speech sounds following theta-burst stimulation of the motor cortex. Frontiers in Psychology, 5 :754. (Cité en page 13.)

Rolf, M. et Steil, J. J. (2012). Goal babbling : a new concept for early sensorimotor exploration. In Humanoid Robots, Workshop on Developmental Robotics : Can developmental robotics yield human-like cognitive abilities ? (Cité en page 57.)

Rolf, M., Steil, J. J., et Gienger, M. (2010). Goal babbling permits direct learning of inverse kinematics. IEEE Transactions on Speech and Audio Processing, 2(3) :216–229. (Cité en pages 57 et 72.)

Rolf, M., Steil, J. J., et Gienger, M. (2011). Online goal babbling for rapid bootstrapping of inverse models in high dimensions. In IEEE International Conference on Developmental and Learning (ICDL 2011), volume 2, pages 1–8. (Cité en page 57.)

Roug, L., Landberg, I., et Lundberg, L.-J. (1989). Phonetic development in early infancy : A study of four Swedish children during the first eighteen months of life. Journal of Child Language, 16(1) :19–40. (Cité en pages vi, 30 et 31.)

Rubin, P., Turvey, M. T., et Van Gelder, P. (1976). Initial phonemes are detected faster in spoken words than in spoken nonwords. Perception & Psychophysics, 19(5) :394–398. (Cité en page 21.)

Saffran, J. R. (2003). Statistical language learning mechanisms and constraints. Current Directions in Psychological Science, 12(4) :110–114. (Cité en page 34.)

Saffran, J. R., Aslin, R. N., et Newport, E. L. (1996). Statistical learning by 8-month-old infants. Science, 274(5294) :1926–1928. (Cité en page 33.)

Saghiran, A. (2017). COSMO WordPhon, une modélisation de l'influence de l'information lexicale dans l'apprentissage des catégories phonémiques. Master's thesis, Université Grenoble-Alpes. (Cité en pages x, 196 et 198.)

Salminen, N. H., Tiitinen, H., et May, P. J. (2009). Modeling the categorical perception of speech sounds : A step toward biological plausibility. Cognitive, Affective, & Behavioral Neuroscience, 9(3) :304–313. (Cité en page 56.)

Samlowski, B. (2016). The syllable as a processing unit in speech production : evidence from frequency effects on coarticulation. PhD thesis, Bielefeld : Universität Bielefeld. (Cité en page 22.)

Samuel, A. G. (2011). Speech perception. Annual Review of Psychology, 62 :49–72. (Cité en page 14.)

Sarma, B. D. et Prasanna, S. M. (2017). Acoustic–phonetic analysis for speech recognition : A review. IETE Technical Review, pages 1–23. (Cité en page 44.)

Sato, M., Grabski, K., Garnier, M., Granjon, L., Schwartz, J.-L., et Nguyen, N. (2013). Converging toward a common speech code : imitative and perceptuo-motor recalibration processes in speech production. Frontiers in Psychology, 4(422) :1–14. (Cité en page 15.)

Sato, M., Grabski, K., Glenberg, A. M., Brisebois, A., Basirat, A., Ménard, L., et Cattaneo, L. (2011). Articulatory bias in speech categorization : Evidence from use-induced motor plasticity. Cortex, 47(8) :1001–1003. (Cité en page 13.)

Sato, M., Tremblay, P., et Gracco, V. L. (2009). A mediating role of the premotor cortex in phoneme segmentation. Brain and Language, 111(1) :1–7. (Cité en page 13.)

Savin, H. B. et Bever, T. G. (1970). The nonperceptual reality of the phoneme. Journal of Verbal Learning and Verbal Behavior, 9(3) :295–302. (Cité en pages 21 et 23.)

Scharenborg, O. (2008). Fine-phonetic variation in a computational model of word recognition. The Journal of the Acoustical Society of America, 123(5) :3072–3072. (Cité en page 45.)

Scharenborg, O. et Boves, L. (2010). Computational modelling of spoken-word recognition processes : Design choices and evaluation. Pragmatics & Cognition, 18(1) :136–164. (Cité en page 44.)

Scharenborg, O., Norris, D., Bosch, L., et McQueen, J. M. (2005). How should a speech recognizer work ? Cognitive Science, 29(6) :867–918. (Cité en page 45.)

Schiller, N. O. (2006). Phonological encoding in speech production. In 2006 ISCA Tutorial and Research Workshop (ITRW) on Experimental Linguistics. (Cité en page 51.)

Schomers, M. R. et Pulvermüller, F. (2016). Is the sensorimotor cortex relevant for speech perception and understanding ? an integrative review. Frontiers in Human Neuroscience, 10(435) :1–18. (Cité en page 20.)

Schroeder, M., Atal, B., et Hall, J. (1979). Objective measure of certain speech signal degradations based on masking properties of human auditory perception. In Björn, L., Öhman, S. E. G., et Fant, G., editors, Frontiers of Speech Communication Research, pages 217–229. London ; New York : Academic Press. (Cité en page 96.)

Schroeter, J. et Sondhi, M. M. (1994). Techniques for estimating vocal-tract shapes from the speech signal. IEEE Transactions on Speech and Audio Processing, 2(1) :133–150. (Cité en page 176.)

Schuerman, W. L., Meyer, A. S., et McQueen, J. M. (2017). Mapping the speech code : Cortical responses linking the perception and production of vowels. Frontiers in Human Neuroscience, 11(161) :1–16. (Cité en page 17.)

Schwartz, J.-L., Basirat, A., Ménard, L., et Sato, M. (2012). The Perception-for-Action-Control Theory (PACT) : A perceptuo-motor theory of speech perception. Journal of Neurolinguistics, 25(5) :336–354. (Cité en page 12.)

Schwartz, J.-L., Boë, L.-J., Vallée, N., et Abry, C. (1997). The dispersion-focalization theory of vowel systems. Journal of Phonetics, 25(3) :255–286. (Cité en page 113.)

Schwartz, J.-L., Escudier, P., et Teissier, P. (2010). Multimodal speech : Two or three senses are better than one. In Mariani, J., editor, Language and Speech Processing, pages 377–415. ISTE, London, UK. (Cité en page 200.)

Schwartz, J.-L., Robert-Ribes, J., Escudier, P., Burnham, B., Campbell, D., et Dodd, R. (1998). Ten years after Summerfield : a taxonomy of models for audio-visual fusion in speech perception. In Campbell, R., Dodd, B., et Burnham, D., editors, Hearing by eye II : Advances in the psychology of speechreading and auditory-visual speech, pages 85–108. Hove, England : Psychology Press/Erlbaum (UK) Taylor & Francis. (Cité en page 200.)

Scott, S. K. et Johnsrude, I. S. (2003). The neuroanatomical and functional organization of speech perception. Trends in Neurosciences, 26(2) :100–107. (Cité en page 14.)

Segui, J., Frauenfelder, U. H., et Mehler, J. (1981). Phoneme monitoring, syllable monitoring and lexical access. British Journal of Psychology, 72(4) :471–477. (Cité en page 21.)

Seidl, A. et Cristia, A. (2012). Infants' learning of phonological status. Frontiers in Psychology, 3 :448. (Cité en page 34.)

Serniclaes, W. et Sprenger-Charolles, L. (2003). Categorical perception of speech sounds and dyslexia. Current psychology letters. Behaviour, brain & cognition, 1(10) :1–8. (Cité en pages vi et 28.)

Sharma, A. et Dorman, M. F. (1999). Cortical auditory evoked potential correlates of categorical perception of voice-onset time. The Journal of the Acoustical Society of America, 106(2) :1078–1083. (Cité en page 24.)

Shiller, D. M., Sato, M., Gracco, V. L., et Baum, S. R. (2009). Perceptual recalibration of speech sounds following speech motor learning. The Journal of the Acoustical Society of America, 125(2) :1103–1113. (Cité en pages 17 et 203.)

Shtyrov, Y., Kujala, T., Ahveninen, J., Tervaniemi, M., Alku, P., Ilmoniemi, R. J., et Näätänen, R. (1998). Background acoustic noise and the hemispheric lateralization of speech processing in the human brain : magnetic mismatch negativity study. Neuroscience Letters, 251(2) :141–144. (Cité en page 24.)

Shtyrov, Y., Kujala, T., Palva, S., Ilmoniemi, R. J., et Näätänen, R. (2000). Discrimination of speech and of complex nonspeech sounds of different temporal structure in the left and right cerebral hemispheres. NeuroImage, 12(6) :657–663. (Cité en page 24.)

Siok, W. T., Jin, Z., Fletcher, P., et Tan, L. H. (2003). Distinct brain regions associated with syllable and phoneme. Human Brain Mapping, 18(3) :201–207. (Cité en pages 25 et 26.)

Skinner, B. F. (1957). Verbal Behavior. BF Skinner Foundation, Cambridge, MA. (Cité en page 33.)

Skipper, J. I., Devlin, J. T., et Lametti, D. R. (2017). The hearing ear is always found close to the speaking tongue : Review of the role of the motor system in speech perception. Brain and Language, 164 :77–105. (Cité en page 14.)

Skipper, J. I., van Wassenhove, V., Nusbaum, H. C., et Small, S. L. (2007). Hearing lips and seeing voices : How cortical areas supporting speech production mediate audiovisual speech perception. Cerebral Cortex, 17(10) :2387–2399. (Cité en pages 12 et 13.)

Smith, B. L., Brown-Sweeney, S., et Stoel-Gammon, C. (1989). A quantitative analysis of reduplicated and variegated babbling. First Language, 9(6) :175–189. (Cité en page 31.)

Stark, R. E. (1980). Stages of speech development in the first year of life. Child phonology, 1 :73–90. (Cité en pages vi, 30, 31 et 36.)

Stasenko, A., Garcea, F. E., et Mahon, B. Z. (2013). What happens to the motor theory of perception when the motor system is damaged ? Language and Cognition, 5(2-3) :225–238. (Cité en page 14.)

Stetson, R. (1951). Motor Phonetics : a study of speech movements in articulation. Amsterdam : North Holland. (Cité en page 21.)

Stevens, K. N. (1989). On the quantal nature of speech. Journal of Phonetics, 17(1) :3–45. (Cité en page 79.)

Stevens, K. N. et Keyser, S. J. (2010). Quantal theory, enhancement and overlap. Journal of Phonetics, 38(1) :10–19. (Cité en page 79.)

Studdert-Kennedy, M. (1975). From continuous signal to discrete message : Syllable to phoneme. In Kavanagh, J. F. et E, C. J., editors, The role of speech in language, pages 113–125. Cambridge, MA : MIT Press. (Cité en page 21.)

Summerfield, Q. (1987). Some preliminaries to a compre hensive account of audio-visual speech perception. In Dodd, B. et Campbell, R., editors, Hearing by eye : The psychology of lip-reading, pages 3–226. Lawrence Erlbaum Associates, Hillsdale, NJ. (Cité en page 200.)

Sussman, H. M. (1984). A neuronal model for syllable representation. Brain and Language, 22(1) :167–177. (Cité en page 52.)

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. Cognitive Psychology, 50(1) :86–132. (Cité en page 33.)

Taniguchi, T., Nagasaka, S., et Nakashima, R. (2016). Nonparametric bayesian double articulation analyzer for direct language acquisition from continuous speech signals. IEEE Transactions on Cognitive and Developmental Systems, 8(3) :171–185. (Cité en page 56.)

Teinonen, T., Aslin, R. N., Alku, P., et Csibra, G. (2008). Visual speech contributes to phonetic learning in 6-month-old infants. Cognition, 108(3) :850–855. (Cité en page 35.)

Thelen, E. (1981). Rhythmical behavior in infancy : An ethological perspective. Developmental Psychology, 17(3) :237–257. (Cité en page 31.)

Thompson, S. P. et Newport, E. L. (2007). Statistical learning of syntax : The role of transitional probability. Language Learning and Development, 3(1) :1–42. (Cité en page 33.)

Toni, I., De Lange, F. P., Noordzij, M. L., et Hagoort, P. (2008). Language beyond action. Journal of Physiology - Paris, 102(1) :71–79. (Cité en page 13.)

Tourville, J. A. et Guenther, F. H. (2011). The DIVA model : A neural theory of speech acquisition and production. Language and Cognitive Processes, 26(7) :952–981. (Cité en pages vi, 48, 49, 50 et 58.)

Trehub, S. E. (1973). Infants' sensitivity to vowel and tonal contrasts. Developmental Psychology, 9(1) :91–96. (Cité en page 27.)

Trehub, S. E. (1976). The discrimination of foreign speech contrasts by infants and adults. Child Development, 47(2) :466–472. (Cité en page 27.)

Treille, A. (2017). Percevoir et agir : La nature sensorimotrice, multisensorielle et prédictive de la perception de la parole. PhD thesis, Université Grenobles-Alpes. (Cité en page 203.)

Tsao, F.-M., Liu, H., Kuhl, P. K., et Tseng, C. (2000). Perceptual discrimination of a Mandarin fricative-affricate contrast by English-learning and Mandarin-learning infants. In Poster presented at the International Meeting of the Society on Infant Studies. Brighton England. (Cité en page 29.)

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., et Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. Proceedings of the National Academy of Sciences, 104(33) :13273–13278. (Cité en pages 55, 59, 60, 70 et 71.)

Varadarajan, B., Khudanpur, S., et Dupoux, E. (2008). Unsupervised learning of acoustic sub-word units. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies, pages 165–168. Association for Computational Linguistics. (Cité en page 56.)

Vaz, M., Brandl, H., Joublin, F., et Goerick, C. (2009). Learning from a tutor : Embodied speech acquisition and imitation learning. In IEEE International Conference on Developmental and Learning (ICDL 2009), pages 1–6. (Cité en pages 59 et 60.)

Vihman, M. M. (2013). Phonological development : The first two years. John Wiley & Sons, Chichester, UK. (Cité en pages vi, 27, 30 et 183.)

Vihman, M. M. et Nakai, S. (2003). Experimental evidence for an effect of vocal experience on infant speech perception. In Proceedings of the 15th International Congress of Phonetic Sciences, pages 1017–1020. (Cité en page 36.)

Villacorta, V. M., Perkell, J. S., et Guenther, F. H. (2007). Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. The Journal of the Acoustical Society of America, 122(4) :2306–2319. (Cité en pages 16 et 19.)

Vinter, A. (1986). The role of movement in eliciting early imitations. Child Development, 57(1) :66–71. (Cité en page 36.)

Walker, G. (2016). Computational Modeling of Speech Production and Aphasia. PhD thesis, University of Irvine. (Cité en page 51.)

Warlaumont, A. S. (2012). A spiking neural network model of canonical babbling development. In The 2nd Joint IEEE International Conference on Developmental and Learning and on Epigenetic Robotics (ICDL-Epirob 2012), pages 1–6. (Cité en pages 58, 60 et 72.)

Warlaumont, A. S., Westermann, G., Buder, E. H., et Oller, D. K. (2013). Prespeech motor learning in a neural network using reinforcement. Neural Networks, 38 :64–75. (Cité en pages 58, 60 et 72.)

Warlaumont, A. S., Westermann, G., et Oller, D. K. (2011). Self-production facilitates and adult input interferes in a neural network model of infant vowel imitation. In AISB 2011 Computational Models of Cognitive Development. Society for the Study of Artificial Intelligence and the Simulation of Behaviour. (Cité en page 58.)

Watkins, K. E., Strafella, A. P., et Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. Neuropsychologia, 41(8) :989–994. (Cité en page 13.)

Weber, A. et Scharenborg, O. (2012). Models of spoken-word recognition. Wiley Interdisciplinary Reviews : Cognitive Science, 3(3) :387–401. (Cité en page 44.)

Werker, J. F. (1994). Cross-language speech perception : Development change does not involve loss. In Goodman, J. C. et Nusbaum, H. C. C., editors, The Development of Speech Perception, pages 95–120. Cambridge, MA : MIT Press. (Cité en pages 33 et 34.)

Werker, J. F., Gilbert, J. H., Humphrey, K., et Tees, R. C. (1981). Developmental aspects of cross-language speech perception. Child Development, 52(1) :349–355. (Cité en page 28.)

Werker, J. F. et Hensch, T. K. (2015). Critical periods in speech perception : new directions. Annual Review of Psychology, 66 :173–196. (Cité en page 182.)

Werker, J. F., Polka, L., et Pegg, J. E. (1997). The conditioned head turn procedure as a method for testing infant speech perception. Early Development and Parenting, 6(34) :171–178. (Cité en page 28.)

Werker, J. F. et Tees, R. C. (1984). Cross-language speech perception : Evidence for perceptual reorganization during the first year of life. Infant Behavior and Development, 7(1) :49–63. (Cité en page 28.)

Westermann, G. et Miranda, E. R. (2004). A new model of sensorimotor coupling in the development of speech. Brain and Language, 89(2) :393–400. (Cité en pages 58 et 60.)

White, K. S., Peperkamp, S., Kirk, C., et Morgan, J. L. (2008). Rapid acquisition of phonological alternations by infants. Cognition, 107(1) :238–265. (Cité en page 34.)

Wilson, S. M. et Iacoboni, M. (2006). Neural responses to non-native phonemes varying in producibility : Evidence for the sensorimotor nature of speech perception. NeuroImage, 33(1) :316–325. (Cité en page 13.)

Wilson, S. M., Saygin, A. P., Sereno, M. I., et Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. Nature Neuroscience, 7(7) :701. (Cité en page 13.)

Wolpert, D. M., Miall, R. C., et Kawato, M. (1998). Internal models in the cerebellum. Trends in Cognitive Sciences, 2(9) :338–347. (Cité en page 67.)

Yeung, H. H. et Werker, J. F. (2009). Learning words' sounds before learning how words sound : 9-month-olds use distinct objects as cues to categorize speech information. Cognition, 113(2) :234–243. (Cité en page 35.)

Yoshida, K. A., Pons, F., Maye, J., et Werker, J. F. (2010). Distributional phonetic learning at 10 months of age. Infancy, 15(4) :420–433. (Cité en page 33.)

Yoshikawa, Y., Asada, M., Hosoda, K., et Koga, J. (2003). A constructivist approach to infants' vowel acquisition through mother–infant interaction. Connection Science, 15(4) :245–258. (Cité en page 59.)

Yu, M., Mo, C., Li, Y., et Mo, L. (2015). Distinct representations of syllables and phonemes in Chinese production : Evidence from fMRI adaptation. Neuropsychologia, 77 :253–259. (Cité en page 26.)

Zekveld, A. A., Heslenfeld, D. J., Festen, J. M., et Schoonhoven, R. (2006). Top–down and bottom–up processes in speech comprehension. NeuroImage, 32(4) :1826–1836. (Cité en pages 13 et 92.)

Zheng, Z. (2012). Perceptual processing of auditory feedback during speech production and its neural substrates. PhD thesis, Queen's University (Canada). (Cité en page 50.)

Ziegler, J. C. et Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages : a psycholinguistic grain size theory. Psychological Bulletin, 131(1) :3. (Cité en page 40.)

Zolnay, A., Schluter, R., et Ney, H. (2005). Acoustic feature combination for robust speech recognition. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2005), pages 457–460. (Cité en page 44.)

Bayesian modeling of joint development of perception, action and phonology

Abstract — Phonetic units can be associated with both auditory and motor representations. In this thesis, we study the development of this double representation and its consequences, mainly in speech perception. We address this set of issues by performing computer simulations with a Bayesian model of communication, named COSMO ("Communicating Objects using Sensory-Motor Operations"). We examine with the model their acquisition, in auditory and motor learning. Simulations suggest that auditory representations are acquired quickly, are based on exogenous processes and are better suited to characterize vowels. On the other hand, motor representations appear to be acquired more slowly, seem to be based on endogenous processes and are better suited to characterize consonants. We observe three consequences of these differences of learning. First, these differences lead to a complementarity during speech perception : auditory representations would be optimally tuned to recognize nominal stimuli, whereas motor representations would have generalization properties and would be able to deal with stimuli typical of adverse conditions. We call this the "auditory-narrow/motor-wide" property. Then, these differences are helpful to better understand the acquisition of variability between representations from one person to another, which is called idiosyncrasies. Simulations suggest that motor representations are acquired thanks to a communicative process rather than a pure sound imitation process. Finally, these differences are used to investigate the development of phonetic units. We show that communication between two agents can occur even if they have different internal representations, and we propose a variant of the model, called COSMO SylPhon, for comparing phoneme and syllable development. Through these three axes, we implemented different versions of the COSMO model based on data from the literature, and discuss them in light of our simulations.

Keywords — Bayesian modeling, auditory and motor representations, distinctive units, learning, perception, idiosyncrasies.

Modélisation bayésienne du développement conjoint de la perception, l'action et la phonologie

Résumé — Les unités phonétiques peuvent être associées à des représentations auditives et motrices. Dans cette thèse, nous étudions le développement de cette double représentation et ses conséquences, principalement durant la perception. Nous abordons ces questions à l'aide de simulations informatiques réalisées à l'aide d'un modèle bayésien de la communication, nommé COSMO ("Communicating Objects using Sensory-Motor Operations"). Nous analysons dans ce modèle les processus d'apprentissage auditif et moteur. A la lumière des simulations, il apparaît que les représentations auditives sont acquises rapidement, sont basées sur des processus exogènes et caractérisent mieux les voyelles. Par contraste, les représentations motrices sont acquises plus lentement, sont basées sur des processus endogènes et caractérisent mieux les consonnes. Nous observons trois conséquences issues de ces différences d'apprentissage. D'abord, elles permettent de mettre en avant l'existence possible de deux voies complémentaires durant la perception : les représentations auditives seraient ajustées de manière optimale pour reconnaître les stimuli standards, tandis que les représentations motrices traiteraient mieux les stimuli inhabituels, dans des conditions de communication « adverses ». Nous appelons ceci la propriété « auditif-bande étroite/moteur-bande large ». Ensuite, ces différences servent à mieux comprendre comment apparaît la variabilité interpersonnelle, ce qui est nommé idiosyncrasies. Les simulations suggèrent que les représentations motrices sont acquises à l'aide d'un processus communicatif plutôt que par un processus purement imitatif. Finalement, ces différences d'apprentissage sont utilisées pour étudier plus spécifiquement le développement des unités phonétiques. Nous montrons que la communication peut s'effectuer même lorsque les deux interlocuteurs possèdent des représentations internes différentes, et nous proposons une version du modèle, intitulée COSMO SylPhon, permettant de mettre en correspondance les développements des syllabes et le développement des phonèmes. A travers ces trois axes, nous avons implémenté différentes versions de notre modèle COSMO en nous basant sur les données de la littérature, et en les discutant en retour à la lumière des simulations.

Mots clés — Modélisation bayésienne, représentations auditives et motrices, unités distinctives, apprentissage, perception, idiosyncrasies.