Rôle des relations perception-action dans la communication parlée et l'émergence des systèmes phonologiques : étude, modélisation computationnelle et simulations

Clément Moulin-Frier

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

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

Arrêté ministérial : 7 août 2006

Présentée par

Clément Moulin-Frier

Thèse dirigée par **Jean-Luc Schwartz**
et codirigée par **Julien Diard et Pierre Bessière**

préparée au sein **du Laboratoire GIPSA**
et de **l'École Doctorale Ingénierie pour la Santé la Cognition et l'Environnement (EDISCE)**

Rôle des relations perception-action dans la communication parlée et l'émergence des systèmes phonologiques : étude, modélisation computationnelle et simulations

Thèse soutenue publiquement le **15 juin 2011**,
devant le jury composé de :

M. Jacques Droulez
DR CNRS, Collège de France, Rapporteur
M. Yves Laprie
DR CNRS, Laboratoire lorrain de recherche en informatique et ses applications, Rapporteur
M. Pierre-Yves Oudeyer
CR INRIA Bordeaux Sud-Ouest, Examinateur
M. Michael A. Arbib
Professor, University of Southern California, Examinateur
M. Augustin Lux
Professeur Grenoble-INP, INRIA Rhône-Alpes, Examinateur
M. Jean-Luc Schwartz
DR CNRS GIPSA-Lab, Directeur de thèse
M. Julien Diard
CR CNRS LPNC, Co-Directeur de thèse
M. Pierre Bessière
DR CNRS LPPA Collège de France, Co-Directeur de thèse

Table des matières

Table des figures

Remerciements

À ceux que j'ai pu oublier dans ces remerciements rédigés rapidement.

À mon équipe d'encadrants, Jean-Luc, Julien et Pierre, qui ont largement contribué à l'avancement de ce boulot lors de nos longues, presque régulières, et toujours passionnantes réunions. Ça a été un plaisir de bosser avec vous et j'espère sincèrement pouvoir continuer à travailler avec des personnes de votre qualité dans la suite de mon parcours.

Jean-Luc, quelques lignes supplémentaires pour mettre en valeur tout ce que tu as pu m'apporter. Tu as compris dès le début de nos interactions ma façon de fonctionner et mes centres d'intérêts, tu les as intégrés dans notre démarche et tu m'as appris à tirer le mieux de cet ensemble un peu brouillon. Ce que tout ceci m'a apporté sur le plan personnel est inestimable, je t'en remercie chaleureusement.

Aux personnels du Gipsa-Lab qui réussissent à faire tourner cette grosse machine, de l'administration (Nadine, Florence, Akila ...) à l'équipe technique (Nino, Laurent, Christian ...).

À tous ceux qui cherchent et qui m'ont fait prendre plaisir à le faire, des permanents du laboratoire (Louis-Jean, Marc, Pascal, Pierre ...) aux doctorants, stagiaires et post-doc (Lucile, Émilie, Olha, Benjamin, Amélie, Raphaël, Audrey ...) avec qui les discussions, mêmes courtes, ont souvent une admirable tendance à créer autant de nouvelles questions qu'elles n'en résolvent.

À l'équipe Parole Cerveau Multimodalité Développement (PCMD) qui m'a accueilli et a fait l'effort de s'intéresser à des travaux de modélisation plus ou moins marginaux.

À mes hôtes de l'USC à Los Angeles qui m'ont accueilli le temps d'un semestre à l'« USC Brain Project ». De Michael A. Arbib, qui m'a intégré dans ses recherches avec une agréable bienveillance, pendant et après mon séjour, aux doctorants du laboratoire, en particulier James et JinYong qui m'ont fait découvrir la vie californienne.

À la région Rhône-Alpes qui m'a permis de réaliser cette belle expérience grâce à son programme Explora-doc.

À toutes ceux qui ont fait partie de ma vie quotidienne et nocturne dans cette ville hallucinante, de mes colocataires Emery et Aziz à celles et ceux qui sont devenus bien plus que de simples rencontres, Valou, Andrea et bien d'autres.

À tous mes potes de Grenoble, Lyon, Saint-Etienne avec qui les apéros, les soirées très sonores, les vacances sur la route, les discussions enflammées et les quelques naissances ont été, et seront toujours, une soupape indispensable pour ne pas péter les plombs dans ce monde compliqué. Vous êtes surtout ce qu'il y a de plus important dans ma vie après ceux qui suivent. Mes amis, je ne cite pas vos noms car votre recensement est un travail à part entière mais vous vous reconnaitrez !

À ma famille bien sûr, ma bande des quatre et les quelques autres qui forment mon cocon certes restreint mais tellement important. Vous m'avez laissé avancer à mon rythme en me permettant toujours de faire mes propres choix et vous avez su me rappeler que faire une thèse était complétement en accord avec mes désirs et délires enfantins : être inventeur comme Gaston Lagaffe (merci donc à lui et à son créateur au passage pour m'avoir inspiré un certain style de vie).

Clément Moulin-Frier, le 8 septembre 2011.

Chapitre 1

Introduction

1.1 Contexte scientifique : un mythe, et des périls

Le processus phylogénétique ayant conduit à l'émergence de la parole et du langage fait partie des énigmes majeures pour la pensée humaine, au même titre que la question de l'origine de l'univers ou de la vie. Mais bien que cette transition essentielle, fondatrice de notre humanité même, soit apparue relativement récemment, les éléments dont nous disposons pour la comprendre ne sont pas nombreux. Contrairement à la vie qui a laissé sur son passage une série d'indices fossilisés, le langage ne laisse pas de traces avant l'apparition de l'écriture et on ne peut pas étudier des formes conservées de « proto-langage ».

La question de l'origine du langage apparaît ainsi comme redoutablement peu contrainte par les faits expérimentaux. C'est ce qui avait conduit en 1866 la Société de Linguistique de Paris à en interdire l'étude ! Depuis, les choses ont certes bien changé, et cette question connaît un indéniable regain d'intérêt depuis quelques décennies, suscitant des recherches nombreuses au carrefour de disciplines variées, comme l'illustre la Figure 1.1 proposée par Christiansen et Kirby (2003).

Reste que, si le manque d'indice direct sur l'émergence du langage a contraint à ouvrir considérablement le champ de recherche pour extraire des éléments de réponse dans l'interaction entre de nombreuses disciplines, les risques de l'extrapolation fantaisiste et de la théorisation débridée demeurent. Nous renvoyons à l'article récent de Boë et collab. (2011) pour une analyse historique et épistémologique de ce paysage complexe de propositions et d'hypothèses sur les mécanismes susceptibles d'avoir contribué à l'émergence du langage humain.

Nous en retiendrons une idée forte : si la question des discontinuités, c'est-à-dire des avantages évolutifs qui pourraient avoir guidé l'émergence de cette spécificité humaine centrale reste mystérieuse et à vrai dire quasiment impossible à tester (d'autant que c'est l'homme lui-même qui s'est mis en charge d'estimer ce que pourraient avoir été ses propres avantages !), la question des continuités, elle, semble beaucoup plus solidement testable. Cette question est celle de ce qui nous rapproche de nos prédécesseurs et cousins primates, plutôt que ce qui nous en éloigne. Elle conduit ainsi, non pas à tenter de définir ce que

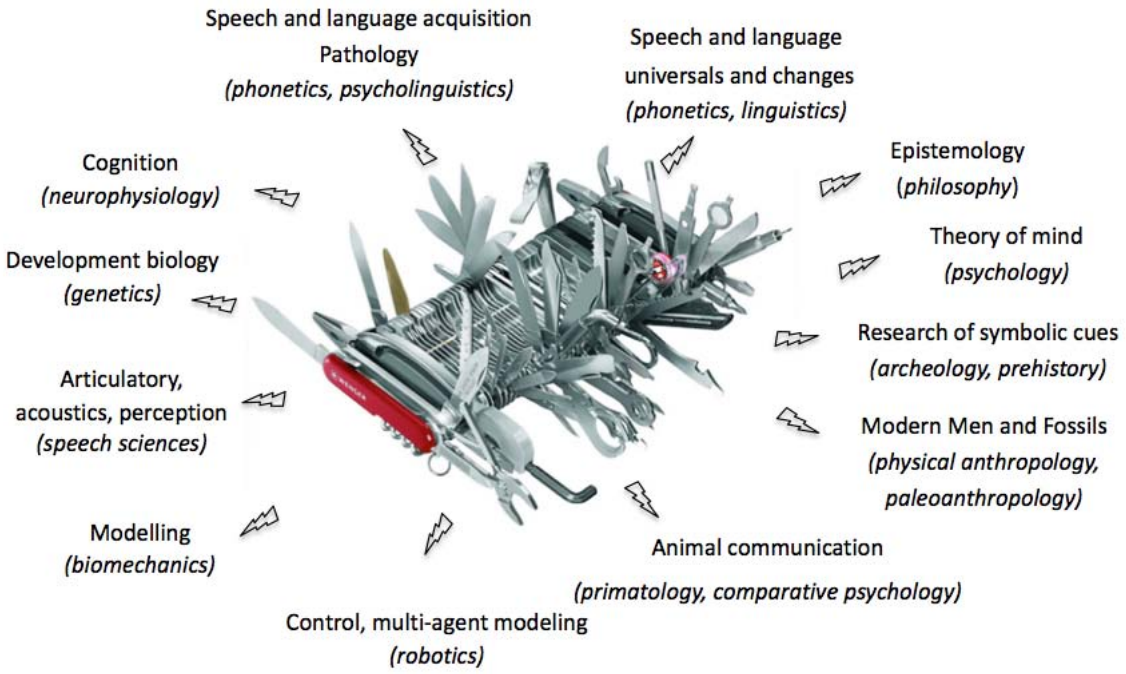


FIGURE 1.1 – Pluridisciplinarité des recherches sur l'émergence de la parole et du langage, d'après Christiansen et Kirby (2003).

le langage humain a de « mieux » que les facultés cognitives des primates non humains (« mieux » étant ici pris au sens précis de « ayant produit un avantage évolutionnaire au sens darwinien »), mais comment il pourrait s'inscrire en continuité de ces facultés cognitives.

1.2 Problématique : morphogenèse des unités du langage

Si la question de l'origine du langage reste d'un abord extrêmement compliqué, il est une question qui semble, elle, plus susceptible de se confronter à la démarche expérimentale, c'est celle de l'origine des formes du langage.

Sous l'infinie variété de ses formes, le langage humain se caractérise en effet par d'évidentes permanences, les « universaux » du langage. Toutes les langues orales se caractérisent par leur « double articulation » en unités de sens (les mots et leurs déclinaisons morphologiques) et en unités de sons (les phonèmes). Toutes les langues présentent au niveau phonémique des alternances plus ou moins régulières de consonnes et de voyelles en syllabes. A l'intérieur même des systèmes sonores des langues du monde, on observe, sinon des universaux, du moins des régularités fortes, avec les voyelles et consonnes « stars »

omniprésentes dans les langues du monde, /i a u/ [1] pour les unes, /p t k/, /f s/ ou /m n/ pour les autres (Schwartz et collab., 2007). La question qui se pose est alors celle de l'origine de la régularité de ces formes. On peut proposer, et suivre dans la littérature, trois types de réponse.

La première est celle initiée par Noam Chomsky avec son « organe du langage » : les propriétés fondatrices et universelles du langage sont contenues dans cet « organe », cet ensemble de compétences linguistiques qui sont contenues toutes entières dans notre patrimoine génétique : « La grammaire universelle et la grammaire de l'état stationnaire (c'est-à-dire final) sont réelles. On s'attend à les trouver physiquement représentées dans le code génétique et le cerveau adulte respectivement, avec les propriétés mises au jour par notre théorie de l'esprit » (Chomsky, 1985). Ainsi, l'invariant linguistique est en réalité un invariant génétique, s'exprimant sous des formes diverses mais unifiées au sein des langues humaines.

La seconde est celle de l'origine commune des langues humaines. Merritt Ruhlen en est le porteur emblématique, avec sa proposition que toutes les langues du monde partagent la même origine, celle d'une langue mère parlée en Afrique, foyer des Hommes modernes, il y a environ 100 000 ans, avant d'évoluer au sein de migrations progressives (Ruhlen, 1996). La forme commune des langues serait ainsi un reflet de cette unicité initiale, évoluant peu à peu jusqu'à présenter des formes culturellement variées, mais conservant des propriétés issues de la forme fondatrice (voir un avatar récent de cette hypothèse d'une origine africaine commune dans Atkinson (2011)).

La troisième hypothèse est celle d'une « mise en forme » des unités du langage par des processus dynamiques qui produiraient des caractéristiques communes issues de mécanismes d'ajustement et d'optimalisation. C'est l'hypothèse de base des théories « substance-based » popularisées par Bjorn Lindblom depuis près de 40 ans, et bien résumée, dans son article de 1984 « Can the models of evolutionary biology be applied to phonetic problems ? » par la formule « Peut-on dériver le langage du non-langage ? »

Cette proposition a ouvert tout un programme de recherche visant à comprendre comment une substance non-langagière, constituée de l'ensemble des mécanismes biologiques, cognitifs et environnementaux présents avant le langage, a pu non seulement en permettre l'émergence mais a également contraint ses propriétés universelles, sa forme. Cette nouvelle approche considère donc les universaux du langage (par exemple les régularités observées dans les systèmes phonologiques des langues du monde) comme le résultat d'un système d'optimisation sous contraintes. On peut proposer la formule de « morphogénèse des unités du langage », par analogie ou extension du sens initial de ce mot venu de la biologie et décrivant l'étude des mécanismes qui permettent à un organisme vivant de développer et contrôler ses formes. Par morphogénèse des unités du langage nous entendons genèse des formes du langage à partir de sa substance prélangagière et dans son environnement spécifique (cognitif, communicatif, social).

1. /u/ comme dans « ch**ou** ».

1.3 Stratégie : des principes dynamiques d'optimisation dérivés de processus d'interaction entre agents cognitifs

Notre stratégie s'organise en quatre points principaux.

1.3.1 Ancrer les théories de la forme dans les théories de l'émergence, et les opérationnaliser par des mécanismes d'interactions locales

En 1972 sont apparues deux « théories de la forme », décrites par Lindblom comme « orientées substance » et qui visaient à expliquer les formes du langage à partir d'une substance non-langagière, en particulier : les contraintes des systèmes articulatoires et auditifs, ainsi que leurs architectures cognitives sous-jacentes ou les processus d'interaction et d'apprentissage entre individus. Nous décrirons plus tard ces deux théories, la « théorie de la dispersion » de Liljencrants et Lindblom (1972) et la « théorie quantique » de Stevens (1972).

Notre première hypothèse fondatrice est que ces théories devraient s'ancrer dans des raisonnements plus généraux sur l'émergence du langage, et notamment dans la recherche de ses précurseurs onto- et phylogénétique, typiquement chez les primates non-humains et les bébés. Ces précurseurs fournissent des éléments comportementaux, sensori-moteurs, neuroanatomiques ou cognitifs qui ne sont pas encore du langage mais qui peuvent constituer les bases à partir duquel il aurait pu émerger. Nous nous proposons de tenter de dériver les théories de la forme (théorie de la dispersion et théorie quantique) d'ingrédients contenus dans ces mécanismes précurseurs.

Pour cela, nous nous appuierons sur les principes de simulation multi-agents, initiée par les travaux pionniers de Luc Steels vers le milieu des années 90 (par exemple Steels, 1996, 1997). Les simulations considérant les formes sonores du langage comme le résultat d'une optimisation sous contraintes se sont longtemps cantonnées à des approches dites globales, cherchant l'équilibre d'un système macroscopique à la manière de la thermodynamique. Ces approches sont maintenant complétées par des approches dites locales qui étudient comment cet équilibre peut émerger de l'interaction entre les constituants microscopiques du système en question, à la manière de la mécanique statistique. Pour cela, ces simulations mettent en jeu des agents prélangagiers en interaction et étudient comment des propriétés du langage humain peuvent en émerger.

Les travaux de cette thèse s'inscrivent dans ce programme de dérivation du langage à partir du non-langage, en considérant la substance comme un système complexe dont l'interaction entre les constituants peut laisser émerger des formes universelles.

1.3.2 Deux hypothèses clés pour ancrer la communication en ligne et en émergence

La communication langagière est ancrée dans un double lien entre locuteur et auditeur, celui qui associe signaux produits et signaux perçus (c'est la question de la parité, qui assure que « ce qui compte » pour le locuteur soit également « ce qui compte » pour l'auditeur) et celui qui associe signaux de communication et objets du monde (c'est la question de la référence).

A chacun de ces liens est associé un possible précurseur, qui a été invoqué comme base possible d'une théorie de l'émergence du langage : les neurones miroir pour la parité, et la deixis pour la référence. Nous aurons l'occasion de présenter plus en détail les théories correspondantes.

Nous appuyons toute notre démarche modélisatrice sur ces deux précurseurs. Ainsi, les deux hypothèses préalables centrales de notre travail, que nous détaillerons, formaliserons et simulerons en détail tout au long de ce document, sont :

une hypothèse d'internalisation d'une situation de communication dans l'architecture cognitive des agents, associant représentations motrices et sensorielles à travers un lien sensori-moteur (parité) ;

une hypothèse de communication prélangagière fournissant une référence d'ordre sémantique et agissant comme une amorce évolutive permettant la morphogénèse de la parole (référence).

Sur ces deux hypothèses, nous bâtissons des modèles de communication qui ont la caractéristique forte (et rare) de s'appliquer à la fois comme modèles d'émergence et comme modèles de communication en ligne. Les recherches sur la phylogenèse de la parole et du langage d'une part, sur la communication parlée d'autre part, sont le plus souvent menées en total cloisonnement. Nos propositions visent à intégrer l'une et l'autre dans un même modèle conceptuel.

1.3.3 De la « construction par le moins » à la « construction par le lien »

Nos travaux s'inscrivent clairement dans la voie ouverte par Steels de simulations de sociétés d'agents prélangagiers, tels que ceux de Berrah (1998); de Boer (2000); Oudeyer (2003). Sans décrire en détail ces travaux dans cette introduction (nous nous y attacherons au cours de la première partie), nous souhaitons seulement esquisser quelques différences d'ordre épistémologique avec l'approche d'Oudeyer. Dans sa thèse, il prend le parti du minimum d'hypothèses préalables pour montrer que certaines formes du langage peuvent apparaître par de simples mécanismes d'auto-organisation et défend l'idée qu'une théorie n'a pas à être proche de la réalité pour être utile à la compréhension du domaine dans lequel elle s'applique, en particulier si celui-ci est encore peu compris. Ses travaux fournissent ainsi à la communauté un résultat important : les systèmes phonologiques peuvent émerger de l'interaction entre agents qui n'ont aucun objectif communicatif, mais simplement la faculté

de produire des vocalisations, conséquences d'un couplage de cartes neurales articulatoire et auditive mises à jour par une règle de Hebb (renforcement des connexions entre les neurones dont les activités sont corrélées). Les théories de l'émergence orientées substance ne sont donc d'aucune utilité à l'élaboration du système, ce qu'Oudeyer assume complètement et argumente dans une réflexion épistémologique sur cette « méthode de l'artificiel » qui occupe un chapitre entier de sa thèse (Oudeyer, 2003, Chapitre 5).

Par rapport à cette approche que l'on pourrait baptiser de « construction par le moins », visant à démontrer qu'un *minimum* de moyens computationnels et d'hypothèses préalables suffisent à produire le « plus » qu'est la naissance d'un système sonore semblable aux systèmes attestés dans les langues humaines, nous utiliserons une approche (complémentaire à nos yeux, plutôt que rivale), de la « construction par le lien ». En effet, dans une démarche similaire visant à la compréhension du phénomène étudié, la morphogénèse des unités du langage, nous recherchons en quelque sorte le pari inverse du *maximum* d'hypothèses préalables. Nous concevons notre démarche comme visant à articuler entre elles, à « lier » des hypothèses plausibles et des mécanismes cognitifs attestés, en voulant faire de notre travail de modélisation un travail de connexion entre des matériaux théoriques variés.

1.3.4 Un outillage théorique et technique indispensable pour formaliser et expliciter systématiquement les hypothèses : la programmation bayésienne

L'enjeu de formalisation et d'explicitation des hypothèses est donc considéré comme central dans ce travail de thèse. Plus nous avons avancé dans cette recherche, plus il nous est apparu que les hypothèses implicites et les mécaniques cachées étaient légion dans un travail de simulation. Nous avons donc cherché en permanence à faire la « chasse à l'implicite », et nous sommes appuyés pour cela sur le formalisme de Programmation Bayésienne des Robots dans lesquels la spécification des connaissances préalables est au contraire érigée en dogme et dont les paramètres s'identifient facilement à des mesures physiques ou cognitives (Bessière et collab., 1999, 1998).

1.4 Mise en œuvre : Plan de lecture

Nos hypothèses centrales nous ont conduit à étudier trois thèmes principaux.

Communication : définition d'une situation de communication parlée, mettant en jeu un agent locuteur capable de transmettre un objet de communication à un auditeur par le biais d'une transformation articulatori-acoustique réalisée par l'environnement ;

Agent communicant : définition d'une architecture cognitive réaliste de production et de perception de la parole internalisant la situation de communication dans un système moteur, un système auditif et un lien sensorimoteur ;

Émergence par interaction entre agents : étude de l'émergence de la parole et du langage, en lien avec les régularités observés dans les langues du monde, s'appuyant

sur des principes d'optimisation issus des théories de la forme ainsi que des amorces prélangagières issues des théories de l'émergence.

Dans notre volonté d'expliciter rigoureusement les hypothèses que nous incluons dans nos modèles, nous proposons d'aborder ces thèmes en trois étapes incrémentales, structurant ce document en trois parties.

De la revue de la littérature aux modèles conceptuels : nous extrayons de la littérature concernant les trois thèmes ci-dessus les connaissances préalables et les données que nous synthétisons et unifions dans des modèles conceptuels.

Des modèles conceptuels aux modèles formels : nous formalisons les connaissances préalables issues des modèles conceptuels obtenus à la partie précédente avec la rigueur imposée par l'outil bayésien.

Des modèles formels à la simulation informatique : nous instancions les modèles formels et étudions la cohérence globale de nos résultats de simulations avec les données phonétiques et avec les prédictions des théories de la forme.

Notre travail s'organise ainsi dans une grille de lecture à deux dimensions. La première concerne les trois étapes de réalisation et constitue les trois parties de ce document. La seconde concerne les trois thèmes et constitue les chapitres à l'intérieur de chaque partie. Cette structure est résumée Table 1.1.

	Communication	Agent communicant	Émergence
De la revue de la littérature aux modèles conceptuels	Chapitre 2	Chapitre 3	Chapitre 4
Des modèles conceptuels aux modèles formels	Chapitre 7	Chapitre 8	Chapitre 9
Des modèles formels à la simulation informatique	Chapitres 10 à 13		

TABLE 1.1 – Plan de lecture du document.

Nous présentons ci-dessous le détail des chapitres.

1.4.1 Première partie : de la littérature en sciences cognitives aux modèles conceptuels

La première partie étudie les trois thèmes de cette thèse sous l'angle des sciences cognitives, en proposant une étude bibliographique de chacun d'entre eux dans le but d'extraire

des connaissances préalables pour en concevoir des modèles conceptuels synthétiques et unifiés.

Nous étudions d'abord une situation de communication parlée (Chapitre 2) mettant en jeu un agent locuteur qui doit transmettre un objet de communication (sémantique ou phonétique) à un agent auditeur, à travers un environnement qui transforme un geste vocal en un stimulus auditif. Nous décrivons simplement les interfaces de cette ligne de communication : le système moteur du locuteur, la transformation articulatori-acoustique effectuée par l'environnement, et le système auditif de l'auditeur. Cette ligne de communication, permettant de lier un objet dans la tête du locuteur à un objet dans celle de l'auditeur constitue notre premier modèle conceptuel.

Puis nous nous intéressons plus précisément aux architectures cognitives de ces agents communicants (Chapitre 3), en passant en revue la littérature des domaines de la production et de la perception de la parole et en identifiant trois grands courants : les théories motrices, auditives et sensori-motrices. Nous argumentons sur la nécessité de concevoir des théories de la communication complètes étudiant conjointement production et perception, en remarquant que ces théories doivent accorder un rôle fondamental à un lien cognitif entre représentations motrices et sensorielles. Nous explicitons alors la première hypothèse centrale de notre travail : une architecture cognitive associant un système moteur, un lien sensori-moteur, et un système auditif dispose de tous les éléments fonctionnels nécessaires à une internalisation de la situation de communication parlée.

Enfin, nous étudions la question de l'émergence des systèmes phonologiques (Chapitre 4), en passant en revue les différentes approches existantes : les théories de la forme des systèmes phonologiques, les théories de l'émergence du langage et les modèles computationnels d'agents interagissant. Dans notre démarche de modélisation de cette émergence, les premières constituent notre objectif, les secondes nos inspirations et les derniers notre moyen. Nous terminons sur notre modèle conceptuel général, dans lequel des agents dotés de l'architecture cognitive proposée au Chapitre 3 interagissent à partir d'une amorce évolutive de communication prélangagière inspirée des théories de l'émergence. Ceci constitue notre deuxième hypothèse centrale : une amorce évolutive de communication prélangagière fournit des principes opérationnels pour la morphogénèse de la parole.

Le Chapitre 5 synthétise cette première partie.

1.4.2 Deuxième partie : des modèles conceptuels aux modèles formels

La deuxième partie étudie les trois thèmes de la thèse sous l'angle de la modélisation mathématique, en utilisant l'outil de Programmation Bayésienne des Robots (présentée au Chapitre 6) qui nous permet de formaliser rigoureusement les connaissances préalables extraites de la première partie.

Nous commençons par définir les variables de la situation de communication parlée et leurs dépendances en termes bayésiens (Chapitre 7).

Puis, reprenant notre hypothèse d'internalisation de cette situation de communication

dans l'architecture cognitive des agents, nous proposons au Chapitre 8 une unification probabiliste des théories de la production et de la perception de la parole, qui exprime les théories motrices, auditives et sensori-motrices de la communication parlée comme différentes utilisations d'un unique modèle bayésien d'agent communicant.

Enfin dans le Chapitre 9, nous spécifions les processus d'évolution des agents par apprentissage dans un algorithme d'interaction déduit de modèle conceptuel général du Chapitre 4.

1.4.3 Troisième partie : des modèles formels aux simulations informatiques

Cette dernière partie étudie nos trois thèmes sous l'angle de la simulation informatique, en instanciant les modèles formels définis dans la partie précédente pour les simuler et analyser les résultats obtenus en regard des données des langues du monde et des principes des théories de la forme.

Nous commençons par instancier notre modèle dans un espace articulatori-auditif très simplifié (Chapitre 10). Ceci nous permet d'une part d'analyser les propriétés générales des différentes architectures cognitives étudiées au Chapitre 3 et formalisées au Chapitre 8, d'autre part de répondre à la problématique exposée à la Section 1.2 : comment des principes de théories de la forme peuvent s'ancrer dans des hypothèses issues de théories de l'émergence, dans un cadre de modélisation quantitative basée sur la simulation multi-agents ?

Les trois derniers chapitres exploitent pleinement les capacités du système dans des simulations d'émergence des systèmes de voyelles (Chapitre 11), de consonnes (Chapitre 12) et de syllabes (Chapitre 13) utilisant un modèle réaliste du conduit vocal humain. Nous montrons que la cohérence globale de nos résultats de simulations avec les données des langues du monde dépend des contraintes articulatori-auditives imposées au système, qui peuvent là encore constituer des connaissances préalables issues de théories de l'émergence.

Nous concluons par une discussion générale qui tire les principales leçons de nos simulations, et analyse les manques et les limitations, et se conclut par un certain nombre de perspectives.

Première partie

De la revue de la littérature aux modèles conceptuels

Chapitre 2

Présentation d'une situation de communication parlée

Nous allons dans ce premier chapitre définir notre situation de communication et poser les objets de base sur lesquels reposera tout ce document. Nous le ferons dans une démarche volontairement simple et synthétique, et renverrons la réflexion théorique dans toute sa complexité bibliographique au prochain chapitre.

La situation que nous considérons est schématisée Figure 2.1.

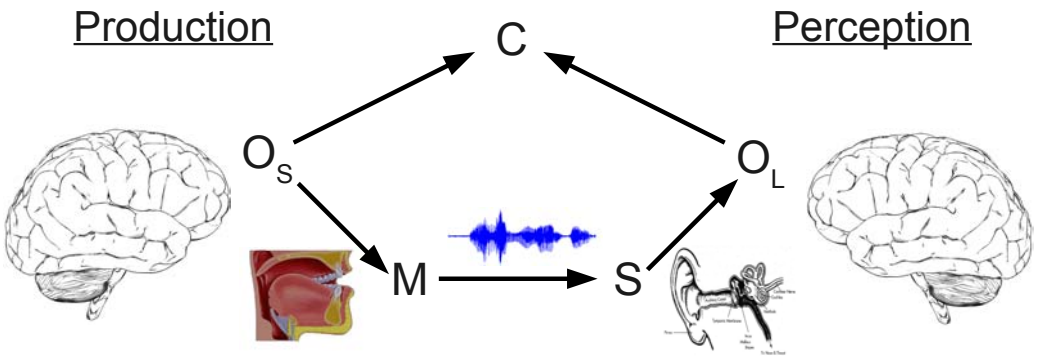


FIGURE 2.1 – Situation de communication parlée. Voir texte pour détail.

Nous définissons une situation de communication parlée de la façon suivante. Un locuteur doit communiquer un certain objet O_S à un agent auditeur [1]. Par objet, nous entendons ici un objet de communication au sens large, quelle que soit sa nature, sémantique ou phonémique par exemple. Pour cela, l'agent locuteur dispose d'un cerveau, doté d'un ensemble

1. O_S pour *Speaker* : nous conservons dans ce texte rédigé en français des notations anglaises que nous avons adoptées dès le départ du travail, et déjà utilisées dans plusieurs articles et communications. Les notations décimales dont également exprimées avec des points (par exemple, nous notons 1.5 la division de 3 par 2.

de représentations et de processus cognitifs de contrôle de la forme d'un conduit vocal à travers différents articulateurs. Nous notons globalement M, pour *Moteur*, cet ensemble (voir Figure 2.1). Ce conduit vocal produit alors une onde sonore, à partir de laquelle l'auditeur doit inférer l'objet de la communication à l'aide de son oreille lui permettant de percevoir le son, ainsi que de son cerveau, doté d'un ensemble de représentations et de processus cognitifs de traitement auditif. Nous notons globalement S, pour *Sensoriel*, cet ensemble et O_L l'objet inféré par l'auditeur [2] (voir Figure 2.1). Finalement, le succès de la communication est défini par la condition $O_S = O_L$, que nous notons C. Nous allons maintenant nous attacher à décrire, toujours brièvement, les étapes permettant de passer de O_S à O_L.

2.1 Le contrôle du conduit vocal

Le conduit vocal est contrôlé à travers un ensemble de muscles reliés à différents articulateurs. Ces derniers sont le larynx (où sont logées les cordes vocales), le pharynx, la mâchoire, la langue et les lèvres (réglant globalement la forme du conduit vocal), et le vélum (permettant l'ouverture de la cavité nasale). Chacun des ces articulateurs possède un ou plusieurs degrés de liberté. Ainsi, dans le modèle articulatoire dont nous disposons au laboratoire GIPSA (VLAM, que nous détaillerons en Partie III), on spécifie une configuration des articulateurs par sept paramètres (Maeda, 1989), représentés Figure 2.2 :

- l'ouverture des lèvres (lip height),
- la protrusion des lèvres (lip protusion),
- la position du corps de la langue (tongue body),
- la position du dos de la langue (tongue dorsum),
- la position de la pointe de la langue (tongue tip),
- l'ouverture de la mâchoire (jaw),
- la hauteur du larynx (larynx height).

Les commandes neurales vers les muscles du conduit vocal proviennent du cortex moteur, situé dans le lobe frontal (voir Figure 2.3). L'organisation des aires motrices est d'une grande complexité, nous ne l'aborderons pas ici (voir une revue dans Kent (1997)). Mais deux principes nous semblent devoir être retenus, pour mieux comprendre la nature de ce système de représentations et de contrôles que nous avons baptisé M.

D'abord, la « tête de pont » de ce système est l'aire motrice primaire, qui est la région qui gère directement, à travers le faisceau complexe des relais sous-corticaux, le chemin vers l'activation des muscles par les neurones moteurs. Une propriété remarquable de cette aire est son organisation cartographique, qui fonde la notion de représentation telle qu'on peut l'imaginer pour des variables motrices. En effet, des stimulations électriques de cette aire déclenchent des activations musculaires très localisées, permettant de dresser une carte somatotopique de cette zone dans laquelle la portion occupée pour une certaine partie du corps est proportionnelle, non pas à sa taille, mais à la complexité des mouvements qu'elle est capable d'effectuer (on parle d'homoncule moteur, Penfield et collab., 1954). Ainsi, les

2. O_L pour *Listener*.

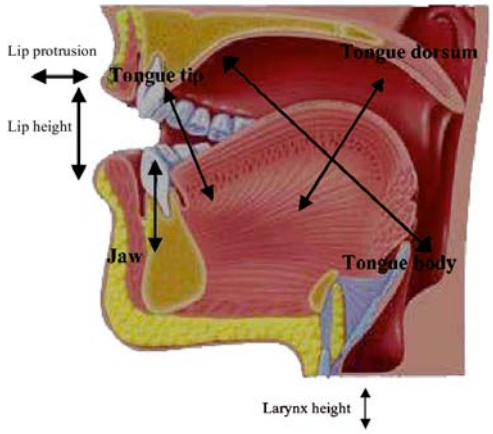


FIGURE 2.2 – Les sept degrés de liberté du conduit vocal.

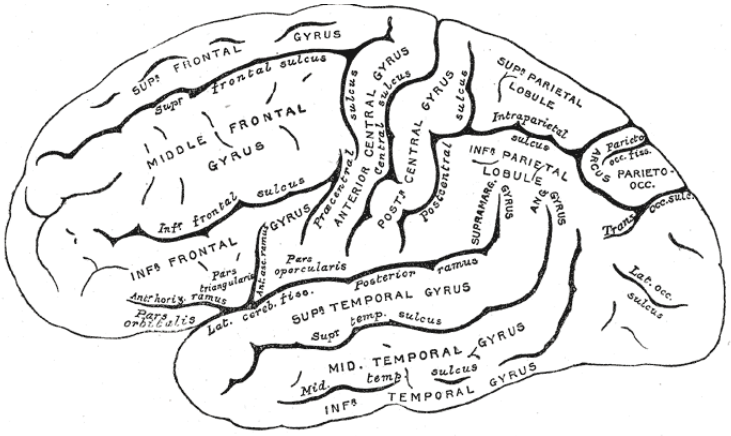
parties correspondant aux muscles manuels et orofaciaux sont largement prépondérantes, traduisant l'importance de deux comportements majeurs pour l'espèce humaine : manipuler et parler.

D'autre part, en amont de ce système cartographié de relais vers la périphérie, le cortex dispose d'une organisation complexe d'aires intégratives qui commandent et organisent la gestion des informations relayées ensuite par l'aire motrice primaire. Ce système est impliqué dans la préparation et la sélection des mouvements complexes, en recevant des flux de différentes aires sensorielles. Il est caractérisé à la fois par des propriétés de hiérarchie et de complexité, et de parallélisme des traitements. Il est situé dans le cortex prémoteur s'étendant vers le bas (direction dite « ventrale ») jusqu'à l'aire de Broca, et vers le haut (direction « dorsale ») jusqu'à l'aire motrice supplémentaire.

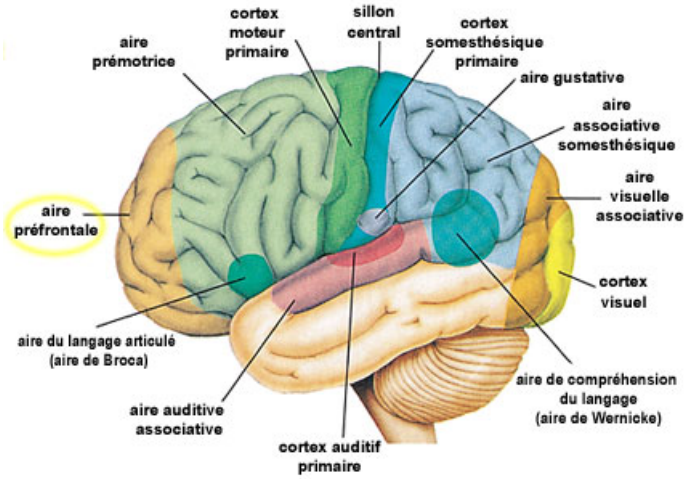
L'aire de Broca est connue pour ses fonctions langagières depuis les observations de Paul Broca (au XIX^{e}siècle) sur un patient aphasique ayant souffert d'une détérioration d'une partie de ses facultés linguistiques suite à une lésion cérébrale que Broca a pu, après le décès du patient, localiser dans le gyrus frontal inférieur gauche. L'aphasie de Broca en particulier se caractérise principalement par des déficits en production de la parole : difficulté à enchaîner des phonèmes ou à construire des structures grammaticales.

L'aire motrice supplémentaire semble également impliquée dans la préparation et la sélection de mouvements complexes, particulièrement pour des mouvements générés intérieurement (génération de mouvements spontanés).

C'est cet ensemble de processus de contrôle relativement périphériques, couplé à la structure des variables de commandes articulatoires, que nous résumons par la notation M.



(a)



(b)

FIGURE 2.3 – Le cortex cérébral. (a) géographie des principaux giri et sulci. (b) localisation de quelques aires usuelles (d'après Seeley et collab. (2006)).

2.2 La transformation articulatori-acoustique

Le conduit vocal joue le rôle d'un résonateur excité par la vibration des cordes vocales, ce qui lui permet de générer des ondes sonores complexes (Figure 2.4). C'est donc la forme du conduit, imposée par la configuration des articulateurs, qui va moduler l'onde produite par les cordes vocales. Cette forme est généralement représentée dans un espace constrictif, une constriction étant une zone d'ouverture minimum dans le conduit vocal, comme illustré Figure 2.5. C'est en grande partie le lieu et le degré d'ouverture de ces constrictions qui déterminent les propriétés de l'onde sonore résultante par les lois de l'aéro-acoustique.

La représentation des sons de parole quant à elle se fait généralement dans l'espace des deux ou trois premiers formants, parfois exprimés en Barks (Bk), une unité percep-

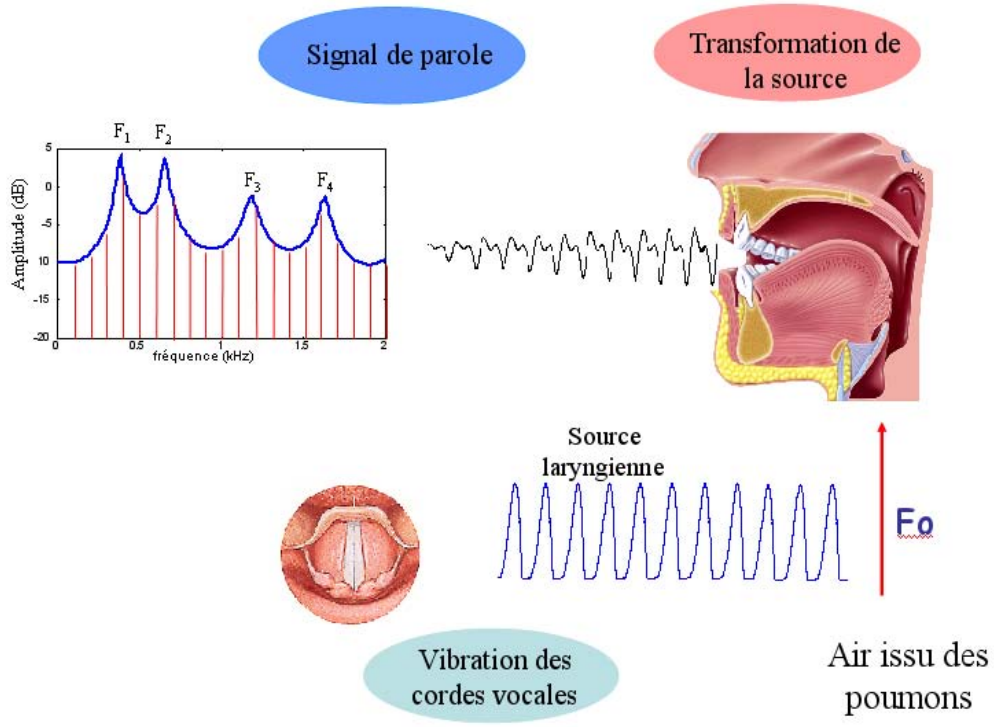


FIGURE 2.4 – La transformation articulatori-acoustique. La vibration des cordes vocales issue du flux d'air en provenance des poumons fournit la source laryngienne, une onde sonore complexe à la fréquence fondamentale F_0. Selon la forme du conduit vocal qui agit comme un résonateur, les harmoniques de la composante fondamentale de la source sont sélectivement amplifiés ou atténués (spectre de raies – lignes verticales sur le diagramme fréquence-amplitude en haut à gauche). Les maxima locaux de ce spectre s'appellent les formants du signal, numérotés du plus bas au plus haut (F1, F2, F3 et F4 sur la figure).

tive (Schroeder et collab., 1979) telle que :

$$F_{Barks} = 7 \sinh^{-1}(F_{Hz}/650).$$

Concernant les voyelles par exemple, il est commun de les représenter dans l'espace des deux premiers formants (F1-F2) qui permet une bonne catégorisation (Figure 2.6). La forme particulière de cet espace lui vaut le nom de triangle vocalique.

De par la structure du conduit vocal, la transformation des paramètres articulatoires (les sept paramètres de la Figure 2.2) aux paramètres acoustiques (les formants) est, dans certaines zones, fortement non-linéaire. La Figure 2.7 expose les espaces constrictifs et acoustiques des consonnes plosives, consonnes définies par une fermeture complète du conduit vocal (/b/ ou /d/ par exemple). L'espace constrictif correspond alors simplement au lieu de la constriction, c'est-à-dire le lieu où le conduit vocal est totalement fermé. On observe sur cette figure comment, lorsque la constriction recule de l'avant du conduit vocal (consonnes labiales comme /b/, puis dentales et alvéolaires comme /d/) vers l'arrière (consonnes vé-

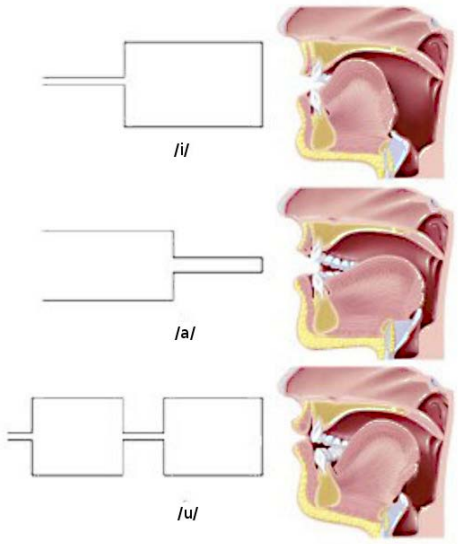


FIGURE 2.5 – Exemple de constrictions. De haut en bas : réalisations des voyelles /i/, /a/ et /u/. À droite : schéma de coupe sagittale du conduit vocal. À gauche : schéma des constrictions. Le /i/ correspond à une constriction à l'avant du conduit vocal ; le /a/ à une constriction à l'arrière, et le /u/ à une constriction au milieu du conduit et une constriction au niveau des lèvres (arrondissement).

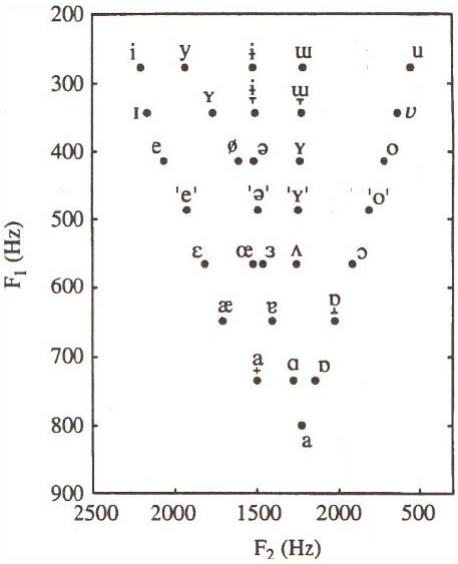


FIGURE 2.6 – Représentation des voyelles dans l'espace des deux premiers formants, F1 et F2.

laires comme /g/ puis uvulaires, pharyngales et enfin épiglottales, c'est-à-dire articulées complètement à l'arrière, vers le pharynx puis l'épiglotte), les formants varient de manière complexe dans les plans F1-F2 et F2-F3.

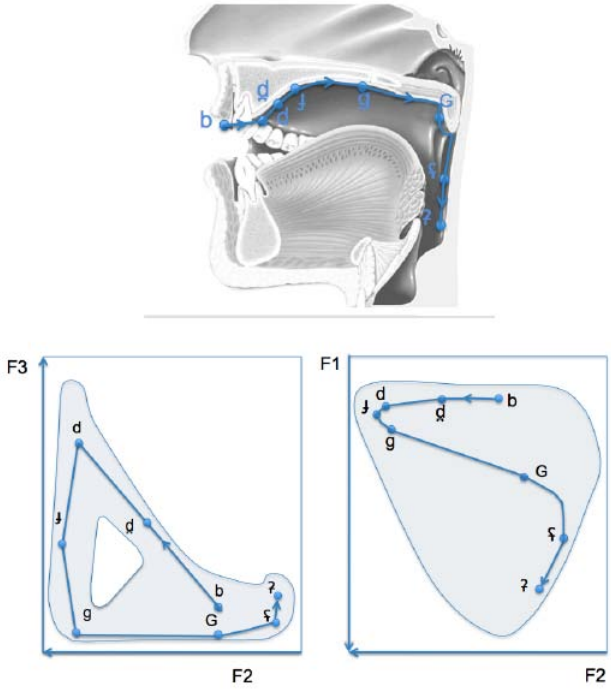


FIGURE 2.7 – Espaces constrictif (en haut) et acoustique (en bas) des consonnes plosives, d'après Schwartz et collab. (2011, en révision). Les formants sont calculés par un modèle articulatoire.

2.3 La perception des sons de parole

La première étape significative dans l'architecture neurocognitive du traitement des sons de la parole est la réalisation d'une analyse temps-fréquence du signal par la cochlée [3] (Figure 2.8). Elle contient une membrane, dite basilaire, fine à sa base et dont l'épaisseur augmente au fur et à mesure que l'on s'approche de sa partie terminale (apex). Ainsi, cette membrane est plus réactive aux hautes fréquences à sa base et plus réactive aux basses fréquences à son apex. Des fibres nerveuses sont connectées à cette membrane (par l'intermédiaire des cellules ciliées, CCx sur la Figure 2.8, qui convertissent le mouvement mécanique de la membrane en courant électrique), la cochlée a alors la propriété de réaliser

3. La perception de la parole est toutefois multimodale, faisant intervenir d'autres sens, en particulier la vision. Nous ne traiterons pas de la perception visuelle ou audiovisuelle de la parole dans cette thèse, bien que les formalismes que nous utiliserons permettraient aisément de passer d'un cadre auditif à un cadre multisensoriel.

une analyse de Fourier du signal acoustique, en permettant d'exciter plus ou moins certains neurones en fonction du spectre de fréquence du signal entrant.

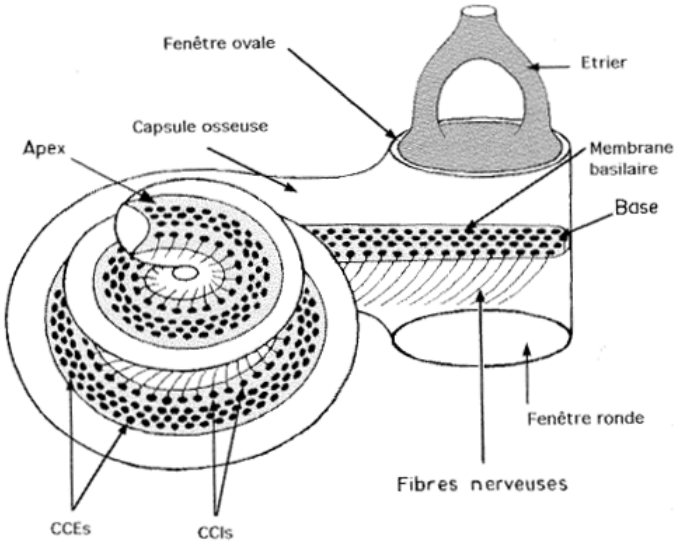


FIGURE 2.8 – Schéma de la cochlée, d'après Romand (2000). Les éléments d'intérêt décrits brièvement dans la Section 2.3 sont la *membrane basiliaire* s'étendant de sa *base* à son *apex*, les cellules ciliées (*CCx*) et les *fibres nerveuses*.

À plus haut niveau (dès le noyau cochléaire) les neurones présentent des caractéristiques de traitement spectro-temporel plus complexes (avec des champs récepteurs eux-mêmes complexes) leur conférant des propriétés de détection de caractéristiques spectrales (pics, formants) et d'événements temporels acoustiques (neurones on/off), tels que le début d'un voisement (vibration des cordes vocales) par exemple. Chistovich (1980) propose ainsi une architecture pour le traitement auditif des sons de parole basée sur deux systèmes, l'un spécialisé dans le domaine temporel (détection des événements), l'autre dans le domaine fréquentiel (détection des caractéristiques fréquentielles), schématisé Figure 2.9.

Une fois parcourus les différents relais sous-corticaux, les nerfs auditifs se projettent ensuite dans l'aire auditive primaire, située dans le lobe temporal. Elle est impliquée dans le traitement bas niveau des sons en général et présente une organisation tonotopique, prenant la forme de cartes neuronales associant fréquences sonores et positions géographiques des fibres nerveuses.

Puis on trouve dans le cortex temporal des aires auditives (et multisensorielles) intégratives multiples (cortex auditif secondaire, aire de Wernicke, et tout un réseau de traitement dans le sillon temporal supérieur et le gyrus temporal supérieur) qui ont des fonctions de plus haut niveau dans la compréhension du langage parlé. Là encore, sans proposer de description plus détaillée (voir une revue dans Kent (1997); Romand (2000)), insistons sur les deux propriétés similaires à celles que nous avons décrites pour le cortex moteur. D'une

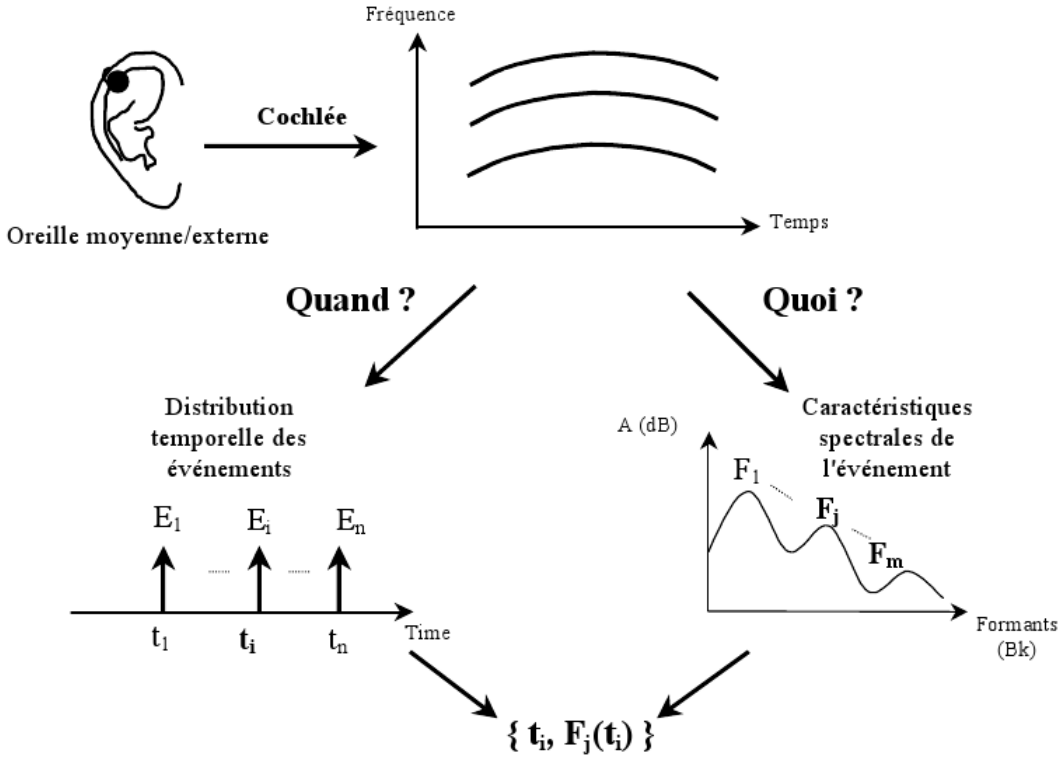


FIGURE 2.9 – Architecture pour le traitement auditif des sons de parole, d'après Serkhane (2005)

part, l'existence d'une « tête de pont » servant de point d'entrée (et non, comme pour le système moteur, de point de sortie) dans le cortex après tout le réseau sous-cortical : il s'agit de l'aire auditive primaire dont les propriétés de tonotopie (cartographie des fréquence, Romand, 2000) rappellent celles des cartes motrices de l'homonculus, décrites précédemment. D'autre part, le traitement perceptif est organisé en un réseau cortical complexe autour de principes de hiérarchie et de parallélisme.

C'est cet ensemble de processus et de représentations corticales, couplé aux processus de traitement périphérique et à la structure des variables acoustiques, qui fournit globalement le système de représentation et de traitement sensoriel que nous avons baptisé S.

2.4 Conclusion

Ce chapitre nous a permis d'introduire globalement notre cadre de recherche, et les ingrédients – très globaux et schématiques – d'une situation de communication parlée, résumés par le schéma de début de chapitre que nous rappelons Figure 2.10. Comme nous l'avons mentionné, le chemin de O_S à M, représentant ici la tâche de production du locuteur, fait intervenir des représentations motrices situées dans le lobe frontal. De même,

le chemin de S à O_L représentant ici la tâche de perception de l'auditeur, fait intervenir des représentions auditives situées dans le lobe temporal.

En détaillant les différents modèles de la littérature concernant la production et la perception de la parole, nous verrons dans le chapitre suivant que les architectures cognitives mises en jeu sont en fait bien plus complexes, et que la question du rôle des interactions entre représentations motrices et auditives dans ces deux tâches est encore aujourd'hui, peut-être plus que jamais, l'objet d'un très vif débat dans la communauté. Puis le Chapitre 4 s'intéressera à la condition C de succès de la communication, en exposant différentes théories et modèles de l'émergence des systèmes phonologiques, concernant d'un part leur optimisation pour la communication, leur amorce évolutive, et leur simulation. Nous allons donc maintenant pénétrer au sein de la ligne de communication décrite dans le présent chapitre, en explorant des situations dans lesquelles chaque agent peut prendre à la fois le rôle de locuteur et d'auditeur et active donc les deux types de représentation (motrices et sensorielles).

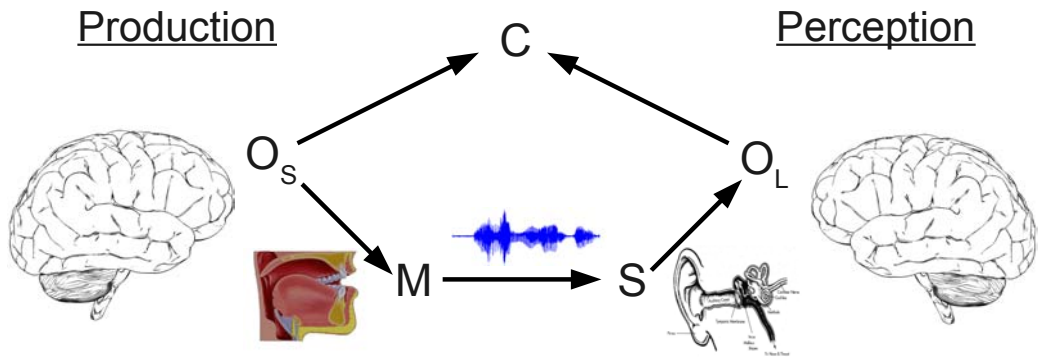


FIGURE 2.10 – Définition de notre situation de communication parlée. Le chemin de O_S à M représente la tâche de production du locuteur, celui de M à S la transformation articulatori-acoustique réalisée par l'environement, et celui de S à O_L la tâche de perception de l'auditeur. C est la condition de succès de la communication (par exemple : $O_S = O_L$)

Chapitre 3

Modèles et théories de la production et de la perception de la parole

Une question centrale dans l'étude de la communication parlée concerne la nature de l'information élémentaire échangée entre deux interlocuteurs. Nous avons vu au chapitre 2 que la nature du signal de communication est essentiellement acoustique [1]. Mais qu'en est-il de la nature de l'information encodée dans ce signal ? Les différents modèles et théories de la production et de la perception de la parole tentent de répondre à cette question, sans réel consensus. Nous proposons de les classer en trois grandes catégories, selon la nature de l'information élémentaire échangée considérée :

- les théories motrices (production et perception de gestes moteurs) ;
- les théories auditives (production et perception de stimuli auditifs) ;
- les théories sensori-motrices (le signal contient des informations à la fois motrices et sensorielles, tant en production qu'en perception).

En reprenant le schéma de situation de communication parlée exposé au chapitre précédent (Figure 2.1), la question de la nature de l'information élémentaire échangée entre deux interlocuteurs peut être posée en ces termes : l'objet de la communication (O_S ou O_L) est-il fondamentalement lié à une représentation motrice (M), sensorielle (S), ou sensori-motrice ? La Figure 3.1 illustre cette formulation en affinant les architectures d'agents de la Figure 2.1. En effet, bien que ceux-ci se trouvent généralement dans un rôle soit de locuteur soit d'auditeur à un instant donné, ils ont joué les deux rôles au cours de leur existence, et en particulier possèdent les deux types de représentation (motrices et sensorielles).

Ce chapitre expose les principaux modèles et théories issus de ces trois grands courants, à la fois dans le domaine de la production et de la perception de la parole. Puis nous nous intéressons à l'existence d'un lien sensori-moteur dans le cerveau, en nous appuyant sur les avancées récentes de neurosciences. Enfin, nous concluons sur le constat que, dans chaque courant théorique, un rôle fonctionnel fort est attribué au lien sensori-moteur et proposons une architecture d'agent récapitulative.

1. Du moins tant que nous ne considérons, comme ce sera le cas dans cette thèse, que des situations de communication acoustique pure et non pas multimodale (audiovisuelle).

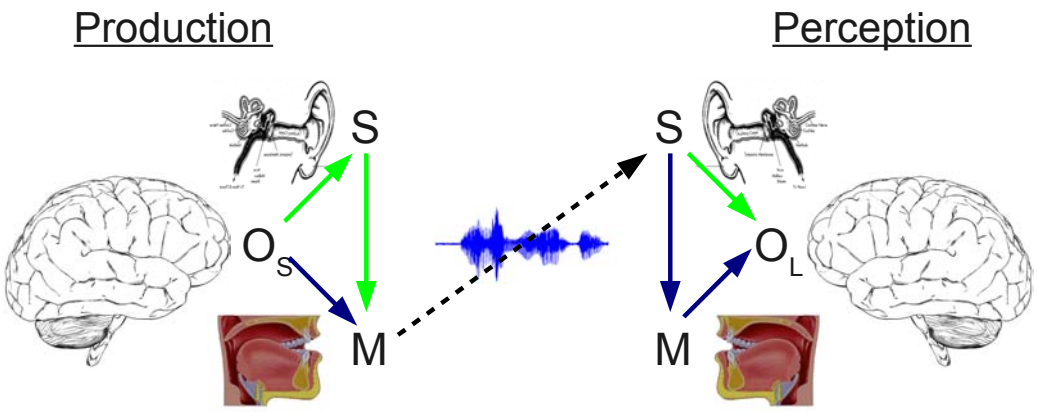


FIGURE 3.1 – Affinage des modèles d'agents en situation de communication parlée. Chacun possède à la fois des représentations motrices et auditives. Les flèches bleues (ou gris foncé) représentent les théories motrices. Dans le cas de la production (à gauche), elles ne nécessitent pas l'utilisation de connaissances sensorielles, alors que dans le cas de la perception (à droite), elles nécessitent d'estimer la représentation motrice à partir de l'entrée auditive. Les flèches vertes (ou gris clair) représentent les théories auditives. Dans le cas de la production, elles nécessitent l'existence d'un lien sensori-moteur afin de transformer une cible auditive en commandes motrices, alors qu'elles ne nécessitent pas l'utilisation de connaissances motrices en perception. Les théories sensori-motrices quant à elles considèrent que l'interaction entre les deux types de représentation, motrices et sensorielles, est au cœur de la production et de la perception de la parole.

3.1 Théories motrices

Les théories motrices considèrent que l'information élémentaire échangée entre deux interlocuteurs est de nature motrice. Certes cette information transite dans un signal essentiellement acoustique, mais qui est la conséquence d'un geste moteur. La représentation motrice de ce geste serait ici à la fois la cible de la production, ainsi qu'une connaissance nécessaire à la perception de la parole.

3.1.1 Production : la Phonologie Articulatoire

La principale théorie motrice de la production de la parole est la Phonologie Articulatoire, conçue aux laboratoires Haskins dans les années 80 (Browman et Goldstein, 1986, 1989, 1992). Elle est implémentée dans un système computationnel constitué de trois étapes, illustrées Figure 3.2 et détaillées ci-après. L'unité phonologique considérée ici est l'action articulatoire, ou geste (gesture), dont l'objectif fonctionnel est une constriction dans le conduit vocal du locuteur (par exemple le mot « bain » commence par une fermeture des lèvres). A chaque geste est associé un ensemble d'articulateurs, décrits Figure 3.3.

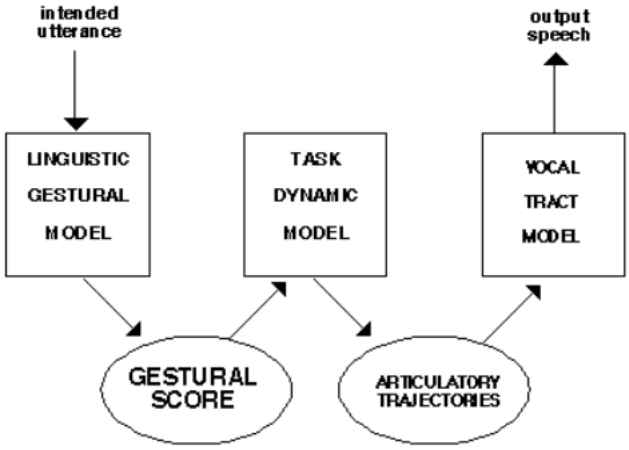


FIGURE 3.2 – Les trois étapes computationnelles de la Phonologie Articulatoire, d'après Browman et Goldstein (1992).

tract variable		articulators involved
LP	lip protrusion	upper & lower lips, jaw
LA	lip aperture	upper & lower lips, jaw
TTCL	tongue tip constrict location	tongue tip, body, jaw
TTCD	tongue tip constrict degree	tongue tip, body, jaw
TBCL	tongue body constrict location	tongue body, jaw
TBCD	tongue body constrict degree	tongue body, jaw
VEL	velic aperture	velum
GLO	glottal aperture	glottis

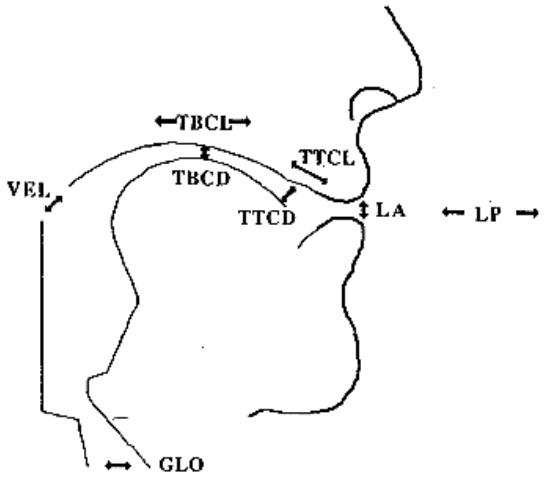


FIGURE 3.3 – Les variables de constrictions (tract variables, colonne de gauche) et leurs articulateurs associés (colonne de droite), d'après Browman et Goldstein (1989).

3.1.1.1 De l'énoncé à la partition de gestes articulatoires

L'entrée du système est un énoncé (*utterance*), qui peut être considéré simplement comme une suite de phonèmes, à partir duquel est calculée une partition de gestes (*gestural score*). Le terme partition s'interprète ici dans son sens musical, c'est-à-dire un ensemble de gestes ordonnés dans le temps, pouvant se chevaucher. Chacun de ces gestes définit un objectif articulatoire, généralement en terme de lieu et de degré de constriction, associé à des contraintes de phasage, de raideur et d'amortissement. Dans une partition, les gestes sont connectés entre eux par des contraintes de phase, comme illustré Figure 3.4.

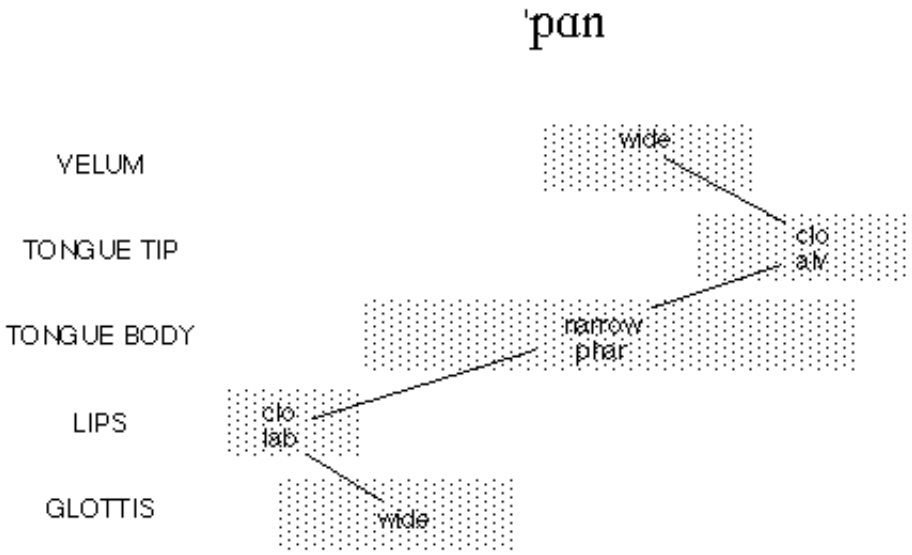


FIGURE 3.4 – La partition de gestes résultant du mot anglais « pan », d'après Browman et Goldstein (1989). Chaque rectangle représente la fenêtre temporelle d'activation du geste sur l'axe horizontal, caractérisé par des paramètres de lieu et/ou de degré de la constriction. Les lignes connectant les différents gestes représentent les contraintes de phase entre deux gestes.

3.1.1.2 De la partition de gestes aux trajectoires d'articulateurs

La partition de gestes obtenue permet alors de calculer les réponses temporelles des variables de constriction, et plus précisément la trajectoire de chacun des articulateurs du conduit vocal, en tenant compte de contraintes de phasage, de raideur et d'amortissement. La Figure 3.5 montre l'évolution temporelle des variables de constriction superposée à la partition de gestes de la Figure 3.4. Du fait de la relation « un-à-plusieurs » entre les gestes et les articulateurs, ainsi que des contraintes appliquées au système, les patrons de trajectoires ainsi obtenus peuvent être relativement complexes, expliquant ainsi certaines

propriétés de la production de la parole telles que la coarticulation ou les variations allophoniques, associées au fait que la réalisation d'un phonème dépend toujours de la réalisation des phonèmes adjacents. On observe par exemple Figure 3.5, avant le geste de fermeture du conduit par l'apex de la langue (tongue tip), que le velum est ouvert afin de préparer la consonne nasale /n/. En effet, la synchronisation parfaite du geste de fermeture de l'apex et de celui d'ouverture du vélum est en pratique impossible. Ce phénomène connu de nasalisation de la voyelle précédant une consonne nasale est ainsi expliqué comme une conséquence de la coordination de gestes articulatoires au sein d'une partition.

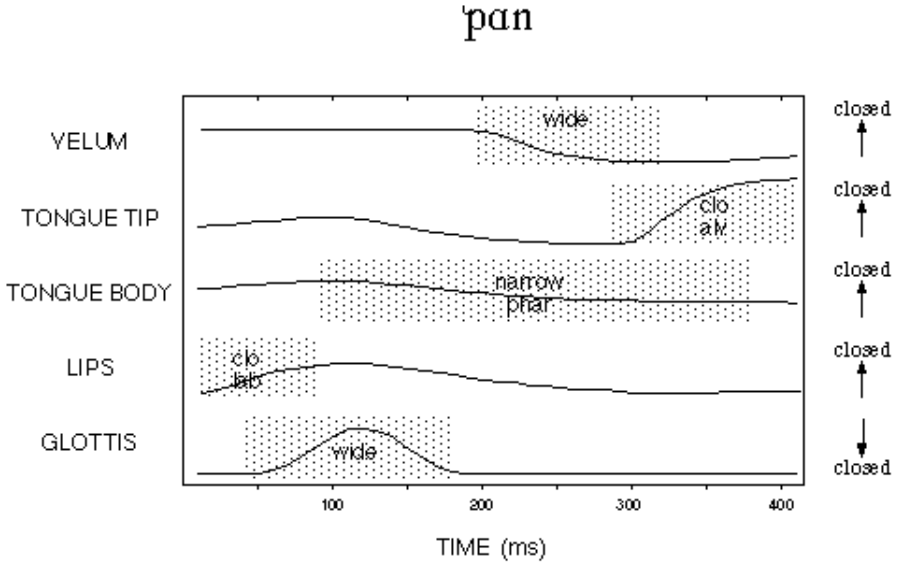


FIGURE 3.5 – Superposition de la partition de gestes résultant du mot anglais "pan" avec la réponse temporelle de certaines variables de constriction calculée par le modèle dynamique de tâches (voir Figure 3.2).

3.1.1.3 Des trajectoires d'articulateurs au signal de parole

Finalement, ces trajectoires sont soumises à un modèle de conduit vocal, qui calcule la forme du conduit vocal, la fonction d'aire et la fonction de transfert pour produire l'onde sonore qui en résulte.

3.1.2 Perception : la Théorie Motrice de la Perception

Également élaborée aux laboratoires Haskins, la Théorie Motrice de la Perception (Liberman et Mattingly, 1985) s'articule autour de trois postulats (Galantucci et collab., 2006) :

- la parole est spéciale, dans le sens où sa production et sa perception seraient intimement liées dans un module cérébral spécifique ;
- les objets de la perception de la parole sont les gestes phonétiques intentionnels du locuteur ;
- le système moteur est recruté en perception de la parole.

L'idée principale est que le lien entre phonèmes et gestes articulatoires est plus direct que celui entre phonèmes et sons. Un exemple classique concerne la réalisation du phonème /d/ dans deux contextes différents : /di/ et /du/. Le geste articulatoire produisant le /d/ est le même dans les deux réalisations : la fermeture du conduit vocal par la pointe de la langue au niveau alvéolaire. Mais par des effets de coarticulation, dus à la préparation de la voyelle suivante, les conséquences acoustiques sont différentes (voir Figure 3.6 pour une illustration).

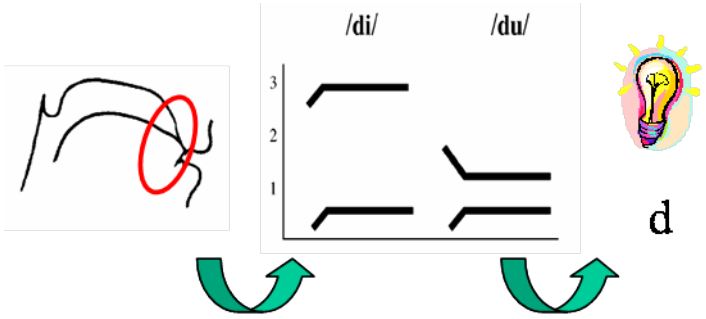


FIGURE 3.6 – Illustration de l'argument principal de la Théorie Motrice, concernant le phonème /d/ dans les syllabes /di/ vs. /du/ : le geste articulatoire est le même (fermeture du conduit vocal par la pointe de la langue), le signal acoustique est différent (2 transitions de formants montantes vs. une montante et une descendante), et le percept est le même (le phonème /d/). L'information invariante ici serait donc mieux représentée dans le geste que dans le son.

De nombreux travaux expérimentaux ont apporté à la fois soutien et critiques à la théorie. Parmi eux, on trouve les expériences de perception catégorielle, qui sont un bon exemple de la vivacité du débat entre les tenants des théories motrices vs. auditives de perception de la parole : pour les premiers, le résultat d'une expérience est interprétée comme la conséquence de la perception d'un événement moteur ; pour les seconds, on montre que ce résultat est récurrent si l'on réplique l'expérience sur des sujets non-langagiers (bébés ou animaux), qui n'ont pas de capacités de production. Nous allons décrire l'effet de perception catégorielle, puis les interprétations des expériences du point de vue des théories motrices, nous verrons dans la section suivante celles du point de vue des théories auditives.

3.1.2.1 Perception catégorielle

La perception catégorielle réfère au phénomène d'invariance de percept catégoriel lorsqu'un stimulus varie le long d'un continuum. Dans le cadre de la perception de la parole, Abramson et Lisker (1970) synthétisent des spectogrammes de syllabes /ba/ et /pa/.

Au niveau acoustique, le paramètre d'intérêt qui les différencie est le délai d'établissement du voisement (Voice Onset Time : VOT), correspondant à l'intervalle de temps entre le relâchement de l'occlusion et le début de voisement. Lors d'une séquence consonne-voyelle, un long VOT est caractéristique d'une consonne non voisée (aucun voisement lors de l'ouverture de l'occlusion), alors qu'un VOT négatif ou court sera caractéristique d'une consonne voisée (le voisement est déjà, ou quasiment présent lors de l'ouverture de l'occlusion). En faisant varier le paramètre auditif de VOT d'une valeur négative (<-50 ms) à une valeur positive (>25 ms), il est donc possible de synthétiser un continuum de stimuli auditifs de la syllabe /ba/ à la syllabe /pa/.

Ces stimuli sont alors présentés à des sujets pour identification ou pour discrimination. Deux effets importants sont remarqués : la tâche d'identification montre une frontière nette entre les deux catégories phonémiques ; la discrimination est difficile au sein d'une même catégorie, presque parfaite lorsque les deux stimuli sont de part et d'autre de la frontière.

Ces résultats peuvent être interprétés en faveur de la théorie motrice. Comme les sujets ne sont capables de produire que des /ba/ et des /pa/, il ne peuvent percevoir que l'une de ces deux catégories, même si le stimulus se trouve très proche de la frontière entre les deux. Des expériences similaires ont été réalisées le long d'un continuum allant de /ba/ à /da/, avec des résultats et une interprétation similaires : bien que le stimulus évolue le long d'un continuum, sa conséquence motrice est catégorielle (il n'existe pas de consonnes plosives pour lesquelles le lieu d'articulation est entre celui du /b/ (labial) et celui de /d/ (alvéolaire), du moins en anglais). De plus, comme cet effet de perception catégorielle n'a pas été trouvé pour certains stimuli sans liens avec la parole, il constitue également un argument en faveur de son caractère spécial.

Nous verrons à la Section 3.2.2 comment les défenseurs des théories auditives de la perception proposent une autre interprétation de cette expérience. Ce seront ces contradictions, révélant le manque d'arguments réellement décisifs en faveur de la théorie motrice, qui conduiront à une diminution de son intérêt au fil des années, intérêt toutefois récemment renouvelé par la découverte des neurones miroirs (voir 3.4.2 plus loin dans ce chapitre).

3.2 Théories auditives

A l'opposé des théories motrices, les théories auditives considèrent que l'information élémentaire échangée entre deux interlocuteurs est de nature acoustico-auditive. La représentation auditive des événements sonores serait ici à la fois la cible de la production, ainsi qu'une connaissance suffisante à la perception de la parole.

3.2.1 Production : le référentiel auditif de Guenther et collab. (1998)

Selon Guenther et collab. (1998), le référentiel utilisé pour la planification de la production de la parole est auditif (à l'inverse de la Phonologie Articulatoire vue en 3.1.1 qui propose un référentiel articulatoire, de séquences de constrictions). Quatre arguments

principaux sont proposés pour étayer cette hypothèse, les trois derniers gravitant autour du même argument principal : la capacité d'équivalence motrice, c'est-à-dire la capacité de produire le même résultat perceptif par des gestes différents. Celle-ci peut être utilisée dans la communication « en ligne » (3.2.1.2), perturbée (3.2.1.3), ou par choix du locuteur (3.2.1.4).

3.2.1.1 L'état de la constriction n'est généralement pas disponible pour le système nerveux central

L'argument ici concerne d'une part le fait que la relation entre la position des muscles contrôlant le conduit vocal et l'état de la constriction (lieu et mode) est très dépendante de la morphologie du locuteur et doit donc être apprise au cours de son développement ; d'autre part que cet apprentissage est rendu très difficile par le manque d'un signal de feedback approprié (le retour tactile n'étant généralement pas suffisant pour déterminer d'une manière unique le lieu et le mode de la constriction, au moins pour les configurations ouvertes de type voyelles).

3.2.1.2 L'invariance du lieu et de la taille de la constriction peut apparaître dans des contrôleurs ne l'utilisant pas comme cible explicite

Une cible auditive permet une plus grande flexibilité dans la sélection de la forme finale du conduit vocal que ne le ferait une cible constrictive. En effet, bien qu'une même constriction puisse correspondre à différentes configurations des articulateurs, une même cible auditive peut elle-même correspondre à différentes constrictions (le cas de la voyelle /u/ par exemple).

Cet avantage de flexibilité du choix de la forme du conduit vocal n'implique pas forcément une plus grande variabilité en production d'une parole non contrainte, si l'on considère que les articulateurs préfèrent certaines configurations plus confortables ou économiques.

3.2.1.3 Des cibles constrictives limiteraient inutilement les capacités d'équivalence motrice du système de production

Les expériences de Savariaux et collab. (1999) montrent comment des locuteurs tendent à réorganiser la configuration de leur conduit vocal pour produire la voyelle /u/ lorsque ce dernier est contraint par un tube rigide placé entre les lèvres. Ne pouvant effectuer correctement le geste d'arrondissement des lèvres, ils ont alors tendance à complètement réorganiser leur conduit vocal. Cette réorganisation motrice a pour effet de modifier radicalement le lieu de la constriction, d'une position vélaire pour le /u/ normal à une position pharyngale pour le /u/ contraint par le tube, afin d'atteindre la même région auditive.

Ces résultats sont en contradiction avec l'hypothèse d'une cible constrictive en production de la parole, qui prédirait alors une réorganisation motrice tendant à se rapprocher le mieux possible de la constriction originale, en dépit d'un changement important de la conséquence auditive.

3.2.1.4 La seule cible invariante pour le /r/ américain semble être une cible de nature acoustique ou auditive

Le /r/ américain peut être produit de deux façons différentes selon le contexte où il apparaît : soit avec le dos de la langue en position palatale, soit avec l'apex de la langue en position alvéolaire. Ces deux configurations ont la particularité de produire des configurations acoustiques très similaires.

Un même locuteur américain aura tendance à produire un /r/ dont la configuration motrice nécessite le moins de mouvement possible par rapport à la configuration imposée par le contexte vocalique. Ces deux /r/ appartiennent pourtant à la même catégorie phonémique, l'invariant ici semble donc être bien plus auditif que moteur.

3.2.2 Perception

Les théories auditives de la perception considèrent la perception de la parole comme un problème classique de traitement du signal, dans lequel il n'y aurait pas de difficulté majeure, et en tout cas pas d'impossibilité théorique, à extraire les invariants de la parole directement à partir du flux sonore, sans nécessité d'accès à des connaissances motrices (Diehl et collab., 2004).

3.2.2.1 Les données de l'aphasie

Les données provenant de patients aphasiques montrent une double dissociation entre les tâches de discrimination phonémique et celles de compréhension de la parole (Hickok et Poeppel, 2004) : il existe des patients avec un déficit en discrimination phonémique mais qui gardent un bon niveau de compréhension des mots, et vice-versa. Or, bien que des patients aphasiques de Broca (dommage du centre moteur de la parole) peuvent manifester un déficit dans des tâches de discrimination de phonèmes, leurs capacités en compréhension des mots n'est généralement pas affectée (Moineau et collab., 2005)[2]. Bien qu'activées lors de tâches de compréhension de la parole, les aires motrices ne semblent donc avoir qu'un rôle fonctionnel limité.

3.2.2.2 Retour sur la perception catégorielle

Nous avons vu à la Section 3.1.2.1 comment des syllabes synthétiques variant de façon continue le long d'un paramètre auditif sont catégorisées de façon nette par des sujets, même lorsque le stimulus se trouve proche de la frontière entre deux catégories. Ce phénomène de perception catégorielle a servi d'argument aux défenseurs des théories motrices : le continuum auditif n'a pas forcément de corrélat moteur de même nature (par exemple, il n'existe pas en anglais un continuum de lieu d'articulation possible entre une plosive bilabiale /b/ et une plosive alvéolaire /d/).

2. Cet argument est toutefois controversé, voir Pulvermüller et Fadiga (2010), p. 354.

Mais lorsque Kuhl et Padden (1982) entraînent des singes chinchillas à répondre différemment à des exemplaires de syllabes synthétisées (/da/ vs. /ta/, différenciées par leur VOT), ceux-ci exhibent finalement le même phénomène de perception catégorielle avec des performances d'identification très proches de celles des sujets humains. Ce phénomène ne semble donc ni découler d'une perception des gestes moteurs (les chinchillas n'ont pas de notions de production), encore moins d'un module spécialisé pour la parole, mais bien de principes opératoires de la perception plus généraux. Diehl et collab. (2004) remarquent qu'un des défis majeurs pour la théorie motrice sera de mettre en évidence des phénomènes de perception spécifiques à la parole, en particulier absents chez les animaux ou les enfants prélangagiers. Toutes leurs recherches depuis plus de vingt ans visent à montrer de façon souvent convaincante comment les problèmes de variabilité contextuelle complexe peuvent être résolus par des mécanismes de traitement auditif classiques.

3.3 Théories sensori-motrices

Les théories sensori-motrices de la parole tentent d'intégrer les différents arguments des théories motrices et auditives de la communication dans des modèles de production et de perception où le couplage sensori-moteur tient un rôle central dans la co-construction des représentations motrices et sensorielles.

3.3.1 Production : le modèle DIVA de Guenther (2006)

Le modèle DIVA (Directions Into Velocities of Articulators) de Guenther (2006) propose une architecture cognitive de production de la parole mettant en jeu deux sous-systèmes :

Un sous-système feedforward qui associe les unités de production (phonème, syllabe ou mot) aux commandes motrices (positions et vitesses des articulateurs) ;

Un sous-système feedback qui associe des cibles auditives et somato-sensorielles aux unités de production ainsi qu'aux commandes motrices via un retour sensoriel.

Ces associations sont implémentées dans des réseaux de neurones organisés en cartes reliées entre elles par des projections synaptiques. L'architecture globale est décrite Figure 3.7.

Avant d'être opérationnel, le modèle requiert deux phases d'apprentissage. La première s'apparente à la période de babillage où des mouvements semi-aléatoires des articulateurs servent à produire les retours auditifs et somato-sensoriels correspondants, afin d'apprendre les associations entre les cartes d'erreurs (*Auditory Error Map, Somatosensory Error Map*) et les commandes motrices. Ces associations ne sont pas spécifiques à une unité de production particulière. La seconde phase d'apprentissage s'apparente à l'exposition d'un enfant à sa langue native. On présente alors au modèle des échantillons de sons de parole variant dans le temps et labellisés, utilisés pour apprendre les régions auditives cibles pour chaque unité de production.

Suite à ces deux phases d'apprentissage, le modèle est capable de reproduire des sons de parole par le sous-système feedback. Au fur et à mesure des ces productions, le sous-système

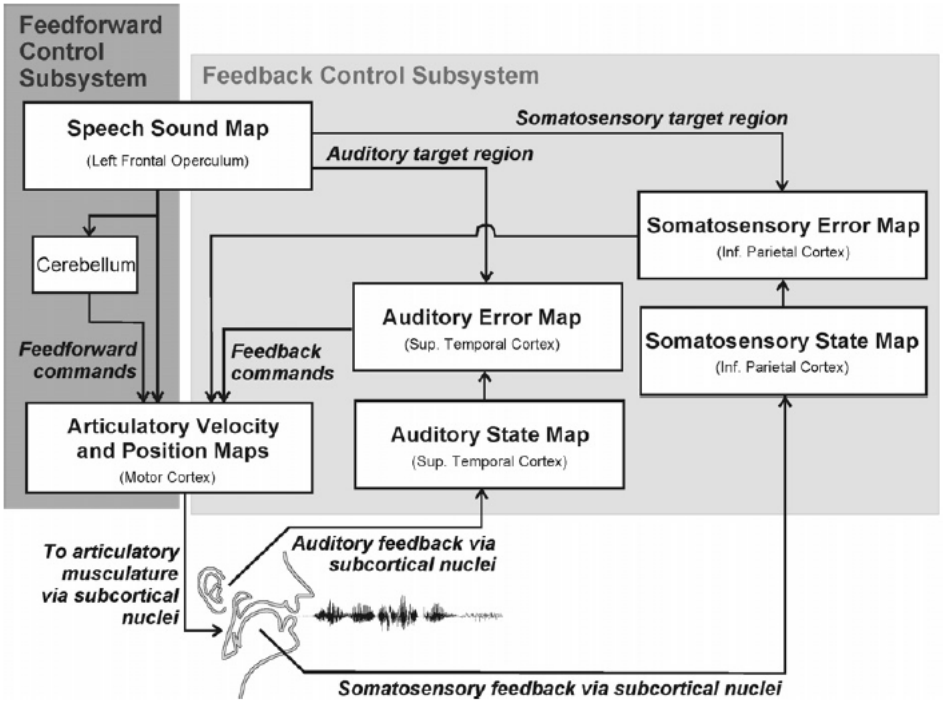


FIGURE 3.7 – Architecture du modèle DIVA. Les rectangles représentent des cartes neurales, les flèches les connections synaptiques entre elles. L'entrée du système correspond à l'activation d'une cellule de la carte des sons (Speech Sound Map), représentant une unité de production (généralement une syllabe). Le conduit vocal est alors contrôlé par deux sous-systèmes. Le sous-système feedforward associe directement les unités de production aux commandes motrices (*feedforward commands*). Le sous-système feedback joue un rôle de correction d'erreurs, d'une part en définissant des régions cibles auditives et somato-sensorielles pour chaque unité de production (*target region*) ; d'autre part en corrigeant les commandes motrices selon les retours auditifs et somato-sensoriels (*feedback commands*).

feedforward apprend les correspondances directes entre unités de production et commandes motrices. Lorsqu'il a acquis suffisamment d'expérience, le sous-système feedforward génère alors peu d'erreurs auditives et peut donc se passer du sous-système feedback, qui n'interviendra plus qu'en cas de perturbations. Au cours de ses productions le modèle apprend également, grâce au retour correspondant, les régions cibles somato-sensorielles qui auront un rôle de corrections d'erreurs similaire à celui des régions cibles auditives.

Ce modèle de production très complet présente un double avantage :

- chacune des unités fonctionnelles du modèle est associée à une localisation cérébrale présumée (Figure 3.7) permettant de comparer le modèle aux données d'imagerie cérébrale (Guenther, 2006) ;
- son architecture à deux sous-systèmes, feedforward et feedback, lui permet de rendre compte d'un grande variété de phénomènes liés à la production de la parole tels que l'équivalence motrice ou la coarticulation (Guenther, 1995).

3.3.2 La théorie de la perception pour le contrôle de l'action

La théorie de la perception pour le contrôle de l'action (PACT : Perception for Action Control Theory (Schwartz et collab., 2002; Schwartz et collab., 2007; Schwartz et collab., 2010), ou Perception-Action Coordination Theory) est centrée sur la co-structuration des systèmes de perception et d'action en relation avec la phonologie. Elle se démarque des positions radicales des théories motrices et auditives en proposant comme objets de la communication parlée :

- des percepts multimodaux régularisés par des connaissances motrices, ou
- des gestes mis en forme par des traitements multisensoriels.

La PACT propose un cadre intégrateur des théories motrices et auditives de la perception, en insistant sur le rôle important de l'influence mutuelle entre perception et action.

3.3.2.1 Des percepts multimodaux régularisés par des connaissances motrices

La PACT ne rejète pas les arguments des théories motrices concernant l'accès au système moteur lors de la perception de la parole, en particulier dans la régularisation des percepts influencés par des phénomènes de coarticulation. Dans le cas de la perception des consonnes plosives par exemple, il semble difficile de ne pas invoquer une spécification articulatoire du lieu d'articulation, tout en remarquant qu'une spécification auditive existe bel et bien si l'on considère non plus la perception des voyelles et des consonnes séparément, mais directement la perception des syllabes. En effet, les trajectoires définies par les syllabes dans l'espace auditif sont facilement séparables malgré les phénomènes de coarticulation. Toutefois, la PACT insiste sur le rôle de la costructuration des représentations motrices et perceptives lors du développement. Avant qu'il ne commence à les produire, le bébé humain est déjà exposé à des stimuli de syllabes qui ne présentent pas de difficultés particulières à être catégorisés dans l'espace auditif si on les considère dans leur ensemble (consonnes + voyelles). Ce ne sera que lors de son apprentissage de la production, qu'il pourra remarquer que certains de ces stimuli ont un invariant moteur en commun (la fermeture des lèvres pour la consonne /b/ par exemple). Le rôle central du lien perceptuo-moteur dans la PACT trouverait son corrélat neuroanatomique dans l'existence de la voie dorsale et des neurones miroirs (Schwartz et collab., 2010), sur lesquels nous reviendrons dans ce chapitre.

3.3.2.2 Des gestes mis en forme par des traitements multisensoriels

Tout en reconnaissant un possible rôle des connaissances motrices dans la régularisation des percepts, la PACT remarque toutefois que les systèmes phonologiques sont optimisés dans l'espace auditif. Il s'agit de la théorie de la dispersion (Liljencrants et Lindblom, 1972) que nous détaillerons à la Section 4.1.1 mais dont l'idée principale est simple : les régularités observées dans les systèmes phonologiques des langues du monde suggèrent que ceux-ci sont optimisés de façon à disperser autant que possible les unités phonémiques dans l'espace acoustique. Ainsi, les trois voyelles utilisées dans presque toutes les langues du monde correspondent aux trois sommets du triangle vocalique (voir Section 2.3), car elles maximisent la facilité à les distinguer par une oreille humaine.

Or, une comparaison rapide des espaces auditifs et articulatoires des voyelles montre bien que cette optimisation n'est effective que dans le premier, pas dans le second, comme illustré Figure 3.8.

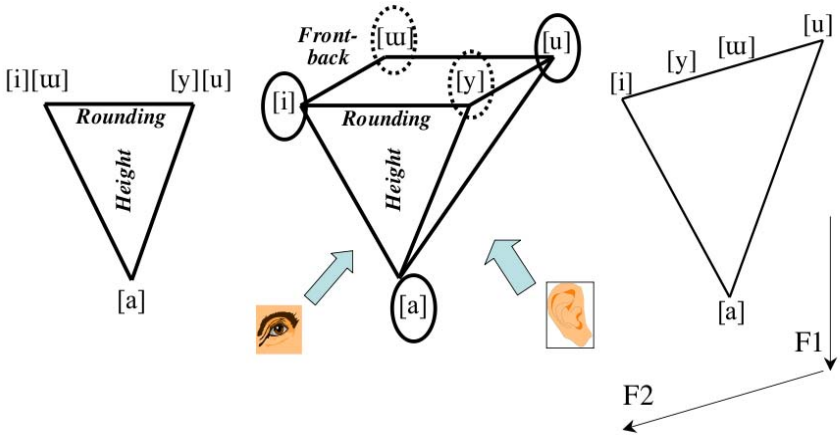


FIGURE 3.8 – Espaces simplifiés des voyelles. L'espace articulatoire (au milieu) comporte trois dimensions : hauteur (height), avant-arrière (front-back), et arrondissement des lèvres (rounding). L'espace visuel (à gauche) en comporte deux : hauteur et arrondissement, la dimension avant-arrière étant difficilement perceptible dans cette modalité. Enfin, l'espace auditif (à droite) comporte deux dimensions, les deux premiers formants. On remarque que deux systèmes sont optimaux en terme de dispersion dans les espaces articulatoire et visuel : [i a u] et [y a ɯ]. Or, les observations dans les langues du monde montrent que, alors que [i a u] est le système privilégié utilisé dans quasiment chacune d'entre elles, [y a ɯ] est une combinaison rare et n'existe jamais seule (aucune langue n'utilise le [y] sans le [ɯ]) (Maddieson, 1984). La raison à cela semble bien être auditive : [i a u] est le meilleur système en terme de dispersion dans l'espace des deux premiers formants, alors que [y a ɯ] est bien plus mauvais (partie droite de la figure).

Ainsi il semble bien que les choix phonémiques des langues du monde soient en partie basés sur une optimisation de l'espace auditif. C'est ce qui conduit la PACT à considérer les gestes comme des unités perceptuo-motrices et non purement motrices, unités « mise en forme » par les traitements multisensoriels qui permettent de leur attribuer une valeur fonctionnelle perceptive.

3.3.2.3 Architecture fonctionnelle

Ces considérations ont mené Schwartz et collab. (2010) à proposer l'architecture fonctionnelle de perception de la parole exposée Figure 3.9, impliquant d'une part une caractérisation et une catégorisation auditives, et d'autre part un double lien perceptuo-moteur, à la fois pour la co-construction des prototypes de catégorisation et pour la structuration des propriétés de la caractérisation.

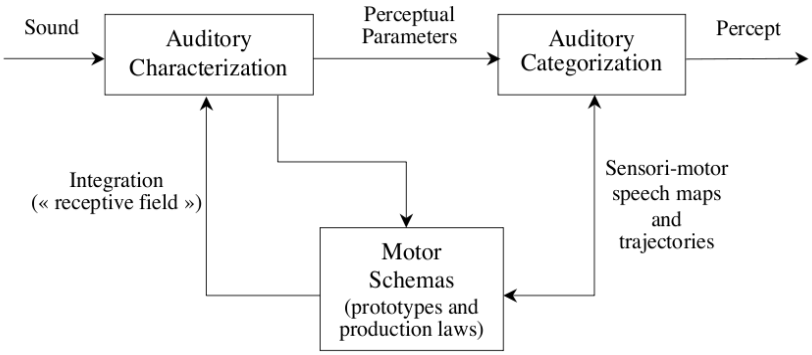


FIGURE 3.9 – Architecture de la PACT pour la perception de la parole, d'après Schwartz et collab. (2010). Le lien sensori-moteur (*Sensori-motor speech maps and trajectories*) permet une co-structuration des représentations motrices et perceptives. La connexion des représentations motrices vers les représentations auditives (*Integration*) pourrait avoir un rôle fonctionnel pour la perception dans certaines situations, par exemple en structurant le flux en environnement bruité.

Notons également l'existence du modèle de Skipper et collab. (2007), qui propose une fonctionnalité un peu différente du lien sensori-moteur en perception. Il propose une boucle sensori-motrice dans laquelle des stimuli multi-sensoriels peuvent être interprétés en termes de commandes motrices qui permettent en retour, par un processus de copie d'efférence, d'affirmer ou d'infirmer des hypothèses sur le percept final. Ce modèle tente d'expliquer certaines illusions perceptives lorsque les stimuli auditifs et visuels ne sont pas congruents, telles que l'effet McGurk (McGurk et MacDonald, 1976).

3.4 Apport des neurosciences

Le débat sur les théories motrices vs. auditives vs. sensori-motrices de la communication parlée a été récemment enrichi par les progrès des neurosciences, notamment avec les avancées extraordinaires des techniques de neuro-imagerie. Globalement, on est ainsi passé d'un localisationnisme un peu réducteur, cantonnant la production de la parole aux lobes frontal et pariétal gérant les représentations motrices, et la perception de la parole au lobe temporal gérant les représentations auditives (voir Chapitre 2), à différents niveaux de complexité, à un connexionnisme peut-être un peu échevelé insistant sur les coactivations frontales, pariétales et temporales lors de tâches de production et de perception de la parole. La neuroimagerie par résonance magnétique fonctionnelle (fMRI) a ainsi démontré largement que les tâches de production, même mentales (« covert speech ») activaient un réseau d'aires temporales associées à la perception (Guenther, 2006; Bohland et Guenther, 2006; Tourville et collab., 2008)), et en retour, que la perception de la parole impliquait des aires pariétales et frontales classiquement associées à la production (Wilson et collab., 2004; Pulvermüller et collab., 2006; Skipper et collab., 2007). On voit, en conséquence,

apparaître des travaux visant à mieux connaître les réseaux communs de la perception et de l'action (Grabski et collab., 2010). Nous allons centrer notre description – volontairement réduite, pour en rester le plus possible aux principes et architectures fonctionnelles – sur quelques grands cadres théoriques qui aident à réfléchir à la manière dont sont reliées aires perceptives et motrices dans le cortex humain.

3.4.1 Modèle interne et copie d'efférence

La première proposition théorique articulant fonctionnellement perception et motricité dans le cerveau humain est la notion de copie d'efférence. Ce concept nous vient des recherches sur la vision dans la première moitié du XX^{e} siècle, avec la question de la stabilisation de la perception visuelle en dépit des saccades oculaires qui devraient, plusieurs dizaines de fois par secondes, modifier l'entrée rétinienne et donc nous exposer à des perceptions constamment changeantes. La solution proposée est celle de la décharge corollaire (Holst, 1954) dans laquelle la commande motrice envoie à un « modèle interne », une copie d'efférence qui permet de prédire la conséquence sensorielle. Cette prédiction peut être alors envoyée, sous forme de « décharge corollaire », au système sensoriel pour qu'il corrige d'autant l'entrée et estime un percept corrigé des mouvements induits par les commandes propres du système (Figure 3.10). La notion de modèle interne, permettant d'estimer les variables sensorielles S à partir des commandes motrices M, joue un rôle central dans toute la littérature moderne sur le contrôle moteur (voir par exemple Wolpert et collab. (1995); Wolpert et Kawato (1998)) et sur la capacité qu'a le sujet humain de séparer les conséquences sensorielles de ses propres actions, de la perception de celles de ses congénères (Frith, 1995).

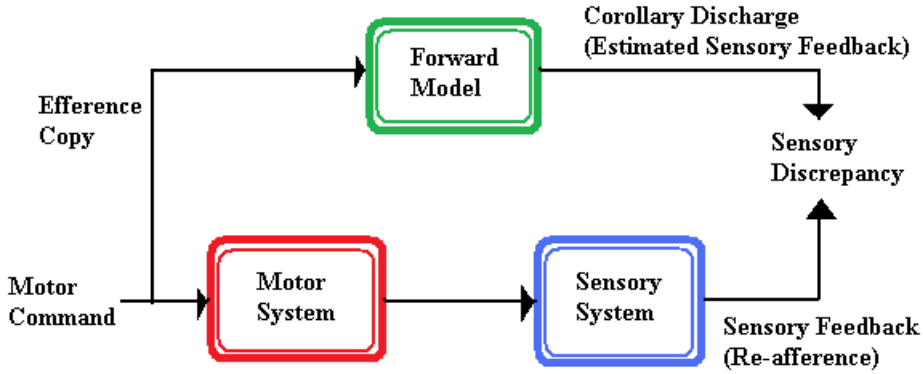


FIGURE 3.10 – La copie d'efférence des commandes motrices est utilisée en entrée d'un modèle direct capable de prédire leurs conséquences sensorielles. Celles-ci peuvent alors être comparées aux conséquences réelles, et le résultat peut éventuellement être utilisé pour corriger les commandes.

3.4.2 Les neurones miroirs

La découverte des neurones miroirs (Rizzolatti et collab., 1996) dans l'aire F5 du cortex moteur du singe est considérée comme majeure dans le domaine des neurosciences. L'aire F5 contient des neurones associés aux mouvements de la main et de la bouche. Les neurones miroirs déchargent à la fois lorsque l'animal exécute une action transitive vers un objet (saisir une pomme par exemple), et lorsque qu'il observe un congénère réaliser la même action. Ils sont à différencier des neurones dits canoniques de la même aire qui présentent également une congruence forte entre leurs propriétés motrices (concernant par exemple le type de prise codée) et leur sélectivité visuelle (concernant par exemple la taille de l'objet). En effet, contrairement à ces derniers les neurones miroirs ne déchargent pas lorsque l'objet seul est perçu, ou lorsque l'action est mimée en l'absence de l'objet. De plus, une grande partie des neurones miroirs sont spécifiques à un type d'action particulier (saisir ou placer un objet, par exemple). Un exemple de neurone miroir est présenté Figure 3.11.

Ainsi, ces neurones miroirs fournissent en quelque sorte un « modèle inverse » parcourant à rebours la boucle de simulation du modèle interne ($M \rightarrow S$) décrite dans la section précédente, pour estimer à partir d'une entrée sensorielle S la commande motrice M susceptible d'en avoir été la cause.

Kohler et collab. (2002) reportent également une classe de neurones miroirs « audio-visuels » dans l'aire F5 du singe qui, en plus de décharger lors de l'exécution ou l'observation d'actions transitives, déchargent également à l'écoute du son produit par l'action (casser une cacahuète ou déchirer une feuille de papier, par exemple).

Il semble qu'un système miroir existe chez les humains. Ainsi, l'observation d'actions effectuées par le bras ou la main d'un agent humain entraîne une excitabilité accrue des neurones du cortex moteur primaire (Fadiga et collab., 1995) et des activations corticales dans le cortex prémoteur ventral et dans le gyrus frontal inférieur (Grezes, 1998; Iacoboni et collab., 1999; Grèzes et collab., 2003). Fadiga et collab. (2002) et Watkins et collab. (2003) ont ainsi montré dans des études de Stimulation Magnétique Transcranienne (TMS) que l'audition ou la vision de stimuli de parole préactivaient les neurones du cortex moteur respectivement associés aux commandes de la langue et des lèvres lorsque les stimuli étaient associés aux mêmes actions implicites.

Ces considérations ont amené Rizzolatti et Arbib (1998) à suggérer que les neurones miroirs représentent le lien entre locuteur et auditeur que Liberman postulait dans sa théorie motrice de la perception (décrite dans ce chapitre en 3.1.2). Cette proposition est appuyée par les similitudes observées entre l'aire F5 du primate (où ont été découverts les neurones miroirs) et l'aire de Broca chez l'humain (connue, nous l'avons vu, pour être un centre du langage et particulièrement de la production de la parole), tant au niveau de leur localisation que de leur cytoarchitectonie. Nous verrons au Chapitre 4 comment ces travaux ont permis de proposer un scénario d'émergence du langage à partir de la communication gestuelle (Arbib, 2005a).

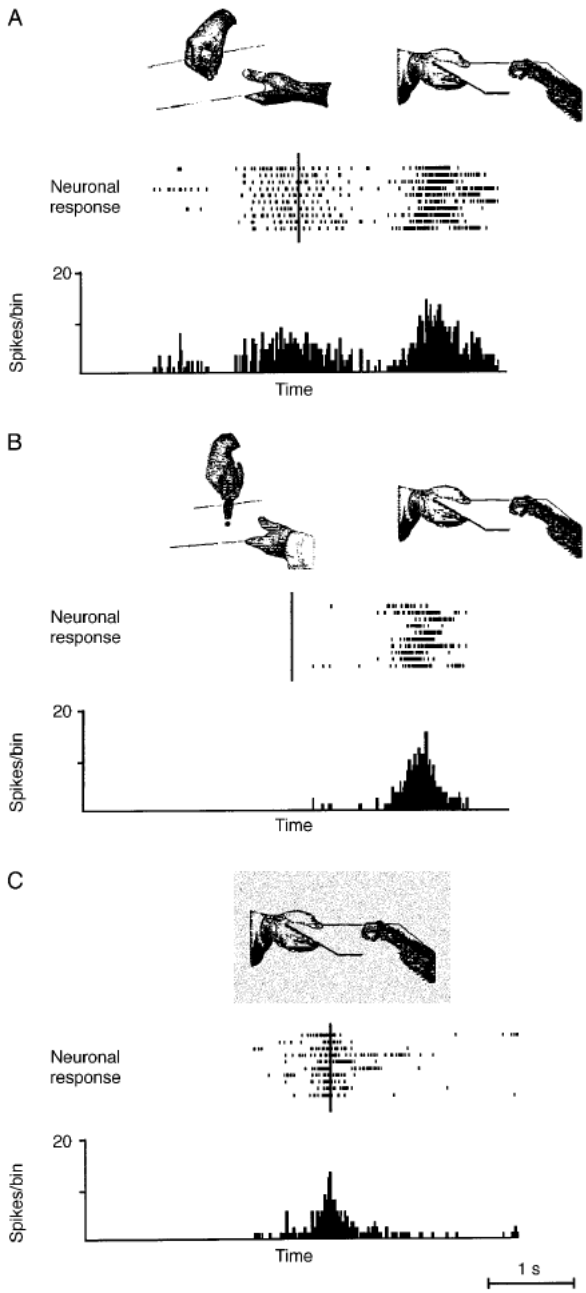


FIGURE 3.11 – Un exemple de neurone miroir chez le singe, d'après Rizzolatti et Arbib (1998). Dans chaque cas (A, B, ou C), la situation est illustrée schématiquement (en haut), et la décharge du neurone est affichée sur 10 essais différents (au milieu) ainsi que sous forme d'histogramme (en bas). En A, l'expérimentateur saisit un morceau de nourriture avec la main, le déplace vers le singe, qui le saisit à son tour. Le neurone décharge pendant l'observation de la prise, ne décharge pas pendant que la nourriture est déplacée, et décharge quand le singe saisit la nourriture. En B, l'expérimentateur saisit le morceau de nourriture avec un outil, puis la suite des événements est la même qu'en A. Le neurone ne décharge pas lorsque la nourriture est saisie avec un outil. En C, le singe saisit la nourriture dans l'obscurité. En A et B, la ligne verticale représente l'instant où la nourriture est saisie par l'expérimentateur, en C le début du geste de saisie.

3.4.3 Voie dorsale et voie ventrale

Ce sont d'abord des travaux sur la vision qui ont permis de distinguer deux voies perceptives distinctes : la voie ventrale et la voie dorsale. La première serait impliquée dans l'accès à la sémantique de l'objet perçu (voie du « quoi ») et impliquerait un circuit du lobe occipital (centre de la vision) au lobe temporal. La seconde serait impliquée dans la localisation des objets (voie du « où ») ainsi que dans l'accès à leurs affordances (les actions qu'il est possible de faire avec, voie du « comment »), et impliquerait un circuit du lobe occipital aux lobes pariétal puis frontal.

Cette dichotomie entre voie ventrale et dorsale a été reprise pour le langage par Hickok et Poeppel (2004, 2007, voir Figure 3.12), en proposant de distinguer une « voie ventrale » impliquée dans l'accès au sens des mots à travers le cortex temporal, et une « voie dorsale » qui aurait une fonction d'association entre sons et gestes articulatoires dans un circuit temporo-pariéto-frontal. Ce modèle à deux voies se justifie par l'existence d'une double dissociation entre les tâches de perception de la parole au niveau sous-lexical (phonèmes) et les tâches de reconnaissance au niveau lexical (sémantique), c'est-à-dire qu'il existe des patients dont la compréhension des mots est normale, mais qui présentent un déficit lors de tâches de discrimination de syllabes, et vice-versa.

Certains tenants des théories auditives (Hickok et Poeppel, 2007) proposent alors que les tâches de reconnaissance de la parole en conditions écologiques ne nécessitent que l'usage de la voie ventrale permettant un accès direct, bien que probablement hiérarchique et parallèle, au sens des mots (voie du « quoi »).

La voie dorsale pourrait quant à elle jouer un rôle fondamental dans l'acquisition de la parole, permettant d'associer sons et gestes articulatoire. Le stimulus d'apprentissage étant essentiellement auditif et visuel, il semble en effet difficile d'apprendre à produire la parole environnante sans apprentissage préalable des relations perception-action. D'autres rôles possibles sont proposés par les théories sensori-motrices de la perception, notamment une aide à la perception en situation perturbée par un accès aux connaissances motrices (Schwartz et collab., 2010), éventuellement au sein d'une boucle sensori-motrice (Skipper et collab., 2007).

Globalement, on le voit, l'architecture « ventral-dorsal » telle que proposée par Hickok et Poeppel (2004) est à la fois compatible avec la notion de neurone miroir (correspondant dans ses grandes lignes au contenu de la voie dorsale) mais aussi très différente, en ce qu'elle cantonne la voie dorsale à des tâches d'apprentissage des relations sensori-motrices, et en récuse l'aspect fonctionnel dans la perception en ligne, du moins en situation non perturbée. L'activation de régions motrices dans la perception en ligne ne serait alors que la conséquence d'un lien organique, au sein de la voie dorsale, mais non nécessairement fonctionnel. Ainsi, ici comme ailleurs, le débat entre théories auditives et motrices n'est, on le voit, pas clos.

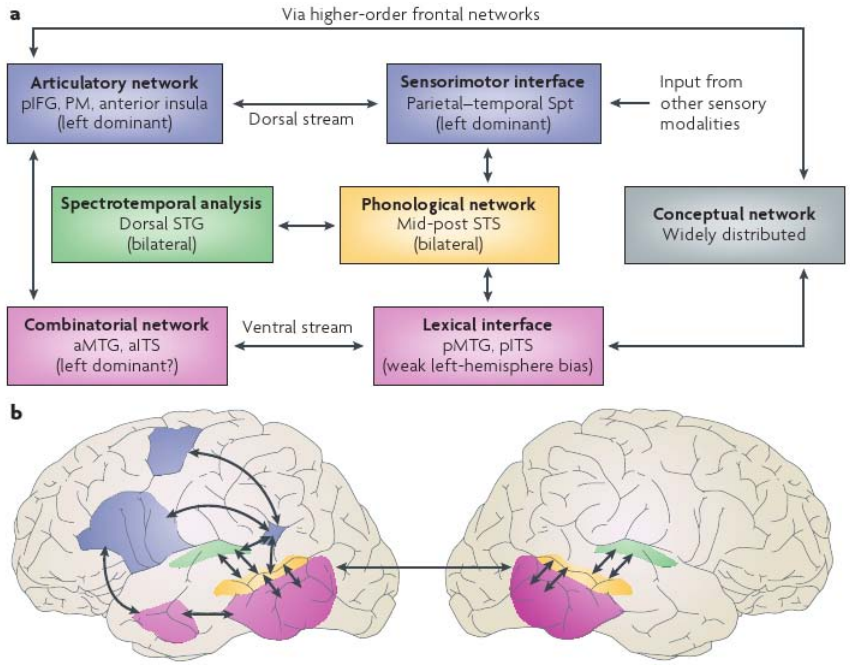


FIGURE 3.12 – Le modèle à deux voies d'anatomie fonctionnelle du langage, d'après Hickok et Poeppel (2007). La première étape est une analyse spectrale de l'entrée auditive dans le cortex auditif (en vert), suivie d'un traitement phonologique (en orange). Ensuite le flux diverge en deux grandes voies : la voie dorsale (en bleu) qui associe les représentations sensorielles ou phonologiques aux représentations motrices à travers un circuit temporo-pariéto-frontal, et la voie ventrale (en violet) qui associe les représentations sensorielles ou phonologiques aux représentations lexicales dans un circuit purement temporal.

3.5 Discussion

Nous allons, pour conclure, reprendre les grandes lignes des différentes théories que nous avons passées en revue, en essayant d'en extraire quelques grands principes ou quelques éléments de structuration fonctionnelle. Pour cela, nous commençons par définir quelques termes.

La nature d'une représentation peut être soit motrice, soit auditive, soit sensori-motrice.

Un système de production a pour but de fournir un geste moteur à partir d'un objet de communication, en se basant sur une représentation de nature motrice, auditive ou sensori-motrice. Toutefois, il passe forcément par une représentation de nature motrice pour produire sa sortie : un geste articulatoire. Il s'agit d'un chemin de O_S à M sur la Figure 3.1.

Un système de perception a pour but de reconstituer un objet de communication à partir d'un stimulus essentiellement auditif, en se basant sur une représentation de nature motrice, auditive ou sensori-motrice. Toutefois, il passe forcément par une représentation de nature auditive pour traiter son entrée : un stimulus auditif. Il

s'agit d'un chemin de S à O_L sur la Figure 3.1.

Un système est dit homogène si la nature des informations qu'il traite est soit exclusivement motrice, soit exclusivement auditive. Les systèmes homogènes sont donc les systèmes de production basés sur une représentation motrice et les systèmes de perception basés sur une représentation auditive. Il s'agit de systèmes dont la représentation sur la Figure 3.1 ne met pas en jeu de lien entre M et S.

On appelle théorie de la communication une théorie du couplage d'un système de production et d'un système de perception.

Une théorie de la communication est dite cohérente si ses systèmes de production et de perception sont basés sur une représentation de même nature. Une théorie de la communication cohérente est donc soit motrice, soit auditive, soit sensori-motrice. Dans les termes de la Figure 3.1, la (les) destination(s) de la flèche partant de O_S dans le système de production doit correspondre à la même représentation (M, S ou les deux) que l' (les) origine(s) de la flèche pointant vers O_L dans le système de perception (les couleurs de flèches de la Figure 3.1 doivent être les mêmes en production et en perception).

3.5.1 Nature des arguments

Nous pensons que le débat sur les théories motrices vs. auditives de la communication souffre d'une certaine ambiguïté des termes employés. En effet, chacun peut se classer dans les catégories motrices, auditives ou sensori-motrices, mais la nature de ces catégories n'est pas toujours très claire. La revue de modèles et théories exposée dans ce chapitre nous permet de distinguer trois thèmes distincts, bien qu'interdépendants :

- celui de la nature des représentations utilisées dans les systèmes de production et de perception : motrices, auditives, ou sensori-motrice.
- celui de l'architecture des systèmes de production et de perception (systèmes homogènes, ou nécessité d'une transformation de la nature des représentations).
- celui de la nature de l'apprentissage (apprentissage séparé des représentations motrices et auditives, ou conjoint ; ou connaissances précablées).

Cette distinction se retrouve en neuroanatomie, où l'on distingue les représentations motrices et sensorielles qui semblent être logées respectivement dans les lobes frontaux et temporaux, des architectures fonctionnelles de production et de perception qui semblent activer les deux types de représentation quelle que soit la tâche. La question de l'apprentissage est également centrale dans l'acquisition des associations sensori-motrices dans la voie dorsale et les neurones miroirs.

Ainsi, il nous semble utile de préciser les thèmes traités par chacun des théories et modèles de ce chapitre (dans l'ordre de présentation).

La phonologie articulatoire traite principalement de la question de la représentation, en proposant un format moteur (séquence de cibles constrictives) pour la production.

La théorie motrice de la perception concerne principalement les représentations et l'architecture. Elle stipule un format moteur traité dans les aires motrices. Ces der-

nières disposeraient d'un mécanisme de transformation sensori-motrice quasi-innée, occultant la question de l'apprentissage.

Le référentiel auditif de Guenther et collab. (1998) pour la production traite essentiellement de représentations et d'apprentissage. Il propose un format auditif pour les cibles de la production en remarquant, entre autres, que l'apprentissage des relations entre commandes motrices et conséquences auditives est mieux réalisable que celui entre commandes motrices et état de la constriction. Bien que le modèle ait son architecture propre, celle-ci n'est pas encore interprétée en termes d'aires corticales motrices et auditives, mais ceci sera réalisé de façon détaillée dans ses versions ultérieures (Guenther, 2006).

Les théories auditives de la perception traitent principalement de représentation et d'architecture. Elles stipulent un traitement direct à partir du flux sonore, dans les aires temporales. L'intervention des aires motrices pendant l'acquisition de la parole n'est pas niée, mais est considéré comme extérieure au domaine de la perception de la parole.

Le modèle DIVA de Guenther (2006) traite chacune de ces questions de façon approfondie. Il est basé sur une architecture neuro-anatomiquement valide (circuit fronto-pariéto-temporal), dans lequel chaque sous-partie traite un type de représentation particulier, et qui nécessite deux phases d'apprentissage (exploration sensori-motrice et exposition à la parole environnante) pour être opérationnel.

La théorie de la perception pour le contrôle de l'action traite également chacune de ces questions. Elle propose une co-construction des représentations motrices et perceptives (représentation et apprentissage), ainsi qu'une architecture distribuée à travers les voies ventrales et dorsales.

La formalisation que nous proposerons dans la deuxième partie de ce manuscrit permettra, dans une certaine mesure, une distinction de ces trois thèmes dans des termes de robotique bayésienne. À ce stade, nous souhaitons juste faire remarquer cette distinction et préciser qu'il convient d'être prudent quant à l'apparente contradiction des arguments employés par chaque camp lorsqu'ils ne traitent pas des mêmes thèmes.

3.5.2 Cohérence dans les théories de la communication

Bien qu'il existe une certaine séparation entre les communautés étudiant la production et la perception de la parole, nous pensons qu'une théorie de la communication se doit de placer les deux domaines à un même niveau, en considérant que l'information partagée entre deux interlocuteurs est de même nature. Ainsi, elle se doit d'être cohérente au sens défini ci-dessus, afin d'assurer que ce qui compte pour le locuteur compte aussi pour l'auditeur (*« what counts for the sender must count for the receiver [...] their representations must, at some point, be the same. »*, (Liberman, 1993)). Bien que relativement cloisonnées dans leurs domaines respectifs, production ou perception, les théories décrites dans ce chapitre semblent respecter cette cohérence. C'est par exemple le cas des théories motrices de la production et de la perception présentées, qui se sont construites autour d'observations

similaires comme les phénomènes de coarticulation, et proviennent des mêmes laboratoires Haskins. Du côté des théories auditives, Guenther (1995) remarque que :

> *« If it turns out that even the speech production process utilizes no invariant articulatory or vocal tract constriction targets, but instead uses only targets that are more directly related to the acoustic signal as suggested [...], then the motor theory claim that the speech perception system utilizes an invariant articulatory gesture representation rests on even shakier ground. »*

De plus, bien qu'il puisse être tentant de proposer une théorie de la communication dont les systèmes de production et de perception seraient tous deux homogènes (au sens défini ci-dessus, c'est-à-dire production purement motrice et perception purement auditive), une telle théorie peinerait à expliquer un bon nombre de phénomènes de la parole où un lien sensori-moteur semble nécessaire, telles que l'équivalence motrice ou la perception unifiée des consonnes.

Il est alors intéressant de noter que toute théorie de la communication cohérente contient nécessairement au moins un système (production ou perception) non-homogène, c'est-à-dire un système qui nécessite une transformation de la nature des représentations. Il en découle que toute théorie de la communication cohérente accorde un rôle fonctionnel fondamental au lien sensori-moteur, dans un sens ou dans l'autre.

Les théories sensori-motrices proposent que ce lien soit bidirectionnel, et au centre de la phylogenèse et de l'ontogenèse de la parole.

3.5.3 De la situation de communication parlée à l'agent communicant

Nous proposons maintenant une hypothèse centrale de notre travail : pour que des agents soient en mesure de communiquer correctement, il faut qu'ils possèdent un modèle interne d'une situation de communication (telle que nous l'avons définie au Chapitre 2). Autrement dit, le schéma à deux cerveaux de la Figure 2.1 doit pouvoir s'internaliser dans un schéma à un seul cerveau, qui sera la base de notre modèle d'agent communicant (Figure 3.13). Cette hypothèse, dont nous verrons qu'elle pourrait renvoyer à une étape critique de l'évolution humaine, conduit à proposer que le cerveau de notre agent communicant maitrise toute la chaîne de communication. D'une part en tant que locuteur, désignant oralement un objet O_S, il doit être capable de simuler toute la chaîne de communication qui doit conduire chez son partenaire auditeur à l'évocation d'un objet O_L. Réciproquement en tant qu'auditeur, percevant la désignation orale d'un objet O_L, il doit être capable d'inférer l'objet O_S qui a conduit initialement, chez un partenaire locuteur, à l'émission des signaux de communication correspondants. Cette chaîne de communication internalisée associe :

- un modèle de l'agent locuteur capable d'associer objets de communication O_S et représentations motrices M ;
- un modèle de la transformation articulatori-acoustique capable d'associer gestes moteurs M et stimuli auditifs S ;

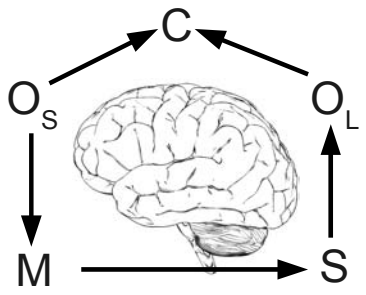


FIGURE 3.13 – Structure du modèle d'agent communicant. Les flèches qui représentaient le sens du flux de données dans le cas du modèle global de situation de communication parlée (Figure 2.10) correspondent maintenant à des processus cognitifs éventuellement inversibles, et peuvent ainsi être ici considérées comme bidirectionnelles. Nous avons toutefois choisi de les garder en l'état, d'une part pour mettre en valeur l'hypothèse d'internalisation de la situation de communication, d'autre part car elles trouveront un sens mathématique lors de la formalisation bayésienne du modèle dans la Partie II. Notons que l'image d'un cerveau ne sert ici qu'à symboliser la notion de modèle interne et n'a aucune vocation à suggérer des localisations neuroanatomiques.

- un modèle de l'agent auditeur capable d'associer stimuli auditifs S et objets de communication O_L.

Nous exposons ici quelques arguments fonctionnels (en lien avec les théories exposées dans ce chapitre) et neuroanatomiques concernant les architectures cognitives disponibles pour traiter de chacun de ces sous-modèles.

L'association directe $O_S - M$ est celle proposée par les théories motrices (Section 3.1). S'il est difficile de proposer une localisation neuroanatomique pour la représentation des objets de communication, en partie car ceux-ci ne sont pas définis précisément dans ce manuscrit (voir Chapitre 2), la localisation des représentations motrices est quant à elle communément admise dans le cortex frontal (Section 2.1).

L'association $M - S$ est, nous l'avons vu à la section précédente (3.5.2), nécessaire à toute théorie de la communication cohérente qu'elle soit motrice, auditive ou sensorimotrice. Elle semble également attestée en neuroanatomie à travers les notions de copie d'efférence, de neurones miroirs et de voie dorsale (Section 3.4).

L'association directe $S - O_L$ est celle proposée par les théories auditives (Section 3.2). Elle se justifie en neuroanatomie par l'existence d'une voie ventrale associant directement stimuli auditifs et objets sémantiques dans le cortex temporal (Section 3.4.3), dans lequel se retrouvent notamment logés les traitements auditifs de base du cortex auditif (Section 2.3).

Enfin, l'hypothèse d'une internalisation de la situation globale de communication, faisant rentrer « deux cerveaux en un », est une composante du développement d'une « théorie

de l'esprit » qui postule que chaque individu est capable de projeter sur ses congénères un modèle de son propre fonctionnement mental, et par là même de pénétrer dans le fonctionnement mental de l'autre, pour mieux le comprendre et agir de façon socialement adaptée (Baron-Cohen et collab., 1985).

Cette interprétation cognitive du modèle nous permet un début de formalisation des concepts de théories motrices, auditives et sensori-motrices de la production et de la perception de la parole (Table 3.1). Pour cela, nous considérons qu'une théorie motrice de la communication est une théorie dans laquelle le système auditif (c'est-à-dire l'association $S - O_L$) ne joue aucun rôle fonctionnel dans la production et la perception et peut donc être désactivé. De même, une théorie auditive de la communication est une théorie dans laquelle le système moteur (c'est-à-dire l'association $O_S - M$) ne joue aucun rôle fonctionnel dans la production et la perception et peut donc être désactivé. Une théorie sensori-motrice quant à elle est une théorie dans laquelle aucun des deux systèmes (moteur et auditif) n'est désactivé.

La Table 3.1 illustre la conception des théories motrices, auditives et sensorimotrices dans le cadre de notre modèle d'agent, en lien avec la littérature décrite dans les différentes sections ce chapitre.

Les théories motrices se définissent alors par l'absence de lien direct entre S et O_L et nécessitent donc un passage préalable par la modalité motrice avant de pouvoir inférer l'objet de communication en perception. C'est la récupération de l'intention du locuteur, O_S, qui tient lieu d'objet perçu. Elles permettent par contre une production indépendante des représentations auditives (association directe $M - O_S$).

Les théories auditives se définissent par l'absence de lien direct entre O_S et M et nécessitent donc un passage par la modalité auditive avant de pouvoir fournir un geste moteur en production. C'est l'inférence du geste adéquat conduisant à la perception de l'objet O_L qui tient lieu de programme moteur. Elles permettent par contre une perception indépendante des représentations motrices (association directe $S - O_L$).

Les théories sensori-motrices possèdent les deux associations directes $O_S - M$ et $S - O_L$ et permettent ainsi à la fois une production et une perception directe, tout en étant capables d'une transformation sensori-motrice si nécessaire.

3.5.4 Conclusion

Nous avons dans ce chapitre exposé les principaux courants théoriques en production et perception de la parole en distingant les théories motrices, auditives et sensori-motrices. Puis nous avons vu comment ce débat déjà ancien a été récemment ravivé par les avancées des neurosciences, sans qu'il ne soit pour autant résolu. Étant donnée la difficulté du problème, traduite par la vivacité du débat malgré son ancienneté, nous pensons que la formalisation et la modélisation peuvent contribuer à le faire avancer, en apportant des éléments d'une autre autre nature, des « expériences de simulation » qui viennent compléter les expériences de comportement et de neurophysiologie.

C'est donc l'objectif de ce chapitre que de proposer un modèle conceptuel intégrateur, faisant l'hypothèse d'une internalisation par les agents de la situation de communication

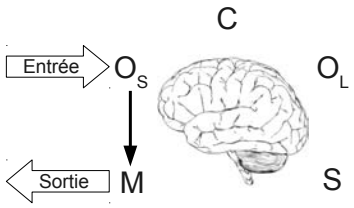
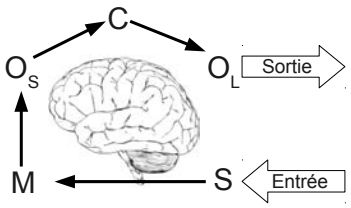
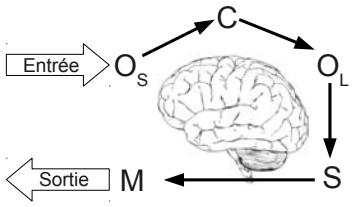
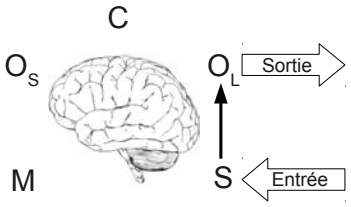
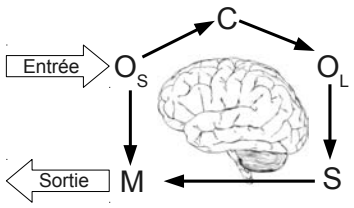
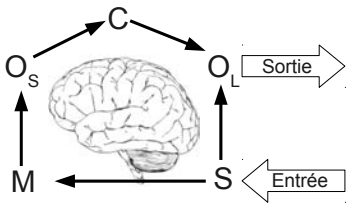
	Production	Perception
Moteur	Section 3.1.1 	Section 3.1.2 
Auditif	Section 3.2.1 	Section 3.2.2 
Sensori-moteur	Section 3.3.1 	Section 3.3.2 

TABLE 3.1 – Synthèse des différents courants théoriques en production et perception de la parole avec renvois aux sections correspondantes de ce chapitre, et conceptions en terme de modèle cognitif d'agent. Les flèches indiquent ici un flux d'information.

parlée décrite au chapitre précédent, et qui regroupe les trois grands courants théoriques dans un schéma unique et cohérent que nous synthétisons Figure 3.14. C'est sur cette base que ce modèle pourra être formalisé au Chapitre 8.

Ce vif débat, historiquement cantonné aux théories motrices et auditives, a produit une grande quantité d'arguments de qualité, mais souvent contradictoires, en faveur des unes et des autres. Les théories sensori-motrices tentent de subsumer ces deux courants théoriques en fournissant un cadre intégrateur. Elles postulent une co-construction des systèmes de production et de perception au cours du développement, qui leur permet d'intégrer de manière optimale les propriétés des formats moteurs et auditifs en terme de

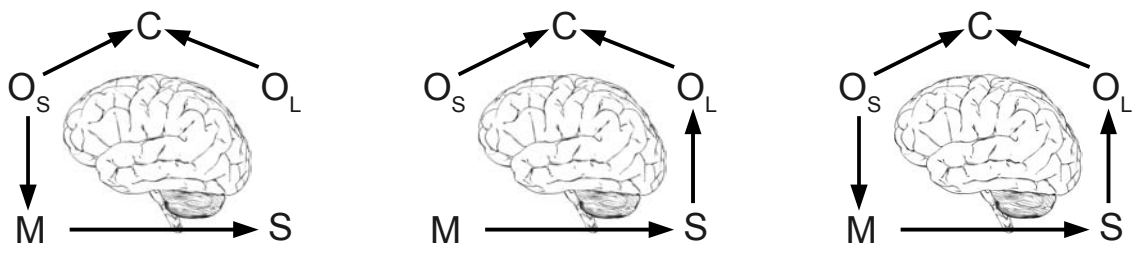
représentation. Le lien sensori-moteur ainsi acquis, bien qu'il ne soit pas nécessairement impliqué en conditions normales de production et de perception, reste toutefois disponible en situations perturbées (production contrainte, bruit en perception, par exemple). C'est en se construisant ainsi de façon intriquée, que les systèmes de production et de perception peuvent à la fois se permettre d'être homogènes en situation normale, tout en s'aidant l'un l'autre lorsque nécessaire.

Les théories sensori-motrices présentent donc les avantages :

- d'être cohérentes ;
- de proposer un cadre intégrateur aux théories motrices et perceptives, ainsi qu'à leurs arguments respectifs qui semblaient paradoxalement recevables mais contradictoires ;
- de s'ancrer dans les observations neuroanatomiques de co-activation des aires motrices et sensorielles dans les tâches de production et de perception, ainsi que de connexions fortes entre ces aires à travers la voie dorsale et les neurones miroirs, tout en permettant un fonctionnement homogène (au sens des définitions) en situations normales.

De plus, la comparaison des Figures 2.10 et 3.14 suggère que les théories sensori-motrices soient les plus à même de fournir une architecture cognitive modélisant dans son ensemble une situation de communication parlée (à la fois le locuteur, la transformation articulatori-acoustique et l'auditeur). Si une réponse appropriée à une situation donnée dépend de la capacité d'un agent à modéliser correctement cette situation, les théories sensori-motrices devraient pouvoir en tirer avantage.

Il s'agit toutefois ici plus d'arguments de rationalité que d'une preuve fonctionnelle décisive. Nous verrons dans la deuxième partie de ce manuscrit comment la modélisation d'une théorie de la communication peut argumenter cette position, et, dans la troisième partie, comment les simulations tirent partie de ces propriétés fonctionnelles.



(a) Théories motrices (b) Théories auditives (c) Théories sensori-motrices

FIGURE 3.14 – Les théories motrices, auditives et sensori-motrices dans le cadre du modèle d'agent. Comme pour la Figure 3.13, les flèches qui représentaient le sens du flux de données dans le cas du modèle global de situation de communication parlée (Figure 2.10) correspondent maintenant à des processus cognitifs éventuellement inversibles, et peuvent ainsi être ici considérées comme bidirectionnelles.

Chapitre 4

Émergence des systèmes phonologiques

Malgré la grande diversité des séquences sonores qui peuvent être produites par le conduit vocal humain, ainsi que celle des langues du monde, les études statistiques montrent que les systèmes phonologiques obéissent à certaines régularités.

Universellement, on retrouve d'abord l'organisation syllabique, alternant consonnes et voyelles avec une large prédominance des syllabes de type consonne-voyelle (CV).

Les inventaires des systèmes phonologiques des langues du monde (Maddieson et Precoda, 1989, UCLA Phonological Segment Inventory Database) recensent 177 voyelles et 654 consonnes. Malgré l'immense variété de systèmes phonologiques possibles induite par une simple règle combinatoire sur ces ensembles, il existe en fait de grandes régularités. Ainsi, le triplet /a,i,u/ est présent dans 97% des langues, et une nette préférence des systèmes vocaliques à 5 voyelles est remarquée (Vallée, 1994). Concernant les consonnes, les phonèmes préférés sont /b,d,g/ ou /p,t,k/ (consonnes bilabiales, coronales, et vélaires), présents dans la quasi-totalité des langues.

Ces observations faites, il reste à expliquer les raisons de ces régularités. Ce sont les travaux pionniers de Björn Lindblom, en particulier sur la théorie de la dispersion (Liljencrants et Lindblom, 1972, exposée ci-dessous), qui ont initié toute une série de recherches qu'il a qualifié plus tard ainsi (Lindblom, 1984) : « dériver le langage du non-langage ». Ces recherches ont débouché sur un ensemble de théories dites « orientées-substance » qui tentent de dériver le langage à partir d'une substance prélangagière. Nous distinguons trois types d'approche.

Les théories de la forme qui abordent le problème « de l'intérieur », en considérant les régularités comme des solutions à des problèmes d'optimisation dans l'espace phonétique, mettant en jeu des contraintes motrices et perceptives.

Les théories de l'émergence qui abordent le problème « de l'extérieur », en considérant ces régularités comme un phénomène émergent de comportements non-langagiers plus primitifs.

Les modèles computationnels d'agents interagissants qui abordent le problème sous l'angle des systèmes complexes, en utilisant la simulation informatique pour laisser émerger les régularités de l'interaction entre agents prélangagiers.

Le point central que nous voulons introduire ici est que toutes ces théories sont, nous le verrons, des *théories de la communication*, faisant dériver des nécessités et contraintes d'une « bonne communication » les propriétés du langage, à travers des chemins divers. C'est en ce sens que les recherches sur l'émergence des formes du langage sont évidemment au cœur de notre projet théorique.

4.1 Théories de la forme

4.1.1 Les théories de la dispersion

La théorie de la dispersion a été initiée par Liljencrants et Lindblom (1972) pour être améliorée en une théorie de la dispersion adaptative (Lindblom, 1990) puis reprise par Schwartz et collab. (1997b) sous le nom de théorie de la dispersion-focalisation. Elle est basée sur l'étude des systèmes vocaliques dans les langues du monde afin d'analyser la manière dont les voyelles s'y distribuent. Le principe de base est relativement simple : les systèmes phonologiques ont tendance à disperser leurs éléments dans l'espace auditif afin d'optimiser leur discrimination, comme illustré Figure 4.1.

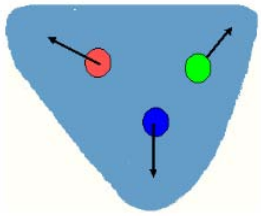


FIGURE 4.1 – Illustration de la théorie de la dispersion : les systèmes phonétiques ont tendance à disperser leurs items dans l'espace auditif afin d'optimiser leur discrimination, à la manière d'un ensemble de charges électriques de même signe qui se repousseraient les unes les autres dans une cuve dont les bords représenteraient les limites de l'espace auditif (typiquement, le triangle vocalique).

Dans la version initiale (Liljencrants et Lindblom, 1972), le postulat de base est donc que les systèmes vocaliques favorisés sont ceux qui parviennent à optimiser une certaine distance perceptive entre les voyelles qui le composent afin de favoriser leur discrimination. La prédiction de certaines voyelles ne peut donc se faire que par rapport à d'autres. Le modèle repose sur un espace vocalique à deux dimensions (le triangle vocalique dans l'espace des deux premiers formants) dans lequel est définie une notion de distance entre deux voyelles. Ainsi, selon cette théorie, les systèmes vocaliques tentent de minimiser l'expression :

$$G = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(\frac{1}{d_{i,j}} \right)^2, \tag{4.1}$$

où n est le nombre total de voyelles et $d_{i,j}$ la distance perceptive entre deux voyelles.

Nous ne détaillerons pas le calcul de cette distance mais on peut la voir simplement comme la distance euclidienne dans l'espace des deux premiers formants. Ainsi, cette expression est minimale lorsque les distances perceptives inter-voyelles sont maximales. Ce principe de dispersion permet de trouver les valeurs phonétiques des systèmes vocaliques en fonction de leur nombre de voyelles. En particulier, il explique pourquoi les voyelles /i/, /a/, /u/ sont présentes dans la quasi-totalité des langues du monde. Elles correspondent aux sommets du triangle vocalique et assurent ainsi une dispersion maximum.

En 1986, Lindblom propose en perspective de ses travaux de prendre en compte les facteurs articulatoires dans sa théorie. Il s'agit de ne plus se focaliser sur les seuls intérêts de l'auditeur (maximiser les distances perceptives), mais également sur ceux du locuteur en minimisant le coût articulatoire. En 1990, cette adaptation simultanée entre les intérêts de l'auditeur et ceux du locuteur est à la base d'une nouvelle théorie : celle de la dispersion adaptative (Lindblom, 1990). L'expression que pourraient minimiser les systèmes vocaliques devient alors :

$$G = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(\frac{A_{i,j}}{d_{i,j}} \right)^2$$

où $A_{i,j}$ représente le coût articulatoire entre les phonèmes i et j, coût dont la définition reste cependant assez délicate.

Des travaux similaires existent également pour les systèmes consonantiques, qui considèrent que le système /b,d,g/ est optimal en terme de dispersion acoustique dans l'espace F2-F3 (Abry, 2003; Schwartz et collab., 2011, en révision), sous certaines contraintes motrices. Les consonnes plosives ont en effet la particularité d'avoir un premier formant F1 bas, autour de 250 Hz, sauf pour celles dont la constriction a lieu complètement à l'arrière du conduit vocal, au niveau de la gorge (voir Figure 2.7 au Chapitre 2). Si l'on se place dans l'espace F2-F3, /d/ et /g/ occupent en effet deux sommets du triangle, alors que le troisième est disputé par /b/ et les plosives arrières. Le fait que ces dernières soient très peu utilisées dans les langues du monde, alors qu'elles sont de très bons candidats en terme de dispersion acoustique (à la fois à un sommet du triangle F2-F3, et avec un F1 élevé), s'expliquerait alors par leur réalisation motrice particulière qui nécessite une ouverture de la mâchoire pour plaquer la langue à l'arrière du conduit vocal. En s'appuyant sur une théorie de l'émergence de la syllabe (Frame/Content, décrite dans ce chapitre en 4.2.2), Schwartz et collab. (2011, en révision) proposent alors des raisons phylogéniques à la difficulté de réaliser des consonnes sur une ouverture de mâchoire, laissant ainsi la place à /b,d,g/ comme système préféré (Figure 4.2).

Notons finalement les travaux de Redford et collab. (2001) qui expliquent la prédominance des syllabes CV en terme d'optimisation sous contraintes de longueur des mots, d'intelligibilité, de mémoire et de distinction perceptive.

4.1.2 La théorie quantique

La théorie quantique (Stevens, 1972, 1989) propose que l'origine des traits phonétiques binaires provient en grande partie des non-linéarités entre les paramètres articulatoires et

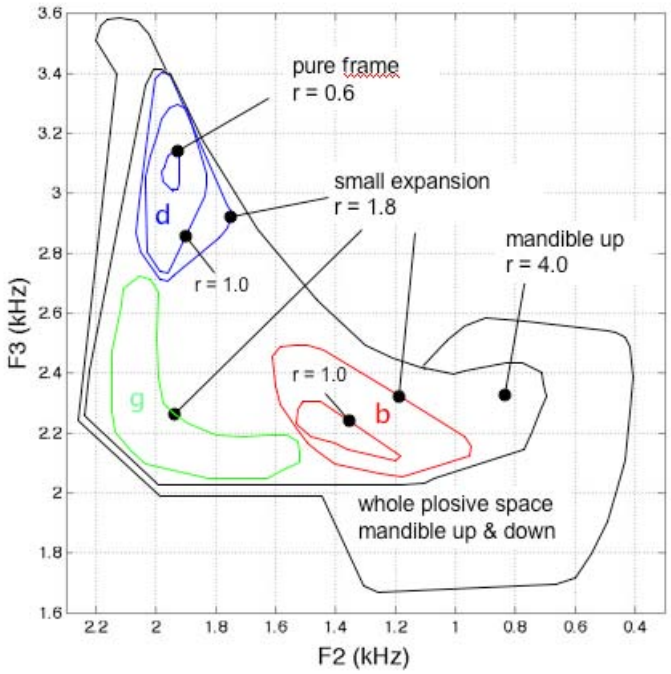


FIGURE 4.2 – Expansion de l'espace des consonnes plosives dans l'espace F2-F3, d'après Schwartz et collab. (2011, en révision). Le paramètre r représente la taille de la zone d'exploration de l'espace articulatoire autour de la position neutre des articulateurs. /b,d,g/ forme un triangle acoustique optimal sous la contrainte d'une mâchoire fermée et d'une exploration limitée. Les formants sont calculés par un modèle articulatoire.

acoustiques dans le système de production des sons. Il existe ainsi deux sortes de régions dans l'espace articulatoire (Figure 4.3) :

- des régions pour lesquelles le résultat acoustique est stable (région I et III sur la Figure 4.3),
- des régions pour lesquelles une faible variation des paramètres articulatoires entraîne une forte variation des paramètres acoustiques (région II sur la Figure 4.3).

Pour Stevens, ces non-linéarités fourniraient un critère de sélection des voyelles et des consonnes. En effet, les régions I et III nécessitent une plus faible précision articulatoire que la région II. Au contraire, cette dernière affecte la perception des sons, qui ne peuvent être produits que par une grande précision articulatoire. Ce dernier type de régions constituerait donc des frontières naturelles de discrimination des phonèmes, qui s'organiseraient de part et d'autres de celles-ci. Un phonème serait d'autant plus fréquent dans les langues du monde qu'il est stable, expliquant ainsi les tendances universelles observées.

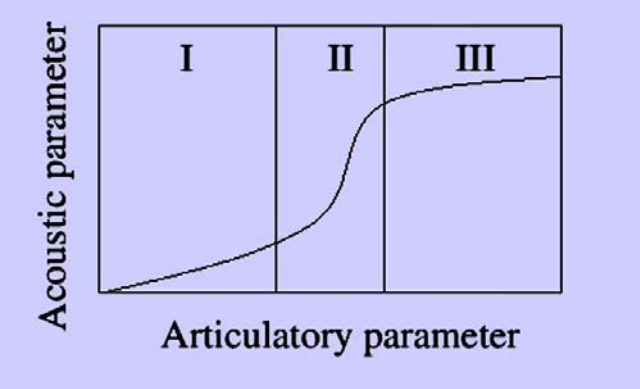


FIGURE 4.3 – Non-linéarités dans le passage des paramètres articulatoires aux paramètres acoustiques, d'après Stevens (1972).

4.2 Théories de l'émergence

Les théories de l'émergence tentent de comprendre l'émergence du langage, en particulier des systèmes phonologiques, à partir de comportements qui ne sont pas encore du langage mais qui peuvent constituer les bases à partir desquelles il aurait pu émerger. Ces théories trouvent leurs inspirations principalement dans l'étude des précurseurs phylo- et ontogénique des humains langagiers : les primates non-humains et les bébés.

Sans nous lancer dans une description exhaustive des théories (fort nombreuses) de l'émergence du langage (voir une analyse historique, épistémologique et critique dans Boë et collab. (2011)), nous allons en décrire quatre, qui ont inspiré nos travaux de modélisation, en nous apportant des éléments de réflexion sur quatre questions principales :

- la question de la référence, c'est-à-dire de l'émergence de comportements à valeurs sémantiques (ce à propos de quoi l'on communique) ;
- la question du média de communication, et essentiellement de la coordination de la main et de la voix ;
- la question de la production et de l'enchaînement de signaux de communication efficaces, et de la création de structures complexes à partir de structures simples ;
- la question de l'internalisation, totale ou partielle, de la situation de communication, allant de simples liens sensori-moteurs à des couplages cognitifs complets.

4.2.1 L'hypothèse du système miroir

L'hypothèse du système miroir (MSH, pour Mirror System Hypothesis (Rizzolatti et Arbib, 1998; Arbib, 2005a)) considère la reconnaissance d'actions bracchio-manuelles transitives vers des objets (saisir une pomme, par exemple) comme *bootstrap* de l'émergence du langage. Cette théorie trouve ses fondements dans la découverte des neurones miroirs (voir 3.4.2), qu'elle considère comme le corrélat neurologique à la base de la reconnaissance d'action. Cette hypothèse est soutenue par les similitudes importantes entre l'aire F5 chez le singe (partie du système moteur où se trouvent des représentations de la main et de la

bouche), et l'aire de Broca chez l'humain (centre moteur du langage), au niveau de leur localisation (cortex frontal inférieur) et de leur cytoarchitectonie.

Selon cette théorie, les propriétés syntaxiques élémentaires du langage sont supposées être déjà présentes dans le système moteur par des associations entre actions et objets (équivalent de la structure syntaxique verbe-argument) (Roy et Arbib, 2005). La reconnaissance d'une action transitive vers un objet exécutée par un semblable, et rendue possible par le système miroir, serait alors la capacité prélangagière nécessaire à la phylogenèse du langage. Puis l'étape majeure dans cette phylogenèse (celle qui différencie les humains des grands singes) est ici le passage de la capacité d'imitation d'actions simples (élémentaires) à l'imitation d'actions complexes (séquence d'actions élémentaires). C'est cette capacité qui engendrerait les propriétés récursives du langage humain, bien avant l'acquisition du contrôle du conduit vocal. Viendrait ensuite le passage d'un répertoire d'actions manuelles à un répertoire de pantomimes, c'est à dire des gestes conventionnalisés pour la communication. Le passage de la communication gestuelle à la communication parlée se serait produit plus tardivement dans l'évolution. Arbib (2005b) propose une évolution conjointe de ces deux types de communication, postérieure à la capacité d'imitation complexe. La Figure 4.4 illustre cette hypothèse.

Figure 4.4 – Les étapes principales de l'émergence du langage selon l'hypothèse du système miroir (d'après Arbib (2009)). La capacité d'imitation simple (*simple imitation*) permet à l'ancêtre commun des grands singes et des humains d'acquérir un petit répertoire de gestes communicatifs bracchio-manuels. C'est ensuite la capacité d'imitation complexe (*complex imitation*), propre à la lignée humaine, qui permet le passage à un répertoire ouvert de signes conventionnalisés, les proto-signes (*protosign*). C'est seulement ensuite que vient l'acquisition progressive du contrôle du conduit vocal qui, en évoluant conjointement avec les proto-signes, permet l'évolution vers le langage humain.

Arbib (2005b) propose ainsi sept grandes étapes vers l'émergence du langage, les trois premières étant antérieures à la lignée humaine.

1. Geste de saisie.
2. Un système miroir pour le geste de saisie, partagé avec l'ancêtre commun des humains et des singes.
3. Un système d'imitation simple pour les gestes de saisie orientés vers des objets, partagé avec l'ancêtre commun des humains et des singes.

Les trois prochaines étapes distinguent la lignée humaine des grands singes.

4. Un système d'imitation complexe pour le geste de saisie, c'est-à-dire la capacité à reconnaître l'action d'un congénère comme un ensemble d'actions familières et à les répéter, ou à reconnaître qu'elle combine des actions nouvelles qui peuvent être approchées par des variantes d'actions déjà présentes dans le répertoire.
5. Le proto-signe, un système de communication manuel qui permet de dépasser le répertoire fixe des vocalisations des primates pour aller vers un répertoire ouvert.
6. La proto-parole, résultant des mécanismes de contrôle ayant évolué pour le proto-signe, transférés vers le contrôle du conduit vocal avec une flexibilité croissante.

   La dernière étape est supposée n'impliquer que peu ou pas d'évolution biologique, mais résulter plutôt de l'évolution culturelle chez les *Homo Sapiens*.
7. Le langage, le passage du cadre action-objet à la structure verbe-argument puis à la syntaxe et à la sémantique, ainsi que la co-évolution de la complexité cognitive et linguistique.

Arbib (2005b) résume l'hypothèse du système miroir ainsi :

> *« The* parity requirement *for language in humans – that what counts for the speaker must count approximately the same for the hearer – is met because Broca's area evolved atop the mirror system for grasping with its capacity to generate and recognize a set of actions. »*

4.2.2 Frame, then Content

La théorie Frame/Content (MacNeilage, 1998) est plus centrée sur l'émergence de la parole articulée que sur celle du langage à proprement parler (impliquant une référence) bien qu'elle trouve ses inspirations dans les comportements communicatifs orofaciaux des primates non-humains (claquements de lèvres, de langue et de dents). Elle propose un chemin évolutif partant des cyclicités ingestives présentes chez tous les mammifères vers des cyclicités communicatives visuofaciales chez les grands singes, puis la parole articulée chez les humains. Cette hypothèse s'appuie sur des données d'acquisition de la parole chez le bébé, notamment au cours du babillage. MacNeilage soutient en effet une version faible de la position d'Haeckel selon laquelle l'ontogénie récapitule la phylogénie. Malgré la discréditation de cette hypothèse, MacNeilage pense qu'elle reste en partie plausible si on la limite aux fonctions motrices humaines.

La théorie Frame/Content distingue deux phases, tant dans l'évolution de l'espèce humaine que dans l'acquisition de la parole chez le bébé.

D'abord le cadre (*frame*), qui consiste en une alternance d'ouvertures et de fermetures de la mâchoire provenant de la mastication, les autres articulateurs (langue, lèvre, etc...) n'étant pas encore correctement contrôlés. Ce cycle mandibulaire serait à l'origine de l'alternance de voyelles (configurations ouvertes) et de consonnes (configurations fermées), en permettant une modulation des conséquences acoustiques qu'elles impliquent. Dans cette

phase, ces deux catégories sont très fortement corrélées puisque seule la mâchoire est contrôlée, impliquant un lieu d'articulation (avant, central ou arrière) similaire en ouverture et en fermeture. La validité de cette phase est soutenue par les données sur le babillage des bébés humains ; elle serait contrôlée par l'aire motrice supplémentaire contrôlant les mouvements cycliques.

Puis, après le cadre vient le contenu (*content*), phase dans laquelle les autres articulateurs, en particulier la langue et les lèvres, acquièrent un contrôle indépendant pendant les phases d'ouverture et de fermeture, permettant ainsi l'enchaînement de séquences voyelle-consonne décorrélées. La validité de cette phase est soutenue par les données sur la chronologie du contrôle des articulateurs chez le bébé humain. Son contrôle s'effectuerait latéralement dans le cortex frontal.

En postulant deux phases successives dans la phylo- et l'ontogenèse de la parole, la théorie Frame/Content implique la prédiction que les vocalisations précédant l'acquisition du contenu soient constituées de séquences consonnes-voyelles fortement corrélées. En effet, si la mâchoire est le seul articulateur contrôlé, tous les autres étant au repos, la syllabe résultante est entièrement déterminée par la morphologie du sujet. C'est ce que MacNeilage appelle le cadre pur, séquence consonne-voyelle obtenue lors d'un cycle de mâchoire lorsque tous les autres articulateurs sont au repos, correspondant à des syllabes de type /ba/ (ou /da/ pour des morphologies différentes selon Vilain et collab. (1999)). En considérant une certaine variabilité de la position des articulateurs au repos, MacNeilage prévoit deux autres types de cadres : le cadre avant où la langue est légèrement avancée, produisant des syllabes de type /di/ (consonne et voyelle « avant »), et le cadre arrière avec la langue légèrement reculée produisant des syllabes de type /gu/ (consonne et voyelle « arrière »). La Figure 4.5 illustre ces dépendances.

Ce n'est qu'avec l'acquisition du contrôle du contenu que la parole peut devenir articulée, avec un contrôle fin des articulateurs portés par la mâchoire, permettant des productions consonne-voyelle décorrélées et un début de compositionalité phonémique.

4.2.3 Vocalize-to-Localize

La théorie Vocalize-to-Localize (Abry et collab., 2004) propose que le comportement prélangagier à l'origine de l'émergence du langage soit la déixis, c'est à dire la capacité de montrer des choses à un congénère avec la main et/ou le regard (du grec *deiktikos*, action de montrer). Ce comportement, bien qu'absent chez les singes dans leur milieu écologique, est attesté en captivité, où il peut être synchronisé avec des vocalisations primitives (non articulées). C'est cette synchronisation entre la main et la voix qui aurait pu fournir les bases d'associations entre une référence par le geste de pointer et une vocalisation. Ainsi, le langage aurait émergé à partir de cette capacité à « montrer de la voix » à distance, ou à « vocaliser pour localiser ».

Un autre argument en faveur de la théorie vient du développement ontogénétique, dans lequel la coordination entre gestes de pointer et vocalisations pourrait être d'une grande importance dans l'acquisition du langage, apparaissant juste avant le développement de la syntaxe et l'explosion de la taille du vocabulaire.

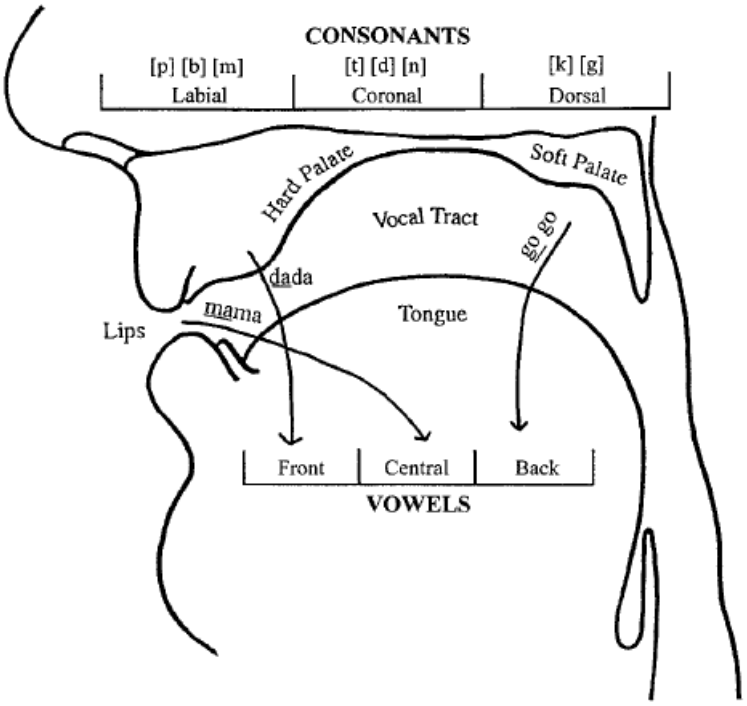


FIGURE 4.5 – Les différents types de cadre prévus par la théorie Frame/Content, d'après MacNeilage et Davis (2000) : le cadre pur (consonne labiale, voyelle centrale), le cadre avant (consonne coronale, voyelle avant) et le cadre arrière (consonne dorsale, voyelle arrière).

Cette théorie permet d'étendre la théorie Frame/Content exposée précédemment en la dotant d'un mécanisme pour la référence dans le geste de pointer. Dans sa thèse, Ducey-Kaufmann (2007) propose que les deux systèmes, de parole précoce par le babillage et de référence précoce par le geste de pointer, se rencontrent dans ce qu'elle appelle un rendez-vous développemental autour de l'âge de un an chez l'enfant (Figure 4.6). De ce point de vue, ce serait donc la synchronisation entre le cycle mandibulaire (le cadre de la parole) et le geste de pointer (le cadre du signe) qui fournirait à l'enfant prélangagier, ou de façon présupposée à l'ancêtre phylogénétique de l'humain prélangagier, la capacité d'associer références et vocalisations. La naissance des mots pourrait ainsi s'ancrer dans cette synchronisation entre geste de pointer et vocalisation. De là pourraient découler des principes de mise en forme des mots, et notamment la tendance à observer majoritairement des mots bisyllabiques, ancrés dans la coordination d'un pointer déictique et de deux oscillations mandibulaires portant deux germes syllabiques naturels. En effet, Ducey-Kaufmann (2007) montre chez le bébé que la durée d'un geste de pointer égale approximativement celle de deux oscillations mandibulaires ; voir aussi Amélie Rochet-Capellan et collab. (2007) pour une mise en évidence de même nature chez l'adulte.

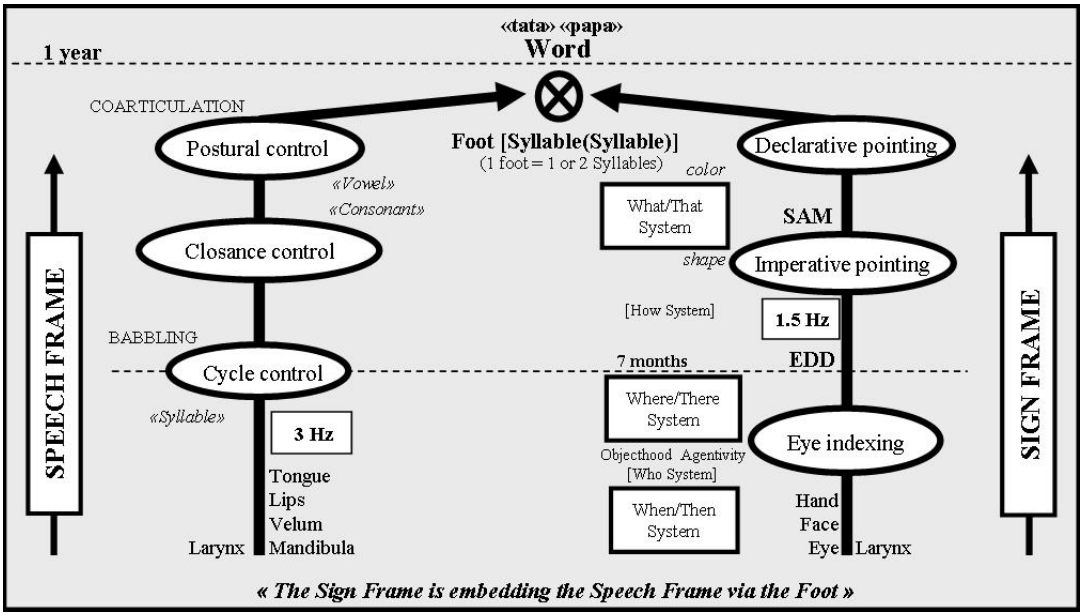


FIGURE 4.6 – La rencontre du cadre de la parole et du cadre du signe, d'après Ducey-Kaufmann (2007). Le mot se construit par la rencontre d'un signifiant et d'un signifié, implémentés physiquement dans un système de vocalisation (*Speech Frame*) et un système de monstration (*Sign Frame*), respectivement. Le premier met en jeu la mâchoire, à une fréquence d'oscillation d'environ 3 Hz, le second le bras à fréquence d'environ 1,5 Hz. Leur rencontre implique donc un rapport de 2 : 1, ce qui expliquerait la prédominance des mots de deux syllabes lors des premières productions autour de l'âge de un an.

4.2.4 Les intentions partagées de Tomasello et collab. (2005)

Tomasello et collab. (2005) proposent que l'émergence du langage (entre autres) dépende d'une motivation à partager des états psychologiques avec ses congénères, qui engendrerait une capacité à participer à des activités collaboratives.

Pour cela, les auteurs définissent un but comme la représentation mentale d'un état de l'environnement désiré (le but d'ouvrir une boîte correspond à la représentation mentale d'une boite ouverte), et une intention comme un plan d'action choisi pour atteindre un but (utiliser un couteau pour ouvrir une boîte, par exemple).

Ils remarquent alors qu'il convient de différencier la compréhension des actions intentionnelles (présente en partie chez les grands singes), du partage des intentions entre individus (activité collaborative spécifique à l'espèce humaine). En effet, la compréhension des actions intentionnelles correspond, dans sa version la plus évoluée, à la capacité de comprendre qu'un individu considère et choisit des plans d'action à effectuer pour ses actions intentionnelles. Le partage d'intentions quant à lui réfère à la capacité à interagir collaborativement avec un autre vers un but partagé, en coordonnant les rôles de chacun. Le passage de la compréhension des actions intentionnelles au partage d'intentions nécessiterait une motivation à partager des états psychologiques propre à l'espèce humaine.

La Figure 4.7 illustre une activité collaborative dans laquelle un but et une intention partagée sont formés.

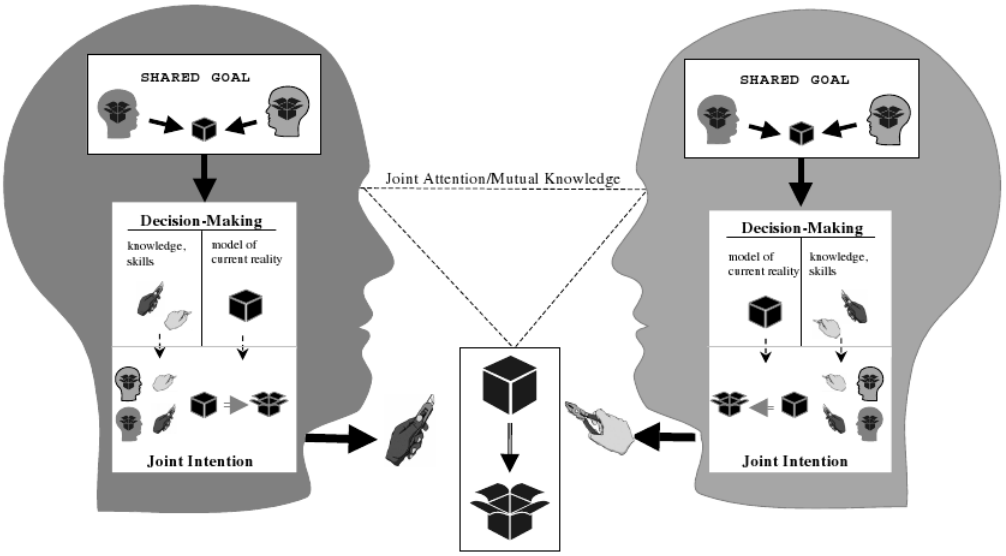


FIGURE 4.7 – Activité collaborative dans laquelle un but et une intention partagée sont formés, d'après Tomasello et collab. (2005). Un individu interagit avec un agent intentionnel vers un but partagé (*shared goal*), par des plans d'actions coordonnées (*joint intention*) et une attention partagée sur une partie de l'environnement (*joint attention/mutual knowledge*). Chaque agent doit ainsi se représenter le but partagé ainsi que les plans d'action incluant des rôles complémentaires.

Dans la recherche des capacités qui ont pu permettre l'évolution vers la cognition et la culture humaine, unique dans le règne animal, une adaptation pour le partage d'intentions semble, pour les auteurs, plus élémentaire que d'autres candidats tels que le langage ou la théorie de l'esprit (pour les auteurs, dire que seuls les humains possèdent le langage revient à dire que seuls les humains construisent des gratte-ciels, alors que, plus simplement, seuls les humains (parmi les primates) construisent des abris qui tiennent debout).

Pour Tomasello et collab. (2005), ce serait cette capacité à participer à des activités collaboratives mettant en jeu des intentions partagées, nécessitant à la fois une compréhension des intentions d'autrui et une motivation à partager des états psychologiques avec ses congénères, qui serait le comportement prélangagier qui aurait permis à l'espèce humaine d'évoluer vers des activités cognitives de plus haut niveau tel que le langage.

4.3 Modèles computationnels d'agents interagissants

Les travaux pionniers de Luc Steels au milieu des années 90 (par exemple Steels, 1996, 1997) ont permis d'aborder la question de l'émergence du langage par une approche computationnelle basée sur la simulation informatique multi-agents. L'idée est d'étudier comment

certaines propriétés du langage (lexique, syntaxe, phonologie ...) peuvent émerger d'interactions locales entre des agents simulés, afin d'évaluer quantitativement certains scénarios d'émergence. Steels (2006) dégage quatre grandes étapes communes à ce type de simulations :

- proposer des hypothèses concernant un lien entre des mécanismes cognitifs préexistants ainsi que des facteurs externes, et l'émergence de certaines propriétés du langage ;
- implémenter computationnellement ces mécanismes dans des agents simulés ;
- définir un scénario d'interaction entre ces agents, éventuellement embarqués dans une simulation du monde extérieur ;
- expérimenter par des simulations informatiques et étudier l'émergence des propriétés attendues.

Steels précise que ce type d'approches ne prouve rien sur l'évolution du langage parce que différents mécanismes peuvent traiter les mêmes objectifs communicatifs, mais qu'il a au moins le mérite de proposer différents chemins possibles.

Depuis, un bon nombre de travaux ont été réalisés dans ce domaine, couvrant une large variété de propriétés du langage. La plupart se sont concentrés sur le partage du lexique, la compositionnalité, l'émergence de la grammaire, ou l'ancrage des symboles. Par exemple, les travaux s'intéressant au partage du lexique (comment des associations consistantes entre mot et sens peuvent émerger d'interactions locales entre agents) considèrent souvent le mot comme une entité abstraite non reliée à des traits articulatoires ou auditifs (Kaplan, 2000, 2005; Griffiths et Kalish, 2005). Un nombre plus restreint ont traité de l'émergence de la phonologie. Dans le cadre de ce manuscrit, qui s'intéresse en particulier à l'émergence des systèmes phonologiques, nous choisissons de décrire trois réalisations computationnelles majeures traitant de l'émergence des systèmes de voyelles.

4.3.1 Le processus d'attraction-répulsion de Berrah et Laboissière (1999)

Berrah et Laboissière (1999) ont mené les premières simulations d'émergence de la phonologie par interactions entre agents. Le modèle est basé sur un procédé d'attraction-répulsion d'items dans le triangle vocalique. Initialement, chaque agent possède un nombre d'items fixé, répartis aléatoirement dans leur espace acoustique. Puis les agents interagissent deux à deux : l'agent locuteur choisit un item au hasard dans son lexique et le produit, et l'agent auditeur le perçoit et le compare à ses propres prototypes. Selon un principe d'attraction, l'item le plus proche est rapproché du son perçu, alors que les autres en sont écartés par un procédé de répulsion (Figure 4.8). Ce système, très relié à la théorie de la dispersion (Liljencrants et Lindblom, 1972), permet de prédire les grandes tendances des systèmes de voyelles des langues du monde. Cependant, il introduit dans les agents un processus d'attraction-répulsion difficile à relier à des fonctions cognitives prélangagières.

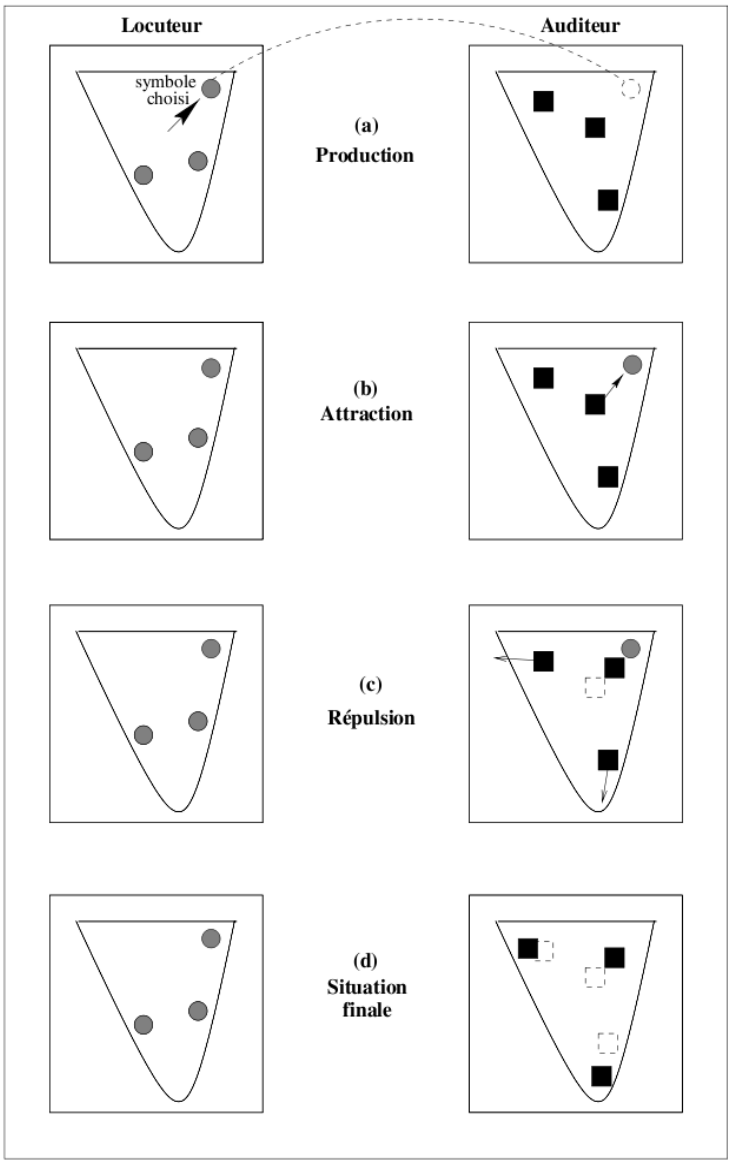


FIGURE 4.8 – Une interaction entre deux agents, d'après Berrah (1998).

4.3.2 Les jeux d'imitation de de Boer (2000)

Les simulations de de Boer (2000) sont plus explicites à ce sujet, en considérant un comportement pré-linguistique plausible : l'imitation. Les agents considérés sont dotés d'un synthétiseur vocal capable de calculer les formants résultants d'une configuration articulatoire. Les agents interagissent toujours deux par deux dans des « jeux d'imitation » qui se déroulent de la façon suivante (illustrés Figure 4.9).

- Le premier agent, appelé initiateur, choisit aléatoirement un prototype de voyelle P1 dans son répertoire et le prononce.
- Le deuxième agent, appelé imitateur, perçoit le son correspondant (les formants,

produits par le synthétiseur vocal), trouve le prototype de son répertoire qui en est le plus proche, et le prononce pour imiter l'agent initiateur.
- À son tour, l'agent initiateur perçoit le son correspondant et trouve le prototype P2 de son répertoire qui en est le plus proche.
- Enfin, par un retour non verbal, les agents sont informés du succès ou de l'échec du jeu d'imitation (succès si P1=P2, échec sinon).

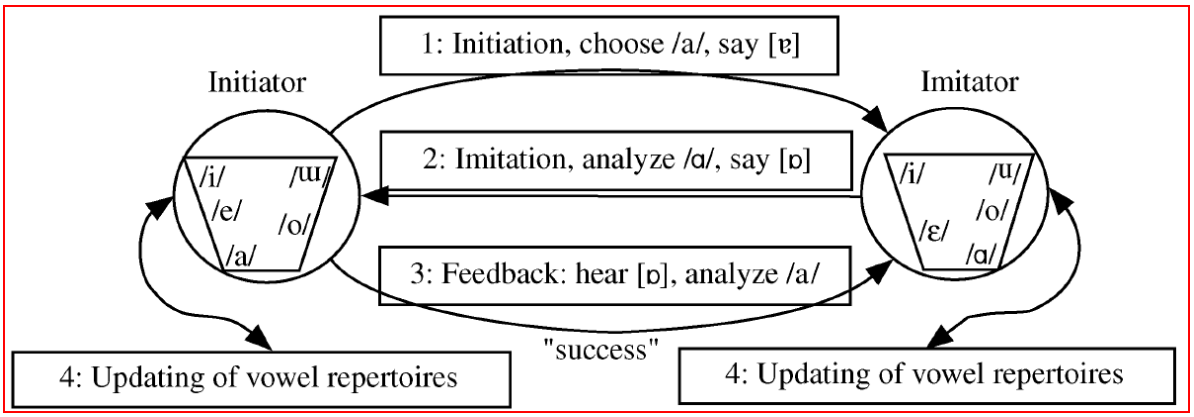


FIGURE 4.9 – Une interaction entre deux agents, d'après de Boer (2000).

Le retour non verbal est utilisé pour établir des scores pour chacune des voyelles des répertoires des agents. Ceux-ci sont utilisés pour promouvoir les « bons prototypes » et éliminer les « mauvais ». Si le jeu d'imitation est un échec (P1 différent de P2), et en fonction du score du prototype utilisé par l'interlocuteur, celui-ci est soit modifié pour ressembler plus au son prononcé par le locuteur, soit un nouveau prototype est créé. Un processus de fusion est également utilisé pour éviter les prototypes trop proches l'un de l'autre. Le nombre de prototypes de voyelles que possèdent les agents n'est donc pas fixé à l'avance, au contraire des travaux de Berrah et Laboissière (1999), mais émerge des simulations. Ainsi, les prédictions obtenues par ce modèle ne concernent plus seulement la disposition des prototypes dans l'espace acoustique mais également leur nombre. Ces prédictions sont bonnes, avec notamment une distribution de la taille des systèmes de voyelles qui se superpose assez bien avec celles des langages humains (avec un pic autour des systèmes à cinq voyelles).

4.3.3 Les cartes neurales couplées de Oudeyer (2004, 2005)

Oudeyer (2004) propose « un système artificiel qui permet de conceptualiser la manière dont une société d'agents, dotés de conduits vocaux et d'oreilles reliés par des réseaux neuronaux, peut former par auto-organisation un code de la parole discret, combinatorial et partagé par tous les agents, sans que l'on présuppose de capacité linguistique ou de capacité de coordination sociale ». L'architecture du système artificiel est décrite Figure 4.10. Celui-ci est constitué d'une oreille artificielle (modèle de la cochlée) qui envoie des impulsions nerveuses à une carte de neurones perceptuels, entièrement interconnectée à

une carte de neurones moteurs, eux-mêmes reliés à un conduit vocal artificiel. La fonction d'activation des neurones est une gaussienne, leur fonctionnement s'apparente à ceux d'une carte auto-organisatrice de Kohonen (Kohonen, 1990). Les neurones moteurs peuvent être activés soit par les neurones de la carte perceptuelle, soit spontanément.

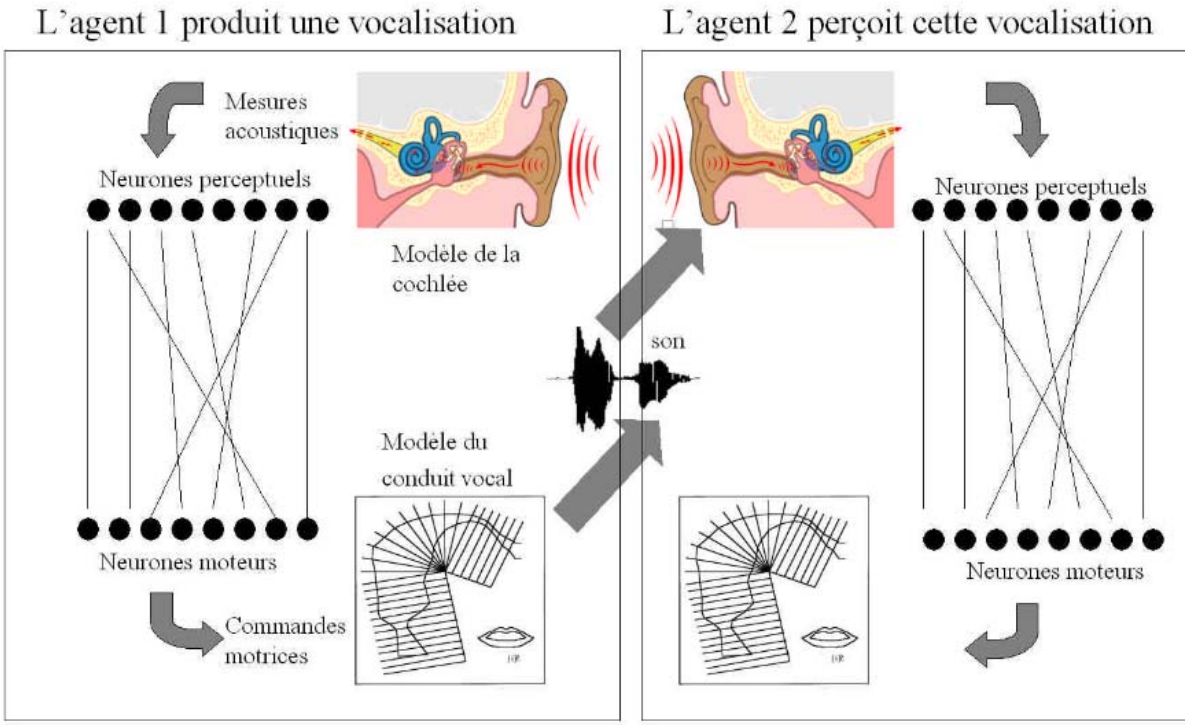


FIGURE 4.10 – Architecture du système, d'après Oudeyer (2004). Voir texte pour détail.

Les agents sont alors plongés dans un environnement virtuel dans lequel ils se déplacent aléatoirement. À certains moments, un agent produit une vocalisation, l'agent le plus proche la perçoit et ces deux agents mettent à jour leurs cartes neurales. La mise à jour des poids des connexions des neurones perceptuels a lieu à chaque fois que les neurones sont sollicités. Ils sont modifiés de manière à ce que les neurones deviennent plus sensibles au stimulus qui les a activés. La mise à jour des neurones moteurs distingue deux cas.

- Si la vocalisation a été produite par l'agent lui-même (c'est-à-dire si les neurones moteurs sont déjà activés au moment où les neurones perceptuels le sont), les poids des connexions entre neurones perceptuels et moteurs sont renforcés si elles relient des neurones dont les activités sont corrélées, affaiblis dans le cas contraire (loi d'apprentissage hebbien). Ceci permet aux agents d'apprendre les correspondances entre stimuli acoustiques et commandes motrices.
- Si la vocalisation a été produite par un autre agent, les poids de connexions entre les deux cartes neurales ne sont pas modifiés. Par contre, l'activation des neurones perceptuels est propagée aux neurones moteurs par les connexions et les poids des connexions entre les cartes motrices et le contrôleur du conduit vocal sont modifiés de

manière à ce que les poids de connexions de sortie de chaque neurone se rapprochent des poids de connexions de sortie du neurone le plus activé.

Cette architecture implique donc un couplage entre les processus de perception et les processus de production. Si un agent entend certains sons plus souvent que les autres, il aura alors tendance à produire ces sons plus souvent que les autres. L'auteur remarque que « ce processus [...] n'est pas réalisé au travers d'une imitation mais est un effet de bord de l'augmentation de la sensitivité des neurones aux stimuli et aux transferts des activations entre les cartes perceptuelles, ce qui est un mécanisme neural générique de très bas niveau ».

En utilisant un synthétiseur articulatoire qui fait correspondre à un point d'un espace articulatoire à trois dimensions (hauteur du corps de la langue, position du corps de la langue, et arrondissement des lèvres), un point dans un espace acoustique à deux dimensions (le premier formant et le second formant effectif), les simulations permettent une prédiction correcte des systèmes de voyelles dans les langues humaines.

L'objectif des travaux de Pierre-Yves Oudeyer est de fournir un système artificiel capable de prédire les tendances universelles des systèmes de voyelles dans les langues du monde dans lequel le nombre d'hypothèses formulées en ce qui concerne les capacités des agents est minimal. Il souhaite ainsi montrer la robustesse du phénomène et sa capacité à émerger en dehors de toute capacité prélangagière.

4.4 Discussion

4.4.1 Tentative d'unification des théories

Les théories Frame/Content et Vocalize-to-Localize sont relativement complémentaires et peuvent s'intégrer dans un scénario cohérent d'émergence du langage à travers la rencontre du cadre de la parole et du cadre du signe (Ducey-Kaufmann, 2007). La première fournit un scénario d'émergence du contrôle du conduit vocal vers la phonologie par la cyclicité mandibulaire, la seconde fournit une référence sur le monde extérieur par la déixis, et la synchronisation des deux constitue un système plausible de précurseur du langage alliant sens et vocalisations.

Frame/Content et l'hypothèse du système miroir quant à elles semblent plus contradictoires. Elles sont en effet concurrentes dans le vif débat entre théories vocales vs. gestuelles de la parole (voir MacNeilage (1998) vs. Rizzolatti et Arbib (1998), et Arbib (2005b)) : le langage a-t-il évolué directement dans la modalité vocale, ou a-t-il d'abord transité par un langage évolué gestuel (brachio-manuel) ?

La première position propose une émergence précoce de la parole à partir des processus d'ingestion puis des mimiques orofaciales de l'ancêtre commun des humains et des singes, qui fournit d'une façon simple et cohérente un système de modulation acoustique robuste par l'alternance de positions ouvertes (voyelles) et fermées (consonnes) du conduit vocal. Elle reste toutefois obscure sur la question de la référence. Sur cette question, l'hypothèse du système miroir est plus explicite : la reconnaissance d'actions brachio-manuelles à travers

le système de neurones miroirs fournit la base d'une communication plus poussée faisant intervenir une phonologie et une syntaxe gestuelles. Le problème devient alors le passage de la communication gestuelle à la communication vocale, qui peut faire apparaître ce chemin évolutif compliqué alors qu'il semble que cette dernière soit déjà présente à l'état primitif chez certains singes (vervets notamment (Cheney et Seyfarth, 1982, 1992)).

Toutefois, Arbib (2005b) propose une transition plus progressive, dans ce qu'il compare à une spirale en expansion (*expanding spiral*) :

> *« Our distant ancestor (*Homo habilis *through to early* Homo sapiens*) had "protosign" which [...] provided essential scaffolding for the emergence of "protospeech", but biological and cultural evolution along the hominid line saw advances in both protosign and protospeech feeding off each other in an expanding spiral so that [...] protosign did not attain the status of a full language prior to the emergence of early forms of protospeech. »*

Vu ainsi, le débat[1] repose moins sur la question « qui de la communication vocale ou gestuelle a émergé en premier ? » que sur la question « comment l'interaction entre communication vocale et gestuelle a pu permettre l'émergence du langage humain ? ». Dans une démarche d'unification, c'est certainement sur ce terrain que les trois théories peuvent apporter conjointement leurs contributions. L'hypothèse du système miroir fournit un cadre cohérent d'évolution de la communication gestuelle à travers le système miroir, Frame/Content fournit celui de la communication vocale à travers le cycle mandibulaire, et Vocalize-to-Localize propose un geste particulier comme bootstrap de la co-évolution entre geste et parole : le geste déictique de pointer.

D'une certaine manière (mais sans en faire part), Tomasello et collab. (2005) affinent l'hypothèse de système miroir en ajoutant à la capacité de comprendre des actions intentionnelles, une motivation à partager des états psychologiques (la première semble en effet présente chez les grands singes de façon élaborée, sans pour autant que ceux-ci aient accès au langage). Toutefois, un système miroir capable de s'activer similairement dans l'action ou dans la perception pourrait être une base neuronale intéressante pour l'émergence d'une capacité à partager des états psychologiques (ou du moins mentaux).

En résumé, et en référence à notre architecture d'agent communicant de la Figure 3.13, nous pouvons récupérer des différents éléments théoriques que nous avons présentés, les ingrédients suivants.

- Sur la question de la référence, ou de la capacité à associer à des objets O des actions et des signaux communicatifs M et S, l'hypothèse du système miroir fournit un scénario moteur complexe, avec une « grammaire d'action », à laquelle la théorie Vocalize-to-Localize fournit un possible bootstrap déictique.
- Sur la question du média de communication et de la production de signaux, si le chemin naturel de l'hypothèse de système miroir est le chemin gestuel, Vocalize-to-Localize propose une « voie orofaciale », complétée et développée par la théorie Frame/Content, qui fournit de puissants mécanismes de combinatoire motrice ; la

1. On trouvera dans l'ouvrage (Vilain et collab., 2011) (auquel nous avons participé (Moulin-Frier et collab., 2011)) une version récente du débat entre théories vocales et gestuelles.

« spirale en expansion » de l'hypothèse du système miroir proposant une synthèse oro-gestuelle convaincante.
- Enfin, sur la question de l'internalisation de la boucle de communication, des éléments nous fournissent une assise théorique pour poursuivre dans le cadre de l'architecture cognitive de la Figure 3.13, du lien sensori-moteur des neurones miroirs à l'émulation intégrale du cerveau de l'autre dans une théorie de l'esprit ou une théorie de l'intention partagée.

Dans la partie II de ce document, nous proposerons une modélisation d'agents prélangagiers doués d'un comportement de déixis très simplifié. Cette modélisation permettra également, dans la partie III, d'intégrer un cycle de mâchoire préexistant aux agents, sous la forme de contraintes motrices. L'extension à un système de référence plus complet, impliquant par exemple un cadre action-objet n'a pas été modélisé jusqu'à aujourd'hui mais pourrait fournir une amorce intéressante à l'émergence de la structure verbe-objet dans la syntaxe, comme nous le discuterons en conclusion générale.

4.4.2 Analyse des modèles

Si le modèle de Berrah et Laboissière (1999) a l'avantage d'être le premier à montrer comment la phonologie, limitée aux systèmes de voyelles, peut émerger d'interactions locales entre agents, il possède les inconvénients suivants.
- Le système cognitif sous-jacent est basé sur des processus d'attraction-répulsion dans l'espace auditif de l'agent, ce qui est difficilement reliable à une architecture cognitive plausible et fait finalement apparaître ce modèle plus comme une version multi-agents de la théorie de la dispersion que comme un modèle plausible d'émergence du langage par interactions locales.
- Il n'opère que sur l'espace auditif des agents, laissant de côté le rôle des contraintes motrices dans la phylogenèse du langage.

Le modèle de de Boer (2000) quant à lui a l'avantage de proposer un candidat plausible comme comportement prélangagier à l'origine de l'émergence du langage : l'imitation. De plus il permet de laisser le nombre de voyelles émerger des simulations. Toutefois :
- il est basé sur des interactions relativement compliquées nécessitant un retour non verbal : l'initiateur produit, l'imitateur catégorise et imite, l'initiateur classe, et les agents sont informés du succès ou de l'échec de l'interaction ;
- Le rôle de la motricité est également plutôt occulté.

Enfin, le modèle de Oudeyer (2005) propose un couplage perception-action intéressant par l'interconnexion de cartes neurales perceptives et motrices et montre la robustesse du phénomène d'émergence des voyelles en limitant au maximum le nombre d'hypothèses concernant les capacités prélangagières des agents. Mais le scénario d'interaction est une fois encore peu plausible d'un point de vue phylogénétique : des agents se déplacent dans un environnement en produisant des vocalisations, rien n'est dit sur la fonction de ces vocalisations ou sur leur association avec une référence.

D'une manière générale, bien que chacun de ces modèles ait apporté une contribution majeure à ce domaine de recherche, nous pensons qu'il est possible d'aller maintenant

plus loin dans la prise en compte de la multitude de travaux expérimentaux et théoriques concernant les systèmes de production et de perception de la parole décrits au Chapitre 3, ainsi que les scénarios d'évolution vers l'émergence du langage décrits dans le présent chapitre (Section 4.2). Cette démarche est pour nous une étape préliminaire essentielle à nos travaux de modélisation, pour lesquels nous souhaitons dégager les grandes tendances des modèles et théories existants, pour les ancrer dans des hypothèses plausibles et argumentées que nous synthétiserons en conclusion de cette partie (Chapitre 5).

4.4.3 Conclusion

Dans le cadre de notre démarche de modélisation, nous retiendrons de la Section 4.2 la proposition de la théorie Vocalize-to-Localize de couplage d'un système de monstration (associé à un comportement prélangagier de déixis pour localiser et porter une attention partagée sur un objet, et pouvant être représenté dans notre modèle par la variable C assurant le lien entre O_S et O_L) et d'un système de vocalisation (hérité de l'alternance de positions ouvertes et fermées de la mâchoire provenant de mécanismes de mastication et d'ingestion, et représenté dans notre modèle d'agent par la variable M). Nous retiendrons de l'hypothèse du système miroir une équivalence des mécanismes de production et de perception, représentée dans notre modèle par l'existence d'un lien sensori-moteur entre variables motrices (M) et sensorielles (S), et qui sera rendu explicite dans la formalisation des comportements de production et de perception des agents au Chapitre 8. La proposition de Tomasello et collab. (2005), de la nécessité d'un partage d'états psychologiques entre agents pour établir des plans d'action communs, se traduit dans notre modèle d'agent communicant par l'existence de deux variables représentant les objets, O_S et O_L, permettant aux agents de se représenter à la fois leur propre état mental ainsi que celui de leur interlocuteur.

Nous proposons alors un paradigme d'interaction inspiré de la notion de « jeux de langage » de Steels (1996), dans lequel des agents sensori-moteurs, instances d'un des modèles conceptuels de la Figure 3.14, procèdent à ce que nous appelons des « jeux déictiques » (Moulin-Frier et collab., 2008, voir Figure 4.11). Un jeu déictique consiste en deux agents (généralement tirés au hasard parmi une population plus grande) se retrouvant devant un objet donné sur lequel ils portent une attention partagée. Cette attention partagée est supposée être assurée par un comportement pré-existant de déixis, permettant aux deux agents d'identifier le même objet de la même façon (grâce au pointage). L'un d'eux prend alors le statut de locuteur, et propose un geste moteur pour nommer l'objet, en fonction de ses connaissances actuelles (exprimées dans les relations entre ses variables O_S, M, S et O_L). L'autre prend le statut d'auditeur, et reçoit l'entrée auditive correspondant au geste produit. Les deux agents mettent alors à jour leurs connaissances en fonction de cette interaction (c'est-à-dire en fonction du geste produit et de l'objet pour le locuteur, et de l'entrée auditive et de l'objet pour l'auditeur). Ces jeux déictiques peuvent se répéter à l'infini, chaque agent pouvant prendre tour à tour le statut de locuteur ou d'auditeur, faisant ainsi évoluer leurs connaissances respectives. L'évolution des gestes et des sons produits par les agents pour chaque objet peut alors être analysée et comparée aux régularités

connues dans les systèmes phonologiques des langues humaines.

La Figure 4.11 est donc une synthèse des Figures 2.10 et 3.14 de la fin des deux chapitres précédents, dans laquelle nous représentons à la fois la situation de communication (augmentée d'une amorce déictique inspirée de la théorie Vocalize-to-Localize décrite dans ce chapitre et représentée par la pomme sur la figure) et son internalisation dans l'architecture cognitive des agents (dans laquelle la variable C permet une parité entre action et perception, ou entre soi et l'autre, telle que décrite dans l'hypothèse du système miroir ou dans la notion de théorie de l'esprit). Elle synthétise ainsi nos deux hypothèses centrales d'internalisation et de communication prélangagière, à partir desquelles nous voulons ancrer les théories de la forme dans les théories de l'émergence dans les deux prochaines parties.

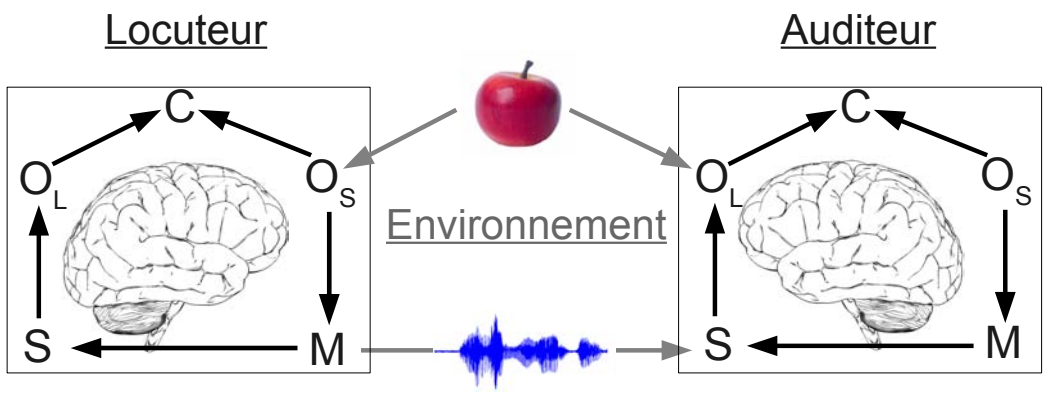


FIGURE 4.11 – Un jeu déictique entre deux agents.

Chapitre 5

Synthèse

Cette première partie constitue une revue de question de nos trois thèmes d'intérêt que sont la communication parlée (Chapitre 2), les architectures cognitives qui la sous-tendent (Chapitre 3) ainsi que les possibles conditions nécessaires à son émergence (Chapitre 4), dans le but d'en obtenir un modèle conceptuel intégrateur. C'est à partir de celui-ci que sera construite la modélisation de ces concepts que nous proposerons dans la Partie II.

Le Chapitre 2 nous a permis de définir une situation de communication parlée comme l'interaction entre un agent locuteur et un agent auditeur, dont le but est de se transmettre un objet de communication, par l'intermédiaire d'un environnement transformant les gestes articulatoires du locuteur en stimuli auditifs pour le locuteur. Nous avons brièvement exposé les corrélats neuronaux des systèmes moteur et auditif des agents, ainsi que la transformation articulatori-acoustique effectuée par l'environnement dans lequel ils évoluent.

Puis, dans le Chapitre 3, nous avons passé en revue différents modèles et théories existants de production et de perception de la parole et proposé une taxonomie en trois catégories distinctes : les théories motrices, auditives et sensori-motrices, à la fois en production et en perception. Ceci nous a permis de différencier la nature des représentations (motrices ou auditives) de l'architecture des systèmes de production et de perception, lesquels peuvent faire intervenir des représentations motrices et/ou auditives à la fois en production et en perception selon les théories considérées. Nous avons alors argumenté sur la nécessité d'un lien sensori-moteur dans l'architecture cognitive des agents, et avons exposé ses différents corrélats neuroanatomiques. Nous avons ensuite présenté une hypothèse centrale de notre travail, basée sur l'analogie entre situation de communication reliant un locuteur et un auditeur à travers un environnement effectuant une transformation articulatori-acoustique, et une architecture cognitive d'agent communicant reliant un système moteur et un système auditif à travers un lien sensori-moteur. Nous avons vu que le modèle conceptuel d'agent communicant ainsi obtenu (Figure 3.13) pouvait rendre compte des différentes théories de la production et de la perception en désactivant éventuellement la partie motrice ou auditive du modèle (Figure 3.14).

Enfin, nous nous sommes intéressés dans le Chapitre 4 à l'émergence des systèmes phonologiques à travers trois grandes approches chronologiquement successives : les théories

de la forme, les théories de l'émergence, et les modèles computationnels d'agents interagissants. Les théories de la forme tentent d'expliquer « de l'intérieur » les régularités des systèmes phonologiques des langues du monde, et constituent en quelque sorte l'objectif des simulations que nous décrirons dans la Partie III. Les théories de l'émergence tentent d'expliquer ces régularités « de l'extérieur » en proposant des comportements prélangagiers qui pourraient être à l'origine de l'émergence de la communication parlée, et constituent une source d'inspiration de nos choix de modélisation. Enfin, nous nous inscrivons dans la tendance des modèles computationnels d'agents interagissants, qui tentent de faire émerger ces régularités par la simulation d'agents communicants en interaction, méthode sur laquelle sera basée notre algorithme de simulation.

La logique de réflexion de cette première partie se résume schématiquement par les figures encadrées en fin de chaque chapitre (Figures 2.10, 3.14 et 4.11) : d'abord nous définissons une ligne de communication du locuteur à l'auditeur (Chapitre 2), puis nous proposons une internalisation de cette ligne de communication dans une architecture cognitive capable de regrouper les différents courants théoriques existants (Chapitre 3), pour finalement intégrer le tout dans un paradigme d'interaction multi-agents évolutif (Chapitre 4). Ce dernier intègre à la fois une hypothèse d'internalisation et de communication prélangagière.

Nous disposons ainsi de tous les éléments conceptuels nécessaires pour nous engager dans une modélisation formelle du problème. C'est l'objectif de la Partie II de ce document dans laquelle, après avoir présenté notre outil de modélisation au Chapitre 6, nous proposons d'abord une modélisation d'une situation de communication parlée (Chapitre 7, en miroir du Chapitre 2), puis une modélisation d'un agent communicant (Chapitre 8, en miroir du Chapitre 3), et enfin un algorithme de simulation d'une société d'agents communicants en interaction par jeux déictiques (Chapitre 9, en miroir du Chapitre 4).

Deuxième partie

Des modèles conceptuels aux modèles formels

Chapitre 6

Programmation bayésienne des robots

L'utilisation des probabilités bayésiennes en modélisation cognitive, en particulier dans le cadre de comportements sensori-moteurs, trouve sa justification dans le fait que tout modèle d'un phénomène réel est par nature incomplet (Lebeltel et collab., 2004) : il existe toujours des variables cachées non prises en compte dans le modèle qui influencent le phénomène. Ainsi, l'utilisation de la logique classique trouve rapidement ses limites dans la modélisation de comportements sensori-moteurs dans lesquels un robot doit agir et percevoir dans un environnement dont le modèle dont il dispose est forcément incomplet et incertain. Jaynes (2003) propose d'utiliser la théorie des probabilités comme une alternative à la logique pour raisonner sur des connaissances incomplètes et incertaines. Dans ce cadre, les probabilités ne sont plus conçues classiquement comme la limite de la fréquence d'occurrence d'un événement, mais plutôt comme un degré de confiance accordé à un état de connaissance. Ce sont ces travaux qui amenèrent Bessière et collab. (1999, 1998), puis Lebeltel et collab. (2004), à proposer une méthode unifiée de modélisation et de programmation de comportements sensori-moteurs : la Programmation Bayésienne des Robots (PBR). Cette interprétation subjective de la théorie des probabilités permet au programmeur de spécifier rigoureusement les connaissances incomplètes et incertaines du robot, de les affiner par apprentissage, et de les traiter par des processus automatisés d'inférence probabiliste.

L'objectif de ce chapitre est de présenter cet outil de modélisation, que nous utiliserons dans toute la suite de ce document. Cette présentation est parfois inspirée de Diard (2003). Nous commençons par la description d'un exemple simple de fusion de capteurs tiré de Lebeltel et collab. (2004) sur lequel nous illustrerons la méthode. Puis nous définissons les concepts mathématiques pour enfin présenter la méthode PBR.

6.1 Exemple : fusion de capteurs

Cet exemple est tiré d'une série d'expériences robotiques sur le robot Khepera (Lebeltel et collab., 2004). Celui-ci est doté de huit capteurs de luminosité disposés autour d'une base circulaire (six à l'avant et deux à l'arrière), comme illustré Figure 6.1. Chacun de ces

capteurs renvoie une valeur comprise entre 0 et 511 dans une relation inverse à la luminosité qu'il mesure (0 pour une luminosité maximale, 511 pour une luminosité minimale). L'objectif du robot sera alors d'estimer la position d'une source lumineuse, determinée par l'angle *Theta* entre son axe frontal et la source, compris entre les valeurs -180 ° et 180 ° .

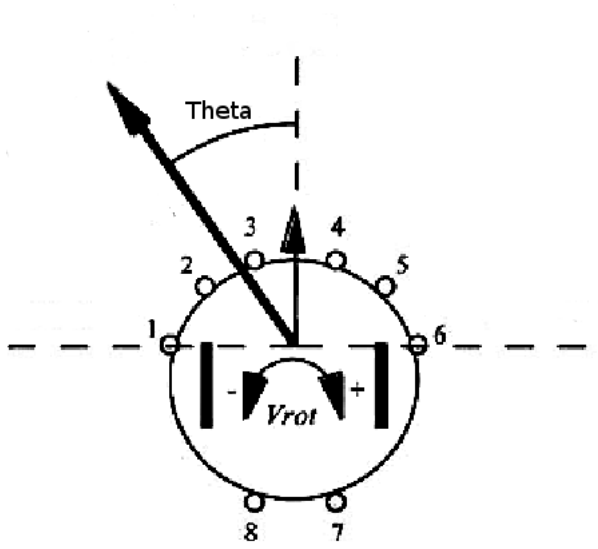


FIGURE 6.1 – Schéma du robot Khepera, adapté de Lebeltel et collab. (2004). $Vrot$ représente la vitesse de rotation du robot, nous ne l'utiliserons pas dans cet exemple.

6.2 Notions mathématiques

6.2.1 Définitions

6.2.1.1 Proposition logique

Une proposition logique $\mathcal{A}$ possède une valeur de vérité (vraie ou fausse) et peut être manipulée par les opérateurs classiques de l'algèbre booléenne. Dans le cadre de notre travail de modélisation, nous nous limitons à l'opérateur de conjonction. La conjonction de deux propositions logiques, $\mathcal{A}$ et $\mathcal{B}$, est notée $\mathcal{AB}$, correspondant à « $\mathcal{A}$ et $\mathcal{B}$ sont vraies ».

6.2.1.2 Variables

Une variable V est définie par son domaine de valeurs $\mathcal{D}_V = \{v_1, \ldots, v_k\}$ et est associée aux k propositions logiques $\mathcal{V}_1, \ldots, \mathcal{V}_k$:

$$\begin{aligned} \mathcal{V}_1 &\equiv [V = v_1] \\ \mathcal{V}_2 &\equiv [V = v_2] \\ &\vdots \\ \mathcal{V}_k &\equiv [V = v_k]. \end{aligned} \tag{6.1}$$

Chacune de ces propositions indique la valeur prise par la variable, une et une seule peut donc être vraie à un moment donné. Notons que nous nous limitons dans le cadre de ce document à des variables discrètes de domaine fini. On note k_V le cardinal du domaine de V, soit $k_V = k$ dans l'Équation 6.1.

La conjonction de deux variables V_1 et V_2 peut elle-même être représentée par une nouvelle variable $V = V_1 \wedge V_2$. Dans ce cas, V est associée aux $k_{V_1} * k_{V_2}$ propositions logiques du type :

$$\mathcal{V}_{1i}\mathcal{V}_{2j} \equiv [V_1 = v_{1i}] \wedge [V_2 = v_{2j}],$$

où $1 \leq i \leq k_{V_1}$ et $1 \leq j \leq k_{V_2}$.

6.2.2 Probabilités

La probabilité que la variable V prenne la valeur v se note $P([V = v])$, ou simplement $P(v)$ s'il n'y a pas d'ambiguïté sur la variable en question. D'une manière générale, les variables seront représentées dans ce document par des lettres majuscules, et les valeurs particulières de ces variables par des lettres minuscules.

6.2.3 Distributions

Une distribution de probabilité discrète sur V est une fonction P de $\mathcal{D}_V$ vers $[0, 1]$ de somme unité sur $\mathcal{D}_V$:

$$\sum_{v \in \mathcal{D}_V} P(v) = 1. \tag{6.2}$$

L'Équation 6.2 est également appelée règle de normalisation. Pour simplifier les notations, nous considérons l'expression suivante comme équivalente à 6.2 :

$$\sum_{V} P(V) = 1.$$

6.2.3.1 Distribution conjointe

La distribution de probabilité de la conjonction de deux variables A et B se note $P(A\ B)$.

6.2.3.2 Distribution conditionnelle

Une distribution de probabilité sur une variable peut dépendre de la valeur d'une autre variable. On parle de distribution conditionnelle et on note $P(A \mid [B = b])$, ou simplement $P(A \mid b)$, la distribution de probabilité sur la variable A sachant que la variable B a la valeur b ($b \in \mathcal{D}_B$). Plus généralement, $P(A \mid B)$ dénote une famille de distributions sur A conditionnée par B (une distribution sur A par valeur de B, soient k_B distributions sur A).

6.2.4 Règles de calcul

L'inférence bayésienne est un processus mathématique de raisonnement sur des distributions de probabilités, basé sur quelques règles simples issues de la théorie des probabilités.

6.2.4.1 Règle du produit

La règle du produit permet de décomposer une distribution conjointe en un produit de distributions plus élémentaires :

$$\begin{aligned} P(A\ B) &= P(A)P(B \mid A) \\ &= P(B)P(A \mid B). \end{aligned} \tag{6.3}$$

La symétrie de cette règle vient de la propriété de commutativité de la conjonction de deux variables. Il en découle le théorème de Bayes :

$$P(B \mid A) = \frac{P(B)}{P(A)} P(A \mid B).$$

6.2.4.2 Règle de marginalisation

De la règle de normalisation 6.2 et de la règle du produit 6.3, on dérive la règle de marginalisation :

$$\sum_B P(A\ B) = P(A). \tag{6.4}$$

Les deux règles 6.3 et 6.4 permettent d'une part de décomposer une distribution conjointe sur un ensemble de variables en un produit de distributions plus simples (par 6.3), d'autre part de calculer toute distribution conditionnelle sur ces variables à partir de la distribution conjointe (par 6.3 et 6.4).

6.3 Programmation bayésienne des robots (PBR)

La PBR permet d'écrire des programmes robotiques en deux phases. La première est une phase déclarative, dans laquelle le programmeur spécifie les connaissances préalables du robot permettant d'exprimer la distribution de probabilité conjointe sur l'ensemble

des variables d'intérêt. La PBR permet la représentation formelle de ces connaissances préalables comme la valeur d'une variable de modèle, généralement noté π. Comme toute description bayésienne du robot est dépendante des connaissances a priori du programmeur, ne serait-ce que par son interprétation des valeurs des capteurs et des actionneurs, toute distribution de probabilité est nécessairement conditionnée par π. Si V est la conjonction des variables d'intérêt du robot, l'objectif de la phase déclarative est donc la définition de la distribution conjointe :

$$P(V \mid \pi).$$

La deuxième phase de la PBR est une phase procédurale, dans laquelle le robot affine ces connaissances par apprentissage et les utilise pour réaliser des comportements. Pour cela, il peut utiliser un ensemble [1] d'apprentissage δ correspondant à un ensemble de valeurs sur un sous-ensemble de ses variables d'intérêt, lui permettant d'estimer certains paramètres de la distribution conjointe $P(V \mid \pi)$ si ceux-ci sont définis en fonction de δ. On note alors :

$$P(V \mid \delta\ \pi).$$

Dans ce qui suit, nous ne notons la valeur δ en partie droite d'une distribution que si celle-ci est effectivement nécessaire dans la définition correspondante. La valeur π quant à elle a un rôle d'identifiant de modèle permettant de distinguer des distributions contenant les mêmes variables mais dont les connaissances préalables sont différentes.

6.3.1 Phase déclarative

La phase déclarative définit l'ensemble des connaissances préalables, notées π, spécifiées par le programmeur. Dans le cadre de notre exemple de fusion de capteurs, nous les noterons π_{fusion}.

6.3.1.1 Définition des variables d'intérêt

Il s'agit d'identifier les variables pertinentes $V = V_1 \wedge \ldots \wedge V_N$ pour le robot, en particulier les variables motrices, les variables sensorielles et les variables internes, puis de spécifier leurs domaines respectifs $\mathcal{D}_{V_1}, \ldots, \mathcal{D}_{V_N}$.

Dans le cadre de notre exemple, nous noterons L_i la variable sensorielle représentant le capteur i ($1 \leq i \leq 8$), et $Theta$ la variable représentant l'angle de la source lumineuse par rapport au robot. Nous choisissons alors :

$$\begin{aligned} \forall 1 \leq i \leq 8 : \mathcal{D}_{L_i} &= \{0, 1 \ldots, 511\}, \\ k_{L_i} &= 512, \\ \mathcal{D}_{Theta} &= \{-170, -160 \ldots, 180\}, \\ k_{Theta} &= 36. \end{aligned}$$

1. Le terme ensemble peut paraître abusif car δ peut comporter plusieurs valeurs d'apprentissage égales qui ne doivent pas pour autant être confondues. Il suffit dans ce cas de considérer qu'un numéro unique est associé chaque valeur d'apprentissage de l'ensemble (par exemple l'indice de temps à laquelle elle a été enregistrée).

Ainsi, L_i correspond à des valeurs entières dans $[0, 511]$ et $Theta$ à des valeurs multiples de 10 dans $[-170, 180]$. L'association entre une variable et un ensemble de propositions logiques permet d'en spécifier une sémantique. Ici par exemple :

$$\mathcal{L}_{ij} \equiv [L_i = j] \equiv \text{« Le capteur } L_i \text{ renvoie la valeur } j \text{ »}.$$

6.3.1.2 Décomposition de la distribution conjointe et hypothèses d'indépendance conditionnelle

Il s'agit ici de décomposer la distribution conjointe sur l'ensemble des variables d'intérêt du robot en un produit de distributions plus simples grâce à la règle du produit 6.3 :

$$P(V_1 \ldots V_N \mid \pi) = P(V_1 \mid \pi)\ P(V_2 \mid V_1\ \pi)\ P(V_3 \mid V_1\ V_2\ \pi)\ \ldots\ P(V_N \mid V_1 \ldots V_{N-1}\ \pi). \quad (6.5)$$

On remarque que la symétrie de la règle 6.3 engendre un grand nombre de décompositions possibles (par commutativité de la conjonction de variables $V_1 \wedge \ldots \wedge V_N$, celles-ci sont interchangeable dans l'équation 6.5). Une fois une décomposition particulière choisie, l'Équation 6.5 peut se simplifier en ajoutant aux connaissances préalables π des hypothèses d'indépendance conditionnelle. On dit qu'une variable A est indépendante d'une variable C conditionnellement à une variable B si l'on a :

$$P(A \mid B\ C) = P(A \mid B). \quad (6.6)$$

La définition d'hypothèses d'indépendance conditionnelle par le programmeur peut alors grandement simplifier 6.5. Considérons notre exemple illustratif pour lequel les variables d'intérêt sont $Theta, L_1, \ldots, L_8$. Une décomposition possible de la distribution de probabilité conjointe est :

$$\begin{aligned} P(Theta\ L_1 \ldots L_8 \mid \pi_{fusion}) \quad = \quad & P(Theta \mid \pi_{fusion}) \\ & P(L_1 \mid Theta\ \pi_{fusion}) \\ & P(L_2 \mid Theta\ L_1\ \pi_{fusion}) \\ & \ldots \\ & P(L_8 \mid Theta\ L_1 \ldots L_7\ \pi_{fusion}). \end{aligned} \quad (6.7)$$

Ensuite, l'hypothèse d'indépendance conditionnelle qui peut être faite ici est la suivante : connaissant la valeur de $Theta$, les valeurs prises par les différents capteurs sont indépendantes. L'idée sous-jacente est que la position de la source de lumière (représentée par $Theta$) est le facteur principal qui influence les mesures des capteurs (les autres sont considérés négligeables et sont absents du modèle). Ainsi, $Theta$ est considérée comme la cause des mesures et connaissant la cause, les conséquences sont indépendantes. On parle de fusion « naïve » de capteurs.

Cette hypothèse d'indépendance conditionnelle permet alors de simplifier 6.7 de façon drastique en :

$$P(Theta\ L_1 \ldots L_8 \mid \pi_{fusion}) = P(Theta \mid \pi_{fusion}) \prod_{i=1}^{8} P(L_i \mid Theta\ \pi_{fusion}). \quad (6.8)$$

6.3.1.3 Formes paramétriques

Une fois choisie une décomposition pour la distribution conjointe sur l'ensemble des variables, il s'agit de spécifier le type de distribution utilisé pour chacun de ses termes. Nous spécifions ici les principales distributions utilisées dans ce document.

Loi uniforme : il s'agit du cas où chaque valeur d'une variable est équiprobable. La loi uniforme sur la variable V de cardinal k_V se note $P(V \mid \pi) = \boldsymbol{U}(V)$ et l'on a alors :

$$\forall v_i \in \mathcal{D}_V, P([V = v_i] \mid \pi) = \frac{1}{k_V}.$$

Loi Dirac : il s'agit du cas où seule une valeur particulière n d'une variable est possible. La loi Dirac sur la variable V se note $P(V \mid \pi) = \boldsymbol{\delta}_n(V)$ et l'on a alors :

$$\forall v_i \in \mathcal{D}_V, P([V = v_i] \mid \pi) = \begin{cases} 1 \text{ si } v_i = n, \\ 0 \text{ sinon.} \end{cases}$$

Il s'agit donc d'une distribution à un paramètre libre : n. Attention toutefois à ne pas confondre la loi Dirac de paramètre n, notée $\boldsymbol{\delta}_n(V)$ (en gras), et un ensemble d'apprentissage, noté δ.

Loi normale ou gaussienne : il s'agit du cas où la répartition des valeurs que peut prendre une variable est résumée par leur moyenne et leur écart-type. Dans le cadre de ce document où nous nous limitons aux variables de domaine discret et fini, nous utiliserons une approximation de la loi gaussienne sur les domaines discrets et bornés à valeurs entières, notée $P(V \mid \pi) = \boldsymbol{G}_{\mu,\sigma}(V)$ et définie par :

$$\forall v_i \in \mathcal{D}_V, P([V = v_i] \mid \pi) = \frac{1}{Z} \int_{v_i - 0.5}^{v_i + 0.5} e^{\frac{(v-\mu)^2}{2\sigma^2}} dv$$

où Z est une constante de normalisation assurant la somme unité sur $\mathcal{D}_V$ (Équation 6.2). Il s'agit donc d'une distribution possédant deux paramètres libres : μ et σ. Dans le cas où ces paramètres dépendent d'un ensemble d'apprentissage δ, on note alors $P(V \mid \delta\ \pi) = \boldsymbol{G}_{\delta}(V) = \boldsymbol{G}_{\mu(\delta),\sigma(\delta)}(V)$ avec :

$$\begin{aligned} \mu(\delta) &= \frac{\sum_{v_i \in \mathcal{D}_V} n_i v_i}{N}, \\ \sigma(\delta) &= \sqrt{\frac{\sum_{v_i \in \mathcal{D}_V} n_i (v_i - \mu(\delta))^2}{N}}, \end{aligned}$$

où n_i est le nombre de fois où $[V = v_i]$ dans l'ensemble d'apprentissage δ et $N = \sum_{i=1}^{k_V} n_i$.

Loi histogramme : il s'agit du cas où la probabilité qu'une variable V prenne la valeur v_i correspond à la fréquence d'occurrence de v_i dans un ensemble de données d'apprentissage δ. La loi histogramme sur la variable V se note $P(V \mid \delta\ \pi) = \boldsymbol{H}_\delta(V) = \boldsymbol{H}_{n_1,\ldots,n_{k_V}}(V)$ et l'on a :

$$\forall v_i \in \mathcal{D}_V, P([V = v_i] \mid \delta\ \pi) = \frac{n_i}{N}$$

où n_i est le nombre de fois où $[V = v_i]$ dans l'ensemble d'apprentissage δ et $N = \sum_{i=1}^{k_V} n_i$. Il s'agit donc d'une distribution possédant k_V paramètres, $n_1, \ldots, n_{k_V}$, dont $k_V - 1$ sont libres.

Loi de succession de Laplace : il s'agit d'un cas proche de le la loi histogramme, mais où l'on considère que chaque valeur v_i d'une variable V a été observée une fois avant l'arrivée des données d'apprentissage δ. L'avantage principal est que la distribution qui en résulte existe même si aucune donnée d'apprentissage n'est présente (et correspond alors à une loi uniforme). Ainsi aucune valeur ne peut avoir une probabilité nulle, évitant de considérer comme impossible une valeur qui n'a jamais été observée. La loi de succession de Laplace sur la variable V se note $P(V \mid \delta\ \pi) = \boldsymbol{L}_\delta(V) = \boldsymbol{L}_{n_1,\ldots,n_{k_V}}(V)$ et l'on a :

$$\forall v_i \in \mathcal{D}_V, P([V = v_i] \mid \delta\ \pi) = \frac{1 + n_i}{k_V + N}$$

où n_i et N sont définis en fonction de δ de la même façon que pour la loi histogramme. Il s'agit donc d'une distribution possédant k_V paramètres, $n_1, \ldots, n_{k_V}$, dont $k_V - 1$ sont libres.

Appel à sous-programme : il s'agit du cas où une distribution sur une variable V dépend d'un autre programme bayésien capable de la calculer. L'appel à sous-programme sur la variable V se note $P(V \mid \pi) = P(V \mid \pi')$, où π' sont les connaissances préalables du sous-programme considéré, et l'on a alors :

$$\forall v_i \in \mathcal{D}_V, P([V = v_i] \mid \pi) = P([V = v_i] \mid \pi').$$

Les paramètres de $P([V = v_i] \mid \pi)$ sont donc ceux de $P([V = v_i] \mid \pi')$.

Chacune de ces définitions s'étend au cas de distributions conditionnelles de type $P(A \mid B)$ pour lesquelles il convient de définir une forme paramétrique sur A pour chaque valeur de $\mathcal{D}_B$. Généralement la forme paramétrique reste identique, mais ses paramètres sont exprimés en fonction de la valeur de B. Dans le cas de distributions conditionnelles dépendant d'un ensemble d'apprentissage δ, on a :

$$P(A \mid [B = b]\ \delta\ \pi) = \boldsymbol{X}_{\delta^b}(A), \tag{6.9}$$

où $\boldsymbol{X} \in \{\boldsymbol{G}, \boldsymbol{H}, \boldsymbol{L}\}$ (lois normale, histogramme et Laplace, respectivement) et δ^b est l'ensemble d'apprentissage δ restreint aux éléments pour lesquels $[B = b]$.

Revenons à notre exemple illustratif dont la distribution conjointe,

$$P(Theta\ L_1 \ldots L_8 \mid \pi_{fusion}),$$

est donnée par l'Équation 6.8, afin de définir les formes paramétriques des termes de la décompostion. Nous considérons que le robot n'a pas d'a priori sur la position de la source de lumière et nous pouvons donc considérer le terme $P(Theta \mid \pi_{fusion})$ comme une distribution uniforme :

$$P(Theta \mid \pi_{fusion}) = \boldsymbol{U}(Theta). \tag{6.10}$$

Concernant chacun des termes $P(L_i \mid Theta\ \pi_{fusion})$, nous considérons que la valeur moyenne de L_i pour une valeur de $Theta$ donnée est spécifiée par une fonction $K(Theta, \theta_i)$ connue et présentée Figure 6.2. Nous choisissons alors une loi gaussienne centrée sur

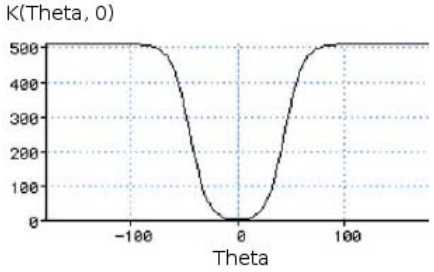


FIGURE 6.2 – $K(Theta, 0)$, d'après Lebeltel et collab. (2004). Le second paramètre de K, θ_i, correspond à l'angle du capteur par rapport à l'axe frontal du robot (ici, $\theta_i = 0$). Il permet de translater cette fonction de façon à ce que son minimum soit en $Theta = \theta_i$.

$K(Theta, \theta_i)$, avec un certain écart-type σ_i représentant la variabilité des valeurs de L_i pour un $Theta$ donné :

$$P(L_i \mid Theta\ \pi_{fusion}) = \boldsymbol{G}_{K(Theta,\theta_i),\sigma_i}(L_i).$$

6.3.2 Phase procédurale

Une fois la distribution conjointe sur l'ensemble des variables d'intérêt spécifiée en phase déclarative, $P(V \mid \pi)$, celle-ci est utilisée en phase procédurale pour définir les comportements du robot.

6.3.2.1 Apprentissage

Les paramètres des termes de la décomposition peuvent être appris à partir de jeux de données fournis au robot. Par exemple, au lieu d'être spécifiés par le programmeur comme proposé précédemment, les moyennes et écarts-types des termes $P(L_i \mid Theta\ \pi_{fusion})$ peuvent être appris à partir de données d'apprentissage δ constituées par exemple d'un ensemble de $T + 1$ triplets $\{(t, L_i(t), Theta(t)) \mid t \in \{0, \ldots, T\}\}$. Cet ensemble d'apprentissage peut provenir d'un programme professeur ou d'une téléopération du robot, capable de mesurer par un moyen quelconque les valeurs de $Theta$ au cours du temps et de les associer aux valeurs correspondantes de L_i. On aurait alors :

$$P(L_i \mid [Theta = \theta]\ \delta\ \pi_{fusion}) = \boldsymbol{G}_{\delta^\theta}(L_i),$$

où δ^θ est la restriction de δ aux éléments pour lesquels $Theta = \theta$ (Équation 6.9). Pour une valeur θ donnée, la loi normale $\boldsymbol{G}_{\delta^\theta}(L_i)$ a alors pour paramètres la moyenne et l'écart-type des valeurs de L_i associées à θ dans l'ensemble d'apprentissage δ.

6.3.2.2 Inférence bayésienne

Une fois les connaissances préalables π spécifiées, et si besoin les paramètres des termes de la décomposition appris au moyen d'un ensemble d'apprentissage δ, la dernière étape de la PBR permet de les utiliser pour définir les comportements du robot grâce à l'inférence bayésienne. On parle de question au programme bayésien : « connaissant la valeur de certaines variables du modèle, quelle est la distribution de probabilité correspondante sur un ensemble d'autres variables ? ». Typiquement, sur notre exemple illustratif, on se pose la question de la distribution sur $Theta$ connaissant une valeur particulière $l_1 \dots l_8$ de la conjonction de variables $L_1 \dots L_8$:

$$P(Theta \mid [L_1 = l_1] \dots [L_8 = l_8]\ \pi_{fusion}),$$

ou, en notation abrégée :

$$P(Theta \mid l_1 \dots l_8\ \pi_{fusion}). \tag{6.11}$$

Plus généralement, une question à un programme bayésien de variables V consiste à calculer une distribution de probabilité sur un sous-ensemble Se de V, connaissant les valeurs des variables d'un autre sous-ensemble Kn de V disjoint de Se. Se (pour *Searched*) représente les variables cherchées et Kn (pour *Known*) les variables connues. Les variables de V n'appartenant ni à Se ni à Kn forment un troisième sous-ensemble Fr (pour *Free*) tel que :

$$V = Se \wedge Kn \wedge Fr, \text{ avec } Se, Kn, Fr \text{ disjoints deux à deux.} \tag{6.12}$$

D'après 6.3 et 6.4, on obtient alors :

$$\begin{aligned} P(Se \mid Kn\ \pi) &= \frac{P(Se\ Kn \mid \pi)}{P(Kn \mid \pi)} \\ &= \frac{\sum_{Fr} P(Se\ Kn\ Fr \mid \pi)}{\sum_{Fr,Se} P(Se\ Kn\ Fr \mid \pi)} \\ &= \frac{\sum_{Fr} P(V \mid \pi)}{\sum_{Fr,Se} P(V \mid \pi)}. \end{aligned} \tag{6.13}$$

L'Équation 6.13 constitue le processus d'inférence bayésienne. Notons que différentes questions peuvent être posées à la distribution conjointe $P(V \mid \pi) = P(Se\ Kn\ Fr \mid \pi)$, tant que le choix de Se, Kn et Fr vérifie l'Équation 6.12.

Sur notre exemple illustratif, nous souhaitons poser la question de l'Équation 6.11 au programme de l'Équation 6.8. Les variables cherchées, connues et libres sont donc :

$$\begin{aligned} Se &= Theta, \\ Kn &= L_1 \dots L_8, \\ Fr &= \emptyset. \end{aligned}$$

Par inférence bayésienne (Équation 6.13), nous avons alors :

$$\begin{aligned} & P(Theta \mid l_1 \dots l_8 \ \pi_{fusion}) \\ =& \frac{P(Theta \ l_1 \dots l_8 \ \pi)}{\sum_{Theta} P(Theta \ l_1 \dots l_8 \ \pi_{fusion})} && \text{(d'après 6.13)} \\ =& \frac{P(Theta \mid \pi_{fusion}) \ \prod_{i=1}^{8} P(l_i \mid Theta \ \pi_{fusion})}{\sum_{Theta} P(Theta \mid \pi_{fusion}) \ \prod_{i=1}^{8} P(l_i \mid Theta \ \pi_{fusion})} && \text{(d'après 6.8)} \\ =& \frac{\prod_{i=1}^{8} P(l_i \mid Theta \ \pi_{fusion})}{\sum_{Theta} \prod_{i=1}^{8} P(l_i \mid Theta \ \pi_{fusion})} && \text{(d'après 6.10).} \end{aligned} \tag{6.14}$$

Toutefois, différentes questions peuvent être posées au programme. Entre autres, nous pouvons calculer la distribution sur $Theta$ connaissant seulement la valeur de quelques capteurs (si les autres sont en panne par exemple). Par exemple :

$$\begin{aligned} & P(Theta \mid l_1 \ l_2 \ \pi_{fusion}) \\ &= \frac{\sum_{L_3 \dots L_8} P(Theta \ l_1 \ l_2 \ L_3 \dots L_8 \ \pi)}{\sum_{Theta \ L_3 \dots L_8} P(Theta \ l_1 \ l_2 \ L_3 \dots L_8 \ \pi_{fusion})} \\ &= \frac{P(l_1 \mid Theta \ \pi_{fusion}) \ P(l_2 \mid Theta \ \pi_{fusion})}{\sum_{Theta} P(l_1 \mid Theta \ \pi_{fusion}) \ P(l_2 \mid Theta \ \pi_{fusion})}. \end{aligned} \tag{6.15}$$

On peut également se demander si le capteur i est en panne. Pour cela, on peut calculer la probabilité de la valeur renvoyée par ce capteur connaissant celles renvoyées par les autres. Par exemple :

$$\begin{aligned} & P(l_1 \mid l_2 \ \dots l_8 \ \pi_{fusion}) \\ &= \frac{\sum_{Theta} P(Theta \ l_1 \dots l_8 \ \pi_{fusion})}{\sum_{Theta \ L_2 \dots L_8} P(Theta \ l_1 \ L_2 \dots L_8 \ \pi_{fusion})} \\ &= \frac{\sum_{Theta} \prod_{i=1}^{8} P(l_i \mid Theta \ \pi_{fusion})}{\sum_{Theta} P(l_1 \mid Theta \ \pi_{fusion})}. \end{aligned} \tag{6.16}$$

Si cette probabilité est très petite, alors la valeur l_1 n'est pas cohérente avec celles des autres capteurs $l_2 \ \dots \ l_8$. Si ceci se vérifie dans différentes situations, et si les valeurs des autres capteurs sont cohérentes entre elles, on pourra envisager une panne du premier capteur.

La Figure 6.3 montre le résultat d'une fusion de capteurs par la question de l'Équation 6.14 pour un exemple de valeurs particulières de $l_1 \ldots l_8$. Les huit figures périphériques indiquent la distribution de probabilité sur $Theta$ dans le cas où la valeur d'un seul capteur, l_i, est disponible, et correspondent donc à la question :

$$P(Theta \mid l_i \; \pi_{fusion}) = \frac{P(l_i \mid Theta \; \pi_{fusion})}{\sum_{Theta} P(l_i \mid Theta \; \pi_{fusion})}. \tag{6.17}$$

On observe que, bien que l'incertitude sur la valeur de $Theta$ soit importante pour chaque

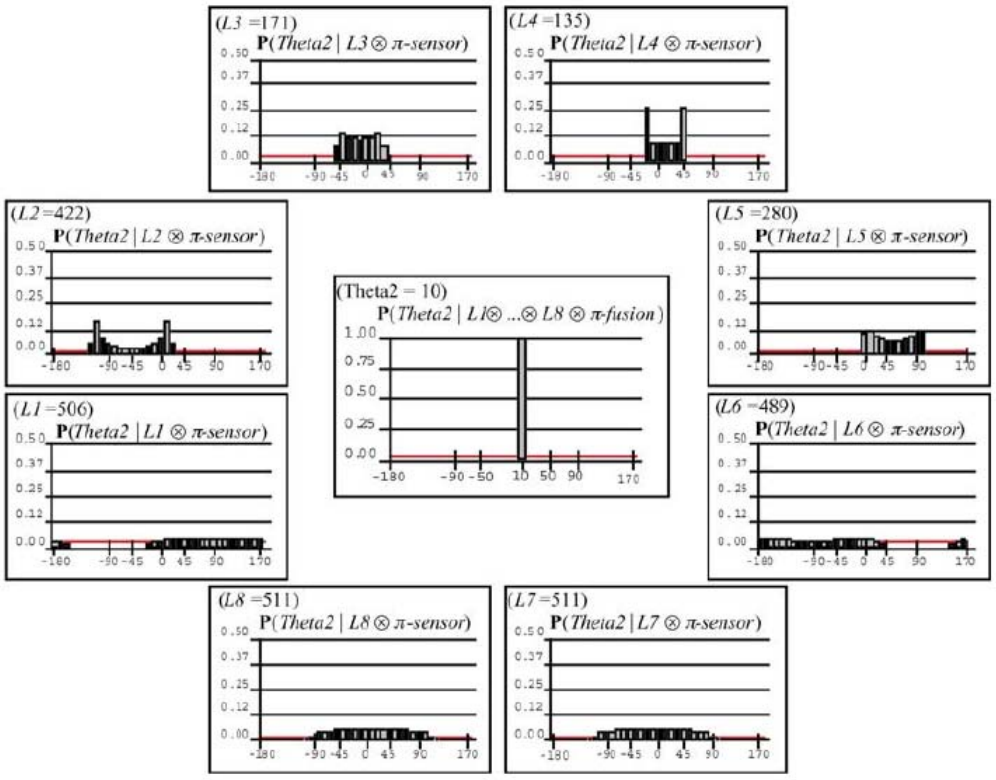


FIGURE 6.3 – Le résultat d'une fusion de capteurs. Les huit figures périphériques indiquent la distribution de probabilité sur $Theta$ dans le cas où la valeur d'un seul capteur, l_i, est disponible (Équation 6.17). La figure centrale montre le résultat de la fusion par la question de l'Équation 6.14.

capteur considéré individuellement, le processus de fusion réalisé par la question de l'Équation 6.14 permet finalement d'obtenir une quasi-certitude (ici, $Theta = 10^\circ$), et ce malgré des hypothèses d'indépendance conditionnelle fortes entre capteurs (fusion dite « naïve »).

6.3.2.3 Décision

Une fois une question posée au programme, il reste à voir comment exploiter le résultat obtenu. Généralement, le résultat d'une question est une distribution sur les variables cher-

chées Se (dans le cadre de la question 6.14 par exemple, une distribution sur $Theta$ connaissant les valeurs $l_1 \dots l_8$). On peut vouloir simplement chercher la valeur de $Theta$ la plus probable : connaissant la valeur des huit capteurs, quelle est la position de la source lumineuse la plus probable ? Il s'agit donc simplement de maximiser l'Équation 6.14 sur $Theta$. Une alternative est de réaliser un tirage selon la distribution résultante. Cette dernière option est souvent avantageuse lorsque l'on élabore des comportements sensori-moteurs dans lesquels le robot doit évoluer en continu dans son environnement, car elle autorise une plus grande variabilité des réponses pour une même valeur des variables connues et peut ainsi permettre de sortir le robot de situations bloquantes. C'est cette méthode que nous retiendrons lors de nos simulations en Partie III.

Notons que le dénominateur de l'Équation 6.13 est en fait une constante de normalisation dont le seul bénéfice est d'assurer la véracité de la règle de normalisation de l'Équation 6.2. Au lieu de le calculer explicitement, ce qui peut être coûteux, on normalise généralement le résultat d'une inférence bayésienne après coup. Quelle que soit l'option choisie pour la décision (maximum ou tirage), le résultat est indépendant de la valeur de ce dominateur et nous pourrons alors alléger l'expression de l'Équation 6.13 par :

$$P(Se \mid Kn\ \pi) \propto \sum_{Fr} P(V \mid \pi) \tag{6.18}$$

où l'opérateur « $\propto$ » signifie « est proportionnel à ». Nous pourrons également utiliser cet opérateur dans les cas où certains facteurs d'une expression correspondent à une loi uniforme.

6.4 Conclusion

Nous avons présenté dans ce chapitre l'outil de modélisation que nous utiliserons dans toute la suite de ce document : la PBR. En résumé, nous considérons que celle-ci se décompose en trois étapes, la première correspondant à la phase déclarative décrite en 6.3.1 et les deux suivantes à la phase procédurale décrite en 6.3.2 :

1. une étape de spécification des connaissances préalables π, contenant :
   - la définition des variables d'intérêt,
   - la décomposition de la distribution de probabilité conjointe sur ces variables, en un produit de termes éventuellement simplifiés par des hypothèses d'indépendance conditionnelle,
   - la définition des formes paramétriques de chacun des termes de la décomposition ;
2. une étape d'identification des paramètres des formes paramétriques, utilisant éventuellement un ensemble d'apprentissage δ ;
3. une étape d'utilisation des connaissances π et δ pour la réalisation du comportement, sous la forme d'une question probabiliste calculée par inférence bayésienne.

6.4.1 Représentation d'un programme bayésien

La définition d'un programme bayésien se résume par la Figure 6.4, appliquée ici à l'exemple de fusion de capteurs utilisé tout au long de ce chapitre. On retrouve les trois étapes ci-dessus dans la structure en accolades où :

- la spécification correspond à l'étape 1,
- la description regroupe les étapes 1 et 2,
- le programme bayésien correspond à l'ensemble des étapes 1, 2 et 3.

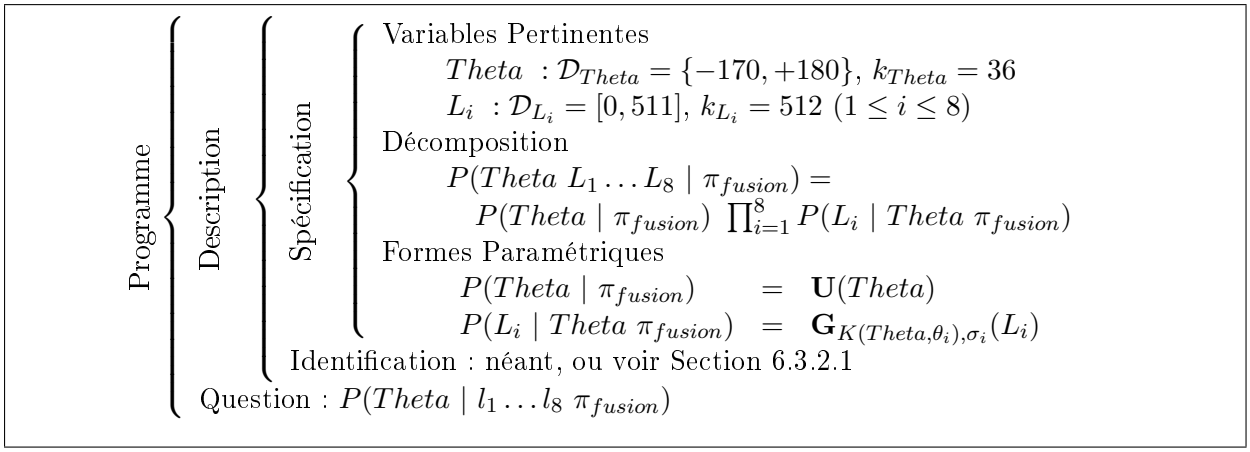


FIGURE 6.4 – Structure d'un programme PBR et exemple de fusion de capteurs.

Chapitre 7

Modélisation d'une situation de communication

Après l'interlude technique du chapitre précédent, nous exploitons dans les trois chapitres à venir toute la réflexion de la Partie I, avec la motivation du modélisateur ayant acquis un outil adéquat.

Pour en saisir correctement la logique, rappelons rapidement l'esprit de la première partie de ce document (pour un résumé plus complet, voir le Chapitre 5). Nous avons décrit une situation de communication parlée dans laquelle un agent locuteur doit transmettre à un agent auditeur un objet de communication à travers une transformation articulatori-acoustique réalisée par l'environnement (Figure 2.10, Chapitre 2). Puis nous avons proposé une conception d'agent communicant selon laquelle cette situation est internalisée sous la forme d'un système moteur et d'un système auditif reliés par un lien sensori-moteur (Figure 3.14, Chapitre 3). Enfin, nous avons réintégré ces agents communicants dans la situation de communication en proposant la notion de jeu déictique (Figure 4.11, Chapitre 4).

Les trois chapitres qui suivent reprennent le même schéma de réflexion sous l'angle de la modélisation bayésienne. La PBR présentée au chapitre précédent opère justement en trois étapes (voir conclusion du chapitre précédent), que l'on propose d'associer aux trois étapes de réflexion de la Partie I.

- Le présent chapitre constitue la phase de spécification des connaissances préalables de la situation de communication parlée définie au Chapitre 2, notées ici π_{Com} : c'est en effet cette situation qui, du point de vue du modélisateur, fournit les variables pertinentes, la décomposition de la distribution conjointe sur ces variables ainsi que leurs formes paramétriques.
- Le Chapitre 8 constitue la phase d'utilisation de ces connaissances préalables, que nous appellerons cette fois π_{Ag}, pour la modélisation d'un agent communicant réalisant des comportements de production et de perception de parole tels que définis au Chapitre 3. L'égalité $\pi_{Ag} = \pi_{Com}$ constitue la traduction formelle de l'hypothèse d'internalisation de la situation de communication dans l'architecture cognitive des agents que nous avons proposée à la Section 3.5.3. Les comportements de production

et de perception des agents sont alors exprimés à partir de ces connaissances comme des questions probabilistes, qui définissent formellement les concepts de théories motrices, auditives et sensori-motrices.
- Le Chapitre 9 constitue la phase d'identification des paramètres des distributions motrices et auditives des agents. Pour cela, nous réintégrons les agents dans la situation de communication par la définition algorithmique d'un paradigme d'interaction par jeux déictiques tel qu'il a été proposé au Chapitre 4, leur permettant de faire évoluer leurs connaissances motrices et auditives par l'accumulation de données d'apprentissage au cours de leurs interactions.

Ces trois chapitres de formalisation ne spécifieront pas le domaine des variables motrices et sensorielles et les paramètres des distributions correspondantes resteront abstraits. Leurs instanciations précises seront en effet l'objet de la Partie III et dépendront des différentes simulations considérées.

Dans ce chapitre, nous exprimons donc la situation de communication parlée décrite dans le Chapitre 2 dans le formalisme PBR que nous avons présenté au chapitre précédent. Nous spécifions pour cela les connaissances préalables, notées ici π_{Com}, fournissant les variables pertinentes de cette situation, la décomposition de la distribution conjointe sur ces variables ainsi que leurs formes paramétriques.

7.1 Rappel du modèle conceptuel d'une situation de communication parlée

Nous reprenons donc la situation décrite au Chapitre 2 composée de deux agents et d'un environnement, dont le schéma est rappelé Figure 7.1. L'agent locuteur doit produire un geste moteur (M) pour un certain objet de communication (O_S), l'environnement transforme le geste du locuteur en un stimulus auditif (S), et l'agent auditeur doit inférer un objet (O_L) à partir de ce stimulus. Finalement, la communication est un succès si et seulement si $O_L = O_S$ (condition C). L'objectif de ce chapitre est de spécifier les connaissances préalables π_{Com} de cette situation en termes de PBR.

7.2 Connaissances préalables π_{Com} de la situation de communication

Nous effectuons ici la décomposition de la distribution conjointe sur l'ensemble des variables d'intérêt $P(O_S\ M\ S\ O_L\ C \mid \pi_{Com})$, où π_{Com} représente l'ensemble des hypothèses préalables du modèle de situation de communication parlée, explicitées ci-dessous (choix et définition des variables, hypothèses d'indépendance et formes paramétriques).

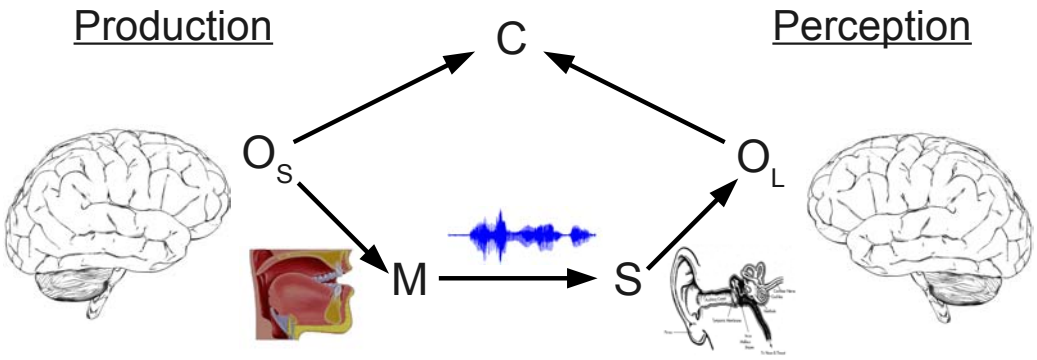


FIGURE 7.1 – Schéma de la situation de communication parlée.

7.2.1 Variables

Les cinq variables d'intérêt de la situation de communication parlée sont O_S, M, S, O_L et C (Figure 7.1).

Les variables O_S et O_L représentent les objets de communication, définis indépendamment de la nature des objets (phonétique ou sémantique par exemple, voir Chapitre 2). Nous les définissons simplement comme un ensemble de catégories. Si N_O est le nombre d'objets communicables par les agents :

$$\begin{aligned} k_{O_S} = k_{O_L} &= N_O, \\ \mathcal{D}_{O_S} = \mathcal{D}_{O_L} &= \{o_1, \ldots, o_{N_O}\}. \end{aligned}$$

Les variables M et S représentent respectivement les commandes motrices et les entrées auditives. Leurs domaines respectifs correspondent aux espaces articulatoires et auditifs des agents tels que nous les avons décrits dans le Chapitre 2. Elles sont généralement multi-dimensionnelles et leurs domaines précis seront spécifiés en Partie III. Considérons seulement pour l'instant que M correspond à un domaine articulatoire (position de la mâchoire, de la langue et des lèvres par exemple, voir Section 2.1) et que S correspond à un domaine auditif (valeurs des trois premiers formants par exemple, voir Sections 2.2 et 2.3).

La variable C représente le succès de la communication et est donc définie comme une variable booléenne :

$$\begin{aligned} k_C &= 2, \\ \mathcal{D}_C &= \{0, 1\}, \end{aligned}$$

où 0 et 1 représentent les valeurs *faux* et *vrai*, respectivement.

7.2.2 Distribution conjointe et hypothèses d'indépendance

La spécification des connaissances préalables π_{Com} de la situation de communication passe ensuite par la définition de la distribution de probabilité conjointe sur l'ensemble des variables du problème. D'après la règle du produit (Équation 6.3), nous avons :

$$\begin{aligned} P(O_S\ M\ S\ O_L\ C \mid \pi_{Com}) &= P(O_S \mid \pi_{Com})P(M \mid O_S\ \pi_{Com})P(S \mid O_S\ M\ \pi_{Com}) \\ & \quad P(O_L \mid O_S\ M\ S\ \pi_{Com})P(C \mid O_S\ M\ S\ O_L\ \pi_{Com}). \end{aligned}$$

Nous simplifions cette expression par l'ensemble d'hypothèses d'indépendance conditionnelle suivantes :

- S est indépendante de O_S conditionnellement à M. En effet, l'entrée auditive S est entièrement déterminée par la connaissance du geste moteur M produit par le locuteur :
$$P(S \mid O_S\ M\ \pi_{Com}) = P(S \mid M\ \pi_{Com}).$$
- O_L est indépendante de O_S et M conditionnellement à S. En effet, l'objet O_L inféré par l'agent auditeur dépend uniquement de l'entrée auditive S qui lui parvient :
$$P(O_L \mid O_S\ M\ S\ \pi_{Com}) = P(O_L \mid S\ \pi_{Com}).$$
- C est indépendante de M et S conditionnellement à O_S et O_L. En effet, le succès de la communication ne dépend que de la véracité de l'expression $O_L = O_L$:
$$P(C \mid O_S\ M\ S\ O_L\ \pi_{Com}) = P(C \mid O_S\ O_L\ \pi_{Com}).$$

On obtient alors la décomposition suivante :

$$\begin{aligned} P(O_S\ M\ S\ O_L\ C \mid \pi_{Com}) &= P(O_S \mid \pi_{Com})P(M \mid O_S\ \pi_{Com})P(S \mid M\ \pi_{Com}) \\ & \quad P(O_L \mid S\ \pi_{Com})P(C \mid O_S\ O_L\ \pi_{Com}). \end{aligned} \tag{7.1}$$

7.2.3 Formes paramétriques

Décrivons plus en détail chacun des termes de la décomposition de l'Équation 7.1.

$P(O_S \mid \pi_{Com})$ est la probabilité *a priori* de l'objet que l'agent locuteur veut communiquer. Nous considérons que chaque objet est équiprobable et la définissons comme une loi uniforme :

$$P(O_S \mid \pi_{Com}) = \boldsymbol{U}(O_S). \tag{7.2}$$

$P(M \mid O_S\ \pi_{Com})$ et $P(O_L \mid S\ \pi_{Com})$ représentent le système de production de l'agent locuteur et le système de perception de l'agent auditeur, respectivement. Leurs définitions s'appuient sur toute la réflexion du Chapitre 3 et sont l'objet du chapitre suivant.

$P(S \mid M\ \pi_{Com})$ représente la transformation articulatori-acoustique (voir section 2.2). Cette distribution est déterminée par la fonction permettant de calculer le stimulus auditif S produit par un geste M (déterminée par les lois de l'aéro-acoustique), par le bruit éventuellement présent dans l'environnement qui pourrait affecter S, ainsi que par l'éventuelle

discrétisation des variables M et S que pourraient nécessiter les outils de modélisation et de simulation en terme de temps de calcul. Nous spécifierons plus en détail cette distribution en Partie III.

$P(C \mid O_S\ O_L\ \pi_{Com})$ est la condition du succès de la communication, que nous définissons simplement comme l'égalité $O_S = O_L$. Elle correspond donc à une loi Dirac :

$$P(C \mid O_S\ O_L\ \pi_{Com}) = \boldsymbol{\delta}_{O_S=O_L}(C) = \begin{cases} 1 \text{ si } C = 1 \text{ et } O_S = O_L, \\ 0 \text{ si } C = 1 \text{ et } O_S \neq O_L. \end{cases} \tag{7.3}$$

Notons que comme C est une variable booléenne, nous avons : $P([C = 0] \mid O_S\ O_L\ \pi_{Com}) = 1 - P([C = 1] \mid O_S\ O_L\ \pi_{Com})$. Techniquement, C est une variable de cohérence (Pradalier et collab., 2003).

7.3 Conclusion

Nous avons spécifié dans ce chapitre une partie des connaissances préalable π_{Com} décrivant la situation de communication parlée décrite au Chapitre 2. Celles-ci peuvent se résumer par le schéma PBR partiel de la Figure 7.2. Nous n'avons pas spécifié les domaines des variables motrice M et auditive S, ni les formes paramétriques des termes de la décomposition dans lesquelles elles interviennent : leurs domaines, ainsi que la distribution $P(S \mid M\ \pi_{Com})$, dépendent en effet du système articulatori-acoustique considéré que nous spécifierons lors de nos simulations en Partie III. Les distributions $P(M \mid O_S\ \pi_{Com})$ et $P(O_L \mid S\ \pi_{Com})$ correspondent quant à elles aux comportements de production et de perception des agents locuteur et auditeur, respectivement. La définition de ces deux comportements est l'objet du chapitre qui suit, à partir de l'hypothèse d'internalisation de la situation de communication par les agents que nous avons proposée au Chapitre 3.

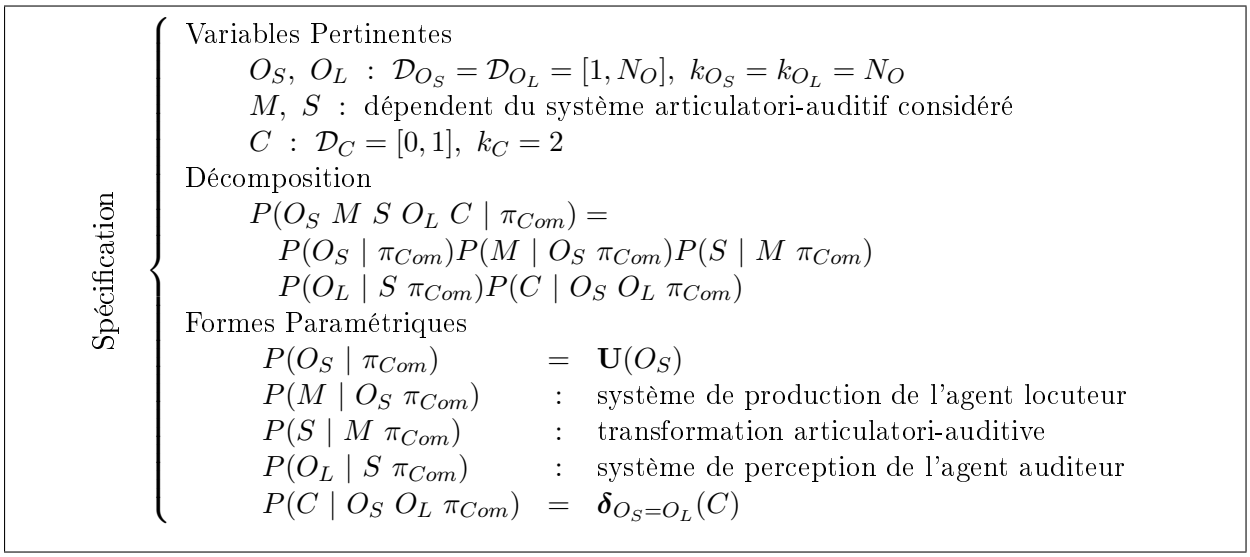


FIGURE 7.2 – Spécification des connaissances préalables π_{Com} de la situation de communication parlée.

Chapitre 8

Modélisation d'un agent communicant

Une des leçons du Chapitre 3 est que les systèmes cognitifs de production et de perception de la parole ne peuvent se cantonner dans des systèmes moteurs et perceptifs, respectivement. L'analyse que nous avons menée du débat entre les différents courants théoriques de ce domaine montre au contraire que les interactions sensori-motrices doivent jouer un rôle nécessaire dans au moins l'une des deux tâches. Cette analyse nous a conduit à proposer une hypothèse, centrale dans notre travail, d'internalisation de la situation de communication dans l'architecture cognitive des agents communicants (Section 3.5.3). Pour pouvoir communiquer correctement, il semble en effet raisonnable de proposer qu'un agent doive posséder un bon modèle de toute la chaîne décrite par la situation de communication. Celle-ci comprend deux agents, locuteur et auditeur, devant se transmettre un objet de communication par le biais d'une transformation articulatori-acoustique réalisée par l'environnement. Notre revue de la littérature sur les théories motrices, auditives et sensori-motrices de la production et de la perception de la parole, ainsi que des travaux de neurosciences sur les associations sensori-motrices, nous a donc amené à proposer un modèle cognitif d'agent communicant comprenant (Figure 8.1) :

- un système moteur, associant les objets de communication aux gestes moteurs de l'agent ;
- un lien sensori-moteur, associant ses gestes moteurs à ses stimuli auditifs ;
- un système auditif, associant ses stimuli auditifs aux objets de communication.

Nous proposons donc que la structure de dépendance de l'architecture cognitive des agents soit identique à celle de la situation de communication. Si π_{Ag} représente l'ensemble des hypothèses préalables du modèle d'agent, nous avons donc :

$$\pi_{Ag} = \pi_{Com}, \tag{8.1}$$

et ainsi, d'après 7.1 :

$$\begin{aligned} P(O_S\ M\ S\ O_L\ C \mid \pi_{Ag}) \;=\; & P(O_S \mid \pi_{Ag})P(M \mid O_S\ \pi_{Ag})P(S \mid M\ \pi_{Ag}) \\ & P(O_L \mid S\ \pi_{Ag})P(C \mid O_S\ O_L\ \pi_{Ag}). \end{aligned} \tag{8.2}$$

Ainsi constitué, un agent communicant dispose de toutes les briques fonctionnelles lui permettant d'avoir un bon modèle interne de la situation de communication. Du point de

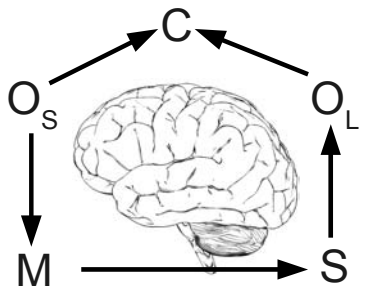


FIGURE 8.1 – Structure du modèle d'agent. Nous rappelons que l'image d'un cerveau ne sert ici qu'à symboliser la notion de modèle interne et n'a aucune vocation à suggérer des localisations neuroanatomiques.

vue des connaissances préalables π_{Com} décrites dans le chapitre précédent et rappelées ci-dessus, le modèle de la situation et le modèle d'agent sont même rigoureusement identiques, à ce stade seule l'interprétation des termes de la décomposition est différente :

- $P(M \mid O_S\ \pi_{Ag})$ est le système moteur de l'agent, associant des objets de communication O_S à ses gestes moteurs M ;
- $P(S \mid M\ \pi_{Ag})$ est le lien sensori-moteur de l'agent, associant ses gestes moteurs M à ses stimuli auditifs S ;
- $P(O_L \mid S\ \pi_{Ag})$ est le système auditif de l'agent, associant ses stimuli auditifs S à des objets de communication O_L.

Ce n'est qu'au chapitre suivant que nous différencierons les termes de π_{Ag} et π_{Com}.

La section suivante utilise les connaissances π_{Ag} pour spécifier des comportements de production et de perception de la parole et formaliser les notions de théories motrices, auditives et sensori-motrices.

8.1 Comportements

À partir de l'architecture générale de l'agent communicant, nous proposions au Chapitre 3 une synthèse des différents courants théoriques en production et perception de la parole, rappelée Table 8.1. Un comportement de production correspond à un chemin de O_S à M, alors qu'un comportement de perception correspond à un chemin de S à O_L. Cette conception permet de définir les notions de théories motrices, auditives et sensori-motrices autour du même modèle général de la Figure 8.1 de la façon suivante :

- une théorie motrice de la communication consiste en une désactivation du système auditif (association $S - O_L$) ;
- une théorie auditive de la communication consiste en une désactivation du système

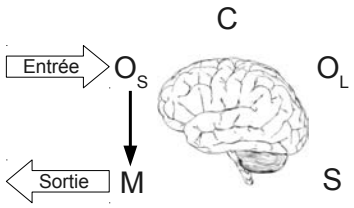
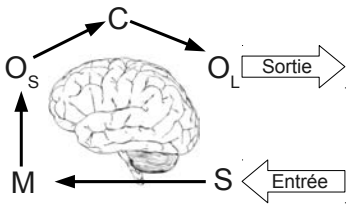
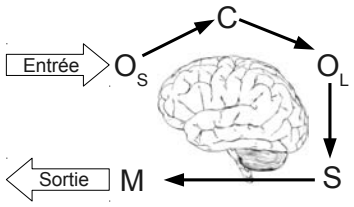
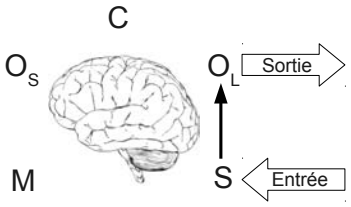
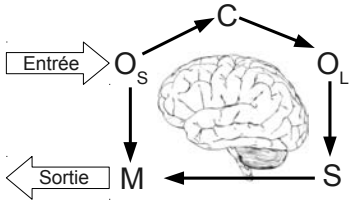
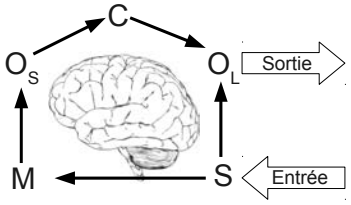
	Production	Perception
Moteur	Section 3.1.1 	Section 3.1.2 
Auditif	Section 3.2.1 	Section 3.2.2 
Sensori-moteur	Section 3.3.1 	Section 3.3.2 

TABLE 8.1 – Rappel de la synthèse des différents courants théoriques en production et perception de la parole avec renvois aux sections correspondantes du Chapitre 3, et conceptions en terme de modèle cognitif d'agent. Les flèches indiquent ici un flux d'information.

moteur (association $O_S - M$) ;
- une théorie sensori-motrice de la communication ne désactive aucun des systèmes moteur et auditif.

En termes probabilistes, la désactivation d'un système consiste en une absence de connaissance sur l'association qu'il représente, et revient donc à fixer la distribution correspondante à une distribution uniforme. Ainsi, les théories motrices et auditives correspondent à des simplifications des connaissances π_{Ag}, que nous appellerons respectivement π_{AgM} et π_{AgA}, dans lesquelles les hypothèses préalables sont identiques à celles du modèle d'agent, à l'exception de :

Théories motrices :

$$P(O_L \mid S\ \pi_{AgM}) = \boldsymbol{U}(O_L). \tag{8.3}$$

Les autres distributions sont identiques à celles du modèle d'agent π_{Ag}.

Théories auditives :

$$P(M \mid O_S\ \pi_{AgA}) = \boldsymbol{U}(M). \tag{8.4}$$

Les autres distributions sont identiques à celles du modèle d'agent π_{Ag}.

Les théories sensori-motrices n'impliquent quant à elles aucune simplification du modèle général. En notant π_{AgSM} le modèle correspondant, on a donc :

Théories sensori-motrices :

$$\pi_{AgSM} = \pi_{Ag}. \tag{8.5}$$

Ainsi, seules les formes des distributions motrices et auditives distinguent les différents courants théoriques. Les questions posées au modèle sont quant à elles identiques :

Un comportement de production consiste à calculer une distribution sur les gestes moteurs M sachant un objet $[O_S = o_i]$ à communiquer par l'agent et sachant que l'on suppose acquis le succès de la communication $[C = 1]$:

$$P(M \mid [O_S = o_i]\ [C = 1]\ \pi_{Th}). \tag{8.6}$$

où π_{Th} correspond à π_{AgM}, π_{AgA} ou π_{AgSM} selon la théorie considérée.

Un comportement de perception consiste à calculer une distribution sur les objets de communication O_L sachant une entrée auditive $[S = s]$ reçue par l'agent et sachant que l'on suppose acquis le succès de la communication $[C = 1]$:

$$P(O_L \mid [S = s]\ [C = 1]\ \pi_{Th}). \tag{8.7}$$

où π_{Th} correspond à π_{AgM}, π_{AgA} ou π_{AgSM} selon la théorie considérée.

8.2 Unification probabiliste des théories de la communication parlée

Ayant formalisé les comportements de production et de perception (Équations 8.6 et 8.7, respectivement), ainsi que les notions de théories motrices, auditives et sensori-motrices (Équations 8.3, 8.4 et 8.5, respectivement), nous réalisons maintenant l'inférence bayésienne nous permettant d'associer à chaque cellule de la Table 8.1 l'expression probabiliste correspondante.

Nous ne précisons pas le développement des expressions exprimées ci-dessous. Le lecteur intéressé pourra utiliser les Équations 6.13, 6.18, 7.2, 7.3, 8.1, 8.2, 8.3, 8.4, 8.5, 8.6 et 8.7 pour les résoudre. Plus précisément, chaque développement consiste à réaliser une inférence bayésienne (Équation 6.18) sur la décomposition de l'Équation 8.2, à partir de l'une des questions de production (Équation 8.6) ou de perception (Équation 8.7). Les expressions obtenues se simplifient alors par 7.2 et 7.3, ainsi que par 8.3, 8.4 et 8.5 selon la théorie considérée.

8.2.1 Théories motrices

8.2.1.1 Production

Le comportement de production d'une théorie motrice est exprimé par la question probabiliste :

$$P(M \mid [O_S = o_i]\ [C = 1]\ \pi_{AgM}) \propto P(M \mid [O_S = o_i]\ \pi_{Ag}). \tag{8.8}$$

Cette expression conduit donc à un comportement favorisant simplement la production de gestes moteurs habituellement exécutés pour l'objet o_i considéré. Ainsi, notre formalisme revient bien à considérer une théorie motrice de la production (telle que la Phonologie Articulatoire exposée en 3.1.1) comme indépendante des représentations auditives, en la définissant comme une distribution sur M dépendant uniquement de l'objet de communication O_S.

8.2.1.2 Perception

Le comportement de perception d'une théorie motrice est exprimé par la question probabiliste :

$$P(O_L \mid [S = s]\ [C = 1]\ \pi_{AgM}) \propto \sum_M P(M \mid O_S = O_L\ \pi_{Ag}) P([S = s] \mid M\ \pi_{Ag}). \tag{8.9}$$

Cette expression conduit donc à un comportement favorisant l'inférence d'objets de communication pour lesquels le système moteur de l'agent aurait produit des gestes avec des conséquences auditives similaires à s. Ainsi, notre formalisme revient bien à considérer qu'une théorie motrice de la perception (telle que celle exposée en 3.1.2) est basée sur une transformation de la modalité auditive à la modalité motrice (par la distribution $P([S = s] \mid M\ \pi_{Ag})$) ainsi qu'une catégorisation dans l'espace moteur (par la distribution $P(M \mid O_S = O_L\ \pi_{Ag})$).

8.2.2 Théories auditives

8.2.2.1 Production

Le comportement de production d'une théorie auditive est exprimé par la question probabiliste :

$$P(M \mid [O_S = o_i]\ [C = 1]\ \pi_{AgA}) \propto \sum_S P(S \mid M\ \pi_{Ag}) P([O_L = o_i] \mid S\ \pi_{Ag}). \tag{8.10}$$

Cette expression conduit donc à un comportement favorisant la production de gestes moteurs dont la conséquence auditive permettra au système auditif de correctement inférer l'objet de communication o_i considéré. Ainsi, notre formalisme revient bien à considérer qu'une théorie auditive de la production (telle que celle exposée en 3.2.1) est basée sur une cible auditive (distribution $P([O_L = o_i] \mid S\ \pi_{Ag})$) et nécessite une transformation de la modalité auditive à la modalité motrice (distribution $P(S \mid M\ \pi_{Ag})$).

8.2.2.2 Perception

Le comportement de perception d'une théorie auditive est exprimé par la question probabiliste :

$$P(O_L \mid [S = s]\ [C = 1]\ \pi_{AgA}) \propto P(O_L \mid [S = s]\ \pi_{Ag}). \tag{8.11}$$

Cette expression conduit donc à un comportement d'inférence d'objets de communication basé uniquement sur la capacité de catégorisation du système auditif (distribution $P(O_L \mid [S = s]\ \pi_{Ag})$). Ainsi, notre formalisme revient bien à considérer une théorie auditive de la perception (telle que celle exposée en 3.2.2) comme indépendante des représentations motrices, en la définissant simplement comme une distribution sur O_L dépendant uniquement de l'entrée sensorielle S.

8.2.3 Théories sensori-motrices

8.2.3.1 Production

Le comportement de production d'une théorie sensori-motrice est exprimé par la question probabiliste :

$$\begin{aligned} &P(M \mid [O_S = o_i]\ [C = 1]\ \pi_{AgSM}) \\ &\quad \propto P(M \mid [O_S = o_i]\ \pi_{Ag}) \sum_S P(S \mid M\ \pi_{Ag}) P([O_L = o_i] \mid S\ \pi_{Ag}). \end{aligned} \tag{8.12}$$

Dans le résultat de cette inférence, on reconnait le produit des expressions des comportements de production des théories motrices et auditives. En effet, d'après les équations correspondantes, 8.8 et 8.10, il se réécrit :

$$\begin{aligned} &P(M \mid [O_S = o_i]\ [C = 1]\ \pi_{AgSM}) \\ &\quad \propto P(M \mid [O_S = o_i]\ [C = 1]\ \pi_{AgM}) P(M \mid [O_S = o_i]\ [C = 1]\ \pi_{AgA}). \end{aligned}$$

Ainsi, dans notre formalisme, une théorie sensori-motrice de la production correspond au produit des expressions des théories motrices et auditives en production. Elle combinera donc les comportements moteurs et auditifs, en produisant un compromis entre un geste habituellement exécuté pour l'objet de communication considéré (terme moteur $P(M \mid O_S = o_i\ \pi_{Ag})$), et un geste dont la conséquence sensorielle permet au système auditif de le catégoriser correctement (terme auditif $\sum_S P(S \mid M\ \pi_{Ag}) P(O_L = o_i \mid S\ \pi_{Ag})$).

8.2.3.2 Perception

Le comportement de perception d'une théorie sensori-motrice est exprimée par la question probabiliste :

$$\begin{aligned} &P(O_L \mid [S = s]\ [C = 1]\ \pi_{AgSM}) \\ &\quad \propto P(O_L \mid [S = s]\ \pi_{Ag}) \sum_M P(M \mid O_S = O_L\ \pi_{Ag}) P([S = s] \mid M\ \pi_{Ag}) \end{aligned} \tag{8.13}$$

Dans le résultat de cette inférence, on reconnait le produit des expressions des comportements de perception des théories auditives et motrices. En effet, d'après les équations correspondantes,8.9 et 8.11, il se réécrit :

$$\begin{aligned} &P(O_L \mid [S=s]\ [C=1]\ \pi_{AgSM}) \\ &\quad \propto P(O_L \mid [S=s]\ [C=1]\ \pi_{AgA}) P(O_L \mid [S=s]\ [C=1]\ \pi_{AgM}). \end{aligned}$$

Ainsi, dans notre formalisme, une théorie sensori-motrice de la perception correspond au produit des expressions des théories motrices et auditives en perception. Elle combinera donc les deux types de comportement, en inférant un compromis entre un objet catégorisé uniquement par le système auditif (terme auditif $P(O_L \mid S = s\ \pi_{Ag})$) et un objet pour lequel le système moteur de l'agent aurait produit des gestes avec des conséquences auditives similaires à s (terme moteur $\sum_M P(M \mid O_S = O_L\ \pi_{Ag}) P([S=s] \mid M\ \pi_{Ag})$).

8.3 Conclusion

Ce chapitre propose une formalisation des comportements de production et de perception de la parole, déclinés dans leurs versions motrices, auditives et sensori-motrices, unifiée autour d'une hypothèse d'internalisation de la situation de communication dans l'architecture cognitive des agents. Nous le résumons par la Table 8.2, qui associe à chaque comportement une expression probabiliste pour chacun des courants théoriques considérés.

Ce formalisme, que nous considérons comme un résultat à part entière de notre travail de conceptualisation des théories de la communication, prolonge ainsi les arguments de rationalité en faveur des théories sensori-motrices que nous avons proposé dans les derniers paragraphes du Chapitre 3. En effet, les spécifications de la situation de communication parlée et celles des théories sensori-motrices sont identiques ($\pi_{AgSM} = \pi_{Com}$ d'après les Équations 8.1 et 8.5) alors que ce n'est pas le cas des spécifications des théories motrices et auditives (Équations 8.3 et 8.4). Le modèle d'agent sensori-moteur est donc le modèle qui contient le plus de connaissance sur la situation de communication et nous verrons à la partie III comment il en dégage ses propriétés fonctionnelles.

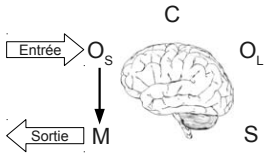
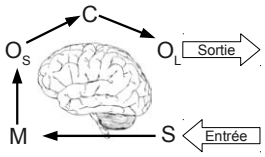
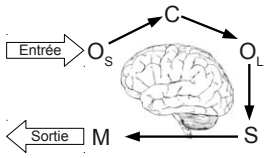
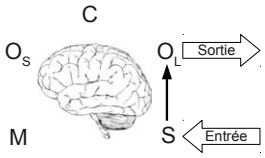
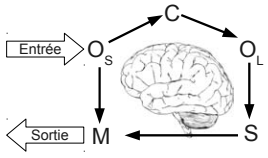
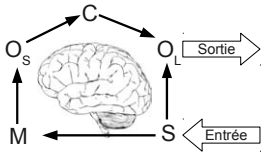
	Production	Perception
Moteur	Section 3.1.1  $P(M \mid [O_S = o_i]\ [C = 1]\ \pi_{AgM})$ $\propto P(M \mid [O_S = o_i]\ \pi_{Ag})$	Section 3.1.2  $P(O_L \mid [S = s]\ [C = 1]\ \pi_{AgM})$ $\propto \sum_M \begin{pmatrix} P(M \mid O_S = O_L\ \pi_{Ag}) \\ P([S = s] \mid M\ \pi_{Ag}) \end{pmatrix}$
Auditif	Section 3.2.1  $P(M \mid [O_S = o_i]\ [C = 1]\ \pi_{AgA})$ $\propto \sum_S \begin{pmatrix} P(S \mid M\ \pi_{Ag}) \\ P([O_L = o_i] \mid S\ \pi_{Ag}) \end{pmatrix}$	Section 3.2.2  $P(O_L \mid [S = s]\ [C = 1]\ \pi_{AgA})$ $\propto P(O_L \mid [S = s]\ \pi_{Ag})$
Sensori-moteur	Section 3.3.1  $P(M \mid [O_S = o_i]\ [C = 1]\ \pi_{AgSM})$ $\propto P(M \mid [O_S = o_i]\ \pi_{Ag}) \sum_S \begin{pmatrix} P(S \mid M\ \pi_{Ag}) \\ P([O_L = o_i] \mid S\ \pi_{Ag}) \end{pmatrix}$	Section 3.3.2  $P(O_L \mid [S = s]\ [C = 1]\ \pi_{AgSM})$ $\propto P(O_L \mid [S = s]\ \pi_{Ag}) \sum_M \begin{pmatrix} P(M \mid O_S = O_L\ \pi_{Ag}) \\ P([S = s] \mid M\ \pi_{Ag}) \end{pmatrix}$

TABLE 8.2 – Synthèse des différents courants théoriques en production et perception de la parole avec renvois aux sections correspondantes du Chapitre 3, conceptions en terme de modèle cognitif d'agent et expressions probabilistes. Les flèches indiquent ici un flux d'information.

Chapitre 9

Modélisation d'une société d'agents prélangagiers

Le chapitre précédent a formalisé la notion d'agent communicant en termes bayésiens (Figure 8.1 et Équation 8.2). Nous avons défini les comportements généraux de production et de perception (Équation 8.6 et 8.7), et montré comment des simplifications de ces comportements basés sur des désactivations des sous-systèmes moteur (Équation 8.4) ou auditif (Équation 8.3) pouvaient rendre compte formellement des théories motrices, auditives et sensori-motrices de la communication exposées au Chapitre 3. Cette formalisation est résumée Table 8.2.

Nous nous intéressons maintenant à la définition algorithmique du scénario d'émergence des systèmes phonologiques retenu au Chapitre 4. Ce chapitre proposait d'unir le modèle de la situation de communication du Chapitre 2 et le modèle d'agent communicant du Chapitre 3, que nous venons de formaliser dans les deux chapitres précédents dans les connaissances préalables π_{Com} et π_{Ag}, dans un modèle d'interaction construit autour de la notion de jeu déictique (rappelé Figure 9.1).

Un jeu déictique consiste en deux agents, un locuteur et un auditeur, partageant leur attention sur un objet de l'environnement. L'agent locuteur propose alors un geste vocal pour nommer l'objet, l'agent auditeur perçoit le stimulus auditif correspondant, et les deux mettent à jour leurs connaissances respectives en fonction de cette interaction. Ces jeux se succèdent au cours du temps, chaque agent pouvant prendre tour à tour le rôle de locuteur ou d'auditeur.

Par le jeu déictique, nous proposons une amorce évolutive vers l'émergence du langage à partir de laquelle des agents placés en situation de communication prélangagière peuvent exploiter leur modèle interne de la situation pour évoluer vers un système phonologique adéquat. Par communication prélangagière, nous comprenons tout système non-vocal permettant d'obtenir une attention partagée entre deux agents sur un objet de communication. Notre choix est celui d'un comportement de déixis, permettant aux deux agents d'identifier le même objet durant leur interaction (en considérant que l'un d'eux le désigne par toute action corporelle non-vocale adéquate).

Nous commençons par étendre la définition d'un agent communicant du chapitre pré-

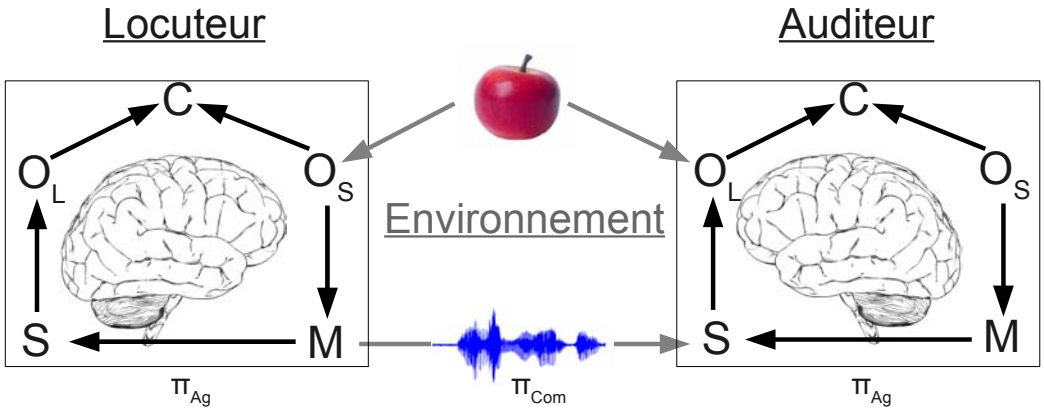


FIGURE 9.1 – Un jeu déictique : deux agents, instances du modèle d'agent communicant π_{Ag} et augmentés d'un comportement de déixis leur permettant une attention partagée sur un objet (représenté par la pomme), interagissent par le biais d'un environnement réalisant la transformation articulatori-acoustique du modèle π_{Com} (représentée par l'onde entre M et S), dans le but de nommer l'objet.

cédent à celle d'un agent apprenant, capable de faire évoluer ses connaissances motrices et sensorielles selon ses interactions avec d'autres agents. Puis exposons l'algorithme de simulation que nous utiliserons tout au long de la Partie III, dans lequel un ensemble d'agents apprenants évoluent par jeux déictiques dans un environnement contenant un ensemble d'objets à nommer.

9.1 Évolution et apprentissage

La définition d'un jeu déictique que nous avons proposée au Chapitre 4 met en jeu des agents capables de faire évoluer leurs connaissances motrices et auditives selon leurs interactions. Le comportement de déixis préexistant dont nous les dotons leur permet en effet d'identifier correctement le même objet avant que l'interaction vocale n'ait lieu. Formellement, ce comportement revient à fixer $[C = 1]$ lors de l'interaction entre deux agents (ce qui est représenté par la pomme identifiée par les deux agents sur la Figure 9.1). Ceci agit comme un signal d'apprentissage de façon à ce que lors d'un jeu déictique :

- les deux agents ont connaissance de l'objet o_i (comportement de déixis) ;
- le locuteur produit un geste m ;
- l'auditeur perçoit un stimulus s.

Ainsi, à la fin d'un jeu, l'agent locuteur dispose d'un couple (m, o_i) et l'agent auditeur d'un couple (s, o_i). Comme ces jeux se succèdent au cours du temps, chaque agent pouvant prendre tour à tour un statut d'auditeur ou de locuteur, un agent a de la société possède alors au temps t un ensemble d'apprentissage $\delta_a(t)$ constitué de triplets :

$$(n, x, o_i) \in (([0,t] \times \mathcal{D}_M \times \mathcal{D}_{O_S}) \cup ([0,t] \times \mathcal{D}_S \times \mathcal{D}_{O_L})) \,.$$

Chacun de ces triplets associe à un jeu déictique $n \in [0, t]$ soit un couple $(m, o_i) \in \mathcal{D}_M \times \mathcal{D}_{O_S}$ si l'agent a un statut de locuteur au temps n, soit un couple $(s, o_i) \in \mathcal{D}_S \times \mathcal{D}_{O_L}$ s'il a un statut d'auditeur. Les jeux dans lesquels l'agent a n'intervient pas ne sont pas enregistrés dans $\delta_a(t)$.

Nous étendons donc la définition d'un agent communicant du chapitre précédent à celle d'un agent apprenant dont les connaissances motrices et auditives évoluent avec son ensemble d'apprentissage $\delta_a(t)$. Le modèle d'un agent apprenant a au temps t devient alors :

$$\begin{aligned} & P(O_S\ M\ S\ O_L\ C \mid \delta_a(t)\ \pi_{Ag}) \\ &= P(O_S \mid \delta_a(t)\ \pi_{Ag}) P(M \mid O_S\ \delta_a(t)\ \pi_{Ag}) P(S \mid M\ \delta_a(t)\ \pi_{Ag}) \\ &\quad P(O_L \mid S\ \delta_a(t)\ \pi_{Ag}) P(C \mid O_S\ O_L\ \delta_a(t)\ \pi_{Ag}) \\ &= P(O_S \mid \pi_{Ag}) P(M \mid O_S\ \delta_a(t)\ \pi_{Ag}) P(S \mid M\ \pi_{Ag}) \\ &\quad P(O_L \mid S\ \delta_a(t)\ \pi_{Ag}) P(C \mid O_S\ O_L\ \pi_{Ag}) \end{aligned} \tag{9.1}$$

Nous considérons en effet que l'évolution de l'état d'un agent a au cours du temps ne dépend que de la mise à jour des connaissances de ses sous-systèmes moteurs $P(M \mid O_S\ \delta_a(t)\ \pi_{Ag})$ et auditifs $P(O_L \mid S\ \delta_a(t)\ \pi_{Ag})$. Les autres distributions n'évoluent pas. En particulier, le lien sensori-moteur $P(S \mid M\ \pi_{Ag})$ est considéré comme déjà appris par les agents (par exemple grâce à une exploration sensori-motrice préalable). Notons que les connaissances préalables π_{Ag} sont identiques pour tous les agents, seuls leurs ensembles d'apprentissage respectifs diffèrent (on pourra parfois parler de « vécu » moteur et auditif des agents).

Le processus de mise à jour d'une distribution contenant $\delta_a(t)$ en partie droite dépend de sa forme paramétrique particulière que nous spécifierons lors des simulations de la Partie III (voir Section 6.3.1.3 pour les différentes formes paramétriques que nous utiliserons). À ce stade, retenons simplement que les paramètres d'une telle distribution sont calculés en fonction des données d'apprentissage accumulées par l'agent au cours des N derniers jeux déictiques auxquels il a participé, et évoluent donc au fur et à mesure du déroulement de la simulation.

9.2 Paramètres et algorithme de simulation

Un environnement est constitué d'un ensemble A de N_A agents, d'un ensemble O de N_O objets et de la transformation articulatori-acoustique $P(S \mid M\ \pi_{Com})$ du Chapitre 7 modélisant le passage d'un geste moteur $m \in M$ à un stimulus sensoriel $s \in S$ dans l'environnement. Notons que nous différencions maintenant la transformation articulatori-acoustique réalisée par l'environnement, $P(S \mid M\ \pi_{Com})$, et le lien sensori-moteur comme connaissance d'un agent, $P(S \mid M\ \pi_{Ag})$.

Les agents sont des instances du modèle d'agent communicant défini au chapitre précédent. Tous les agents implémentent la même version de ce modèle, notée π_{Th}, selon la théorie considérée : motrice ($\pi_{Th} = \pi_{AgM}$), auditive ($\pi_{Th} = \pi_{AgA}$), ou sensori-motrice ($\pi_{Th} = \pi_{AgSM}$). En d'autres termes, nous souhaitons simuler et comparer des sociétés d'agents « tous moteurs », « tous auditifs » ou « tous sensori-moteurs ». La Table 8.2 du

chapitre précédent synthétise les questions probabilistes des comportements de production et de perception dans chacun des courants théoriques. L'algorithme de simulation est alors :

- t=0 ;
- Tant que $t < N_{JD}$ (le nombre de jeux déictiques avant l'arrêt de la simulation) :
  - tirage aléatoire d'un objet $o_i \in O$, d'un agent locuteur $speaker \in A$ et d'un agent auditeur $listener \in A$, avec $listener \neq speaker$;
  - l'agent locuteur $speaker$ tire un geste m pour l'objet o_i selon son comportement de production $P(M \mid [O_S = o_i]\ [C = 1]\ \delta_{speaker}(t)\ \pi_{Th})$ (dans sa version motrice, auditive ou sensori-motrice selon π_{Th}) ;
  - l'environnement transforme le geste moteur m en un stimulus sensoriel s en tirant selon $P(S \mid [M = m]\ \pi_{Com})$;
  - l'agent auditeur $listener$ reçoit l'entrée sensorielle s ;
  - $\delta_{speaker}(t) = \delta_{speaker}(t-1) \cup (t, m, o_i)$;
  - $\delta_{listener}(t) = \delta_{listener}(t-1) \cup (t, s, o_i)$;
  - $t \leftarrow t + 1$.

C'est donc l'évolution des données d'apprentissage dans les ensembles $\delta_a(t)$ au cours du temps qui permet aux agents de faire évoluer leurs systèmes moteurs et auditifs. En effet, les paramètres des distributions correspondantes (contenant $\delta_a(t)$ en partie droite dans le modèle d'agent apprenant (Équation 9.1)) sont calculés en fonction des données de cet ensemble (Section 6.3.1.3). Ainsi, l'évolution des ensembles d'apprentissage des agents dans les dernières étapes de l'algorithme ci-dessus induit implicitement la mise à jour des distributions concernées (mais parfois sur des pas de temps différents, nous le verrons en Partie III). L'évolution des connaissances de chaque agent dépend donc des interactions qu'il a vécu avec ses congénères.

Afin de calculer l'évolution du taux de reconnaissance dans la société au cours du temps, on ajoutera parfois une dernière étape dans laquelle l'agent auditeur infère un objet o_L selon son comportement de perception (Équation 8.9, 8.11 ou 8.13 selon que l'agent implémente une version motrice, auditive ou sensori-motrice du comportement, respectivement). Cette étape n'est pas nécessaire dans le sens où elle n'agit pas sur la mise à jour des agents, mais elle peut permettre de comparer o_L à o_i pour savoir si la communication aurait été un succès s'il n'y avait pas eu d'attention partagée préalable à l'interaction vocale, et ainsi mesurer l'évolution de taux de reconnaissance dans la société au cours des jeux déictiques. En d'autres termes, ce taux permet l'évaluation de la capacité des agents à évoluer d'une communication déictique, nécessitant une attention partagée sur un objet présent dans l'environnement, à une communication uniquement vocale dans laquelle l'objet ne serait pas connu des deux agents avant l'interaction.

9.3 Discussion

9.3.1 De l'optimisation de la situation de communication à l'apprentissage par jeux déictiques

L'algorithme d'interaction par jeux déictiques alterne donc deux étapes. La première est l'inférence d'une distribution sur M sachant $[O_S = o_i]$ et $[C = 1]$ par un comportement de production. Elle mène au tirage d'un geste moteur $m \in M$, puis d'un stimulus $s \in S$ (par la transformation de l'environnement). La deuxième est l'estimation des paramètres des distributions motrices et auditives des agents à partir de données d'apprentissage issues de la première étape.

Bien que ces étapes soient ici décentralisées dans une société d'agents alternant des rôles de locuteur et d'auditeur, nous pensons qu'elles sont analogues aux étapes d'une procédure itérative d'estimation des paramètres des systèmes de production et et de perception des agents qui maximisent le succès de la situation de communication π_{Com}.

Sans montrer d'équivalence formelle, nous souhaitons proposer une telle procédure inspirée de l'algorithme Espérance-Maximisation (Dempster et collab., 1977) pour mettre en valeur ses points communs avec l'algorithme d'interaction par jeux déictiques. Pour cela, nous reprenons la décomposition de la situation de communication de l'Équation 7.1 et supposons que les distributions représentant le système de production de l'agent locuteur et celui de perception de l'auditeur dépendent d'un ensemble de paramètres Θ tels que :

$$\begin{aligned} P(O_S\ M\ S\ O_L\ C \mid \Theta\ \pi_{Com}) &= P(O_S \mid \pi_{Com})P(M \mid O_S\ \Theta\ \pi_{Com})P(S \mid M\ \pi_{Com}) \\ &\quad P(O_L \mid S\ \Theta\ \pi_{Com})P(C \mid O_S\ O_L\ \pi_{Com}). \end{aligned} \tag{9.2}$$

Nous nous intéressons alors à l'approximation de l'expression :

$$argmax_{\Theta}\,(P(C = 1 \mid \Theta\ \pi_{Com})).$$

En d'autres termes : quels paramètres des systèmes de production et de perception des agents maximisent le succès de la communication dans la situation π_{Com} ? Le calcul exacte de cette expression est trop complexe pour permettre l'utilisation de méthodes analytiques. En effet, par inférence bayésienne sur 9.2, nous avons :

$$\begin{aligned} P(C = 1 \mid \Theta\ \pi_{Com}) &= \sum_{O_S} P(O_S \mid \pi_{Com}) \sum_{M} P(M \mid O_S\ \Theta\ \pi_{Com}) \sum_{S} P(S \mid M\ \pi_{Com}) \\ &\quad \sum_{O_L} P(O_L \mid S\ \Theta\ \pi_{Com})P(C = 1 \mid O_S\ O_L\ \pi_{Com}) \end{aligned} \tag{9.3}$$

Nous proposons donc une procédure itérative inspirée de l'algorithme Espérance-Maximisation, alternant une étape de génération de données d'apprentissage δ sachant $C = 1$, et une étape de calcul des paramètres θ de la décomposition à partir de ces données d'apprentissage par une fonction $\mathcal{A} : \delta \rightarrow \Theta$ (correspondant généralement à des estimateurs statistiques).

En partant de paramètres initiaux $\theta = \theta_0$, la première étape consiste à générer, pour chaque objet $o_i \in \mathcal{D}_{O_S}$ $(= \mathcal{D}_{O_L})$, deux jeux de données sachant $[C = 1]$ tirés selon les distributions (calculées par inférence bayésienne à partir de l'Équation 9.2) :

$$P(M \mid O_S = o_i \; C = 1 \; \theta \; \pi_{Com}) \propto P(M \mid O_S = o_i \; \theta \; \pi_{Com}) \sum_S P(S \mid M \; \pi_{Com}) P(O_L = o_i \mid S \; \theta \; \pi_{Com}) \quad (9.4)$$

et

$$P(S \mid O_L = o_i \; C = 1 \; \theta \; \pi_{Com}) \propto P(O_L = o_i \mid S \; \theta \; \pi_{Com}) \sum_M P(M \mid O_S = o_i \; \theta \; \pi_{Com}) P(S \mid M \; \pi_{Com}) \quad (9.5)$$

En notant δ l'ensemble de ces deux jeux de données, on raffine alors les paramètres des distributions par :

$$\theta = \mathcal{A}(\delta). \quad (9.6)$$

Cette procédure est répétée autant de fois que nécessaire pour obtenir une valeur satisfaisante de $P(C = 1 \mid \theta \; \pi_{Com})$ (calculée par 9.3).

L'algorithme d'optimisation décrit ici n'a pas vocation à être mis œuvre dans ce document et reste marginal dans notre démarche de modélisation. Nous l'introduisons seulement dans le but de relier l'algorithme d'interaction entre agent que nous proposerons au Chapitre 9 à une procédure classique d'estimation de paramètres d'un modèle probabiliste par l'alternance d'une étape de génération de données et d'une étape d'estimation des paramètres. On voit en effet :

- que l'étape de génération de données motrices sachant $[C = 1]$ dans π_{Com} (Équation 9.4) utilise une inférence similaire à celle du comportement sensori-moteur de production des agents (Équation 8.12) ;
- que l'étape d'estimation des nouveaux paramètres θ est comparable à la procédure d'apprentissage des agents à partir de leur ensemble d'apprentissage $\delta_a(t)$.

Ainsi, nous pensons que l'algorithme d'interaction par jeux déictiques, lorsqu'il met en jeu des agents en comportement sensori-moteur, réalise en fait une version décentralisée d'une procédure itérative d'estimation des paramètres des systèmes de production et de perception des agents qui maximisent le succès de la communication dans π_{Com}, avec l'avantage d'une interprétation cognitive et évolutionnaire réaliste des éléments du modèle fondée sur toute la réflexion de la Partie I.

9.3.2 Conclusion

Le modèle d'agent apprenant et l'algorithme de simulation décrits dans ce chapitre finalisent le travail de formalisation de la réflexion que nous avons menée en Partie I. Ils seront mis en œuvre tout au long des simulations de la Partie III dans le but :

- d'étudier et de comparer la capacité des comportements de production issus des théories motrice, auditive ou sensori-motrice à faire émerger un système phonologique efficace dans notre paradigme d'interaction par jeux déictiques (Chapitre 10) ;
- d'analyser comment le comportement le plus efficace dans l'étape précédente peut faire émerger des prédictions correctes des données des langues du monde concernant les systèmes de voyelles (Chapitre 11), de consonnes plosives (Chapitre 12) et de syllabes (Chapitre 13).

Il conviendra alors, pour chaque type de simulation mise en œuvre, de définir :

- les espaces et la transformation articulatori-auditive considérés, comprenant :
  - les domaines des variables motrices et sensorielles : $\mathcal{D}_M$ et $\mathcal{D}_S$;
  - la distribution de transformation articulatori-auditive de l'environnement : $P(S \mid M\ \pi_{Com})$;
- l'instanciation du modèle d'agent apprenant considéré (synthétisé dans la notation PBR à la Figure 9.2), comprenant :
  - la théorie considérée : motrice, auditive ou sensori-motrice ;
  - la définition des formes paramétriques des systèmes moteur et auditif, ainsi que du lien sensori-moteur des agents.
- le nombre d'agents dans la société : N_A ;
- le nombre d'objets dans l'environnement : N_O ;
- le nombre de jeux déictiques avant l'arrêt de la simulation : N_{JD}.

Programme

Description

Spécification

Variables Pertinentes

$O_S,\ O_L\ :\ \mathcal{D}_{O_S} = \mathcal{D}_{O_L} = [1, N_O],\ k_{O_S} = k_{O_L} = N_O$

$M,\ S$: dépendent du système articulatori-auditif considéré

$C\ :\ \mathcal{D}_C = [0,1],\ k_C = 2$

Décomposition

$$P(O_S\ M\ S\ O_L\ C \mid \delta_a(t)\ \pi_{Ag}) = P(O_S \mid \pi_{Ag})P(M \mid O_S\ \delta_a(t)\ \pi_{Ag})P(S \mid M\ \pi_{Ag})\ P(O_L \mid S\ \delta_a(t)\ \pi_{Ag})P(C \mid O_S\ O_L\ \pi_{Ag})$$

Formes Paramétriques

$P(O_S \mid \pi_{Ag})$	$=$	$\mathbf{U}(O_S)$
$P(M \mid O_S\ \delta_a(t)\ \pi_{Ag})$	:	système moteur (appris)
$P(S \mid M\ \pi_{Ag})$	:	lien sensori-moteur
$P(O_L \mid S\ \delta_a(t)\ \pi_{Ag})$	:	système auditif (appris)
$P(C \mid O_S\ O_L\ \pi_{Ag})$	$=$	$\boldsymbol{\delta}_{O_S=O_L}(C)$

Identification : par l'ensemble d'apprentissage $\delta_a(t)$ de l'agent acquis au cours des jeux déictiques

Question : $P(M \mid [O_S = o_i]\ [C = 1]\ \delta_a(t)\ \pi_{Ag})$

FIGURE 9.2 – Programme bayésien d'un agent apprenant a au temps t.

Troisième partie

Des modèles formels à la simulation informatique

Chapitre 10

Comparaison des différents courants théoriques en communication parlée pour l'émergence des systèmes phonologiques

Ce chapitre débute la Partie III de ce document, dans laquelle nous exposons nos résultats de simulations d'émergence de systèmes phonologiques dans le paradigme d'interaction par jeux déictiques dans des sociétés d'agents communicants. Ce paradigme, formalisé à la partie précédente, est défini par :

- un modèle d'agent communicant π_{Ag}, définissant les comportements de production et perception des agents, et déclinable dans ses versions motrice (π_{AgM}), auditive (π_{AgA}) ou sensori-motrice (π_{AgSM}) (Chapitre 8).
- L'extension de ce modèle à un modèle d'agent apprenant, dont les paramètres des distributions de π_{Ag} sont appris à partir du « vécu » moteur et auditif de l'agent, enregistré dans un ensemble d'apprentissage (Section 9.1 du Chapitre 9).
- Un algorithme d'interaction par jeux déictiques d'une société d'agents apprenants dans un environnement d'objets, dans lequel l'ensemble d'apprentissage de chaque agent évolue au cours des jeux déictiques de la simulation (Section 9.2 du Chapitre 9).

Tous les agents de la société sont des instances du même modèle d'agent communicant (π_{AgM}, π_{AgA} ou π_{AgSM}). Chaque agent évolue par contre selon sa propre expérience, accumulée pendant les jeux déictiques dans son ensemble d'apprentissage $\delta_a(t)$, où a est un agent de la société et t le nombre de jeux déictiques depuis le début de la simulation. Ces ensembles associent à chaque jeu déictique dans lequel l'agent a a participé les données collectées (un geste moteur et un objet si l'agent avait un statut de locuteur, un stimulus auditif et un objet en statut d'auditeur).

Nous nous intéressons alors tout au long de cette partie à l'étude des systèmes articulatori-auditifs émergeant de telles simulations, d'abord à partir d'un modèle articulatori-auditif simplifié dans le présent chapitre, puis dans des simulations réalistes d'émergence des systèmes de voyelles, de consonnes plosives et de syllabes dans les trois suivants.

Dans ce chapitre, nous intéressons particulièrement à la comparaison des trois comportements d'agents : moteur, auditif et sensori-moteur. Nous commençons par définir un modèle simplifié de transformation articulatori-auditive, dans lequel les variables motrice et auditive sont définies sur un espace mono-dimensionnel borné, ainsi que le modèle d'agent apprenant correspondant. Puis nous étudions les propriétés générales de chaque comportement en fournissant des éléments permettant d'appréhender intuitivement et formellement la dynamique de chacun au cours d'une simulation. Enfin, nous nous intéressons à l'effet des variations de certains paramètres sur les systèmes phonologiques émergeant de nos simulations, en particulier concernant la dégradation des conditions de communication et l'introduction d'une non-linéarité dans la fonction de transformation articulatori-auditive. Nous conclurons sur l'analyse des propriétés de notre modèle en lien avec certains points de la réflexion de la Partie I.

10.1 Modèle de transformation articulatori-auditive

Nous considérons dans ce chapitre un modèle simplifié de transformation articulatori-auditive dans lequel :

$$\begin{aligned} \mathcal{D}_M = \mathcal{D}_S &= \{-10, \ldots, 10\}, \\ k_M = k_S &= 21, \\ \text{TransMS}(m) &= S_{max} \left(\frac{\arctan(NL\ (m - D))}{\arctan(NL\ M_{max})} \right). \end{aligned}$$

Les domaines respectifs des variables articulatoires (M) et auditives (S), $\mathcal{D}_M$ et $\mathcal{D}_S$, correspondent simplement aux 21 valeurs entières de l'intervalle $\{-10, \ldots, 10\}$.

M_{max} et S_{max} sont les bornes supérieures respectives des domaines $\mathcal{D}_M$ et $\mathcal{D}_S$ (soit ici $M_{max} = S_{max} = 10$).

La fonction de transformation articulatori-auditive associe à chaque geste moteur $m \in M$ le stimulus auditif correspondant dans S. Il s'agit d'une fonction sigmoïdale dépendant de deux paramètres : NL et D (Figure 10.1). NL permet de passer d'une fonction quasi-linéaire (si $D = 0$, TransMS tend vers la fonction identité quand NL tend vers 0) à une fonction fortement non-linaire (TransMS tend vers un créneau quand NL tend vers l'infini). Si la fonction est non-linéaire ($NL = 5$ par exemple), le paramètre D permet de déplacer la position du point d'inflexion de la courbe horizontalement entre $M = -10$ et $M = 10$ (Figure 10.2).

La transformation articulatori-auditive réalisée par l'environnement est alors définie par :

$$\forall m \in \mathcal{D}_M \ : P(S \mid [M = m]\ \pi_{Com}) = \boldsymbol{G}_{\text{TransMS}(m),\sigma_{Env}}(S), \tag{10.1}$$

où TransMS est la fonction de transformation décrite ci-dessus et σ_{Env} est appelé le bruit de l'environnement. Plus précisément, chaque distribution $\{P(S \mid [M = m]\ \pi_{Com}), m \in \mathcal{D}_M\}$ est une loi gaussienne (approximée sur un domaine discret, voir Section 6.3.1.3) de moyenne TransMS(m) et d'écart-type σ_{Env}. Nous modélisons ainsi une transformation articulatori-auditive bruitée, comme illustrée Figure 10.3.

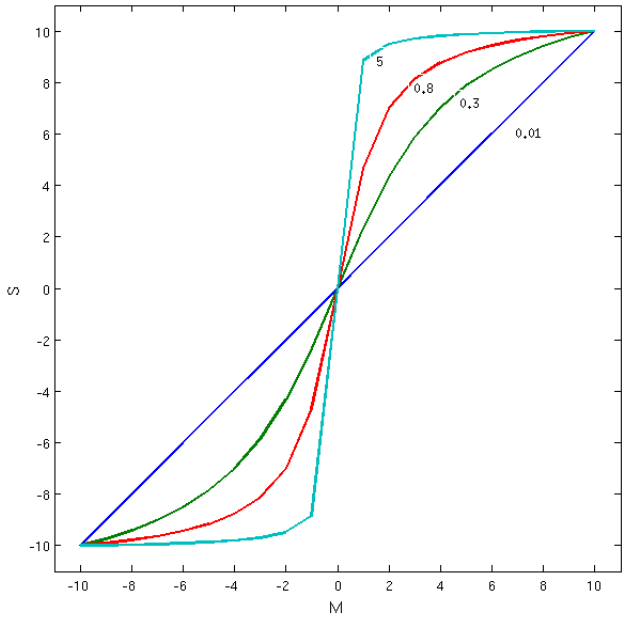


FIGURE 10.1 – Fonction de transformation articulatori-auditive pour les différentes valeurs de NL utilisées dans ce chapitre, indiquées sous chacune des courbes correspondantes. La position du point d'inflexion (paramètre D) est fixé à 0.

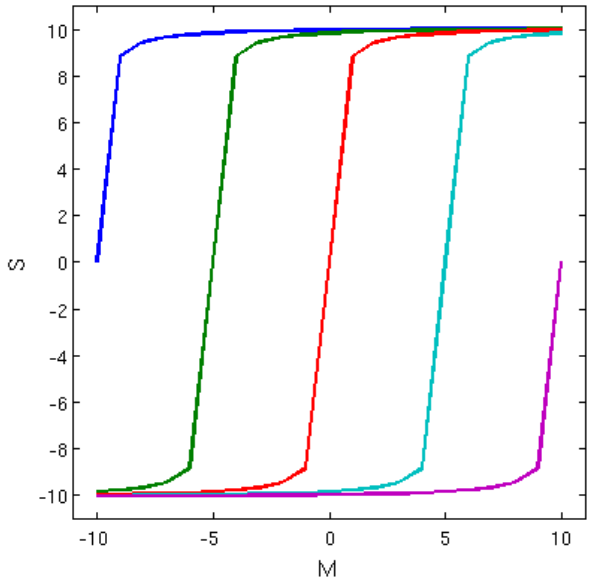


FIGURE 10.2 – Fonction de transformation articulatori-auditive pour des valeurs de D dans $\{-10, -5, 0, 5, 10\}$, de gauche à droite. Le paramètre NL est fixé à 5.

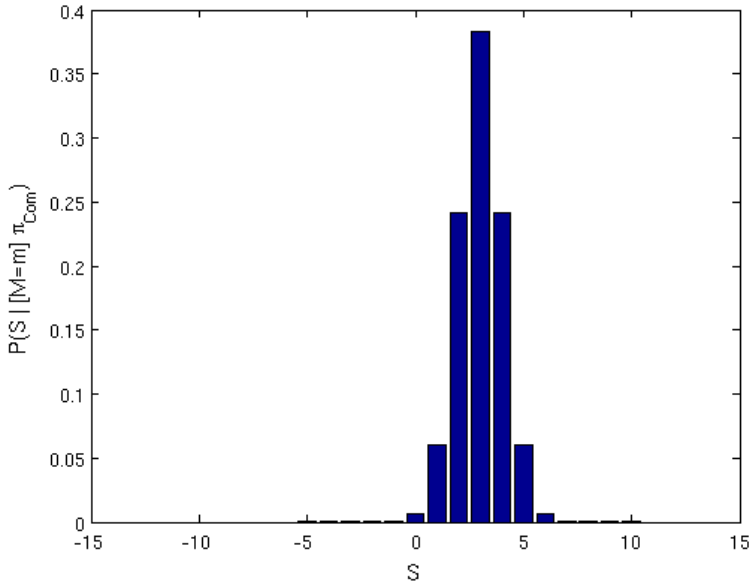


FIGURE 10.3 – Exemple de distribution $P(S \mid [M = m]\ \pi_{Com})$ dans le cas où TransMS$(m) = 3$ et $\sigma_{Env} = 1$. Au plus σ_{Env} sera grand, au plus les valeurs possibles pour S pourront s'écarter de la valeur TransMS(m). Ceci modélise la notion de bruit auditif dans l'environnement.

10.2 Modèle d'agent

Le modèle bayésien d'un agent apprenant $a \in A$, où A est l'ensemble des agents de la société, est défini à la Figure 9.2 du Chapitre 9 en notation PBR. La distribution conjointe d'un agent a au temps t correspond à la décomposition :

$$\begin{aligned} P(O_S\ M\ S\ O_L\ C \mid \delta_a(t)\ \pi_{Ag}) \quad = \quad & P(O_S \mid \pi_{Ag}) P(M \mid O_S\ \delta_a(t)\ \pi_{Ag}) P(S \mid M\ \pi_{Ag}) \\ & P(O_L \mid S\ \delta_a(t)\ \pi_{Ag}) P(C \mid O_S\ O_L\ \pi_{Ag}) \end{aligned} \tag{10.2}$$

où π_{Ag} représente les connaissances préalables générales du modèle d'agent communicant décrit au Chapitre 8, et $\delta_a(t)$ les données d'apprentissage collectées par l'agent a au temps t au cours des jeux déictiques auxquels il a participé en tant que locuteur et qu'auditeur. Rappelons que les ensembles $\delta_a(t)$ sont constitués de triplets :

$$(n, x, o_i) \in (([0,t] \times \mathcal{D}_M \times \mathcal{D}_{O_S}) \cup ([0,t] \times \mathcal{D}_S \times \mathcal{D}_{O_L}))\,.$$

Chacun de ces triplets associe à un jeu déictique $n \in [0,t]$ dans lequel l'agent a a participé, soit un couple $(m, o_i) \in \mathcal{D}_M \times \mathcal{D}_{O_S}$ si l'agent a un statut de locuteur au temps n, soit un couple $(s, o_i) \in \mathcal{D}_S \times \mathcal{D}_{O_L}$ s'il a un statut d'auditeur. Les jeux dans lesquels l'agent a n'intervient pas ne sont pas enregistrés dans $\delta_a(t)$. Nous notons $\delta_a^M(t)$ (respectivement $\delta_a^S(t)$) l'ensemble des $N_{App} = 200$ derniers éléments de $\delta_a(t)$ enregistrés en statut de locuteur (respectivement auditeur).

C'est à partir de cet ensemble d'apprentissage $\delta_a(t) \supseteq (\delta_a^M(t) \cup \delta_a^S(t))$, spécifique à chaque agent, que nous allons définir chacune des formes paramétriques de π_{Ag} qui évoluent aux cours des jeux déictiques.

10.2.1 Système moteur

Le système moteur d'un agent a correspond à une famille de distributions sur ses gestes moteurs M conditionnée par les objets de l'environnement O_S. On parle de prototypes moteurs. L'état initial de la distribution est défini par :

$$P(M \mid O_S \ \delta_a(0) \ \pi_{Ag}) = \boldsymbol{U}(M).$$

Ceci revient à considérer une absence de connaissances motrices en début de simulation. Puis, tous les N_{App} jeux déictiques dans lesquels l'agent a a été en situation de locuteur, nous définissons l'apprentissage de la distribution motrice par :

$$P(M \mid [O_S = o_i] \ \delta_a(t) \ \pi_{Ag}) = \boldsymbol{G}_{\delta_a^{M,o_i}(t)}(M), \tag{10.3}$$

où $\delta_a^{M,o_i}(t)$ est l'ensemble $\delta_a^M(t)$ restreint aux éléments pour lesquels $[O_S = o_i]$ et la loi gaussienne apprise $\boldsymbol{G}$ est définie à la Section 6.3.1.3.

Pour un objet o_i donné, les paramètres de la distribution $P(M \mid [O_S = o_i] \ \delta_a(t) \ \pi_{Ag})$ correspondent donc à la moyenne et l'écart-type des N_{App} derniers gestes moteurs tirés par l'agent a en situation de locuteur devant l'objet o_i.

10.2.2 Lien sensori-moteur

La distribution définissant le lien sensori-moteur dans le modèle d'un agent a est considérée comme déjà acquise par les agents et est donc indépendante de l'ensemble d'apprentissage $\delta_a(t)$. Elle est définie par :

$$\forall m \in \mathcal{D}_M \ : P(S \mid [M = m] \ \pi_{Ag}) = G_{\mathrm{TransMS}(m),\sigma_{Ag}}(S), \tag{10.4}$$

où σ_{Ag} correspond à incertitude de l'agent concernant la transformation articulatori-auditive. Nous considérons ce paramètre, ainsi que la fonction TransMS comme identiques pour tous les agents de la société.

10.2.3 Système auditif

Le système auditif est un classifieur défini par un appel à un sous-modèle :

$$P(O_L \mid S \ \delta_a(t) \ \pi_{Ag}) = P(O_L \mid S \ \delta_a(t) \ \pi_{Class}),$$

où π_{Class} spécifie la décomposition suivante :

$$P(S \ O_L \mid \delta_a(t) \ \pi_{Class}) = P(O_L \mid \pi_{Class}) P(S \mid O_L \ \delta_a(t) \ \pi_{Class}). \tag{10.5}$$

Le premier terme de cette décomposition est la probabilité a priori d'un objet de l'environnement, définie par une loi uniforme :

$$P(O_L \mid \pi_{Class}) = \boldsymbol{U}(O_L).$$

Le second terme représente les prototypes auditifs pour chaque objet de l'environnement. Initialement, ils sont définis par :

$$P(S \mid O_L\ \delta_a(0)\ \pi_{Class}) = \boldsymbol{U}(S).$$

Ceci revient à considérer une absence de connaissances auditives en début de simulation. Puis, tous les N_{App} jeux déictiques dans lesquels l'agent a a été en situation d'auditeur, nous définissons l'apprentissage des prototypes auditifs par :

$$P(S \mid [O_L = o_i]\ \delta_a(t)\ \pi_{Class}) = \boldsymbol{G}_{\delta_a^{S,o_i}(t)}(S), \tag{10.6}$$

où $\delta_a^{S,o_i}(t)$ est l'ensemble $\delta_a^S(t)$ restreint aux élément pour lesquels $[O_L = o_i]$ et la loi gaussienne apprise $\boldsymbol{G}$ est définie à la Section 6.3.1.3.

Pour un objet o_i donné, les paramètres de la distribution $P(S \mid [O_L = o_i]\ \delta_a(t)\ \pi_{Class})$ correspondent donc à la moyenne et l'écart-type des N_{App} derniers stimuli auditifs perçus par l'agent a en situation d'auditeur devant l'objet o_i.

Par inférence bayésienne sur la distribution conjointe de π_{Class} (Équation 10.5), on a alors :

$$P(O_L \mid S\ \delta(t)\ \pi_{Ag}) = \frac{P(S \mid O_L\ \delta_a(t)\ \pi_{Class})}{\sum_{o_i \in O_L} P(S \mid [O_L = o_i]\ \delta_a(t)\ \pi_{Class})}.$$

Pour alléger les notations, nous considérons les connaissances représentées par π_{Class} comme incluses dans π_{Ag} et nous permettons donc de noter :

$$P(O_L \mid S\ \delta(t)\ \pi_{Ag}) = \frac{P(S \mid O_L\ \delta_a(t)\ \pi_{Ag})}{\sum_{o_i \in O_L} P(S \mid [O_L = o_i]\ \delta_a(t)\ \pi_{Ag})}. \tag{10.7}$$

Nous disposons maintenant de la définition complète du programme bayésien d'un agent apprenant, que nous synthétisons Figure 10.4.

Notons que ce programme bayésien correspond au modèle général d'agent π_{Ag} et peut donc se décliner dans ses versions implémentant une théorie motrice, auditive ou sensori-motrice par les simplifications des Équations 8.3, 8.4 et 8.5, respectivement.

Notons également qu'il convient de bien différencier la distribution définissant la transformation articulatori-auditive de l'environnement (Équation 10.1) et celle définissant la connaissance des agents quant à cette transformation (Équation 10.4). La première calcule le stimulus reçu par l'agent auditeur à partir du geste produit par le locuteur dans un jeu déictique et est paramétrée par σ_{Env} (le bruit de l'environnement), alors que la seconde est utilisée dans les inférences bayésiennes définissant les comportements des agents dans le Chapitre 8 et est paramétrée par σ_{Ag} (l'incertitude des agents sur cette transformation).

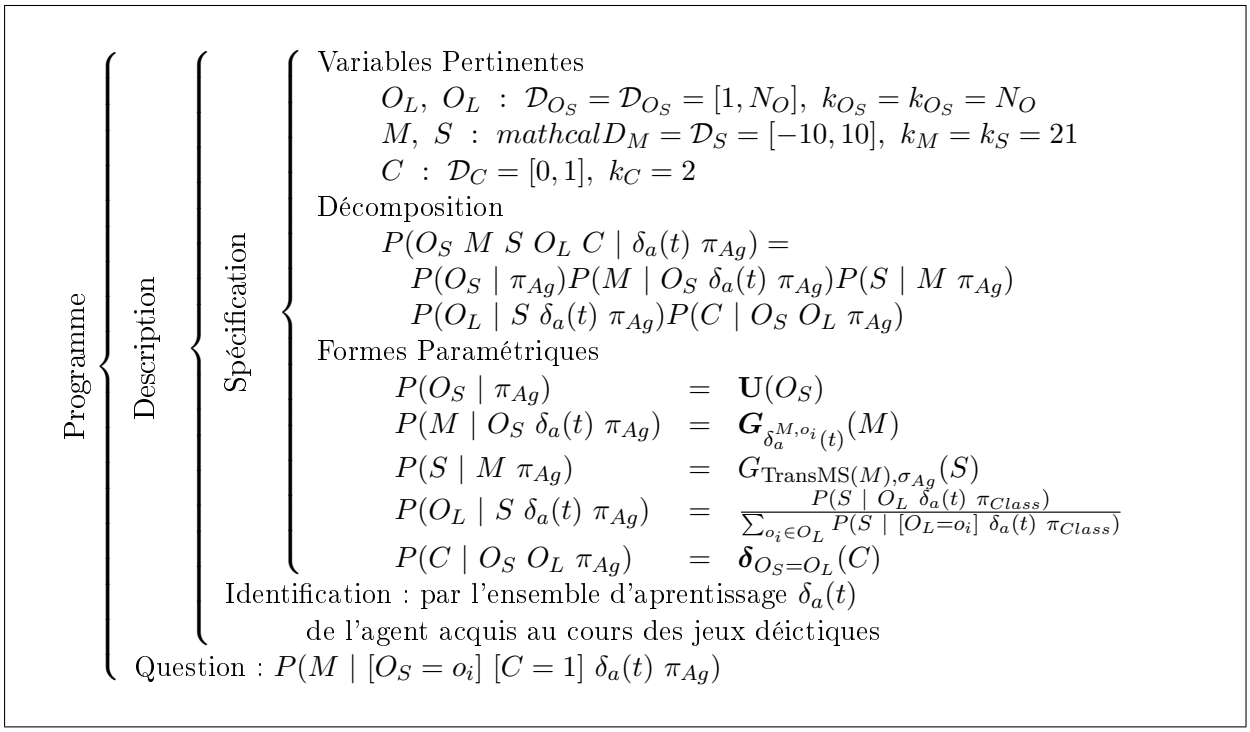


FIGURE 10.4 – Programme bayésien d'un agent apprenant a au temps t ($t \geq N_{App}$).

Dans la suite de ce chapitre, nous considérons donc une société d'agents apprenants tels que définis par la Figure 10.4, déclinés dans leur version motrice, auditive ou sensori-motrice et évoluant selon l'algorithme d'interaction par jeux déictiques du Chapitre 9.

10.3 Propriétés générales des comportements

Pour nous permettre de dégager des propriétés générales des comportements moteur, auditif et sensori-moteur, nous commençons par étudier le cas particulier défini par :

$$\begin{aligned} NL &= 0, \\ D &= 0, \\ \sigma_{Env} = \sigma_{Ag} &= \varepsilon. \end{aligned}$$

Nous nous intéressons donc au cas où la fonction de transformation articulatori-auditive TransMS est une fonction linéaire (la fonction identité, voir Figure 10.1), la communication n'est pas bruitée ($\sigma_{Env} = \varepsilon$, et ε est une très petite valeur) et où les agents ont une connaissance parfaite de la transformation articulatori-auditive ($\sigma_{Ag} = \sigma_{Env}$). Pour simplifier, nous considérons une loi gaussienne d'écart-type ε comme une loi Dirac, les Équations 10.4 et 10.1 deviennent donc :

$$\forall m \in \mathcal{D}_M\ : P(S \mid [M = m]\ \pi_{Env}) = P(S \mid [M = m]\ \pi_{Ag}) = \boldsymbol{\delta}_m(S). \tag{10.8}$$

10.3.1 Comportement moteur

Rappelons l'Équation du comportement moteur de production, définie à la Section 8.2.1 du Chapitre 8 :

$$P(M \mid [O_S = o_i]\ [C = 1]\ \delta_a(t)\ \pi_{AgM}) \propto P(M \mid [O_S = o_i]\ \delta_a(t)\ \pi_{Ag}). \tag{10.9}$$

Comme nous l'avions déjà remarqué au Chapitre 8, le comportement moteur revient simplement à favoriser la production de gestes M habituellement exécutés pour chacun des objets o_i considérés. Ici, un agent particulier a tire ses gestes moteurs selon une famille de distributions gaussiennes (Équation 10.3) :

$$P(M \mid [O_S = o_i]\ \delta_a(t)\ \pi_{Ag}) = \boldsymbol{G}_{\delta_a^{M,o_i}(t)}(M).$$

Celles-ci sont mises à jour à partir des moyennes et des écarts-types des N_{App} derniers jeux déictiques dans lesquelles l'agent a participé en tant que locuteur. Les informations auditives collectées par l'agent lorsqu'il a le rôle d'auditeur n'interviennent pas dans le processus de mise à jour car le système auditif est absent de l'Équation 10.9. Ainsi, la seule source d'évolution du comportement de production moteur au cours des jeux déictiques correspond au biais des estimateurs de moyennes et d'écart-types sur l'ensemble d'apprentissage, et on prédit donc qu'il ne peut laisser émerger un système phonologique efficace. C'est ce que vérifie la simulation observée Figure 10.5. On observe que le taux de reconnaissance reste constant autour de 25% tout au long de simulation (le niveau du hasard pour 4 objets). En effet, dans le comportement moteur, les différences entre les distributions pour chacun des objets ne proviennent que des approximations réalisées par le processus d'apprentissage (estimations des moyennes et écart-types sur les N_{App} derniers jeux déictiques). Ces gestes n'ont donc aucune raison de former un système phonologique efficace, car l'état des distributions d'un agent ne dépend pas des productions de ses congénères. Le comportement moteur de production ne permet donc pas l'émergence d'un système phonologique efficace dans la société, l'évolution de chaque agent étant indépendante de celle des autres.

Notons toutefois que ce résultat est dépendant du paradigme d'interaction par jeux déictiques que nous avons choisi et argumenté au Chapitre 4. Ce paradigme, dans lequel les deux agents sont doués d'un comportement prélangagier de déixis leur assurant d'identifier correctement le même objet, exclut l'existence d'un retour d'information de l'auditeur au locuteur pour qu'ils jugent du succès ou de l'échec de leur communication. Un tel retour serait en effet inutile car nous considérons que les agents ont tous deux connaissance de l'objet avant leur interaction vocale. Dans ce cadre, le comportement moteur ne peut permettre l'émergence d'un système phonologique efficace pour les raisons évoquées plus haut.

Ce comportement moteur peut pourtant permettre une telle émergence si, au lieu d'assurer la communication *avant* l'interaction vocale, on fournit aux agents un moyen de vérification du succès de leur communication *après* cette interaction, conditionnant la mise à jour de leurs connaissances. Nous avons mené ce type de simulations et en effet, seuls les

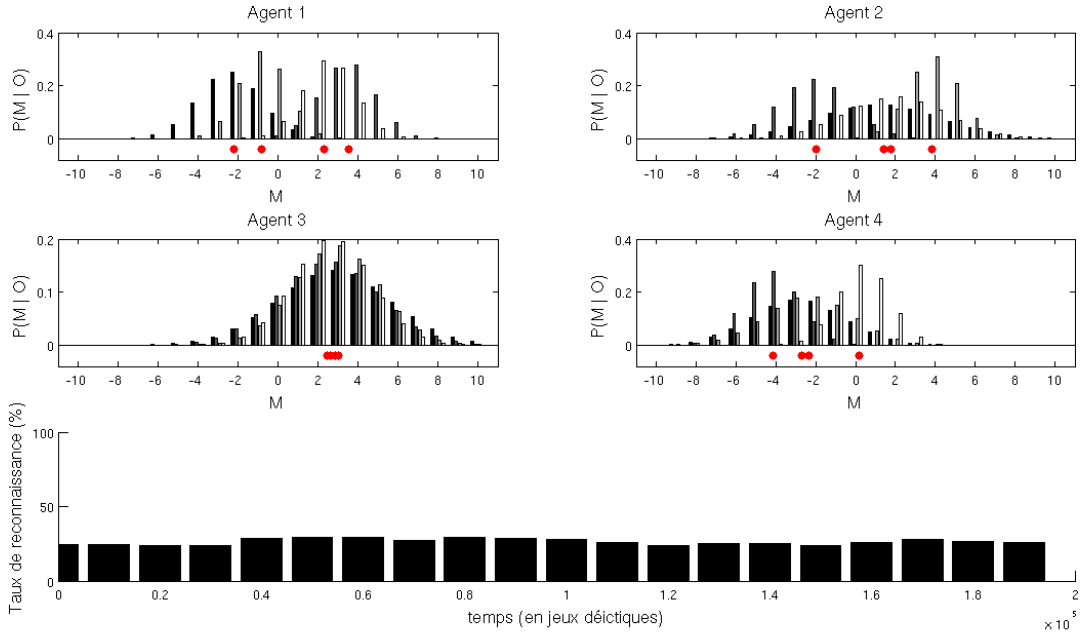


FIGURE 10.5 – Exemple de simulation de 4 agents en comportement moteur dans un environnement de 4 objets. Les 4 fenêtres supérieures indiquent les distributions des gestes moteurs produits par chacun des agents durant les 10 000 derniers jeux déictiques de la simulation (une fenêtre par agent). Dans chacune de ces fenêtres, chaque histogramme représente la distribution de l'agent correspondant pour chacun des 4 objets de l'environnement (différenciés par le niveau de gris des barres, un par objet). Les points sous l'axe des abscisses indiquent la moyenne de chaque distribution. La fenêtre inférieure montre l'évolution du taux de reconnaissance durant la simulation sur des intervalles de 10 000 jeux déictiques, qui reste ici constant autour de 25%. Ce taux est défini à la fin de la Section 9.2.

gestes moteurs permettant un succès de la communication sont alors sélectionnés, ce qui permet l'émergence d'un système phonologique efficace. Mais nous tenons à faire remarquer que ce type de paradigme nous éloigne de la notion de théorie motrice car c'est alors bel et bien par la perception d'un agent auditeur, mettant obligatoirement en jeu des valeurs auditives, que les gestes moteurs d'un agent locuteur sont sélectionnés. D'ailleurs, ce type de simulations produit des résultats équivalent à ceux du comportement sensori-moteur que nous étudierons plus loin dans ce chapitre.

Dans cette thèse, nous considérons qu'une théorie motrice de la production ne doit en aucun cas faire intervenir de connaissances auditives, et un paradigme impliquant un retour vers le locuteur de la perception de l'auditeur n'est donc pas satisfaisant. Ce type de paradigme a été bien étudié dans de Boer (2000).

Terminons cette analyse du comportement moteur en remarquant que ce résultat négatif sur une théorie motrice « pure » reprend sous une forme particulière liée à notre paradigme

de simulation, un résultat théorique utilisé par Schwartz et collab. (2007) indiquant qu'il n'y a pas de capacité prédictive des systèmes phonologiques dans la théorie motrice, résultat d'ailleurs repris par les tenants de la « la théorie motrice de la production » (la phonologie articulatoire, présentée à la Section 3.1.1) eux-mêmes (Studdert-Kennedy et Goldstein, 2003).

10.3.2 Comportement auditif

Rappelons l'Équation du comportement auditif de production, définie à la Section 8.2.2 du Chapitre 8 :

$$\begin{aligned} & P(M \mid [O_S = o_i]\ [C = 1]\ \delta_a(t)\ \pi_{AgA}) \\ & \quad \propto \sum_S P(S \mid M\ \pi_{Ag}) P([O_L = o_i] \mid S\ \delta_a(t)\ \pi_{Ag}) \\ & \quad \propto \sum_S P(S \mid M\ \pi_{Ag}) \frac{P(S \mid [O_L = o_i]\ \delta_a(t)\ \pi_{Ag})}{\sum_{o_j \in O_L} P(S \mid [O_L = o_j]\ \delta_a(t)\ \pi_{Ag})} \quad \text{(d'après 10.7).} \end{aligned} \tag{10.10}$$

Comme nous l'avons remarqué au Chapitre 8, le comportement auditif revient à favoriser la production de gestes moteurs dont la conséquence auditive permettra au système auditif de correctement inférer l'objet de communication o_i considéré. De plus, contrairement au comportement moteur, les données d'apprentissage utilisées dans la distribution auditive de chaque agent de la société proviennent de la production de ces congénères. La Figure 10.6 montre que ce comportement permet l'émergence d'un système phonologique cohérent. On observe que le taux de reconnaissance passe de 25% en début de simulation (le niveau du hasard pour 4 objets) à environ 80% en fin de simulation (c'est-à-dire lorsque les agents produisent les gestes moteurs indiqués par les 4 fenêtres supérieures). Le code de parole émergeant de cette simulation est à la fois efficace, les prototypes étant bien distinguables dans leur espace, et partagé, les agents utilisant des prototypes similaires pour chaque objet.

Afin d'appréhender plus précisément la dynamique cognitive des agents induite par ce comportement au cours de la simulation, intéressons-nous particulièrement à leur système auditif. Celui-ci est un classifieur gaussien (Équation 10.7) dont certaines propriétés sont illustrées Figure 10.7.

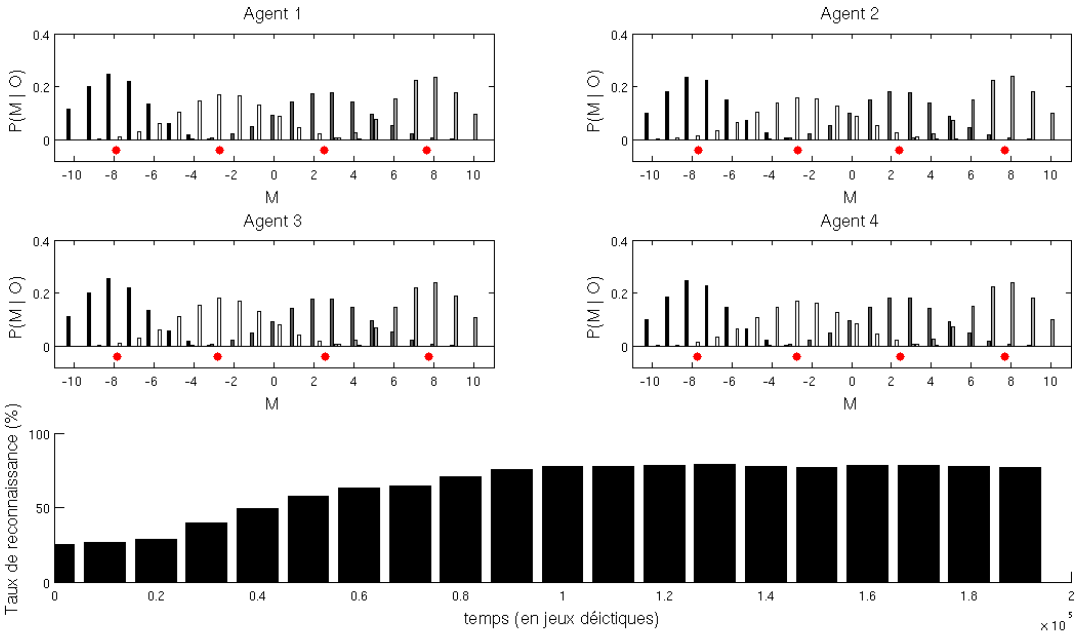


FIGURE 10.6 – Exemple de simulation de 4 agents en comportement auditif dans un environnement de 4 objets, dans les mêmes conventions que la Figure 10.5.

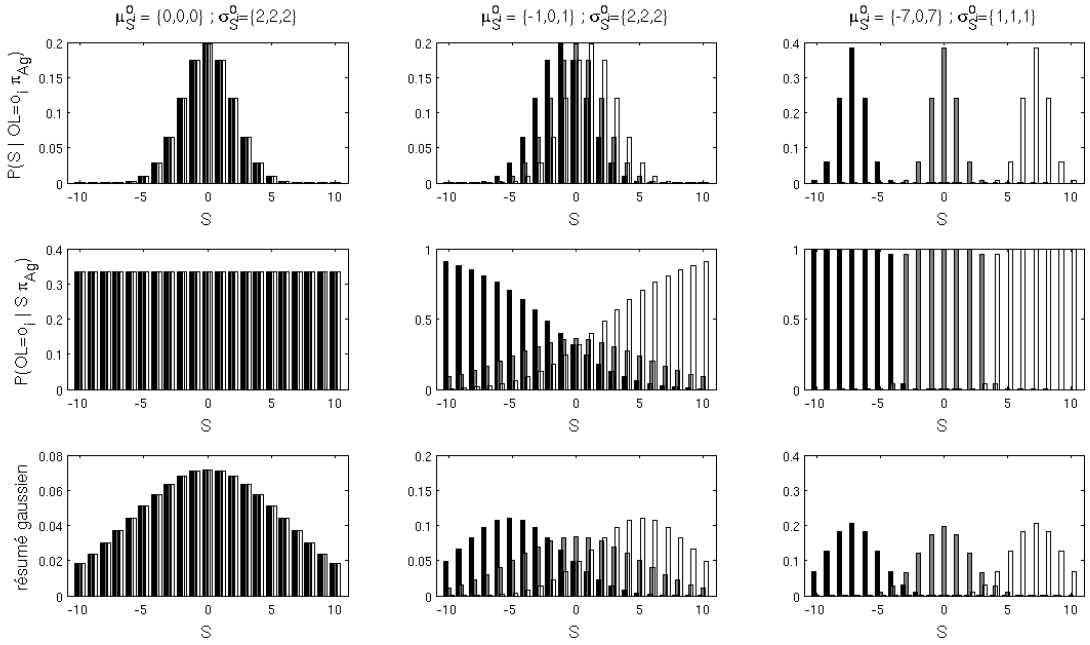


FIGURE 10.7 – Propriétés d'un classifieur gaussien à 3 objets tel qu'exprimé par l'Équation 10.7. Chaque fenêtre de la première ligne montre les distributions gaussiennes $P(S \mid [O_L = o_i]\ \delta_a(t)\ \pi_{Ag})$ pour chacun des 3 objets (3 histogrammes par fenêtre, distingués par leur niveau de gris). On distingue trois conditions : à gauche, les distributions pour chaque objet sont confondues, au milieu elles sont légèrement séparées et à droite, elles sont très séparées. La deuxième ligne montre les classifieurs $P([O_L = o_i] \mid S\ \delta_a(t)\ \pi_{Ag})$ correspondants calculés par l'Équation 10.7, associant à chaque stimulus $s \in S$ la probabilité de reconnaitre l'objet o_i (un niveau de gris par objet). On observe, selon les trois conditions, que l'on passe d'une famille de distribution uniforme (colonne de gauche), pour lesquelles la classification d'un stimulus $s \in S$ est aléatoire, à une famille de « créneaux » (colonne de droite), pour lesquelles la classification est quasi-déterministe. La troisième ligne montre les résumés gaussiens des classifieurs $P([O_L = o_i] \mid S\ \delta_a(t)\ \pi_{Ag})$ [1]. On remarque que la légère séparation de la deuxième colonne est amplifiée dans le résumé gaussien, alors que la forte séparation de la troisième colonne est atténuée.

Étant donné le cas particulier que nous étudions pour l'instant, dans lequel la distribution représentant la connaissance du lien sensori-moteur des agents est déterministe (Équation 10.8) et la fonction TransMS est la fonction identité ($NL = D = 0$), nous

1. Il s'agit de lois gaussiennes dont les moyennes et les écarts-types sont calculés sur l'ensemble des valeurs $s \in S$ pondérées par la probabilité $P([O_L = o_i] \mid [S = s]\ \delta_a(t)\ \pi_{Ag})$.

pouvons considérer la simplification suivante :

$$\begin{aligned} P([M=m] \mid [O_S=o_i]\ [C=1]\ \delta_a(t)\ \pi_{AgA}) \\ \propto \sum_S \boldsymbol{\delta}_m(S) P([O_L=o_i] \mid [S=m]\ \delta_a(t)\ \pi_{Ag}) \\ \propto \frac{P([S=m] \mid [O_L=o_i]\ \delta_a(t)\ \pi_{Ag})}{\sum_{o_j \in O_L} P([S=m] \mid [O_L=o_j]\ \delta_a(t)\ \pi_{Ag})}, \end{aligned} \tag{10.11}$$

où la somme sur S se simplifie grâce à la loi Dirac $\boldsymbol{\delta}_m(S)$ (définie Section 6.3.1.3). Comme l'apprentissage des prototypes auditifs $P(S \mid [O_L = o_i]\ \delta_a(t)\ \pi_{Ag})$ consiste ici en une loi gaussienne estimée à partir des données d'apprentissage auditives dans $\delta_a(t)$ (Équation 10.6), les propriétés du classifieur gaussien exposé à la Figure 10.7 fournissent une bonne représentation du comportement d'un agent pour la simplification considérée dans l'Équation 10.11. Nous pouvons alors réinterpréter la Figure 10.7 en termes cognitifs :

- La première ligne correspond à l'état des prototypes auditifs $P(S \mid [O_L = o_i]\ \delta_a(t)\ \pi_{Ag})$ à un instant donné t de la simulation.
- Le classifieur gaussien de la deuxième ligne résulte de la question du comportement de production auditive simplifiée de l'Équation 10.11.
- Le résumé gaussien de la troisième ligne correspond au processus d'apprentissage, dans lequel les agents ré-estiment les moyennes et les écarts-types des prototypes auditifs $P(S \mid [O_L = o_i]\ \delta_a(t)\ \pi_{Ag})$ à partir des couples (s, o_i) de l'ensemble d'apprentissage.

Ce dernier point revient à considérer que les données d'apprentissage sont uniquement issues du comportement de production de l'agent lui-même, alors qu'elles proviennent dans la simulation de celui de ses congénères. Ceci nous permet de résumer la dynamique complexe d'un système constitué d'un ensemble d'agents probabilistes en interaction à celle d'une séquence d'opération sur un unique classifieur, plus facile à appréhender. Nous allons voir que ces simplifications aident à la compréhension du résultat de la Figure 10.6.

Ainsi, le début de la simulation correspond à la première colonne, où les distributions $P(S \mid [O_L = o_i]\ \delta_a(0)\ \pi_{Ag})$ sont identiques pour chaque objet o_i. Le classifieur renvoie alors des distributions uniformes car il ne possède pas d'information lui permettant de distinguer les objets à partir d'un stimulus.

Le processus d'apprentissage ne faisant que résumer le vécu auditif d'un agent par une loi gaussienne sur une fenêtre de temps limitée, son résultat ne sera en pratique jamais identique à la troisième ligne de la première colonne de la Figure 10.7. Les moyennes et écart-types des gaussiennes se seront certainement légèrement écartés de leurs valeurs initiales, ce qui nous emmène au cas de la deuxième colonne.

Les distributions $P(S \mid [O_L = o_i]\ \delta_a(t)\ \pi_{Ag})$ étant maintenant légèrement différentes (deuxième colonne, première ligne), on voit apparaître dans le classifieur correspondant de la deuxième ligne des zones auditives de « confiance » aux extrémité de l'espace, où il sait classifier correctement un objet.

On observe alors que le résumé gaussien (troisième ligne, deuxième colonne), représentant le processus d'apprentissage, a écarté les prototypes dans l'espace auditif par rapport à son état précédent (première ligne, deuxième colonne).

On voit ainsi que l'état initial des agents est instable. Les approximations nécessairement présentes dans le processus d'apprentissage permettent au classifieur de dégager des zones auditives dans lesquelles il est plus efficace et réalise ainsi une dispersion des prototypes auditifs dans l'espace correspondant.

La troisième colonne illustre la stabilisation de l'état des agents. On considère des prototypes très bien séparés (première ligne, troisième colonne). Dans ce cas, le classifieur est quasiment déterministe : l'espace auditif est divisé en trois zones très nettes pour chacun des objets (deuxième ligne, troisième colonne). On observe alors que les résumés gaussiens (troisième ligne, troisième colonne) issus de ce classifieur en forme de créneaux sont moins dispersés dans l'espace auditif que les prototypes initiaux (première ligne, troisième colonne).

Ainsi, on voit que l'état initial (première colonne) est instable par le processus d'apprentissage, car les approximations inhérentes au processus d'apprentissage impliquent forcément une légère séparation des prototypes qui entraine par la suite leur dispersion (deuxième colonne). Toutefois, cette dispersion trouve sa limite dans les résumés gaussiens des classifieurs déterministes de la troisième colonne. On voir ainsi poindre un état stable assez bien dispersé.

Cette propriété d'évolution du comportement auditif de production vers une dispersion des prototypes peut également être appréhendée plus formellement par l'Équation 10.11. On peut en effet considérer le comportement comme stable lorsque les tirages successifs de gestes moteurs selon la distribution :

$$P([M=m] \mid [O_S=o_i]\ [C=1]\ \delta_a(t)\ \pi_{AgA}) \propto \frac{P([S=m] \mid [O_L=o_i]\ \delta_a(t)\ \pi_{Ag})}{\sum_{o_j \in \mathcal{D}_{O_L}} P([S=m] \mid [O_L=o_j]\ \delta_a(t)\ \pi_{Ag})}$$

ne font plus évoluer les prototypes auditifs $P(S \mid [O_L=o_i]\ \delta_a(t)\ \pi_{Ag})$ (ils sont le seul terme qui évolue dans ce comportement). Comme nous considérons ici une transformation articulatori-auditive déterministe et égale à la fonction identité, les stimuli auditifs $s \in S$ de l'ensemble d'apprentissage correspondent justement aux gestes moteurs $m \in M$ générés par le comportement. Ainsi, le comportement est stable si tirer selon $P(M \mid [O_S=o_i]\ [C=1]\ \delta_a(t)\ \pi_{AgA})$ revient à tirer selon $P(S \mid [O_L=o_i]\ \delta_a(t)\ \pi_{Ag})$. En termes probabilistes, cela revient à considérer les deux distributions comme proportionnelles et donc le dénominateur de l'Équation précédente comme une distribution uniforme, soit :

$$\sum_{o_j \in \mathcal{D}_{O_L}} P(S \mid [O_L=o_j]\ \delta_a(t)\ \pi_{Ag}) \propto \boldsymbol{U}(S). \tag{10.12}$$

Une solution pour satisfaire cela est de répartir au mieux les prototypes auditifs dans leur espace, et donc de les disperser. La Figure 10.8 illustre ce principe.

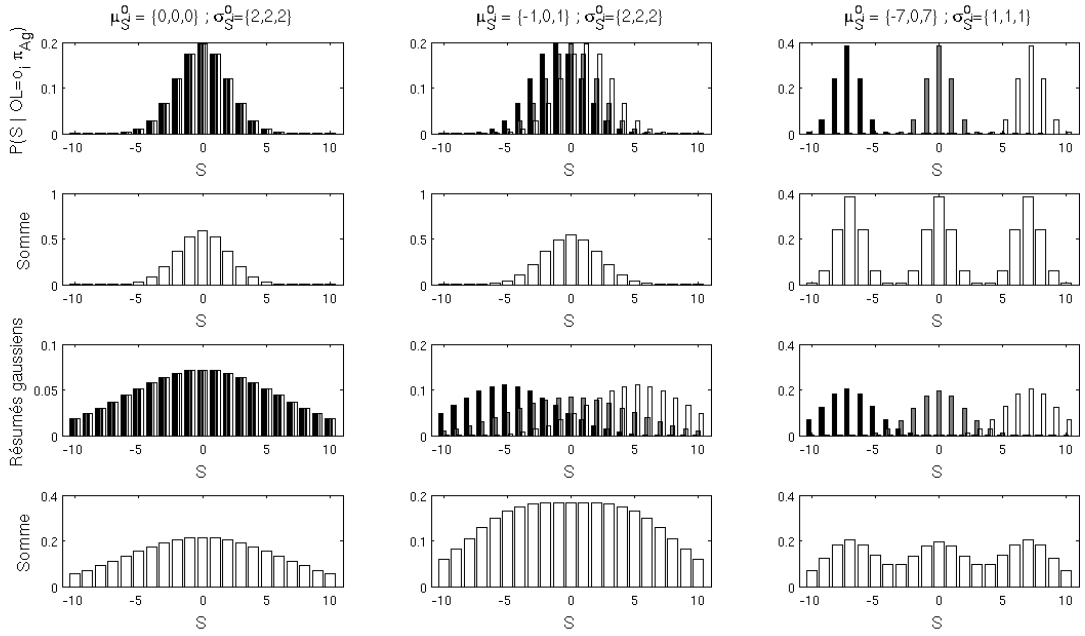


FIGURE 10.8 – Illustration de la condition de stabilité du comportement auditif exprimée par l'Équation 10.12. La première ligne est identique à celle de la Figure 10.7 et montre les distributions gaussiennes $P(S \mid [O_L = o_i]\ \delta_a(t)\ \pi_{Ag})$. On distingue toujours trois conditions : à gauche, les distributions pour chaque objet sont confondues, au milieu elles sont légèrement séparées et à droite, elles sont très séparées. La deuxième ligne montre les sommes des distributions de la première ligne (Équation 10.12). La troisième ligne est identique à celle de la Figure 10.7 et montre les résumés gaussiens des distributions $P([O_L = o_i] \mid S\ \delta_a(t)\ \pi_{Ag})$, représentant le pas suivant du processus d'apprentissage par rapport à la première ligne. La quatrième ligne montre la somme des résumés de la troisième ligne. On constate que dans chaque colonne, la distribution « somme » est plus proche d'une loi uniforme après le résumé gaussien (représentant ici un pas d'apprentissage) qu'avant. Il semble donc bien que le comportement auditif cherche à vérifier l'Équation 10.12, une solution étant la dispersion uniforme des prototypes auditifs dans leur espace.

10.3.3 Comportement sensori-moteur

Rappelons l'Équation du comportement sensori-moteur de production, définie à la Section 8.2.3 du Chapitre 8 :

$$\begin{aligned} & P(M \mid [O_S = o_i]\ [C = 1]\ \delta_a(t)\ \pi_{AgSM}) \\ & \quad \propto \quad P(M \mid [O_S = o_i]\ \delta_a(t)\ \pi_{Ag}) \sum_S P(S \mid M\ \pi_{Ag}) P([O_L = o_i] \mid S\ \delta_a(t)\ \pi_{Ag}) \\ & \quad \propto \quad P(M \mid [O_S = o_i]\ [C = 1]\ \delta_a(t)\ \pi_{AgM}) \\ & \qquad\quad P(M \mid [O_S = o_i]\ [C = 1]\ \delta_a(t)\ \pi_{AgA}). \end{aligned} \tag{10.13}$$

Comme nous l'avions remarqué au Chapitre 8, le comportement sensori-moteur correspond au produit des comportements moteur et auditif (comme l'indiquent les deux dernières lignes de l'Équation ci-dessus). Il réalise ainsi un compromis entre un comportement moteur de nature conservative (qui favorise les gestes moteurs habituellement utilisés pour un objet donné, mais sans principe d'optimisation) et un comportement auditif de nature dispersive (qui cherche à satisfaire un classifieur en répartissant les prototypes auditifs dans leur espace). La Figure 10.9 montre que ce comportement permet l'émergence rapide d'un système phonologique très efficace dans la société d'agents. On observe en effet que dès les premiers 10 000 jeux déictiques, le taux de reconnaissance à déjà dépassé celui du hasard (25% pour 4 objets) et atteint 100% lors des 20 000 suivants.

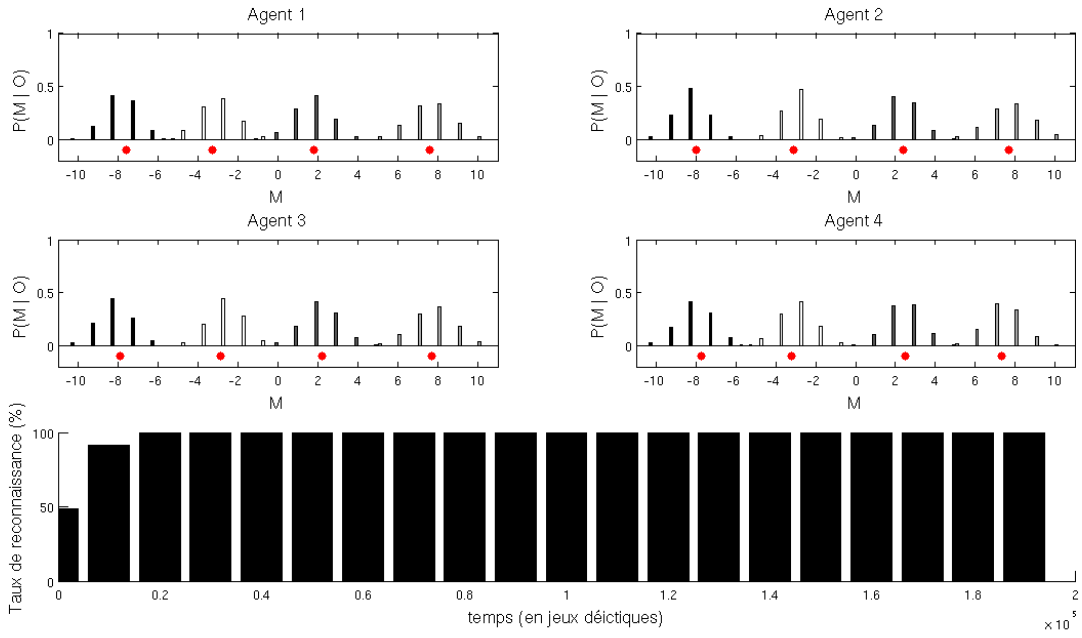


FIGURE 10.9 – Exemple de simulation de 4 agents en comportement sensori-moteur dans un environnement de 4 objets, dans les mêmes conventions que la Figure 10.5.

Afin d'appréhender le principe d'optimisation à l'œuvre dans ce comportement, considérons de nouveau la simplification articulatori-auditive de l'Équation 10.8. Sous cette

hypothèse, l'Équation 10.13 devient :

$$\begin{aligned}
&P([M=m] \mid [O_S=o_i]\ [C=1]\ \delta_a(t)\ \pi_{AgSM}) \\
&\quad\propto\ P([M=m] \mid [O_S=o_i]\ \delta_a(t)\ \pi_{Ag}) \sum_S \boldsymbol{\delta}_m(S) P([O_L=o_i] \mid S\ \delta_a(t)\ \pi_{Ag}) \\
&\quad\propto\ P([M=m] \mid [O_S=o_i]\ \delta_a(t)\ \pi_{Ag}) P([O_L=o_i] \mid [S=m]\ \delta_a(t)\ \pi_{Ag}) \\
&\quad\propto\ \underbrace{P([M=m] \mid [O_S=o_i]\ \delta_a(t)\ \pi_{Ag})}_{\text{terme moteur}} \underbrace{\frac{P([S=m] \mid [O_L=o_i]\ \delta_a(t)\ \pi_{Ag})}{\sum_{o_j \in \mathcal{D}_{O_L}} P([S=m] \mid [O_L=o_j]\ \delta_a(t)\ \pi_{Ag})}}_{\text{terme auditif}}
\end{aligned} \tag{10.14}$$

La Figure 10.10 peut permettre de rendre compte de la dynamique cognitive des agents de la société menant au résultat de la Figure 10.9. Chaque colonne de la Figure 10.10 représente un instant différent de la simulation : à gauche l'état initial, à droite l'état après convergence, au milieu un état intermédiaire. La première ligne est le terme moteur de l'Équation 10.14, la seconde le terme auditif et la troisième le produit des deux. Nous traduisons donc la simplification de l'Équation 10.8 par une égalité des prototypes moteurs (première ligne) et auditifs (à partir desquels sont calculés les classifieurs de la deuxième ligne).

La première colonne semble stable, la distribution issue du comportement sensorimoteur étant égale aux prototypes initiaux (égalité de la première et de la troisième colonne). Elle ne l'est en réalité pas pour la même raison que dans notre analyse du comportement auditif : les prototypes proviennent d'estimations sur des données d'apprentissage et ne sont donc en pratique jamais parfaitement égaux, ce qui nous emmène au cas de la deuxième colonne. Les prototypes étant maintenant légèrement séparés, on observe que les gestes issus du comportement amplifient cette séparation (comparaison de la première et de la troisième ligne de la deuxième colonne). Mais contrairement au comportement auditif étudié plus haut, on constate ici que la troisième colonne semble stable : la première ligne représentant les prototypes moteurs (et auditifs, d'après notre simplification) est identique à la troisième représentant le comportement de l'Équation 10.14. Autrement dit, les tirages successifs de gestes moteurs réalisés dans cet état ne devraient plus faire évoluer les prototypes moteurs et auditifs de la troisième colonne, et le comportement peut alors être considéré comme stable.

Plus formellement, nous pouvons considérer ce comportement comme stable lorsque les tirages successifs selon $P(M \mid [O_S = o_i]\ [C = 1]\ \delta_a(t)\ \pi_{AgSM})$ ne font plus évoluer les prototypes moteurs $P(M \mid [O_S = o_i]\ \delta_a(t)\ \pi_{Ag})$. D'après l'Équation 10.14, il suffit que le terme auditif :

$$\frac{P([S=m] \mid [O_L=o_i]\ \delta_a(t)\ \pi_{Ag})}{\sum_{o_j \in \mathcal{D}_{O_L}} P([S=m] \mid [O_L=o_j]\ \delta_a(t)\ \pi_{Ag})}$$

n'ait pas d'influence sur les prototypes moteurs. Une condition pour cela est que, dans la zone d'un prototype moteur donné, le classifieur se comporte comme une loi uniforme. C'est

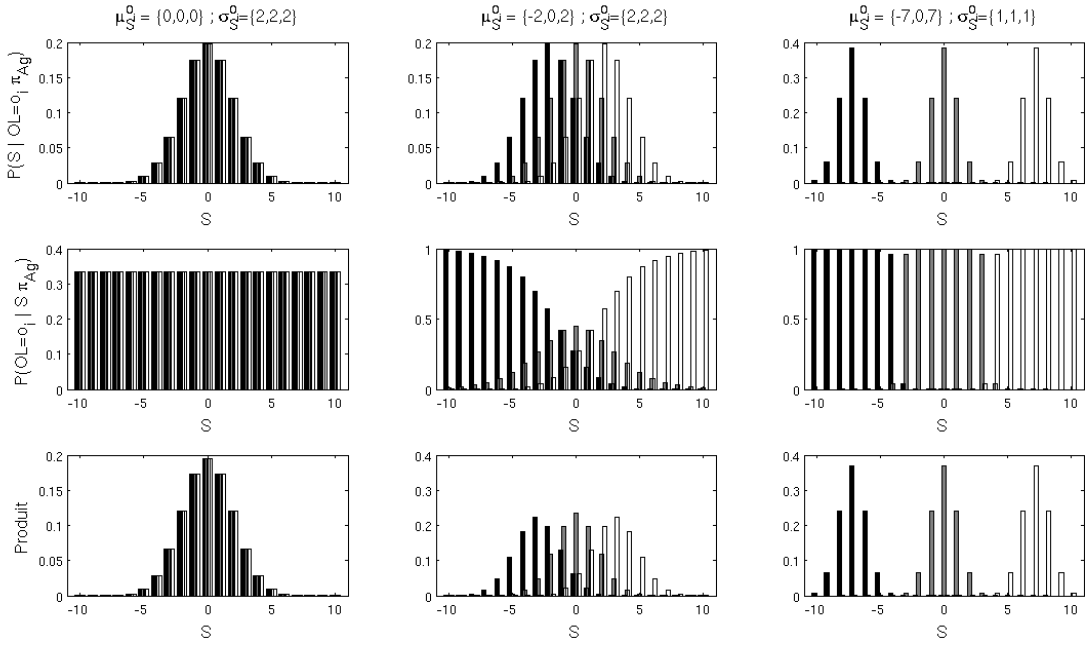


FIGURE 10.10 – Illustration du comportement sensori-moteur de l'Équation 10.13. La première ligne montre les distributions gaussiennes $P(S \mid [O_L = o_i]\ \delta_a(t)\ \pi_{Ag})$. On distingue toujours trois conditions : à gauche, les distributions pour chaque objet sont confondues, au milieu elles sont légèrement séparées et à droite, elles sont très séparées. La deuxième ligne montre les distributions $P([O_L = o_i] \mid S\ \delta_a(t)\ \pi_{Ag})$ calculées par l'Équation 10.7. On observe, selon les trois conditions, que l'on passe d'une famille de distribution uniforme (colonne de gauche) pour lesquelles la classification est très incertaine à une famille de « créneaux » (colonne de droite) pour lesquelles la classification est quasi-déterministe. La troisième ligne correspond au résumé gaussien du produit des deux premières. On remarque que la légère séparation de la deuxième colonne est amplifiée par le produit, alors que la forte séparation de la troisième colonne reste inchangée. Voir texte pour une interprétation en terme de comportement sensori-moteur.

le cas lorsque les lois gaussiennes des prototypes moteurs et les créneaux du classifieur, pour un objet donné, se trouvent dans la même zone de l'espace. C'est bien ce que l'on observe à la troisième colonne de la Figure 10.10. Ainsi, le comportement sensori-moteur va favoriser des prototypes moteurs calés sur des « bons » classifieurs (idéalement déterministes en forme de créneaux). Ceci fournit des éléments permettant de comprendre comment ce comportement obtient les meilleurs scores de reconnaissance.

10.3.4 Conclusion sur les propriétés générales

Cette première analyse des principes d'optimisation à l'œuvre dans chacun des trois comportements nous permet de dégager des caractéristiques générales de chacun.

Dans le paradigme d'interaction par jeux déictiques que nous considérons, le comportement moteur ne peut mener à l'émergence d'un système phonologique cohérent. Ce comportement ne permet en effet pas aux agents de prendre en compte les besoins de l'auditeur lorsqu'ils sont en situation de locuteur car les seules données d'apprentissage disponibles sont les gestes moteurs qu'ils ont eux-même produits. Il est d'ailleurs intéressant de remarquer qu'à notre connaissance, il n'existe pas de théorie de la forme des systèmes phonologiques qui ne prennent en compte que des connaissances motrices. Par exemple, dans les deux théories que nous avons présentées à la Section 4.1 du Chapitre 4, l'optimisation auditive joue un rôle prépondérant. Il en est de même pour les modèles computationnels d'agents interagissants présentés à la Section 4.3 du même chapitre dans lesquels, si les connaissances motrices peuvent en être absentes, ce n'est jamais le cas des connaissances auditives. Dans ce qui suit, nous ne nous intéresserons donc plus au comportement moteur et concentrerons nos analyses sur les comportements auditifs et sensorimoteurs.

Le comportement auditif permet l'émergence de systèmes phonologiques relativement efficaces. Ce comportement cherche en effet à disperser les prototypes auditifs uniformément dans leurs espaces à partir de données sensorielles d'apprentissage provenant de l'ensemble de la société d'agents. Ses performances sont toutefois limitées en termes de taux de reconnaissance et de temps de convergence.

Le comportement sensori-moteur quant à lui permet l'émergence de systèmes phonologiques très efficaces et converge très rapidement. L'ajout de connaissances motrices permet en effet de fixer les prototypes moteurs sur les bons prototypes auditifs en terme de classification et permet ainsi de relâcher la contrainte de dispersion uniforme à l'œuvre dans le comportement auditif. Les gestes moteurs choisis sont ainsi confinés dans les zones où le classifieur auditif a les meilleurs performances, évitant d'utiliser des zones de classifications ambiguës.

10.4 Évaluation détaillée des comportements

Dans la section précédente, nous nous limitions à étudier les propriétés générales des trois comportements de production (moteur, auditif et sensori-moteur) en simplifiant le lien sensori-moteur des agents par une fonction déterministe égale à la fonction identité (Équation 10.8). Ceci nous a permis d'étudier les propriétés des systèmes moteurs et auditifs indépendamment (comportement moteur et auditif, respectivement) ainsi que de leur produit (comportement sensori-moteur), et de relier leur dynamique à des principes généraux d'optimisation. Nous allons maintenant procéder à une analyse plus complète dans laquelle nous étudierons les propriétés des comportements auditif et sensori-moteur (en laissant de côté le comportement moteur non-dispersif et donc ne permettant pas de simuler une communication de manière adéquate) lorsque nous faisons varier les paramètres de la transformation articulatori-auditive. Pour cela, nous commençons par définir les paramètres variés dans nos simulations ainsi que les mesures permettant leur évaluation.

10.4.1 Paramètres variés dans les simulations

Pour chaque du modèle de théorie de la communication (auditif π_{AgA}, ou sensori-moteur π_{AgSM}), nous nous intéressons :

- aux variations conjointes du bruit de l'environnement σ_{Env} et de l'incertitude des agents σ_{Ag} sur la transformation articulatori-auditive, lorsque celle-ci est linéaire (fonction identité) ;
- aux variations conjointes de la force de la non-linéarité NL présente dans la fonction de transformation articulatori-auditive et du bruit de l'environnement σ_{Env} ;
- à la variation de la position du point d'inflexion D dans le cas d'une transformation fortement non-linéaire.

10.4.2 Évaluation

Afin de comparer les performances et propriétés des comportements moteur et auditif, nous utilisons les mesures suivantes :

- le taux de reconnaissance dans la société d'agents en fin de simulation ;
- la dispersion des stimuli auditifs choisis par les agents pour chaque objet en fin de simulation.

Le taux de reconnaissance est défini comme le pourcentage de succès de communication sur les 10 000 derniers jeux déictiques d'une simulation, c'est-à-dire le pourcentage de jeux déictiques pour lesquels l'agent auditeur a correctement inféré l'objet de la communication à partir de son comportement de perception, sur la seule base du stimulus reçu de l'agent locuteur. Cette mesure permet d'évaluer la capacité des agents à évoluer d'une communication déictique, nécessitant une attention partagée sur un objet présent dans l'environnement, à une communication uniquement vocale dans laquelle l'objet ne serait pas connu avant l'interaction.

Une mesure de la dispersion des stimuli auditifs choisis par les agents pour chaque objet nous est fournie par Liljencrants et Lindblom (1972) (voir Section 4.1.1) :

$$G = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(\frac{1}{d_{i,j}} \right)^2 \tag{10.15}$$

où n est le nombre d'éléments dans le système phonologique considéré (ici le nombre d'objets N_O) et $d_{i,j}$ est la distance auditive entre deux éléments i et j. Nous définissons cette dernière comme la distance entre les moyennes des stimuli auditifs produits par l'ensemble des agents de la société pour les objets o_i et o_j sur les 10 000 derniers jeux déictiques d'une simulation particulière.

La mesure de Lindblom renvoie une valeur en relation inverse de la dispersion globale du système. Elle a en particulier la propriété de vérifier si tous les éléments sont correctement séparés. Si c'est le cas, les inverses des carrés des distances auront toutes de petites valeurs. Mais dans le cas où il existe deux éléments proches, l'inverse du carré de leur distance renverra une grande valeur, qui tend rapidement vers l'infini. Cette mesure est

donc performante pour détecter les mauvais systèmes, mais n'est pas très appropriée pour comparer les dispersions au sein de différents bons systèmes. Pour atténuer cet effet, nous observerons son logarithme.

Pour chaque jeu de paramètres, nous calculons les moyennes et écarts-types des mesures ci-dessus sur un ensemble de 10 simulations indépendantes.

10.4.3 Effet du bruit de l'environnement σ_{Env} et de l'incertitude des agents σ_{Ag}

Nous nous intéressons aux variations conjointes du bruit de l'environnement σ_{Env} et de l'incertitude des agents σ_{Ag} et analysons leurs effets sur le taux de reconnaissance et la mesure de Lindblom. Les résultats présentés proviennent d'un jeu de 490 simulations par type de comportement (auditif ou sensori-moteur) : 10 simulations indépendantes pour chaque combinaison possible de $\sigma_{Env} \in \{0.1, 0.4, 1, 2, 5, 9, 15\}$ et $\sigma_{Ag} \in \{0.1, 1, 3, 7, 12, 25, 40\}$. Ces simulations concernent une société de $N_A = 4$ agents évoluant dans un environnement de $N_O = 4$ objets sur une succession de $N_{JD} = 200\ 000$ jeux déictiques.

10.4.3.1 Comportement auditif

La Figure 10.11 montre l'effet du bruit de l'environnement et de l'incertitude des agents sur les taux de reconnaissance en fin de simulations.

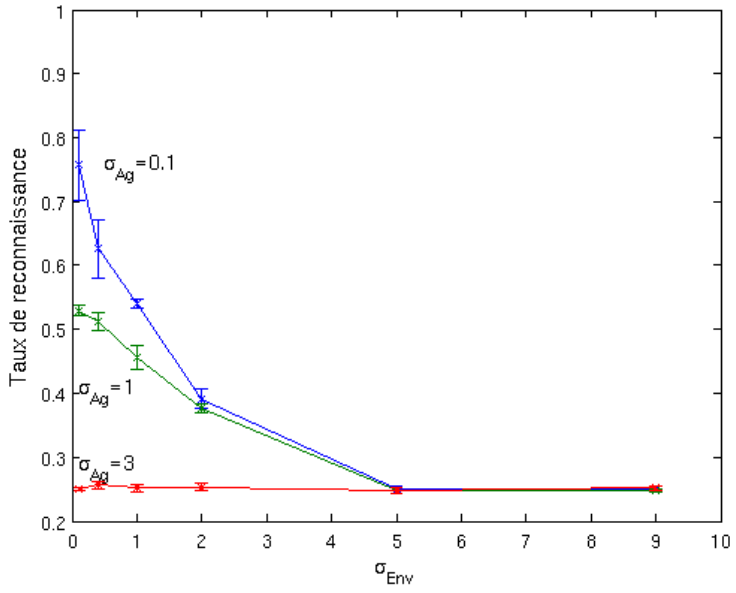


FIGURE 10.11 – Taux de reconnaissance en fin de simulation dans une société de 4 agents en comportement auditif, et un environnement de 4 objets, en fonction du bruit de l'environnement et de l'incertitude des agents. Moyennes et écarts-types sur 10 simulations indépendantes.

On observe que même dans le meilleur des cas où le bruit de l'environnement et l'incertitude des agents sont quasi-nuls ($\sigma_{Env} = \sigma_{Ag} = 0.1$), le taux de reconnaissance dans la société en fin de simulation n'atteint qu'environ 75%. L'explication vient de l'analyse du comportement auditif à laquelle nous avons procédé à la section précédente : pour des raisons d'optimisation des performances du classifieur, les prototypes auditifs émergeant de ce comportement tendent à occuper uniformément tout l'espace auditif, et ainsi à se chevaucher (voir Figure 10.6).

La deuxième observation est que ce comportement est très sensible à l'augmentation de la valeur de ces paramètres. En particulier, les taux de reconnaissance chutent rapidement lorsque l'on augmente la valeur de σ_{Ag}. L'explication vient de l'Équation 10.10 régissant ce comportement, dans laquelle la distribution représentant le lien sensori-moteur (paramétrée par σ_{Ag}) permet le passage des « bons » stimuli auditifs aux gestes correspondants. Si cette connaissance est dégradée, les performances du comportement en pâtissent radicalement. L'augmentation de la valeur du paramètre σ_{Env} quant à elle a pour effet d'augmenter la variance des prototypes auditifs, ce qui entraîne une dégradation des performances du classifieur. Comme le comportement auditif repose en grande partie sur ce dernier (Section 10.3.2), cherchant à produire des gestes pour lesquels le stimulus auditif correspondant permet de correctement classifier l'objet, ces performances en sont également rapidement affectées.

La Figure 10.12 montre l'effet du bruit de l'environnement et de l'incertitude des agents sur la mesure de Lindblom du système obtenu en fin de simulation. Les distances $d_{i,j}$ intervenant dans l'Équation 10.15 correspondent à la distance entre les moyennes des stimuli auditifs produits par l'ensemble des agents sur les 10 000 derniers jeux déictiques d'une simulation, avant ajout du bruit dans l'environnement. L'observation faite ici est très liée à celle concernant les taux de reconnaissance : l'augmentation des valeurs de σ_{Env} et σ_{Ag} dégrade rapidement le pouvoir de dispersion du comportement auditif, ce qui fait chuter les taux de reconnaissance. Pour l'illustrer, la Figure 10.13 montre trois simulations particulières où l'on fait varier σ_{Env} pour $\sigma_{Ag} = 0.1$.

10.4.3.2 Comportement sensori-moteur

La Figure 10.14 montre l'effet du bruit de l'environnement et de l'incertitude des agents sur les taux de reconnaissance en fin de simulation. On observe des performances bien meilleures que celles du comportement auditif. Premièrement, le taux de reconnaissance est bien supérieur et peut atteindre 100%. Deuxièmement, les performances du système sont bien plus robustes à l'augmentation des deux paramètres. Les performances sont en effet similaires pour des valeurs de σ_{Ag} entre 0.1 et 3 (cette dernière valeur n'étant pas négligeable étant donnée la taille de l'espace auditif). Même pour une connaissance très incertaine du lien sensori-moteur ($\sigma_{Ag} = 12$), le taux atteint 70% pour de petites valeurs de σ_{Env}. Concernant ce dernier paramètre, le plateau observé au début des courbes témoigne également d'une meilleure robustesse que pour le comportement auditif.

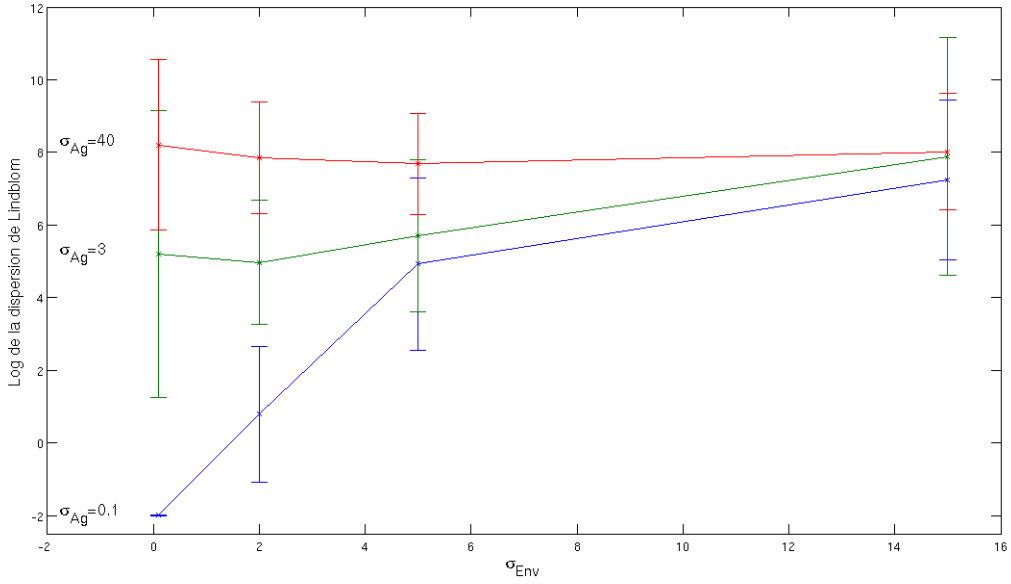


FIGURE 10.12 – Logarithme de la mesure de Lindblom en fin de simulation dans une société de 4 agents en comportement auditif, et un environnement de 4 objets, en fonction du bruit de l'environnement et de l'incertitude des agents. Moyennes et écart-types sur 10 simulations indépendantes. Les mesures de dispersion sont réalisées sur les moyennes des prototypes auditifs de l'ensemble des agents avant l'ajout du bruit par l'environnement.

La Figure 10.15 montre l'effet du bruit de l'environnement et de l'incertitude des agents sur la dispersion des moyennes des prototypes auditifs en fin de simulations. Outre le fait qu'ici encore le comportement sensori-moteur semble bien plus robuste que son homologue auditif, un point intéressant est la décroissance au début de la courbe correspondant à $\sigma_{Ag} = 0.1$. Le maximum de dispersion (c'est-à-dire le minimum de la mesure de Lindblom) n'est donc plus ici lié au minimum de bruit de l'environnement. Les prototypes auditifs semblent moins se disperser à faible bruit qu'à bruit plus fort, pour s'approcher de nouveau ensuite.

Ce phénomène est illustré à la Figure 10.16 pour une valeur de σ_{Ag} constante à 0.1, sur des simulations particulières dont les valeurs de dispersion sont proches des moyennes de la Figure 10.15. On observe que les moyennes des prototypes auditifs sont un peu mieux dispersées pour $\sigma_{Env} = 2$ que pour $\sigma_{Env} = 0.1$, et très mal dispersées pour $\sigma_{Env} = 5$. L'interprétation de ce phénomène vient des éléments d'optimisation que nous avions fournis à la section précédente : le comportement sensori-moteur cherche à placer ses prototypes moteurs pour obtenir de bons classifieurs (en forme de créneaux) sur leurs conséquences auditives : si peu de bruit est présent dans l'environnement, il n'est pas nécessaire de beaucoup séparer les prototypes auditifs pour obtenir de bons classifieurs. En ajoutant du bruit, on augmente le besoin de dispersion pour compenser. On observe un effet similaire concernant la diminution de la mesure de Lindblom avec l'augmentation de la valeur de

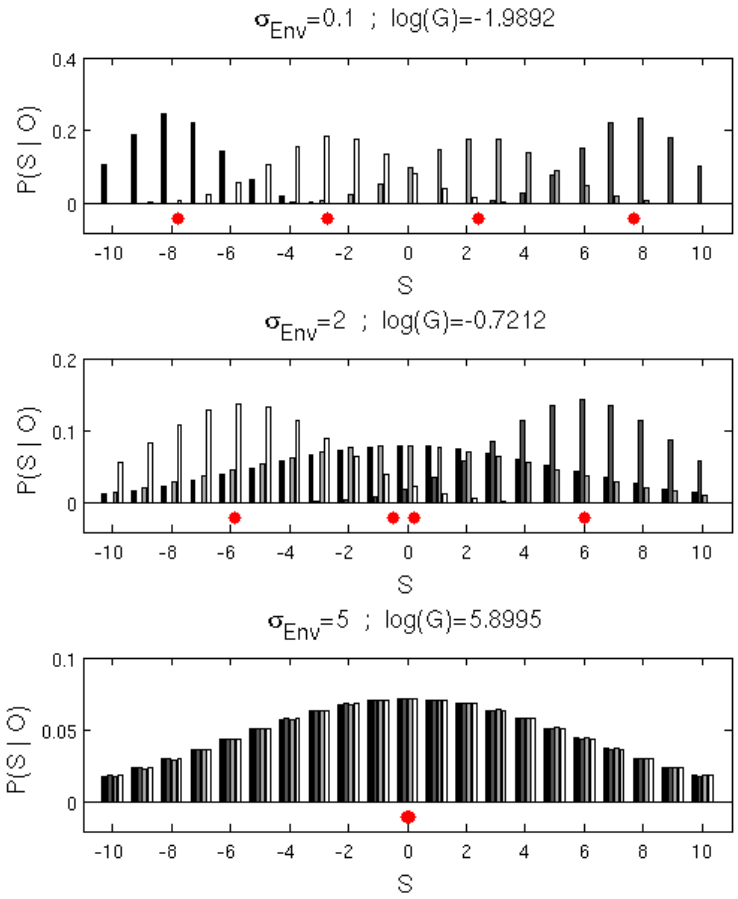


FIGURE 10.13 – Exemple de fins de simulations (comportement auditif, 4 agents, 4 objets) dont les valeurs de la mesure de Lindblom sont proches des moyennes de la Figure 10.12 (pour les valeurs de σ_{Env} indiquées et $\sigma_{Ag} = 0.1$). Les distributions observées sont calculées à partir des stimuli auditifs produits par l'ensemble des agents sur les 10 000 derniers jeux déictiques de la simulation, avant ajout du bruit de l'environnement. Les disques sous les axes des abscisses indiquent les moyennes des stimuli auditifs produits par les agents.

σ_{Ag}, pour σ_{Env} constant.

Cet effet qui reste léger dans ces simulations où l'espace auditif est relativement restreint, sera bien mieux observable dans les simulations d'émergence de voyelles du chapitre suivant, nous le verrons.

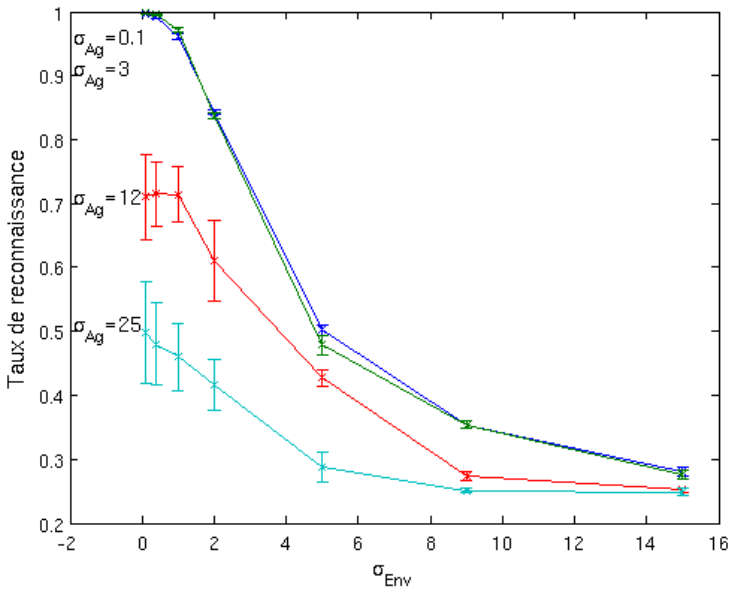


FIGURE 10.14 – Taux de reconnaissance en fin de simulation dans une société de 4 agents en comportement sensori-moteur, et un environnement de 4 objets, en fonction du bruit de l'environnement et de l'incertitude des agents. Moyennes et écarts-types sur 10 simulations indépendantes.

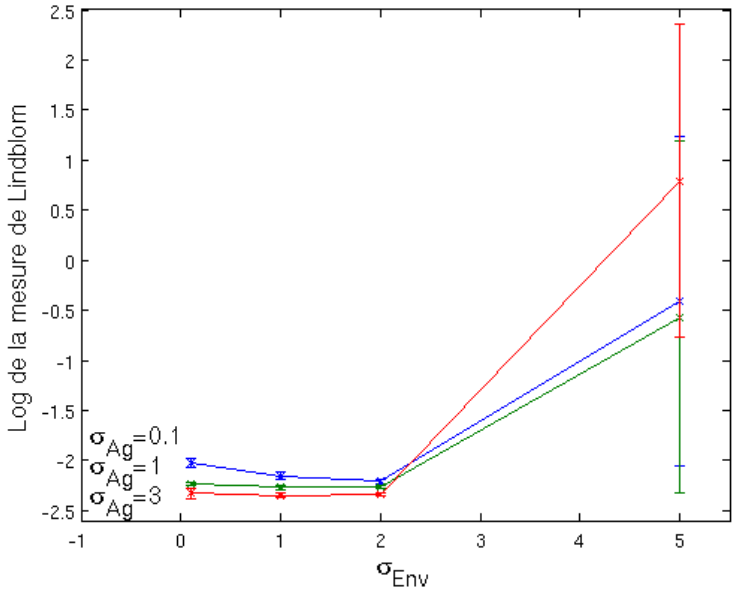


FIGURE 10.15 – Logarithme de la mesure de Lindblom en fin de simulation dans une société de 4 agents en comportement sensori-moteur, et un environnement de 4 objets, en fonction du bruit de l'environnement et de l'incertitude des agents. Moyennes et écart-types sur 10 simulations. Les mesures de dispersion sont réalisées sur les moyennes des prototypes auditifs de l'ensemble des agents avant l'ajout du bruit par l'environnement.

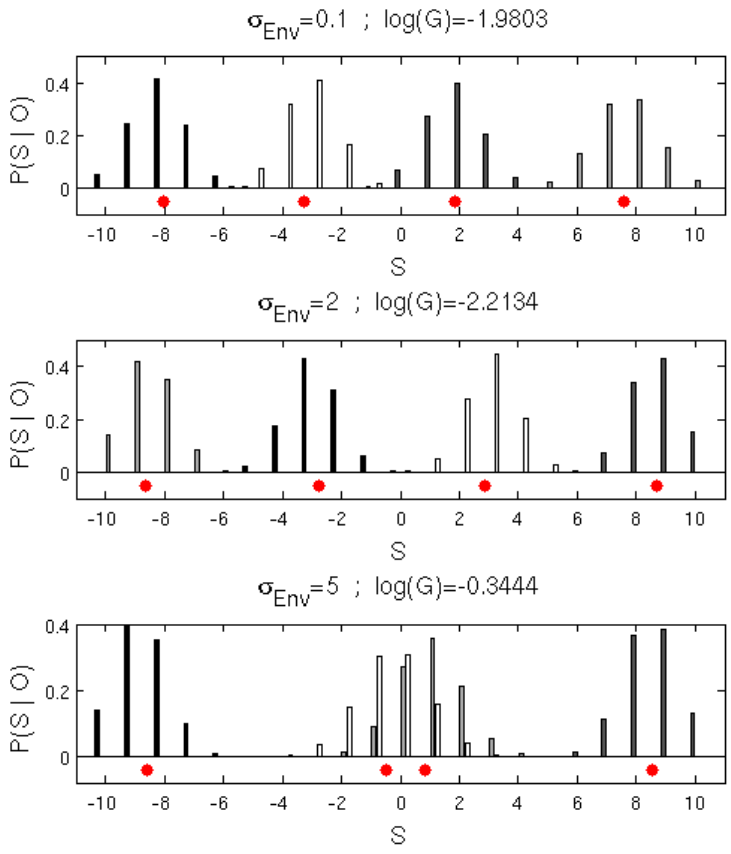


FIGURE 10.16 – Exemple de fins de simulations (comportement sensori-moteur, 4 agents, 4 objets) dont les valeurs de la mesure de Lindblom sont proches des moyennes de la Figure 10.15 (pour les valeurs de σ_{Env} indiquées et $\sigma_{Ag} = 0.1$). Les distributions correspondent aux stimuli auditifs produits par l'ensemble des agents en fin de simulation avant ajout du bruit dans l'environnement. Les disques sous les axes des abscisses indiquent les moyennes de ces distributions.

10.4.4 Effet d'une non-linéarité dans la fonction de transformation articulatori-acoustique TransMS

Nous étudions l'influence d'une non-linéarité dans la fonction de transformation articulatori-auditive TransMS sur les systèmes émergeant de nos simulations. Pour cela, nous nous intéressons d'abord aux variations conjointes de la force de la non-linéarité de la transformation (NL) et du bruit de l'environnement (σ_{Env}), et analysons leurs effets sur le taux de reconnaissance. Puis, pour une non-linéarité forte, nous nous intéressons à la variation de la position du point d'inflexion (D) sur la structuration des systèmes émergeant de nos simulations.

Les simulations mises en œuvre concernent une société de $N_A = 4$ agents évoluant dans un environnement de $N_O = 2$ objets sur une succession de $N_{JD} = 200\ 000$ jeux déictiques. Ce choix de 2 objets est dû à la forme de la fonction TransMS (sigmoïdale, voir Figure 10.1). Nous cherchons en effet à comparer nos résultat avec les prédictions de la théorique quantique de Stevens (1972, 1989) exposée à la Section 4.1.2, qui propose que les non-linéarités structurent les systèmes phonologiques de façon à utiliser préférentiellement des zones stables de la transformation. L'aspect sigmoïdale de TransMS induit 2 zones stables, plus ou moins larges selon la valeur du paramètre D (Figure 10.2), d'où notre choix d'un environnement à 2 objets. Nous reviendrons sur la théorie quantique en conclusion de ce chapitre. L'incertitude des agents est fixée à la valeur $\sigma_{Ag} = 1$ pour toutes les simulations de cette section.

10.4.4.1 Effets conjoints de la force de la non-linéarité NL et du bruit de l'environnement σ_{Env}

Ces résultats proviennent d'une série de 240 simulations par type de comportement (auditif et sensori-moteur) : 10 simulations indépendantes pour chaque combinaison possible de $NL \in \{0.01, 0.3, 0.8, 5\}$ et $\sigma_{Env} \in \{0.1, 1, 2, 4, 5, 8\}$.

Comportement auditif La Figure 10.17 montre l'effet de la force de la non-linéarité NL et du bruit de l'environnement σ_{Env} sur le taux de reconnaissance en fin de simulation dans une société d'agents en comportement auditif.

Pour un bruit d'environnement très faible ($\sigma_{Env} = 0.5$), les performances mitigées du comportement auditif, qui n'atteint jamais 100% de reconnaissance dans le cas d'une fonction de transformation linéaire (voir l'interprétation du comportement auditif à la Section 10.3.2, certes sur des simulations à 4 objets mais dont les arguments sont transposables à des simulations à 2 objets), sont compensées par l'effet bénéfique de l'augmentation de la force de la non-linéarité (NL). Pour un bruit moyen ($\sigma_{Env} = 2$), l'augmentation de NL permet également de compenser l'effet néfaste du bruit, pour finalement atteindre 100% comme dans le cas précédent. Dans le cas d'un bruit fort ($\sigma_{Env} = 5$), alors que le taux de reconnaissance ne dépasse pas le niveau du hasard sans non-linéarité (50% pour 2 objets), l'augmentation de NL permet d'obtenir rapidement un code efficace. Finalement, lorsque

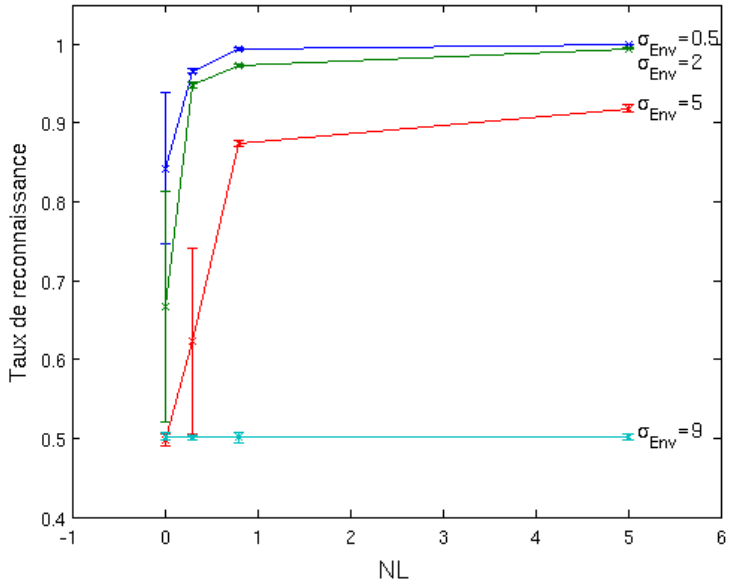


FIGURE 10.17 – Taux de reconnaissance en fin de simulation dans une société de 4 agents en comportement auditif, et un environnement de 2 objets, en fonction de l'importance de la non-linéarité dans TransMS (NL) et du bruit de l'environnement (σ_{Env}). Moyennes et écarts-types sur 10 simulations.

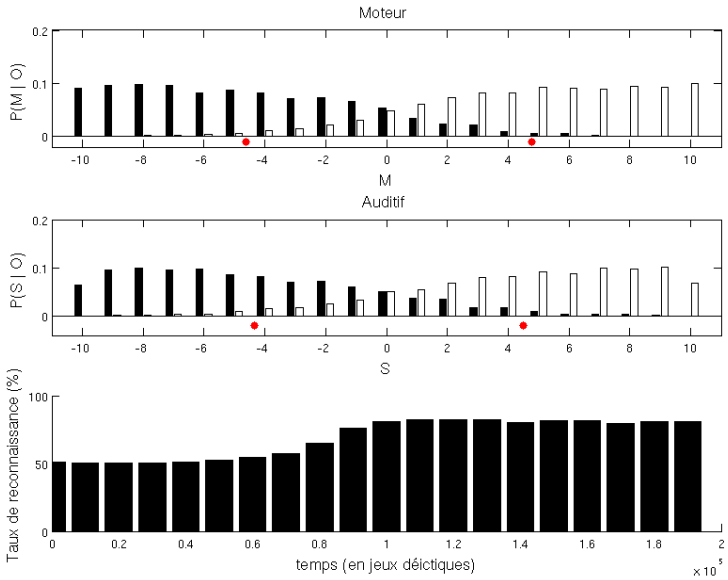
le bruit est trop fort ($\sigma_{Env} = 9$), l'introduction d'une non-linéarité ne peut plus compenser la dégradation des conditions de communication.

La Figure 10.18 montre deux exemples tirés de l'ensemble des simulations dont est issue la Figure 10.17, correspondant aux valeurs extrêmes de NL pour $\sigma_{Env} = 2$. On observe que l'introduction de la non-linéarité améliore d'une part le taux de reconnaissance en fin de simulation (de 80% à 100%), d'autre par le temps de convergence (de 110 000 jeux déictiques à 40 000).

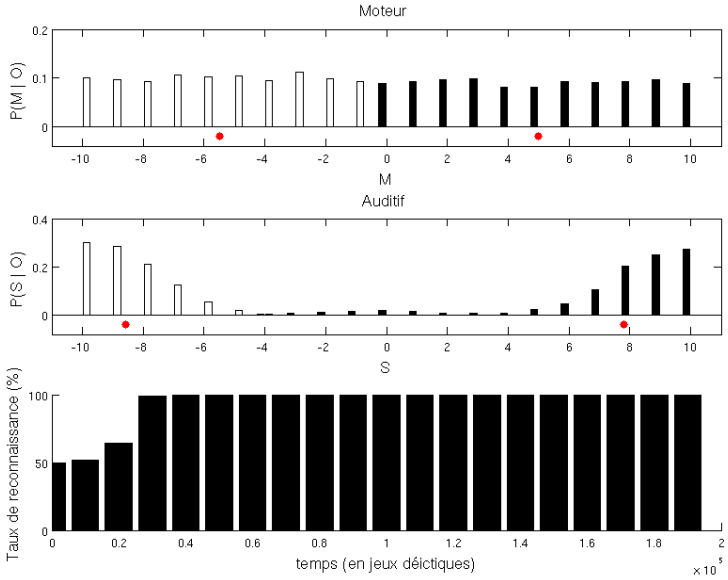
Ainsi, la non-linéarité « déforme » les conséquences auditives des gestes moteurs produits par les agent et les rend plus facilement distinguables. Les simulations bénéficient de cet effet et convergent plus rapidement vers un code plus efficace.

Comportement sensori-moteur La Figure 10.19 montre l'effet de la force de la non-linéarité NL et du bruit de l'environnement σ_{Env} sur le taux de reconnaissance en fin de simulation dans des sociétés d'agents en comportement sensori-moteur.

On observe de façon générale que l'introduction d'une non-linéarité permet d'obtenir un code de parole sensiblement plus efficace, d'autant plus que la valeur de NL est élevée. La Figure 10.20 montre deux exemples tirés de l'ensemble des simulations dont est issue la Figure 10.19, correspondant aux valeurs extrêmes de NL pour $\sigma_{Env} = 5$. On observe que l'introduction de la non-linéarité améliore légèrement le taux de reconnaissance en fin de



(a) Exemple de simulation à 4 agents en comportement auditif et 2 objets pour $NL = 0.01$ et $\sigma_{Env} = 2$.



(b) Exemple de simulation à 4 agents en comportement auditif et 2 objets pour $NL = 5$ et $\sigma_{Env} = 2$.

FIGURE 10.18 – Deux exemples de simulation à 4 agents en comportement auditif et 2 objets pour $\sigma_{Env} = 2$. (a) Cas linéaire : $NL = 0.01$. (b) Cas fortement non-linéaire : $NL = 5$. Les fenêtres titrées « Moteur » et « Auditif » correspondent aux distributions, sous forme d'histogrammes, des gestes moteurs produits et stimuli auditifs perçus par l'ensemble des agents sur les 10 000 derniers jeux déictiques de la simulation.

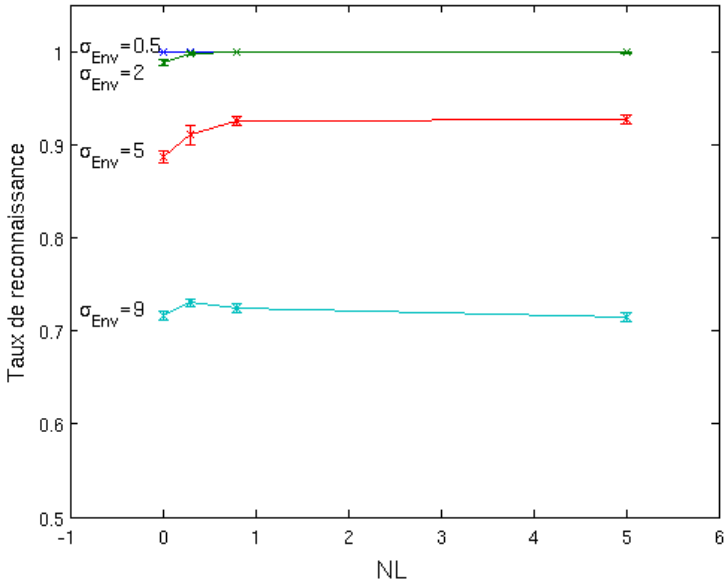


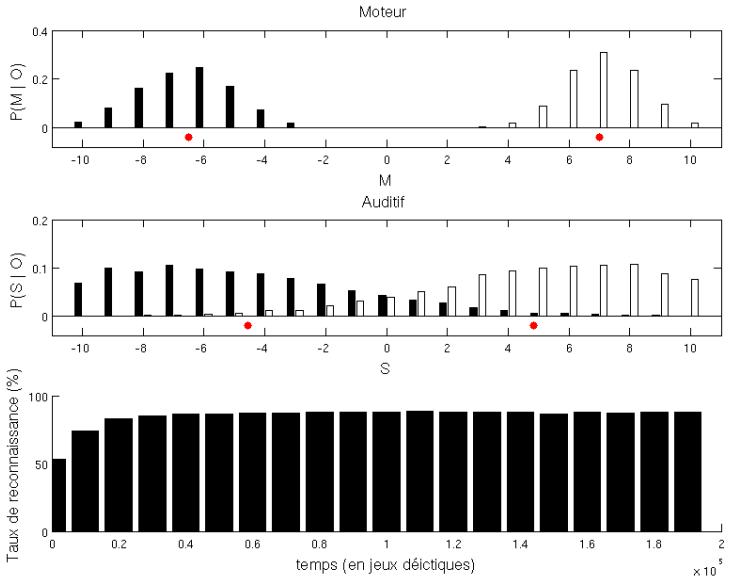
FIGURE 10.19 – Taux de reconnaissance en fin de simulation dans une société de 4 agents en comportement sensori-moteur, et un environnement de 2 objets, en fonction de l'importance de la non-linéarité dans TransMS (NL) et du bruit de l'environnement (σ_{Env}). Moyennes et écarts-types sur 10 simulations.

simulation (de 88% à 93%) et le temps de convergence (de 40 000 jeux déictiques à 20 000).

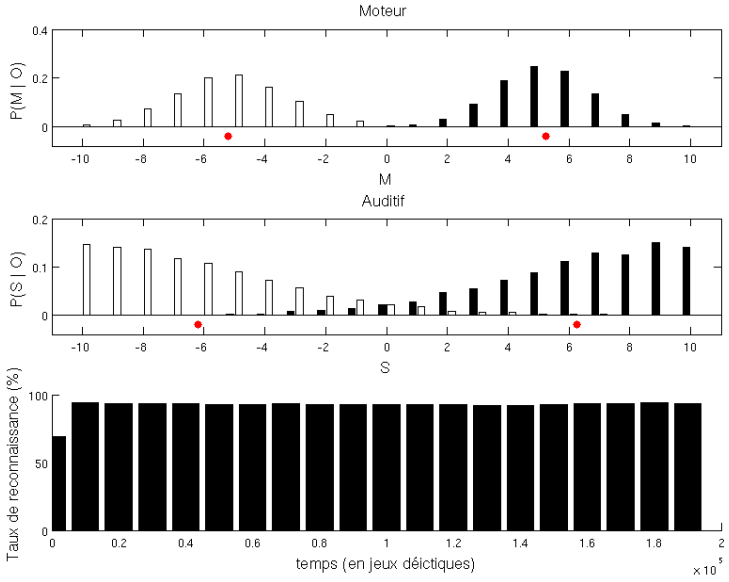
L'introduction d'une non-linéarité a donc plus d'effet en comportement auditif qu'en comportement sensori-moteur. En effet, alors que le premier disperse les éléments de façon uniforme dans leur espace, le deuxième cherche à les séparer suffisamment pour une bonne communication. Ainsi, le comportement sensori-moteur compense déjà l'effet du bruit dans le cas d'une séparation linéaire en dispersant mieux ses éléments (Section 10.4.3.2) et tire donc un avantage limité de l'introduction d'une non-linéarité.

10.4.4.2 Effet de la position du point d'inflexion D

Nous étudions maintenant l'effet de la position du point d'inflexion (paramètre D) d'une fonction TransMS fortement non-linéaire ($NL = 5$) sur la structuration du code de parole émergeant des simulations (Figure 10.2). Si les non-linéarités structurent les systèmes phonologiques comme proposé par la théorique quantique, la prédiction est une forte corrélation de la position du point d'inflexion avec des gestes moteurs choisis par les agents, afin de se placer de part et d'autre de celui-ci. Ces résultats proviennent d'une série de 210 simulations par type de comportement (auditif et sensori-moteur) : 10 simulations indépendantes pour chaque valeur de $D \in [-10, 10]$.



(a) Exemple de simulation à 4 agents en comportement sensori-moteur et 2 objets pour $NL = 0.01$ et $\sigma_{Env} = 5$.



(b) Exemple de simulation à 4 agents en comportement sensori-moteur et 2 objets pour $NL = 5$ et $\sigma_{Env} = 2$.

FIGURE 10.20 – Deux exemples de simulation à 4 agents en comportement sensori-moteur et 2 objets pour $\sigma_{Env} = 5$. (a) Cas linéaire : $NL = 0.01$. (b) Cas fortement non-linéaire : $NL = 5$. Les fenêtres titrées « Moteur » et « Auditif » correspondent aux distributions, sous forme d'histogrammes, des gestes moteurs produits et stimuli auditifs perçus par l'ensemble des agents sur les 10 000 derniers jeux déictiques de la simulation.

Comportement auditif La Figure 10.21 expose les résultats issus de nos simulations dans une société d'agents en comportement auditif. Pour des valeurs de $D \in [-7, 7]$,

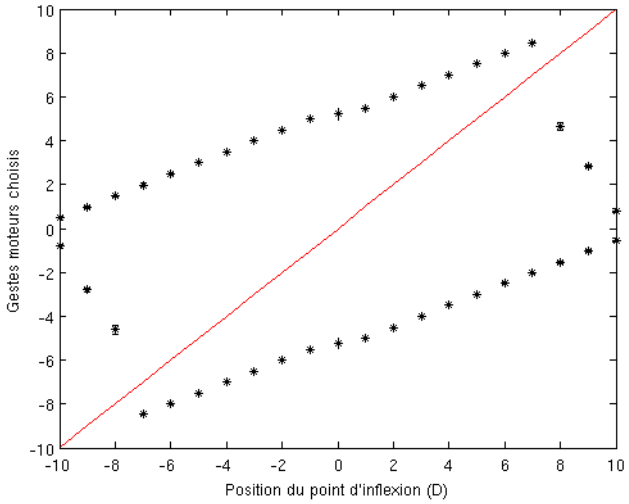
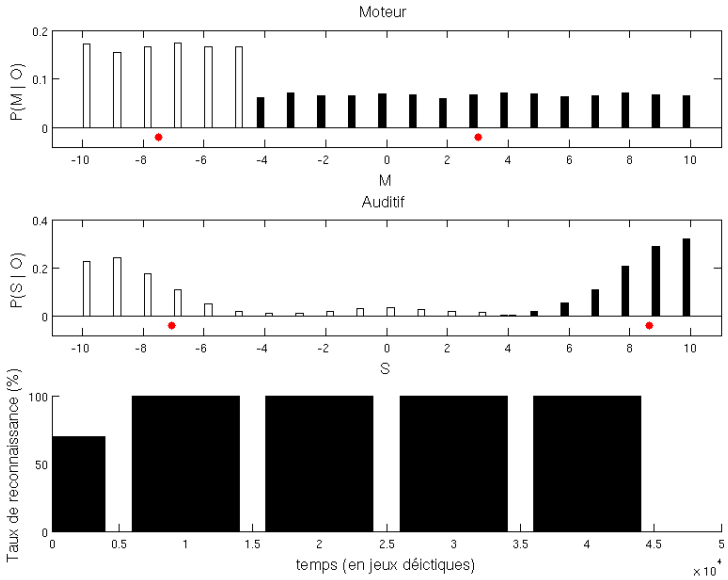


FIGURE 10.21 – Gestes moteurs choisis en fin de simulation dans une société de 4 agents en comportement auditif, et un environnement de 2 objets, en fonction de la position du point d'inflexion de TransMS. Pour chaque valeur de $D \in [-10, 10]$, les deux points indiquent la moyenne et l'écart-type sur 10 simulations indépendantes des moyennes des gestes moteurs produits par l'ensemble des agents en fin de simulations (les écarts-types sont très faibles et donc peu visibles). La ligne rouge est la fonction identité indiquant la position du point d'inflexion (abscisse) dans l'espace des gestes moteurs (ordonnée).

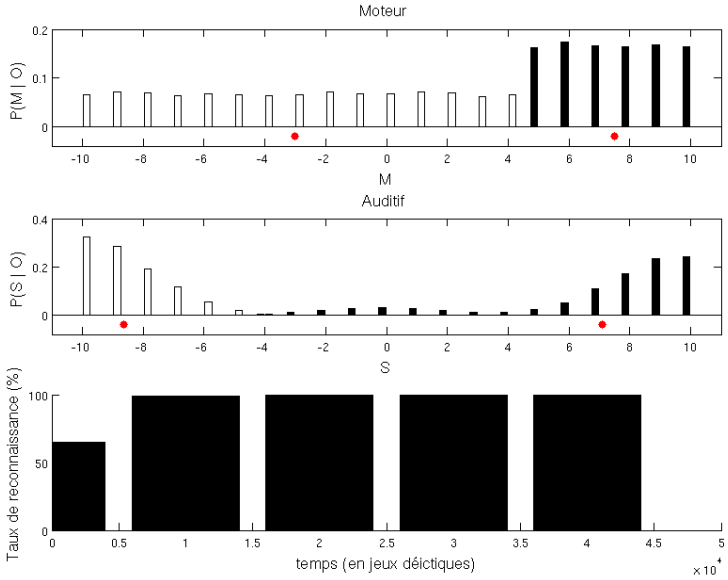
on observe que les gestes moteurs choisis par les agents se placent de part et d'autre du point d'inflexion de la fonction TransMS. Plus précisément, les gestes se trouvent à equidistance du point d'inflexion et de l'une des bornes de l'espace, avec une très faible variabilité. Pour des valeurs de D inférieures à -7 ou supérieures à 7, les gestes sont dans la même région car l'espace laissé pour l'autre est trop restreint. La Figure 10.22 montre deux exemples de simulation tirés de l'ensemble des 210 simulations dont est issue la Figure 10.21, correspondant aux valeurs $D = -5$ et $D = 5$. On observe que la position du point d'inflexion sert de frontière entre les gestes moteurs choisis par les agents, ceux-ci étant uniformément répartis dans leur demi-espace. Autrement dit, les gestes moteurs sont choisis en fonction de leurs conséquences auditives, identiques quelque soit la valeur de D.

Comportement sensori-moteur La Figure 10.21 expose les résultats issus de nos simulations dans une société d'agents en comportement sensori-moteur.

Les observations sont similaires à celles du comportement auditif. On remarque toutefois que les gestes moteurs peuvent se placer de part et d'autre du point d'inflexion pour des valeurs de D plus extrêmes (entre -8 et 8, jusqu'à -9 et 9 pour certaines simulations)



(a) Exemple de simulation à 4 agents en comportement auditif et 2 objets pour $NL = 5$ et $D = -5$.



(b) Exemple de simulation à 4 agents en comportement auditif et 2 objets pour $NL = 5$ et $D = 5$.

FIGURE 10.22 – Deux exemples de simulation à 4 agents en comportement auditif et 2 objets pour $NL = 5$. (a) Non-linéarité à gauche : $D = -5$. b) Non-linéarité à droite : $D = 5$. Les fenêtres titrées « Moteur » et « Auditif » correspondent respectivement aux distributions des gestes moteurs produits et stimuli auditifs perçus par l'ensemble des agents sur les 10 000 derniers jeux déictiques de la simulation.

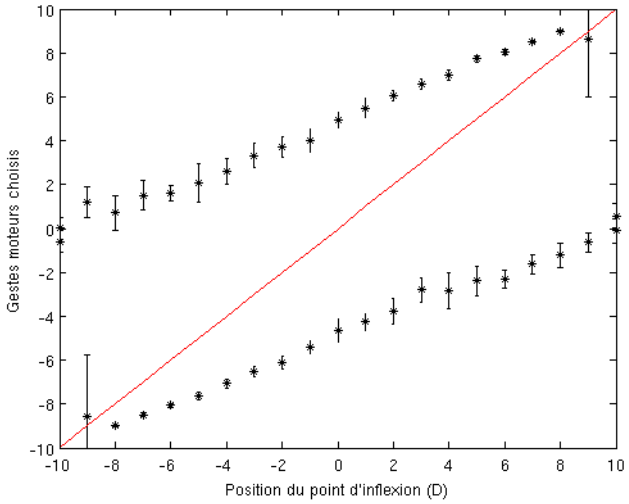
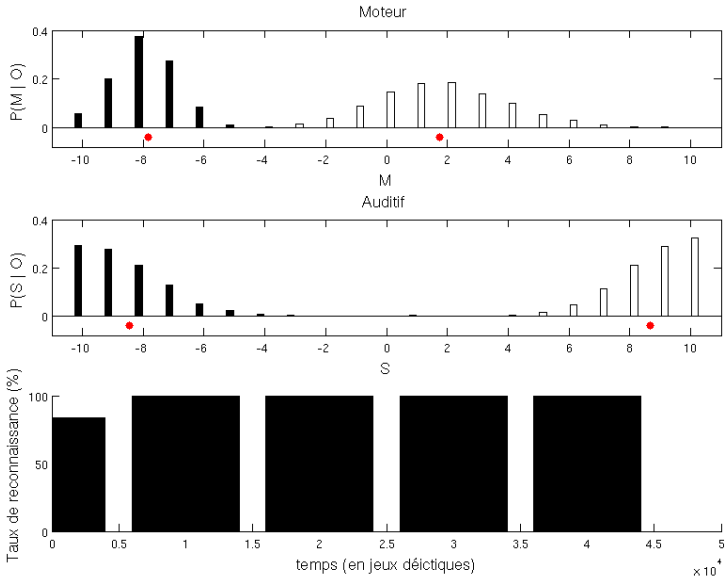
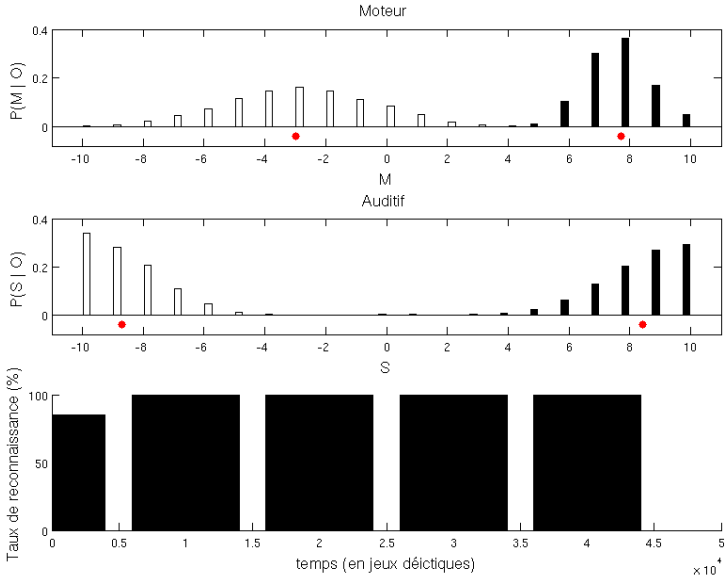


FIGURE 10.23 – Gestes moteurs choisis en fin de simulation dans une société de 4 agents en comportement sensori-moteur, et un environnement de 2 objets, en fonction de la position du point d'inflexion de TransMS. Moyennes sur 10 simulations. Les éléments du graphique sont présentés Figure 10.21.

qu'en comportement auditif. On remarque également une plus grande variabilité des gestes choisis dans les régions où la distance entre le point d'inflexion et la borne de l'espace est large. La Figure 10.24 montre deux exemples de simulations tirés de l'ensemble des 210 simulations dont est issue la Figure 10.23, correspondant aux valeurs $D = -5$ et $D = 5$. Alors que la position du point d'inflexion sert toujours de frontière entre les gestes moteurs choisis par les agents, ceux-ci ne sont plus uniformément répartis dans leur demi-espace comme en comportement auditif. Le comportement sensori-moteur cherchant simplement à suffisamment séparer les éléments du système, les gestes moteurs choisis sont plus variables d'une simulation à une autre.



(a) Exemple de simulation à 4 agents en sensori-moteur et 2 objets pour $NL = 5$ et $D = -5$.



(b) Exemple de simulation à 4 agents en comportement sensori-moteur et 2 objets pour $NL = 5$ et $D = 5$.

FIGURE 10.24 – Deux exemples de simulation à 4 agents en comportement sensori-moteur et 2 objets pour $NL = 5$ et $\sigma_{Env} = 2$. (a) Non-linéarité à gauche : $D = -5$. (b) Non-linéarité à droite : $D = 5$. Les fenêtres titrées « Moteur » et « Auditif » correspondent respectivement aux distributions des gestes moteurs produits et stimuli auditifs perçus par l'ensemble des agents sur les 10 000 derniers jeux déictiques de la simulation.

10.4.5 Conclusion sur l'évaluation des comportements

Cette analyse d'une grande série de simulations montre que les codes de parole émergeant de sociétés d'agents en comportements auditif ou sensori-moteur vérifient certaines prédictions des théories de la forme des systèmes phonologiques présentées à la Section 4.1 du Chapitre 4.

La première série de simulations (Section 10.4.3) montre le lien entre la dispersion des éléments des systèmes émergeant de nos simulations et le taux de reconnaissance dans la société d'agent. Pour cela nous avons fait varier les conditions de communication (bruit de l'environnement σ_{Env} et incertitude des agents σ_{Ag}) et étudié leurs effets sur la dispersion et la reconnaissance. Le lien entre ces deux mesures est très clair en comportement auditif, comme le montrent les Figures 10.11 et 10.13 : plus la dispersion est faible (mesure de Lindblom élevée), plus le taux de reconnaissance est bas. Ce lien est par contre plus subtil en comportement sensori-moteur, pour lequel les meilleurs scores de reconnaissance ne sont pas forcément associés aux plus fortes dispersions. Pour des contraintes de communication acceptables, c'est-à-dire des valeurs de σ_{Env} et σ_{Ag} qui ne font pas trop chuter les taux de reconnaissance, les deux mesures diminuent lorsque les conditions se dégradent. En effet, la dispersion nécessaire à une bonne communication entre les agents dépend des contraintes de communication. Le comportement sensori-moteur, qui a la propriété de ne disperser les éléments du système que d'une façon suffisante, laisse alors émerger des systèmes plus « compacts » lorsque les contraintes le permettent. On retrouve ici un des principes de la théorie de la dispersion adaptative de Lindblom (1990) qui conçoit la phonologie comme un compromis entre les contraintes d'articulation du locuteur et celles de perception de l'auditeur, en fonction des contraintes de communication.

La seconde série de simulations (Section 10.4.4) montre comment l'introduction d'une non-linéarité dans la fonction de transformation articulatori-auditive peut améliorer et structurer le code de parole émergeant de la société d'agent. Nos résultats concernant les variations de la force de la non-linéarité (paramètre NL, Figures 10.17 et 10.19) montre que celle-ci permet de compenser la baisse de reconnaissance due au bruit de l'environnement, en particulier pour le comportement auditif (le comportement sensori-moteur adaptant déjà par lui-même sa dispersion pour atteindre un bon taux de reconnaissance). Les prédictions de la théorie quantique de Stevens (1972, 1989), selon laquelle les systèmes phonologiques doivent tirer partie de la séparation de l'espace sensori-moteur en zones stables et instables induite par les non-linéarités, se vérifient donc dans nos simulations. D'autre part, les résultats concernant les variations de la position du point d'inflexion (paramètre D, Figures 10.21 et 10.23) montrent comment ces non-linéarités peuvent structurer le code parole, quel que soit le comportement considéré. Il s'agit là encore d'une des prédictions de la théorie quantique selon laquelle les systèmes phonologiques privilégient les zones stables induites par les non-linéarités, les zones instables constituant alors des barrières « naturelles » entre les phonèmes qui se placent de part et d'autre.

10.5 Conclusion

Ce chapitre a constitué la validation préliminaire de toute la démarche de modélisation effectuée dans les deux parties précédentes. Nous avons implémenté pour cela un modèle de transformation articulatori-auditive simple à une dimension afin d'évaluer et de comparer les résultats issus de différentes société d'agents : tous moteurs, tous auditifs ou tous sensori-moteurs.

Nous avons d'abord exposé les propriétés générales de notre paradigme d'interaction par jeux déictiques dans une société d'agents dans le cas d'une transformation articulatori-auditive linéaire et déterministe. Nous avons vu que le comportement moteur, qui ne prend pas en compte les besoins de l'auditeur, ne permettait pas l'émergence d'un code de parole efficace. Le comportement auditif, qui cherche à produire des gestes moteurs dont les conséquences auditives permettent une bonne classification, permet l'émergence d'un code de parole efficace en dispersant les éléments du système uniformément dans leur espace. Quant au comportement sensori-moteur, produit des deux comportements précédents, il permet l'émergence d'un code de parole plus rapidement et plus efficacement que son homologue auditif en limitant les productions des agents dans les zones facilitant la catégorisation.

Puis nous avons évalué chaque comportement dispersif (auditif et sensori-moteur) sur des séries de simulations. Nous avons vu qu'ils étaient en accord avec des théories de la forme des systèmes phonologiques : la théorie de la dispersion et la théorie quantique. Concernant la théorie de la dispersion, nous avons tisser des premiers liens formels entre comportement et optimisation (Équation 10.12).

Nos prédictions des Parties I et II concernant la supériorité du modèle sensori-moteur, qui dispose du modèle le plus complet de toute la situation de communication, sont maintenant validées par la simulation. Dans les trois chapitres qui suivent, dans lesquels nous traitons successivement de l'émergence des systèmes de voyelles, de consonnes plosives et de syllabes dans des simulations utilisant un modèle réaliste de la transformation articulatori-acoustique réalisée par le conduit vocal humain, nous nous concentrerons donc uniquement sur ce comportement sensori-moteur.

Chapitre 11

Emergence des systèmes de voyelles

Le chapitre précédent nous a permis d'étudier en détail les propriétés générales de chacun des comportements de production dans le cadre de simulations d'émergence des systèmes phonologiques, en utilisant un modèle simplifié de transformation articulatori-auditive reposant sur des variables motrice M et auditive S à une dimension. Nous avons retenu le comportement sensori-moteur comme le plus efficace dans ce cadre. C'est donc exclusivement sur ce comportement que nous nous concentrons dans les trois chapitres à venir, dans lesquels nous effectuons le passage à une échelle réaliste en utilisant un modèle de la transformation articulatori-auditive réalisée par le conduit vocal et l'oreille humains. Nous serons alors en mesure d'étudier les capacités du modèle de société d'agents en comportement sensori-moteur à faire émerger des systèmes phonologiques composés uniquement de voyelles (le présent chapitre), uniquement de consonnes (Chapitre 12) et finalement de syllabes (Chapitre 13).

Ce chapitre commence par exposer rapidement les données sur les tendances des systèmes de voyelles des langues du monde, ainsi que certains travaux visant à les expliquer. Puis nous présentons le modèle de conduit vocal utilisé, à partir duquel nous définissons le modèle de la transformation articulatori-auditive de l'environnement. Nous définissons ensuite le modèle d'agent dans le cadre d'émergence de systèmes de voyelles. Enfin nous présentons nos résultats sur un grand nombre de simulations et les comparons aux régularités des systèmes de voyelles des langues du monde.

11.1 Données et prédictions des systèmes de voyelles des langues du monde

Comme nous l'avions déjà remarqué au Chapitre 4, les systèmes de voyelles présents dans les langues du monde obéissent à certaines régularités. La base de données UPSID (UCLA Phonological Segment Inventory Database, Maddieson et Precoda, 1989) recense les systèmes phonologiques des langues du monde, parmi lesquels sont identifiées 177 voyelles. Loin d'être uniformément répartis selon une simple règle combinatoire, ils se trouvent que la distribution des systèmes identifiés laisse apparaître certaines préférences (voir Schwartz

et collab. (1997a) pour une analyse détaillée). La Figure 11.1 montre ces tendances.

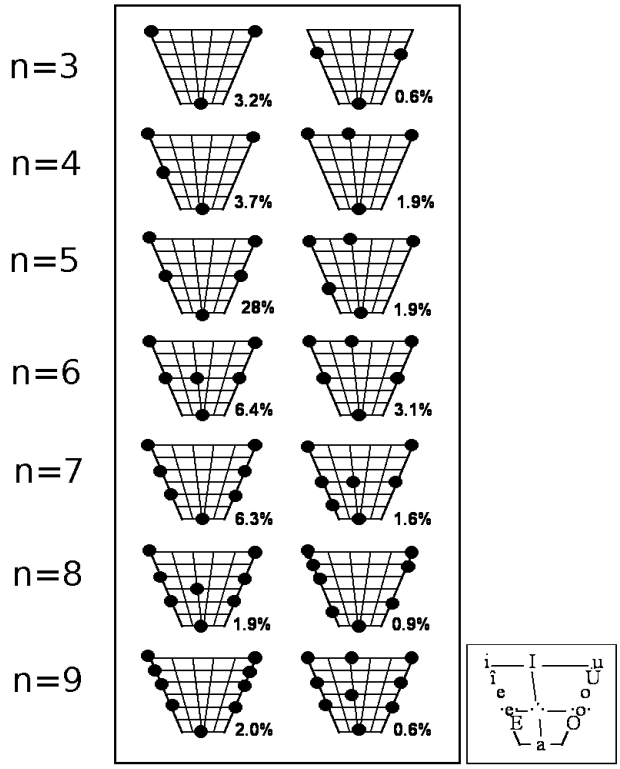


FIGURE 11.1 – Pourcentage des systèmes de voyelles majoritaires dans les langues du monde de la base UPSID, d'après (Vallée, 1994). Chaque ligne correspond aux deux systèmes majoritaires comportant un certain nombre de voyelles (de $n = 3$ à $n = 9$). Chaque pourcentage indique la proportion d'un système particulier dans l'ensemble des systèmes vocaliques des langues recensées (quel que soit leur nombre d'éléments). Les deux dimensions de l'espace visualisé, compartimenté en 6 sections horizontales (F1) et 5 sections verticales (F2), sont interprétable en termes articulatoires ou acoustiques. Dans le premier cas la dimension horizontale représente la position horizontale de la langue (de l'avant (gauche) à l'arrière (droite)) et la dimension verticale le degré d'ouverture du conduit vocal (de quasi-fermé (haut) à très ouvert (bas). Dans le deuxième cas il s'agit du plan des deux premiers formants du signal produit, la valeur de F1 croissant vers le bas sur l'axe vertical, celle de F2 vers la gauche sur l'axe horizontal (les deux axes sont inversés pour permettre cette double interprétation, les deux espaces correspondants étant topologiquement proches dans le cas des voyelles).

L'observation générale est que les systèmes vocaliques ont tendance à disperser leurs éléments dans le triangle vocalique. C'est la théorie de la dispersion (Liljencrants et Lindblom, 1972; Lindblom, 1990) (présentée à la Section 4.1.1 du Chapitre 4), selon laquelle les systèmes vocaliques cherchent à disperser au maximum leurs éléments dans l'espace auditif, maximisant ainsi leurs distinguabilité (voir Schwartz et collab. (2007) ou la Section 3.3.2

de ce document pour la justification d'une dispersion auditive plutôt qu'articulatoire).

Plus précisément, Schwartz et collab. (1997b) montrent que les prédictions de systèmes de voyelles issues de la théorie de la dispersion sont fortement dépendantes du poids accordé à chaque dimension formantique. Ils définissent la notion de second formant effectif, noté F'2, comme une combinaison de F2, F3 et F4 dans laquelle F3 a deux fois moins de poids que F2. Ils montrent alors que les prédictions les mieux en accord avec les systèmes des langues du monde attribuent à F1 un poids 3 fois plus grand qu'à F'2 dans le calcul de distance et de dispersion perceptive. Ainsi, il semble que F1 ait 3 fois plus de poids que F2, qui a lui-même 2 fois plus de poids que F3, pour que la théorie de la dispersion fournisse des prédictions correctes des systèmes vocaliques.

Dans la suite de ce chapitre, nous souhaitons montrer la capacité de notre modèle à faire émerger des systèmes de 3 et 5 voyelles globalement cohérents avec ces données. Les données montrent que les deux systèmes largement majoritaires sont les systèmes /a, i, u/ (3 voyelles) et /a, i, e, o, u/ (5 voyelles). Les systèmes majoritaires de 3 et 5 voyelles sont d'intérêt particulier : le premier est le système « minimal » (tous les systèmes ont /a,i,u/, quelque soit le nombre d'éléments) et le second est le plus fréquent de l'ensemble des systèmes vocaliques recensés des langues du monde. Topologiquement, il s'agit dans les deux cas de systèmes « de la forme d'un V » (une voyelle basse, les autres régulièrement réparties sur les bords gauche et droit), cette tendance étant d'ailleurs vérifiée pour tous les systèmes à nombre impair d'éléments.

11.2 Modèle de transformation articulatori-auditive : VLAM

Notre laboratoire dispose d'un modèle réaliste de la transformation articulatori-acoustique réalisée par le conduit vocal humain : VLAM (Variable Linear Articulatory Model, (Boë, 1999)), dérivé du modèle articulatoire de Maeda (1989). Ce dernier est conçu à partir d'une analyse statistique de 519 contours sagittaux provenant de films radiographiques et labiographiques de phrases prononcées par un locuteur français. Ces contours sont divisés en 28 segments allant de la glotte (1) jusqu'aux lèvres (28), à partir desquels les aires correspondantes du conduit vocal sont calculées. Une analyse en composantes principales (ACP) montre que 7 paramètres expliquent 88 % de la variance des données. Une équation linéaire combinant ces 7 paramètres permet de regénérer les contours sagittaux du conduit vocal. De façon intéressante, chacun des paramètres issus de l'ACP est facilement interprétable en terme de commande motrice comme le montre la Figure 11.2. Chacun de ces paramètres prend généralement des valeurs entre -3 et +3 écarts-types des données enregistrées, la position neutre étant fixée à 0.

Jaw est la hauteur de mâchoire, de -3 pour une valeur ouverte à +3 pour une valeur fermée ;

Body est la position du corps de la langue, de -3 pour une valeur avant (vers les lèvres, à gauche de la Figure 11.2) à +3 pour une valeur arrière (vers la gorge, à droite de la

figure).

Drsm (pour dorsum) est la position du dos de la langue, de -3 pour une valeur basse (vers le bas de la figure) à +3 pour une valeur haute (vers le haut de la figure).

Apex est la position de la pointe de la langue, de -3 pour une valeur basse à +3 pour une valeur haute (amenant la pointe de la langue vers le palais, comme pour un [d]).

LipH est la hauteur des lèvres, de -3 pour une valeur fermée à +3 pour une valeur ouverte (dans le sens inverse de la mâchoire, LipH s'interprétant comme le degré d'ouverture des lèvres).

LipP est la protrusion des lèvres, de -3 pour une petite protrusion (comme dans un [i]) à +3 pour une grande protrusion (comme dans un [u]).

Lx est la hauteur du larynx, de -3 pour une valeur basse à +3 pour une valeur haute.

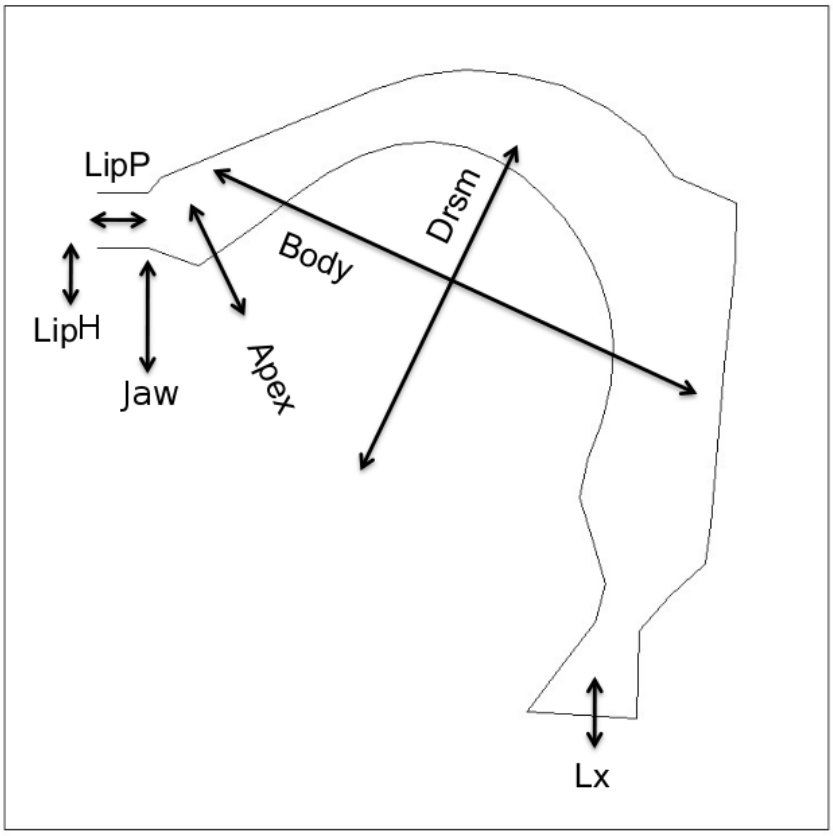


FIGURE 11.2 – Les 7 paramètres du modèle VLAM.

À partir des valeurs de ces paramètres, VLAM génère le contour sagittal correspondant (Figure 11.3a) et calcule la fonction d'aire associant à chacune des 28 sections du conduit vocal l'aire correspondante (Figure 11.3b, fenêtre en haut à droite). Celle-ci est enfin utilisée pour calculer la fonction de transfert (associant à chaque fréquence du signal

d'excitation son gain en intensité en sortie du conduit vocal, Figure 11.3b, fenêtre en bas à droite). On considère que cette fonction de transfert est bien caractérisée par ses résonances, ou formants, qui correspondent à ses maxima (notés F1, F2 et, F3 pour les quatre formants de plus basse fréquence, visualisables dans les plans F1-F2 et F2-F3 à gauche de la Figure 11.3b). On peut supposer que les formants sont estimés au niveau du système auditif (Serkhane et collab., 2005).

À partir de la fonction d'aire, VLAM calcule également des valeurs de constriction, c'est-à-dire des zones où le conduit vocal est le plus étroit, et qui déterminent pour l'essentiel les valeurs des résonances acoustiques :

Xc est la position de la constriction, c'est-à-dire le numéro de la section d'aire minimale ;

Ac est la taille de la constriction, c'est-à-dire l'aire de la section d'aire minimale ;

Al est l'aire aux lèvres, c'est-à-dire l'aire de la dernière section (dépendant à la fois des paramètres LipH et Jaw).

Ces valeurs de constriction n'entrent pas en compte dans le modèle mais nous seront utiles pour diverses sélections ou analyses de données.

11.2.1 Génération du dictionnaire

VLAM nous fournit donc un modèle réaliste de la fonction articulatori-acoustique réalisée par le conduit vocal (des paramètres articulatoires à la fonction de transfert) jusqu'à l'extraction des formants par le système auditif. Dans le but de réduire la dimensionnalité de nos variables motrices, nous ne retiendrons que les paramètres suivants nous permettant d'obtenir une bonne représentativité de l'ensemble des voyelles (et des consonnes plosives, nous le verrons dans le chapitre suivant) : la mâchoire (Jaw), le corps et le dos de la langue (Body et Drsm) et la hauteur des lèvres (LipH). Nous fixons les autres paramètres articulatoires à la valeur neutre (0).

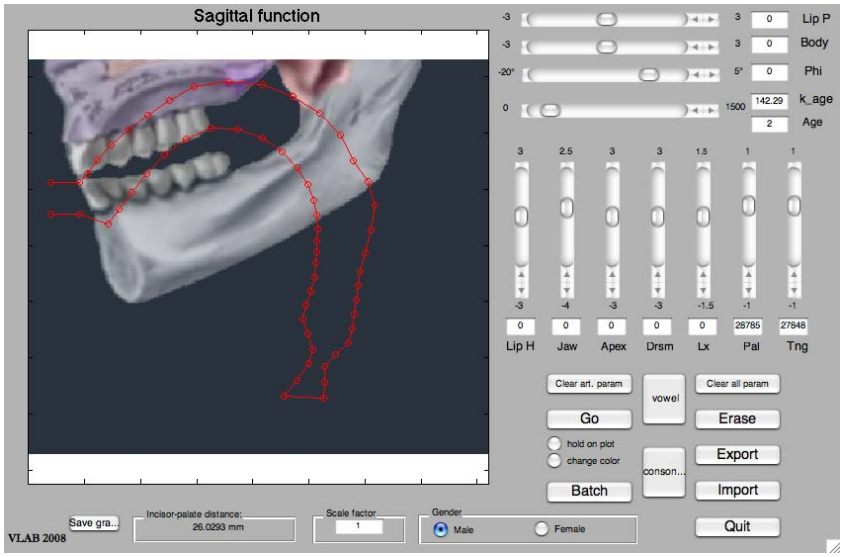
Pour capturer la fonction de transformation articulatori-acoustique, nous réalisons un dictionnaire de 100 000 enregistrements associant les données suivantes :

- les valeurs articulatoires : Jaw, Body, Drsm et LipH ;
- les valeurs constrictives : Xc, Ac et Al ;
- les valeurs formantiques : F1, F2 et F3.

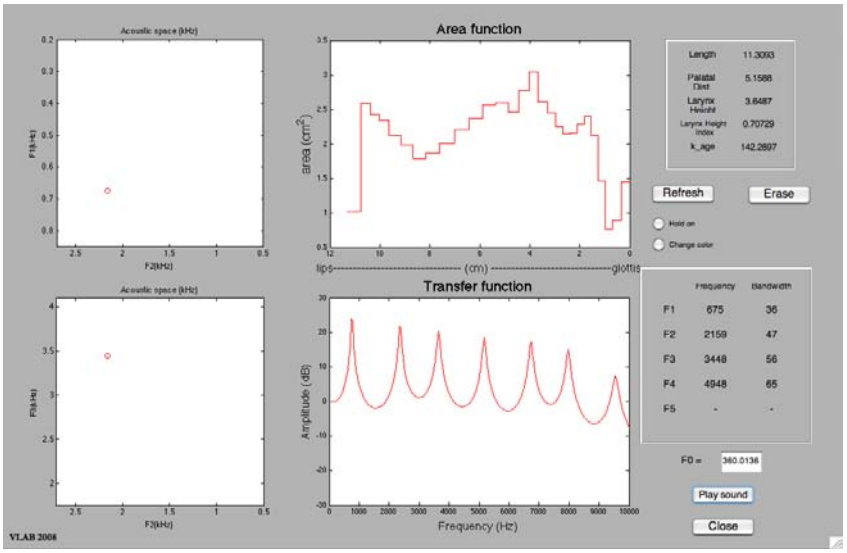
En pratique, on tire de façon aléatoirement uniforme des valeurs dans l'espace articulatoire. Pour chacune, VLAM calcule les valeurs constrictives et formantiques correspondantes. Les valeurs constrictives permettent de sélectionner les voyelles en ne retenant que les configurations suffisamment ouvertes. Pour cela nous fixons l'aire minimale aux lèvres (Al) et de la constriction (Ac) à 0.15 cm^2 et enchaînons les tirages jusqu'à obtenir 100 000 enregistrements. Les valeurs formantiques sont converties en Barks, une unité perceptive pour les formants (Schroeder et collab., 1979) telle que :

$$F_{Barks} = 7 \sinh^{-1}(F_{Hz}/650).$$

L'espace auditif généré par ce dictionnaire est représenté Figure 11.4.



(a) Partie articulatoire.



(b) Partie acoustico-auditive.

FIGURE 11.3 – L'interface de VLAM.

11.2.2 Espaces et transformation articulatori-auditive

Les paramètres articulatoires que nous retenons sont : la mâchoire, le corps et le dos de la langue ainsi que la hauteur des lèvres. Nous notons les variables probabilistes correspondantes J (pour Jaw), TB (pour Tongue Body), TD (pour Tongue Dorsum) et LH (pour Lip Height). Chaque variable est discrétisée sur un nombre de valeurs adapté pour pouvoir générer correctement tout l'espace, en limitant le plus possible l'explosion combinatoire.

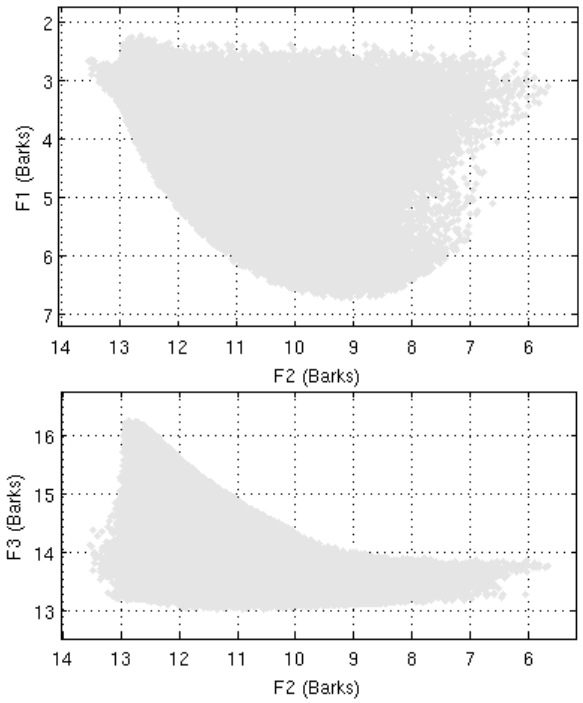


FIGURE 11.4 – Espace auditif issu du dictionnaire de voyelles généré avec VLAM dans les plans F1-F2 et F2-F3.

Les variables motrices correspondent alors à la conjonction :

$$M = J \wedge TB \wedge TD \wedge LH, \tag{11.1}$$

où

$$\begin{aligned} \mathcal{D}_J &= \{0, \ldots, 3\} \\ k_J &= 4 \\ \mathcal{D}_{TB} &= \{0, \ldots, 8\} \\ k_{TB} &= 9 \\ \mathcal{D}_{TD} &= \{0, \ldots, 8\} \\ k_{TD} &= 9 \\ \mathcal{D}_{LH} &= \{0, \ldots, 3\} \\ k_{LH} &= 4. \end{aligned}$$

Nous discrétisons ainsi chaque variable articulatoire $X \in \{J, TB, TD, LH\}$ en k_X valeurs telles que :

$$[X = x]$$
$$\equiv$$
« La variable X a une valeur comprise entre $-3 + \frac{6x}{k_X}$ et $-3 + \frac{6(x+1)}{k_X}$ dans VLAM ». (11.2)

Par exemple, $J = 1$ (valeur d'une variable discrète du modèle) correspond aux valeurs du paramètre VLAM Jaw comprises entre -1 et 1 (c'est-à-dire autour de la valeur neutre). Remarquons que la borne supérieur du domaine de J correspond à des valeurs du paramètre VLAM entre 3 et 5, ce qui est possible car ceux-ci sont exprimés en écarts-types et nous sera nécessaire pour rendre compte de certaines configurations fermées du conduit vocal au chapitre suivant.

Les variables auditives correspondent aux trois premiers formants calculés par VLAM, $F1$, $F2$, $F3$, exprimés en Barks :

$$S = F1 \wedge F2 \wedge F3, \tag{11.3}$$

où

$$\begin{aligned} \mathcal{D}_{F1} &= \{2.5, \ldots, 6.5\} \\ k_{F1} &= 5 \\ \mathcal{D}_{F2} &= \{5.5, \ldots, 13.5\} \\ k_{F2} &= 9 \\ \mathcal{D}_{F3} &= \{13.5, \ldots, 16.5\} \\ k_{F3} &= 4. \end{aligned}$$

Nous discrétisons ainsi chaque variable formantique $X \in \{F1, F2, F3\}$ en k_X valeurs telles que :

$$[X = x]$$
$$\equiv$$
« La variable X a une valeur comprise entre $x - 0.5$ et $x + 0.5$ Barks dans VLAM »,

où x est de la forme $2i+0.5$ (i entier). Par exemple, $F1 = 4.5$ (valeur d'une variable discrète du modèle) correspond aux valeurs de premier formant calculées par VLAM comprises entre 4 et 5 Barks.

Les 100 000 enregistrements du dictionnaire défini plus haut forment un ensemble d'apprentissage δ_{Voy}, dans lequel les valeurs articulatoires et formantiques sont discrétisées dans les domaines des variables correspondantes.

La distribution représentant la transformation articulatori-auditive réalisée par l'environnement est alors définie par :

$$\forall m \in M \ : \tag{11.4}$$

$$P(S \mid [M = m]\ \delta_{Voy}\ \pi_{Com}) \ = \ \boldsymbol{H}_{\delta^m_{Voy}}(S), \tag{11.5}$$

où δ^m_{Voy} correspond à l'ensemble δ_{Voy} restreint aux enregistrements pour lesquels $M = m$. Nous utilisons une loi histogramme car celle-ci nous permet de capturer finement toute la complexité de la transformation réalisée par le conduit vocal humain. L'utilisation d'une loi gaussienne n'est pas appropriée à cause de la discrétisation des variables motrices : certains hypercubes de $M = J \wedge TB \wedge TD \wedge LH$ correspondant à une distribution multi-modale dans $S = F1 \wedge F2 \wedge F3$ (c'est-à-dire avec plusieurs maxima) que la loi gaussienne ne peut capturer correctement.

La Figure 11.5 résume les conséquences auditives de certains hypercubes de M sous la formes d'ellipses de dispersion. Les non-linéarités de la transformation articulatori-acoustique s'expriment alors à la fois dans la variabilité intra-hypercubes (dispersions variables d'une ellipse à une autre) et inter-hypercubes (répartition des ellipses non-homogène). La Figure 11.6 superpose au dictionnaire VLAM une représentation des probabilités de chaque région dans l'espace discrétisé F1-F2. Plus précisément, en considérant chaque hypercube de l'espace moteur comme équiprobable ($P(M \mid \pi_{Com}) = \boldsymbol{U}(M)$), chaque pavé représente la probabilité d'une valeur de formants dans $\mathcal{D}_{F1} \times \mathcal{D}_{F2}$, calculée par :

$$P(F1\ F2 \mid \delta_{Voy}\ \pi_{Com}) = \sum_{M,F3} P(S \mid M\ \delta_{Voy}\ \pi_{Com}), \tag{11.6}$$

où $S = F1 \wedge F2 \wedge F3$ d'après l'Équation 11.3. On observe que des tirages uniformes selon les variables articulatoires n'impliquent pas une distribution uniforme sur les variables formantiques correspondantes, effet dû à l'aspect fortement non-linéaire de la transformation articulatori-auditive considérée (Figure 11.5).

Dans l'algorithme de simulation, après avoir tiré le stimulus auditif correspondant au geste moteur m de l'agent locuteur selon la distribution $P(S \mid [M = m]\ \delta_{Voy}\ \pi_{Com})$, nous ajoutons un bruit gaussien sur chaque dimension formantique avant de transmettre ce stimulus à l'agent auditeur. Ces bruits gaussiens indépendants sur $F1$, $F2$ et $F3$ ont pour écarts-types respectifs σ_{F1}, σ_{F2}, σ_{F3}. Dans la présentation des résultats de nos simulations, ces stimuli seront enregistrés avant ajout du bruit de l'environnement, à la manière d'un micro qui serait placé au plus près de la sortie du conduit vocal de l'agent locuteur, avant que le signal soit bruité puis transmis à l'agent auditeur (les paramètres σ_{F1}, σ_{F2}, σ_{F3} ont donc un effet sur les stimuli reçus par les agents en situation d'auditeur).

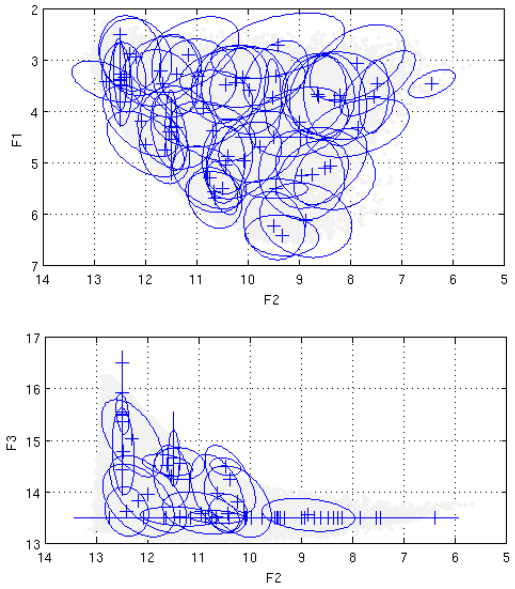


FIGURE 11.5 – Ellipses de dispersion à 1.5 écarts-types des conséquences auditives des hypercubes de M correspondant au produit cartésien des espaces restreints $\mathcal{D}_J = \{1\}$ $\mathcal{D}_{TD} = \mathcal{D}_{TB} = \{0, 2, 4, 6, 8\}$ et $\mathcal{D}_{LH} = \{1, 2, 3\}$ (les espaces sont restreints de façon à éviter de saturer la figure en nombre d'ellipses, en restant toutefois représentatif de la transformation).

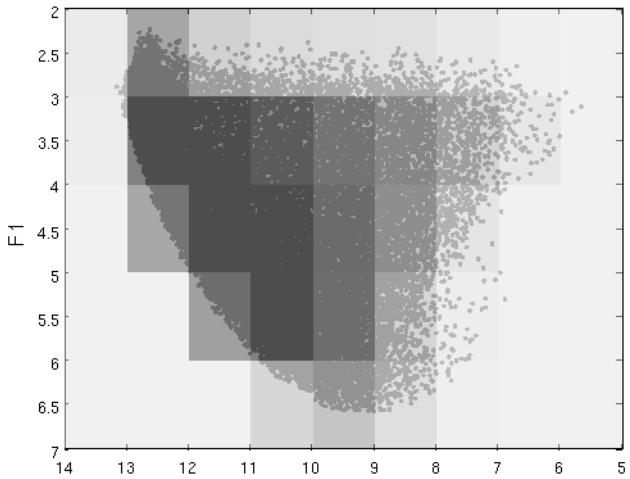


FIGURE 11.6 – Superposition du dictionnaire VLAM et de la probabilité de chaque couple de valeurs de $\mathcal{D}_{F1} \times \mathcal{D}_{F2}$. Plus un pavé est foncé, plus la probabilité dans F1-F2 calculée selon l'Équation 11.6 est grande.

11.3 Modèle d'agent

11.3.1 Ensemble d'apprentissage

Comme précédemment, l'ensemble d'apprentissage d'un agent a au temps t de la simulation se note $\delta_a(t)$ et est consitué de triplets :

$$(n, x, o_i) \in (([0,t] \times \mathcal{D}_M \times \mathcal{D}_{O_S}) \cup ([0,t] \times \mathcal{D}_S \times \mathcal{D}_{O_L}))\,.$$

Chacun de ces triplets associe à un jeu déictique $n \in [0,t]$ soit un couple $(m, o_i) \in \mathcal{D}_M \times \mathcal{D}_{O_S}$ si l'agent a un statut de locuteur au temps n, soit un couple $(s, o_i) \in \mathcal{D}_S \times \mathcal{D}_{O_L}$ s'il a un statut d'auditeur. Les jeux dans lesquels l'agent a n'intervient pas ne sont pas enregistrés dans $\delta_a(t)$. Nous notons $\delta_a^M(t)$ (respectivement $\delta_a^S(t)$) l'ensemble des $N_{App} = 200$ derniers éléments de $\delta_a(t)$ enregistrés en statut de locuteur (respectivement auditeur).

11.3.2 Système moteur

De la même façon que dans le chapitre précédent, le système moteur d'un agent a correspond à une famille de distributions sur ses gestes moteurs M conditionnée par les objets de l'environnement O_S et l'ensemble d'apprentissage $\delta_a t$. La différence est que la variable M est maintenant une conjonction $J \wedge TB \wedge TD \wedge LH$. On définit alors le système moteur par :

$$\begin{aligned} & P(M \mid O_S\ \delta_a(t)\ \pi_{Ag}) \\ &= P(J\ TB\ TD\ LH \mid O_S\ \delta_a(t)\ \pi_{Ag}) \\ &= P(J \mid \pi_{Ag}) P(TB \mid O_S\ \delta_a(t)\ \pi_{Ag}) P(TD \mid O_S\ \delta_a(t)\ \pi_{Ag}) P(LH \mid O_S\ \delta_a(t)\ \pi_{Ag}) \end{aligned}$$

où

$$P(J \mid \pi_{Ag}) = \boldsymbol{\delta}_1(J),$$

et

$$\begin{aligned} \forall X \in \{TB, TD, LH\}, o_i \in \mathcal{D}_{O_S}\ : & \\ P(X \mid [O_S = o_i]\ \delta_a(t)\ \pi_{Ag}) &= \boldsymbol{L}_{\delta_a^{M,o_i}(t)}(X), \end{aligned}$$

où $\delta_a^{M,o_i}(t)$ correspond à l'ensemble $\delta_a^M(t)$ restreint aux enregistrements pour lesquels $[O_S = o_i]$.

Ainsi, le système moteur correspond à une configuration neutre de la mâchoire à $J = 1$ (valeurs VLAM entre -1 et 1, d'après 11.2) indépendante de l'ensemble d'apprentissage et de l'objet considéré. Concernant les autres variables motrices (TB, TD, LH), chaque distribution correspondante est conditionnée par les objets et est une loi de succession de Laplace apprise à partir données des N_{App} derniers jeux déictiques dans lesquels l'agent a avait le statut de locuteur. Ceci signifie que l'ensemble des simulations vocaliques sont faites sur 3 degrés de liberté articulatoire (corps et dos de langue, et hauteur des lèvres) en prenant une mâchoire dans une position autour de la valeur neutre, pour réduire la dimensionnalité sans réduire les capacités génératives du modèle.

11.3.3 Lien sensori-moteur

Le lien sensori-moteur est identique pour tous les agents de la société et est appris préalablement à la simulation comme une loi de succession de Laplace à partir du dictionnaire VLAM δ_{Voy} :

$$\begin{aligned} &\forall m \in M \ : \\ P(S \mid [M = m]\ \delta_{Voy}\ \pi_{Ag}) \ &= \ \boldsymbol{L}_{\delta^m_{Voy}}(F1\ F2\ F3), \end{aligned}$$

où δ^m_{Voy} correspond à l'ensemble δ_{Voy} restreint aux enregistrements pour lesquels $M = m$.

11.3.4 Système auditif

De même que dans le modèle simplifié à une dimension du chapitre précédent, le système auditif est un classifieur gaussien. Les prototypes auditifs correspondent ici à un produit de loi gaussiennes pour chaque variable formantique :

$$P(O_L \mid F1\ F2\ F3\ \delta_a(t)\ \pi_{Ag}) = \frac{P(F1 \mid O_L\ \delta_a(t)\ \pi_{Ag})\ P(F2 \mid O_L\ \delta_a(t)\ \pi_{Ag})\ P(F3 \mid O_L\ \delta_a(t)\ \pi_{Ag})}{\sum_{O_L} P(F1 \mid O_L\ \delta_a(t)\ \pi_{Ag})\ P(F2 \mid O_L\ \delta_a(t)\ \pi_{Ag})\ P(F3 \mid O_L\ \delta_a(t)\ \pi_{Ag})}$$

où

$$\begin{aligned} &\forall X \in \{F1, F2, F3\}, o_i \in \mathcal{D}_{O_L} \ : \\ &P(X \mid [O_L = o_i]\ \delta_a(t)\ \pi_{Ag}) \ = \ \boldsymbol{G}_{\delta_a^{S,o_i}}(X), \end{aligned}$$

où $\delta_a^{S,o_i}(t)$ correspond à l'ensemble $\delta_a^S(t)$ restreint aux enregistrements pour lesquels $O_L = o_i$. L'utilisation d'une loi gaussienne permet de modéliser une notion de distance dans l'espace auditif (la probabilité d'un stimulus diminuant lorsque la distance à la moyenne de la distribution augmente). Ceci est nécessaire pour obtenir un comportement dispersif.

11.4 Simulations

Nous souhaitons montrer la cohérence globale des systèmes de voyelles émergeant de nos simulations aux systèmes de 3 et 5 voyelles des langues du monde décrits à la section 11.1. Les objets sémantiques issus du comportement de déixis, et phonologiques issus des prototypes articulatori-auditifs sont confondus, nous nous limiterons donc à des environnements de 3 et 5 objets (conduisant à l'émergence de systèmes de 3 et 5 voyelles, respectivement).

Dans le but de réduire le temps de simulation, nous nous concentrons sur des sociétés de 2 agents procédant à une série de 150 000 jeux déictiques. Rappelons qu'à partir du présent chapitre, toutes les simulations considérées concernent des sociétés d'agents en

comportement sensori-moteur.

$$\begin{aligned} N_A &= 2, \\ N_O &= 3 \text{ ou } 5 \text{ selon les simulations}, \\ N_{JD} &= 150\,000. \end{aligned}$$

11.4.1 Paramètres variés dans les simulations

Les écarts-types des bruits gaussiens de l'environnement sur chaque dimension formantique, σ_{F1}, σ_{F2}, σ_{F3}, nous permettent de tester différentes hypothèses quant au poids relatif de chacune d'entre elles, dans le but de comparer nos résultats de simulation à ceux de Schwartz et collab. (1997b) décrits à la Section 11.1. Ainsi, σ_{F2} et σ_{F3} sont définis comme des multiples de σ_{F1}. Nous avons donc :

$$\begin{aligned} \sigma_{F2} &= \alpha_{F2}\, \sigma_{F1}, \\ \sigma_{F3} &= \alpha_{F3}\, \sigma_{F1}, \end{aligned}$$

où α_{F2} et α_{F3} sont les coefficients de bruit sur les dimensions F2 et F3 par rapport à F1.

Pour chaque cardinalité de système vocalique (3 ou 5 éléments), nous distinguons alors deux cas :

- Les trois formant F1, F2, F3 ont des poids identiques, correspondant à des bruits relatifs de rapports 1-1-1 :
$$\alpha_{F2} = \alpha_{F3} = 1 \ ;$$
- F1 a 3 fois plus de poids que F2, qui a 2 fois plus de poids que F3, soient des bruits relatifs de rapports 1-3-6 (selon Schwartz et collab. (1997b)) :
$$\alpha_{F2} = 3 \ ; \ \alpha_{F3} = 6 \ ;$$

Pour la première hypothèse, nous faisons varier σ_{F1} de 0.1 à 4.1 par pas de 0.5. Pour la seconde, nous faisons varier σ_{F1} de 0.1 à 0.9 par pas de 0.1. Ces deux échelles nous permettent une variation sur F3 relativement similaire pour les deux hypothèses.

11.4.2 Évaluation

L'évaluation des résultats de simulation est un problème compliqué qui fait en général appel dans les travaux de la littérature, tels que ceux exposés à la Section 4.3 du Chapitre 4, à des procédures partiellement manuelles et en partie arbitraires. Comme le note Oudeyer (2003), ce qui compte avant tout dans ces simulations est la cohérence globale avec les données plus que des simulations précises des systèmes existants. Nous avons opté dans ce travail pour des outils automatiques et donc dénués de critères subjectifs difficiles à contrôler.

La discrétisation de nos variables motrices et auditives implique dans nos simulations des approximations qui ne nous permettent pas d'utiliser la grille de classification UPSID de

la Figure 11.1 de façon satisfaisante. Comme nous l'avons précisé plus haut, nous souhaitons seulement montrer la cohérence globale de nos résultats de simulation avec les données, en particulier une nette préférence pour les systèmes « en forme de V ». Ainsi, nous divisons le plan F1-F2 en 7 régions comme indiqué Figure 11.7.

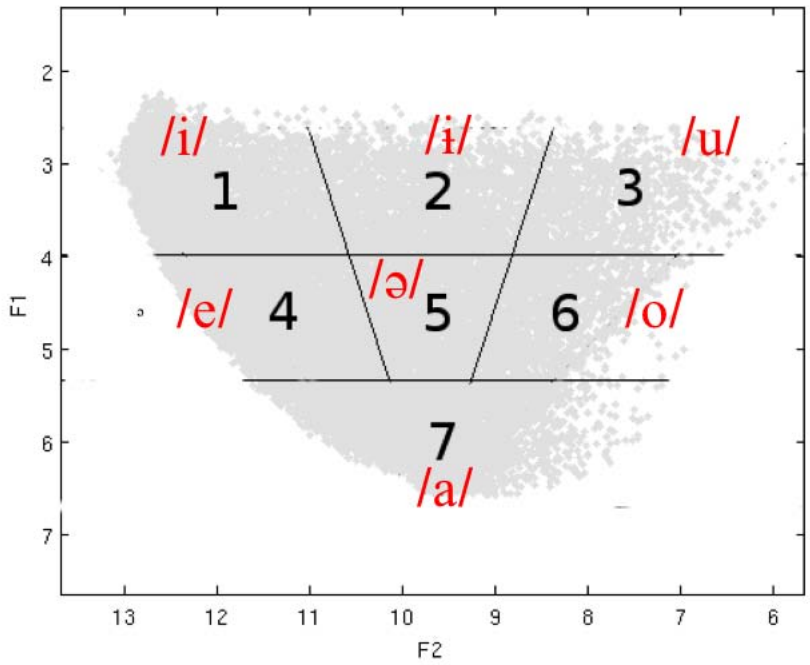


FIGURE 11.7 – Classification des voyelles émergeant de nos simulations dans le triangle vocalique et correspondance avec des symboles phonétiques usuels.

Pour chaque simulation, la voyelle correspondant à chaque objet est classifiée dans une des 7 régions de la Figure 11.7 selon la moyenne sur F1-F2 des stimuli produits par l'ensemble des agents de la société avant ajout du bruit de l'environnement pendant les 3000 derniers jeux déictiques.

Combien de systèmes différents sont alors possibles selon cette classification ? Un système de N_O voyelles correspond à une combinaison avec répétition de N_O chiffres dans un ensemble de 7 (les répétitions sont possibles pour certains jeux de paramètres conduisant à l'émergence de « mauvais » systèmes). Le nombre de systèmes de N_O voyelles possibles est donc égal à :

$$\binom{7+N_O-1}{N_O} = \frac{(7+N_O-1)!}{N_O!(7-1)!}.$$

Ainsi, il y a 84 systèmes possibles à $N_O = 3$ éléments et 462 systèmes à $N_O = 5$ éléments. Parmi ceux-ci, les systèmes majoritaires dans UPSID correspondent dans notre classification à /1, 3, 7/ (pour /i, u, a/, dans l'ordre) et /1, 3, 4, 6, 7/ (pour /i, u, e, o, a/).

Ce processus de classification, bien que très simplifié par rapport à ceux utilisés pour

les systèmes des langues du monde, nous permet d'analyser les fréquences d'occurrence de chaque système pour un jeu de paramètres donné.

11.5 Résultats

Nous exposons nos résultats pour les quatre jeux de paramètres suivants :

- rapports de bruits 1-1-1 ($\alpha_{F2} = \alpha_{F3} = 1$) :
  - systèmes de 3 voyelles ($N_O = 3$) : Section 11.5.1.1,
  - systèmes de 5 voyelles ($N_O = 5$) : Section 11.5.1.2,
- rapports de bruits 1-3-6 ($\alpha_{F2} = 3, \alpha_{F3} = 6$) :
  - systèmes de 3 voyelles ($N_O = 3$) : Section 11.5.2.1,
  - systèmes de 5 voyelles ($N_O = 5$) : Section 11.5.2.2.

Pour chacun, 10 simulations indépendantes sont effectuées pour 9 valeurs différentes de σ_{F1}, soient 360 simulations dont le temps d'exécution est de l'ordre de 10 minutes chacune.

11.5.1 Rapports de bruit 1-1-1

Ces résultats proviennent de simulations dont les poids de chaque dimension formantique sont égaux, soient $\alpha_{F2} = \alpha_{F3} = 1$ (le bruit de l'environnement est identique sur chaque dimension). Nous réalisons des séries de 10 simulations indépendantes pour chaque valeur de σ_{F1} allant de 0.1 à 4.1 par pas de 0.5.

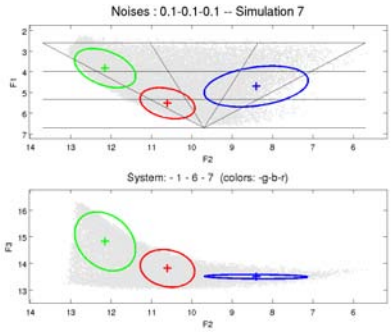
11.5.1.1 Systèmes de 3 voyelles

Nous exposons les résultats de simulations dans un environnement de 3 objets, menant à l'émergence de systèmes de 3 voyelles. Rappelons que le système de trois voyelles le plus fréquent dans les langues du monde est /i, u, a/ (Figure 11.1), dont l'équivalent dans notre système de classification (Figure 11.7) est le système /1, 3, 7/.

Les résultats sont exposés dans les 9 tables qui suivent. Chacune correspond à un niveau de bruit, pour lequel on donne le pourcentage de chaque système dans 10 simulations, ainsi qu'une visualisation d'une simulation particulière appartenant au système le plus fréquent (ou à l'un de ces systèmes dans le cas d'ex-aequo) et dont la valeur de dispersion est proche de la moyenne des 10 simulations de ce niveau de bruit. Cette visualisation montre les ellipses de dispersion à 1.5 écarts-types (soit 68% des données enregistrées) dans les plans F1-F2 et F2-F3 des stimuli auditifs produits par les agents pour chaque objet sur les 3000 derniers jeux déictiques, avant ajout du bruit dans l'environnement. Les ellipses de la même couleur correspondent à la même voyelle. Au milieu des deux plans est indiqué la classification (*System*) et le code de couleur correspondants (*colors*) : r (rouge), g (vert), et b (bleu).

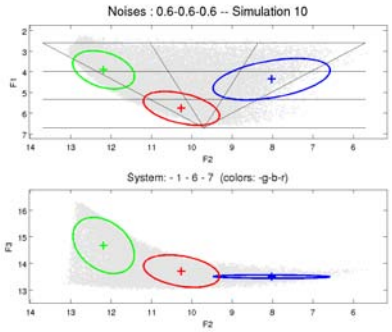
$\sigma_{F1} = 0.1$; $\sigma_{F2} = 0.1$; $\sigma_{F3} = 0.1$

Système	Pourcentage sur 10 simulations
1 6 7	50 %
1 4 6	30 %
4 6 7	20 %



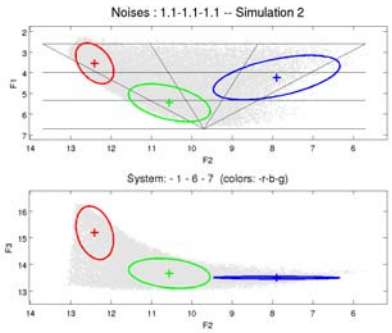
$\sigma_{F1} = 0.6$; $\sigma_{F2} = 0.6$; $\sigma_{F3} = 0.6$

Système	Pourcentage sur 10 simulations
1 4 6	30 %
1 3 7	30 %
1 6 7	30 %
4 6 7	10 %



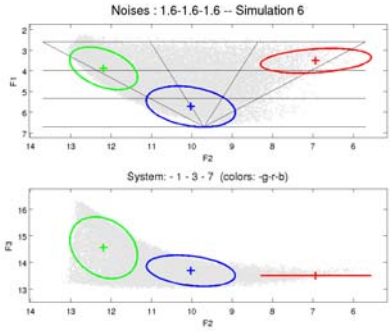
$\sigma_{F1} = 1.1$; $\sigma_{F2} = 1.1$; $\sigma_{F3} = 1.1$

Système	Pourcentage sur 10 simulations
1 3 7	50 %
1 6 7	50 %



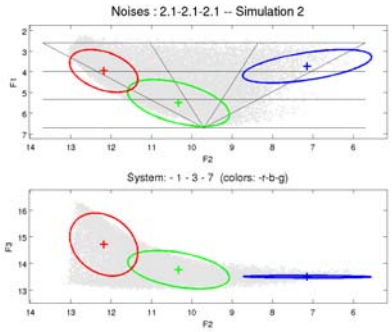
$\sigma_{F1} = 1.6$; $\sigma_{F2} = 1.6$; $\sigma_{F3} = 1.6$

Système	Pourcentage sur 10 simulations
1 3 7	90 %
3 4 7	10 %



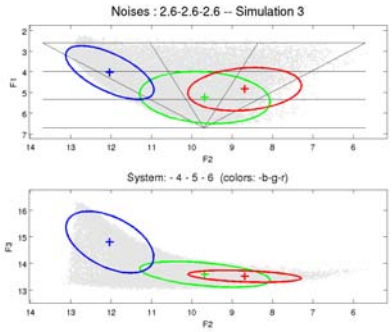
$\sigma_{F1} = 2.1$; $\sigma_{F2} = 2.1$; $\sigma_{F3} = 2.1$

Système	Pourcentage sur 10 simulations
1 3 7	50 %
3 4 7	20 %
4 6 7	10 %
1 6 7	10 %
4 4 6	10 %



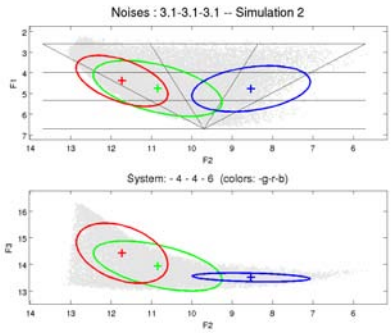
$\sigma_{F1} = 2.6$; $\sigma_{F2} = 2.6$; $\sigma_{F3} = 2.6$

Système	Pourcentage sur 10 simulations
4 5 6	30 %
3 4 7	20 %
4 4 6	20 %
1 5 6	20 %
4 6 7	10 %



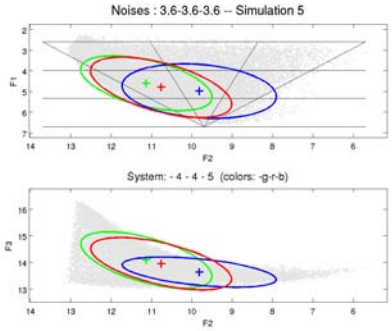
$\sigma_{F1} = 3.1$; $\sigma_{F2} = 3.1$; $\sigma_{F3} = 3.1$

Système	Pourcentage sur 10 simulations
4 4 6	70 %
4 5 6	20 %
1 4 6	10 %



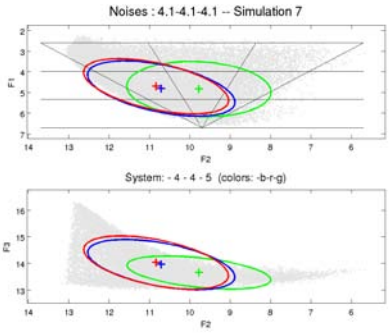
$\sigma_{F1} = 3.6$; $\sigma_{F2} = 3.6$; $\sigma_{F3} = 3.6$

Système	Pourcentage sur 10 simulations
4 5 5	50 %
4 4 5	50 %



$\sigma_{F1} = 4.1$; $\sigma_{F2} = 4.1$; $\sigma_{F3} = 4.1$

Système	Pourcentage sur 10 simulations
4 4 5	60 %
4 5 5	30 %
4 4 4	10 %



On observe ici encore l'effet du bruit sur la dispersion dans une société d'agents en comportement sensori-moteur, que nous avions constaté au chapitre précédent. Pour de faibles valeurs de bruits, les systèmes se dispersent peu mais suffisamment pour que leurs éléments soient distinguables. Puis en augmentant les valeurs de bruit, les systèmes sont de plus en plus dispersés jusqu'à obtenir le système /1, 3, 7/ (9 des 10 simulations pour $\sigma_{F1} = 1.6$), correspondant au système /a, i, u/ majoritaire dans les langues du monde. Enfin, l'excès de bruit entraine des systèmes moins dispersés pour lesquels les agents n'ont pas pu converger vers un système efficace. La Figure 11.8 expose cet effet par la courbe de dispersion moyenne des systèmes obtenus à chaque niveau de bruit. Le logarithme de la mesure de Lindblom est calculé à partir des distances euclidiennes en Barks dans l'espace F1-F2-F3.

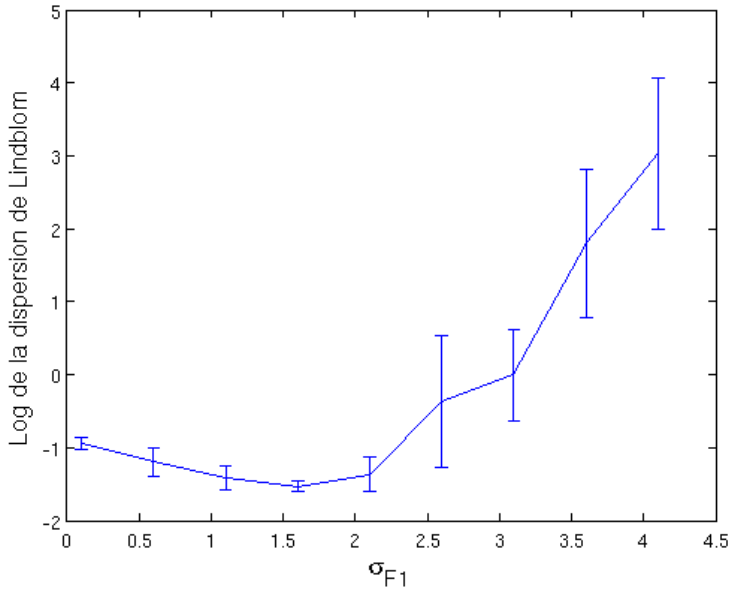


FIGURE 11.8 – Logarithme de la mesure de Lindblom des systèmes à 3 voyelles en fonction du bruit sur F1 pour des rapports 1-1-1. Moyennes et écarts-types sur 10 simulations indépendantes. Nous rappelons que plus cette mesure est petite, plus les systèmes sont dispersés.

À partir de la Figure 11.8 et des 9 tables ci-dessus, on extrait les niveaux de bruit « convenables », permettant l'émergence de systèmes dont les éléments sont correctement distinguables (ici, jusqu'à $\sigma_{F1} = 2.1$).

Sur la Figure 11.9, nous faisons figurer l'évolution des pourcentages d'apparition des systèmes avec le niveau de bruit en ne conservant, pour améliorer la lisibilité, que les systèmes qui apparaissent avec un pourcentage non négligeable aux niveaux dits convenables définis ci-dessus. On observe que des systèmes à dispersion non optimale dominent à faible bruit et que le système optimalement dispersé /i,u,a/ s'impose aux bruits moyens. Quant aux bruits forts, les 9 tables ci-dessus montrent que les conditions de communication ne

permettent plus de dispersion.

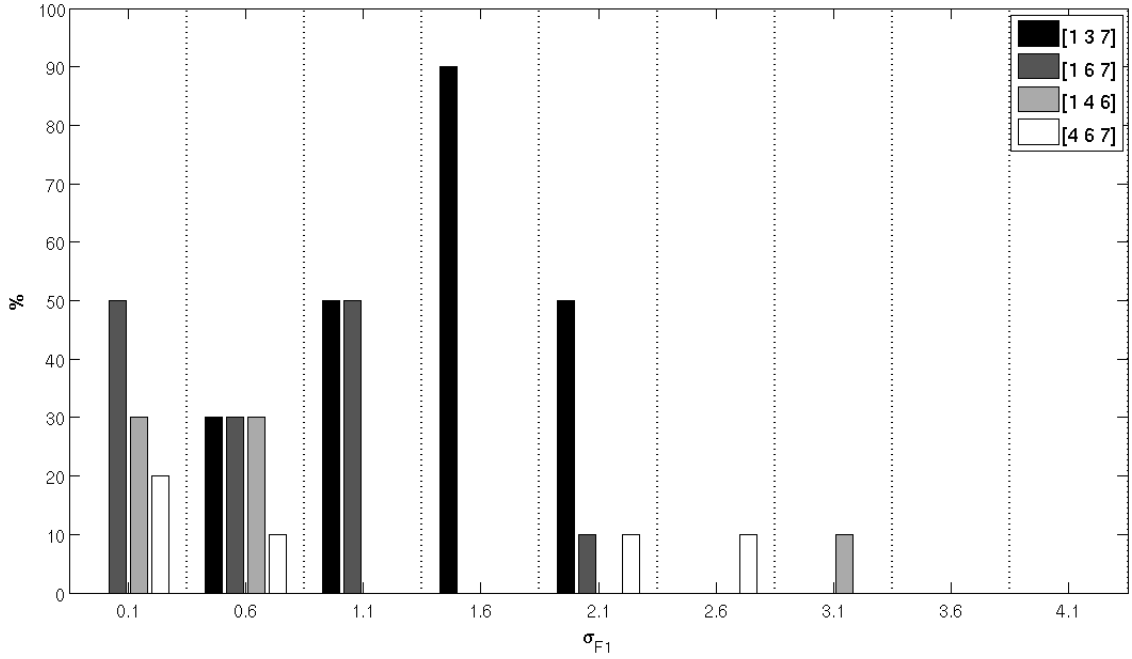


FIGURE 11.9 – Pourcentage de différents systèmes de 3 voyelles obtenus avec des rapports de bruit 1-1-1.

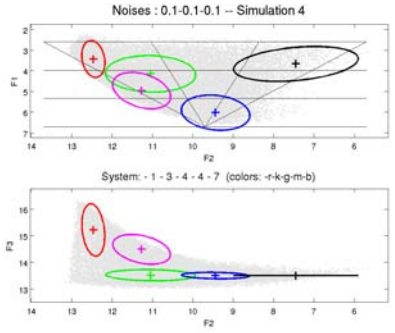
11.5.1.2 Systèmes de 5 voyelles

Nous exposons les résultats de simulations dans un environnement de 5 objets, menant à l'émergence de systèmes de 5 voyelles (les bruits de l'environnement sur chaque dimension formantique sont toujours identiques).

Dans les langues du monde, le système de 5 voyelles le plus fréquent correspond à /i, u, e, o, a/ (Figure 11.1), dont l'équivalent dans notre système de classification (Figure 11.7) est le système /1, 3, 4, 6, 7/. Les résultats sont exposés dans les 9 tables qui suivent, avec les mêmes conventions que pour les systèmes de 3 voyelles exposés précédemment.

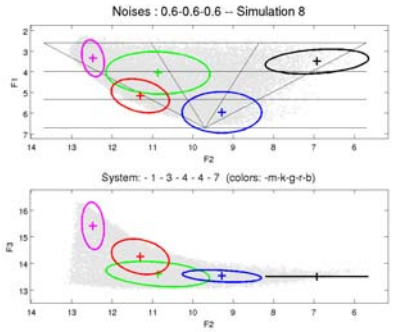
$\sigma_{F1} = 0.1$; $\sigma_{F2} = 0.1$; $\sigma_{F3} = 0.1$

Système	Pourcentage sur 10 simulations
1 3 4 7 7	40 %
1 3 4 4 7	40 %
1 3 4 6 7	10 %
1 3 4 5 7	10 %



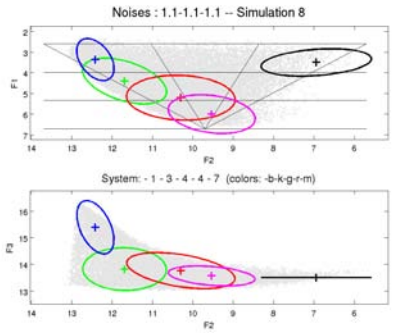
$\sigma_{F1} = 0.6$; $\sigma_{F2} = 0.6$; $\sigma_{F3} = 0.6$

Système	Pourcentage sur 10 simulations
1 3 4 4 7	50 %
1 1 3 4 7	30 %
1 3 4 7 7	10 %
1 3 4 6 7	10 %



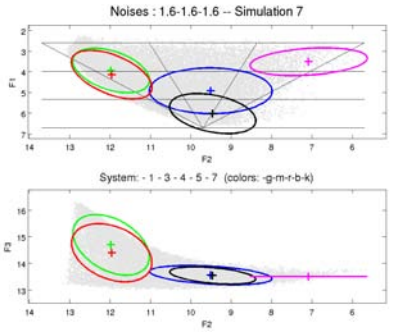
$\sigma_{F1} = 1.1$; $\sigma_{F2} = 1.1$; $\sigma_{F3} = 1.1$

Système	Pourcentage sur 10 simulations
1 3 4 4 7	60 %
1 3 4 6 7	30 %
1 3 4 5 7	10 %



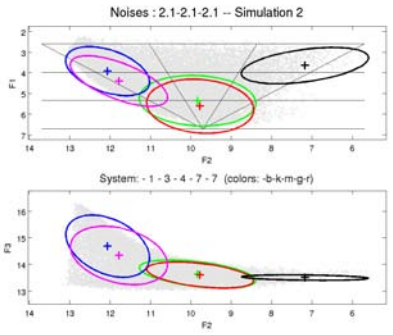
$\sigma_{F1} = 1.6$; $\sigma_{F2} = 1.6$; $\sigma_{F3} = 1.6$

Système	Pourcentage sur 10 simulations
1 3 4 5 7	50 %
1 3 4 7 7	20 %
1 3 4 4 7	20 %
1 1 3 4 7	10 %



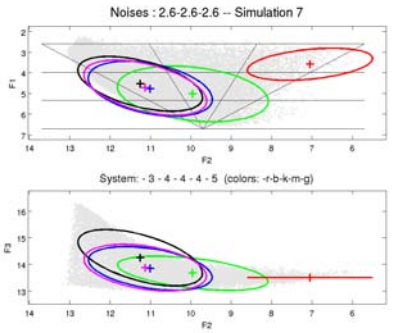
$\sigma_{F1} = 2.1$; $\sigma_{F2} = 2.1$; $\sigma_{F3} = 2.1$

Système	Pourcentage sur 10 simulations
1 3 4 7 7	20 %
3 4 4 4 7	20 %
1 3 4 5 7	20 %
1 3 4 4 6	10 %
3 4 4 5 7	10 %
4 4 5 6 6	10 %
1 3 4 5 6	10 %



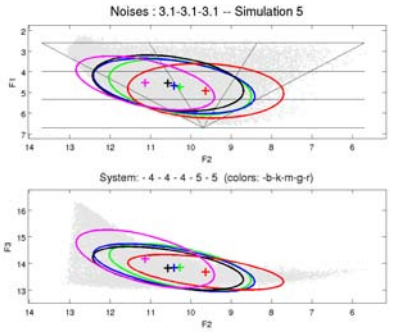
$\sigma_{F1} = 2.6$; $\sigma_{F2} = 2.6$; $\sigma_{F3} = 2.6$

Système	Pourcentage sur 10 simulations
3 4 4 4 5	20 %
4 4 5 5 6	20 %
1 4 4 5 6	10 %
1 4 4 6 6	10 %
4 4 4 5 6	10 %
4 5 5 5 6	10 %
4 4 4 6 6	10 %
3 4 4 4 4	10 %



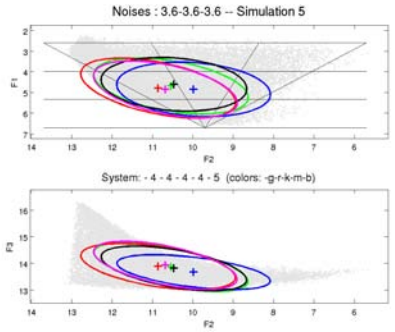
$\sigma_{F1} = 3.1$; $\sigma_{F2} = 3.1$; $\sigma_{F3} = 3.1$

Système	Pourcentage sur 10 simulations
4 4 4 5 5	70 %
4 4 4 4 5	10 %
4 4 5 5 5	10 %
4 4 4 5 6	10 %



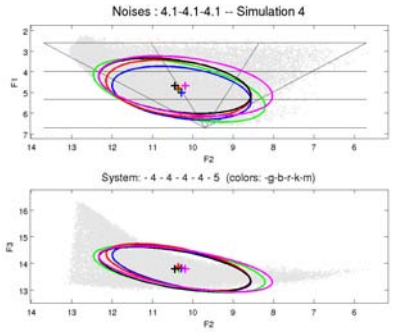
$\sigma_{F1} = 3.6$; $\sigma_{F2} = 3.6$; $\sigma_{F3} = 3.6$

Système	Pourcentage sur 10 simulations
4 4 4 5 5	50 %
4 4 4 4 5	50 %



$\sigma_{F1} = 4.1$; $\sigma_{F2} = 4.1$; $\sigma_{F3} = 4.1$

Système	Pourcentage sur 10 simulations
4 4 4 4 5	50 %
4 4 4 5 5	30 %
4 4 4 4 4	10 %
4 5 5 5 5	10 %



La courbe de dispersion des simulations à 5 voyelles en rapport de bruit 1-1-1 est exposée Figure 11.10. D'après celle-ci et les 9 tables ci-dessus, on extrait les niveaux de bruit « convenables », pour lesquels les éléments des systèmes sont correctement distinguables (ici, jusqu'à $\sigma_{F1} = 1.1$).

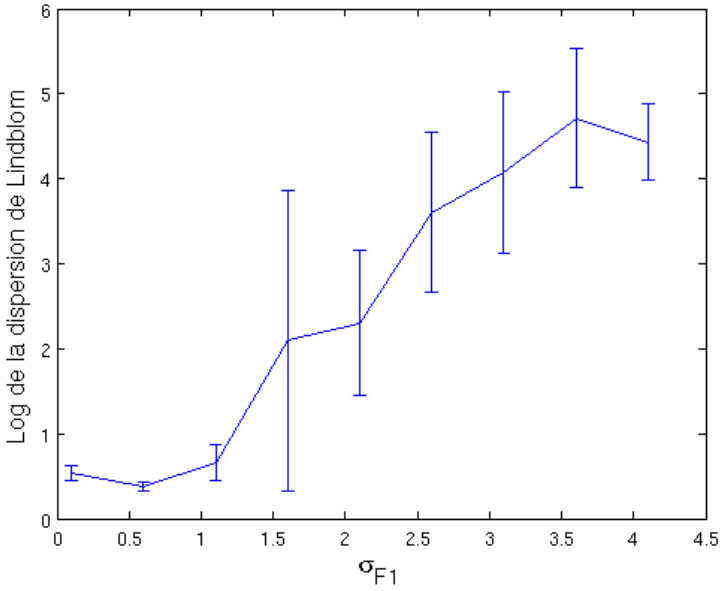


FIGURE 11.10 – Logarithme de la mesure de Lindblom des systèmes à 5 voyelles en fonction du bruit sur F1 pour des rapports 1-1-1. Moyennes et écarts-types sur 10 simulations indépendantes.

Sur la Figure 11.11, nous faisons figurer l'évolution des pourcentages d'apparition des systèmes avec le niveau de bruit en ne conservant, pour améliorer la lisibilité, que les systèmes qui apparaissent avec un pourcentage non négligeable aux niveaux dits convenables définis ci-dessus. Contrairement aux systèmes de 3 voyelles, on observe que le système de 5 voyelles majoritaire dans les langues du monde (/1, 3, 4, 6, 7/) n'est pas privilégié dans nos simulations pour des rapports de bruits 1-1-1. Le gagnant est ici /1, 3, 4, 4, 7/, système dans lequel les deux éléments de la classe /4/ se différencient sur la dimension F3 (voir les trois premières tables ci-dessus ($\sigma_{F1} \in \{0.1, 0.6, 1.1\}$)). En effet, les poids identiques des trois dimensions formantiques permettent une utilisation efficace de F3 en terme de dispersion qui limite en conséquence celle dans le plan F1-F2.

Ainsi, si des poids identiques sur chaque dimension formantique (rapports de bruits 1-1-1) laissent émerger dans nos simulations des systèmes de 3 voyelles relativement en accord avec les systèmes correspondants des langues du monde (dispersés plutôt dans le plan F1-F2 dont la surface vocalique est supérieure à celle du plan F2-F3), ce n'est pas le cas des systèmes de 5 voyelles (dont le plus grand nombre d'éléments à communiquer tire avantage de l'utilisation de F3).

On remarque également dans nos simulations une tendance vers les voyelles avant

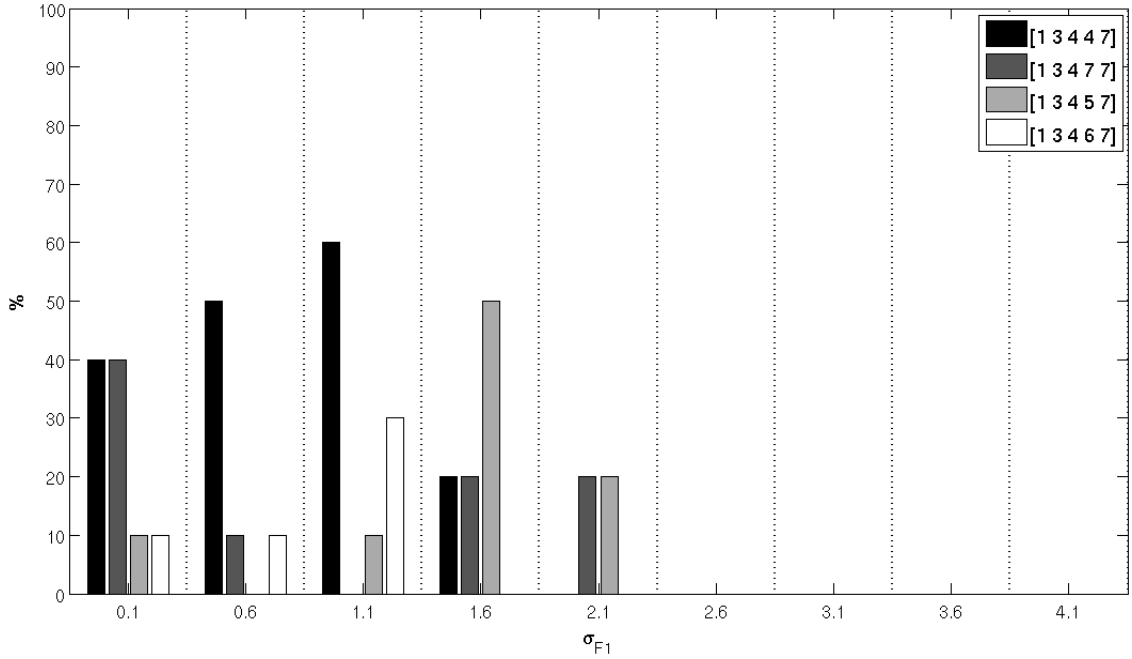


FIGURE 11.11 – Pourcentage de différents systèmes de 5 voyelles obtenus avec des rapports de bruit 1-1-1.

(à gauche dans le triangle vocalique). Celle-ci, également présente dans les langues du monde (légèrement visible sur les systèmes à nombre pair d'éléments de la Figure 11.1, voir Schwartz et collab., 1997a, pour une analyse plus détaillée), peut s'expliquer ici par la prédominance des stimuli auditifs avec une valeur élevée de F2 (Figure 11.6), dûe à l'aspect fortement non-linéaire de la transformation articulatori-auditive considérée.

11.5.2 Rapports de bruit 1-3-6

Les systèmes de 5 voyelles émergeant de nos simulations avec des poids identiques sur chaque dimension formantique n'étant pas en accord avec les données des langues du monde, nous testons à présent l'hypothèse de Schwartz et collab. (1997b) correspondant à des bruits de l'environnement de rapports 1-3-6 :

$$\begin{aligned} \alpha_{F2} &= 3, \\ \alpha_{F3} &= 6. \end{aligned}$$

Nous réalisons des séries de 10 simulations indépendantes pour chaque valeur de σ_{F1} allant de 0.1 à 0.9 par pas de 0.1.

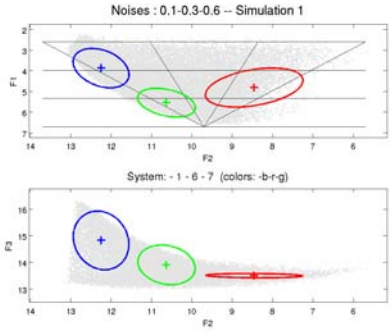
11.5.2.1 Systèmes de 3 voyelles

Nous exposons les résultats de simulations dans un environnement de 3 objets, menant à l'émergence de systèmes de 3 voyelles.

Rappelons que le système de trois voyelles le plus fréquent dans les langues du monde est /i, u, a/, dont l'équivalent dans notre système de classification (Figure 11.7) est le système /1, 3, 7/. Les 9 tables qui suivent exposent nos résultats, avec les mêmes conventions que précédemment.

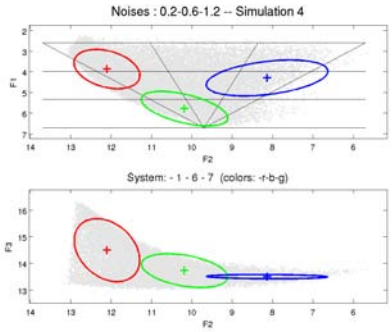
$\sigma_{F1} = 0.1$; $\sigma_{F2} = 0.3$; $\sigma_{F3} = 0.6$

Système	Pourcentage sur 10 simulations
1 6 7	50 %
1 4 6	50 %



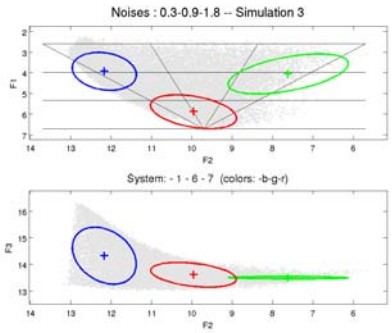
$\sigma_{F1} = 0.2$; $\sigma_{F2} = 0.6$; $\sigma_{F3} = 1.2$

Système	Pourcentage sur 10 simulations
1 6 7	80 %
4 6 7	10 %
1 3 7	10 %



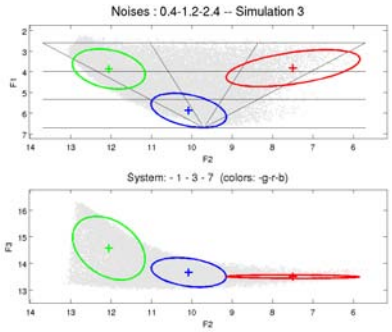
$\sigma_{F1} = 0.3$; $\sigma_{F2} = 0.9$; $\sigma_{F3} = 1.8$

Système	Pourcentage sur 10 simulations
1 6 7	60 %
1 3 7	30 %
4 6 7	10 %



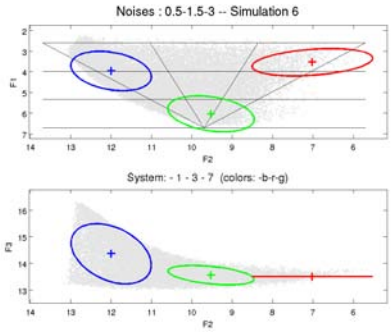
$\sigma_{F1} = 0.4$; $\sigma_{F2} = 1.2$; $\sigma_{F3} = 2.4$

Système	Pourcentage sur 10 simulations
1 3 7	70 %
1 6 7	20 %
3 4 7	10 %



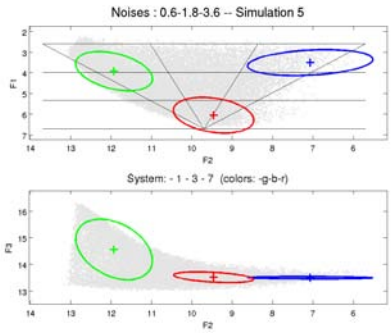
$\sigma_{F1} = 0.5$; $\sigma_{F2} = 1.5$; $\sigma_{F3} = 3$

Système	Pourcentage sur 10 simulations
1 3 7	70 %
3 4 7	30 %



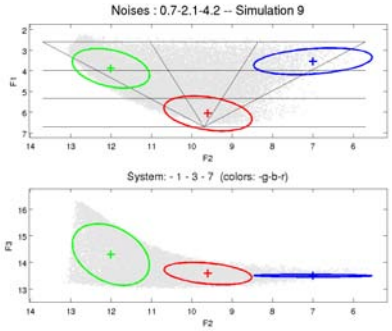
$\sigma_{F1} = 0.6$; $\sigma_{F2} = 1.8$; $\sigma_{F3} = 3.6$

Système	Pourcentage sur 10 simulations
1 3 7	100 %



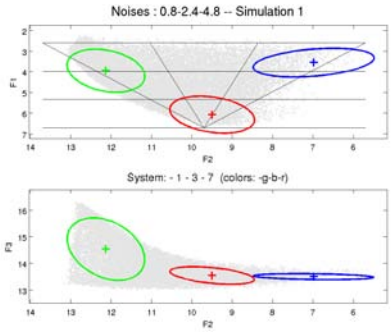
$\sigma_{F1} = 0.7$; $\sigma_{F2} = 2.1$; $\sigma_{F3} = 4.2$

Système	Pourcentage sur 10 simulations
1 3 7	60 %
3 4 7	40 %



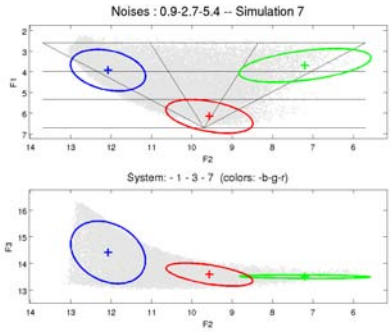
$\sigma_{F1} = 0.8$; $\sigma_{F2} = 2.4$; $\sigma_{F3} = 4.8$

Système	Pourcentage sur 10 simulations
1 3 7	90 %
3 4 7	10 %



$\sigma_{F1} = 0.9$; $\sigma_{F2} = 2.7$; $\sigma_{F3} = 5.4$

Système	Pourcentage sur 10 simulations
1 3 7	90 %
1 5 7	10 %



La Figure 11.12 montre la courbe de dispersion moyenne des systèmes obtenus à chaque niveau de bruit. Le logarithme de la mesure de Lindblom est toujours calculée à partir des distances en Barks dans l'espace F1-F2-F3, mais cette fois pondérées selon les rapports de bruit 1-3-6 (distances sur F2 divisées par 3, sur F3 divisées par 6).

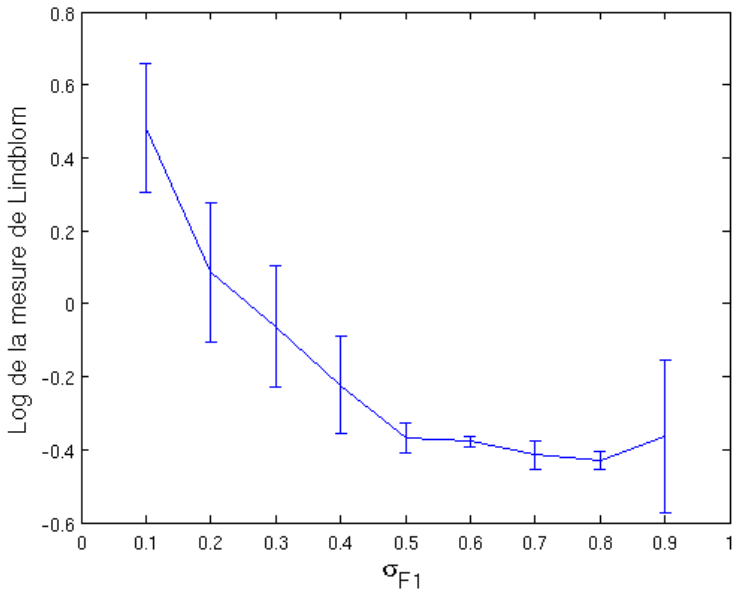


FIGURE 11.12 – Logarithme de la mesure de Lindblom en fonction du bruit sur F1 pour des systèmes de 3 voyelles avec des rapports de bruit 1-3-6. Moyennes et écarts-types sur 10 simulations indépendantes.

À partir de la Figure 11.12 et des 9 tables ci-dessus, on extrait les niveaux de bruit « convenables », permettant l'émergence de systèmes dont les éléments sont correctement distinguables. Il s'agit en fait ici de tout l'intervalle de σ_{F1}, les variations utilisées ne permettant pas d'observer clairement une concentration des éléments en environnement trop bruité (c'est toutefois déjà le cas de quelques simulations à $\sigma_{F1} = 0.9$, comme en témoigne le grand écart-type du dernier niveau de bruit de la Figure 11.12).

La Figure 11.13 montre les pourcentages des systèmes majoritaires. Les systèmes moins dispersés (différents de /1,3,7/) sont présents sur des intervalles de bruit relativement restreints, soit pour des niveaux faibles, soit pour niveaux plus forts. Le système majoritaire, qui couvre le plus large intervalle, est /1,3,7/.

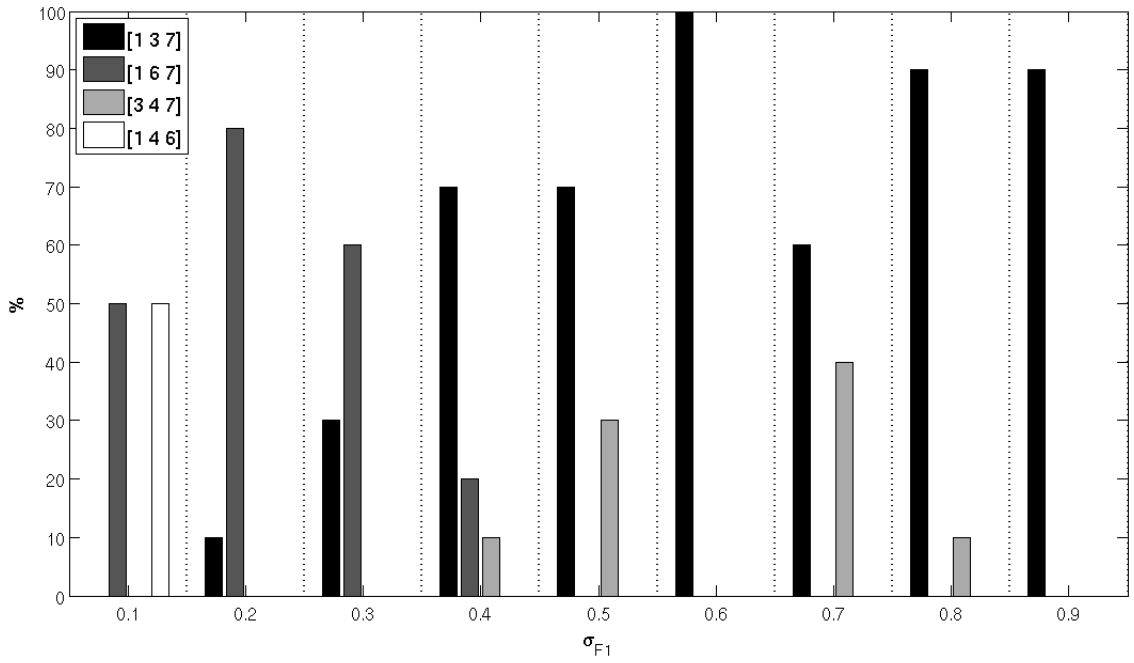


FIGURE 11.13 – Pourcentage de différents systèmes de 3 voyelles obtenus avec des rapports de bruit 1-3-6.

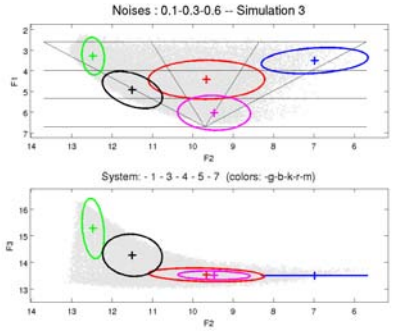
11.5.2.2 Systèmes de 5 voyelles

Nous exposons les résultats de simulations dans un environnement de 5 objets, menant à l'émergence de systèmes de 5 voyelles (toujours avec des rapports de bruit 1-3-6).

Rappelons que le système de 5 voyelles le plus fréquent dans les langues du monde correspond à /i, u, e, o, a/, dont l'équivalent dans notre système de classification (Figure 11.7) est le système /1,3,4,6,7/. Les résultats sont exposés dans les 9 tables qui suivent, avec les mêmes conventions que précédemment.

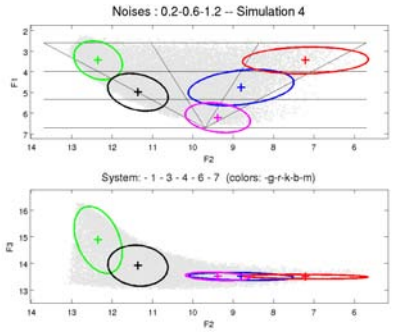
$\sigma_{F1} = 0.1$; $\sigma_{F2} = 0.3$; $\sigma_{F3} = 0.6$

Système	Pourcentage sur 10 simulations
1 3 4 5 7	50 %
1 3 4 6 7	30 %
1 3 4 7 7	20 %



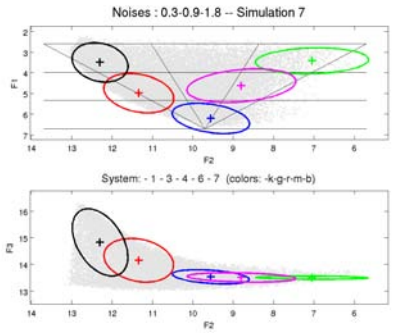
$\sigma_{F1} = 0.2$; $\sigma_{F2} = 0.6$; $\sigma_{F3} = 1.2$

Système	Pourcentage sur 10 simulations
1 3 4 6 7	40 %
1 3 4 5 7	30 %
1 3 4 7 7	30 %



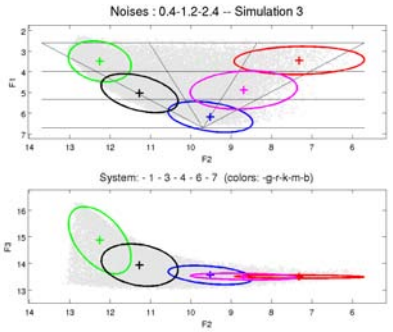
$\sigma_{F1} = 0.3$; $\sigma_{F2} = 0.9$; $\sigma_{F3} = 1.8$

Système	Pourcentage sur 10 simulations
1 3 4 6 7	70 %
1 3 4 5 7	20 %
1 2 3 4 7	10 %



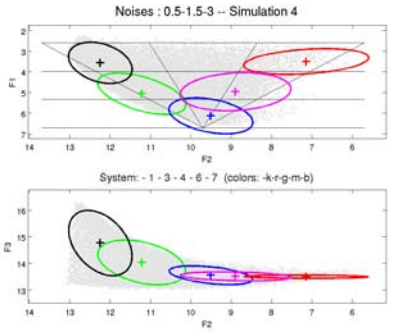
$\sigma_{F1} = 0.4$; $\sigma_{F2} = 1.2$; $\sigma_{F3} = 2.4$

Système	Pourcentage sur 10 simulations
1 3 4 6 7	90 %
1 3 4 5 7	10 %



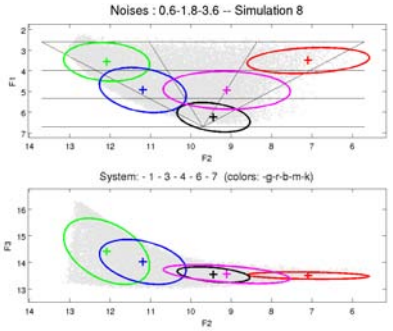
$\sigma_{F1} = 0.5$; $\sigma_{F2} = 1.5$; $\sigma_{F3} = 3$

Système	Pourcentage sur 10 simulations
1 3 4 6 7	70 %
1 3 4 4 7	10 %
1 3 5 7 7	10 %
1 3 4 5 7	10 %



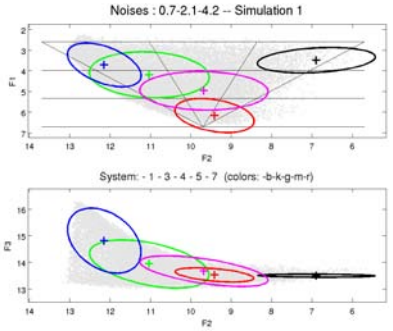
$\sigma_{F1} = 0.6$; $\sigma_{F2} = 1.8$; $\sigma_{F3} = 3.6$

Système	Pourcentage sur 10 simulations
1 3 4 6 7	50 %
1 3 4 5 7	30 %
1 3 4 4 7	20 %



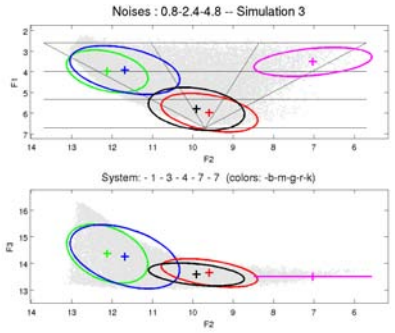
$\sigma_{F1} = 0.7$; $\sigma_{F2} = 2.1$; $\sigma_{F3} = 4.2$

Système	Pourcentage sur 10 simulations
1 3 4 5 7	50 %
1 3 4 7 7	40 %
1 3 4 4 7	10 %



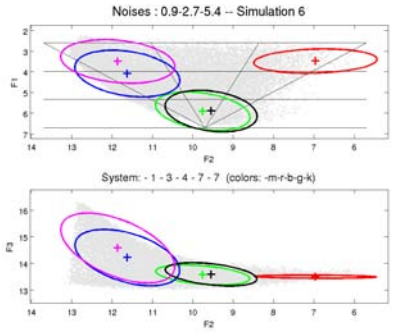
$\sigma_{F1} = 0.8$; $\sigma_{F2} = 2.4$; $\sigma_{F3} = 4.8$

Système	Pourcentage sur 10 simulations
1 3 4 7 7	70 %
1 1 3 7 7	20 %
1 3 4 4 7	10 %



$\sigma_{F1} = 0.9$; $\sigma_{F2} = 2.7$; $\sigma_{F3} = 5.4$

Système	Pourcentage sur 10 simulations
1 3 4 7 7	70 %
1 1 3 7 7	30 %



La Figure 11.14 montre la courbe de dispersion moyenne des systèmes obtenus à chaque niveau de bruit. Le logarithme de la mesure de Lindblom est calculée à partir des distances en Barks dans l'espace F1-F2-F3 pondérées selon les rapports de bruit 1-3-6 (distances sur F2 divisées par 3, sur F3 divisées par 6).

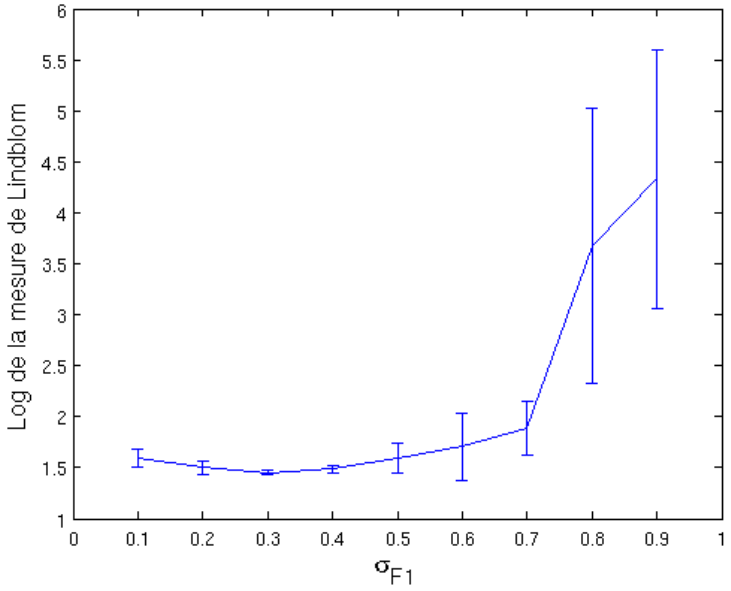


FIGURE 11.14 – Logarithme de la mesure de Lindblom en fonction du bruit sur F1 pour des rapports 1-3-6. Moyennes et écarts-types sur 10 simulations indépendantes.

À partir de la Figure 11.14 et des 9 tables ci-dessus, on extrait les niveaux de bruit « convenables », permettant l'émergence de systèmes dont les éléments sont correctement distinguables (ici jusqu'à $\sigma_{F1} = 0.7$). La Figure 11.15 montre les pourcentages de ces systèmes sur l'ensemble des niveaux de bruit, affichés sur tout l'intervalle de σ_{F1}. Pour $\sigma_{F1} \in \{0.1, \ldots, 0.7\}$ (niveaux de bruit convenables), le système le plus dispersé /1,3,4,6,7/ est majoritaire. Les autres, moins dispersés, sont plus fréquents aux bruits faibles (car suffisants pour une communication correcte) et très forts (car la communication est saturée, empêchant la dispersion).

Nous observons donc que le système majoritaire dans les langues du monde, /i, u, e, o, a/ l'est également dans nos simulations à des niveaux de bruit convenables (système /1, 3, 4, 6, 7/ dans les tables ci-dessus). Ceci s'explique d'une part par le fort bruit sur F3 qui empêche d'utiliser cette dimension pour distinguer des éléments, mais surtout par le rapport de bruit 1 :3 appliqué au dimensions F1 et F2. Celui-ci permet de mieux utiliser F1 et ainsi favoriser les systèmes « de la forme d'un V » dans le triangle vocalique. C'est d'ailleurs précisément pour cette raison que Schwartz et collab. (1997b) ont introduit ce rapport 3 :1 entre les poids des dimension F1 et F2, qu'ils justifient par des considérations sur le « masquage auditif » conduisant à une réduction de la représentation auditive des hautes fréquences (comme F2) « masquées » dans le système auditif par les basses fréquences

(comme F1).

Nous remarquons également et comme précédemment une légère tendance pour les voyelles avant lorsque les systèmes ne sont pas symétriques, et l'expliquons par les mêmes raisons de fortes non-linéarités dans la fonction de transformation articulatori-auditive.

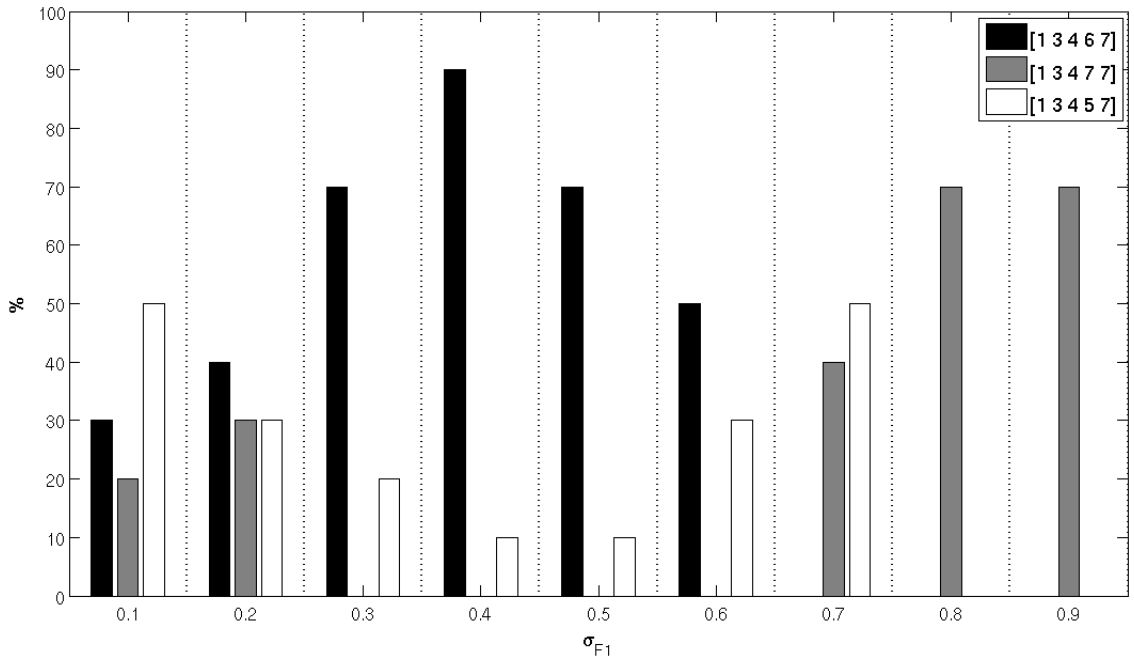


FIGURE 11.15 – Pourcentage de différents systèmes de 5 voyelles obtenus avec des rapports de bruit 1-3-6.

11.6 Conclusion

Ce chapitre a montré l'efficacité du comportement sensori-moteur de production à faire émerger, dans notre paradigme d'interaction par jeux déictiques, des systèmes de voyelles en accord avec les systèmes majoritaires des langues du monde à certains niveaux de bruit dont l'interprétation est compatible avec les travaux de Schwartz et collab. (1997b). Pour cela, nous avons utilisé un modèle réaliste du conduit vocal, VLAM, pour définir la transformation articulatori-auditive réalisée par l'environnement ainsi que la connaissance qu'en ont les agents.

Nous avons exposé les résultats d'une série de simulations pour différents niveaux de bruit de communication et différentes tailles de systèmes vocaliques, et observé que le comportement sensori-moteur disperse les éléments du système en fonction de ces contraintes, en utilisant préférentiellement les dimensions qui permettent une bonne distinguabilité. On retrouve de plus les courbes convexes (en U) des mesures de Lindblom observées au chapitre précédent : les systèmes obtenus sont dispersés d'une manière à pouvoir assurer une bonne distinguabilité des éléments, ni trop ni trop peu, selon les conditions de communication.

Les simulations de ce chapitre fournissaient un point de passage obligatoire pour notre projet. Elles nous permettent de transférer nos résultats obtenus au chapitre précédent sur des situations simplifiées, vers des situations plus réalistes, et de les confronter à des données réelles sur les langues du monde. Elles aboutissent à des résultats relativement prévisibles, et parfaitement en phase avec les travaux préalables de Schwartz et collab. (1997b) en ce qui concerne les poids respectifs des formants dans les calculs de distance, et de Berrah (1998); de Boer (2000); Oudeyer (2003) montrant comment des paradigmes d'interaction entre agents peuvent conduire à la simulation de systèmes vocaliques.

Il reste que la nouveauté de ce travail est d'adosser les mécanismes de dispersion perceptive à des principes de communication intégrés dans des connaissances préalables (sur les théories sensori-motrices de la communication parlée), et d'articuler les espaces moteurs et auditifs autour d'un modèle du conduit vocal, VLAM, relativement réaliste. Le véritable défi de ce travail était en réalité de valider la possibilité d'effectuer des simulations malgré les risques d'explosion combinatoire liés à la grande dimensionnalité des modèles en jeu. Nous voyons ici que c'est possible, même s'il a fallu pour cela geler certaines dimensions, et accepter des temps de simulation non négligeables : typiquement de l'ordre de 10 minutes pour une simulation de 150 000 jeux déictiques entre 2 agents avec 5 objets, sur un total de 360 simulations (voir début de la Section 11.5).

On peut évidemment se demander quelles sont les conséquences possibles de nos choix ayant conduit à des réductions de dimension articulatoire. Les trois variables que nous avons utilisées permettent d'explorer les dimensions majeures pour le contrôle des voyelles : positionnement antéro-postérieur (via TB) et haut-bas (via TD) de la langue et dimensionnement de l'aire aux lèvres (via LH). Ainsi, nous avons généré un espace acoustique satisfaisant, ce qui a assuré la qualité de nos simulations et leur bon accord avec les données phonétiques des langues du monde. Les dimensions articulatoires non utilisées, et notamment celle de la mâchoire, permettraient aux agents d'exprimer, pour des objectifs acoustiques partagés par une société d'agents, leurs possibilités de choisir des variantes

individuelles, des idiosynchrasies bien décrites dans la littérature (voir par exemple Ménard et collab. (2008) montrant des variations d'utilisation de la hauteur de la langue à l'intérieur des possibilités offertes pour contrôler le système vocalique de base du français).

Dans ce chapitre, nous n'avons que peu exploité et discuté du rôle des contraintes articulatoires des agents, si ce n'est pour souligner son effet dans l'orientation des prédictions vers l'avant du triangle vocalique. Dans le prochain chapitre, nous allons explorer des espaces moins travaillés dans la littérature et pour lesquels ces connaissances motrices auront un plus grand rôle : l'espace des consonnes plosives.

Chapitre 12

Emergence des systèmes de consonnes plosives

Après avoir validé la capacité de notre modèle à faire émerger des systèmes de 3 et 5 voyelles relativement en accord avec les données des systèmes vocaliques des langues du monde, nous nous intéressons maintenant à l'émergence de systèmes de consonnes plosives. Ces consonnes, que nous avons déjà mentionnées à la Section 2.2 du Chapitre 2, se définissent par une occlusion complète du conduit vocal, de façon à ce que le flux d'air soit totalement coupé durant leur réalisation.

L'objectif de ce chapitre est de montrer que les paramètres retenus pour les voyelles concernant les poids relatifs de chaque dimension formantique permettent également l'émergence de systèmes de plosives globalement cohérents avec les données des langues du monde. Cette nouvelle instanciation ne diffère de celle du chapitre précédent que sur deux points : la génération du dictionnaire VLAM et le système moteur des agents.

Ainsi, ce chapitre reprend toute la structure du précédent en ne détaillant que les différences entre les deux instanciations. Nous commençons donc par exposer rapidement les données sur les tendances des systèmes de plosives des langues du monde, ainsi que certains travaux visant à les expliquer. Puis nous définissons le modèle de transformation articulatori-auditive et le modèle d'agent pour les consonnes plosives, en insistant seulement sur les points qui diffèrent des voyelles du chapitre précédent. Enfin nous présentons nos résultats sur un grand nombre de simulations et les comparons aux régularités des données des langues du monde.

12.1 Données et prédictions des systèmes de consonnes plosives des langues du monde

Les consonnes plosives se définissant par une occlusion complète du conduit vocal, elles se caractérisent articulatoirement uniquement par la position de cette fermeture (haut de la Figure 12.1). On considère généralement 8 lieux d'articulation, des lèvres à la glotte : les bilabiales, les dentales, les (post)-alvéolaires, les palatales, les vélaires, les uvulaires, les

pharyngeales et épiglottales. Du fait de la relation « un à plusieurs » entre le lieu d'occlusion et et les configurations articulatoires, les conséquences acoustiques d'une plosive sont généralement très dépendantes du contexte dans lequel elle est réalisée. Il s'agit généralement d'un contexte vocalique imposant des contraintes de coarticulation. Par exemple, la façon de réaliser un /b/ dans le contexte de la voyelle /i/ est différente dans le contexte de voyelle /u/ : la langue sera plutôt vers l'avant dans le premier cas pour préparer le /i/, plutôt vers l'arrière dans le deuxième pour préparer le /u/.

Le bas de la Figure 12.1 montre les conséquences acoustiques de chaque classe de plosives réalisées en contexte /a/ dans les plans F2-F3 et F1-F2. On observe une transformation articulatori-acoustique fortement non-linéaire. Les plosives « hautes » (de /b/ à /g/, réalisées avec la mâchoire plutôt fermée) correspondent à des valeurs faibles de F1, et se différencient donc principalement sur les dimensions F2 et F3. À l'inverse, les plosives « arrière » (de /ɢ/ à /ʔ/, réalisées mâchoire ouverte) correspondent à des valeurs élevées de F1 et sont relativement regroupées dans le plan F2-F3.

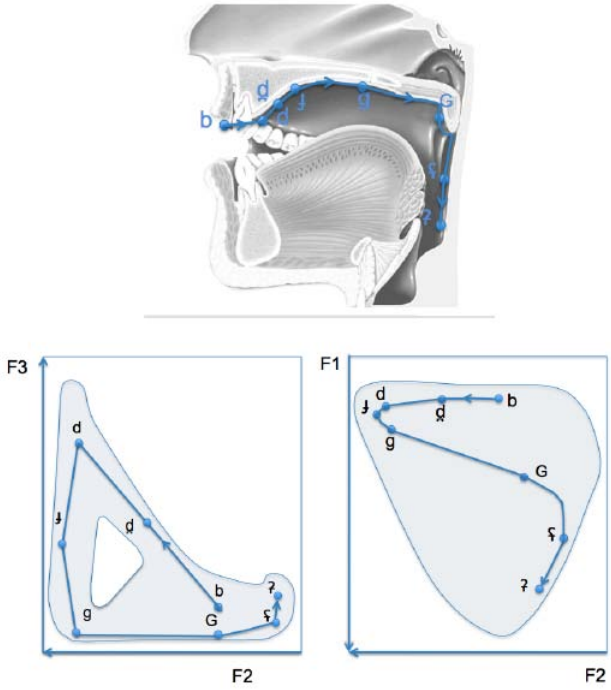


FIGURE 12.1 – Espaces constrictif (en haut) et acoustique (en bas) des consonnes plosives, d'après Schwartz et collab. (2011, en révision). Les formants sont calculés par le modèle VLAM, par une légère ouverture au point de constriction.

La base de données UPSID (Maddieson et Precoda, 1989) atteste d'une forte prédominance des consonnes plosives /b,d,g/ dans les langues du monde, présentes dans la quasi-totalité des systèmes (Boë et collab., 2000). C'est cette tendance globale dont nous souhaitons rendre compte dans nos simulations. Alors que la théorie de la dispersion, associée à certaines pondérations de chaque dimension formantique, permet une bonne

prédiction des systèmes de voyelles, celle des systèmes de consonnes plosives est moins évidente. Certes les trois consonnes privilégiées /b,d,g/ sont bien dispersées dans l'espace F2-F3, occupant les trois sommets de ce « triangle consonantique » de façon similaire à /a,i,u/ dans le triangle vocalique (Figure 12.1). Il reste toutefois que les consonnes arrières (de /ɢ/ à /ʔ/) sont également de très bons candidats en terme de dispersion perceptive, occupant à la fois un sommet du triangle dans F2-F3 et se distinguant de plus par une valeur élevée de F1 (cette dernière dimension jouant de plus un rôle prépondérant dans la dispersion des voyelles). Or, ces consonnes arrières sont peu présentes dans les langues du monde.

Schwartz et collab. (2011, en révision) proposent une explication phylogénétique à cette observation en lien avec la théorie Frame/Content décrite à la Section 4.2.2 du Chapitre 4, selon laquelle les cyclicités mandibulaires induites par les processus d'ingestion sont un précurseur phylogénétique de la structure syllabique universelle du langage. Ces cyclicités, impliquant l'alternance de positions ouvertes et fermées du conduit vocal, fourniraient ainsi un système puissant de modulation vocale alternant voyelles et consonnes, respectivement. Dans ce cadre évolutif, l'émergence des consonnes est donc liée aux fermetures de la mâchoire, privilégiant ainsi les consonnes hautes et expliquant la sous-représentation des consonnes arrières dans les langues du monde.

De ce point de vue, l'émergence des systèmes de consonnes plosives des langues du monde est donc à la fois dirigée par des principes de dispersion perceptive similaire aux systèmes de voyelles et par des principes de contraintes articulatoires provenant du comportement prélangagier d'ingestion (contrainte de mâchoire fermée).

La Figure 12.2, d'après Schwartz et collab. (2011, en révision), illustre ce chemin évolutif. Les auteurs utilisent le modèle VLAM pour explorer progressivement l'espace articulatoire à partir de la position neutre des articulateurs et d'une mâchoire haute, conformément à la théorie Frame/Content. Les consonnes /b,d,g/ apparaissent relativement rapidement dans cette exploration et sont correctement dispersées dans l'espace auditif du plan F2-F3 des plosives réalisées mâchoire fermée. Le reste de l'espace est généré en explorant de plus en plus largement, puis en relâchant la contrainte de mâchoire haute.

12.2 Modèle de transformation articulatori-auditive

Nous utilisons le modèle VLAM décrit au chapitre précédent.

12.2.1 Génération du dictionnaire

Nous retenons les mêmes paramètres articulatoires de VLAM pour la génération du dictionnaire de consonnes plosives : la mâchoire (Jaw), le corps et le dos de la langue (Body et Drsm) et la hauteur des lèvres (LipH). Nous fixons les autres paramètres articulatoires à la valeur neutre (0).

Pour capturer la fonction de transformation articulatori-acoustique, nous réalisons un dictionnaire de 100 000 enregistrements associant les données suivantes :

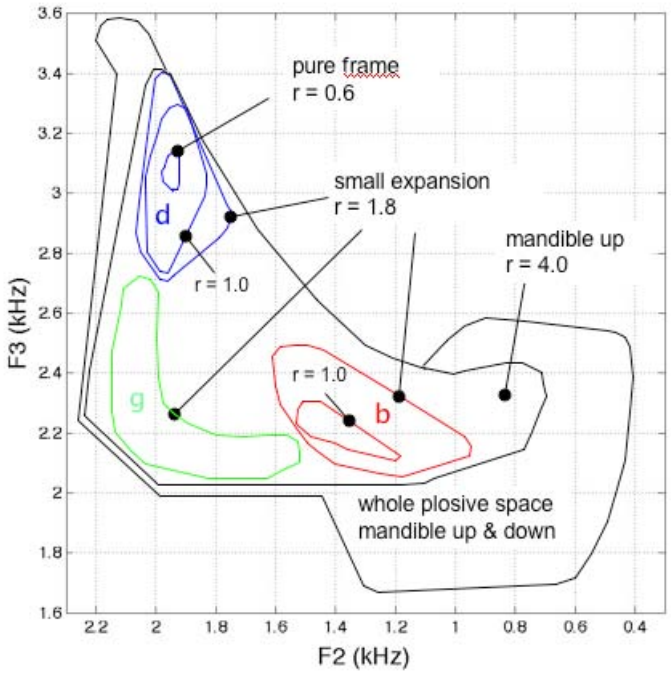


FIGURE 12.2 – Expansion de l'espace des consonnes plosives dans l'espace F2-F3, d'après Schwartz et collab. (2011, en révision). Le paramètre r représente la taille de la zone d'exploration de l'espace articulatoire autour de la position neutre des articulateurs. /b,d,g/ forme un triangle acoustique optimal sous la contrainte d'une mâchoire fermée et d'une exploration limitée.

- les valeurs articulatoires : Jaw, Body, Drsm et LipH ;
- les valeurs constrictives : Xc, Ac et Al ;
- les valeurs formantiques : F1, F2 et F3.

En pratique, on tire de façon aléatoirement uniforme des valeurs dans l'espace articulatoire. Pour chacune, VLAM calcule les valeurs constrictives et formantiques correspondantes. Les valeurs constrictives permettent de sélectionner les consonnes plosives en ne retenant que les configurations suffisamment fermées. Trois cas se présentent.

- Ac et Al sont tous deux supérieurs à 0.15 cm^2, l'enregistrement correspond alors à une voyelle et est rejeté.
- Ac est compris entre 0.05 et 0.15 cm^2 et Al est supérieur à 0.6 cm^2, ou l'inverse. On considère alors l'enregistrement comme étant une plosive et on l'ajoute au dictionnaire. Ces seuils proviennent de Schwartz et collab. (2011, en révision) : 0.05 cm^2 est l'aire minimale de la constriction pour que VLAM puisse calculer les formants correctement, 0.15 cm^2 est l'aire à partir de laquelle on considère un enregistrement comme une voyelle. Le seuil minimal de 0.6 cm^2 permet d'éviter les doubles occlusions (c'est-à-dire à la fois aux lèvres et dans le conduit).
- Soit Ac soit Al a une valeur nulle : le conduit vocal est complétement fermé. Dans

ce cas, nous exécutons un algorithme de recherche dichotomique qui, à partir de la position de mâchoire courante, cherche la position la plus proche qui vérifie les conditions du cas précédent. Si la recherche aboutit, nous calculons les valeurs constrictives et formantiques de la nouvelle configuration et ajoutons l'enregistrement au dictionnaire, sinon nous le rejetons.

Par ce dernier point, nous considérons la mâchoire comme le seul articulateur contrôlé en début de simulation, en accord avec la théorie Frame/Content. Ainsi, une configuration complètement fermée correspond à une plosive si elle s'ouvre lors de la réalisation du cycle de mâchoire pour produire des formants. Ceci permet de ne pas imposer un contrôle fin de l'aire de constriction pour vérifier le deuxième cas. Notons que nous ne spécifions pas ici si le relâchement de l'occlusion a lieu durant une phase d'ouverture ou de fermeture de la mâchoire.

Nous enchaînons alors les tirages jusqu'à accepter 100 000 enregistrements. Les valeurs formantiques sont converties en Barks. L'espace auditif généré par ce dictionnaire est représenté Figure 12.3.

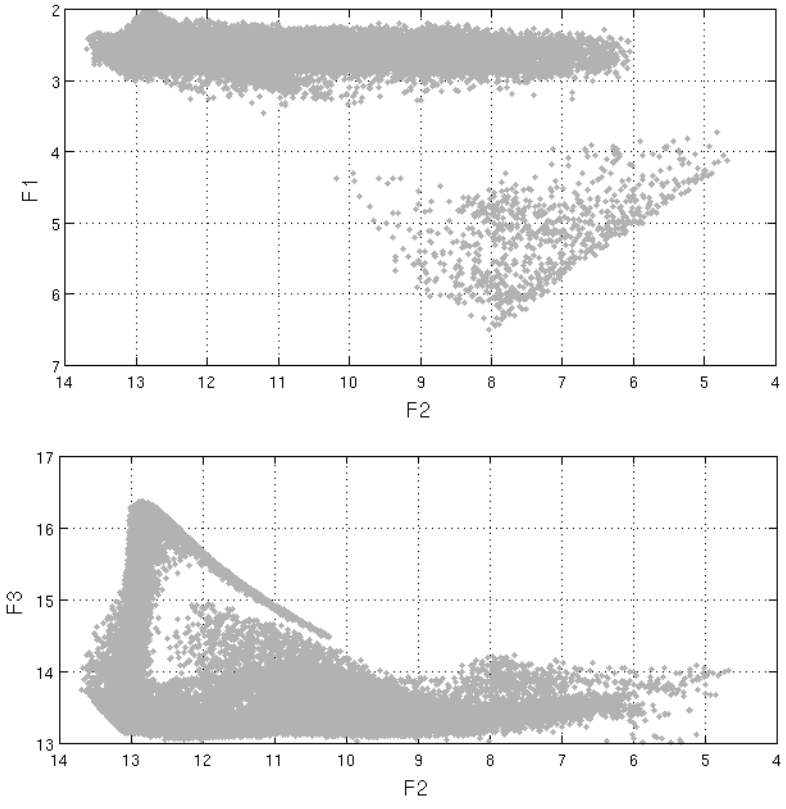


FIGURE 12.3 – Espace auditif issu du dictionnaire de consonnes plosives généré avec VLAM dans les plans F1-F2 et F2-F3.

12.2.2 Espaces et transformation articulatori-auditive

La définition des variables articulatoires et auditives est identique à celle des voyelles (Section 11.2.2 du chapitre précédent).

Les 100 000 enregistrements du dictionnaire de consonnes défini plus haut forment un ensemble d'apprentissage δ_{Cons}, dans lequel les valeurs articulatoires et formantiques sont discrétisées dans les domaines des variables correspondantes.

La distribution représentant la transformation articulatori-auditive réalisée par l'environnement est alors définie par :

$$\begin{array}{rcl} \forall m \in M \ : & & \\ P(S \mid [M = m]\ \delta_{Cons}\ \pi_{Com}) & = & \boldsymbol{H}_{\delta^m_{Cons}}(S), \end{array} \tag{12.1}$$

où δ^m_{Cons} correspond à l'ensemble δ_{Cons} restreint aux enregistrements pour lesquels $M = m$. Le choix d'une loi histogramme répond aux mêmes justifications que pour les voyelles.

Toujours de façon similaire au chapitre précédent, des bruits gaussiens indépendants sur $F1$, $F2$ et $F3$, d'écarts-types respectifs σ_{F1}, σ_{F2}, σ_{F3}, sont ajoutés après le tirage du geste moteur m de l'agent locuteur selon la distribution $P(S \mid [M = m]\ \delta_{Cons}\ \pi_{Com})$ dans l'algorithme de simulation. Dans la présentation des résultats de nos simulations, les stimuli sont enregistrés avant ajout du bruit de l'environnement, à la manière d'un micro qui serait placé au plus près de la sortie du conduit vocal de l'agent locuteur, avant que le signal soit bruité puis transmis à l'agent auditeur (les paramètres σ_{F1}, σ_{F2}, σ_{F3} ont donc un effet sur les stimuli reçus par les agents en situation d'auditeur).

12.3 Modèle d'agent

12.3.1 Ensemble d'apprentissage

L'ensemble d'apprentissage d'un agent a au temps t de la simulation se note $\delta_a(t)$ et est défini de la même façon que dans le chapitre précédent. Nous notons toujours $\delta^M_a(t)$ (respectivement $\delta^S_a(t)$) l'ensemble des $N_{App} = 200$ derniers éléments de $\delta_a(t)$ enregistrés en statut de locuteur (respectivement auditeur).

12.3.2 Système moteur

Avec la génération du dictionnaire, le système moteur des agents est la seule différence existante entre le modèle d'émergence de voyelles du chapitre précédent et celui de consonnes plosives défini ici. Alors que nous considérions la mâchoire (variable J) « bloquée » en configuration neutre pour les voyelles, nous la libérons maintenant au même titre que les trois autres variables articulatoires (TB, TD, LH). Le système moteur est donc

maintenant défini par :

$$
\begin{aligned}
&P(M \mid O_S\ \delta_a(t)\ \pi_{Ag}) \\
&\quad = \quad P(J\ TB\ TD\ LH \mid O_S\ \delta_a(t)\ \pi_{Ag}) \\
&\quad = \quad P(J \mid O_S\ \delta_a(t)\ \pi_{Ag}) P(TB \mid O_S\ \delta_a(t)\ \pi_{Ag}) P(TD \mid O_S\ \delta_a(t)\ \pi_{Ag}) P(LH \mid O_S\ \delta_a(t)\ \pi_{Ag})
\end{aligned}
$$

où

$$
\begin{aligned}
\forall X \in \{J, TB, TD, LH\}, o_i \in \mathcal{D}_{O_S}\ : \\
P(X \mid [O_S = o_i]\ \delta_a(t)\ \pi_{Ag}) \quad = \quad \boldsymbol{L}_{\delta_a^{M,o_i}(t)}(X),
\end{aligned}
$$

et $\delta_a^{M,o_i}(t)$ correspond à l'ensemble $\delta_a^M(t)$ restreint aux enregistrements pour lesquels $[O_S = o_i]$.

Les agents sont ainsi maintenant capables d'explorer tout leur espace articulatoire. La variable J est en effet d'une importance particulière pour l'émergence des consonnes plosives, généralement effectuées avec la mâchoire dans une position extrême (soit très fermée lorsque l'occlusion a lieu sur le haut du conduit vocal (lèvres, dents, palais), soit très ouverte si elle a lieu à l'arrière (gorge)). Nous reviendrons sur ce point lors de l'exposé des principes de simulation.

12.3.3 Lien sensori-moteur

Le lien sensori-moteur est identique pour tous les agents de la société et est appris préalablement à la simulation comme une loi de succession de Laplace à partir du dictionnaire VLAM δ_{Cons} :

$$
\begin{aligned}
\forall m \in M\ : \\
P(S \mid [M = m]\ \delta_{Cons}\ \pi_{Ag}) \quad = \quad \boldsymbol{L}_{\delta_{Cons}^m}(F1\ F2\ F3),
\end{aligned}
$$

où δ_{Cons}^m correspond à l'ensemble δ_{Cons} restreint aux enregistrement pour lesquels $M = m$.

12.3.4 Système auditif

Le système auditif est identique à celui du chapitre précédent.

12.4 Simulations

Dans le but de réduire le temps de simulation, nous nous concentrons sur des sociétés de 2 agents. L'environnement comporte 3 objets, conduisant à l'émergence de systèmes de 3 consonnes. Le temps de simulation est de 150 000 jeux déictiques. Toutes les simulations considérées concernent des sociétés d'agents en comportement sensori-moteur.

$$
\begin{aligned}
N_A &= 2, \\
N_O &= 3, \\
N_{JD} &= 150\ 000.
\end{aligned}
$$

12.4.1 Paramètres variés dans les simulations

La variation des paramètres σ_{F1}, σ_{F2} et σ_{F3} correspond aux poids relatifs des dimensions formantiques retenus pour les voyelles. Ainsi, F1 a 3 fois plus de poids que F2, qui a 2 fois plus de poids que F3, soient des bruits relatifs de rapports 1-3-6 (selon Schwartz et collab. (1997b)) :

$$\sigma_{F2} = 3\sigma_{F1},$$
$$\sigma_{F3} = 6\sigma_{F1}.$$

Nous faisons varier σ_{F1} de 0.1 à 0.9 par pas de 0.1.

Nous souhaitons vérifier par la simulation les prédictions de Schwartz et collab. (2011, en révision) selon lesquelles /b,d,g/ est le système de plosives optimal sous la condition d'une contrainte de mâchoire fermée. Pour cela nous distinguons deux conditions :

- la distribution du système moteur sur la variable de mâchoire, $P(J \mid O_S\ \delta_a(t)\ \pi_{Ag})$ est libre, telle que défini à la Section 12.3.2 ;
- la distribution contraint la mâchoire en position fermée :

$$P(J \mid O_S\ \delta_a(t)\ \pi_{Ag}) = P(J \mid \pi_{Ag}) = \boldsymbol{\delta}_3(J)$$

Dans le premier cas, que nous appelons la condition « mâchoire libre », les agents sont capables d'explorer tout leur espace articulatoire. Dans le second, que nous appelons la condition « mâchoire fermée », la variable de mâchoire est bloquée à la valeur $J = 3$, correspondant à des valeurs du paramètre VLAM Jaw supérieures à 3 (mâchoire haute) d'après l'Équation 11.2.

12.4.2 Évaluation

Nous souhaitons vérifier la cohérence globale des systèmes de 3 consonnes plosives émergeant de nos simulations avec les données issues des langues du monde. Comme dans le chapitre précédent, nous optons pour une méthode automatisée dénuée de tout caractère subjectif, facilitée ici par la bonne caractérisation des consonnes plosives en terme de lieu de constriction.

12.4.2.1 Classification des systèmes de consonnes issus de nos simulations

Nous classons les consonnes plosives selon les 8 classes usuelles (Figure 12.1), à partir des valeurs de constriction Xc, Ac et Al issues du dictionnaire VLAM. Nous utilisons pour cela la correspondance de Schwartz et collab. (2011, en révision) illustrée Figure 12.4.

Notons que la taille de ces régions est très variables d'un cas à l'autre : par exemple, la zone 4 (palatales) est très étroite par rapport à la zone 5 (vélaires). Ce point est largement commenté et justifié par Schwartz et collab. (2011, en révision) à partir de données phonétiques, nous ne reprendrons pas cette discussion. Notons cependant que la dispersion correspondante dans l'espace acoustique est beaucoup plus régulière, comme nous l'atteste les ellipses 4 et 5 de la Figure 12.5.

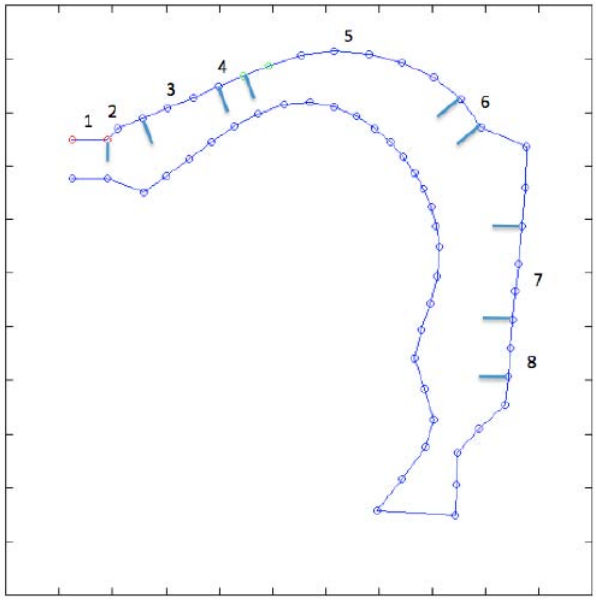


FIGURE 12.4 – Correspondance entre les sections VLAM du lieu de constriction (Xc) et les 8 classes usuelles de plosives. 1 : bilabiales ; 2 : dentales ; 3 : (post)-alvéolaires ; 4 : palatales ; 5 : vélaires ; 6 : uvulaires ; 7 : pharyngeales ; 8 : épiglottales. Les sections qui ne sont associées à aucune classe ne peuvent correspondre à un lieu de constriction dans la morphologie de VLAM.

Pour chaque simulation effectuée, le processus de classification extrait sur les 3000 derniers jeux déictiques la plosive la plus fréquente pour chaque objet. La variabilité au sein d'une même classe d'objet pourra être visualisée dans l'espace auditif sur des exemples de simulation.

Un système particulier correspond donc à un vecteur de trois chiffres entre 1 (bilabiales) et 8 (epiglottales). Par exemple, le système /b,d,g/ le plus fréquent dans les langues du monde correspond au vecteur /1,3,5/.

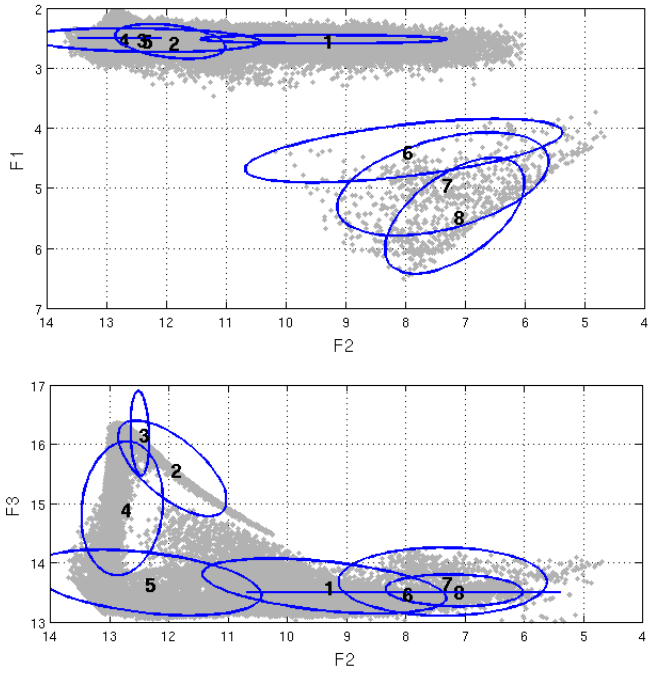


FIGURE 12.5 – Conséquences auditives des 8 classes de consonnes plosives dans les plans F1-F2 et F2-F3. Les ellipses de dispersion (1.5 écarts-types) sont calculées à partir du dictionnaire VLAM et identifiées par le chiffre en leur centre d'après la correspondance de la Figure 12.4.

12.5 Résultats

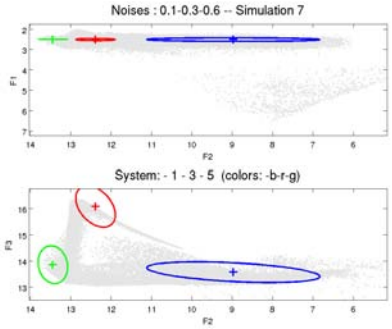
Nous exposons d'abord nos résultats pour la condition mâchoire libre, puis pour la condition mâchoire fermée.

12.5.1 Mâchoire libre

Nous réalisons des séries de 10 simulations indépendantes en condition mâchoire libre pour chaque valeur de σ_{F1} allant de 0.1 à 0.9 par pas de 0.1. Les 9 tables qui suivent exposent nos résultats, avec les mêmes conventions qu'au chapitre précédent.

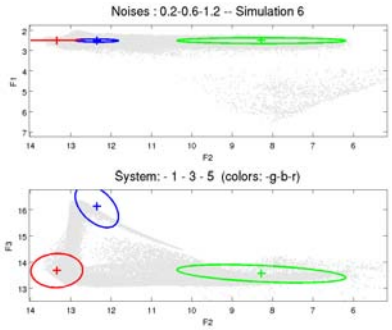
$\sigma_{F1} = 0.1$; $\sigma_{F2} = 0.3$; $\sigma_{F3} = 0.6$

Système	Pourcentage sur 10 simulations
1 3 5	40 %
1 2 5	20 %
1 1 3	20 %
1 1 2	10 %
1 3 4	10 %



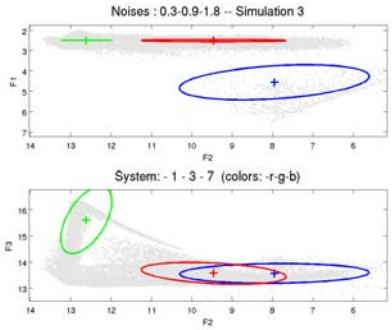
$\sigma_{F1} = 0.2$; $\sigma_{F2} = 0.6$; $\sigma_{F3} = 1.2$

Système	Pourcentage sur 10 simulations
1 3 5	60 %
1 1 3	20 %
1 3 7	10 %
1 2 5	10 %



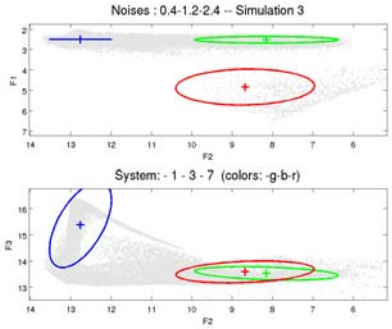
$\sigma_{F1} = 0.3$; $\sigma_{F2} = 0.9$; $\sigma_{F3} = 1.8$

Système	Pourcentage sur 10 simulations
1 5 7	30 %
1 3 7	30 %
1 2 5	20 %
1 1 3	10 %
1 3 4	10 %



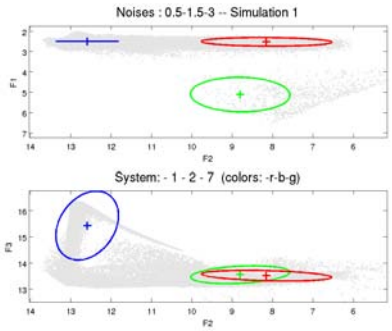
$\sigma_{F1} = 0.4$; $\sigma_{F2} = 1.2$; $\sigma_{F3} = 2.4$

Système	Pourcentage sur 10 simulations
1 3 7	40 %
1 1 5	20 %
1 4 7	20 %
1 2 7	10 %
1 3 5	10 %



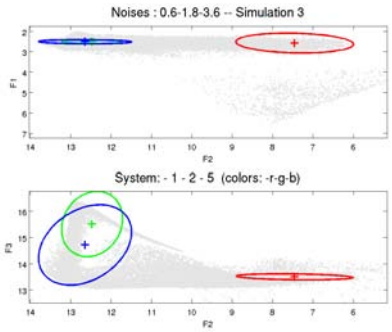
$\sigma_{F1} = 0.5$; $\sigma_{F2} = 1.5$; $\sigma_{F3} = 3$

Système	Pourcentage sur 10 simulations
1 2 7	20 %
1 2 5	20 %
1 4 7	10 %
1 3 3	10 %
1 2 3	10 %
1 5 7	10 %
1 3 7	10 %
1 3 5	10 %



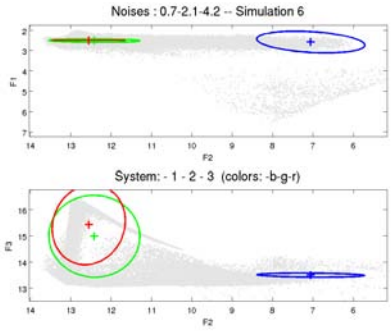
$\sigma_{F1} = 0.6$; $\sigma_{F2} = 1.8$; $\sigma_{F3} = 3.6$

Système	Pourcentage sur 10 simulations
1 2 5	30 %
1 2 3	30 %
1 3 7	20 %
1 3 5	10 %
1 2 7	10 %



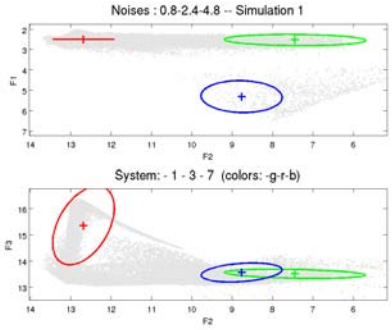
$\sigma_{F1} = 0.7$; $\sigma_{F2} = 2.1$; $\sigma_{F3} = 4.2$

Système	Pourcentage sur 10 simulations
1 2 3	30 %
1 2 4	20 %
1 3 4	10 %
1 1 3	10 %
1 2 5	10 %
1 5 7	10 %
1 2 2	10 %



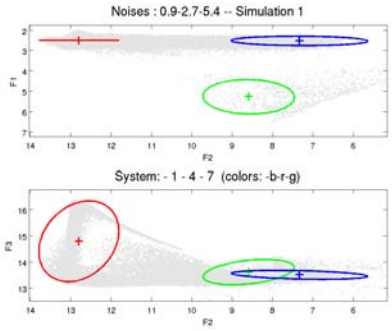
$\sigma_{F1} = 0.8$; $\sigma_{F2} = 2.4$; $\sigma_{F3} = 4.8$

Système	Pourcentage sur 10 simulations
1 3 7	40 %
1 5 7	20 %
1 2 4	10 %
1 2 5	10 %
1 3 5	10 %
1 4 7	10 %



$\sigma_{F1} = 0.9$; $\sigma_{F2} = 2.7$; $\sigma_{F3} = 5.4$

Système	Pourcentage sur 10 simulations
1 4 7	30 %
1 3 7	20 %
1 5 7	20 %
1 3 5	10 %
1 2 7	10 %
1 3 4	10 %



La Figure 12.6 montre la courbe de dispersion moyenne des systèmes obtenus à chaque niveau de bruit en condition mâchoire libre. Le logarithme de la mesure de Lindblom est calculé à partir des distances en Barks dans l'espace F1-F2-F3, pondérées selon les rapports de bruit 1-3-6 (distances sur F2 divisées par 3, sur F3 divisées par 6).

Contrairement aux courbes de dispersion en comportement sensori-moteur présentées précédemment, on observe ici une décroissance pour les bruits forts, que nous expliquons par l'aspect discontinu de l'espace auditif des plosives. Les consonnes hautes et arrières se trouvent en effet dans deux régions séparées du plan F1-F2 (Figure 12.5). Les mesures de dispersion entre les systèmes avec ou sans consonnes arrières sont très différentes et peuvent expliquer cet effet.

À partir de la Figure 12.6 et des 9 tables ci-dessus, on extrait les niveaux de bruits « convenables », permettant l'émergence de systèmes dont les éléments sont correctement distinguables (ici jusqu'à $\sigma_{F1} = 0.4$). La Figure 12.7 montre les pourcentages de ces systèmes sur l'ensemble des niveaux de bruit.

De façon générale, on observe une certaine variabilité des systèmes émergeant de ces simulations. Plus précisément, le système majoritaire à bruits faibles ($\sigma_{F1} \in \{0.1, 0.2\}$) est /1,3,5/, correspondant à /b,d,g/. À bruit moyen ($\sigma_{F1} \in \{0.3, 0.4\}$) et fort ($\sigma_{F1} \in \{0.8, 0.9\}$), il s'agit plutôt de /1,3,7/, correspondant à /b,d,ʕ/.

L'observation principale de cette condition mâchoire ouverte est la forte présence de plosives pharyngeales /ʕ/, qui s'explique naturellement par leur bonne distinguabilité par rapport aux consonnes hautes et est en accord avec les prédictions de Schwartz et collab.

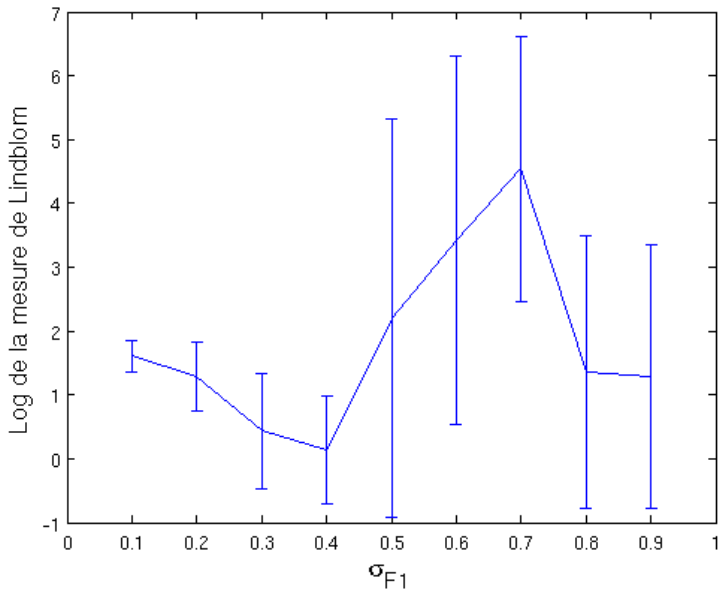


FIGURE 12.6 – Logarithme de la mesure de Lindblom en fonction du bruit sur F1 pour des systèmes de 3 consonnes en condition mâchoire libre avec des rapports de bruit 1-3-6. Moyennes et écarts-types sur 10 simulations indépendantes.

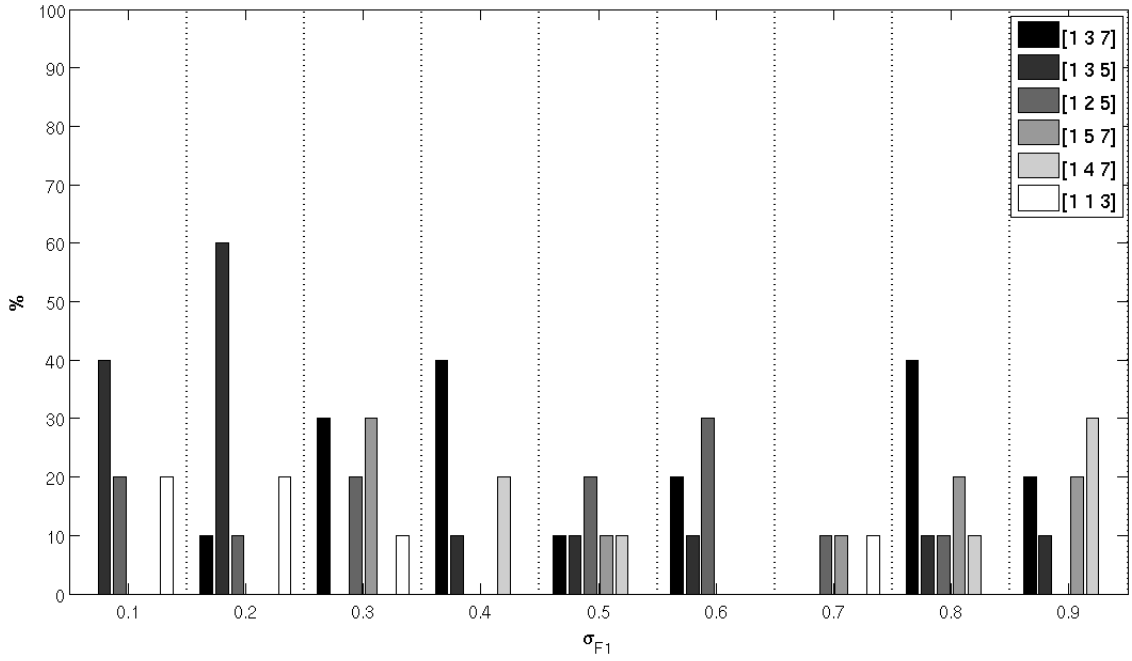


FIGURE 12.7 – Pourcentage de différents systèmes de 3 consonnes obtenus avec des rapports de bruit 1-3-6.

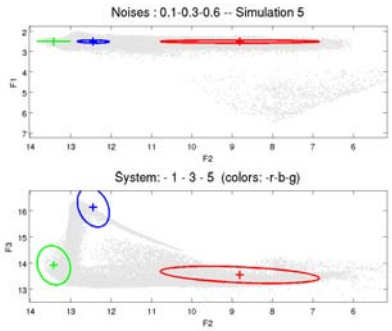
(2011, en révision).

12.5.2 Mâchoire fermée

Nous réalisons des séries de 10 simulations indépendantes en condition mâchoire fermée pour chaque valeur de σ_{F1} allant de 0.1 à 0.9 par pas de 0.1. Les 9 tables qui suivent exposent nos résultats, avec les mêmes conventions qu'au chapitre précédent.

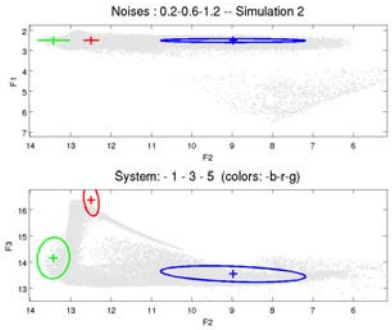
$\sigma_{F1} = 0.1$; $\sigma_{F2} = 0.3$; $\sigma_{F3} = 0.6$

Système	Pourcentage sur 10 simulations
1 3 5	70 %
1 1 2	10 %
1 3 4	10 %
1 1 3	10 %



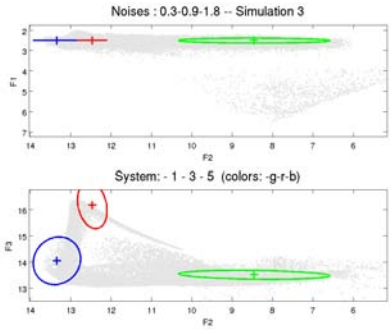
$\sigma_{F1} = 0.2$; $\sigma_{F2} = 0.6$; $\sigma_{F3} = 1.2$

Système	Pourcentage sur 10 simulations
1 3 5	60 %
1 1 3	30 %
1 1 2	10 %



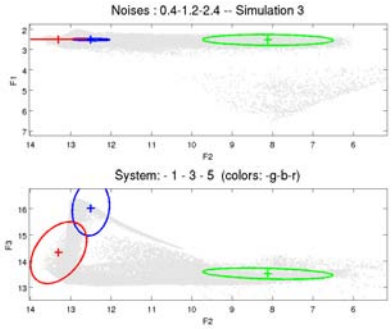
$\sigma_{F1} = 0.3$; $\sigma_{F2} = 0.9$; $\sigma_{F3} = 1.8$

Système	Pourcentage sur 10 simulations
1 3 5	50 %
1 1 2	20 %
1 1 5	10 %
1 1 4	10 %
1 1 3	10 %



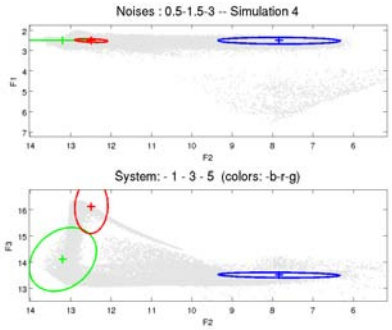
$\sigma_{F1} = 0.4$; $\sigma_{F2} = 1.2$; $\sigma_{F3} = 2.4$

Système	Pourcentage sur 10 simulations
1 3 5	60 %
1 1 3	20 %
1 1 2	10 %
1 1 5	10 %



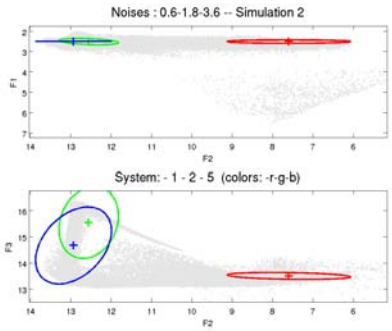
$\sigma_{F1} = 0.5$; $\sigma_{F2} = 1.5$; $\sigma_{F3} = 3$

Système	Pourcentage sur 10 simulations
1 3 5	40 %
1 1 2	20 %
1 2 5	20 %
1 1 3	10 %
1 4 5	10 %



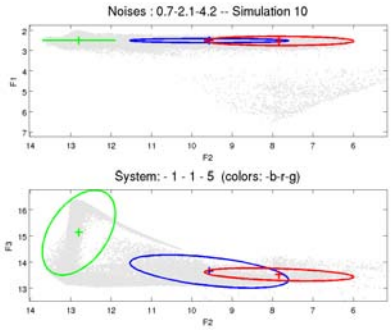
$\sigma_{F1} = 0.6$; $\sigma_{F2} = 1.8$; $\sigma_{F3} = 3.6$

Système	Pourcentage sur 10 simulations
1 2 5	30 %
1 3 5	20 %
1 1 5	20 %
1 1 3	10 %
1 2 4	10 %
1 4 5	10 %



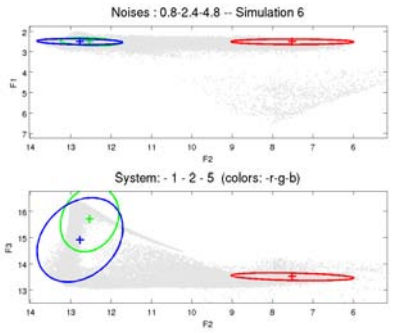
$\sigma_{F1} = 0.7$; $\sigma_{F2} = 2.1$; $\sigma_{F3} = 4.2$

Système	Pourcentage sur 10 simulations
1 5 5	20 %
1 1 5	20 %
1 2 5	20 %
1 3 4	10 %
1 1 3	10 %
1 2 3	10 %
1 3 3	10 %



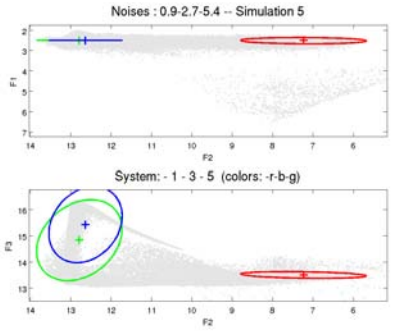
$\sigma_{F1} = 0.8$; $\sigma_{F2} = 2.4$; $\sigma_{F3} = 4.8$

Système	Pourcentage sur 10 simulations
1 2 5	30 %
1 2 4	20 %
1 5 5	10 %
1 3 4	10 %
1 4 4	10 %
1 3 5	10 %
1 4 5	10 %



$\sigma_{F1} = 0.9$; $\sigma_{F2} = 2.7$; $\sigma_{F3} = 5.4$

Système	Pourcentage sur 10 simulations
1 3 5	40 %
1 5 5	20 %
1 4 5	20 %
1 2 2	10 %
1 2 5	10 %



La Figure 12.8 montre la courbe de dispersion moyenne des systèmes obtenus à chaque niveau de bruit. Le logarithme de la mesure de Lindblom est calculé à partir des distances en Barks dans l'espace F1-F2-F3, pondérées selon les rapports de bruit 1-3-6 (distances sur F2 divisées par 3, sur F3 divisées par 6).

À partir de la Figure 12.8 et des 9 tables ci-dessus, on extrait les niveaux de bruits « convenables », permettant l'émergence de systèmes dont les éléments sont correctement dispersés (ici jusqu'à $\sigma_{F1} = 0.4$). La Figure 12.9 montre les pourcentages de ces systèmes majoritaires pour l'ensemble des niveaux de bruits. On observe cette fois une nette prédominance du système /1,3,5/, en accord avec les prédictions de Schwartz et collab. (2011, en révision) selon lesquels /b,d,g/ est le système optimal sous la contrainte de mâchoire fermée. On remarque également fréquemment des systèmes comportant deux /b/, dont un exemple est exposé Figure 12.10. On observe que les deux /b/ correspondent à deux éléments distinguables du système. Comme nous l'avons remarqué à la Section 12.1, les conséquences auditives d'une plosive dépendent en effet de l'ensemble de la configuration articulatoire dans laquelle elle est réalisée. Il existe ainsi différents gestes moteurs correspondant à une occlusion labiale, dont les conséquences auditives sont distinguables. Il est donc normal que notre modèle laisse émerger des systèmes comme celui de la Figure 12.10.

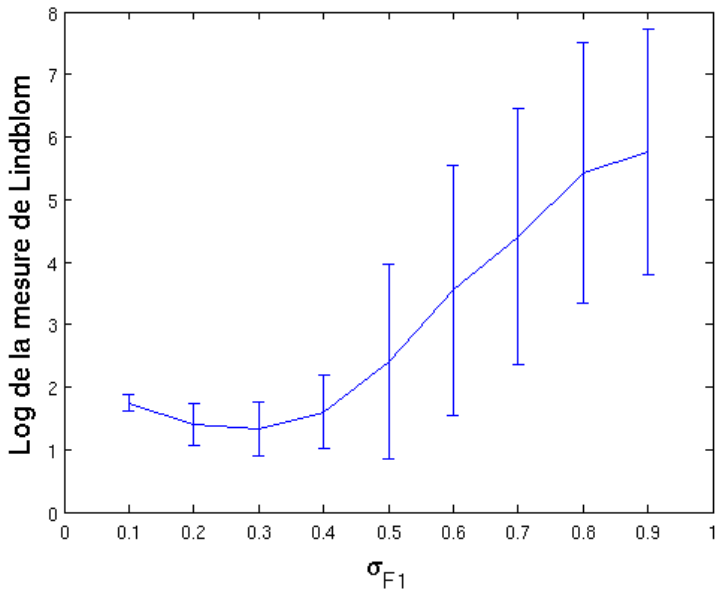


FIGURE 12.8 – Logarithme de la mesure de Lindblom en fonction du bruit sur F1 pour des systèmes de 3 consonnes avec des rapports de bruit 1-3-6. Moyennes et écarts-types sur 10 simulations indépendantes.

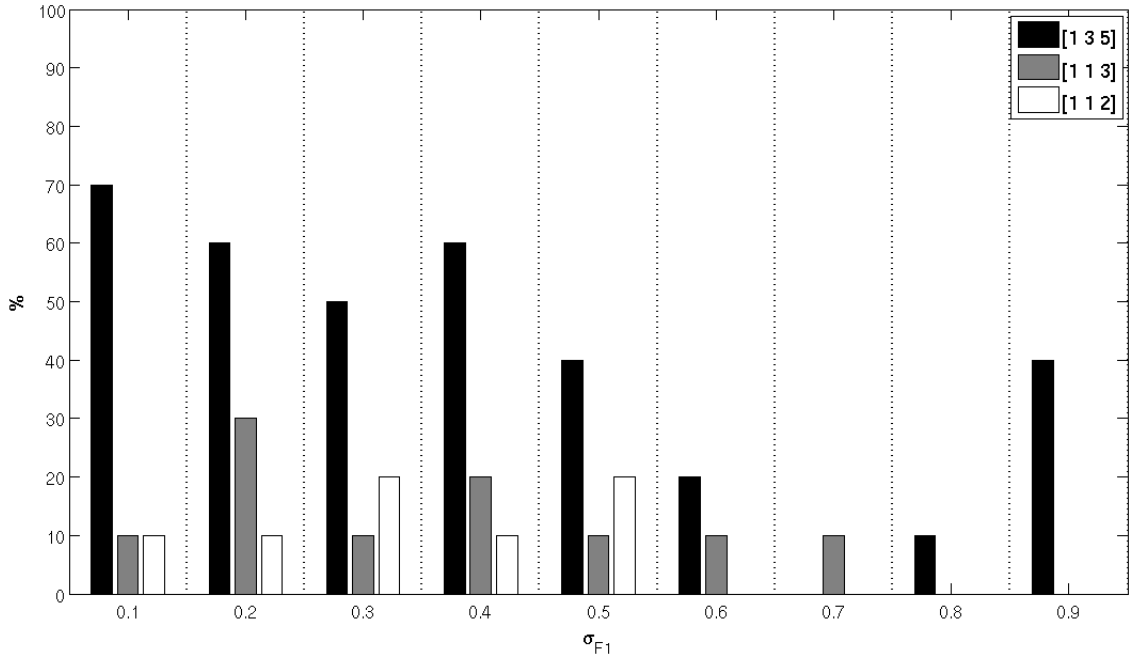


FIGURE 12.9 – Pourcentage de différents systèmes de 3 consonnes obtenus avec des rapports de bruit 1-3-6.

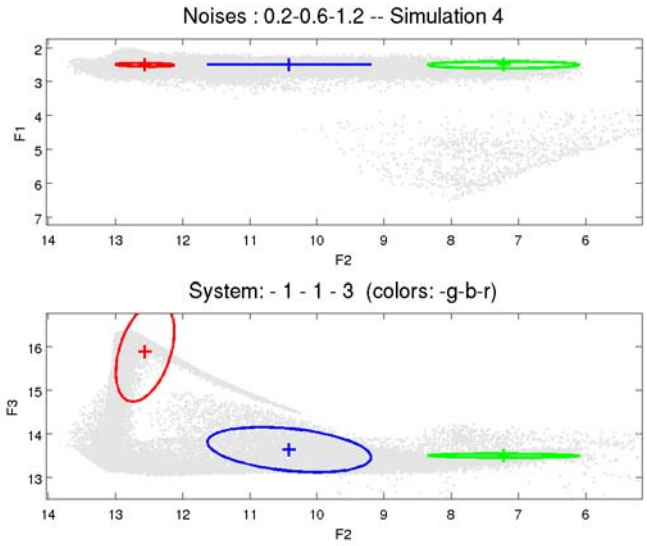


FIGURE 12.10 – Exemple de système /b,b,d/.

12.6 Conclusion

Ce chapitre nous fournit ainsi un jeu de simulations de systèmes de consonnes plosives émergeant du paradigme d'interaction par jeux déictiques entre agents dotés d'un comportement sensori-moteur de production.

Si les simulations du chapitre précédent venaient confirmer une série de résultats précédents obtenus notamment par Berrah (1998), de Boer (2000) et Oudeyer (2003) sur les simulations de systèmes vocaliques à base de sociétés d'agents en interaction, les résultats de ce chapitre constituent en quelque sorte une première. Nous montrons ici pour la première fois comment des principes implicites de dispersion perceptive, dérivés de mécanismes de communication entre agents, conduisent à des systèmes réalistes, et comment des contraintes additionnelles sur la position de la mâchoire, émanant de la théorie Frame/Content, s'avèrent nécessaires pour parvenir à des simulations réalistes du système « star » /b,d,g/. Nous avons été guidés dans ces simulations par l'étude de Schwartz et collab. (2011, en révision), fournissant en quelque sorte le « portrait de phase » dans lequel nos simulations devaient venir s'inscrire (tout comme les simulations de Lindblom (1984) ou de Schwartz et collab. (1997b) fournissaient le portrait de phase pour les voyelles). Nous utilisons à dessein ce terme de « portrait de phase », en reprenant une analogie classique entre d'une part les approches globales « à la Lindblom » et dérivés, et les approches locales « à la Berrah » ou successeurs ; et d'autre part la relation entre thermodynamique et mécanique statistique, que nous avons introduite dans notre Introduction générale. Les simulations multi-agents telles que celles que nous présentons ici, fournissant en quelque sorte l'instanciation par la physique statistique des prédictions et des lois globales de la thermodynamique.

Le point fort de nos simulations est que ces prédictions, relativement en accord avec

les systèmes majoritaires des langues du monde et les prédictions de Schwartz et collab. (2011, en révision), ont été obtenues à partir d'un modèle qui ne diffère de celui du chapitre précédent que par la génération du dictionnaire et le système moteur des agents. Par contre, nous avons pu obtenir des simulations satisfaisantes de systèmes consonantiques avec exactement les mêmes paramètres de simulation, notamment en terme de rapport entre formants F1-F2-F3 dans les calculs de distance, et de zones de bruit adéquates.

Évidemment, beaucoup de travail reste à faire sur les systèmes consonantiques. Nous avons utilisé ici 4 paramètres articulatoires, abandonnant notamment le paramètre de pointe de la langue (Tongue tip, TT) qui est pourtant le plus adapté pour produire des dentales et alvéolaires : nous avons utilisé pour cela la versatilité du modèle articulatoire, qui peut coller la langue sur l'avant du palais près des dents avec le paramètre de contrôle de la masse de la langue, TD. Là encore, cela n'impacte guère la validité de nos simulations, puisque les générations acoustiques sont tout à fait réalistes (l'espace acoustique généré par nos 4 paramètres, Figure 12.3, correspond bien à l'espace réel sur les 7 paramètres articulatoires, voir Figure 12.1). Mais les formes articulatoires ne sont pas parfaitement adéquates, et d'autres simulations, portant sur un jeu plus étendu de paramètres, seront nécessaires pour fortifier nos prédictions.

Il conviendra également à l'avenir d'élargir le champ des études consonantiques, en s'attaquant notamment au voisement des plosives (on sait que les systèmes sonores utilisent plus souvent les plosives non voisées, comme /p,t,k/ que voisées comme /b,d,g/, et des contraintes aérodynamiques pourraient être introduites pour tenter de simuler ce fait, voir Boë et collab. (2000), puis en allant vers d'autres sous-systèmes sonores, nasales ou fricatives notamment.

Reste que nous disposons maintenant d'un modèle qui semble suffisamment puissant pour traiter simultanément, et dans le même jeu de mécanismes, des consonnes plosives et des voyelles.

Cette homogénéité des modèles d'émergence de voyelles et de consonnes nous permet d'étudier leur couplage dans le chapitre suivant, en visant l'émergence de systèmes de syllabes.

Chapitre 13

Emergence de la syllabe

Ce chapitre clôt cette partie de simulations. Les deux chapitres précédents ont montré la cohérence globale de nos résultats de simulations avec les systèmes de voyelles et de consonnes des langues du monde, sous certaines contraintes articulatori-auditives. Nous disposons ainsi d'un modèle homogène d'émergence des systèmes de voyelles et de consonnes, que nous couplons dans ce chapitre vers un modèle d'émergence de syllabes.

Nous commençons par exposer quelques données et prédictions de la littérature. Puis nous définissons les espaces et la transformation articulatori-auditive. Nous nous intéressons ensuite à la définition du modèle d'agent en justifiant de lourdes simplifications par des raisons de complexité calculatoire. Nous exposons enfin nos simulations et résultats et étudions leur cohérence globale avec les données.

13.1 Données et prédictions des systèmes de syllabes

Une simple règle combinatoire sur les systèmes majoritaires de voyelles et de consonnes des langues du monde prédit des systèmes syllabiques composés de /ba/ /bi/ /bu/ /da/ /di/ /du/ /ga/ /gi/ /gu/. Bien que cette combinatoire soit vérifiée d'un point de vue phonémique, chaque langage comportant autant de syllabes qu'il est possible d'en composer par le produit de ses systèmes de voyelles et de consonnes, la simulation des systèmes de syllabes pose deux types de problèmes.

D'abord, l'émergence de la combinatoire elle même est un problème majeur, que nous n'avons malheureusement pu aborder en détail dans cette thèse : nous y reviendrons dans le chapitre suivant.

Ensuite, la combinatoire semble en fait en partie contrainte, et cela sous deux formes liées mais bien distinctes. La première contrainte porte sur le choix des objets dans les systèmes phonologiques. C'est le phénomène de cooccurrence, qui suggère que certaines associations de plosives et de syllabes soient plus probables que d'autres. Ces cooccurrences sont mises en avant par les tenants de la théorie Frame/Content, qui propose que la précédence du contrôle de la mâchoire sur les autres articulateurs pourrait orienter la morphogénèse des systèmes de syllabes vers ces associations préférentielles.

La seconde contrainte porte elle sur les propriétés phonétiques des syllabes au sein des systèmes sonores : les réalisations phonétiques des plosives et des voyelles sont jusqu'à un certain point corrélées. C'est le phénomène de coarticulation, qui peut être résumé par une influence de la configuration motrice d'une voyelle sur celle d'une consonne adjacente. Ainsi, le même phonème /b/ peut mener à des réalisations phonétiques différentes selon le contexte vocalique : /ba/ vs /bu/ par exemple.

13.1.1 Cooccurrences

La théorie Frame/Content présentée à la Section 4.2.2 du Chapitre 4 fournit des données et prédictions intéressantes sur les systèmes de syllabes. La Figure 13.1 rappelle les prédictions : la précédence du contrôle de la mâchoire sur les autres articulateurs implique des cooccurrences entre consonne labiale et voyelle neutre (de type /ba/), consonne coronale et voyelle avant (de type /di/) et consonne dorsale et voyelle arrière (de type /gu/).

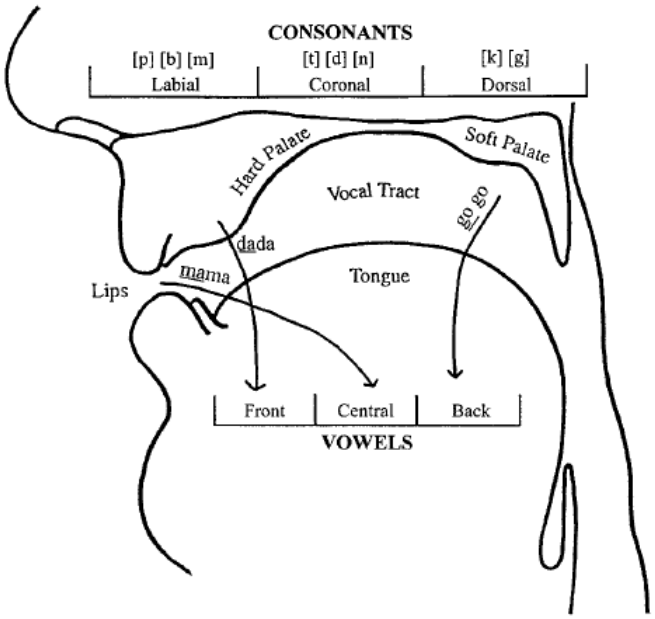


FIGURE 13.1 – Rappel des prédictions de la théorie Frame/Content, d'après MacNeilage et Davis (2000). Les associations préférentielles correspondent à un geste principalement de mâchoire entre la plosive et la voyelle, la configuration de la langue restant, elle, proche entre les deux configurations, ce qui conduit à des liens entre lieu d'articulation de la plosive et de la voyelle.

Ces prédictions se vérifient globalement sur des données provenant de babillage de bébés, de premiers mots d'enfants et de corpus de mots (Figure 13.2). Ces données fournissent le rapport entre les occurrences de certaines associations, prédites par la théorie, et les occurrences attendues par une simple combinatoire des plosives et des voyelles. Le fait que ces

rapports soient systématiquement plus grands que 1 exprime la tendance des productions à s'inscrire plus systématiquement dans ces cooccurrences que si les associations étaient libres (ce qui correspondrait au rapport 1). On peut noter que les rapports sont élevés pour le babillage (babbling) et plus encore pour les premiers mots, qui imposeraient au jeune enfant une simplification accrue du contrôle. Par contre, les rapports diminuent dans la langue adulte, bien qu'ils restent significativement supérieurs à 1, et ce sur un large ensemble de langues (Rousset, 2004).

Data set	Pattern		
	Labial-central	Coronal-front	Dorsal-back
Babbling	1.34	1.28	1.22
First words	1.29	1.48	1.39
Languages	1.10	1.16	1.27

FIGURE 13.2 – Rapports des fréquences observées sur les fréquences attendues des patterns de la Figure 13.1, d'après MacNeilage et Davis (2000).

Toutefois, sans contredire les prédictions de Frame/Content, Vilain et collab. (1999); Serkhane (2005) les relativisent en montrant qu'elles sont contraintes par des variabilités intra- et inter-individuelles, dues à la variabilité de productions vocaliques similaires d'un individu et aux différences de morphologie du conduit vocal entre individus, respectivement. À partir du modèle de conduit vocal VLAM et de quelques variantes, les auteurs étudient les configurations consonantiques atteintes par une simple montée de la mâchoire à partir de différentes configurations vocaliques. Leurs résultats montrent qu'une configuration /a/ aboutit sur ce modèle plus souvent à un /d/ qu'à un /b/, qu'une configuration /i/ aboutit clairement à un /d/, et qu'une configuration /u/ aboutit plus souvent à un /b/ qu'à un /g/ (l'aire aux lèvres étant petite pour un /u/). Ainsi, la théorie Frame/Content adaptées au modèle VLAM prédirait majoritairement des systèmes de type /da, di, bu/ si aucun principe de dispersion n'est appliqué pour contrebalancer l'effet de ces contraintes motrices.

13.1.2 Coarticulation

La coarticulation est exprimée par Sussman et collab. (1998) comme une corrélation entre les seconds formants d'une voyelle et d'une consonne adjacentes. F2 reflète en effet grossièrement la position avant-arrière de la constriction, cette corrélation indique donc un lien entre les deux lieux d'articulation. La Figure 13.3 illustre ce « paradigme du locus », tel que le nomme les auteurs. L'omniprésence de ces phénomènes de coarticulation dans la parole (et ce dans toutes les langues du monde) indique que les configurations articulatoires d'une plosive et d'une voyelle au sein d'une syllabe sont toujours régies par des principes

d'économie articulatoire qui fournissent en quelque sorte une version « souple » du principe de cooccurrence de la théorie Frame/Content, ce principe d'économie indiquant une proximité articulatoire entre configurations labiale et linguale de la plosive à la voyelle.

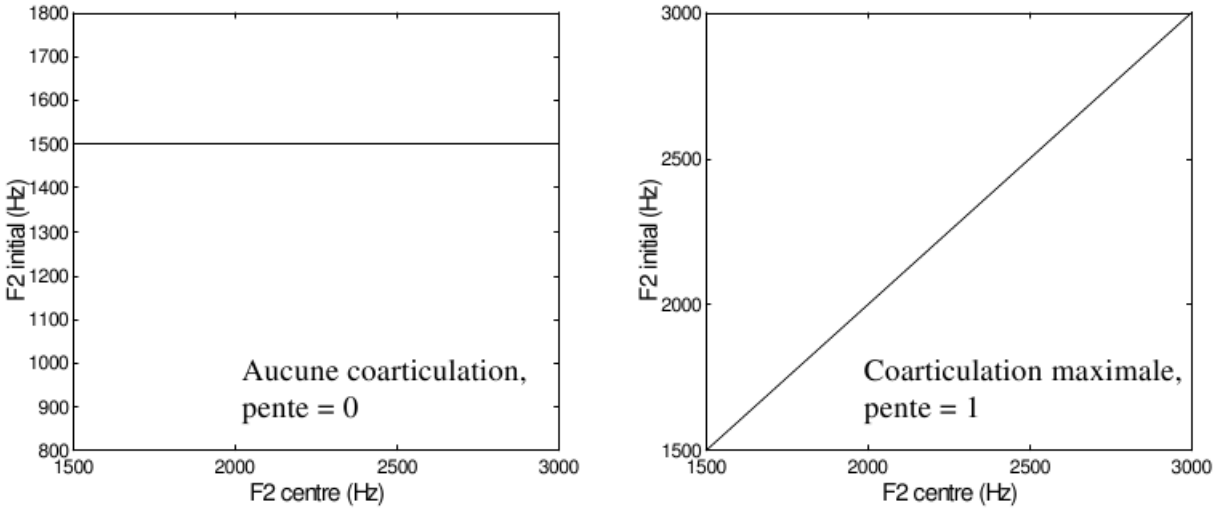


FIGURE 13.3 – Représentation schématique du paradigme du locus, d'après Sussman et collab. (1999) (figure tirée de Ménard (2002)). F2 centre et F2 initial sont les valeurs en Hertz de second formant de la voyelle et de la consonne, respectivement.

Le cas du /b/ est un bon exemple, la seule fermeture labiale laissant la langue très libre. Pour un /bi/, la langue peut rester en position avant pendant toute la réalisation de le syllabe, de même pour un /a/ en position centrale. Cette variabilité a un effet sur les conséquences acoustiques correspondantes, comme le montre la Figure 13.4.

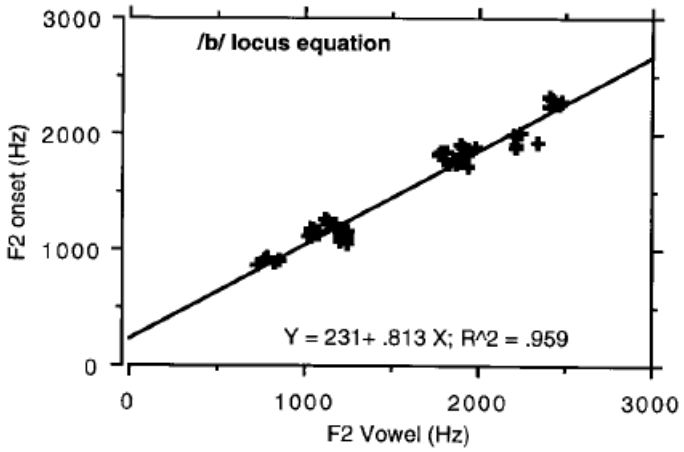


FIGURE 13.4 – Second formand d'une consonne /b/ (ordonné) dans différents contextes vocaliques différenciés sur F2 (abscisse), d'après Sussman et collab. (1998) .

13.2 Modèle de transformation articulatori-acoustique

Nous considérons une syllabe comme la succession d'une voyelle et d'une consonne. Les variables motrices $M = J \wedge TB \wedge TD \wedge LH$ et auditives $S = F1 \wedge F2 \wedge F3$ ont les mêmes domaines que dans les deux chapitres précédents, mais nous les indiçons maintenant respectivement par V pour les voyelles et par C pour les consonnes. Par exemple, la configuration motrice d'une voyelle correspond désormais à une valeur de $M_V = J_V \wedge TB_V \wedge TD_V \wedge LH_V$.

La transformation articulatori-auditive réalisée par l'environnement correspond à la succession des transformations des deux chapitres précédents. Si un agent a produit deux gestes m_V et m_C, les stimuli auditifs correspondant sont respectivement tirés selon $P(S_V \mid [M_V = m_V]\ \delta_{Voy}\ \pi_{Com})$ et $P(S_C \mid [M_C = m_C]\ \delta_{Cons}\ \pi_{Com})$ (Équations 11.4 et 12.1 sur lesquelles on applique un simple changement de variable).

13.3 Modèle d'agent

Nous commençons par exposer la complexité d'un modèle de syllabe complet. Nous montrons que l'explosion de la taille de l'espace articulatori-auditif nécessite d'en considérer des simplifications pour être simulable. Puis nous définissons notre modèle simplifié.

13.3.1 Complexité du modèle complet et pistes de simplification

Le modèle complet est donné par l'Équation 9.1 avec l'instanciation des variables motrices et auditives suivante :

$$\begin{aligned} M &= M_V \wedge M_C, \\ M_V &= J_V \wedge TB_V \wedge TD_V \wedge LH_V, \\ M_C &= J_C \wedge TB_C \wedge TD_C \wedge LH_C \end{aligned}$$

et

$$\begin{aligned} S &= S_V \wedge S_C, \\ S_V &= F1_V \wedge F2_V \wedge F3_V, \\ S_V &= F1_C \wedge F2_C \wedge F3_C. \end{aligned}$$

Pour un objet o_i donné, le comportement sensori-moteur de production produit un geste $m \in M$ selon la question probabiliste de l'Équation 8.12. Il s'agit donc ici d'un couple de gestes moteurs (m_V, m_C) pour la voyelle et la consonne. La Figure 13.5 illustre ce comportement dans le modèle complet de syllabe. La structure de dépendance visualisée correspond à la structure générique de la Figure 8.1 adaptée à la définition des variables M et S ci-dessus.

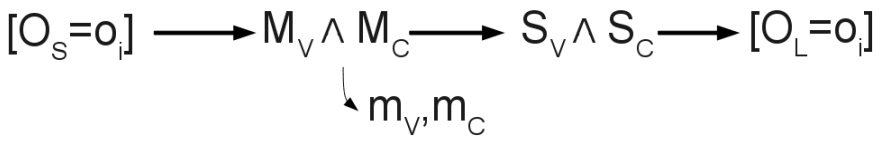


FIGURE 13.5 – Comportement sensori-moteur de production dans le modèle complet de syllabes. La condition de communication $[C = 1]$ est ici représentée de façon équivalent par $[O_S = o_i]$ et $[O_L = o_i]$.

La complexité des processus de tirage et d'inférence dans un tel modèle devient problématique. En effet, la cardinalité des variables M et S correspond maintenant au produit de celles définies pour les voyelles et les consonnes, soient $k_M = (1 * 9 * 9 * 4)^2 = 104\ 076$ et $k_S = (5 * 9 * 4)^2 = 32\ 400$ (voir Section 11.2.2 ; nous ne comptabilisons pas le cardinal de la variable de mâchoire qui est bloquée soit en position neutre pour la voyelle, soit en position fermée pour la plosive). À chaque jeu déictique, un agent doit tirer une valeur de M, ce qui nécessite typiquement une somme de k_M probabilités et une recherche parmi k_M éléments. De plus, le processus d'inférence du comportement sensori-moteur de production de l'Équation 8.12 nécessite de calculer, pour les k_M valeurs de M, une somme sur les k_S valeurs de S. Cette inférence doit avoir lieu à chaque mise à jour des paramètres des distributions motrices et auditives par le processus d'apprentissage, soit tous les N_{App} jeux déictiques dans lesquels l'agent a participé (en général $N_{App} = 200$). La convergence des systèmes nécessitant de l'ordre de $N_{JD} = 100\ 000$ jeux déictiques, on comprend l'obligation de considérer des versions simplifiées du modèle.

Une solution serait l'optimisation du processus d'inférence par des méthodes de type Monte-Carlo qui permettent d'approximer des sommes par des tirages aléatoires favorisant les valeurs de poids fort. Nous n'avons pas étudié en détail ce type de méthodes dans cette thèse. Nous choisissons plutôt de casser la complexité du processus d'inférence par la séparation du modèle complet en deux sous-modèles couplés :

- la voyelle est produite indépendamment de la consonne par le modèle du Chapitre 11 ;
- le système moteur du modèle de consonne est contraint par la position des articulateurs de langue et de lèvres de la voyelle (la mâchoire reste contrainte en position haute).

13.3.2 Séparation en deux sous-modèles couplés

La simplification du modèle complet que nous avons choisie est illustrée Figure 13.6. Le comportement de production se décompose maintenant en deux étapes. La première (en haut) correspond exactement au modèle de production de voyelles du Chapitre 11 et

produit un geste m_v. La seconde (en bas) utilise un modèle similaire à celui du Chapitre 12, mais augmenté d'une variable M_V permettant une influence de la configuration motrice vocalique m_V sur la production de la consonne. C'est ainsi que s'effectue le lien entre le modèle de voyelles et le modèle de consonnes au sein de ce modèle de syllabes.

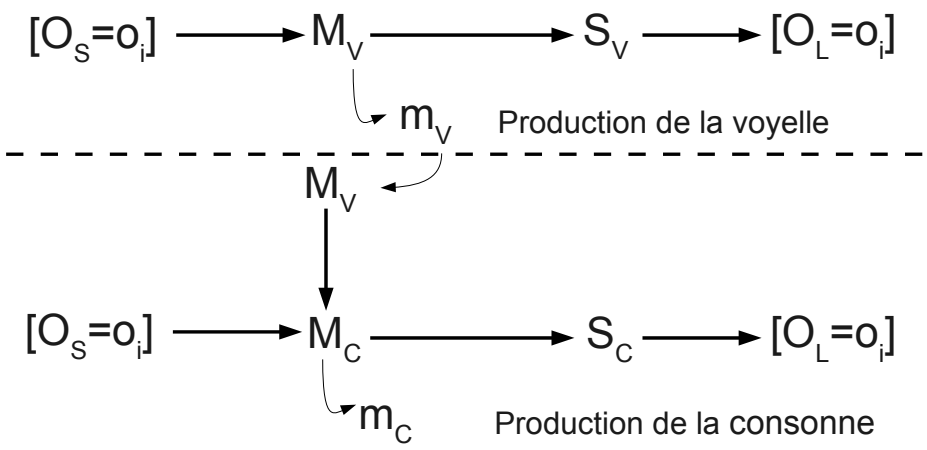


FIGURE 13.6 – Comportement sensori-moteur de production dans le modèle simplifié de syllabes. La condition de communication $[C = 1]$ est ici représentée de façon équivalente par $[O_S = o_i]$ et $[O_L = o_i]$.

Avant de détailler plus formellement ce comportement, rappelons que la variable de mâchoire est contrainte à la fois dans le modèle de voyelles (en position neutre $J_V = 1$, voir Section 11.3.2) et dans le modèle de consonnes en condition mâchoire fermée (en position haute $J_C = 3$, voir Section 12.4.1). Le modèle de syllabe conserve ces contraintes et correspond ainsi à un demi-cycle de mâchoire.

13.3.2.1 Ensemble d'apprentissage

L'ensemble d'apprentissage $\delta_a(t)$ d'un agent a au temps t associe donc à chaque jeu déictique $n \in [0, t]$ soit un triplet $(m_V, m_C, o_i) \in (\mathcal{D}_{M_V} \times \mathcal{D}_{M_C} \times \mathcal{D}_{O_S})$ si l'agent a un statut de locuteur au temps n, soit un triplet $(s_V, s_C, o_i) \in (\mathcal{D}_{S_V} \times \mathcal{D}_{S_C} \times \mathcal{D}_{O_L})$ s'il a un statut d'auditeur. Les jeux dans lesquels l'agent a n'intervient pas ne sont pas enregistrés dans $\delta_a(t)$. Nous notons $\delta_a^M(t)$ (respectivement $\delta_a^S(t)$) l'ensemble des N_{App} derniers éléments de $\delta_a(t)$ enregistrés en statut de locuteur (respectivement auditeur).

13.3.2.2 Production de la voyelle

Le modèle de voyelles est strictement identique à celui du Chapitre 11. Pour un objet donné o_i : il produit un geste moteur $m_V \in M_V$ selon le comportement sensori-moteur de production.

13.3.2.3 Production de la consonne

Pour modéliser l'influence d'une dépendance motrice entre les deux réalisations, le geste m_V produit pour la voyelle est également transmis au modèle de consonnes. Celui-ci est similaire à celui du Chapitre 12 mais inclut maintenant la variable M_V représentant la configuration articulatoire de la voyelle précédente (bas de la Figure 13.6). La distribution conjointe correspondante s'écrit alors :

$$\begin{aligned} & P(O_S\ M_V\ M_C\ S_C\ O_L\ C \mid \delta_a(t)\ \pi_{Ag}) \\ & = P(O_S \mid \pi_{Ag})P(M_V \mid \pi_{Ag})P(M_C \mid M_V\ O_S\ \delta_a(t)\ \pi_{Ag})P(S_C \mid M_C\ \pi_{Ag}) \\ & \quad P(O_L \mid S_C\ \delta_a(t)\ \pi_{Ag})P(C \mid O_S\ O_L\ \pi_{Ag}). \end{aligned} \tag{13.1}$$

Nous considérons donc que seul le système moteur, $P(M_C \mid M_V\ O_S\ \delta_a(t)\ \pi_{Ag})$, dépend de la production de la voyelle. Les autres termes de la décomposition sont définis de la même façon qu'au chapitre précédent, seul l'indiçage des variables articulatori-auditives change pour les différencier du modèle de voyelles. Nous considérons que la position de chaque articulateur pour la consonne dépend de la position du même articulateur pour la voyelle, excepté pour la mâchoire qui reste contrainte en position haute. Le système moteur se décompose donc en :

$$\begin{aligned} & P(M_C \mid M_V\ O_S\ \delta_a(t)\ \pi_{Ag}) \\ & = P(J_C\ TB_C\ TD_C\ LH_C \mid J_V\ TB_V\ TD_V\ LH_V\ O_S\ \delta_a(t)\ \pi_{Ag}) \\ & = P(J_C \mid \pi_{Ag})P(TB_C \mid TB_V\ O_S\ \delta_a(t)\ \pi_{Ag}) \\ & \quad P(TD_C \mid TD_V\ O_S\ \delta_a(t)\ \pi_{Ag})P(LH_C \mid LH_V\ O_S\ \delta_a(t)\ \pi_{Ag}). \end{aligned}$$

La position haute de la mâchoire correspond à :

$$P(J_C \mid O_S\ \delta_a(t)\ \pi_{Ag}) = P(J_C\ |\pi_{Ag}) = \boldsymbol{\delta}_3(J).$$

Les distributions correspondant aux trois autres variables articulatoires sont définies initialement par :

$$\begin{aligned} & \forall X \in \{TB, TD, LH\},\ x_V \in X_V,\ o_i \in O_S\ : \\ & \quad P(X_C \mid [X_V = x_V]\ [O_S = o_i]\ \delta_a(0)\ \pi_{Ag}) \ = \ \boldsymbol{G}_{x_V,1}(X_C). \end{aligned} \tag{13.2}$$

La configuration consonantique de chaque articulateur dans $\{TB, TD, LH\}$ est donc très contrainte par la configuration vocalique qui la précède. Plus précisément, la distribution

d'un articulateur pour la consonne est une loi gaussienne centrée sur la position de l'articulateur pour la voyelle et d'écart-type 1.

Puis, tous les N_{App} jeux déictiques dans lesquels l'agent a a été en situation de locuteur, nous définissons l'apprentissage des termes de la distribution motrice par :

$$\begin{aligned} &\forall X \in \{TB, TD, LH\},\ x_V \in X_V,\ o_i \in O_S\ : \\ &\quad P(X_C \mid [X_V = x_V]\ [O_S = o_i]\ \delta_a(t)\ \pi_{Ag}) \ = \ \boldsymbol{G}_{\delta_a^{M,o_i,x_V}}(X_C). \end{aligned} \tag{13.3}$$

où $\delta_a^{M,o_i,x_V}(t)$ correspond à l'ensemble $\delta_a^M(t)$ restreint aux enregistrement pour lesquels $[O_S = o_i]$ et $[X_V = x_V]$. Nous disposons donc d'un système moteur initialement très contraint par la position précédente des articulateurs, mais capable d'évolution par apprentissage au cours des jeux déictiques.

Le comportement de production de consonnes correspond finalement au tirage d'un geste moteur m_C connaissant le geste produit pour la voyelle m_V et l'objet o_i (le même objet que pour la production de la voyelle) selon la question suivante, posée à la distribution conjointe de l'Équation 13.1 :

$$\begin{aligned} &P(M_C \mid [M_V = m_V]\ [O_S = o_i]\ [C = 1]\ \delta_a(t)\ \pi_{AgSM}) \\ &\quad \propto \ P(M_C \mid [M_V = m_V]\ [O_S = o_i]\ \delta_a(t)\ \pi_{Ag}) \\ &\qquad \sum_{S_C} P(S_C \mid M_C\ \pi_{Ag}) P([O_L = o_i] \mid S_C\ \delta_a(t)\ \pi_{Ag}) \end{aligned}$$

La seule différence avec le modèle de production de consonnes du chapitre précédent est donc l'introduction d'une variable motrice supplémentaire, dont la valeur connue indique le geste produit par la voyelle et contraint initialement le système moteur consonantique.

13.4 Simulations

Ces simulations concernent des sociétés de 2 agents évoluant dans un environnement de 3 objets, conduisant à l'émergence de systèmes de 3 syllabes. Le temps de simulation est de 150 000 jeux déictiques. Toutes les simulations considérées concernent des sociétés d'agents en comportement sensori-moteur.

$$\begin{aligned} N_A &= 2, \\ N_O &= 3, \\ N_{JD} &= 150\ 000. \end{aligned}$$

13.4.1 Paramètres variés

Nous héritons dans ce dernier travail de simulation d'un ensemble de paramètres défini et validé dans les deux chapitres précédents, et notamment :

- Des poids identiques sur chaque dimension formantique ne permettent pas de bonnes prédictions des systèmes de 5 voyelles. Des rapports 1-3-6, modélisés par un bruit sur F2 trois fois supérieur à F1 et un bruit sur F3 six fois supérieur, semblent être de bons candidats pour la prédiction des systèmes de 3 et 5 voyelles.
- Ces mêmes rapports 1-3-6 permettent également la cohérence des résultats avec les données sur les systèmes de consonnes. On montre que l'absence de contraintes sur la mâchoire laisse émerger des consonnes arrières, pourtant peu représentées dans les langues du monde. Une contrainte de mâchoire fermée les élimine et privilégie nettement le système majoritaire des langues du monde : /b,d,g/.

Ce sont ces paramètres que nous utilisons ici. En conséquence, nous ne faisons pas à proprement parler varier de paramètres dans ces simulations mais nous voulons plutôt montrer comment l'introduction d'une dépendance entre les gestes moteurs vocalique et consonantique agit sur la forme des systèmes de syllabes obtenus, en les comparant à deux cas extrêmes :

- la production de la consonne ne dépend en rien de celle de la voyelle,
- la production de la consonne est complétement déterminée par celle de la voyelle.

Nous n'avons pas lancé de simulations sur ces deux cas extrêmes mais nous en prédisons les résultats d'après des travaux précédents.

Le premier cas correspond simplement à un comportement de production qui enchaine le tirage d'une voyelle selon le modèle du Chapitre 11 et d'une consonne selon celui du Chapitre 12. Le premier tirage n'aurait alors aucune influence sur le second et on obtiendrait donc des systèmes syllabiques composés d'un système vocalique de la Section 11.5.2.1, majoritairement /a,i,u/, et d'un système consonantique de la Section 12.5.2, majoritairement /b,d,g/. Les systèmes syllabiques obtenus correspondraient donc majoritairement à des associations aléatoires de {/a/,/i/,/u/} avec {/b/,/d/,/g/}, soient 6 systèmes majoritaires équiprobables : /ba, di, gu/, /ba, gi, du/, /da, bi, gu/, /da, gi, bu/, /ga, bi, du/ et /ga, di, bu/ (nous adoptons la notation CV bien que le modèle produise des séquence VC pour des comparaisons plus aisées avec les données). Pour se rendre compte de l'effet de l'introduction de la dépendance motrice entre la voyelle et la consonne, il suffit donc de comparer la distribution des systèmes syllabiques composés de /a,i,u/ et /b,d,g/ émergeant du modèle défini à la section précédente avec une loi uniforme.

Concernant le deuxième cas, où la production de la consonne est complétement déterminée par celle de la voyelle, nous utilisons les données de Vilain et collab. (1999); Serkhane (2005) présentées à la Section 13.1 qui prédisent plutôt des systèmes de type /da, di, bu/.

Nous voulons donc comparer à ces deux cas nos résultats de simulations, dans lesquelles les réalisations motrices ne sont ni complètement contraintes ni complétement dépendantes.

13.4.2 Évaluation

Toutes les simulations sont réalisées aux mêmes niveaux de bruit, choisis à partir des simulations des deux chapitres précédents pour obtenir de bonnes dispersions. D'après les Figures 11.12 et 12.8, $\sigma_{F1} = 0.4$ semble être un bon compromis pour une dispersion correcte des voyelles et des consonnes.

La classification des systèmes de syllabes émergeant de nos simulations concatène les classifications de systèmes de voyelles et de consonnes présentés aux Sections 11.4.2 et 12.4.2. Ainsi, un système syllabique est composé de trois éléments, un par objet, chacun d'entre eux associant une classe de voyelles parmi 7 et une classe de consonne parmi 5 (on ne compte pas les consonnes arrières rendues impossibles par la contrainte de mâchoire fermée), soit 35 configurations possibles pour chacune des trois syllabes d'un système simulé.

13.5 Résultats

Nous avons lancé 33 simulations indépendantes. Les résultats sont présentés Table 13.1.

Système	Pourcentage
/da, di, bu/	39.4%
/ba, di, bu/	21.2%
/ba, di, gu/	15.1%
/da, d'e', bu/	6.1%
/da, bi, gu/	6.1%
/ba, d'e', gu/	3%
/da, bi, bu/	3%
/ga, di, bu/	3%
/ba, di, g'o'/	3%

TABLE 13.1 – Pourcentages des systèmes de syllabes émergeant de nos simulations.

La Table 13.2 expose ces systèmes de syllabes. Les voyelles sont visualisées dans le plan F1-F2 dans lequel on les catégorise, et les consonnes dans le plan F2-F3, F1 étant quasiment constant pour les consonnes mâchoire fermée (voir Figure 12.5). Dans chaque simulation (chaque case du tableau), les ellipses de la même couleur correspondent à la même syllabe.

Le système /da,di,bu/ prévu dans le cas de dépendance totale reste majoritaire. Il semble donc que la force de dispersion du comportement sensori-moteur ne soit pas toujours suffisante pour casser la contrainte motrice qui impose ce système pourtant mauvais en terme de dispersion auditive consonantique. L'atteinte d'une configuration /g/ à partir d'un /a/, qui permettrait l'émergence d'un système à bonne dispersion, semble rendue très difficile par la contrainte motrice car le passage de l'une à l'autre nécessite un important mouvement du dos de la langue TD du bas vers le haut. On observe d'ailleurs que le seul /ga/ obtenu correspond en fait à une piètre réalisation de la consonne (système /ga, di, bu/ de la Table 13.2.

La dispersion auditive a toutefois souvent raison des contraintes motrices, menant majoritairement aux systèmes /ba, di, bu/ et /ba, di, gu/. Pour le premier on remarque que, de même qu'au chapitre précédent, les systèmes avec deux exemplaires différents de /b/

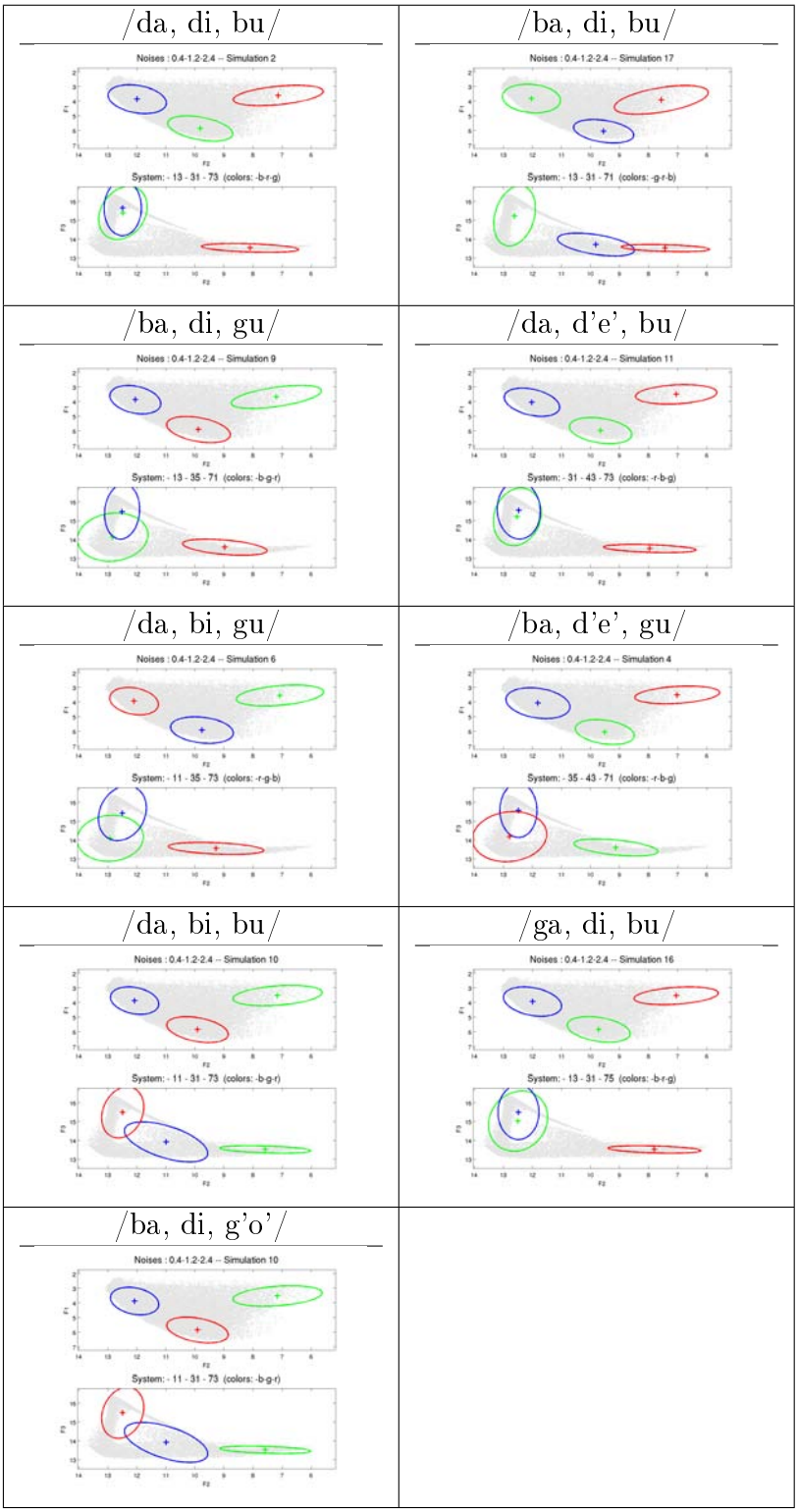


TABLE 13.2 – Visualisation des systèmes de syllabes, dans les plans F1-F2 pour la voyelle et F2-F3 pour la consonne.

sont bien représentés, ce qui s'explique naturellement par la bonne dispersion qu'ils permettent. Le second est celui prédit par la théorie Frame/Content, associant voyelle centrale et consonne labiale, voyelle avant et consonne alvéolaire, et voyelle arrière et consonne vélaire. On en trouve également deux variantes dans les systèmes /ba, d'e', gu/ et /ba, di, g'o'/, qui n'en diffèrent que par une dispersion légèrement moindre d'une des voyelles (voir Table 13.2).

On note également l'absence de deux syllabes, /du/ et /gi/. Ces deux réalisations nécessitent en effet d'importants mouvements du corps de langue sur la dimension avant/arrière, leur absence est en accord avec la théorie Frame/Content.

Ces résultats conduisent donc à de fortes tendances de cooccurrences, associées à un lien sans doute trop étroit entre plosive et voyelle. Elles devraient donc également produire des traces fortes de coarticulation, que nous vérifions sur la Figure 13.7. Celle-ci rend compte des phénomènes de coarticulation en lien avec le paradigme du locus de Sussman et collab. (1998) présenté à la Section 13.1. On observe une corrélation, et donc une coarticulation, maximale pour le système /di,ba,bu/, montrant que les deux exemplaires de /b/ sont coarticulés avec la voyelle. Il est également intéressant d'observer comment les réalisations du /b/ en contexte /ba/, /bi/ ou /bu/ s'inscrivent clairement dans une droite de régression proche de l'équation du locus proposée par Sussman et collab. (1998) (voir Figure 13.4).

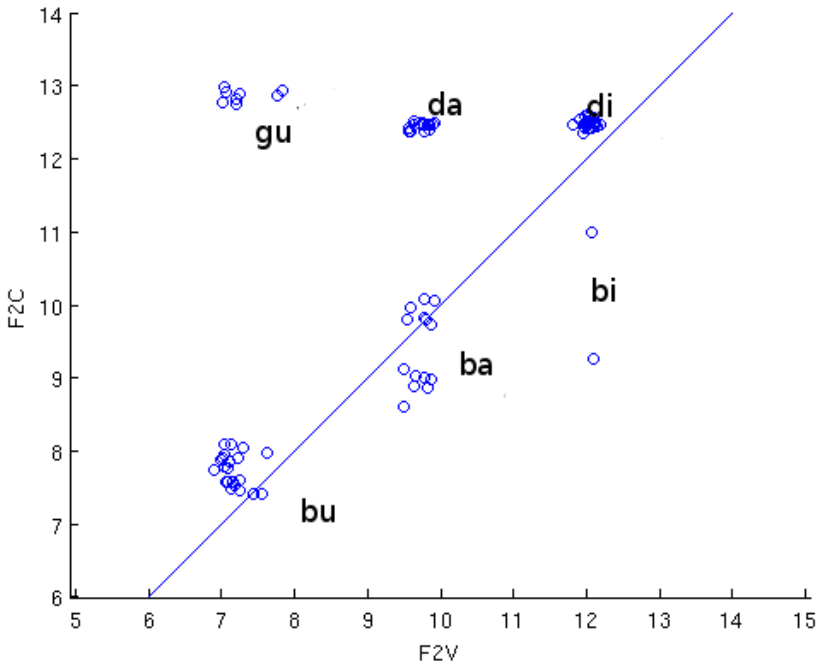


FIGURE 13.7 – Second formant de la consonne (F2C) en fonction de celui de la voyelle (F2V), en Barks. La droite est la fonction identité.

13.6 Conclusion

Nous avons exposé dans ce chapitre nos premiers résultats de simulation d'émergence de systèmes syllabes. Malgré les limitations imposées par la complexité calculatoire, que nous avons pu casser par une séparation en deux sous-modèles couplés, c'est la première fois que l'émergence de systèmes de ce type est simulée sur VLAM. Nous avons ainsi réussi à combiner des principes de dynamique motrice inspirés de la théorie Frame/Content et des principes de dispersion auditive inhérents à notre comportement de production sensori-moteur.

Globalement, ces résultats montrent que l'introduction d'une dépendance motrice entre la production de la voyelle et de la consonne influe sur la forme des systèmes de syllabes obtenus. Plus précisément, nous nous plaçons quelque part dans le continuum entre une dépendance totale dans laquelle la consonne est complétement déterminée par la voyelle qui la précède, dont la prédiction pour la morphologie VLAM est le système /da, di, bu/ ; et une indépendance totale des deux productions, dont la prédiction est l'équiprobabilité des 6 systèmes possibles associant /a,i,u/ et /b,d,g/. L'écart-type initial des gaussiennes de l'Équation 13.2 peut jouer le rôle curseur pour se déplacer dans ce continuum, bien que nous n'ayons pas étudié les effets de sa variation dans ce document (sa valeur est fixée à 1). En effet, une valeur infiniment petite correspondrait à une dépendance totale (le geste consonantique serait en conséquence indépendant de l'objet conditionnellement à la voyelle précédente), alors qu'une valeur infinie correspondrait au cas où le système moteur du modèle de consonnes ne dépendrait pas du geste produit pour la voyelle (et ne dépendrait donc que de l'objet). En rapport avec la théorie Frame/Content, dont les bases onto- et phylogénétiques sont présentées à la Section 4.2.2, nous sommes en quelque sorte en mesure de modéliser le passage du « Frame » (le cadre, imposant une dépendance forte motrice) au « Content » (le contenu, dans lequel les deux productions sont contrôlées et donc indépendantes).

Par contre, ce modèle ne semble pas être adapté pour rendre compte de la compositionalité des voyelles et des consonnes, c'est-à-dire l'émergence de toutes les combinaisons possibles au sein d'un même système : /ba/ /bi/ /bu/ /da/ /di/ /du/ /ga/ /gi/ /gu/. En effet, 9 éléments nécessitent dans notre paradigme 9 objets. Le modèle de syllabe défini ici évoluerait donc vers un système associant 9 voyelles à 9 consonnes différentes en les dispersant autant que possible dans leurs espaces respectifs, au lieu d'une composition de 3 par 3. Il nous donc manque ici un ingrédient, soit parce que nous l'avons perdu dans la simplification du modèle, soit parce qu'il nécessite d'être introduit explicitement dans le modèle sur la base d'éléments de théorie de l'émergence traitant de la compositionalité. Nous discuterons de ce point dans la conclusion générale qui suit.

Quatrième partie

Conclusion

Chapitre 14

Conclusion générale

Nous avons traité dans cette thèse la problématique de la morphogénèse des unités de langage. Nous l'avons abordée par une approche que nous qualifions de « construction par le lien », dans une démarche d'unification consistant à ancrer les théories de la forme dans des théories plus générales d'émergence du langage. Ces travaux ont été tirés par une volonté de modélisation computationnelle et de simulations dans un paradigme d'interaction entre agents prélangagiers. Nous nous sommes appuyés sur deux hypothèses centrales.

La première est une hypothèse d'internalisation d'une situation de communication dans l'architecture cognitive des agents, associant représentations motrices et sensorielles à travers un lien sensori-moteur. Nous l'avons justifiée par une revue détaillée des modèles existants de production et de perception de la parole en ligne, pour dégager des éléments pertinents du « produit fini ». Nous avons conclu sur la nécessité d'un lien entre le système moteur et le système auditif dans une théorie de la communication complète étudiant conjointement production et perception et avons proposé un premier cadre unificateur dans lequel les théories motrices et auditives sont conçues comme des cas particuliers des théories sensori-motrices, cadre formalisé dans un modèle bayésien d'agent communicant.

La seconde est une hypothèse de communication prélangagière fournissant une référence d'ordre sémantique et agissant comme une amorce évolutive permettant la morphogénèse des unités de langage. Nous l'avons justifiée par une revue des théories de l'émergence dans lesquelles cette hypothèse semble nécessaire pour traiter la question de la référence, que ce soit par de la reconnaissance d'action, de la deixis ou des intentions partagées. Cette hypothèse est modélisée par un paradigme d'interaction entre agents par jeux déictiques, agissant comme un signal d'apprentissage vers l'évolution d'un code de parole permettant de communiquer dans son état final en l'absence d'un objet physique à montrer.

À partir de ces deux hypothèses, nous avons construit des simulations d'émergence des systèmes phonologiques qui montrent la cohérence globale des résultats avec des principes de théories de la forme (dispersion, théorie quantique) et avec les données des langues du monde, et valident ainsi un modèle homogène et réaliste d'émergence des systèmes de voyelles, de consonnes plosives et de syllabes.

Ce sont les grandes lignes de nos contributions principales, que nous explicitons plus en détail dans la première section de cette conclusion générale. Puis nous présentons ra-

pidement deux contributions complémentaires, issues de travaux satellites à la thèse : l'encadrement d'un projet de master d'ouverture de ce modèle à des problématiques de perception de la parole (projet maintenant poursuivi en thèse), et des travaux réalisés lors d'une visite de 6 mois chez Michael Arbib [1], de conception d'un nouveau modèle computationnel d'adaptation à un accent étranger proposant de nouveaux rôles des neurones miroirs en perception de la parole. Nous terminons sur une discussion et des perspectives d'améliorations, d'ouvertures et d'extensions théoriques du modèle, ainsi qu'une liste de nos publications.

Chaque contribution est mise en valeur par une mise en forme dédiée.

14.1 Contributions principales

14.1.1 Unification théorique

Notre première contribution est l'unification des théories de la production et de la perception de la parole. Les théories motrices et auditives sont conçues comme des cas particuliers des théories sensori-motrices, par la désactivation du système auditif ou moteur, respectivement. Cette unification est formalisée dans un modèle bayésien associant une expression probabiliste à chacune des six classes de théories. Elle met en valeur les enjeux computationnels de chacune et n'a pas d'équivalent à notre connaissance dans la littérature.

Contribution 1 *Les théories motrices et auditives de la communication peuvent être conçues comme des cas particuliers de théories sensori-motrices qui couplent un système moteur et un système auditif à travers un lien sensori-moteur. Cette conception permet d'exprimer les enjeux computationnels de chaque type de théorie sous forme probabiliste à partir d'un modèle bayésien unifié.*

Notre deuxième contribution valide notre approche de construction par le lien. Nous avons montré que les théories de la forme pouvaient s'ancrer dans les théories de l'émergence. En particulier, la dispersion et l'aspect quantique de la parole émergent du système à partir des hypothèses d'internalisation et de communication. Contrairement à la contribution précédente, celle-ci trouve de nombreux équivalents dans la littérature sur les modèles d'émergence par simulation multi-agents (Section 4.3). Ainsi, bien qu'utilisant au moins implicitement l'hypothèse d'internalisation par le couplage de cartes neurales articulatoire et auditive, Oudeyer (2003) avait déjà montré un résultat plus fort : des principes des théories de la forme peuvent émerger sans hypothèse de communication. Mais il faut remarquer que si la communication n'est pas une hypothèse du système, elle n'en est pas non plus une propriété émergente : certes une phonologie apparaît, mais elle n'est pas opérationnelle pour une tâche communicative car rien ne la relie à du sens. Il est intéressant de noter une analogie avec les théories de l'émergence. Trois de celles que nous avons présentées au Chapitre 4 proposent une hypothèse de communication prélangagière : sous la forme

1. Directeur de l'USC Brain Project à Los Angeles. Financement Explora-Doc de la région Rhône-Alpes.

d'imitation simple puis complexe dans l'hypothèse du système miroir, de déixis dans Vocalize to Localize ou d'intentions partagées pour Tomasello. La seule qui n'en propose pas est la théorie Frame/Content, mais elle ne propose pas non plus de chemin évolutif vers la construction du sens.

À l'opposé, notre approche nous incite à aborder la question de la référence et nous permet de concevoir un système qui n'est plus réalisé sur la seule base de mesures topologiques (mesure de dispersion par exemple) mais également par sa capacité à réaliser son objectif fonctionnel : communiquer. Nous montrons ainsi comment des agents peuvent évoluer d'une communication déictique, les dotant d'une capacité d'attention partagée sur un même objet, à une communication vocale permettant la communication d'un objet cognitif du locuteur à l'auditeur en l'absence de son corrélat physique. En d'autres termes, l'hypothèse de communication impose le succès de la communication $[C = 1]$ dans des interactions en présence de l'objet, au cours desquelles les agents apprennent à l'internaliser pour évoluer vers une communication uniquement vocale « de cerveau à cerveau ».

Contribution 2 *Une architecture cognitive couplant un système moteur et un système auditif à travers un lien sensori-moteur dispose de tous les éléments fonctionnels nécessaires à une internalisation de la situation de communication parlée. Une société d'agents plongés dans une amorce évolutive de communication prélangagière de type déixis, et implémentant une telle architecture cognitive évoluent vers la morphogénèse d'un code de parole optimal.*

14.1.2 Réalisation computationnelle et résultats de simulation

La comparaison des comportements de production moteur, auditif et sensori-moteur montre que :

- le comportement moteur ne permet pas l'émergence d'un code de parole efficace dans notre paradigme d'interaction car il produit pour lui-même, sans tenir compte des autres agents,
- le comportement auditif le permet car il cherche à optimiser un classifieur dont les prototypes sont appris à partir des stimuli reçus des autres agents,
- le sensori-moteur est le plus efficace car il permet de focaliser les gestes produits sur les zones stables du classifieur.

Ces deux derniers comportements permettent l'émergence de systèmes en accord avec les prédictions de la théorie de la dispersion de Lindblom et de la théorie quantique de Stevens.

Contribution 3 *La prise en compte des besoins de l'auditeur, par exemple par leur internalisation dans le système auditif, est nécessaire à l'émergence d'un code de parole. L'ajout des besoins du locuteur, c'est-à-dire l'internalisation complète de la chaine de communication, favorise et accélère l'émergence de codes optimaux. Le couplage d'un système moteur et d'un système auditif dans une théorie sensori-motrice de la communication permet cette internalisation et conduit aux systèmes de communication les plus efficaces.*

Contribution 4 *L'émergence d'un code de parole dans le modèle est en accord avec les prédictions de la théorie de la dispersion et de la théorie quantique. Nous montrons que ces prédictions dépendent de contraintes articulatori-auditives et de communication.*

Nous avons également effectué le passage à l'échelle du modèle formel vers son instanciation réaliste complète et homogène pour l'émergence de systèmes de voyelles, de consonnes plosives et de syllabes. Ceci nécessite un cadrage articulatoire et auditif complexe entre voyelles et consonnes permettant l'évolution du modèle vers la syllabe, ainsi que l'exécution d'un nombre important de simulations pour en analyser les propriétés. Ce travail de réalisation computationnelle, couplant les représentations motrices et auditives entre elles et avec une référence sémantique, puis temporellement dans des séquences voyelle-consonne, est à notre connaissance pas ou peu rencontré dans la littérature. Ceci nécessite un long travail de choix de modélisation pour contenir autant que possible l'explosion combinatoire induite par les nombreux couplages.

Le cadrage paramétrique permettant la cohérence globale de nos résultats de simulation avec les données est le suivant :

- des variables articulatori-auditives de mêmes domaines pour les voyelles et les consonnes,
- F1 doté de trois fois plus de poids que F2 et 6 fois plus que F3,
- la mâchoire en position neutre pour la réalisation de la voyelle, en position haute pour celle de la consonne.

Concernant les poids respectifs des dimensions formantiques, le rapport 1-3-6 permet une prédiction cohérente des systèmes de voyelles et de consonnes plosives. Plus précisément, la nature convexe de la courbe de la mesure de Lindblom en fonction du bruit de l'environnement suggère une valeur autour de $\sigma_{F1} = 0.4$ qui permet à la fois une bonne dispersion des systèmes de 3 voyelles et 3 consonnes (Figure 11.12 et 12.8). La signification précise de cette valeur est la suivante : des dimensions formantiques comprises entre 2 et 7 Barks pour F1, 5 et 14 pour F2 et 13 et 17 pour F3, discrétisées par pas de 1 Barks, et sur lesquelles on applique des bruits gaussiens d'écarts-types respectifs 0.4, 1.2 et 2.4 Barks, permettent une prédiction cohérente à la fois des systèmes de 3 voyelles et de 3 consonnes plosives.

Ces valeurs de 0.4, 1.2 et 2.4 dépendent certainement de différents facteurs : la taille des dimensions formantiques, l'unité dans laquelle elles sont exprimées, leur discrétisation, le type de bruit appliqué et la cardinalité des systèmes phonologiques considérée. Il est donc difficile d'en fournir une interprétation physique étant donnée leur forte corrélation avec ces nombreux facteurs, mais elles peuvent servir de cadrage paramétrique initial à prendre en compte pour de futurs travaux de prédiction, permettant ainsi d'en accélérer la recherche souvent longue et fastidieuse (à la manière des travaux préalables de Schwartz et collab. (1997b) mais généralisés aux consonnes plosives et aux syllabes).

Contribution 5 *Il est possible de prédire les tendances globales des systèmes de voyelles, de plosives et de syllabes dans un unique espace articulatori-auditif. Celui-ci comprend 4 variables articulatoires, la hauteur de mâchoire, la position du corps et du dos de la langue et la hauteur des lèvres ; et 3 variables auditives, les trois premiers formants. Ces variables*

peuvent être discrétisées à l'identique pour les voyelles, les plosives et les syllabes. Des rapports de type 1-3-6 sur les poids de chaque dimension formantique sont cohérents pour les trois types de systèmes.

14.2 Contributions complémentaires

Le modèle s'applique aussi bien à des questions d'émergence qu'à des questions de communication en ligne. Nous n'avons pas mis en œuvre de simulations de communication en ligne dans le strict de cadre de cette thèse. Ces travaux sont en cours depuis le master, maintenant poursuivi en thèse, de Raphaël Laurent. Le modèle bayésien développé au Chapitre 8 est exploité pour des expériences de simulation de communication en ligne. Le paradigme considère un système phonologique préalablement spécifié dans les connaissances des agents, pour tester les performances des comportements de perception moteur, auditif et sensori-moteur. Le premier résultat est que les trois types de comportements sont indistinguables en condition de communication parfaite, c'est-à-dire lorsque les connaissances motrices et auditives des agents sont identiques et que leur connaissance du lien sensori-moteur correspond exactement à la transformation articulatori-auditive de l'environnement. Un corollaire de ce résultat est que des différences de performances entre les comportements ne peuvent apparaitre qu'en conditions dégradées, par exemple par des différences de connaissances préalables entre les agents (pouvant s'interpréter comme des différences d'accent) ou l'introduction de bruit dans l'environnement. La suite de ces travaux pourrait permettre de fournir des prédictions expérimentales à la psychologie cognitive.

Contribution 6 *Le modèle constitue une base unifiée pour traiter de nombreuses questions de communication parlée et s'applique aussi bien à des questions d'émergence du langage qu'à des questions de communication en ligne. Le premier point à été traité en détail dans cette thèse. Le second est en cours dans la thèse de Raphaël Laurent que je co-encadre avec Pierre Bessière, Julien Diard et Jean-Luc Schwartz. Dans ce cadre, le modèle permet de mener des expériences de simulation utilisant les mêmes connaissances sensori-motrices pour des tâches de production et de perception.*

Dans ce même esprit d'apport des modèles computationnels pour la compréhension des processus de communication en ligne, et en parallèle de cette thèse, je travaille depuis maintenant plus de deux ans avec Michaël Arbib, sur un modèle d'adaptation perceptuelle à un accent étranger servant de point d'appui à une réflexion sur les corrélats neurolinguistiques de la perception de la parole, en lien avec la théorie des neurones miroirs. Ces travaux ont mené à une publication en congrès et une soumission en revue internationale.

14.3 Discussion et perspectives

Ce travail, qui a largement consisté à réfléchir à des formalismes et à des implémentations acceptables, ouvre certainement plus de pistes de développement qu'il ne résout

de problèmes, et appelle de nombreuses perspectives d'améliorations, de développements et d'extensions théoriques (nous confrontant ainsi à la dure réalité de tout ce que nous aurions aimé faire et n'avons pu mener à bien faute de temps).

14.3.1 Améliorations du modèle

14.3.1.1 Simulation

Nous avons pris le soin de définir le modèle indépendamment du domaine des variables articulatori-auditives. Cette séparation entre définition et instanciation permet d'imaginer de nombreuses améliorations possibles. Ainsi, le modèle est extensible pour la prise en compte d'autres articulateurs du modèle VLAM, par exemple la pointe de la langue et la protrusion des lèvres qui permettraient respectivement un meilleur contrôle des consonnes plosives avant de type /d/ et des voyelles arrondies de type /u/. En ce qui concerne les variables auditives, il est possible d'ajouter d'autres dimensions comme par exemple le voisement ou la frication des consonnes. Pour aller plus loin, on peut également considérer d'étendre l'espace perceptif, y incluant l'espace visuel pour rendre compte de phénomènes de parole audio-visuelle, par exemple le célèbre effet McGurk (McGurk et MacDonald, 1976).

14.3.1.2 Formalisation

Un travail de preuve mathématique de l'équivalence entre l'algorithme d'interaction entre agents en comportement sensori-moteur et un algorithme itératif d'estimation des paramètres maximisant le succès dans le modèle de situation de communication, de type Espérance-Maximisation est une perspective intéressante. Elle pourrait déboucher sur des éléments de liens formels entre communication, internalisation et optimisation.

Nous avons également essayé de montrer une équivalence entre la condition de stabilité du comportement auditif (Équation 10.12) et le principe d'optimisation de la théorie de la dispersion (Équation 4.1) mais avons buter sur des problèmes d'analyse mathématique qui dépassaient nos compétences (si tant est que cette preuve soit possible).

Globalement, il y a donc, on le voit, tout un travail de formalisation mathématique, qui nous semble nécessaire pour progresser maintenant dans les recherches computationnelles sur la morphogenèse des systèmes sonores.

14.3.1.3 Validation

Tout modélisateur, on le sait bien, est confronté à des problèmes de validation de ses simulations, et ce problème est à la fois important et complexe. La validation passe toujours par la confrontation avec des données expérimentales, et les seules données auxquelles se confrontent les spécialistes de simulation de systèmes linguistiques sont les statistiques et observations sur les systèmes réels, qui sont le produit de séquences extraordinairement complexes de processus dynamiques multiples.

Or les jeux déictiques auxquels jouent nos agents se prêteraient en théorie volontiers à des confrontations expérimentales, si l'on trouvait le moyen d'y faire jouer des agents réels, qu'ils soient humains ou robots.

Dès le départ de cette thèse, nous avons eu l'idée d'une mise en œuvre d'expériences impliquant humains ou robots et devant servir de validation expérimentale complémentaire. Le temps nous a manqué pour passer à l'acte, mais nous souhaitons tout de même en mentionner le concept. L'idée est inspirée des jeux de communication de Galantucci (2005) qui fait jouer des sujets à des jeux de communication complexes dans lesquels ils ne disposent pas directement des signaux émis par leurs partenaires, et doivent donc, confrontés à des conditions de communication très dégradées, inventer un système de signes (manuels, dans son paradigme) dont Galantucci décrit les propriétés en relation avec celles des systèmes sémiotiques traditionnels. Notre projet consisterait à faire jouer des agents humains, soit entre eux, soit avec des simulations informatiques, à des jeux déictiques dans lesquels les productions vocales qu'ils réaliseraient ou devraient percevoir seraient fortement dégradées (par exemple, en transformant systématiquement les sons par un système d'analyse-synthèse modifiant les formants), et en observant le type de systèmes sonores de communication qu'ils chercheraient alors à mettre en place. On peut alors se demander s'il est possible de voir émerger de nouveaux systèmes sonores, par exemple des systèmes vocaliques dont on contrôlerait le nombre de phonèmes par les situations de communication induites ; puis de voir comment ces systèmes s'adaptent par exemple au bruit de communication (avec une dispersion plus ou moins grande selon le niveau de bruit) ; ou si, pour un trop grand nombre d'objets à désigner, des stratégies de composition de vocalisations peuvent apparaître.

14.3.2 Ouvertures du modèle et illustration de l'approche de construction par le lien

Une limitation majeure de notre travail (par rapport aux espoirs initiaux) est de n'avoir pas dépassé le niveau du phonème et de la syllabe, et notamment de n'avoir pas traité de deux extensions majeures : celle de la compositionnalité phonologique (comment combiner les phonèmes ?), et celle de la compositionalité référentielle vers la syntaxe (comment combiner les mots ?).

Nous fournissons quelques éléments théoriques qui pourraient permettre des ouvertures du modèle dans ce sens, et respectant notre approche de « construction par le lien » entre théories de l'émergence et de la forme.

14.3.2.1 Émergence de la compositionalité

Le problème de la compositionalité phonologique a été abordé au Chapitre 13. Il s'agit de comprendre comment, à partir de systèmes sonores, construire des séquences adéquates, et surtout, ce qui conduit les agents à découvrir cette possibilité de composition et à l'exploiter optimalement. La question centrale est ici celle de l'émergence du principe dit de « Maximal Use of Available Features » (MUAF, voir Chapitre 13) qui permet par

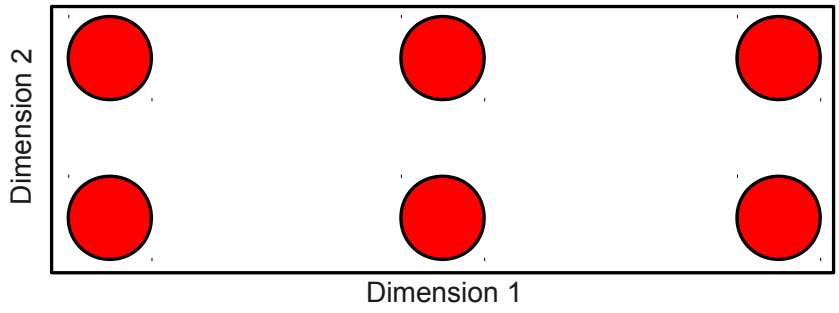
exemple, disposant de voyelles /i a u/ et de plosives /b d g/, à chercher à les combiner systématiquement (même si, on l'a vu, certaines combinaisons sont au départ préférables à d'autres).

Nous pensons que l'émergence d'une compositionalité des voyelles et consonnes nécessite au moins l'implémentation du modèle complet de syllabe, dont la complexité calculatoire est problématique (Section 13.3.1). Toutefois, nous avons mené ce type de simulations sur des espaces simplifiés à deux dimensions et n'avons pas obtenu de résultats concluants. Le problème semble venir d'une incompatibilité entre les principes de compositionalité et de dispersion, comme l'illustre la Figure 14.1. En effet, la composition peut se résumer comme la réutilisation d'un système existant : par exemple les trois éléments du bas de l'espace rectangulaire de la Figure 14.1a, le long de la dimension 1, sont réutilisés à l'identique sur le haut de l'espace. Or, l'optimisation de la dispersion dans l'espace à deux dimensions favorise plutôt des systèmes du type de la Figure 14.1b, en particulier si la dimension 2 est peu discriminante. Les propriétés dispersives de notre modèle mènent soit à des résultats de simulation similaires à ce dernier cas si les poids des dimensions sont très différents, soit à des résultats similaires au premier mais dont l'analyse de la dynamique montre qu'il n'y a pas duplication, mais simplement deux optimisations dispersives débouchant sur les mêmes solutions. Le modèle ne semble donc pas approprié pour rendre compte de l'émergence d'une compositionalité. Une des raisons possibles est celle d'un « manque de substance » : rien dans nos connaissances préalables ne le favorise. Nous proposons ici de fournir quelques éléments à considérer pour traiter ce problème et en profitons pour illustrer les étapes de notre approche de construction par le lien sur un cas nouveau.

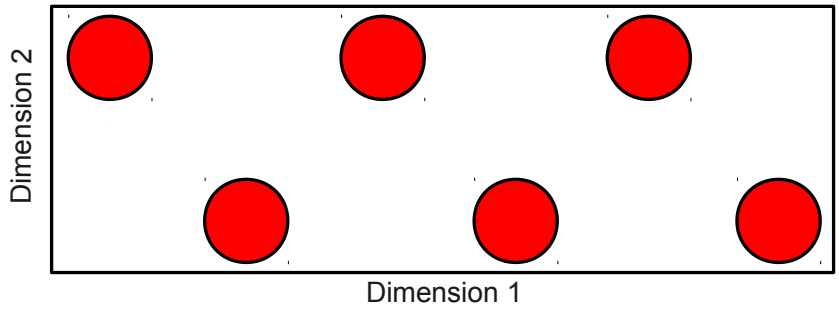
De la revue de la littérature aux modèles conceptuels. Dans l'esprit de la première partie de ce document, cette étape nécessite une revue de littérature traitant de la compositionalité, ainsi que la collecte de données sur les tendances globales dans les données des langues du monde.

Ohala (1979) s'interroge sur la validité de la théorie de la dispersion, en remarquant que celle-ci mène à des prédictions fausses des systèmes utilisant plusieurs traits distinctifs (dans l'esprit de la Figure 14.1). En effet, pour préserver la distinctivité dans de grands systèmes phonologiques, il peut y avoir recours à des traits phonétiques supplémentaires tels que la durée ou la nasalisation. Lorsque ces traits phonétiques sont présents dans une langue, on observe souvent que le système existant a été parfaitement dupliqué pour y ajouter le trait supplémentaire. De plus, ce phénomène n'apparaît que sur les grands systèmes sur lesquels la distinctivité n'est plus suffisante (« Maximum Use of Available Features », (Clements, 2003a,b)).

Lindblom (1998) suggère l'introduction de contraintes développementales pour expliquer pourquoi les systèmes phonétiques obéissent au principe MUAF. L'acquisition de nouveaux phonèmes serait facilitée quand les commandes motrices associées recouvrent celles déjà apprises. Ainsi, l'effort d'apprentissage serait moindre qu'en acquérant un nouveau système n'ayant rien en commun avec le précédent.



(a) Composition : les éléments (disques rouges) dispersés sur la dimension 1 sont composés sur la dimension 2.



(b) Dispersion : si la dimension 2 est petite par rapport à la dimension 1, les systèmes les mieux dispersés disposent leurs éléments à la manière de charges électriques se repoussant les unes des autres.

FIGURE 14.1 – Illustration de l'incompatibilité entre les principes de composition (en haut) et de dispersion (en bas).

Schwartz et collab. (2007) proposent de lier la théorie de la dispersion et le principe MUAF dans une théorie de la perception pour le contrôle de l'action.

Nous pensons que ces quelques références peuvent fournir les éléments nécessaires pour lier des hypothèses prélangagières d'économie d'apprentissage à l'émergence de la compositionalité dans la morphogénèse du langage.

Des modèles conceptuels aux modèles formels. Dans l'esprit de la seconde partie de ce document, cette étape nécessite la spécification précise des connaissances préalables engendrées par les hypothèses prélangagières et leurs influences éventuelles sur l'interaction entre les agents.

Il conviendra d'abord de formaliser l'hypothèse d'économie d'apprentissage proposée par Lindblom dans le modèle d'agent. La question centrale est ici : comment un apprentis-

sage préalable sur certaines dimensions d'un espace articulatori-auditif peut être réutilisé pour un apprentissage ultérieur combinant d'autres dimensions ? Les capacités de discrimination sur chaque dimension sont peut-être ici un point important, comme le proposait déjà Berrah (1998).

Puis, comme la composition implique une duplication de connaissances déjà apprises, il conviendra de définir un paradigme d'interaction dans lequel les agents « découvrent » une nouvelle dimension après avoir déjà convergé vers un premier système. Une solution pourrait-être l'ajout d'objets dans l'environnement après une première convergence de la simulation.

Des modèles formels aux simulations informatiques. Dans l'esprit de la troisième partie de ce document, cette étape nécessite la simulation du modèle obtenu et sa comparaison avec les données des langues du monde, débouchant idéalement sur une mise en valeur du lien entre des hypothèses prélangagières et des principes de morphogénèse. Il conviendra alors de lancer des simulations réalistes d'interaction entre les agents spécifiés à l'étape précédente et d'analyser la cohérence globale des résultats avec les données pour répondre à la question : l'hypothèse prélangagière d'économie articulatoire (et d'autres, peut-être) contraint-elle la morphogénèse des codes de parole vers la composition d'éléments déjà appris ?

14.3.2.2 Émergence de la syntaxe

La seconde ouverture majeure concerne la compositionnalité de la référence. En effet, si la proposition d'une amorce évolutive déictique est convaincante, elle reste tout de même limitée à une attention partagée sur un objet et donc à une référence « élémentaire ».

De la revue de la littérature aux modèles conceptuels. L'hypothèse du système miroir (Section 4.2.1) peut fournir une ouverture, proposant un précurseur prélangagier d'attention partagée sur une action transitive vers un objet, et donc vers une référence plus complète d'ordre syntaxique. Roy et Arbib (2005) proposent en effet que les propriétés syntaxiques élémentaires du langage soient déjà présentes dans le système moteur par des associations entre actions et objets (équivalent de la structure syntaxique verbe-argument).

En d'autres termes, le comportement prélangagier de saisie d'un objet et son instanciation communicative potentielle proposée par Arbib comme pivot vers un protolangage (mime d'une action de saisie) présente la propriété majeure de référer à la fois à un objet et à une action sur cet objet (« ce grain de raisin, je le prends », d'où « je prends le raisin » ; « cette pomme, je la lance », d'où « je lance la pomme »). C'est cette étape d'une référence enrichie (à partir de la référence « pauvre » que constitue la deixis, amorce d'une communication référentielle dans notre hypothèse inspirée de la théorie « Vocalize to Localize ») qu'il s'agira de simuler dans de futurs modèles d'interaction entre agents communicants.

Des modèles conceptuels aux modèles formels. On peut proposer l'extension des variables d'objets O_S et O_L de notre modèle d'agent vers des variables plus complexes

associant objets et actions. Le pouvoir d'expression de la modélisation bayésienne suggère alors de nombreuses perspectives. Pour donner un exemple, l'équivalent du premier terme de la décomposition π_{Ag} (Équation 8.2), $P(O_S \mid \pi_{Ag})$, pourrait alors être augmenté d'une variable d'action A pour donner :

$$P(O_S\ A \mid \pi_{Ag}) = P(O_S \mid \pi_{Ag})P(A \mid O_S\ \pi_{Ag}).$$

Le terme introduit, $P(A \mid O_S\ \pi_{Ag})$ correspond à la distribution des actions connaissant un objet, et semble bien approprié pour la modélisation de la notion d'affordance (l'invocation des actions qu'il est possible de faire à la vue d'un objet). Celle-ci est centrale dans les travaux de modélisation du système miroir (Fagg et Arbib, 1998; Bonaiuto et collab., 2007) pour rendre compte de la nécessité de la présence de l'objet dans la réponse des neurones à l'observation d'une action (les neurones miroirs ne déchargent pas lors de l'observation d'une action mimée, voir Section 3.4.2).

Des modèles formels aux simulations informatiques. L'analyse des résultats de simulations incorporant actions et objets permettra peut-être de fournir des éléments de réponse à la question : une hypothèse prélangagière de reconnaissance d'actions transitives vers des objets par le système miroir peut-elle influer la morphogénèse d'une syntaxe verbe-objet ?

14.3.3 Propositions d'extensions théoriques

Les deux propositions qui suivent sont plus personnelles. Il convient de les aborder avec précautions, peu de documentation sérieuse ayant été effectuée par nos soins sur ces thèmes.

14.3.3.1 Vers un modèle de neurosciences computationnelles

Il est admis en neuroanatomie que les représentations motrices et auditives sont stockées dans le lobe frontal et temporal, respectivement (voir Chapitre 2). Les récents progrès en neurosciences témoignent également d'importants flux d'informations entre ces deux zones par les voies ventrales et dorsales (voir Section 3.4). Ainsi, notre modèle d'agent construit autour d'une interaction entre un système moteur et un système auditif à travers un lien sensori-moteur semble cohérent avec ces données.

Sans vouloir exposer plus en détail cette architecture, nous pensons qu'il s'agit là d'une piste intéressante qui pourrait permettre d'étudier la correspondance entre les flux d'information dans le modèle et les données de neuroimagerie dans des tâches de parole.

Reste qu'une difficulté théorique non négligeable, sur laquelle nous n'avons que peu réfléchi, est la question de l'analogie entre des mesures calculatoires dans des processus d'inférence bayésienne, et des mesures physiques d'activités cérébrales localisées. Pour résumer, la question centrale est de savoir sous quelle forme neuronale est implémentée une distribution de probabilités telle que $P(M \mid O_S)$, $P(S \mid M)$ ou $P(O_L \mid S)$.

L'enjeu pour la suite de ces recherches sera de s'attaquer à l'explicitation des localisations neuroanatomiques. De nombreuses propositions détaillées et argumentées sont disponibles pour cela (avec notamment les modèles de Guenther (2006) pour la production ; et de Skipper Skipper et collab. (2007) pour la perception).

14.3.3.2 Vers un modèle de curiosité artificielle

La dernière perspective que nous souhaitons proposer s'éloigne du cadre strict de la parole pour aborder la curiosité artificielle (Oudeyer et Kaplan, 2006), thème de recherche émergeant d'une nouvelle approche de la robotique : la robotique développementale.

Si l'on définit la curiosité comme une tendance comportementale à explorer un espace sensori-moteur vers la découverte de zones stables (pour lesquelles les conséquences sensorielles des gestes moteurs deviennent prévisibles, mais ne l'étaient pas auparavant) et bien différenciées (c'est-à-dire en préférant des zones éloignées de celles que l'on connaît déjà), alors l'analogie avec l'aspect quantique et les tendances dispersives de la morphogénèse du langage est tentante (il convient toutefois de considérer cette définition de la curiosité avec précautions).

En effet, la curiosité incite un agent sensori-moteur à explorer son espace sensori-moteur à la recherche de comportements informatifs, donc différenciés et stables. Une piste de recherche qui nous semble intéressante consisterait à évaluer jusqu'à quel point ces comportements s'organisent, de manière similaire aux comportements communicatifs que nous avons étudiés dans cette thèse, mais dans des instanciations sensori-motrices plus générales.

14.4 Publications

Chapitres d'ouvrage

Moulin-Frier, C., Schwartz, J., Diard, J., and Bessière, P. (2011b). *Primate communication and human language : Vocalisations, gestures, imitation and deixis in humans and non-humans*, chapter Emergence of articulatory-acoustic systems from deictic interaction games in a "Vocalize to Localize" framework. Advances in Interaction Studies' series by John Benjamins Pub. Co.

Publications en cours en revues internationales

Moulin-Frier, C. and Arbib, M. A. (2010, submitted). Recognizing speech in a novel accent : The motor theory of speech perception reframed.

Moulin-Frier, C., Laurent, R., Bessière, P., Schwartz, J., and Diard, J. (2010, in revision). Adverse conditions improve distinguishability of auditory, motor and perceptuo-motor theories of speech perception : an exploratory bayesian modeling study.

Conférences internationales à comité de lecture

Moulin-Frier, C., Laurent, R., Bessière, P., Schwartz, J., and Diard, J. (2011a). Noise and inter-speaker variability improve distinguishability of auditory, motor and perceptuo-motor theories of speech perception : An exploratory bayesian modeling study. In *9th International Seminar on Speech Production, ISSP'11*, Montreal, Canada.

Moulin-Frier, C., Schwartz, J., Diard, J., and Bessière, P. (2010). A unified theoretical bayesian model of speech communication. In *1st conference on Applied Digital Human Modeling, Miami, USA*.

Arbib, M. A. and Moulin-Frier, C. (2010). Recognizing speech in a novel accent : The motor theory of speech perception reframed. In *Neurobiology of Language Conference, San Diego, USA*.

Moulin-Frier, C., Schwartz, J., Diard, J., and Bessière, P. (2008b). Emergence of a language through deictic games within a society of sensori-motor agents in interaction. In *8th International Seminar on Speech Production, ISSP'08*, Strasbourg France.

Moulin-Frier, C., Schwartz, J., Diard, J., and Bessière, P. (2008a). Emergence du langage par jeux déictiques dans une société d'agents sensori-moteurs en interaction. In *27e Journées d'Etudes sur la Parole, JEP'2008*, Avignon France.

Workshops

Moulin-Frier, C., Schwartz, J., Diard, J., and Bessière, P. (2008c). Emergence of a language through deictic games within a society of sensori-motor agents in interaction. In *International Workshop on "Speech and Face to Face Communication"*, Grenoble France.

Schwartz, J., Rochet-Capellan, A., and Moulin-Frier, C. (2007). Speech at reach of hand and mouth : Theoretical arguments, experimental facts and computational advances. In *Workshop " Vocoid – Vocalization, COmmmunication, Imitation and Deixis in adult and infant human and non human primates "*, Grenoble France.

Bibliographie

Abramson, A. et L. Lisker. 1970, «Discrimination along the voicing continuum : cross language tests», dans *6th Int Congr of Phonetic Science. Prague : Academia*, p. 569–573.

Abry, C. 2003, «[b]-[d]-[g] as a universal triangle as acoustically optimal as [i]-[a]-[u]», dans *Proceedings of the 15th international congress of phonetic sciences, Barcelona*, p. 727—-730.

Abry, C., A. Vilain et J.-L. Schwartz. 2004, «Vocalize to localize ? a call for better crosstalk between auditory and visual communication systems researchers», *Interaction Studies : social behaviour and communication in biological and artificial systems*, vol. 5(3), p. 313–325.

Arbib, M. 2009, «Invention and community in the emergence of language : A perspective from new sign languages», *Foundations in evolutionary cognitive neuroscience : Introduction to the discipline*, p. 117–52.

Arbib, M. A. 2005a, «From monkey-like action recognition to human language : an evolutionary framework for neurolinguistics», *Behavioral and Brain Sciences*, vol. 28, p. 105—167.

Arbib, M. A. 2005b, «Interweaving protosign and protospeech : Further developments beyond the mirror», *Interaction Studies*, vol. 6, doi :10.1075/is.6.2.02arb, p. 145–171.

Atkinson, Q. D. 2011, «Phonemic diversity supports a serial founder effect model of language expansion from africa», *Science*, vol. 332, n° 6027, doi :10.1126/science.1199295, p. 346 –349.

Baron-Cohen, S., A. Leslie et U. Frith. 1985, «Does the autistic child have a "theory of mind" ?», *Cognition*, vol. 21, n° 1, p. 37–46, ISSN 0010-0277.

Berrah, A. 1998, *Evolution d'une société artificielle d'agents de parole : un modéle pour l'émergence des structures phonétiques*, thèse de doctorat, Institut National Polytechnique de Grenoble.

Berrah, A. et R. Laboissière. 1999, «SPECIES : an evolutionary model for the emergence of phonetic structures in an artificial society of speech agents», dans *Advances in Artificial Life*, p. 674–678.

Bessière, P., E. Dedieu, O. Lebeltel, E. Mazer et K. Mekhnacha. 1998, «Interprétation ou Description (II) : Fondements mathématiques de l'approche F+D», *Intellectica*, vol. 26-27, p. p. 313–336.

Bessière, P., E. Dedieu, O. Lebeltel, E. Mazer et K. Mekhnacha. 1999, «Interprétation versus Description (I) : Proposition pour une théorie probabiliste des systèmes cognitifs sensori-moteurs», *Intellectica*, vol. 26-27, p. pp 257–311.

de Boer, B. 2000, «Self-organization in vowel systems», *Journal of Phonetics*, vol. 28, n° 4, doi :10.1006/jpho.2000.0125, p. 441–465, ISSN 0095-4470.

Bohland, J. W. et F. H. Guenther. 2006, «An fMRI investigation of syllable sequence production», *NeuroImage*, vol. 32, n° 2, doi :10.1016/j.neuroimage.2006.04.173, p. 821–841, ISSN 1053-8119.

Bonaiuto, J., E. Rosta et M. Arbib. 2007, «Extending the mirror neuron system model, i : Audible actions and invisible grasps», *Biological Cybernetics*, vol. 96, doi :10.1007/s00422-006-0110-8, p. 9–38, ISSN 0340-1200. ACM ID : 1229746.

Boë, L. 1999, «Vowel spaces of newly-born infants and adults consequences for ontogenesis and phylogenesis», dans *14th Internation Congress of Phonetic Sciences*, p. 2501–2504.

Boë, L., J. Schwartz, J. Granat, J. Heim, A. Serrurier, P. Badin, G. Captier, P. B. J. Schwartz, L. Boë, P. Badin et T. R. Sawallis. 2011, «L'émergence de la parole : aspects historiques et épistémologiques d'une nouvelle réarticulation», *Faits de langue (à paraître)*.

Boë, L., N. Vallée, P. Badin, J. Schwartz et C. Abry. 2000, «Tendencies in phonological structures : the influence of substance on form», *Bulletin de la Communication Parlée*, vol. 5, p. 35–55.

Browman, C. P. et L. Goldstein. 1989, «Articulatory gestures as phonological units», *Phonology*, vol. 6, n° 02, doi :10.1017/S0952675700001019, p. 201–251.

Browman, C. P. et L. Goldstein. 1992, «Articulatory phonology : an overview», *Phonetica*, vol. 49, n° 3-4, p. 155–180, ISSN 0031-8388. PMID : 1488456.

Browman, C. P. et L. M. Goldstein. 1986, «Towards an articulatory phonology», *Phonology Yearbook*, vol. 3, p. 219–252, ISSN 02658062. ArticleType : primary_article / Full publication date : 1986 / Copyright © 1986 Cambridge University Press.

Cheney, D. L. et R. M. Seyfarth. 1982, «How vervet monkeys perceive their grunts : Field playback experiments», *Animal Behaviour*, vol. 30, n° 3, doi :10.1016/S0003-3472(82) 80146-2, p. 739–751, ISSN 0003-3472.

Cheney, D. L. et R. M. Seyfarth. 1992, *How Monkeys See the World : Inside the Mind of Another Species*, University of Chicago Press, ISBN 9780226102467.

Chistovich, L. 1980, «Auditory Processing of Speech.», *Language and Speech*, vol. 23, n° 1, p. 67–73.

Chomsky, N. 1985, *Règles et représentations. .*

Christiansen, M. H. et S. Kirby. 2003, «Language evolution : Consensus and controversies», *TRENDS IN COGNITIVE SCIENCES*, vol. 7, p. 300—307.

Clements, N. 2003a, «Feature economy as a phonological universal», dans *Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona*, vol. [CD ROM], p. 371–374.

Clements, N. 2003b, «Feature economy in sound systems», *Phonology*, vol. 3, p. 287–333.

Dempster, A. P., N. M. Laird et D. B. Rubin. 1977, «Maximum likelihood from incomplete data via the EM algorithm», *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, n° 1, p. 1–38, ISSN 00359246. ArticleType : research-article / Full publication date : 1977 / Copyright © 1977 Royal Statistical Society.

Diard, J. 2003, *La carte bayésienne : un modèle probabiliste hiérarchique pour la navigation en robotique mobile*, thèse de doctorat, Institut National Polytechnique de Grenoble - INPG.

Diehl, R. L., A. J. Lotto et L. L. Holt. 2004, «Speech perception», *Annual Review of Psychology*, vol. 55, doi :10.1146/annurev.psych.55.090902.142028, p. 149–179, ISSN 0066-4308. PMID : 14744213.

Ducey-Kaufmann, V. 2007, *Le cadre de la parole et le cadre du signe : un rendez-vous développemental*, thèse de doctorat, Université Stendhal - Grenoble III. Département Parole et Cognition.

Fadiga, L., L. Craighero, G. Buccino et G. Rizzolatti. 2002, «Speech listening specifically modulates the excitability of tongue muscles : a TMS study», *European Journal of Neuroscience*, vol. 15, n° 2, p. 399–402.

Fadiga, L., L. Fogassi, G. Pavesi et G. Rizzolatti. 1995, «Motor facilitation during action observation : a magnetic stimulation study», *J Neurophysiol*, vol. 73, n° 6, p. 2608–2611.

Fagg, A. H. et M. A. Arbib. 1998, «Modeling parietal-premotor interactions in primate control of grasping», *Neural Networks*, vol. 11, n° 7-8, doi :10.1016/S0893-6080(98) 00047-1, p. 1277–1303, ISSN 0893-6080.

Frith, C. D. 1995, *The cognitive neuropsychology of schizophrenia*, Psychology Press, ISBN 9780863773341, 188 p..

Galantucci, B. 2005, «An experimental study of the emergence of human communication», *Cognitive Science*, vol. 29, p. 737–767.

Galantucci, B., C. A. Fowler et M. T. Turvey. 2006, «The motor theory of speech perception reviewed», *Psychonomic Bulletin & Review*, vol. 13, n° 3, doi :VL-13, p. 361–377.

Grabski, K., L. Lamalle, J. Schwartz, C. Vilain, N. Vallée, I. Tropres, M. Baciu, J.-F. L. Bas et M. Sato. 2010, «Corrélats neuroanatomiques des systèmes de perception et de production des voyelles du français», dans *28e Journées d'Etudes sur la Parole, JEP'2010*, Avignon France.

Grezes, J. 1998, «Top down effect of strategy on the perception of human biological motion : A PET investigation», *Cognitive Neuropsychology*, vol. 15, n° 6, p. 553–582.

Griffiths, T. L. et M. L. Kalish. 2005, «A bayesian view of language evolution by iterated learning», dans *27th Annual Conference of the Cognitive Science Society*.

Grèzes, J., J. L. Armony, J. Rowe et R. E. Passingham. 2003, «Activations related to "mirror" and "canonical" neurones in the human brain : an fMRI study», *NeuroImage*, vol. 18, n° 4, doi :10.1016/S1053-8119(03)00042-9, p. 928–937, ISSN 1053-8119.

Guenther, F. H. 1995, «Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production.», *Psychological Review*, vol. 102, n° 3, doi : 10.1037/0033-295X.102.3.594, p. 594–621, ISSN 0033-295X.

Guenther, F. H. 2006, «Cortical interactions underlying the production of speech sounds», *Journal of Communication Disorders*, vol. 39, n° 5, doi :10.1016/j.jcomdis.2006.06.013, p. 350–365, ISSN 0021-9924.

Guenther, F. H., M. Hampson et D. Johnson. 1998, «A theoretical investigation of reference frames for the planning of speech movements», *Psychological Review*, vol. 105, n° 4, p. 611–633, ISSN 0033-295X. PMID : 9830375.

Hickok, G. et D. Poeppel. 2004, «Dorsal and ventral streams : a framework for understanding aspects of the functional anatomy of language», *Cognition*, vol. 92, n° 1-2, doi :10.1016/j.cognition.2003.10.011, p. 67–99, ISSN 0010-0277. PMID : 15037127.

Hickok, G. et D. Poeppel. 2007, «The cortical organization of speech processing», *Nat Rev Neurosci*, vol. 8, n° 5, doi :10.1038/nrn2113, p. 393–402, ISSN 1471-003X.

Holst, E. 1954, «Relations between the central nervous system and the peripheral organs», *British Journal of Animal Behaviour*, vol. 2, p. 89–94.

Iacoboni, M., R. P. Woods, M. Brass, H. Bekkering, J. C. Mazziotta et G. Rizzolatti. 1999, «Cortical mechanisms of human imitation», *Science*, vol. 286, n° 5449, doi :10.1126/science.286.5449.2526, p. 2526–2528.

Jaynes, E. T. 2003, *Probability Theory : The Logic of Science*, Cambridge University Press.

Kaplan, F. 2000, «Semiotic schemata : Selection units for linguistic cultural evolution», dans *Artificial life VII : proceedings of the seventh International Conference on Artificial Life*, The MIT Press, p. 372.

Kaplan, F. 2005, «Simple models of distributed co-ordination», *Connection Science*, vol. 17, n° 3, p. 249–270.

Kent, R. D. 1997, *The Speech Sciences*, Singular Publishing Group.

Kohler, E., C. Keysers, M. A. Umiltà, L. Fogassi, V. Gallese et G. Rizzolatti. 2002, «Hearing sounds, understanding actions : Action representation in mirror neurons», *Science*, vol. 297, n° 5582, doi :10.1126/science.1070311, p. 846 –848.

Kohonen, T. 1990, «The self-organizing map», *Proceedings of the IEEE*, vol. 78, n° 9, p. 1464–1480.

Kuhl, P. K. et D. M. Padden. 1982, «Enhanced discriminability at the phonetic boundaries for the voicing feature in macaques», *Perception & Psychophysics*, vol. 32, n° 6, p. 542–550, ISSN 0031-5117. PMID : 7167352.

Lebeltel, O., P. Bessiere, J. Diard et E. Mazer. 2004, «Bayesian robot programming», *Autonomous Robots*, vol. 16, p. 49–79. See http ://www.springerlink.com/content/lh76585657n21244/ See http ://emotion.inrialpes.fr/bibemotion/2004/LBDM04/.

Liberman, A. M. 1993, «Some assumptions about speech and how they changed», *Haskins Laboratories Status Report on Speech Research*, vol. 113, p. 1–32.

Liberman, A. M. et I. G. Mattingly. 1985, «The motor theory of speech perception revised», *Cognition*, vol. 21, n° 1, p. 1–36, ISSN 0010-0277. PMID : 4075760.

Liljencrants, J. et B. Lindblom. 1972, «Numerical simulation of vowel quality systems : The role of perceptual contrast», *Language*, vol. 48, n° 4, p. 839–862, ISSN 00978507. ArticleType : primary_article / Full publication date : Dec., 1972 / Copyright © 1972 Linguistic Society of America.

Lindblom, B. 1984, «Can the models of evolutionary biology be applied to phonetic problems», dans *Proceedings of the tenth international congress of phonetic sciences*, Foris Pubns USA, p. 67–81.

Lindblom, B. 1990, «On the notion of 'possible speech sound'», *Journal of phonetics*, vol. 18, n° 2, p. 135–152.

Lindblom, B. 1998, «Systemic constraints and adaptive change in the formation of sound structure», dans *Approaches to the Evolution of Language : Social and Cognitive Bases*, édité par J. R. Hurford, M. Studdert-Kennedy et K. C., Cambridge University Press, Cambridge.

MacNeilage, P. F. 1998, «The frame/content theory of evolution of speech production», *Behavioral and Brain Sciences*, vol. 21, p. 499–511.

MacNeilage, P. F. et B. L. Davis. 2000, «On the origin of internal structure of word forms», *Science*, vol. 288, doi :10.1126/science.288.5465.527, p. 527–531.

Maddieson. 1984, *Patterns of Sounds*, Cambridge University Press, ISBN 0521265363.

Maddieson, I. et K. Precoda. 1989, «Updating UPSID», *The Journal of the Acoustical Society of America*, vol. 86, n° S1, doi :10.1121/1.2027403, p. S19.

Maeda, S. 1989, «Compensatory articulation during speech : Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model», *Speech production and speech modelling*, p. 131–149.

McGurk, H. et J. MacDonald. 1976, «Hearing lips and seeing voices», *Nature*, vol. 264, n° 5588, doi :10.1038/264746a0, p. 746–748.

Moineau, S., N. F. Dronkers et E. Bates. 2005, «Exploring the processing continuum of Single-Word comprehension in aphasia», *J Speech Lang Hear Res*, vol. 48, n° 4, doi : 10.1044/1092-4388(2005/061), p. 884–896.

Moulin-Frier, C., J. Schwartz, J. Diard et P. Bessière. 2008, «Emergence of a language through deictic games within a society of sensori-motor agents in interaction», dans *8th International Seminar on Speech Production, ISSP'08*, Strasbourg France. Département Parole et Cognition.

Moulin-Frier, C., J. Schwartz, J. Diard et P. Bessière. 2011, *Primate communication and human language : Vocalisations, gestures, imitation and deixis in humans and non-humans*, chap. Emergence of articulatory-acoustic systems from deictic interaction games in a "Vocalize to Localize" framework, Advances in Interaction Studies' series by John Benjamins Pub. Co.

Ménard, L. 2002, *Production et perception des voyelles au cours de la croissance du conduit vocal : variabilité, invariance et normalisation*, thèse de doctorat, Université Stendhal, Grenoble.

Ménard, L., J. Schwartz et J. Aubin. 2008, «Invariance and variability in the production of the height feature in french vowels», *Speech Communication*, vol. 50, p. 14–28.

Ohala, J. 1979, «Moderator's introduction to symposium on phonetic universals in phonological systems and their explanation», dans *Proceedings of the 9th International Congress of Phonetic Sciences*, vol. 3, p. 181–185.

Oudeyer, P. 2003, *L' auto-organisation de la parole*, Thèse doctorat, Université Pierre et Marie Curie (Paris), [S.l.].

Oudeyer, P. 2004, «Aux sources du langage : l'auto-organisation de la parole», *Cahiers Romans de Sciences Cognitives, In Cognito*, vol. 2(2), p. 1–24.

Oudeyer, P. 2005, «The self-organization of speech sounds», *Journal of Theoretical Biology*, vol. 233, n° 3, doi :10.1016/j.jtbi.2004.10.025, p. 435–449, ISSN 0022-5193.

Oudeyer, P. et F. Kaplan. 2006, «Discovering communication», *Connection Science*, vol. 18, n° 2, doi :10.1080/09540090600768567, p. 189, ISSN 0954-0091.

Penfield, O., M. CMG et M. Jasper. 1954, «Epilepsy and the functional anatomy of the human brain.», *JAMA*, vol. 155, n° 1, p. 86.

Pradalier, C., F. Colas et P. Bessière. 2003, «Expressing bayesian fusion as a product of distributions : Applications in robotics», dans *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS03)*, vol. 2, p. 1851–1856.

Pulvermüller, F. et L. Fadiga. 2010, «Active perception : sensorimotor circuits as a cortical basis for language», *Nature Reviews. Neuroscience*, vol. 11, n° 5, doi :10.1038/nrn2811, p. 351–360, ISSN 1471-0048. PMID : 20383203.

Pulvermüller, F., M. Huss, F. Kherif, F. M. del Prado Martin, O. Hauk et Y. Shtyrov. 2006, «Motor cortex maps articulatory features of speech sounds», *Proceedings of the National Academy of Sciences*, vol. 103, n° 20, doi :10.1073/pnas.0509989103, p. 7865 –7870.

Redford, M. A., C. C. Chen et R. Miikkulainen. 2001, «Constrained emergence of universals and variation in syllable systems», *Language and Speech*, vol. 44, p. 27–56.

Rizzolatti, G. et M. A. Arbib. 1998, «Language within our grasp», *Trends in Neurosciences*, vol. 21, n° 5, p. 188–194, ISSN 0166-2236. PMID : 9610880.

Rizzolatti, G., L. Fadiga, V. Gallese et L. Fogassi. 1996, «Premotor cortex and the recognition of motor actions», *Brain Research. Cognitive Brain Research*, vol. 3, n° 2, p. 131–141, ISSN 0926-6410. PMID : 8713554.

Amélie Rochet-Capellan, Jean-Luc Schwartz, Rafael Laboissière et Arturo Galvan. 2007, «Two CV syllables for one pointing gesture as an optimal ratio for jaw-arm coordination in a deictic task : A preliminary study», dans *Proceedings of the 2nd European Cognitive Science Conference, EuroCogSci07 2nd European Cognitive Science Conference, EuroCogSci07*, 23-27 May 2007, Delphi, Greece, S. Vosniadou, D. Kayser, & A. Protopapas, Delphi Greece, p. 608–613.

Romand, R. 2000, *La parole*, chap. Introduction au fonctionnement du système auditif, Hermès, p. 101–133.

Rousset, I. 2004, *Structures Syllabiques et Lexicales des Langues du Monde : Typologies, Tendances Universelles et Contraintes Substancielle*, thèse de doctorat, Université Stendhal, Grenoble.

Roy, A. C. et M. A. Arbib. 2005, «The syntactic motor system», *Gesture*, vol. 5, n° 1, doi :10.1075/gest.5.1.03roy, p. 7–37, ISSN 15681475.

Ruhlen, M. 1996, *The Origin of Language : Tracing the Evolution of the Mother Tongue*, New York : John Wiley & Sons, ISBN 0471159638.

Savariaux, C., P. Perrier, J. P. Orliaguet et J. L. Schwartz. 1999, «Compensation strategies for the perturbation of french [u] using a lip tube. II. perceptual analysis», *The Journal of the Acoustical Society of America*, vol. 106, n° 1, p. 381–393, ISSN 0001-4966. PMID : 10420629.

Schroeder, M., B. Atal et J. Hall. 1979, *Frontiers of Speech Communication Research*, chap. Objective measure of certain speech signal degradations based on masking properties of human auditory perception, London Academic Press, p. 217–229.

Schwartz, J., A. Basirat, L. Ménard et M. Sato. 2010, «The Perception-for-Action-Control theory (PACT) : a perceptuo-motor theory of speech perception», *Journal of Neurolinguistics*, vol. In Press, doi :10.1016/j.jneuroling.2009.12.004, ISSN 0911-6044.

Schwartz, J., L. Boë, N. Vallée et C. Abry. 1997a, «Major trends in vowel system inventories», *Journal of Phonetics*, vol. 25, n° 3, doi :10.1006/jpho.1997.0044, p. 233–253, ISSN 0095-4470.

Schwartz, J., L. Boë, P. Badin et T. R. Sawallis. 2011, en révision, «Grounding stop place features in the perceptuo-motor substance of speech communication», .

Schwartz, J., L. Boë, N. Vallée et C. Abry. 1997b, «The Dispersion-Focalization theory of vowel systems», *Journal of Phonetics*, vol. 25, n° 3, doi :10.1006/jpho.1997.0043, p. 255–286, ISSN 0095-4470.

Schwartz, J. L., C. Abry, L. J. Bo\ë et M. Cathiard. 2002, «Phonology in a theory of perception-for-action-control», *Phonetics, phonology and cognition*, p. 244–280.

Schwartz, J.-L., L.-J. Boë et C. Abry. 2007, «Linking the Dispersion-Focalization Theory (DFT) and the Maximum Utilization of the Available Distinctive Features (MUAF) principle in a Perception-for-Action-Control Theory (PACT)», dans *Experimental Approaches to Phonology*, édité par M. O. M.J. Solé, P.S. Beddor, Oxford University Press, p. 104–124.

Seeley, R. R., T. D. Stephens et P. Tate. 2006, *Essentials of Anatomy & Physiology*, 6e éd., McGraw-Hill Science/Engineering/Math, ISBN 0073228052.

Serkhane, J. 2005, *Un bébé androïde vocalisant : Etude et modélisation des mécanismes d'exploration vocale et d'imitation orofaciale dans le developpement de la parole*, thèse de doctorat, Institut Polytechnique de Grenoble.

Serkhane, J., J.-L. Schwartz et P. Bessiere. 2005, «Building a talking baby robot : A contribution to the study of speech acquisition and evolution», *Interaction Studies*, vol. 6, no 2, p. 253–286, ISSN 1572-0373.

Skipper, J. I., V. van Wassenhove, H. C. Nusbaum et S. L. Small. 2007, «Hearing lips and seeing voices : How cortical areas supporting speech production mediate audiovisual speech perception», *Cereb. Cortex*, doi :10.1093/cercor/bhl147, p. bhl147.

Steels, L. 1996, «Emergent adaptive lexicons», dans *SAB96*, édité par P. Maes, M. Mataric, J.-A. Meyer, J. Pollack et S. W. Wilson, MIT Press, Cambridge, MA.

Steels, L. 1997, «The synthetic modeling of language origins», *Evolution of Communication*, vol. 1, no 1, p. 1–34.

Steels, L. 2006, «How to do experiments in artificial language evolution and why», dans *Proceedings of the 6th International Conference on the Evolution of Language*, p. 323–332.

Stevens, K. 1972, «The quantal nature of speech : Evidence from articulatory-acoustic data», dans *Human communication : A unified view*, édité par E. David et P. Denes, McGraw-Hill, p. 51–66.

Stevens, K. 1989, «On the quantal nature of speech», *Journal of phonetics*, vol. 17, no 1, p. 3–45.

Studdert-Kennedy, M. et L. Goldstein. 2003, «Launching language : The gestural origin of discrete infinity», dans *Language Evolution : The States of the Art*, édité par M. Christiansen et S. Kirby, Oxford University Press.

Sussman, H. M., C. Duder, E. Dalston et A. Cacciatore. 1999, «An acoustic analysis of the development of CV coarticulation : A case study», *J Speech Lang Hear Res*, vol. 42, no 5, doi : , p. 1080–1096.

Sussman, H. M., D. Fruchter, J. Hilbert et J. Sirosh. 1998, «Linear correlates in the speech signal : the orderly output constraint», *The Behavioral and Brain Sciences*, vol. 21, no 2, p. 241–259 ; discussion 260–299, ISSN 0140-525X. PMID : 10097014.

Tomasello, M., M. Carpenter, J. Call, T. Behne et H. Moll. 2005, «Understanding and sharing intentions : the origins of cultural cognition», *The Behavioral and Brain Sciences*, vol. 28, no 5, doi :10.1017/S0140525X05000129, p. 675–691 ; discussion 691–735, ISSN 0140-525X. PMID : 16262930.

Tourville, J. A., K. J. Reilly et F. H. Guenther. 2008, «Neural mechanisms underlying auditory feedback control of speech», *NeuroImage*, vol. 39, n° 3, doi :10.1016/j.neuroimage.2007.09.054, p. 1429–1443, ISSN 1053-8119.

Vallée, N. 1994, «Systèmes vocaliques : de la typologie aux prédictions», *Grenoble, Université Stendhal : Thèse de Doctorat en Sciences du Langage.*

Vilain, A., C. Abry, P. Badin et S. Brosda. 1999, «From idiosyncratic pure frames to variegated babbling : Evidence from articulatory modelling», dans *Proceedings of the 14th International Congress of Phonetic Sciences*, vol. 3, p. 2497–2500.

Vilain, D. A., J. Schwartz, P. C. Abry et D. J. Vauclair. 2011, *Primate Communication and Human Language : Vocalisation, gestures, imitation and deixis in humans and non-humans*, John Benjamins Publishing Company, ISBN 9027204543.

Watkins, K. E., A. P. Strafella et T. Paus. 2003, «Seeing and hearing speech excites the motor system involved in speech production», *Neuropsychologia*, vol. 41, n° 8, doi : 10.1016/S0028-3932(02)00316-0, p. 989–994, ISSN 0028-3932.

Wilson, S., A. Saygin, M. Sereno et M. Iacoboni. 2004, «Listening to speech activates motor areas involved in speech production», *Nature Neuroscience*, vol. 7, n° 7, p. 701–702.

Wolpert, D., Z. Ghahramani et M. Jordan. 1995, «An internal model for sensorimotor integration», *Science*, vol. 269, n° 5232, doi :10.1126/science.7569931, p. 1880–1882.

Wolpert, D. M. et M. Kawato. 1998, «Multiple paired forward and inverse models for motor control», *Neural Networks*, vol. 11, n° 7-8, doi :10.1016/S0893-6080(98)00066-5, p. 1317–1329, ISSN 0893-6080.

Résumé

Si la question de l'origine du langage reste d'un abord compliqué, celle de l'origine des formes du langage semble plus susceptible de se confronter à la démarche expérimentale. Malgré leur infinie variété, d'évidentes régularités y sont présentes : les universaux du langage. Nous les étudions par des raisonnements plus généraux sur l'émergence du langage, notamment sur la recherche de précurseurs onto- et phylogénétiques.

Nous abordons trois thèmes principaux : la situation de communication parlée, les architectures cognitives des agents et l'émergence des universaux du langage dans des sociétés d'agents.

Notre première contribution est un modèle conceptuel des agents communicants en interaction, issu de notre analyse bibliographique. Nous en proposons ensuite une formalisation mathématique Bayésienne : le modèle d'un agent est une distribution de probabilités, et la production et la perception sont des inférences bayésiennes. Cela permet la comparaison formelle des différents courants théoriques en perception et en production de la parole. Enfin, nos simulations informatiques de société d'agents identifient les conditions qui favorisent l'apparition des universaux du langage.

Abstract

If the origin of language is difficult to properly study, the origin of its forms appears to be accessible to the experimental method. Languages, despite their large variety, display obvious regularities, the linguistic universals. We study them through more general reasoning about language emergence, in particular in the search of its precursors, both in ontogeny and phylogeny.

We study three main themes : the communication situation, the agent's cognitive architectures and the emergence of linguistic universals in agent societies.

Our first contribution is a conceptual model of communicating agents in interaction, emanating from our bibliographic survey. We then cast it into the Bayesian mathematical formalism : an agent model is a probability distribution, and production and perception are defined by Bayesian inference. This allows a formal comparison of speech perception and production theoretical trends. Finally, computer simulations of agent societies help identify the conditions that favor the appearance of linguistic universals.